Census TopDown: The Impacts of Differential Privacy on Redistricting

Aloni Cohen ✉
Hariri Institute for Computing and School of Law, Boston University, MA, USA

Moon Duchin ✉
Department of Mathematics, Tufts University, Medford, MA, USA

JN Matthews ✉
Tisch College of Civic Life, Tufts University, Medford, MA, USA

Bhushan Suwal ✉
Tisch College of Civic Life, Tufts University, Medford, MA, USA

Abstract

The 2020 Decennial Census will be released with a new disclosure avoidance system in place, putting *differential privacy* in the spotlight for a wide range of data users. We consider several key applications of Census data in redistricting, developing tools and demonstrations for practitioners who are concerned about the impacts of this new noising algorithm called TopDown. Based on a close look at reconstructed Texas data, we find reassuring evidence that TopDown will not threaten the ability to produce districts with tolerable population balance or to detect signals of racial polarization for Voting Rights Act enforcement.

2012 ACM Subject Classification Security and privacy; Applied computing → Law; Applied computing → Voting / election technologies

Keywords and phrases Census, TopDown, differential privacy, redistricting, Voting Rights Act

Digital Object Identifier 10.4230/LIPIcs.FORC.2021.5

Supplementary Material *Text (Documentation)*: `https://mggg.org/DP`
Software (Source Code): `https://github.com/mggg/census-diff-privacy`
archived at `swh:1:dir:aeb9390dfc99f5bc755546b21a1e9ddaecabb65e`

Funding This project was supported on NSF OIA-1937095 (Convergence Accelerator) and by a grant from the Alfred P. Sloan Foundation.
Aloni Cohen: NSF CNS-1414119; NSF CNS-1915763; DARPA HR00112020021.
Moon Duchin: NSF DMS-2005512

Acknowledgements Authors are listed alphabetically. We thank Denis Kazakov, Mark Hansen, and Peter Wayner. Kazakov developed the reconstruction algorithm as a member of Hansen's research group. Wayner guided our deployment of TopDown in AWS and was an invaluable team member for the technical report. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our funders.

1 Introduction

A new disclosure avoidance system is coming to the Census: the 2020 Decennial Census releases will use an algorithm called TopDown to protect the data from increasingly feasible *reconstruction attacks* [2]. Census data is structured in a nesting sequence of geographic units covering the whole country, from nation at the top to small *census blocks* at the bottom. TopDown starts by setting a *privacy budget* $\varepsilon > 0$ which is allocated to the levels of a designated hierarchy, then adding noise at each level in a *differentially private* way [12]. When $\varepsilon \to \infty$, the data alterations vanish, while $\varepsilon \to 0$ yields pure noise with no fidelity to the input data. The algorithm continues with a post-processing step that leaves an output dataset that is designed to be suitable for public use.

2nd Symposium on Foundations of Responsible Computing (FORC 2021).
Editors: Katrina Ligett and Swati Gupta; Article No. 5; pp. 5:1–5:22
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Redistricting is the process of dividing a polity into territorially delimited pieces in which elections will be conducted. The Census has a special release – named the PL 94-171 after the law that requires it – that reports the number of residents in every geographic unit in the country by race, ethnicity, and the number of voting-age residents [9]. The 2020 release is slated to occur by September 2021, after which many thousands of district lines will be redrawn: not only U.S. Congressional districts, but those for state legislatures, county commissions, city councils, and many more.

Many user groups have expressed concerns about the effects of differential privacy on redistricting. They largely but not exclusively concern two issues. First, "One Person, One Vote" case law calls for balancing population across the electoral districts in a jurisdiction, whether small like city council districts or large like congressional districts. Most states balance congressional districts to within one person based on Census counts. Second, the most reliable legal tool against gerrymandering has been the Voting Rights Act of 1965 (VRA), which requires a demonstration of racially polarized voting (RPV). This RPV analysis is typically performed by statistical techniques that infer voting by race from precinct-level returns. Many voting rights advocates worry that noising of Census data will confuse population balancing practices, and others worry that it will attenuate RPV signals, making it harder to press valid claims.

