Selecting Good Redistricting Plans from a Large Pool of Available Plans Using the Efficient Frontier

Ram Gopalan
Rutgers University, Camden
ram.gopalan@rutgers.edu

Lee Hachadoorian
Temple University
lee.hachadoorian@temple.edu

Steven O. Kimbrough
University of Pennsylvania
kimbrough@wharton.upenn.edu

Frederic H. Murphy
Temple University
fmurphy@temple.edu

Abstract

As part of a widespread frustration with partisan gerrymandering, many states have considered or implemented redistricting reforms—and others will eventually have to—that include a higher degree of citizen participation in proposing and evaluating redistricting plans. In some states without redistricting reform, public interest groups have created shadow commissions that encourage citizens to submit their own maps. For example, the new map for Pennsylvania Congressional districts, chosen by the state Supreme Court, was proposed by a citizens group.

As citizen participation grows, analytical methods for rating plans that recognize the different mapping criteria are needed to sort through multiple maps, both for highlighting good maps and for providing measures that allow courts to rule that a map is gerrymandered. Using a modified version of a model called *data envelopment analysis* (DEA), we present a nonpartisan approach that can score maps while not imposing any prior weights on the criteria. Our modification measures how close a plan is to the convex hull of the Pareto frontier when bigger is better for some criteria and smaller is better for others. Thus, we provide a novel and scalable way to filter out poor plans from large corpora of redistricting plans.

Keywords: political redistricting, data envelopment analysis.

1 Introduction

The districting problem consists of partitioning a geographic region into territories such that the resulting plan is "good" on some dimension or dimensions. When the goal of this partitioning is to create districts for the purpose of political representation, the process is usually known as " political redistricting," or "redistricting" for short, as a former plan of electoral districts is changed to account for population growth and migration. The United Kingdom and its former colonies typically use representation by single-member geographic districts, while many other democracies sidestep the problem of redistricting through systems of proportional representation. However, the Commonwealth countries (e.g., New Zealand, India) have all taken redistricting out of the political arena by adopting independent nonpartisan redistricting commissions. The United States stands alone in its process, heavily influenced by layers of politics, legislation, and evolving jurisprudence, and varying widely between states (Butler and Cain, 1985).

American redistricting is guided by several traditional redistricting criteria, including contiguity, compactness, population equality, racial equity, the avoidance of splitting administrative divisions such as school districts, counties, and townships (which this paper will refer to generically as "political units" or "units"), and the maintenance of communities of interest (Webster, 2013). Newer criteria, such as competitiveness and partisan bias, can also be used. Gerrymandering refers to using the redistricting process for the purpose of partisan advantage, that is, ignoring these criteria (often while pretending to honor them) and subverting the goal of equal representation.

Very little about American redistricting is prescribed in the U.S. Constitution. Even districting into single-member geographic territories is not required, and in the first several decades of U.S. history several states elected their Congressional representatives at large (Engstrom, 2016). Furthermore, even the "one person one vote" rule was not established until a series of U.S. Supreme Court decisions in the 1960s, first for state legislatures and only then for the House of Representatives. On the eve of the "redistricting revolution", state legislative chambers were severely malapportioned, in the worst cases with population ratios of hundreds to one from the largest representative district to the smallest (Bullock, 2010). Population ratios in the House of Representatives were also bad by today's standards, though not as extreme as in the state legislatures. Throughout, naked political manipulation of the redistricting process has been common.

In a majority of states, the legislature fully controls the redistricting process. This has meant that they have strong incentives to design districts that keep themselves in office. This could mean the majority party protecting its majority, but it can also mean both parties collaborating to protect their incumbent seats. In the states where the legislature does the redistricting, a legislative committee, typically consisting of the most senior legislators, draws the districts. They then pass a bill designating the new districts. If the governor is from the same party as the majority in the legislature, the bill is signed into law. If of the opposing party, the governor either vetoes the bill or, having already negotiated a deal with the legislature, signs it into law.

Several states, typically through citizen referenda, have changed the rules to have a nonpartisan citizen committee designing the map. This has lead to maps that better represent the citizens of those states. The process of drawing the maps takes several months and generally includes public hearings, comment periods, and redrawing.

The core problem in demonstrating gerrymandering in the redistricting of political districts is that there are no simple measures that can prove a redistricting plan (map) is gerrymandered (Grofman (2019)). Since *Davis v. Bandemer* (1986)[1] the Supreme Court of the United States has failed to reject allegedly gerrymandered maps for lack of a "judicially discernable and manageable standard." Attempts to create a single number that could be a judicially manageable standard, such as the efficiency gap, died in *Gill v. Whitford* (2018)[2] and *Benisek v. Lamone* (2018).[3] *Rucho v. Common Cause* (2019)[4] went further, as the majority ruled that partisan gerrymandering claims were "non-justiciable"–that is, not subject to judicial review–in *federal* courts, while allowing that state legislatures and courts were better equipped to handle such claims. Going further, the problem is actually two problems: No single measure can capture all of the criteria used to evaluate redistricting plans without imposing value judgments on the relative importance of the criteria; and for any given criterion no single cutoff can be established nationally as a bright line indicating that a map is gerrymandered.

Consider two traditional redistricting criteria, population deviation and split political subdivisions. Population equality (low to no deviation in the range of

[1]Davis v. Bandemer, 478 U.S. 109 (1986). https://www.oyez.org/cases/1985/84-1244

[2]Gill v. Whitford, 585 U.S. ___ (2018). https://www.oyez.org/cases/2017/16-1161

[3]Benisek v. Lamone, 585 U.S. ___ (2018). https://www.oyez.org/cases/2017/17-333

[4]Rucho v. Common Cause, 588 U.S. ___ (2019). https://www.oyez.org/cases/2018/18-422

district populations) is required by the jurisprudence of one person, one vote. Minimizing the number of times political units are divided, termed splits, is important for local governments (such as municipalities and counties) to be represented at the state and national level. But these two criteria are in tension with each other, and with other criteria. In their 2011 map for Pennsylvania, Republicans used a map that was good only on population deviation as cover for a highly noncompact and arguably extremely gerrymandered map. Loose population targets make it easier to avoid splits, but abet other forms of gerrymandering or incumbent advantage. For example, as population shifts, district lines can be kept intact (altered minimally or not at all) during a redistricting round in order to advantage districts and incumbents in areas losing population. During the rapid urbanization of the early 20th century, before one person, one vote was established, this was done to advantage rural areas that were losing population to urban areas. More recently, and under a requirement of a maximum 10% population deviation in state and local legislative districts, this was done in Philadelphia in the 2022 redistricting round to lessen the voice of the growing Hispanic population on City Council.

For many criteria, no single number can be used as a cutoff or standard nationally. Consider compactness, which, since the original "Gerry-mander", has been used as an eyeball test of gerrymandering. First, the geometries of states vary wildly. Kansas is almost rectangular while Maryland has a long western neck extending into the Appalachians, and a larger eastern portion that is split by a body of water (Chesapeake Bay). It is quite easy to construct highly compact districts in Kansas, which can nonetheless be drawn for partisan advantage. Mathematically, Maryland can have a map with more partisan fairness (less gerrymandered) that is less compact than Kansas, or less than some nationally applied compactness threshold. Second, an uneven distribution of party affiliation makes it hard or impossible to achieve partisan fairness with a rigid adherence to the compactness criterion. In many states, urban areas have Democratic supermajorities (80% or 90% Democratic share is not uncommon). Rural areas trend Republican, but at lower ratios (generally only achieving 60% to 70% vote share). Highly compact districts tend to dilute Democratic voting power by collecting Democrats into "self-packed" urban districts. In this regard, minimizing splits can have the same effect as compactness by keeping heavily Democratic urban municipalities whole.

Choosing a good map is a multiple criteria decision problem. The chosen map needs to balance the redistricting criteria, and the nature of the balance—and even what the criteria are—is a judgment call. Since the number of possible maps is vast, it is impossible to generate all maps. A given map can only be compared to a finite number of alternatives sampled from this vast space.

Thus, no single measure or an index of measures constructed with a fixed weights should be used to compare plans. Ultimately, the final map should be chosen using human judgment, trading off relevant criteria and including non-quantitative judgments.

Recently, algorithms have been developed that generate large numbers of legally valid plans with the intention of showing the extent to which enacted plans are outliers relative to the population of possible plans. The courts have yet to use this approach as a valid statistical test of the extent to which a redistricting plan is an outlier.

Given the need to sort through large numbers of maps and the importance of not imposing a priori weights on criteria, we need a tool that can extract a subset of maps for human evaluation that rates maps using the weights that give the maps their best possible rating. That is, we want maps that are on or near the Pareto frontier defined over agreed upon criteria. We show how data envelopment analysis (DEA) can be used to find a collection of plans on and near the Pareto frontier.

At this time we know of no other method that can sort through a large number of maps and provide a measure of proximity to the convex hull of the Pareto frontier. In fact, this last round of redistricting was the first to generate large numbers of maps because the quality of the free internet-based software, e.g. Dave's Redistricting App (https://davesredistricting.org) has improved greatly since the 2010 redistricting. This has created the opportunity for citizen engagement such as the successful project of the Committee of 70 in Pennsylvania that led to the production of over 1000 maps. Given the level of political interest and the ability to generate large numbers of random (and spam) maps, DEA has the potential to be an important tool for managing maps in the future.

Here we propose a two-step process that separates maps that should be contenders from unsuitable maps. The first step can be skipped if either there are not many maps to compare (fewer, say, than 1000) or it is acceptable for a plan to perform badly on some or all criteria.

The first step for eliminating proposed redistricting plans is to eliminate maps that are decidedly inferior on any criterion. The criteria can include standard measures for traditional redistricting criteria, e.g., population deviation, the number of split political units, and compactness, and newer measures such as competitiveness and partisan fairness. For each measure, a consensus-based cutoff on the distribution of values can be applied to determine which plans should be immediately rejected. This method is useful for rejecting plans that score highly on one dimension but are terrible in others. As an example, Gopalan et al. (2013) won a redistricting contest in Philadelphia by producing the most compact map. The authors recommended against using their own winning map because of the way it divided neighborhoods/communities and ignored geographical boundaries.

The cutoffs for the first pass cannot be applied too rigidly in that there are clear trade-offs among redistricting criteria, e.g., between minimizing population deviation and the number of split units, or the number of split units and compactness. Consequently, accepting or rejecting a large percentage of plans by using each criterion *separately* has the potential to reject interesting plans that are good compromises on conflicting criteria, but not among the best on any one criterion.

At the same time, applying minimal standards for each criterion is computationally cheap, and reduces the total number of plans for further evaluation, whether by commission staff or third parties. We incorporate minimal standards for criteria as part of our approach for producing a set of plans for further consideration. The key is to keep the minimums loose, so that the eliminated plans are highly unlikely to be of interest. The imposed criteria thresholds are a configurable aspect of our proposed approach.

In the second step we propose a method for selecting from among a very large pool of proposed plans a comparatively small number that deserve further consideration, and which includes plans that are good compromises among the criteria using data envelopment analysis (DEA).

The principal goal in this paper is to motivate, describe, and demonstrate this application of DEA, both in its practicability and in its scalability. For the sake of demonstration, we consider three important traditional redistricting criteria: population equality, split avoidance, and compactness. The proposed method, however, is criteria-neutral. Any measurable criterion of interest can be added to the consideration set, e.g., political competitiveness, core preservation, or number of minority opportunity districts.

