

M³ Challenge Champions, Summa Cum Laude Team prize of \$20,000

Montgomery Blair High School - Team #141

Silver Spring, MD

Coach: David Stein

Students: Andrew Das Sarma, Jacob Hurwitz, David Tolnay, Scott Yu

Making Sense of the 2010 Census

Team #141

April 19, 2010

Executive Summary

With so much political and economic interest behind accurate results for the United States Census, the Census Bureau has implemented several strategies for dealing with the particularly pesky problem of undercounting, or the exclusion of certain individuals from the Census: These include sampling the population after the Census to gauge how many people were excluded, guessing values for missing data, and examining public records to estimate the breakdown of the population. Of these, we found that only the last two are sufficiently helpful to merit use, whereas the first strategy of post-Census sampling can lead to error greater than what it was intended to remedy.

Of course, even with perfectly reliable Census results, proper political representation cannot be attained without a system that distributes seats in the House of Representatives in a manner that addresses the particularities of the population. Evaluating six methods (Hill, Dean, Webster, Adams, Jefferson, and Hamilton–Vinton) that Congress has historically considered for dividing seats in the House, we found one (Hamilton–Vinton) that surpasses the others in the arena of fair apportionment.

After Congress, the next bearers of responsibility are the fifty states of the Union, who are constitutionally charged with drawing district lines that demarcate regions for their representatives. In regard to this process, we suggest that states commit to a system that impartially divides the state, according to population density. Such a system, we hold, will ultimately best serve the common good of the state and strengthen the faithfulness with which our representative democracy reflects the sentiments of the American people.

1 Introduction

“The actual Enumeration shall be made...every subsequent Term of ten Years, in such Manner as they shall by Law direct.” — United States Constitution, Article I, Section 2

In 1787, the Founding Fathers set in law the requirement for a decennial enumeration of the country’s population. Today, we call this once-a-decade process the Census, named after a similar practice in Ancient Rome. The original purpose of the Census results was to assist in the apportionment of seats in the House of Representatives and in the collection of direct taxes. The earliest Censuses tried to collect all data on a single day, and, for the most part, they succeeded. However, with a population now over 300 million, the United States’ 2010 Census will take weeks, if not months. When will it end? When can the Census Bureau be sure that everybody has responded?

At some point, the Census Bureau will have to declare data collection to be over and start tabulating results. Anybody who has not replied yet will be left out. Since the Census will record fewer people than actually live in the United States, this is known as **undercounting**. Undercounting may not sound like a big deal, but in today’s world, the stakes are *huge*. For instance, suppose Wyoming has a population great enough to merit two seats in the House of Representatives. There are no well-defined addresses in rural areas, and the sparse layout makes it difficult for an interviewer to travel from house to house. If enough of Wyoming’s citizens go uncounted, the Census Bureau’s “official” statistic for Wyoming’s population may only be large enough to award one seat to the state. In other words, all Wyomingites will be denied *half* of their voice in the House.

Though the Census’s second original purpose, which was to determine how much money the government should *charge* the citizens of each state, has since reached obsolescence with the passing of the 16th Amendment, in its place has arisen a new, more important, purpose: The determination of how much money the government should *give* the citizens of each state. Every year, Congress appropriates more than \$200 billion in federal funding to states based on Census estimates. Money targeted toward black youths, for example, is allocated based on the number of black youths in each state, as recorded by the Census. Since many populations in certain geographic areas are undercounted chronically, there can sometimes be a mismatch between funding and the need for it in these areas. Statistically adjusting for undercounting can help alleviate this problem, as we describe in Section 2.

Even if the Census were adjusted to perfection, some potential problems would still remain. With regard to apportionment of Representatives, the algorithm that Congress uses may not be optimal, even with perfect data. Finding a better method for apportioning these seats is a key component in giving each citizen his or her fair voice. Our suggested method is given in Section 3. Furthermore, once states learn how many seats (and consequently how many Congressional districts) are needed, they still have to draw the actual district lines. Sometimes, states **gerrymander** by intentionally drawing districts to concentrate a particular interest in one district, or to spread a certain interest thinly across all the districts. Congress must eschew inequity by encouraging states to adopt new, fairer methods, creating impartial standards for **boundary delimitation**, as we recommend in Section 4.

2 Census adjustments

2.1 Introduction

The U.S. Census is supposed to be a complete tabulation of every person residing the United States; however, the omission of persons from the Census (known as **undercounting**) has been a perennial (or rather, decennial) problem. Undercounting is of interest to policy-makers and civil society because it afflicts groups at different rates; if poor people, for instance, are more likely to be undercounted than wealthier people, then poorer states will be politically underrepresented compared to their wealthier counterparts. Census data is critical to allocating funding for big-ticket federal programs, so regions with undercounted populations will also receive less federal money than they deserve. Thus, what matters is not the absolute magnitude of the undercount, but rather the differential in undercounting between various groups. The size and nature of the undercount is not precisely known; indeed, if it were, then the entire matter would be moot. In the most recent Census, overcounting due to multiple entries for single persons was also found to be a significant problem. Several statistical methods are currently available to aid in identifying and rectifying erroneous Census counts. So, if undercounting introduces unfairness into the allocation of domestic political and monetary resources, and if methods are available to rectify it, then why is there any debate at all about the advisability of using them?