The Census Bureau has been commendably transparent about the development of TopDown, making working code publicly available along with documentation and research papers describing the algorithm. The complexity of the algorithm makes it extremely difficult to study analytically, so many people have sought to run it on realistic data. However, since person-level Census data remain confidential for 72 years after collection, detailed input data for TopDown is not public. Data users who would like to understand its impacts are left with two options: decades-old data or a limited demonstration data product.

In this paper, we get around the empirical obstacle by use of reconstructed block-level 2010 microdata for the state of Texas, and we try to understand the algorithm through theoretical analysis of a much-simplified toy algorithm, ToyDown, that retains the two-stage, top-down structure of TopDown but is much easier to analyze symbolically. We investigate three questions about the count discrepancies created by TopDown in units of census geography and "off-spine" aggregations like districts and precincts.

Hierarchical budget allocation. We derive easy-to-evaluate expressions for ToyDown errors as a function of the privacy budget allocation. Error at higher levels of the geographic hierarchy impacts lower-level counts with a significant discount, suggesting that bottom-heavy allocations may be optimal for accuracy on small geographies. This is consistent with the small-district errors in our experiments with TopDown. For larger districts, a tract-heavy allocation gives greatest accuracy. Equal allocation over the levels is a strong performer in both cases, making this a good choice from the point of view of multi-scale redistricting.

District construction. From there, we create further tests to study the impacts of district design. We compare hierarchically greedy to geometrically greedy district-generation schemes, where the former attempt to keep large units whole and the latter attempt to build districts with short boundaries. We find that the ToyDown model gives errors very closely keyed to the fragmentation of the hierarchy, but that spatial factors damp out the primary role of fragmentation in the shift to the TopDown setting.

Robustness of linear regression. Finally, we consider the unweighted linear regressions commonly used to assess racial polarization in voting rights cases. We find that the noise from both ToyDown and TopDown introduces an attenuation bias that seems alarming at first. However, unweighted linear regression on precincts is already vulnerable to major skews imposed by the inclusion of very small precincts. For any reasonable way of counteracting that – trimming out the tiny precincts or weighting the regression by the number of votes cast – the instability introduced by ToyDown and TopDown all but vanishes.

Our investigation is set up to answer questions about the status quo workflow in redistricting. As usual with studies of differential privacy, a finding that DP unsettles the current practices might lead us to call to refine the way it is applied, but might equally lead us to interrogate the traditional practices and seek next-generation methods for redistricting. In particular, it is clear that the practice of *one-person* population deviation across districts was never reasonably justified by the accuracy of Census data nor required by law, and the adoption of differential privacy might give redistricters occasion to reconsider that practice. We make a similar observation about the way that racially polarized voting analysis is commonly performed in expert reports. On the other hand, by focusing on decisions still to be announced like the privacy budget and its allocation over the hierarchy, we are able to make recommendations that can assist the Bureau in protecting privacy while attending to the important concerns of user groups.

2 Background on Census and redistricting

2.1 The structure of Census data and the redistricting data products

Every ten years the U.S. Census Bureau attempts a comprehensive collection of person-level data – called *microdata* – from every household in the country. The microdata are confidential, and are only published in aggregated tables subject to disclosure avoidance controls. The Decennial Census records information on the sex, age, race, and ethnicity for each member of each household, using categories set by the Office of Management and Budget [8]. The 2020 Census used six primary racial categories: White, Black, American Indian, Asian, Native Hawaiian/Pacific Islander, and Some Other Race. An individual can select these in any combination but must choose at least one, creating $2^6 - 1 = 63$ possible choices of race. Separately, *ethnicity* is represented as a binary choice of Hispanic/Latino or not.

The 2010 Census divided the nation into over 11 million small units called *census blocks* which nest in larger geographies in a six-level “central spine”: nation – state – county – tract – block group – block. Counts of different types are provided with respect to these geographies. This tabular data is then used in an enormous range of official capacities, from the apportionment of seats in the U.S. House of Representatives to the allocation of many streams of federal and state funding. The redistricting (PL 94-171) data includes four such tables: H1, a table of housing units whose types are occupied/vacant; and four tables of population, P1 (63 races), P2 (Hispanic, and 63 races of non-Hispanic population), and P3/P4 (same as P1/P2 but for voting age population). Each table can be thought of as a *histogram*, with each included type constituting one histogram *bin*. For instance, in table P1 there is 1 person in the $t =$White+Asian bin in the Middlesex County, MA, block numbered 31021002.