The rest of the paper is organized as follows. §2, "Literature Review," identifies prior methods for generating pluralities of redistricting plans. These methods lead to the problem of processing very large collections of plans, an issue that this paper addresses. The section also discusses redistricting as a multiple criteria decision making (MCDM) problem. §4 describes our proposed solution methodology to the problem, which is based upon our adaptation of *data envelopment analysis* (DEA). Our intention is that this solution will be widely adopted for screening out poor redistricting plans. In sections §5 and §6 we choose our measures for evaluating plans and perform a preliminary screen using these measures. We begin with a corpus of 11,206 plans for the Commonwealth of Pennsylvania generated by a stochastic algorithm, and reduce it to 2,355 plans meriting further consideration. The plans are publicly available and were generated by a published and open-source algorithm (Haas et al., 2020). §7 studies the *distributions* of DEA scores and also the *correlations* between plan measures. §8, "Results," applies our solution method to the subset of 2,355 plans. §9 addresses future research directions. §10 concludes with a perspective on our goals in studying redistricting and a discussion of future research.

2 Literature Review

In an attempt to provide evidence of gerrymandering to the courts, researchers have developed methods for stochastically generating plans for the purpose of providing probability distributions of various measures such as compactness and counts of split political units (e.g., counties, townships). Examples include Chen and Cottrell (2016), Chen and Rodden (2015), Cho and Liu (2016), and Liu et al. (2016). The starting point for this literature is Altman and Mcdonald (2011). They developed an automated redistricting program titled BARD (Better Automated Redistricting). In their opinions up to now, the courts have not used the evidence that redistricting plans are clear outliers when compared against the probability distributions of plans. Whether or not the courts see fit to accept this argument, plans generated by computer or by citizens can be useful for providing alternatives to politically generated plans. To facilitate this, however, a methodology is needed to extract reasonably good candidate plans from the large number of plans generated.

One of our goals is to facilitate the understanding of the trade-offs that must be made in selecting a final plan. For example, there is a clear trade-off between how closely the population target must be met (population equality) and the number of splits of political units. With compactness, there are multiple measures of compactness to choose from as reviewed in Niemi et al. (1990); MacEachren (1985); Young (1988); Horn et al. (1993). Some are redundant, but some measure different aspects of compactness that could be considered as distinct criteria. Determining a good plan requires trade-offs, which is the subject of multiple attribute utility theory and multiple criteria decision making. See (Keeney, 1992; Keeney and Raiffa, 1993; Wallenius, Jyrki et al., 2008)

An egregiously gerrymandered plan can be bad by all measures, as demonstrated by the 2011 enacted plan in Pennsylvania in Cervas and Grofman (2020). They compare that plan to three alternatives created after the Pennsylvania Supreme

Court declared the 2011 enacted plan unconstitutional: the joint legislative plan proposed by Republican leaders from both chambers (but not passed by the legislature), the governor's plan, and the plan produced by a Special Master appointed by the Court. They found the 2011 enacted plan to be the worst of the four plans. However, they describe the joint legislative plan as a stealth gerrymander, as it was acceptable by compactness and splits criteria, but still a gerrymander based on four measures of partisan asymmetry.

2.1 Multiple criteria decision making

Our approach to designing districts and our solution methodology relies on the concept of the Pareto frontier: i.e., solutions that are not dominated on all objectives (or criteria or attributed) by other solutions. The multiple objective approach leaves the tradeoff decisions to the stakeholders, who can choose any solution on the Pareto frontier knowing that it is not dominated on all criteria by any available alternative solution. See Geneletti (2019), chapters 4 and 5, for an example. See Greco et al. (2016) for a thorough overview of the field of multiple criteria decision making (MCDM).

Typically, the set of solutions on or near the Pareto frontier is very much smaller than the set of all feasible solutions. This allows decision makers to focus on a smaller consideration set of solutions, knowing that feasible solutions far from the frontier are inferior on all criteria to some solution on the frontier. In the specific instance of redistricting, a smaller consideration set consisting of points on or near the frontier lowers the burden on commissions evaluating maps. This fact is key to our proposed approach.

Constructing a fixed-weight index and choosing the plans with high index values as in MCDM is problematical because different actors may assign different weights to different criteria, depending on their personal values. People living in rural areas might not want their counties split because they identify with their county of residence, and small cities and counties are concerned that splitting will make it hard for them to get the attention of any single legislator (**?**). People in large metropolitan areas might care about compactness more, because without enforcing compactness, it is possible to assemble gerrymandered districts within political units given the heterogeneity of different neighborhoods in cities and suburbs. Differences in demographics, socioeconomic status, residential geography, and other characteristics will lead individuals and groups to value criteria differently and choose different weights for the criteria. Thus, analytical methods for redistricting should find a range of points on and near the Pareto frontier.

The perceived costs of reducing the level of a specific criterion is also subject to diminishing returns. For example, in a plan that performs very well on splits (i.e., few splits) but has moderate to high population deviation, a small increase in splits would be acceptable to achieve a smaller population deviation. A plan that is broadly acceptable balances the criteria, recognizing that weights can differ among different political actors. This very real possibility, if it occurs, violates preferential independence, which is assumed in a weighted aggregation function. It constitutes a reason for taking a multiple objective approach instead of a multiple attribute approach.

The standard approach to solving MCDM problems involves choosing sets of weights and then selecting the best alternative for each set set of weights. (See https://www.mcdmsociety.org/content/software-related-mcdm-0 for a list of MCDM tools.) By varying the weights and finding the best plan associated with each set of weights, we can begin to determine the Pareto frontier. However, this approach is computationally inefficient.

2.2 Data envelopment analysis and MCDM

As an alternative to varying the weights a priori or interactively as is done in MCDM, each plan can be evaluated using the weights that make it look the *best* it can. This produces a collection of plans selected for further evaluation that effectively covers a range of weights.

The starting point for our approach to sorting plans in this way is *data envelopment analysis* (DEA). See (Cooper et al., 2007; Cook and Seiford, 2009) for textbook treatments of this well-established analysis tool. For a simple and clear introduction see Winston (2003). For a broad overview of the area see Cooper et al. (2011).

Doyle and Green (1993) and Belton and Vickers (1993) show how DEA effectively subsumes MCDM in that weights on preferences can be specified a priori or varied interactively in a DEA model or be part of the decision process, the standard DEA approach. As Doyle and Green (1993) say "DEA is flexible since it may be done with preferences articulated a priori, progressively or not at all." That is, rather than doing an ad hoc variation of weights using MCDM tools, DEA provides a methodology for a more systematic exploration of the consequences of using different weights on preferences in MCDM.

This also gives DEA an advantage over the only other application of MCDM to electoral redistricting uncovered in our literature search. de-los-Cobos-Silva et al. (2017) propose a new MCDM fuzzy numbers algorithm for selecting a single "best" plan from a universe of plans created by simulated annealing for legislative districts in Mexico. Again, Doyle and Green (1993) and Belton and Vickers (1993) show that all such MCDM methods can be replaced by the DEA approach to finding weights. Furthermore, while we agree that MCDM can be fruitfully applied to the redistricting problem, we do not think that an algorithm should select a single best plan.

2.3 DEA articles related to redistricting

Although there are standard methods for finding all Pareto points without the overhead of building a DEA model, our goal is to find maps that are near-Pareto as well as the maps on the Pareto frontier, because near-Pareto maps can be superior to Pareto maps on aspects that are not captured in the standard, quantitative measures of map quality. The DEA solutions identify thes near-Pareto maps.

A similar problem is constructing indices measuring the relative performance of different countries. Cherchye et al. (2004), Cherchye et al. (2007), show how to build performance indicators for the European Union. The problem with the standard construction of indices using fixed weights is that different countries can place different values on the individual components of an index. Their position is that the index for a country should use weights that are consistent with the values of that country. As a proxy for value-consistent weights, they use a DEA model that assigns weights that maximizes DEA scores, i.e., proximity to the convex hull of the Pareto frontier of possible index values. Our approach to valuing maps is consistent with theirs.

Lovell and Pastor (1999) show that the DEA model without inputs is the same as the Banker, Charnes and Cooper model with one input. Liu et al. (2011) provides a very thorough analysis of the properties of DEA models without inputs. They have index data without easy access to the components of the index. Thus, they could not use the model in Cherchye et al. (2004). They then expanded on Lovell and Pastor (1999) to provide a formal analysis of models without inputs, as the inputs were embedded in the index that had been calculated for the DMU's. They then examine the situation of evaluating student grades with constraints on the weights that can constrain the possible values of the weights.

Our situation corresponds to a model without inputs and no constraints on the weights. The theoretical results for their model (1) apply to our model.

Unlike our situation, all of the papers referenced so far presume that for *all* measures bigger is better, while we have mixed outputs in districting: for some measures smaller is better and for others, bigger is better.

One potential way for a mix of both bigger-is-better and smaller-is-better to be modeled is the literature on including outputs that are negative contributions, e.g. pollution, since smaller is better for pollutants and the free disposal assumption does not apply. See Murty and Russell (2018) and Murty et al. (2012).]. See also, Førsund (2018). Dakpo et al. (2016) provides a survey of the DEA and environmental modeling literature. The models presented in these environmental papers focus on the problem of simultaneously modeling inputs, useful outputs, and problematic outputs. Much of the discussion focuses on laws of conservation in physics leading to material balances that must be modeled in a production function. The analogy with production functions for our problem breaks down here because each plan is a unit and all measures are indicators of quality, not measures of scalable outputs. Furthermore, unlike pollution models, it is possible to have a measure for which larger is better that is negatively correlated with a smaller-is-better measure, which means both tend improve at the same time, unlike the situation with pollution where more goods production with a given technology implies more pollution. Consequently, we cannot use pollution models as an analogy for representing smaller-is-better measures and we need an approach that does not presume a production function with inputs positively correlated with outputs. We detail why this is the case as we formulate our model.

Another issue we face is that the Pareto frontier is not convex, which is why we refer to the convex hull of the Pareto frontier. This is due to two factors that differentiate our situation from the standard DEA model in which the DMU's are presumed to be on the efficiency frontier and the underlying production function is convex. Since the number of possible maps is astronomical and all of them cannot be generated with current technology, we cannot know if any map that is on the Pareto frontier of the ones available is actually on the Pareto frontier of all maps. Second, since no one has enumerated the space of maps for a realistic problem and because the number of maps is finite, we cannot presume convexity.

Kerstens et al. (2019) examine the issue of heterogeneous DMU's not defining a convex production function. Their example is of DMU's located in different countries with different laws that affect production costs and technologies.

Our problem is not the same as theirs. First, even the automated map generators that generate billions of maps as described above generate a miniscule fraction of potential maps. Thus, whenever a Pareto-optimal map has a low DEA score, this is an indication that likely a new map can be generated that pushes out the frontier, making it closer to convex. Second, we are using DEA to eliminate obviously inferior maps, as measured by the DEA score and retain a significant number of maps for more detailed examination. We also can check that all Pareto maps are retained by simply using a standard dominance calculation to find all Pareto maps.

2.4 Literature on fast methods for finding the Pareto frontier

Another stream of literature on finding the Pareto frontier or an approximation to it is from computer science, in particular database querying. We review a methodology termed *skyline methods*. For example, Kung et al. (1975) provide a fast search algorithm for finding the Pareto points in a finite set. Kumar et al. (2019) provide a survey of the literature on "skyline" problems, i.e., finding a quick approximation to the Pareto frontier so that companies such as Airbnb can almost instantaneously provide a list of Pareto or near Pareto choices to customers. For example, rentals with a range of prices, amenities, and proximity to a desired location can be returned by a skyline method. Xiao et al. (2002) develop a genetic algorithm for finding points on the Pareto frontier over a continuous set of feasible solutions. An example of a use for their approach is facility location for health clinics where there are multiple goals and communities. Later on we show how these methods can be used to enhance the approach described here.

We have provided the context for the uses of these different models to explain why we choose to use DEA. Redistricting is a political process that extends over months with multiple draft plans. In states where citizen commissions draw the maps, the process includes citizen participation and is more open. In states without citizen commissions, the process is adversarial with political parties seeking advantage through gerrymandering. In these states citizen groups and losing political parties providing maps to counter gerrymandered maps. The goal of this research is to improve the selection of maps to enhance open discussions where they are possible and offer alternative maps in adversarial situations.