The various methods of Census data improvement attempt to squeeze more accuracy out of inherently inaccurate numbers on a huge scale, mitigating some sources of error but introducing others. With each method, one must question whether the introduced error is a fair tradeoff for the eliminated error. The potential for minor imperfections with individual entries in small data inputs to snowball into multithousand person population fluctuations is a frightening prospect that is compounded by the inherent impossibility of perfectly surveying each individual even in a fairly small area. Consequently, here we weigh each method's potential risk against its expected reward.

2.2 Assumptions

We make our recommendations without regard to current law. Our findings with respect to the Post-Enumeration Survey (PES) (which uses the sampling method), for instance, were not influenced by the current antisampling legal environment. In general, we also assume that Census data is *reasonably* complete and accurate. Obviously the data is not perfect, since the point of this exercise is to attempt to exorcise errors from the data. However, it is impossible to paint an accurate picture of the U.S. population without fairly complete initial input data.

2.3 Analysis

We consider three principal techniques: Sampling, demographic analysis, and imputation. Each of these techniques finds wide international use; the U.S. Census Bureau currently uses demographic analysis and imputation, but not sampling. Some statisticians claim that the addition of sampling would help make the Census more accurate. However, we agree with

Freedman and Wachter [1], who caution, “Despite what you read in the newspapers, the census is remarkably accurate. Statistical adjustment is unlikely to improve on the census, because adjustment can easily put in more error than it takes out.” We agree that the Census Bureau’s decision to use demographic analysis and imputation but not sampling results in a more accurate Census.

2.3.1 Sampling

Perhaps most prominent among the currently available Census rectification methods is **sampling**, a technique which has been on the political radar since the 1970s [8]. In 1999, the Supreme Court held that federal law prohibits the Census Bureau from using sampling to determine population counts [9], barring the use of sampling to produce the official 2000 population count.

Broadly, sampling involves surveying a relatively small sample of the population after the main Census and cross-referencing survey responses with Census data to approximate the rate of incorrect inclusions or exclusions of people. The sample, which serves as a “second pass” after the Census, is the essence of the technique commonly referred to as capture-recapture. In the United States, the survey (known as the Post-Enumeration Survey, or PES) is implemented as a blocked stratified cluster survey of more than one million people [10]. By essentially surveying a region twice and comparing the results, sampling supposedly gives additional accuracy to results.

To derive an equation for this sampling methodology, we define some variables:

N_F is the final count.

N_C is the initial census count.

N_E is the number of extraneous persons identified during the sampling process who should not have been included in the Census for the region.

N_S is the number of people identified in the sampling survey.

N_M is the number of matches (i.e., people who completed both the Census and the sampling survey).

$P(C)$ is the probability that a person was properly included in the Census.

$P(S)$ is the probability that a person was included in the survey.

We start by rewriting some of these variables in terms of other variables:

$$\begin{aligned} N_M &= P(C \wedge S) \cdot N_F = P(C) \cdot P(S) \cdot N_F, \\ N_C - N_E &= P(C) \cdot N_F, \\ N_S &= P(S) \cdot N_F. \end{aligned} \tag{1}$$

Then, by substituting, we find

$$\begin{aligned}
 N_F &= N_F \cdot \frac{N_F}{N_F} \cdot \frac{P(C)}{P(C)} \cdot \frac{P(S)}{P(S)} \\
 &= \frac{\left(N_F \cdot P(C)\right) \cdot \left(N_F \cdot P(S)\right)}{\left(N_F \cdot P(C) \cdot P(S)\right)}, \\
 N_F &= \frac{(N_C - N_E) \cdot N_S}{N_M}. \tag{2}
 \end{aligned}$$

Equation (2) is a macroscopic approximation of the sampling methodology.

First, the good news: Sampling can be highly effective at estimating population sizes from incomplete data. Running a computer simulation with actual population size $N_A = 1,000,000$ and $P(C) = P(S) = 90\%$ gives $N_F = 999,932$. In other words, even when each individual had only a 90% probability of responding to the Census or the PES, sampling was still able to estimate the population size with 99.99% accuracy. A similar result was obtained for $P(C) = 10\%$ and $P(S) = 20\%$, indicating that sampling is robust with respect to changes in the response rate between the two surveys.