Treating the 2010 tables as accurate, it is easy to infer information not explicitly presented in the tables. For instance, the same bin in the P3 table (race for voting age population) also has a count of 1, implying that there are no White+Asian people under 18 years old in block 31021002. This is the beginning of a *reconstruction* process that would enable an attacker, in principle, to learn much of the person-level microdata behind the aggregate releases.

2.2 Disclosure avoidance

Title 13 of the U.S. Code requires the Bureau to take measures to protect the privacy of respondents' data [1]. In the 2010 Census, this was largely achieved by an ad hoc mechanism called *data swapping*: a Bureau employee manually swapped data between small census blocks to thwart re-identification. In 2020, swapping is no longer considered adequate to protect against more sophisticated (but mathematically straightforward) data attacks that seek to reconstruct the individual microdata. An internal Census Bureau study concluded that data swapping was unacceptably vulnerable: Census staff were able to reconstruct the 2010 Census responses of – and correctly reidentify – tens of millions of people.

With the reconstruction/reidentification threat in mind, the Bureau has developed an algorithm called TopDown [2], which begins with a noising step that is *differentially private*, following a mathematical formalism that provides rigorous guarantees against information disclosure [12]. Differentially private algorithms obey a quantifiable limit to how much the output can depend on an individual record in the input. The relationship of output to input is specified by a tuneable parameter, ε, often called the *privacy budget*. When $\varepsilon \to \infty$, the output approaches equality to the input (high risk of disclosure). When $\varepsilon \to 0$, the output bears no resemblance to the input whatsoever (no risk of disclosure). Like a fiscal budget, the privacy budget can be allocated until it is fully spent, in this case by spending parts of the budget on particular queries and on levels of the hierarchy.

TopDown takes an individual-level table of census data and creates a "synthetic" dataset that will be used in its place to generate the PL 94-171 tables. It can be thought of as taking as input a histogram with a bin for each person type (i.e., a combination of race, sex, ethnicity, etc.) and outputting an altered version of the same histogram. It proceeds in two stages. First, it privatizes the input histogram counts: it adds enough random noise to get the required level of differential privacy (according to the budget ε). At this stage, it also allocates a portion of the total privacy budget for generating additional noisy histograms of data of particular importance to the Census Bureau. Second, TopDown does post-processing on the noisy histograms to satisfy a handful of additional plausibility constraints. Among other things, post-processing ensures that the resulting histograms contain only non-negative integers, are self-consistent, and agree with the raw input data on a handful of *invariants* (e.g., total state population).

The overall privacy guarantees of TopDown are poorly understood. In this paper, we design a simpler cousin of TopDown nicknamed ToyDown and we explore the properties of both ToyDown and TopDown, primarily focusing on reconstructed Texas data from 2010.

2.3 The use of Census products for redistricting

The PL 94-171 tables are the authoritative source of data for the purposes of apportionment to the U.S. House of Representatives, and with a very small number of exceptions also for within-state legislative apportionment. The most famous use of population counts is to decide how many members of the 435-seat House of Representatives are assigned to each state. In "One person, one vote" jurisprudence initiated in the *Reynolds v. Sims* case of 1964, balancing Census population is required not only for Congressional districts within a state but also for districts that elect to a state legislature, a county commission, a city council or school board, and so on [17, 18, 3].

Today, the Congressional districts within a state usually balance total population extremely tightly: each of Alabama's seven Congressional districts drawn after the 2010 Census has a total population of either 682,819 or 682,820 according to official definitions of districts and the Table P1 count, while Massachusetts districts all have a population of 727,514 or 727,515. Astonishingly, though no official rule demands it, more than half of the states maintain this "zero-balancing" practice (no more than one person deviation) for Congressional districts [16]. This ingrained habit of zero-balancing districts to protect from the possibility of a malapportionment challenge is the first source of worry in the redistricting sphere. If disclosure avoidance practices introduce some systematic bias – say by creating significant net redistribution towards rural and away from urban areas – then it becomes hard to control overall malapportionment, which could in principle trigger constitutional scrutiny. In the end, redistricters may not care very much how many people live in a single census block, but it could be quite important to have good accuracy at the level of a district.