In our case, unlike the typical situation for finding the Pareto frontier, we do not have the complete collection of possible redistricting plans because there are in general too many to enumerate, see (Altman and Mcdonald (2011)). Among the available maps, we want to know the Pareto frontier, not an approximation. The frontier provides the best performing maps, under the mapping criteria, for comparing with commission-drawn maps and court challenges.

We also want to know which Pareto points are far from the convex hull of the Pareto frontier to direct a search for more Pareto plans that fill out the frontier. In our case we envision having a set of maps on the order of a few thousand, mainly produced by human mappers, which was done in Pennsylvania during the last round of redistricting. If the maps are computer generated, the numbers would be smaller than in the MCMC literature because almost all randomly generated maps are very inferior to human-drawn maps and need to go through a repair process that moves them towards the frontier.

Another issue we face is that it may be possible for a dominated plan off of the Pareto frontier to be superior to a plan that dominates it because there are *non-quantitative* aspects of the dominated plan that make it superior in some way. Examples of such non-quantitative attributes include keeping together communities of interest beyond those stated in the Voting Rights Act and recognizing geographic barriers, such as mountain ranges and large bodies of water. Thus, we want a method that also highlights near Pareto plans.

Because we want the Pareto frontier of *available* maps and a method for scoring near Pareto maps, we prefer DEA to skyline methods, although DEA is slower. Skyline methods may only provide an approximation to such a Pareto frontier. They do not provide the radial measure of proximity to the frontier. The Supreme Court has already rejected using statistical methods to evaluate maps based on MCMC methods. DEA provides a viable alternate measure by showing how far a map is from the convex hull of the Pareto frontier and it provides instances of maps that dominate the contested map.

Once a districting commission proposes a map, the DEA score of this proposed map, along with the Pareto frontier of other existing maps, provides a starting point for mappers in advocacy groups and political parties to then search for new, dominating maps that have the potential to demonstrate the weaknesses of the proposed map and/or provide a useful guide to improvement. These Pareto or near-Pareto maps can also be considered to be superior alternatives in a court challenge to a gerrymandered map.

If state law requires a well functioning independent commission, the commission can take advantage of community-drawn maps to present samples of Pareto and near-Pareto maps in hearings and in revising its map. Such Pareto maps can also be made available online prior to determining the final choice for redistricting.

Another requirement for any analytic approach is that it has to deal with incommensurate measures. Some districting plan measures are in the single digits while others (e.g., a Herfindhal index to be discussed later) number in the thousands. Thus, we want a measure of proximity to the frontier that is independent of scale and based on percentages. The ratio scale of DEA does this, while skyline methods don't, unless the data is scaled beforehand.

Thus, DEA is useful for finding maps on or near the convex hull of the Pareto frontier and can be useful in the context of the extended and highly contested process of redistricting both by commissions and by advocacy groups.

3 Contributions of the paper

The paper makes the following specific contributions:

1. To the best of our knowledge, this is the first paper applying a DEA methodology to the application area of political districting.

2. We provide a new radial DEA formulation to measure the proximity of a districting plan to the Pareto frontier when output measures are mixed in the following sense: some output measures satisfy bigger is better, while other output measures satisfy smaller is better. We also provide an efficient solution algorithm to our DEA formulation, which is non-linear and non-convex.

3. Finally we discuss several avenues for future research that can extend our DEA model. In particular, the Skyline literature stream from database query management can be effectively combined with our DEA approach to characterize the Pareto frontier for any districting project.

We now describe our implementation of DEA, which is the principal tool we use to generate our results.

4 Solution Methodology

In standard DEA one measures the inputs and outputs of multiple units of an organization or industry. One then finds the most efficient units for producing the outputs, given their mixes of inputs. The result is a production function. A classic use of this tool is comparing various bank branches. The inputs are factors such as staff skills, number of employees and equipment. Outputs (evaluation criteria) include the numbers of loans closed, deposits garnered, ATM fees and other similar factors. See Cooper et al. (2011).

In the redistricting process, we have multiple output measures for a plan but no inputs as in the models of Belton and Vickers (1993) and Liu et al. (2011). Furthermore, higher scores for the output measures can indicate either a bad or good outcome. As examples, the Polsby-Popper measure for compactness, which we define later, is a measure where a larger value is better, while the number of split counties in any districting plan is an output measure where smaller is better. A central feature of our application is that the number of potential plans is too large to be generated in a reasonable amount of time and samples of billions plans have been generated. Thus, when we refer to the Pareto frontier, we have the frontier for a *sample* of plans, not the underlying frontier for the entire universe of plans. This is analogous to the "best practice frontier" described by Cook et al. (2014). One of the features of the DEA models in the papers referenced above is that all measures improve in the same direction, e.g., either bigger is better for all measures *or* smaller is better. Thus, the models referenced here do not apply directly. The models in the bulk of the literature assume bigger is better, consistent with the free disposal assumption in economics, which states that any output produced in excess can be discarded at no cost.

One necessary condition for the Pareto frontier to be convex is that the consideration set is continuous. Since we are dealing with a finite set of maps and some measures are discrete, the Pareto frontier is not convex. Instead, it has the shape of a step function (hence the term "skyline"). Our DEA model produces the *convex hull* of the Pareto frontier of the consideration set of plans. Thus, it can miss plans interior to the convex hull that are on the Pareto frontier. We introduce additional processing to capture those and other potentially good plans. The result is a collection of plans close to the convex hull of the Pareto frontier. We term those"good plans" for our subsequent discussion because proximity to the Pareto frontier means the plan does well on one or more measures that are of interest to some groups of citizens.

We include plans that are near (but not on) the Pareto frontier for two reasons. First, we want to lower the odds of discarding a good plan when that plan has features that are of interest but not included in the measures used in calculating DEA scores. Second, the goal is not for the tool to provide a clear demarcation between good and bad plans. Instead, it is to discard obviously bad plans, reducing the final numbers that must be evaluated by districting commissions or researchers in the search for good plans. The method for retaining and eliminating plans has to balance the time it takes to evaluate a plan that is retained and the possibility of discarding a very good plan that would appeal to voters.

Dividing the plans into two groups, the ones for further examination and the ones to be discarded, resembles hypothesis testing in statistics. Since there is no perfect discriminator between the two groups, good plans can be rejected, a type I error, or bad plans can be accepted, a type II error. A preliminary sort, such as the one presented here, should have a low type I error, which means the type II error can be larger than standard examples of hypothesis testing. The dividing line between acceptance and rejection can be chosen based on the mix of good and bad plans around the cutoff point and the number of plans to retain for further consideration. This is a user decision.

4.1 A graphical overview of the augmented DEA method for selecting plans

The basic idea can be illustrated graphically using a small example. To simplify the discussion, we use just two of the criteria, split political/administrative units and population deviation from strict equality among districts. We note that both measures used in the illustration are of the type smaller-is-better. The methodology can be applied with any number of criteria. The data used in the illustration are hypothetical and not connected to any real plan.

Figure 1 plots the hypothetical number of splits and the population-deviation scores for 7 plans. The bigger the number on either dimension, the worse a plan is. Non-dominated plans are ones for which there is no plan that is better on both splits and population deviation. These plans form the Pareto frontier, the line 1-5-2-3-7-4. The line connecting plans 1, 2, 3, and 4 form the convex hull of the efficient frontier. If there were a continuum of plans, with concave objective functions (when maximizing) the Pareto frontier would likely be convex.

Plan 1 has many splits but is perfect on population deviation, while plan 4 has no splits but has a large population deviation. Plans 2 and 3 fall in the middle, and no plans are better than them in both dimensions. The dominated plan, 6, is worse than plan 3 on both dimensions and, presumably, can be gotten rid of immediately. All points are either above or on the Pareto frontier because smaller is better for the two criteria. No points fall below the Pareto frontier. Plans 5 and 7 are above the convex hull of the Pareto frontier. However, they are not dominated. If they are too far above the convex hull of the Pareto frontier, they might be discarded, as we did with 7. Plan 7 has more than double the population deviation as plan 3. If they are close to the frontier, as with 5, they likely should be kept because they might have some desirable properties that are not part of the criteria and they fill out the Pareto frontier. Furthermore, plans close to the convex hull have the potential for adjustment to improve their quality, e.g., by taking into consideration criteria other than those used to construct the convex hull. An example of a desirable property that is not an expressed criterion in the example is to keep a district on one side of a ridge line or a large river.

We have connected the points on the Pareto frontier as if it were possible to

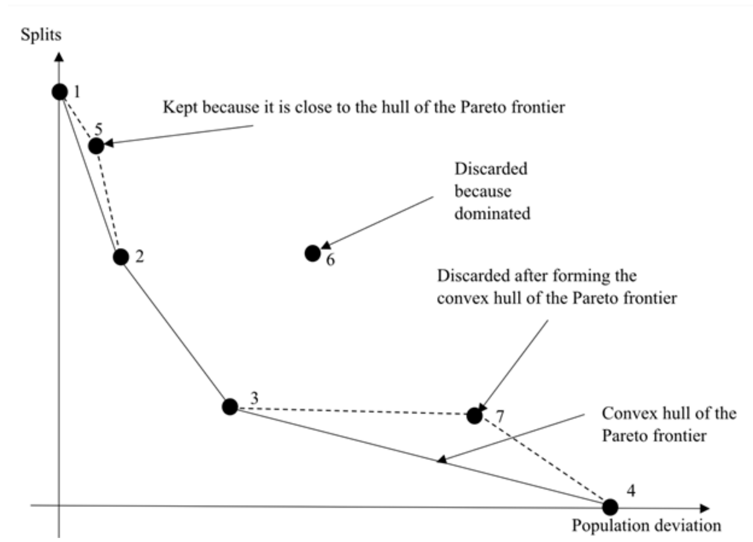

Figure 1: A plot of two properties of seven plans.

generate a plan for all points on the lines that make up the Pareto frontier and convex hull. However, there are a discrete number of plans, not a continuum of possible plans. A plan that is above the convex hull of the Pareto frontier can be a Pareto point and can be a potentially good plan as a compromise among the criteria. This figure illustrates a formal process for keeping good plans that are close to the Pareto frontier for serious evaluation. (In this graph for both measures, smaller is better. For an illustration of a graph with both bigger-is-better and smaller-is-better measures and an illustrative solution, go to the online appendix at Gopalan et al. (2023).

Finally, it should be remembered that the convex hull is constructed with respect to a particular consideration set of plans. Subsequent additional plans can dominate plans initially on the convex hull, resulting in a change in the hull. When that happens, the subset of plans kept for further evaluation is altered, with the newly found good plans added and highly-dominated plans removed. (Previously removed plans are never returned to the consideration set, as the DEA scores are nonincreasing as plans are added.) Consequently, constructing the convex hull is materially useful for identifying newly discovered plans of good quality during an ongoing submission process. The augmented DEA analysis thus serves as a tool for ongoing deliberation.

4.2 Finding plans on the convex hull of the Pareto frontier and ones nearby

Since we are dealing with cases (and not production functions with scalable inputs and outputs), we cannot use a DEA model that presumes homogeneity. Our starting point is the BCC model. A feature of this model is that it has a convexity constraint, which leads to an important property: the solution and objective value are invariant when scaling any of the output measures. This is an important requisite feature when dealing with districting plans because there is no natural common unit for all measures and moreover, their ranges are radically different.

We begin with the formulations for the BCC input and output models and then combine the two types of models into a single model that is appropriate for our context. The first of the two BCC models finds the most efficient set of inputs for producing the outputs of the technology of interest and the second finds the best possible production of outputs given the inputs of the technology of interest.