However, there is a flaw: In Equation (1), we assumed that $P(C \wedge S) = P(C) \cdot P(S)$. To see why this is a problem, let us write $P(C) = P(A \vee X)$ and $P(S) = P(A \vee Y)$, where A represents the probability that a person is highly unlikely to complete either survey and X and Y are random noise variables. Obviously, C and S are not independent because A implies both C and S . Using simple Boolean identities and the laws of probability, we can write

$$\begin{aligned}
 P(C \wedge S) &= P\left((A \vee X) \wedge (A \vee Y)\right) \\
 &= P\left(A \vee (X \wedge Y)\right) \\
 &= P(A) + P(X \wedge Y) - P(A \wedge X \wedge Y) \\
 &= P(A) + P(X) \cdot P(Y) - P(A) \cdot P(X) \cdot P(Y), \tag{3}
 \end{aligned}$$

$$\begin{aligned}
 P(C) \cdot P(S) &= P(A \vee X) \cdot P(A \vee Y) \\
 &= \left(P(A) + P(X) - P(A) \cdot P(X)\right) \cdot \left(P(A) + P(Y) - P(A) \cdot P(Y)\right). \tag{4}
 \end{aligned}$$

But note that Equation (3) does not equal Equation (4). Therefore, by the transitive property, $P(C \wedge S) \neq P(C) \cdot P(S)$. In other words, if C and S are not independent, then we can no longer use the identity that we took for granted in Equation (1), and our approximation in Equation (2) no longer holds! Logically, we should assume that C and S are not independent because there is likely to be a good deal of correlation between failing to complete the Census and failing to complete the PES. People in difficult socioeconomic conditions and radical subscribers to antigovernment philosophies, for instance, are both significantly less likely to complete either survey than a typical person. Even though Equation (2) is not perfect, it is the best approximation that we have.

Running a simulation with “correlation parameter” $A = .05$, actual population size $N_A = 1,000,000$, and noise parameter equal to 3%, we get $N_F = 949,896$ using the above formula. It appears that $N_F = (1 - A) \cdot N_A$; indeed, this makes sense since repeat survey nonrespondents appear the same to the Census as nonexistent people. However, the math only becomes more complicated when additional real-world factors, such as the disinclination of some people to respond (but no guarantee), come into play. Sampling-based results are consequently guaranteed to be inaccurate by at least the number of people who purposefully avoid Census surveys, a population that remains uncounted even after sampling is conducted. This effect is known as correlation bias. If the proportion of nonrespondents is known, then the true population can be calculated from N_F ; however, a .001 error in the estimation of A corresponds directly to a .1% change in the estimated population, or 300,000 people. One paper estimates the size of this “doubly uncounted” population as 3 million, but any approximation of this elusive group’s size is highly speculative [11].

A final problem from the use of sampling comes from details of real-life implementation. In order to obtain PES results that are representative of the entire population, it is necessary to conduct the survey on a sample that is carefully controlled in terms of diversity. (Typically, the Bureau has used a stratified block cluster sample [10].) In such an arrangement, there may be over 1,000 strata representing different slices of the populace; assuming about a million participants, the average stratum works out to 1,000 individuals. The “leverage factor” N_C/N_S , the number of individuals across the country represented by each individual in the sample, would be about 300. In such an arrangement, small pieces of questionable data could have large sway over the final outcome. For instance, suppose that there are 100 individuals in the relatively small Montana block of the African-American female stratum. Great Falls, home to more than 10% of Montanan African-Americans, is a cluster site for the PES. An unscrupulous field worker fabricates half of his forms, marking on most of them that the supposed interviewee did not participate in the Census. The N_M for this stratum-block are consequently depressed (and the N_P number may be inflated as well), leading to a potentially drastic overstatement of the Montanan African-American female population, potentially numbering tens of thousands of individuals, as a result of a few dozen fabricated PESs. Such an extreme situation is unlikely in real life, but it demonstrates the pitfalls of adjusting the results of the broad main Census based on data from a relatively small post-survey.

Sampling is a very powerful statistical technique in an ideal survey, but the real-world risk of small pieces of false data causing large swings in population estimates, coupled with sampling’s inability to discern the existence of people wholly averse to a government-operated Census of the entire population, makes it unsuitable for use in producing official Census statistics. A Berkeley statistics professor identifies a 1990 instance in which 13 PES responses caused a 50,000 swing in the undercount estimate [12]. Such situations are uncommon but evidently do exist. The risk of throwing off the entire U.S. population estimate, particularly in smaller demographic categories, does not stack up against the reward of potentially identifying and rectifying some count errors.