The second major locus of concern for redistricting practitioners is the enforcement of the Voting Rights Act (VRA). Here, histogram data is used to estimate the share of voting age population held by members of minority racial and ethnic groups. Voting rights attorneys must start by satisfying three threshold tests without which no suit can go forward.

- **Gingles 1**: the first "Gingles factor" in VRA liability is satisfied by creating a demonstration district where the minority group makes up over 50% of the voting age population.
- **Gingles 2-3**: the voting patterns in the disputed area must display *racial polarization*. The minority population is shown to be cohesive in its candidates of choice, and bloc voting by the majority prevents these candidates from being elected. In practice, inference techniques like linear regression or so-called "ecological inference" are used to estimate voting preferences by race.

Since the VRA has been a powerful tool against gerrymandering for over 50 years, many worry that even where the raw data would clear the Gingles preconditions, the noised data will tend towards uniformity – blocking deserving plaintiffs from a cause of action.

3 Census TopDown and ToyDown

3.1 Setup and notation

For the Census application, the data universe is a set of *types*: for instance, the redistricting data (the PL 94-171) has the types $T = T_R \times T_E \times T_{VA} \times T_H$, where T_R is the set of 63 races, T_E is binary for ethnicity (Hispanic or not), T_A is binary for age (voting age or not), and T_H is the set of housing types. (The fuller decennial Census data has more types.)

A *hierarchy* H is a rooted tree of some depth d, so that every leaf has distance $\leq d-1$ from the root. We will usually assume the hierarchy has uniform depth, so that every leaf is exactly $d-1$ away from the root. For node $h \in H$, let $n(h) \in \mathbb{N}$ be the number of children of h in the tree, and let $\ell(h)$ be the level of node h. A hierarchy is called *homogeneous* if each node at level ℓ has the same number of children, denoted n_ℓ. Let H_ℓ denote the set of nodes at level ℓ, so that the set of leaves is H_d in the uniform-depth case. Label the root of the tree $h = 1$. We adopt an indexing of the tree and refer to the ith child of h as h_i; the parent of any non-root node h is denoted $\hat{h}$. In Census data, the hierarchy represents the large and complicated set of nested geographical units, from the nation at the root down to the census blocks at the leaves. The standard hierarchy has the six levels (nation – state – county – tract – block group – block) described above.

We associate with hierarchy H and types T a set of *counts* $A_{H,T} = \{a_{h,t} \in \mathbb{N}\}_{h \in H, t \in T}$, where $a_{h,t}$ is the population of type t in unit h of census geography. We say $A_{H,T}$ is *hierarchically consistent* if the counts add up correctly: for every non-leaf h and every t, we require $a_{h,t} = \sum_{i \in [n(h)]} a_{h_i,t}$. For a singleton T, we write $A_H = \{a_h\}$. We set an *allocation* $(\varepsilon_1, \ldots, \varepsilon_d)$ breaking down the privacy budget $\varepsilon = \sum \varepsilon_i$ to the different levels of the hierarchy.

Our *queries* will always be counting queries, so that for instance $q_{F,44}(h)$ returns the number of 44-year-old females in geographic unit h. This particular query is part of a "sex by age" *histogram* $Q_{sex,age} = \{q_{s,a} : s \in T_S, a \in T_A\}$, which partitions T into *bins* by sex and age. In this language, $q_{F,44}$ is a bin of the sex-by-age histogram. By slight abuse of notation, we will use the same terminology for the queries and their outputs, so that the histogram can be thought of as the collection of queries or the collection of counts. Similarly, the "voting age by ethnicity by race" histogram consists of a query for each combination of the $2 \times 2 \times 63$ possible combinations of the three attributes.

3.2 ToyDown and TopDown

The Bureau's TopDown and our simplified ToyDown are both algorithms for releasing privatized population counts for every $h \in H$. That is, these algorithms protect privacy by noising the data histograms. TopDown releases not just total population counts, but counts by type. We will define *single-attribute* and *multi-attribute* versions of ToyDown that noise A_H and $A_{H,T}$, respectively, where consistency must hold for each type t.