Let

i index inputs

j index outputs

k index technologies

k' the technology being evaluated in a given model

a_{ik} be the technology coefficients for inputs

a_{jk} be the technology coefficients for outputs

λ_k be the level of technology k in the solution

θ be the radial efficiency measure

DEA Input model:

$$\min \theta \tag{1}$$

Subject to

$$\theta a_{ik'} - \sum_k a_{ik} \lambda_k \geq 0 \tag{2}$$

$$\sum_k a_{jk}\lambda_k \geq a_{jk'} \tag{3}$$

$$\sum_k \lambda_k = 1 \tag{4}$$

This model finds the smallest fraction that leads to at least one input matching the most efficient combination of production choices for that input while exceeding all other inputs for the efficient combination. At the same time it meets or exceeds the outputs of the most efficient combination of inputs.

DEA Output model:

$$\max \theta \tag{5}$$

$$\sum_k a_{ik}\lambda_k \leq a_{ik'} \tag{6}$$

$$\sum_k a_{jk}\lambda_k \geq \theta a_{jk'} \tag{7}$$

$$\sum_k \lambda_k = 1 \tag{8}$$

This model finds the largest scaling of outputs that leads to one output matching the most efficient combination of production choices for that output while falling below all other outputs for the efficient combination of outputs. At the same time it meets or uses more of the inputs of the most efficient combination. Note that $\theta \geq 1$ in this model. If $\theta = 1$, then this technology is efficient.

These models are classified as “radial” models in that they measure the percentage change in inputs or outputs, not the deviations from the best levels of inputs and outputs. In our case, we do not have inputs. However, we have some outputs for which smaller is better.

Neither of these models applies directly because they do not allow variable weights between smaller-is-better outputs and larger-is-better outputs. What we want is the equivalent of the radial measure for reaching the convex hull. This means we are looking for the smallest percentage shift in the increasing and decreasing measures such that a combination of measures touches the convex hull with a mix of smaller-is-better and bigger-is-better measures.

We now examine the model of Dakpo et al. (2017) because it captures the features of the environmental models.

We extend our previous notation to include:

ip index polluting inputs

in index non-polluting inputs

jg index good outputs

jb index bad outputs

μ shares of reference set for polluting inputs

Then the formulation in Dakpo et al. (2017), stated in our notation is:

$$\min \theta \tag{9}$$

Subject to:
(Note: The first four constraints are from the radial input model with the input constraints separated into polluting and nonpolluting inputs. Also, the first term in the first two equations has no θ as in the input model.)

$$a_{ipk'} - \sum_k a_{ipk}\lambda_k \geq 0 \tag{10}$$

$$a_{ink'} - \sum_k a_{ink}\lambda_k \geq 0 \tag{11}$$

$$\sum_k a_{jgk}\lambda_k \geq a_{jgk'} \tag{12}$$

$$\sum_k \lambda_k = 1 \tag{13}$$

The next two constraints are from the output model and applied only to polluting inputs. Note that the direction of the inequalities are reversed to measure the efficiency with respect to pollution and that μ replaces λ.

$$\sum_k a_{ipk}\mu_k \geq a_{ipk'} \tag{14}$$

$$\sum_k a_{jbk}\mu_k \leq \theta a_{jbk'} \tag{15}$$

The convexity constraint of the output model is replaced by a control total requiring the same level of polluting inputs.

$$\sum_k a_{ipk}\lambda_k = \sum_k a_{ipk}\mu_k \tag{16}$$

By introducing μ rather than using λ in these constraints the model remains a linear program but sacrifices the precision of using the same mix of the reference set for all inputs and outputs. We, however, need to have the same weights for all outputs to truly compare plans.

In the following formulation we simultaneously meet the input/smaller-is-better and output/greater-is-better constraints using the input constraint from the input model and a modified version for the output constraint. The formulation finds the minimum percentage deviation from 1, i.e., the convex hull. If $\theta = 1$, the solution is on the convex hull of the Pareto frontier. The composite DEA formulation for evaluating technology (districting plan) k' becomes (parametrized by θ):

MathProg(θ):

$$\min \theta \tag{17}$$

Subject to:
Constraint for "smaller is better" outputs:

$$\sum_k a_{ik}\lambda_k \leq \theta a_{ik'} \tag{18}$$

Constraint for "bigger is better" outputs:

$$\theta \sum_k a_{jk}\lambda_k \geq a_{jk'} \tag{19}$$

Constraint for convexity:

$$\sum_k \lambda_k = 1 \tag{20}$$

$$\lambda_k \geq 0 \tag{21}$$

This model is nonlinear and non-convex because of the product of two variables in (19). Thus, a standard linear-programming solver cannot be used and a nonlinear programming solver does not guarantee an optimal solution. We address this issue simply.

Note that giving θ a fixed value results in a linear program where the solver finds a feasible solution for the given θ if there is one and reports that the model is infeasible if there is no solution. Since $\theta = 1$ is feasible and $\theta = 0$ is infeasible, $\theta > 0$. Since θ is unidimensional, we can use interval bisection search to solve the model. The algorithm is as follows.

Algorithm-Bisection-Search:
Let ϵ = a small tolerance parameter (in our runs this was set to 0.0001).
Let ν = a small positive number , say 0.001.
Let θ_{UB} =1 and $\theta_{LB} = \nu$.

Do while (θ_{UB} - θ_{LB}) $> \epsilon$:
    Fix current θ_{CUR} = (θ_{UB} + θ_{LB})/2.0
    Solve MathProg(θ) after fixing θ to θ_{CUR}.
    If MathProg(θ_{CUR}) is feasible, update θ_{UB} = θ_{CUR}
    If MathProg(θ_{CUR}) is infeasible, update θ_{LB} = θ_{CUR}
End While Loop
Return θ_{UB} as the desired efficiency .

Theorem 4.1 *The interval bisection algorithm finds the global minimum of MathProg(θ).*

Proof: Note that in MathProg(θ) if θ is feasible and $\theta' \geq \theta$, then, θ' is feasible in constraints (18) and (19). Thus, there is only one boundary point between feasible and infeasible values of θ. We know $\theta = 1$ is feasible and $\theta = 0$ is infeasible. Thus, the interval (0,1] brackets the optimal solution. Interval bisection narrows the upper and lower limits of the potential feasible values of θ and finds the smallest feasible θ, which is also the optimal solution.

Moreover, since the search interval is cut by a factor of 2 in each iteration, the bisection search terminates in at most $\log_2(1/\epsilon)$ iterations. If ϵ = 0.0001, the algorithm terminates in about 14 iterations. We have formulated and solved this model using Python and PYOMO, which is open source software (`http://www.pyomo.org`). Inexpensive (under $1000) commercial software is also readily available.Thus, we have a formulation for models with no inputs (from the DEA model class DEA-WEI, see Liu et al. (2011)), where some output measures can be "smaller is better" while other output measures are "bigger is better". We have posted a small example showing the progress of iterations in the online appendix that can be found at Gopalan et al. (2023). If we add back the input constraints, this formulation can also be used for environmental models whenever there is a need for greater precision (Dakpo et al. (2017)).

5 Eliminating weak plans before applying the DEA model

In this and next few sections we walk through an example of using DEA to evaluate plans using a set of randomly generated plans. Our goal is to illustrate the process and explore the nature of the results from using the model using randomly generated plans.

When processing the plans in this paper, we make a set of choices on what criteria we use, the cutoff levels, etc. None of these are cast in stone, as this part of the paper just illustrates the use of our methodology. In actual implementations, choices on cutoff levels, the criteria, and the details of the steps should be determined by the specifics of the situation.

Plans on or near the convex hull of the Pareto frontier do well using the subset of measures with significant weight (larger dual variables) in the optimal solution. A feature of our approach is that a plan with a good score can do badly on some measure when calculating its proximity to the Pareto frontier because the weights chosen in the model are the ones most favorable to the plan and a criterion in which a plan does badly can be assigned a low weight.

To make sure a plan selected for further evaluation performs reasonably well on all criteria, we first reject plans that perform very badly on any one measure. If used, the cutoff should be relatively loose so that plans that are good at blending criteria are not discarded. In the example given later in the paper, we filter on the compactness and split district measures, using the medians of the distributions as the cutoffs. No cutoff is applied to the population deviation measures as all population deviations for the 11,206 plans are within an acceptable threshold by design. With many states enforcing a requirement that the maximum deviation from the target population be one-person, most redistricters would make the same choice. However, future research could investigate the degree to which rigid population equality rules force states to adopt plans that are dominated on all other redistricting criteria (i.e., have low DEA scores compared to plans with slightly higher population deviations). This would be useful because issues in data collection guarantee that no maps meet the strict population-equality goal even when required.

In our example analysis we take the following steps when deciding which plans to retain for further evaluation:

1. For every plan calculate the numerical values for each criterion (plan measure) such as population deviation, splits, and compactness.

2. Remove obviously weak plans, defined as plans that are worse on any single criterion by more than some amount, e.g., remove plans below some percentile value for compactness.

3. For every remaining plan find the value of the DEA efficiency θ, which brings the plan closest to the convex hull of the Pareto frontier as defined by (1)-(6). The output of the mathematical model is a measure of how far a plan is from the convex hull of the Pareto frontier.

4. Rank the plans from the lowest to highest cost and establish a θ^{min} cutoff for plans, discarding plans with a θ below the specified threshold value or alternatively, keeping a specified number of the plans with the the highest values for θ.

Next we examine the behavior of this model using a collection of 11,206 stochastically generated plans for Pennsylvania made publicly available by (Haas et al., 2020).

6 Preparing the plans for evaluation in the DEA model

We begin by summarizing our proposed overall procedure. The first step is to select the measures we use to evaluate plans. Here we use population deviation, compactness and county splits as the criteria. Other common redistricting criteria that we do not use for this study include the number of majority-minority districts, competitiveness, and political fairness. Any measurable criterion can be incorporated in the DEA methodology.

For each criterion, there can be multiple measures of that single criterion, which we address below. It is also possible that a criterion would benefit from using multiple measures to capture its critical features. For example, when addressing the requirement for majority minority districts, the number of such districts is clearly an important measure. Since a majority minority district can be packed to lessen the influence of the minority in a legislature, the excess minority population beyond 50% can be included as an additional measure. If the desired number of such districts is low due to the size of the minority population, then it is better to set a minimum number of districts as a constraint in the initial sort. One can add a measure of the compactness of the majority minority districts to assess the extent to which these districts are distorted to achieve their majority.

We start with three ways of measuring each of the criteria used in this illustration, i.e., three measures of compactness, etc. As some of the measures are highly correlated, we retain only some of them, illustrating the kinds of choices one can make in implementing our model. For each plan we calculate the following values:

1. Population deviation measures:

    - Absolute difference in population between the largest and smallest district. This is the primary population measure used in our analysis.
    - Maximum absolute deviation over all districts from the *target population*, defined as the population which would equalize population among the districts in the plan.
    - Absolute deviation of a district's population from the target population, summed over all districts.

2. Compactness measures:

    - Polsby-Popper, PP, or the isoperimetric quotient, is the ratio of the area of the district to the area of a circle with the same perimeter as the district.
    $$PP = \frac{4\pi \mathit{Area}}{\mathit{Perimeter}^2} \tag{22}$$
    - Convex hull ratio is the ratio of the area of the district to the area of the district's convex hull; this measure is sensitive to indentedness of the perimeter, a crescent having a particularly bad convex hull ratio.

- Moment of inertia-based compactness uses the definition in Li et al. (2013). Let I_g be the area moment of inertia about the centroid g for a district. Let z_g be the distance from the centroid g of the district to an infinitesimal area in the district, dA. Then calculate

$$I_g = \int z_g^2 dA. \tag{23}$$

The compactness shape index, C_{MI}, is the moment of inertia relative to the moment of inertia of a circle with the same area. It is measured as the square of the area of the district, $Area$, divided by $2\pi I_g$:

$$C_{MI} = \frac{\mathit{Area}^2}{2\pi I_g} \tag{24}$$

.