2.3.2 Demographic analysis

A second method of calculating discrepancies between Census counts and actual count is Demographic Analysis (DA). Used for every Census since 1960 [13], DA estimates undercount/overcount by comparing Census data with existing aggregate data sets held by the government. These data sets are usually independent of the Census and often include administrative statistics on births, deaths, authorized international migration, healthcare, welfare, etc. [13]. Also included are estimates of legal emigration and net illegal immigration. By analyzing the DA, the Census Bureau develops population benchmarks, with which, via subtraction, we can find the Census net undercount [13]. The internal consistency of DA and its high fidelity with population changes make DA a preferable alternative over statistic sampling. One important distinction must be made, however: DA and sampling are, by nature, different techniques for estimating count discrepancies. Whereas sampling estimates undercount directly with the capture-recapture strategy, DA simply offers baselines from which we can draw conclusions about the completeness of the Census data.

2.3.3 Imputation

The final method, and also the technique most extensively used by the 2000 Census, is **imputation**. Imputation deals with the attribution of data (both demographic and population count) to households or sites that are known to exist but have not been surveyed by the Census. The fundamental premise of imputation is essentially, “I am like my neighbors.” In other words, the idea behind imputation assumes that we can make an educated guess about a household based on information already known about the household’s neighbors. A central question of imputation, then, is how information about a person or housing unit should be guessed based on its surroundings.

There are two categories of imputation: Count imputation and whole-person characteristics imputation. Count imputation affects the Census count of the actual population, while characteristics imputation affects the demographic information of individual people. In the 2000 Census, count imputation added 1.2 million people to the Census total, while characteristics imputation was necessary to complete the personal information of an additional 4.6 million people [5]. Count imputation was needed when, for instance, the Postal Service returned Census materials as undeliverable. Even if a housing unit could not be located, its expected occupancy could be imputed and the personal information of its theoretical residents completed. Imputation differs from sampling because it involves extrapolating known regional demographic data to known (or previously known) residential locations; sampling, on the other hand, necessitates the use of statistical and mathematical methods to estimate a final population from the overlap of two samples. The people added to a Census count via count imputation are added for some reason (such as the Census Bureau knowing that their residences exist but not being able to locate them). The people added to a Census count via sampling are added for no reason other than the prediction of their existence by statistical models. In this way, count imputation is more based in reality than sampling and thus less prone to catastrophic error.

Count imputation is implemented in three stages: First, if a housing unit’s status (i.e., whether it exists and is in habitable condition) is unknown, it is imputed. Next, if the housing

unit is habitable and its occupancy status (occupied or vacant) is unknown, it too is imputed. Finally, if the household is vacant but its size is unknown, the number of residents is imputed. Characteristics imputation is more straightforward than count imputation: essentially, if a person is known (or projected, based on count imputation) to exist but their demographic information is incomplete, the information is imputed based on the demographic trends around the person's home.

It is difficult to come by information that identifies demographic information and residences for individual people; since imputation is all about correlating data on a house-to-house scale, it is consequently difficult to evaluate the effectiveness of imputation on real-world test cases. Larger data sets, such as the map in Figure 1 depicting counties shaded based on the African-American portion of the population, tend to display the same associative clustering (on a macroscopic scale) that imputation assumes on a microscopic scale:

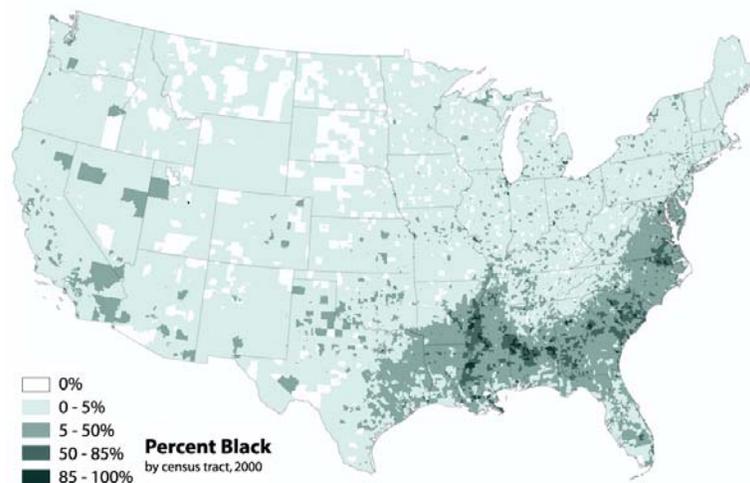


Figure 1: This map depicts counties shaded based on the African-American portion of the population [7].

A potential pitfall of imputation is the repetition of “oddball” individuals: for instance, if the Census Bureau imputes Bob’s personal data by replicating the data of his nearest neighbor, but Bob happens to live next to William Gates, the imputation process will have introduced a good deal of error. As a less extreme example, a neighborhood with an atypically low response rate and a correspondingly high imputation rate may have the same person’s data imputed multiple times to represent several neighbors. Even if that person is fairly typical, the repetition of a single person’s data several times is not good design and is indicative of potential problems. (What if Bill Gates’s data were repeated four times? What kind of per capita income would that ZIP code have?) Fairly simplistic imputation methods, such as sequential hot deck (which is essentially nearest-neighbor matching, possibly with correlation of known demographic data), suffer from potential multiple repetition of individuals and also potential repetition of oddball individuals. However, a more advanced technique that examines multiple neighbors and repeats a median value could mitigate such concerns [6].