TopDown and ToyDown share the same two-stage structure. Starting with hierarchically consistent raw counts a, the *noising stage* generates differentially private counts $\widehat{a}$. The *post-processing stage* solves a constrained optimization problem to find noisy counts α that are close to the $\widehat{a}$ values while satisfying hierarchical consistency and other requirements. TopDown is named after the iterative approach to post-processing: one geographic level at a time, starting at the top (nation) and working down to the leaves (blocks). We sketch the noising and post-processing here, and we describe them in Appendix A in more detail.

The simple ToyDown model can be run in a single-attribute version (only counts A_H), a multi-attribute version (counts by type $A_{H,T}$), or in multi-attribute form enforcing non-negativity. The single-attribute version is easy to describe: level by level, random noise values are selected from a Laplace distribution with scale $1/\varepsilon_\ell$ and added to each count, replacing each a_h with $\widehat{a}_h = a_h + L_h$. Then, working from top to bottom, the noisy $\widehat{a}_h$ are replaced with the closest possible real numbers α_h satisfying hierarchical consistency. Multi-attribute ToyDown is defined analogously, but using $A_{H,T}$ instead of A_H and requiring hierarchical consistency within each type $t \in T$. Non-negative ToyDown adds the inequality requirement that $\alpha_h \geq 0$.

TopDown is structurally similar but much more complex, with more kinds of privatized counts in the noising stage and a great many more constraints in the post-processing stage, including integrality. The privatized counts computed by TopDown are specified by a collection of histograms (or complex queries) called a *workload* W. For each bin of each histogram in the workload and for each node h in the geographic hierarchy, TopDown adds geometric noise to the count. The post-processing step finds the closest integer point that satisfies the requirements given by hierarchical consistency, non-negativity, as well as additional conditions given as invariants and structural inequalities. For example, any block with zero households in the raw counts must have zero households and zero population in the output adjusted counts. Together, the invariants, structural inequalities, integrality, and non-negativity make this optimization problem very hard. The problem is NP-hard in the worst case and TopDown cannot always find a feasible solution. There is a sophisticated secondary algorithm for finding approximate solutions that is beyond the scope of this paper.

ToyDown is simple enough that solutions can often be obtained symbolically. ToyDown simplifies the noising stage by fixing the workload to be the detailed workload partition $Q_{detailed} = \{\{t\}\}_{t \in T}$ consisting of all singleton sets and using the continuous Laplace Mechanism instead of the discrete Geometric Mechanism. It simplifies the post-processing stage by dropping invariants, structural inequalities, integrality, and non-negativity. When negative answers are permitted, multi-attribute ToyDown is equivalent to executing $|T|$ independent instances of single-attribute ToyDown on inputs $A_{H,t} = \{a_{h,t}\}_{h \in H}$ for each $t \in T$. As a result, many of our analytical results for single-attribute ToyDown extend straightforwardly to multi-attribute ToyDown (allowing negative answers) by scaling by a factor of $|T|$ in appropriate places.

4 Methods

We use both analytical and empirical techniques in this work. This section describes our high-level empirical approach: what algorithms and raw data we used and how we used them. See Appendix B for more details. We repeatedly ran TopDown and ToyDown in various configurations on a reconstructed person-level Texas dataset created by applying a reconstruction technique to the block-level data from the 2010 Census, following [15] based on [11]. The reconstructed microdata records – obtained from collaborators – contain block-level sex, age, ethnicity, and race information consistent with a collection of tables from 2010 Census Summary File 1.

We executed 16 runs of TopDown with each of 20 different allocations of the privacy budget across the five lower levels of the national census geographic hierarchy: $\varepsilon = \varepsilon_2+\varepsilon_3+\varepsilon_4+\varepsilon_5+\varepsilon_6$. The 20 allocations consist of five different splits across the levels (Table 1) for each of four total budgets $\varepsilon \in \{0.25, 0.5, 1.0, 2.0\}$. TopDown operates on the six-level Census hierarchy and requires specifying ε_1. In our experiments, we ran TopDown with a fixed total privacy budget $\varepsilon_{total} = 10$, with $\varepsilon_1 = 10 - \varepsilon$. Because the nation-level budget is so much higher than the lower level budgets, we omit further discussion of it. The TopDown workload was modeled after the workload used in the 2018 End-to-End test release, omitting household invariants and queries.