While we make use of the area moment of inertia, it is possible to use the population of district building blocks (e.g., precincts, census tracts) to calculate the population moment of inertia.

3. Measures for splits in political units (e.g., counties):

   - Splits: This is the sum of splits of political units (counties) over the entire state. For each unit, this is equal to N-1 where N is the number of pieces that comprise the unit. That is, for a unit that is kept whole, its contribution to the score is 0. A unit in 3 pieces contributes 2 (it has been split twice).
   - Population spill ratio: The share of political unit population that is split away from the single *most populous* territory of the unit.

   $$S_2 = 1 - p_1 \tag{25}$$

   where p_1 is the share of population in the most populous single territory fragment assigned to any district. If the political unit falls entirely within one district, $p_1 = 1$ and the spill ratio $S_2 = 0$.
   - Herfindahl index of concentration for each political unit (used in economics to measure concentration in an industry). If the Herfindahl index for a political unit is higher, there are fewer splits. For each political unit

   $$H = \sum_{i=1}^{N} p_i^2 \tag{26}$$

where N is the number of fragments the county is split into and p_i is the *percentage* share of county population in each district, i. For a single political unit that is not split at all, H attains a maximum value of 10,000 (= 100$\times$100). For a complete districting plan, this measure can take a maximum value = (Number of politcal units)$\times$10,000.

The population deviation measures are straightforward and not discussed here. While many states require "compactness" for legislative or Congressional districts, most states do not define what compactness means. At the same time, people have a natural sense of what compactness is, Chou et al. (2014). A few quantitative measures have been used in redistricting court cases, and many more have been proposed, e.g., Niemi et al. (1990); MacEachren (1985); Young (1988); Horn et al. (1993). Useful properties of a compactness measure include scale invariance and, to facilitate comparison among measures, being bounded by 0 (lowest compactness) and 1 (highest compactness) (Niemi et al. (1990)). We apply three such measures, described briefly above. [5] For each of the three measures, the compactness of the 17 districts is averaged to generate a score for the overall plan. Larger is better for all three compactness measures.

The first of the definitions for splits is the standard one (smaller is better). The second and third measures attempt to correct for the fact that some stochastic processes used to generate pools of redistricting plans (including the pool of plans used in this article) will generate large numbers of political unit splits that could be easily corrected by human intervention. The second measure calculates the proportion of population that spills over into neighboring districts (lower is better). The third measure uses the Herfindahl index(H), which was originally constructed to measure the concentration of firms in an industry. At its maximum value (10,000), the entire population of a political unit is contained, undivided, within the same political unit. The H values are summed for all political units to provide a composite H value for the districting plan as a whole. If a number between 0 and 1 is desired for H, the composite H derived previously can be divided by (number of political units*10,000). Finally, we note that when the population spill ratio is small, or when Herfindahl H is large, the political unit splits could be repaired fairly easily by trading building blocks (e.g. election precincts) between neighboring areas.

[5] All three measures are calculated using a Python module available at https://github.com/leehach/geocompactness.

7 Exploring the Distributions of DEA Scores and Correlations between Plan Measures

To explore the behavior of the DEA model, we successively ran the model with just one measure from each category for a total of 3 criteria, two measures from each of the categories for a total of 6 possible criteria, and finally all three measures in each category for a total of 9 criteria. The effects of increasing the number of criteria are to increase the number of plans on the convex hull of the Pareto frontier and compress the distributions of the DEA scores. This is known as the curse of dimensionality in the DEA context. With just three criteria, only 5 plans are on the convex hull; with 6 criteria 24 plans are; and with 9 criteria 47 plans are on the convex hull of the efficiency frontier.

Figure 2 shows the densities of the distribution functions of the efficiency scores for plans using 3, 5, 6, and 9 criteria. The first property to notice is that there is a small bump at 1 because of the plans on the Pareto frontier. This bump increases as more plans are added to the DEA model because the added plans create more opportunities to be on the Pareto frontier. Second, the density function of DEA scores narrows as the number of measures increases and, outside of the bump, is unimodal. Third, the distribution of scores also tends to have a higher peak with more measures, but not always. Third, the distribution of scores shifts to the right as more measures are added. The third feature implies a narrower range of scores.

The reason one would expect the distribution with 6 measures to be shifted right relative to 5 measures is that adding a variable (measure) to a linear program increases the flexibility in choosing the optimal solution. With a variable added to any linear program where the constraints are unchanged, the prior solution is still feasible. Thus, the value of the objective can only be improved. This is akin to getting a better fit in a regression when more independent variables are added. That is, there is a tradeoff between the number of measures included and the ability of the model to discriminate among plans. Beyond the next lemma, this is an issue we do not explore further here.

Lemma 7.1 *Adding a new measure to our DEA model results in a nondecreasing DEA score for a decision making unit.*

Proof: Note that adding a new measure adds a constraint and reduces the size of the feasible region. If the original solution is still feasible, there is no better

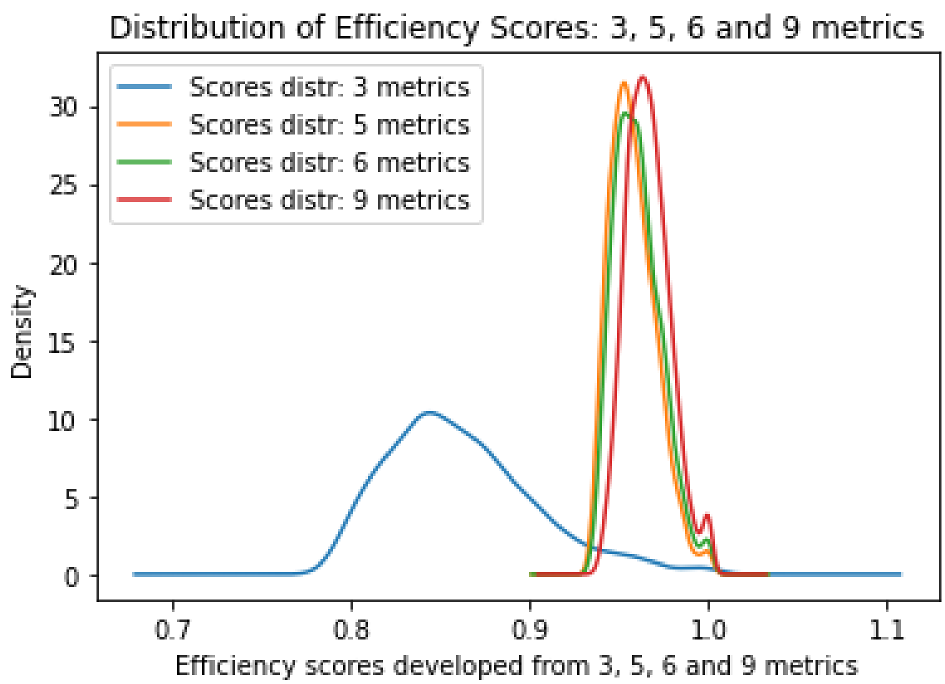

Figure 2: Distributions of efficiency scores.

solution in a subset of the feasible region and θ is unchanged. Assume the original optimal solution is not feasible in the added constraint. If the original optimal solution is unique, θ must increase to achieve a feasible solution and the value of the objective function for the new optimal solution is higher than the original solution. Alternatively, there are multiple solutions with the original value for θ and one of those solutions is still feasible.

To decide which criteria to include, we examined the correlation matrix of the 9 measures with the 11,206 plans. See Table 1.

Within each group of criteria some of the correlations are high. The population measures are especially highly correlated and very similar in what they are measuring. Consequently, we use the standard measure of the maximum spread from largest to smallest as the lone population measure in the model. The compactness measures are also highly correlated. They operationalize different concepts of compactness and are each sensitive to different deviations from compactness, such as irregular boundaries (PP), indentedness (convex hull ratio), and dispersion (C_{MI}). We retain PP and C_{MI}, the two least correlated of the compactness measures. We do not use the standard measure for splits because we want to preserve the ability to slightly modify plans, reducing splits by selecting plans that are amenable to that repair. Thus, we keep the measure of splits (S) that does not count small separations from political districts, and the Herfindahl index, which measures a different aspect of splits.

The virtue of this model is that it finds plans that are good using combinations of measures. However, by allowing the model to choose the weights applied to the measures, the result can be a significant underweighting of a measure in which a plan does badly. Since a plan should be reasonable on all measures, we introduce the rule that every plan kept for further evaluation should be better than the median of all plans for each of the chosen measures. Thus, we find the median value for each measure (except for population deviation) over the full set of plans and then drop all plans that fall below the median on any measure. This reduces the number of plans from 11,206 to 2,355. (Note that this changes the correlations of the measures. The correlations are much higher with the original 11,206 plans because the larger set spans a wider range of values. For example, the correlation of Polsby-Popper drops to 0.39 with 2,355 plans from 0.62 with 11,206 plans.)

Table 1: Correlations among the measures of goodness of maps

Measure	Pop. deviation	Truncated pop. spill	Herfindahl	Polsby-popper	Moment of Inertia	Convex hull ratio	County splits	Sum pop dev	Max pop. Dev
Pop. deviation	1.00	-0.00	-0.00	-0.04	-0.02	-0.03	-0.03	0.62	0.88
Truncated pop. spill	-0.00	1.00	0.88	0.14	0.01	0.03	0.09	-0.01	-0.00
Herfindahl	-0.00	0.88	1.00	0.22	0.06	0.11	0.40	-0.02	-0.01
Polsby-popper	-0.04	0.14	0.22	1.00	0.39	0.64	0.27	-0.04	-0.04
Moment of Inertia	-0.02	0.01	0.06	0.39	1.00	0.55	0.13	-0.00	-0.03
Convex hull ratio	-0.03	0.03	0.11	0.64	0.55	1.00	0.23	-0.01	-0.02
County splits	-0.03	0.09	0.40	0.27	0.13	0.23	1.00	-0.02	-0.03
Sum pop dev	0.62	-0.01	-0.02	-0.04	-0.00	-0.01	-0.02	1.00	0.52
Max pop dev	0.88	-0.00	-0.01	-0.04	-0.03	-0.02	-0.03	0.52	1.00

8 Results

The purpose of this section is to observe the behavior of the plan measures as a function of the DEA scores. What we expected, and did observe, is that high (DEA) scoring plans tended to do better than low (DEA) scoring plans on every plan measure. However, as expected, not all plans with a high DEA score do well on *every* measure because the model can give a low weight to a measure on which the plan performs poorly. Plans with low scores generally did badly on all measures. This makes sense because if a plan has a low DEA score, it does not do well on any measure. Thus, the model accomplishes what we are looking for: a way to filter out plans near the Pareto frontier for further evaluation.

To test the discriminatory power of our model, we adopt the following process:

1. Create three comparison data groupings (datasets) as follows. Extract the plans falling in the top quintile (best 20% of the DEA scores for the 2,355 plans) and worst quintile (lowest 20% of DEA scores). This gives three comparison datasets: top DEA quintile, bottom DEA quintile, and the entire corpus of 2,355 plans.

2. Construct histograms of the distribution of each measure (e.g. population deviation) for each of these three datasets, using the following process:

   (a) Create five equal-interval bins for each measure.

   (b) Scale the y-axis to show the proportion of plans within each bin *in each of the three datasets*. For example, Figure 3 shows that 0.48 (48%) of plans in the bottom quintile fall into the highest bin on the population deviation measure. This is 48% of plans in the 471 (=2355/5) plans in the bottom quintile of DEA scores, not 48% of all 2,355 plans.

   (c) For each measure, display histograms for all three datasets in one graph, to allow visual comparison.

3. Examine whether the top DEA quintile outweighs the bottom DEA quintile in the "appropriate" bin. For instance, for the population deviation measure, lower population deviation is a better measure for the plan. Therefore, in bin 1 of the population deviation graph, the top DEA quintile of plans should outweigh the bottom DEA quintile. For the Herfindahl measure however, higher is better and therefore the top DEA quintile should outweigh the bottom DEA quintile in bin 5.