More advanced forms of imputation use more complex models that also allow imputation error to be estimated and bounded. Since imputation extrapolates information based on the existence of real people (or real housing structures, at least), it is less subject to error from statistical and survey artifacts than sampling. In terms of manpower and cost, imputation is quite efficient since it is mostly accomplished by computer analysis. The greatest danger from imputation is the repetition of “oddball” individuals or the multiple repetition of individuals in areas with spotty data; modern, high-power computers are capable of using advanced cross-sampling spatial analysis algorithms to mitigate these effects. Imputation is a cost-effective way to improve the detail and accuracy of spotty Census data within a known error bound and is thus advisable for use within the Census.

3 Apportionment method

3.1 Introduction

As described in Section 1, one of the primary purposes of the U.S. Census is to determine the number of people in each state so that seats in the House of Representatives can be apportioned among the states fairly. (How the states then choose to district their seats is detailed in Section 3.) After every Census, Congress uses the population of each state to compute the number of seats each state will receive. However, what algorithm they should use is under contentious debate. Currently, Congress uses what is known as the equal proportions (or Hill) method. A 2001 report by the Congressional Research Service [14] suggested five alternative methods without reaching a conclusion as to which is the “best” or “most fair.” In this section, we examine these six methods (the Hill method plus the five alternatives) quantitatively to determine which method is statistically superior.

3.2 Goals

In order to create a model, we must precisely define what the goal of our model is. In other words, what constitutes a fair reapportionment method?

Goal 3.1. *We want to choose a reapportionment method that, in the long run, is indistinguishable from assigning seats proportionally to states’ populations.*

What does this mean? In one reapportionment cycle, it is impossible to assign seats exactly proportionally to populations because seats come in whole numbers. Imagine a simple country with only two states of equal population and 435 seats to distribute. Each state should receive the same number of seats, but there are an odd number of seats total. One state will be cheated out of half a seat and receive only 217, while the other will receive a bonus half seat to make 218 total. This isn’t fair! Luckily, we can make it fair. The next time the seats are reapportioned, switch the “bonus” seat to the other state. If we repeat this process over many reapportionment cycles, it will be fair *on average*.

Unfortunately, this goal alone is not enough. What if we assigned all 435 seats to the same state (and 0 to the other state), and switched all 435 seats to the other state every 10 years? This is perfectly fair in the long run but horribly unfair in the short term. This leads us to another goal.

Goal 3.2. We want to choose a reapportionment method that, in the short term, is as close as possible to assigning seats proportionally to states' populations.

How can we quantify whether a method achieves these goals? Imagine making a scatterplot of all 50 states, where the x -coordinate represents the proportion of the total U.S. population living in that state, and the y -coordinate represents the proportion of the total House of Representatives seats apportioned to that state. A state with 1% of the total population should receive 1% of all seats, and a state with 10% of the total population should receive 10% of all seats. In other words, our target is the line $\hat{y} = x$. However, the actual y values will not lie on this line. If we have n data points (that is, n states), then the standard error of the actual statistics about the line is

$$SE = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}. \quad (5)$$

We know that $n = 50$ and $\hat{y} = x$. Additionally, $\text{pop}_{\text{state}}/\text{pop}_{\text{total}}$ and $y = \text{seats}_{\text{state}}/435$.

$$SE = \sqrt{\frac{\sum \left(\frac{\text{seats}_{\text{state}}}{435} - \frac{\text{pop}_{\text{state}}}{\text{pop}_{\text{total}}} \right)^2}{50}}. \quad (6)$$

For instance, we can calculate the standard error for the actual 2000 apportionment. Using Census-provided data for the apportionment population and number of apportioned representatives in each state [15], we find that $SE \approx 0.000637$. Figure 2 confirms this small standard error because we can visually see that all of the data points lie very close to the line $\hat{y} = x$.

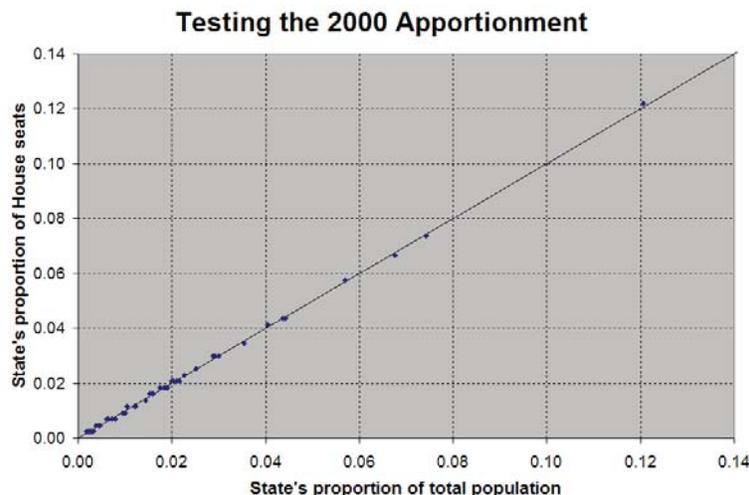


Figure 2: A scatterplot of the 2000 apportionment. The data lies very close to the expected line, meaning there is a small standard error.