We also ran three variants of ToyDown (single-attribute, multi-attribute, and non-negative) on a simplified version of the same data 2010 data. We executed 16 runs of each variant with each of five different splits of the privacy budget across the five lower levels of the census geographic hierarchy (Table 1), fixing the total budget for those five levels at $\varepsilon = 1$. The data was derived from the reconstructed Texas data simplified to include only seven distinct types: one for the total Hispanic population and one for each of six subgroups of the non-Hispanic population based on race (White; Black; American Indian; Asian; Native Hawaiian/Pacific Islander; and Some Other Race or multiple races). Post-processing for single-attribute ToyDown was implemented in NumPy, while post-processing for multi-attribute and non-negative ToyDown used a Gurobi solver.

5 Hierarchical budget allocation

The relationship of the hierarchical allocation $(\varepsilon_1, \ldots, \varepsilon_d)$ to various measures of output accuracy is not obvious. On one hand, it might seem that higher values of ε_d (the block-level budget) will best promote accuracy at the block level, for a fixed ε. But on the other hand, imposing hierarchical consistency forces lower levels to be consistent with the totals at higher levels, which means that noise at higher levels can trickle down to lower levels. These competing effects create tradeoffs that are hard to balance without further analysis.

Table 1 Names of designated budget splits used in ToyDown and TopDown runs below, each with a budget of $\varepsilon_1 = 9$ on the nation and a total of 1 allocated below the national level.

Split name	state ε_2	county ε_3	tract ε_4	BG ε_5	block ε_6
equal	0.2	0.2	0.2	0.2	0.2
state-heavy	0.5	0.25	0.083	0.083	0.083
tract-heavy	0.083	0.167	0.5	0.167	0.083
BG-heavy	0.083	0.083	0.167	0.5	0.167
block-heavy	0.083	0.083	0.083	0.25	0.5

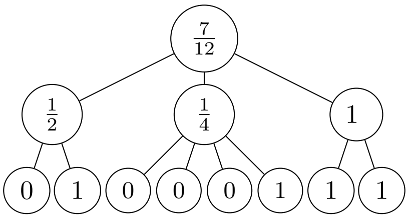


Figure 1 A district in a three-level hierarchy. The 0/1 weight of a leaf indicates its membership in the district; each non-leaf weight is the average of the node's children.

5.1 ToyDown error expressions

▶ **Definition 1** (District, weights, error). *A* district $D \subseteq H_d$ *is a subset of the leaves (blocks) of the hierarchy H. For hierarchy H, a district D induces* weights $w_h \in [0,1]$ *on the hierarchy nodes, defined recursively as follows:*

- *For each leaf $h \in H_d$, let $w_h = 1$ if $h \in D$ and $w_h = 0$ otherwise.*
- *For $\ell \le d-1$ and $h \in H_\ell$, let $w_h = \frac{1}{n(h)} \cdot \sum_{i\in[n(h)]} w_{h_i}$ be the average of the weights of the children.*

In a homogeneous hierarchy, we can observe that each w_h equals the fraction of the leaves descended from h that belong to D. In particular, the root weight is $w_1 = |D|/|H_d| = 1/k$ if there are k districts of equal population made from nodes of equal population.

For node $h \in H$, we record the *error* $E_h = \alpha_h - a_h$ introduced by ToyDown to the count a_h. The total error over district D is $E_D = \sum_{h\in D} E_h$. Let $\hat{h}$ denote the parent of node h.

▶ **Theorem 2** (Error expressions). *$E_1 = L_1$. For $\ell \in \{2,\dots,d\}$ and non-root node $h_i \in H_\ell$, and for every district D with associated weights w_h on the nodes,*

$$E_{h_i} = L_{h_i} + \frac{1}{n(h)}\left(E_h - \sum_{j\in[n(h)]} L_{h_j}\right), \qquad E_D = w_1 L_1 + \sum_{h\in H\setminus\{1\}} (w_h - w_{\hat{h}}) L_h. \tag{1}$$

We make several observations. First, our intuition that error at higher levels trickles down to lower levels is correct, but this effect is rather weak. The error at a child h_i is determined by the parent error E_h discounted by the degree $n(h)$, the number of siblings. This suggests that placing more budget at level ℓ is an efficient way to secure accuracy at that level, until a fairly extreme level of error at higher levels overwhelms the degree-based "discount."