4. Assess the ability of each plan measure to be a strong contributor to the discriminatory power of the DEA procedure.

We now present the binning behavior of the 2,355 plans in the top quintile of DEA scores versus the bottom quintile for 5 metrics: population deviation, population spill ratio, Herfindahl, Polsby-Popper and moment of inertia. In the following histograms the bins on the horizontal axis are the range of values of that measure divided into 5 equal-width intervals, *not* quintiles. The vertical bars are the proportion of the plans in the top quintile of plans, the bottom quintile of plans, and all plans, in the relevant ranges.

One of the features that stands out is that the percentage change in the range from the bin with the smallest value in the histograms to the largest is small. That, is, on the chosen measures the plans drawn using stochastic (randomized) methods have similar values. This is partly due to cutting out the plans that are below the median in any measure before running the plans through the model, as we can see for a few of the measures in the histograms below. Another factor is that if a plan is very good on one or more measures it can be in the top DEA quintile while being weak in other measures.

Since plans can be strong on some measures and weak on others, for any one measure highly rated plans can be among the worst performers in that measure. Thus, when examining the distribution of outcomes on a single measure one would not expect to see a clear demarcation between higher and lower rated plans on any single measure, and we do not see a clear demarcation.

For population, we use the maximum deviation measure. See Figure 3. Note that since the plans are generated to meet the criterion of a maximum population deviation of 2%, the range is very narrow. Still, the results show the pattern that the top quintile of plans skew towards lower deviations and the bottom quintile skew towards the the higher deviations. All 9 plans in the bin with the lowest deviation are included in the top quintile.

This graph illustrates the basic pattern resulting from the DEA model giving high ratings to all or almost all of the plans that are best on some measure or combination of measures: some plans that are weak on one measure but good on others are retained. That is, this approach retains a variety of plans, which is good because the idea with this approach is to evaluate plans based on multiple measures *simultaneously* and have a portfolio of plans on or near the Pareto frontier for further human evaluation. Note that the number of the bottom quintile of plans peaks in the bin for the highest (worst) deviation and decreases with lower

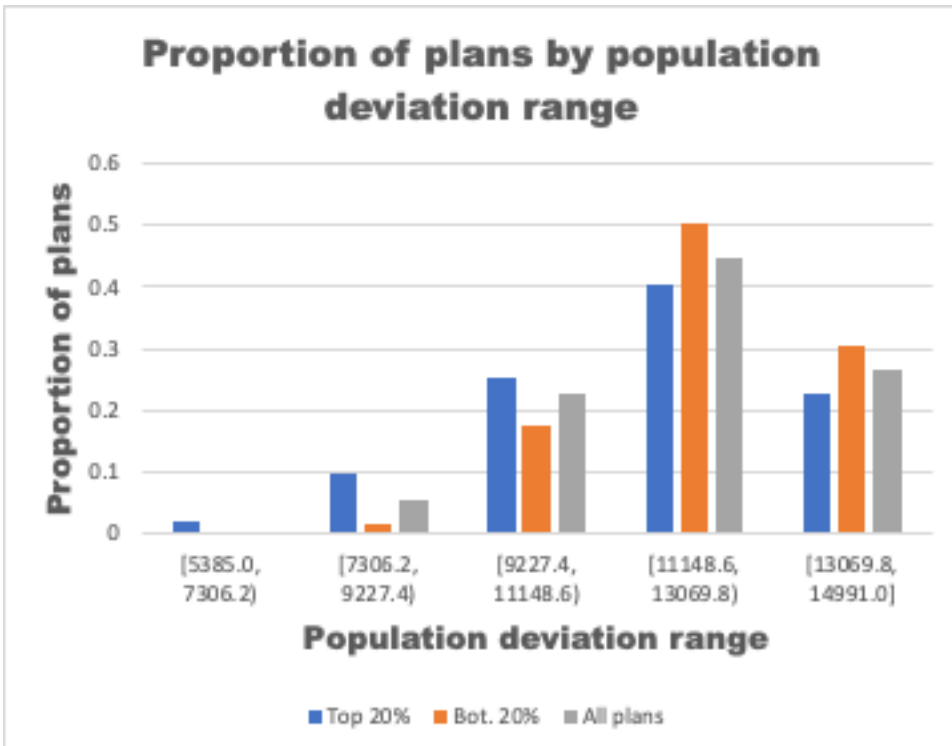

Figure 3: Plan distributions for population deviation comparing the top quintile in score with all 2,355 plans and the bottom quintile of plans.

deviations. At the same time the top quintile of plans has an almost symmetric unimodal distribution. This reflects the removal of the bottom 50% of plans in each measure and that the values of the measures are correlated. This pattern results from the values of the measures coming from a unimodal and symmetric distribution of all plans, such as the normal distribution. In our situation the histograms are unimodal, except for the bump at 1.0, when all plans are included.

The moment of inertia results, Figure 4, show clear discrimination between the top quintile and the bottom quintile. Sixty-two percent of the plans grouped into the bottom quintile of DEA scores are in the bottom bin of plans as measured by moment of inertia. The eight best plans on moment of inertia have DEA scores in the top quintile. The pattern described above appears with this and all other measures. For the top quintile of plans the distribution of scores is unimodal and relatively symmetric, while the distribution of scores for the bottom quintile is skewed with the mode at the worst range of outcomes.

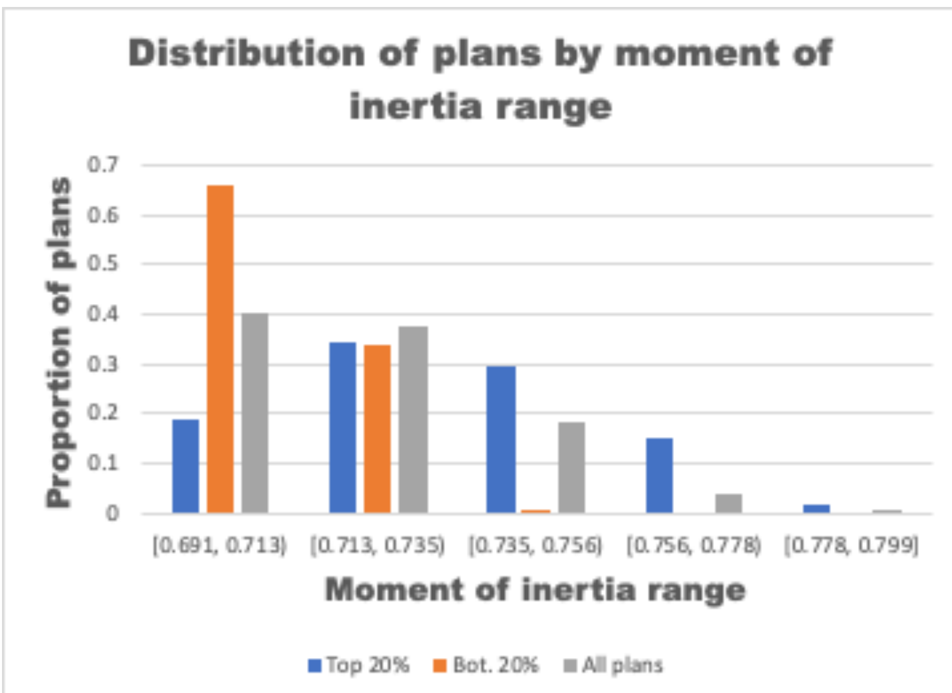

Figure 4: The distributions of moment of inertia for the top quintile of plans, all plans, and the bottom quintile of plans.

The Polsby-Popper measure, Figure 5, strongly differentiates between the top and bottom quintiles and shows the same pattern as moment of inertia with roughly the same differentiation among the quintiles. This is surprising since the Polsby-Popper measure uses the perimeter in calculating this measure, which is sensitive to wiggles in the boundary, while the moment-of-inertia measure does not increase with minor wiggles to the boundary.

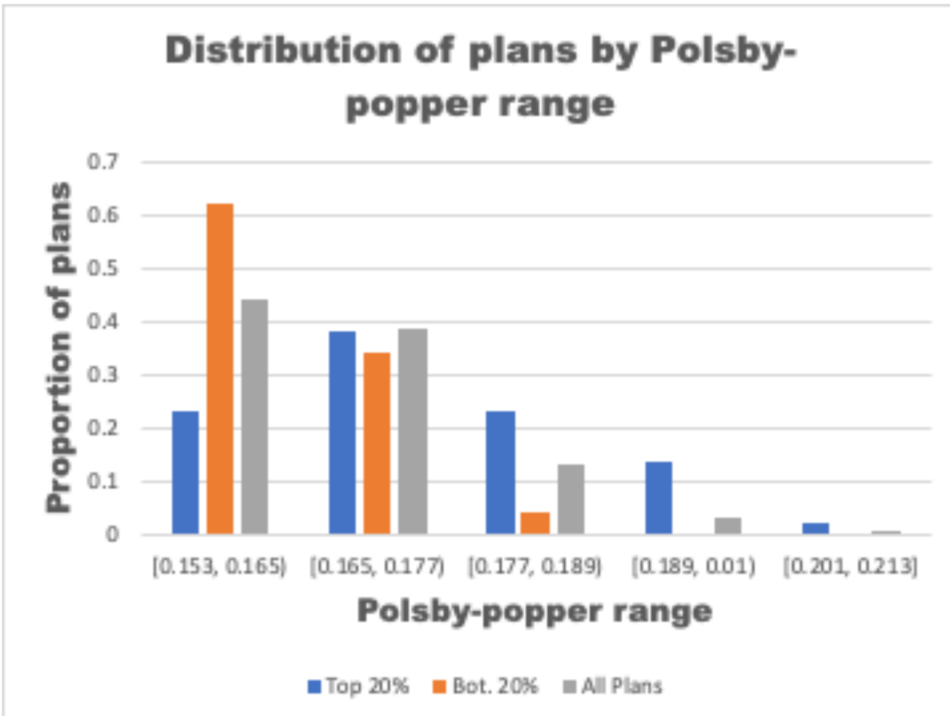

Figure 5: The distributions of the Polsby-Popper values for the top quintile of plans, all plans, and the bottom quintile of plans.

The results with the population spill ratio parallel the other results in that the top quintile of efficiency scores do better than the bottom quintile and the full population as seen in Figure 6. The differences between the top quintile and the bottom quintile are dramatic. The bottom quintile does not have any plans in the best two bins and only 1 in the third best bin. The vast majority of the bottom-quintile plans are in the worst bin.