3.3 Model

Rather than working with real Census data, we worked with a model of the United States. This model made several simplifying assumptions.

Assumption 3.1. *The number of states in the United States will remain fixed at 50.*

Assumption 3.2. *The number of seats in the United States House of Representatives will remain fixed at 435, as it has for 9 of the last 10 decades [16].*

Assumption 3.3. *The total population of the United States will remain between 300 million and 400 million.*

One trial run of the program consists of randomly picking a population between 300 million and 400 million and then distributing these among the 50 states. For this step, we used two methods: A uniform distribution that assigns roughly the same number of people to each state, and a power-law distribution that assigns a small number of people to many states and a large number of people to very few states. We calibrated our power-law distribution so that the expected values for the most and least populous states resemble the actual populations of the most and least populated states in the United States, as of 2000.

After distributing the population into states, we apportioned the 435 House of Representatives seats to the 50 states using 6 different methods. The following definitions are adapted from [14].

Method 3.1 (Adams). *Use a binary search to find a number so that, when it is divided into each state's population and resulting quotients are rounded up for all fractions, the total number of seats will sum to 435. In all cases where a state would be entitled to less than one seat, it receives one anyway because of the constitutional requirement.*

Method 3.2 (Dean). *... are rounded at the harmonic mean, the total number of seats...*

Method 3.3 (Hill). *... are rounded at the geometric mean, the total number of seats...*

Method 3.4 (Webster). *... are rounded at the arithmetic mean, the total number of seats...*

Method 3.5 (Jefferson). *... are rounded down for all fractions, the total number of seats...*

Method 3.6 (Hamilton–Vinton). *First, the population of 50 states is divided by 435 in order to find the national “ideal size” district. Next this number is divided into each state's population. Each state is then awarded the whole number in its quotient (but at least one). If fewer than 435 seats have been assigned by this process, the fractional remainders of the 50 states are rank-ordered from largest to smallest, and seats are assigned in this manner until 435 are allocated.*

For each of the two population distribution functions, we ran the program over many trials and calculated the standard error for the apportionment determined by each of the six methods.

3.4 Results

For the uniform distribution, the averages of the standard errors over 1,000,000 trials were as follows:

Method	Mean Std Err
Hamilton–Vinton:	0.000782851
Adams:	0.001348290
Dean:	0.000839477
Hill:	0.000812758
Webster:	0.000795882
Jefferson:	0.001124880

To create a more manageable data set, we reduced the number of trials to 1,000 and computed standard deviation as well:

Method	Mean Std Err	Std Dev
Hamilton–Vinton:	0.0007854	0.0000799
Adams:	0.0013489	0.0001859
Dean:	0.0008423	0.0001020
Hill:	0.0008150	0.0000945
Webster:	0.0007987	0.0000884
Jefferson:	0.0011269	0.0001314

An ANOVA test at the 5% significance level showed significant evidence that these means are *not* the same ($p \approx 0.000$). It appears that the Hamilton–Vinton has the lowest mean, so we performed five 2-sample matched-pair one-sided t-tests, one comparing Hamilton–Vinton to each of the other five methods. All five showed significant evidence at the 5% significance level ($p \approx 0.000$) that the Hamilton–Vinton test has a lower standard error than the other tests. This means that, if the Hamilton–Vinton method were not the best and we conducted many many trials, we would expect to see results as extreme as these only 5% of the time.

For the power-law distribution, the data after 1,000,000 trials was as follows:

Method	Mean Std Err
Hamilton–Vinton:	0.000802677
Adams:	0.001531700
Dean:	0.000884385
Hill:	0.000848855
Webster:	0.000824826
Jefferson:	0.001214880

The results of the ANOVA test and t-tests were similar. In other words, the Hamilton–Vinton method appears to be the best method for apportioning seats fairly, given the restraints of our model. However, not all real-world situations will obey the assumptions we made. To truly test this model, we would want to test the six methods on actual state population data.