The Herfindahl index, Figure 7, performance has the same character as the spill ratio. Still it has only 1 plan in the bottom quintile outside the worst bin. Only 6% of the plans in the top quintile of plans appear in the worst bin and

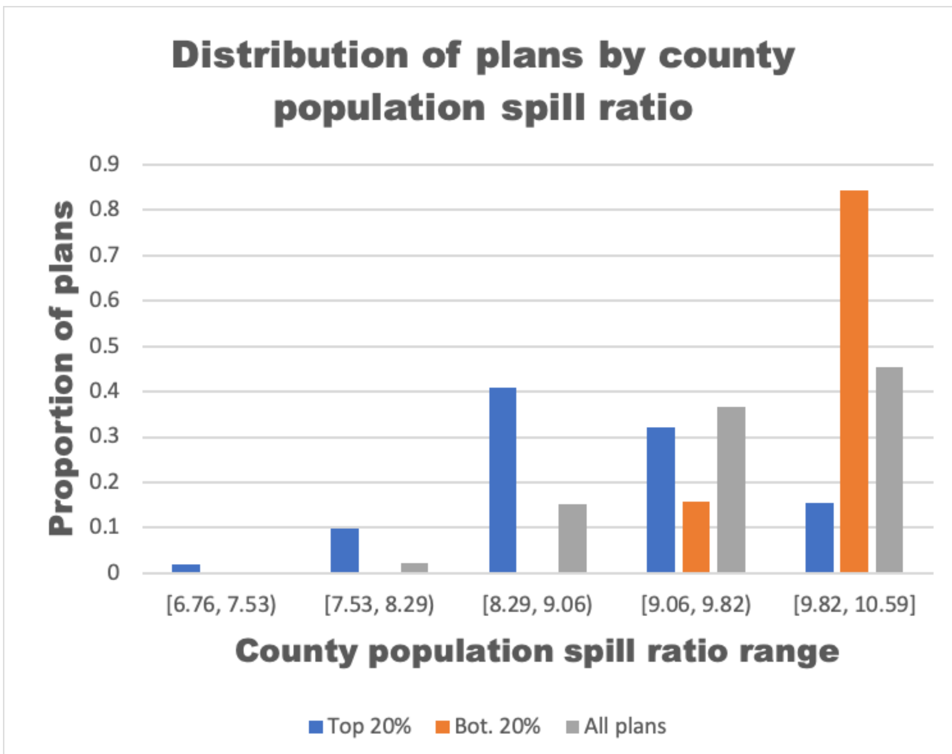

Figure 6: The distributions of the spill ratio as a proxy for the number of split counties.

more in the best bin. Note that larger is better with this measure. Because the two measures of splits look so similar, we calculated the percentage of overlap between the measures among the higher scoring plans for the two measures. Of the 471 plans in the the top 20% on the Hirfindahl measure alone, 362 plans are in the top 20% using the split measure alone, or 77%.

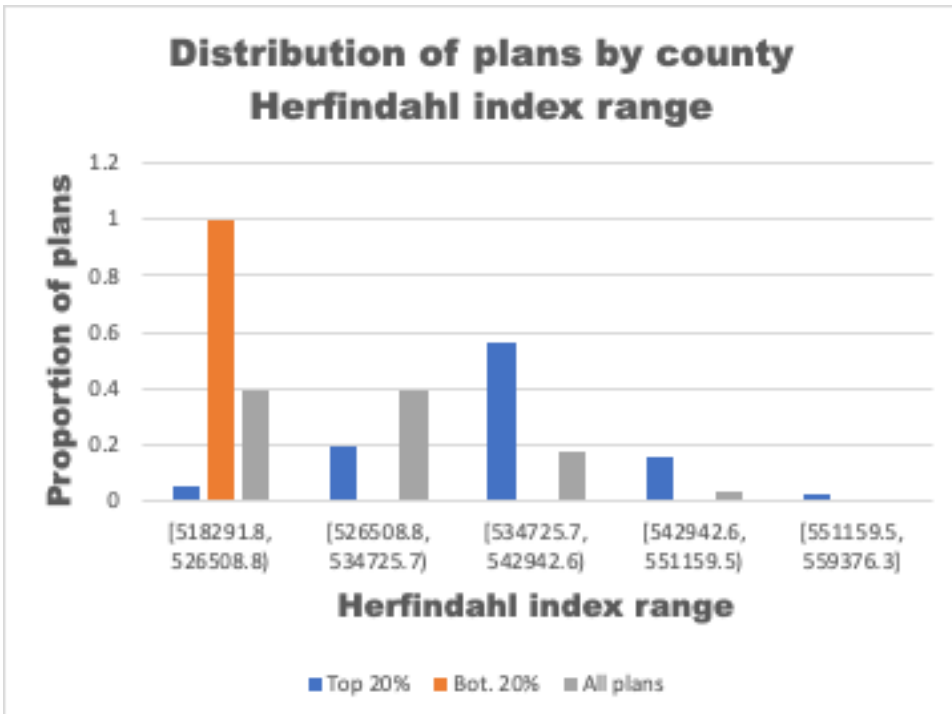

Figure 7: The distributions of the Herfindahl index as a measure of split counties.

Table 2: Average of plan metrics for the five DEA tiers

	DEA TIERS				
Metric average	Bottom 20% DEA	20-40% DEA	40-60 % DEA	60-80 % DEA	Top 20 % DEA
Polsby-poper	0.163	0.165	0.166	0.169	0.175
Moment of inertia	0.709	0.714	0.719	0.726	0.733
Convex hull ratio	0.704	0.707	0.709	0.712	0.717
County splits	65.301	63.781	63.662	62.504	61.282
County population spill ratio	10.181	9.968	9.715	9.476	8.966
Herfindahl index	521960.943	525860.911	528810.776	531877.939	538107.838
Population deviation	12339.355	12167.940	11961.157	11831.790	11563.380
Max. deviation from target population	6655.736	6571.655	6492.303	6438.592	6310.131
Sum of deviations from target population	48884.128	48435.194	46831.945	47322.986	46415.414

Given the level of overlap between the top quintile of plans and the bottom quintile in the previous figures, we have to ask if the DEA model is just imprecise or is actually discriminating between potentially interesting plans and ones that should be discarded. Our final test of the efficacy of the DEA methodology is a table of the *average* value of each of the 9 measures for the DEA-score bins of the plans (developed using just the 5 metrics). *All* of the metrics show improvement as the DEA-score ranges increase. That is, even the metrics that are *not* used in constructing the DEA score improve as the DEA score tier improves. Table (2) of DEA tiers summarizes.

8.1 The relationship between the Pareto frontier and its convex hull

Since the DEA model provides the convex hull of the Pareto frontier, we now examine the extent to which the plans on the Pareto frontier make the cut in the DEA calculations and how the plans on the Pareto frontier can be exploited to speed up computations. A related issue is how close the Pareto frontier is to the convex hull. To examine the relationship between the convex hull and the actual Pareto frontier, we determined all of the plans on the Pareto frontier and compared their DEA scores with the plans in the top quintile. Using the 2,355 plans, we have 79 plans on the Pareto frontier. Included in this total are the plans with DEA scores of 1, as plans on the convex hull of the Pareto frontier are also on the frontier. The DEA cutoff score for the top quintile of the 2,355 plans is 0.96827. All of the 79 Pareto plans have DEA scores above this number.

The worst DEA score is 0.9740 and its DEA rank position is 79 in the set of 2,355 plans, or well within the top 5% of plans. This DEA score is close to the convex hull and shows that the points on the Pareto frontier are close to the convex hull. Thus, retaining plans in the top 20% does not cut off Pareto plans in our collection of plans. Furthermore, the narrow range of DEA scores does not seem to reduce the discriminatory power of the DEA score.

9 Further directions for research

If a plan is on the Pareto frontier, there is no existing plan that is better in all measures. Consequently, there is a cone starting at this plan with faces parallel to the axes that contains no plans. For example, in Figure 1 there are no points in the quadrant that runs below and to the left of plan 7. This quadrant is the cone in two dimensions. Consequently, the Pareto frontier does not have sharp peaks or valleys. When a Pareto point has a low DEA score and is far from the surface of the convex hull, there is a large area on and near the surface of the convex hull with no viable plans in the generated set.

One can use this information to bias a heuristic search for more plans to fill this hole if it exists. The plans that are in the *basis* of the DEA linear program surround the hole and can be used as starting points in a heuristic search to construct plans that reduce the size of the hole or guide humans in constructing new maps. Developing the search heuristics is a potential research area. Potentially good methods would include genetic algorithms that breed solutions as in Xiao et al. (2002) or tabu search.

The Pareto plans can be useful for accelerating the determination of the near Pareto points as follows. The only constraints that affect the DEA score of any plan are the ones defined by the plans on the convex hull and these are also on the Pareto frontier. We can therefore find all plans on the Pareto frontier and form the DEA constraints with just those plans, solving a smaller model to find the plans that define the convex hull of the Pareto frontier. Once the plans on the convex hull are known, we can use the DEA model with just these plans in the constraints. We then solve this DEA model for every plan using the same constraints and objective function as in the original formulation but with far fewer activities.This reduced model determines the same DEA scores as the larger model above, but with much less computer time. Furthermore, this formulation makes the solution time linear in the number of plans and moreover, the linear programs can be run in parallel, allowing for the processing of large numbers of plans simultaneously. Thus, the literature on fast methods for finding the Pareto frontier can also accelerate the construction of the DEA scores for *all* plans. Moreover, Pareto plans can also be the basis for judging mapping contests and for comparing human and computer judgment of maps (Chou et al. (2014)).

Currently, a major research area in redistricting uses Monte Carlo Markov chains (MCMC) to generate billions of maps. The idea has been to use the probability distributions of these maps to estimate the probability a map would have certain characteristics, such as the number of majority minority seats in Congress. If a map is an outlier in the generated distributions, the researchers contend the map is gerrymandered, and if it is not an outlier, it is not gerrymandered. The recent Supreme Court decision in the Alabama case (21-1086 Allen v. Milligan (06/08/2023)) has added to the discussion of the MCMC approach and may limit any future use. Here is what Justice Roberts wrote:

“Here, Alabama contends that because HB1 sufficiently ‘resembles’ the ‘raceneutral' maps created by the State's experts—all of which lack two majority-black districts—HB1 does not violate §2. Alabama's reliance on the maps created by its experts ... is misplaced because those maps do not accurately represent the districting process in Alabama. Regardless, the map-comparison test that Alabama proposes is flawed in its fundamentals. Neither the text of §2 nor the fraught debate that produced it suggests that 'equal access' to the fundamental right of voting turns on technically complicated computer simulations. Further, while Alabama has repeatedly emphasized that HB1 cannot have violated §2 because none of plaintiffs' two million odd maps contained more than one majority-minority district, that (albeit very big) number is close to irrelevant in practice, where experts estimate the possible number of Alabama districting maps numbers is at least in the trillion trillions."

That is, the Court rejects at this time the use of probability distributions of randomly generated maps in districting. This large population of maps can, however, be used to construct an approximation to the Pareto frontier using the methods described in the literature review. The resulting frontier could be used as is. The selected points can also be run through a set of heuristics described in some MCMC papers and other papers to incrementally improve the maps and the estimate of the frontier.

Alternatively, the points generated could be used as the initial points in the algorithm presented above to remove all the points they dominate, getting rid of almost all of the non-Pareto points early on as part of applying a standard algorithm for finding Pareto points.

If the MCMC methods and improvement heuristics generate points along the Pareto surface and a skyline algorithm is used to sort them out, a superior estimate of the Pareto frontier of the space of maps could be generated without needing human-generated maps, For this approach to work, a few hundred of the billions of maps generated have to be near the Pareto frontier. Determining whether this is the case requires having a frontier from a large collection of human generated maps for comparison with the MCMC generated frontier.

We have made the assumption that a few thousand plans would be submitted to a redistricting commission's website. However, a spammer could submit large numbers of maps to clog the system and hinder real public comment. This already happens at the Federal Energy Regulatory Commission during the commenting phase of regulatory rule making. One possibility is to apply clustering methods. If a very large portion of maps are clustered together, spamming is likely. If not, other methods to ferret out the spam maps are necessary. An arms race between spammers and spam filters is likely.

10 Conclusions

Our goal is to develop efficient computational methods for filtering good plans for further evaluation, not to select a “best plan.” Thus, we are not taking a position on what criteria should be used or applying value judgments on the relative weights of various criteria. We are looking for ways to mitigate the potential impact of spammers on an open redistricting process and support the evaluation of citizen-submitted maps if shadow commissions and other advocacy groups succeed in generating significant citizen participation, as has, for example, the Committee of 70 in Pennsylvania (https://drawthelinespa.org/). Gopalan (2020) also outlines a framework for increasing citizen participation in general.

The method for forming a consideration set of good plans should respect the different definitions of “good” among different groups of citizens. Consequently, we want to use multiple measures when evaluating plans and we want to allow for different weights among the measures to make sure that we include plans that represent the multifarious views of diverse citizens.

We have developed a filtering mechanism by identifying plans for further consideration that are close to the convex hull of the Pareto frontier. How close each plan is to the frontier is a key feature of our method. Calculating the proximity of each plan requires discovering the weights on each measure that provide the best score, which would be difficult without using a linear program that is similar to ours.

The fundamental problem in designing a plan that is good by any reasonable definition is that doing so requires tradeoffs of the different criteria to balance redistricting goals. Requiring strict population equality increases the number of political-unit splits. Requiring more compact districts can also lead to more splits, especially if the political units are not compact. One can even find tradeoffs with contiguity, because contiguity actually has multiple definitions. It can be defined as “rook” contiguity, requiring a shared edge of nonzero length, or “queen” contiguity, which also allows touching at only one point. Stricter rook contiguity, which we use here, makes it more likely other measures would have to be relaxed.

The overall pattern is that our method seems to work well in reducing the number of plans that need to be examined, keeping plans that are highly rated in some combination of measures even when not strong in other measures. The ability to keep some plans that are not strong on any measure but strong on a mix of measures is useful for understanding the tradeoffs that need to be made when constructing a good plan. Because our method allows the weights of a measure to be extremely low, a plan can be extremely low on a given measure but considered a good candidate for further evaluation. To compensate for this, in our example we have chosen to eliminate all plans that are below the median in any single measure. This has the advantage of getting the benefits of two polar opposite ways of selecting plans.

Analytical criteria for evaluating plans have traditionally been used to ferret out gerrymandered plans. When looking for gerrymandering, Young (1988) showed through examples that the common measures for the criteria can all be subverted through clever designs of gerrymandered districts. Yet, none of his gerrymandered plans can pass all of the proposed criteria. Nagle and Ramsay (2020) also find that using multiple measures improves the ability to assess partisan fairness. Consequently, using multiple definitions for some criterion could be useful for determining a measure of gerrymandering.

We have found a different phenomenon. The more measures used for a criterion the narrower the range of the scores on plans. Thus, the discriminatory power of the DEA model is probably diminished as more measures are added. The narrowing is rapid even when the measures are highly correlated. The reason is that each measure adds an extra dimension to the surface of the convex hull and noise in the measures can create opportunities for a plan to be on the frontier.

Since this method is new, it needs further testing and development. None of the parameters we have used, such as the 20% cutoff, should be taken as fixed. The appropriate parameters are probably specific to each situation. That is, when this approach is applied, the cutoff parameters should be determined by experimenting with an initial set of plans and testing a sample of plans around various cutoffs. Furthermore, what steps should be taken depends on the number of maps that are submitted. An initial screen using a minimum for each criterion is optional but could be a helpful step for dealing with extremely large numbers of submissions.

On a final note, our model can be adapted to the environmental literature as follows: when separate input constraints are added to our formulation, the DEA model can provide a more precise measure of the efficiencies of polluting facilities by applying the *same* convex combination of DMU's to both the positive and negative outputs.

Acknowledgement: John Nagle provided useful comments that helped us improve a previous draft. Note: We have also provided a small numerical example in an online appendix. Please refer to Gopalan et al. (2023).

References

Altman, M. and Mcdonald, M. (2011). Bard: Better automated redistricting. *Journal of Statistical Software*, 42:1 – 28.

Belton, V. and Vickers, S. (1993). Demystifying DEA-A Visual Interactive Approach Based on Multiple Criteria Analysis. *The Journal of the Operational Research Society*, 44(9):883–896.

Bullock, III, C. S. (2010). *Redistricting: The Most Political Activity in America*. Rowman & Littlefield Publishers, Inc., Lanham, MD.

Butler, D. and Cain, B. (1985). Reapportionment: A Study in Comparative Government. *Electoral Studies*, 4(3):197–213.

Cervas, J. and Grofman, B. (2020). Tools for identifying partisan gerrymandering with an application to congressional districting in pennsylvania. *Political Geography*, 26:1–12.

Chen, J. and Cottrell, D. (2016). Evaluating partisan gains from Congressional gerrymandering: Using computer simulations to estimate the effect of gerrymandering in the U.S. House, , 44 329-340. *Electoral Studies*, 44:329–340.

Chen, J. and Rodden, J. (2015). Cutting Through the Thicket: Redistricting Simulations and the Detection of Partisan Gerrymanders. *Election Law Journal*, 14(4).

Cherchye, L., Moesen, W., Rogge, N., and Van Puyenbroeck, T. (2007). An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research*, 82:111–145.

Cherchye, L., Moesen, W., and Van Puyenbroeck, T. (2004). Legitimately diverse, yet comparable: on synthesizing social inclusion performance in the EU. *Journal of Common Market Studies*, 42:919–955.

Cho, W. K. and Liu, Y. Y. (2016). Toward a Talismanic Redistricting Tool: A Computational Method for Identifying Extreme Redistricting Plans. *Election Law Journal*, 15(4).

Chou, C., Kimbrough, S., Murphy, F., Sullivan-Fedock, J., and Woodard, C. (2014). On empirical validation of compactness measures for electoral redistricting and its significance for application of models in the social sciences. *Social Science Computer Review*, 32(4):534–542.

Cook, W. and Seiford, L. (2009). Data envelopmenet analysis (DEA)-30 years on. *European Journal of Operational Research*, 192(1):1–17.

Cook, W., Tone, K., and Zhu, J. (2014). Data envelopment analysis: Prior to choosing a model. *Omega*, 44:1–4.

Cooper, W., Seiford, L. M., and Zhu, J. (2011). *Handbook on data envelopment analysis*. Springer, second edition.

Cooper, W. W., Seiford, L. M., and Tone, K. (2007). *Data Envelopment Analysis*. Springer, second edition.

Dakpo, K., Jeanneaux, P., and Latruffe, L. (2016). Modeling pollution-generating technologies in performance benchmarking: recent developments, limits and future prospects in the nonparametric framework. *Eur. J. Oper. Res.*, 250:347–359.

Dakpo, K., Jeanneaux, P., and Latruffe, L. (2017). Greenhouse gas emissions and efficiency in french sheep meat farming: A nonparametric framework of pollution adjusted technologies. *European Review of Agricultural Economics*, 44(1):33– 65.

de-los-Cobos-Silva, S. G., Gutiérrez-Andrade, M. A., Rincón-García, E. A., Mora-Gutiérrez, R. A., Lara-Velázquez, P., and Ponsich, A. (2017). FuGA, a Fuzzy Greedy Algorithm for Redistricting in Mexico. *Fuzzy Economic Review*, 22(2):5–27.

Doyle, J. and Green, R. (1993). Data envelopment analysis and multiple criteria decision making. *Omega*, 21(6):713–715.

Engstrom, E. J. (2016). *Partisan Gerrymandering and the Construction of American Democracy*. Legislative Politics & Policy Making. University of Michigan Press, Ann Arbor, MI.

Førsund, F. (2018). Multi-equation modelling of desirable and undesirable outputs satisfying the materials balance. *Empir. Econ.*, 54(1):67–99.

Geneletti, D. (2019). *Muticriteria Analysis for Environmental Decision-Making*. Anthem Press, London.

Gopalan, R. (2020). Designing information ecosystems to process citizen input and improve public sector decision making. *International Journal of Applied Decision Sciences*, 13(4):387–416.

Gopalan, R., Hachadoorian, L., Kimbrough, S. O., and Murphy, F. H. (2023). Bisection search algorithm:. *Online Appendix*. `https://drive.google.com/drive/folders/1tZ1FLmNJ0KHRr-3uSPV6hDYXaEsFJpr1?usp=sharing`.

Gopalan, R., Kimbrough, S. O., Murphy, F. H., and Quintus, N. (2013). The Philadelphia districting contest: Designing territories for City Council based upon the 2010 census. *Interfaces*, 43(5):477–489. `http://dx.doi.org/10.1287/inte.2013.0697`.

Greco, S., Ehrgott, M., and Figueira, J. R., editors (2016). *Multiple Criteria Decision Making: State of the Art Surveys*. Springer, 2nd edition.

Grofman, B. (2019). Tests for unconstitutional partisan gerrymandering in a post-gill world. *Election Law Journal*, 18(2):93 – 115.

Haas, C., Hachadoorian, L., Kimbrough, S. O., Miller, P., and Murphy, F. (2020). Seed-fill-shift-repair: A redistricting heuristic for civic deliberation. *PLoS ONE*, 15(9):e0237935. `https://doi.org/10.1371/journal.pone.0237935`.

Horn, D. L., Hampton, C. R., and Vandenberg, A. J. (1993). Practical application of district compactness. *Political Geography*, 12(2):103–120.

Keeney, R. L. (1992). *Value-Focused Thinking: A Path to Creative Decisionmaking*. Harvard University Press, Cambridge, MA.

Keeney, R. L. and Raiffa, H. (1993). *Decisions with multiple objectives : preferences and value tradeoffs*. Cambridge University Press, Cambridge, UK.

Kerstens, K., O'Donnell, C., and Van de Woestyne, I. (2019). Metatechnology frontier and convexity: A restatement. *European Journal of Operational Research*, 275(2):780–792.

Kumar, N., Raichel, B., Sintos, S., and Buskirk, G. V. (2019). Approximating Distance Measures for the Skyline. In Barcelo, P. and Calautti, M., editors, *22nd International Conference on Database Theory (ICDT 2019)*, volume 127 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 10:1–10:20, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Kung, H.T., K., Luccio, F., and Preparata, F. (1975). On finding the maxima of a set of vectors. *Journal of the ACM*, 22(4):469–476.

Li, W., Goodchild, M. F., and Church, R. (2013). An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. *International Journal of Geographical Information Science*, 27(6):1227–1250.

Liu, W., Zhang, D., Meng, W., Li, X., and Xu, F. (2011). A study of dea models without explicit inputs. *Omega*, 39(5):472 – 480.

Liu, Y. Y., Tam Cho, W., and Wang, S. (2016). “PEAR: A Massively Parallel Evolutionary Computational Approach for Political Redistricting Optimization and Analysis. *Swarm and Evolutionary Computation 30 (October): 78–92*, 30(October):78–92.

Lovell, C. and Pastor, J. (1999). Radial DEA models without inputs or without outputs. *European Journal of Operational Research*, 118:46–51.

MacEachren, A. M. (1985). Compactness of Geographic Shape: Comparison and Evaluation of Measures. *Geografiska Annaler. Series B, Human Geography*, 67(1):53–67.

Murty, S. and Russell, R. (2018). Modeling emission-generating technologies: reconciliation of axiomatic and by-production approaches. *Empir. Econ.*, 54(1):7–30.

Murty, S., Russell, R., and Levkoff, S. (2012). On modeling pollution-generating technologies. *J. Environ. Econ. Manage.*, 64:117–135.

Nagle, J. and Ramsay, A. (2020). On measuring two-party partisan bias in unbalanced states. Working paper.

Niemi, R. G., Grofman, B., Carlucci, C., and Hofeller, T. (1990). Measuring compactness and the role of a compactness standard in a test for partisan and racial gerrymandering. *The Journal of Politics*, 52(4):1155–1181.

Wallenius, Jyrki, Dyer, James, Fishburn, Peter, Steuer, Ralph, Zionts, Stanley, and Deb, Kalyanmoy (2008). Multiple Criteria Decision Making, Multiattribute Utility Theory: Recent Accomplishments and What Lies Ahead. *Management Science*, 54(7):1336–1349.

Webster, G. R. (2013). Reflections on current criteria to evaluate redistricting plans. *Political Geography*, 32:3–14.

Winston, W. L. (2003). *Operations Research: Applications and Algorithms*. Duxbury Press, 4th edition.

Xiao, N., Bennett, D., and Armstrong, M. (2002). Using evolutionary algorithms to generate alternatives for multiobjective site-search problems. *Environment and Planning A*, 34(4):639– 656.

Young, H. P. (1988). Measuring the compactness of legislative districts. *Legislative Studies Quarterly*, 13(1):105–115.