4 Fair redistricting

When the Founding Fathers crafted the U.S. Government, they purposefully constructed our institutions to guard against the unrestrained despotism that had driven the first pilgrims

to flee the Old World. With the doctrine of “dual spheres of sovereignty,” the Framers decentralized power, splitting governing duties between the federal government and the states. Thus, although the government apportions the seats in the House of Representatives, the responsibility of boundary delimitation rests with the state legislatures. Redistricting, therefore, has always been politically contentious because the party in power has the ability to fix the political makeup of the state, possibly for the next ten years. State legislatures have often proven to be incredibly nimble in their redistricting, carving districts of bizarre shape in order to concentrate a particular interest (a practice called **gerrymandering**). Since the whole point of a representative democracy is that interests find a voice through the legislature, and since so much political tension invariably accompanies the decennial reapportionment, fairer standards for redistricting would not only produce a more complete representation of American citizens in government but also obviate the occasion for partisan bickering and needless tying up of our judicial system by politically motivated suits.

The particular algorithm we proffer is an impartial drawing of district lines, based on where people live, but without regard to who they are. The reason for this is simple: Redistricting has become a huge, involved task that, in many states, drastically changes existing boundaries and local constituencies, without regard to the needs of the local communities. An impartial algorithm is an equitable and appealing alternative because, in the long run, no party benefits or suffers as the demographic composition changes with time.

A moral justification for accepting impartial boundary delimitation comes from the concept of the “original position,” a philosophy first outlined by John Rawls in his 1971 magnum opus *A Theory of Justice*. The original position specifies a hypothetical situation in which all parties are behind what Rawls called a “veil of ignorance” [2]. Rawls’ principle argument, then, was that the choices made in such a disinterested scenario comprise the truly equitable and just resolution.

Accordingly, in such a Rawlsian situation, political parties have no knowledge of any factors that would contribute to any partiality in decision-making; that is, they know nothing of the racial composition of the state, the distribution of income, geographic party affiliation, political majority, etc. In such a situation, neither party would be so foolish as to gerrymander and opt for specifically shaped districts because such an arrangement can just as likely help as hurt. In fact, understanding that the demographics of the state are, in a sense, subject to chance, each party would *want* to subscribe to an impartial standard, or, at the very least, hold no objection to one; for impartial redistricting will tend to mirror the state’s sociopolitical composition in the long run, so a party would possess rightful clout with the support of a majority, yet still retain representation if it should draw support from a minority.

Moving away from philosophy, one reasonable question is whether an impartial mechanism for redistricting is at all feasible in the real world. To address this, we created a model that forms districts for Texas, based on population data by county from the 2000 Census [3, 4] and assuming homogenous distribution inside each county. The districting algorithm moves laterally across the state, dividing the region into columns. During this process, two factors are under consideration: The number of residents encompassed by the column and the population density of each district. Columns contain a multiple of the number of individuals in a district that would be found in each district if the state’s populace were to be uniformly divided. However, the size of districts is bounded by population density, which we

prevent from deviating too much from that of the entire state. Thus, in “thicker” regions of the state, the algorithm establishes more districts per column; in areas that are rural and sparsely populated, the algorithm creates districts that are generally larger in size.

Consequently, in Figure 3, the rectangles correspond to the districts created via our algorithm; each point corresponds to approximately 3000 people. As expected, smaller districts correspond to higher population densities, and larger districts to lower population densities. In particular, notice that some of the smallest districts correspond to the Lone Star State’s largest urban centers, namely, Austin, Dallas, and Houston.

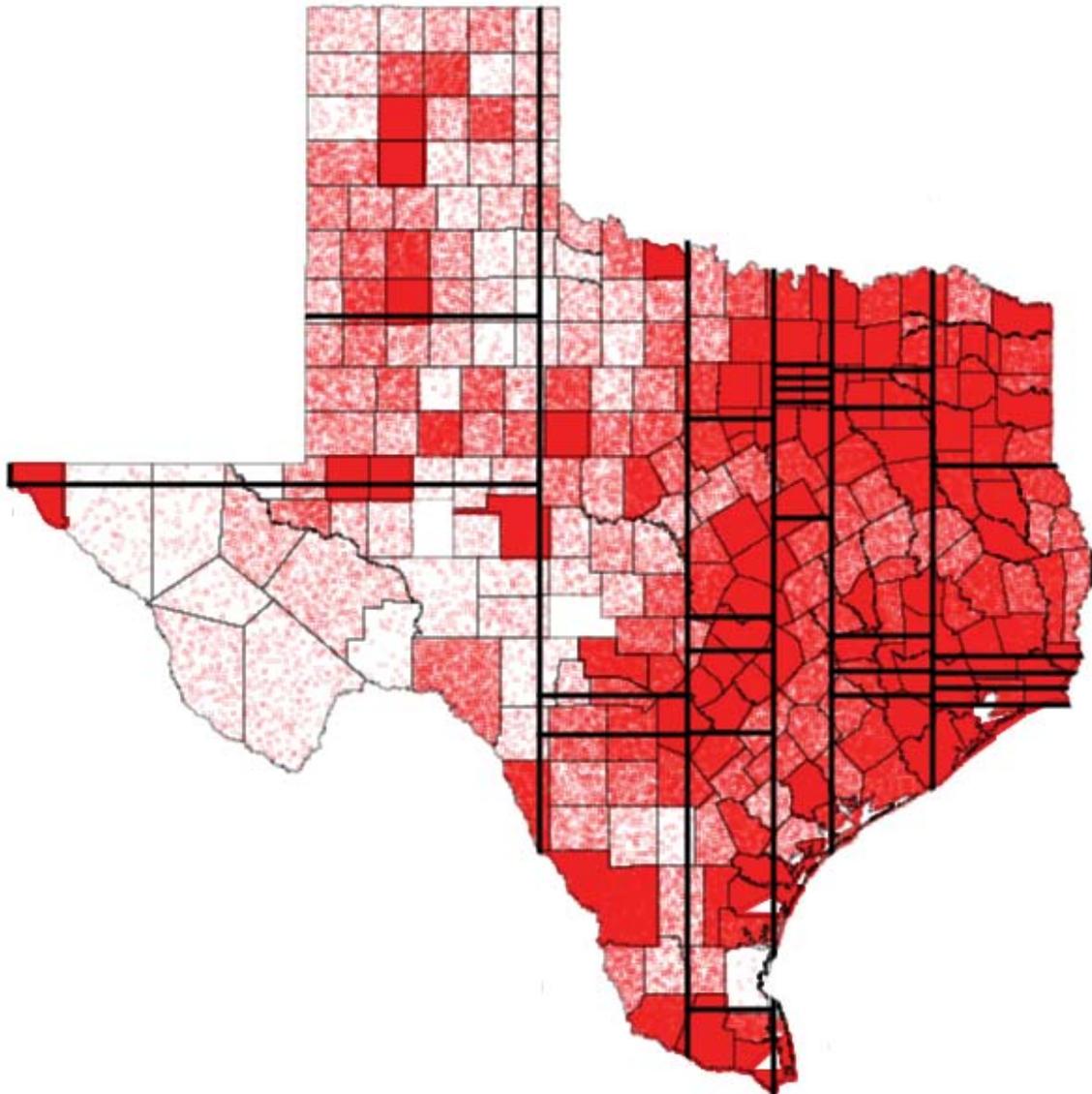


Figure 3: An illustration of the proposed districting method applied to the state of Texas. Red illustrates population density based on the 2000 Census.

References

- [1] D. A. Freedman and K. W. Wachter. *On the likelihood of improving the accuracy of the census through statistical adjustment*, in *Statistics and Science: A Festschrift for Terry Speed*, IMS Lecture Notes Monogr. Ser. 40, Inst. Math. Statist., Beachwood, OH, 2003, pp. 197–230.
- [2] J. Rawls, *A Theory of Justice*, Belknap Press, Cambridge, MA, 1971.
- [3] County boundary data from Texas Tech University.
http://cgstftp.gis.ttu.edu/Texas/TxCountyData/Shapefiles/TxCntyLayers/TxCnty_Bndry_ESRI/
- [4] County population data from University of Texas at San Antonio.
<http://txsdc.utsa.edu/txdata/dp1/>
- [5] Analysis of 2000 Census imputations, U.S. Census Bureau.
<http://www.census.gov/dmd/www/pdf/Report21.PDF>
- [6] Research to improve census imputation methods, ASA section on survey research methods. <http://www.census.gov/dmd/www/pdf/Report21.PDF>
- [7] Image of the distribution of the U.S. African-American population.
http://image.absoluteastronomy.com/images/encyclopediainages/n/ne/new_2000_black_percent.gif
- [8] Information about the history of sampling in the U.S. Census.
<http://www.scienceclarified.com/dispute/Vol-2/Should-statistical-sampling-be-used-in-the-United-States-Census.html>
- [9] Dept. of Commerce vs. U.S. House of Representatives (1999) Supreme Court case summary. http://www.oyez.org/cases/1990-1999/1998/1998_98_404/
- [10] An illuminating report on various sampling techniques and their applications from the Statistics Department at Stanford University. <http://www.stat.berkeley.edu/~census/sample.pdf>
- [11] A paper by Wachter and Freedman addressing double-undercounting resulting from correlation bias. <http://erx.sagepub.com/cgi/content/abstract/24/2/191>
- [12] A statistician's perspective on why sampling in the Census is a bad idea, full of provocative figures. http://www.stat.berkeley.edu/~stark/Seminars/bbc98.htm#why_adjust
- [13] Demographic analysis results. <http://www.census.gov/dmd/www/pdf/Report1.PDF>
- [14] House of Representatives apportionment formula. <http://www.rules.house.gov/archives/RL31074.pdf>
- [15] 2000 Census data. <http://www.census.gov/population/www/cen2000/maps/files/tab01.xls>
- [16] Size of the House of Representatives. http://www.nationalatlas.gov/articles/boundaries/a_conApport.html