Second, because the L_h are all independent random variables with $\mathbb{E}(L_h) = 0$ and $\mathrm{Var}(L_h) = 8/\varepsilon^2_{\ell(h)}$, the theorem provides the following expression for variance that we use repeatedly.

▶ **Corollary 3** (Error expectation and variance). *For all $D \subseteq H_d$ and associated weights w_h, the expected error and error variance produced by ToyDown satisfy $\mathbb{E}(E_D) = 0$ and*

$$\mathrm{Var}(E_D) = \frac{8w_1^2}{\varepsilon_1^2} + \sum_{\ell=2}^{d}\left(\frac{8}{\varepsilon_\ell^2} \cdot \sum_{h\in H_\ell} (w_h - w_{\hat{h}})^2\right). \tag{2}$$

Third, we get a more explicit expression if restricting to homogeneous hierarchies H. Consider the case of a singleton district $\{h\}$ made of a single census block $h \in H_d$.

▶ **Corollary 4** (Error variance, homogeneous case)**.** *The ToyDown error for a single block $h \in H_d$ satisfies*

$$\mathrm{Var}(E_h) = \frac{8}{\varepsilon_1^2(n_1 \cdots n_{d-1})^2} + \sum_{\ell=2}^{d} \frac{8n_{\ell-1}(n_{\ell-1}-1)}{\varepsilon_\ell^2(n_{\ell-1}\cdots n_{d-1})^2}. \tag{3}$$

Figure 2 plots this expression for various ways of splitting a total privacy budget of $\varepsilon = 1$ across a three-level hierarchy with $n_1 = n_2 = 10$. The minimum of $f(x_1, \dots, x_d) = \sum_{\ell=1}^{d} a_\ell / x_\ell^2$ subject to $\sum_\ell x_\ell = \varepsilon$ and $x_\ell \geq 0$ is achieved at $x_\ell = \varepsilon a_\ell^{1/3} / \sum_i a_i^{1/3}$ for all ℓ. For the example in Figure 2, the minimum-variance split is $(\varepsilon_1, \varepsilon_2, \varepsilon_3) = (0.038, 0.171, 0.791)$ with variance 14.52. (See accompanying CoLab notebook.) One important note in interpreting Figure 2 is that these variance numbers are absolute and don't depend on knowing population counts for the nodes of the hierarchy. They are simply based on sampling Laplace noise with the given parameters. If a variance of about 15 in the bottom-level counts is too high to be tolerated in an application, one would have to increase ε to achieve lower variance.

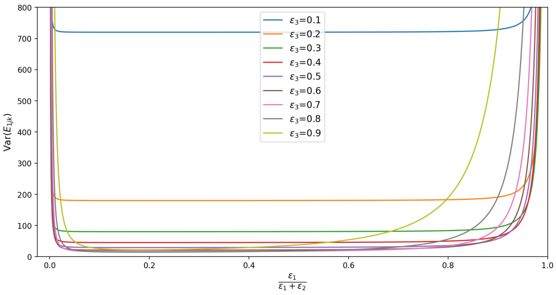


Figure 2 ToyDown error variance for a leaf node in the three-level hierarchy with $n_1 = n_2 = 10$ and $\varepsilon = 1$. The curves show varying ε_3 (colors) and the relative balance of ε_1 and ε_2 (x-axis).

Table 2 L^1 error measurements from selected TopDown runs on reconstructed Texas data. The allocation $(\varepsilon_1, \dots, \varepsilon_6)$ goes from the nation $\ell = 1$ down to census blocks at $\ell = 6$.

ε	Allocation	L^1 error
1.0	(.16, .16, .16, .16, .16, .2)	0.03
1.0	(.2, .16, .16, .16, .16, .16)	0.03
1.0	(.1, .1, .1, .1, .1, .5)	0.02
1.0	(.02, .02, .02, .02, .02, .9)	0.03
1.0	(.66, .30, .01, .01, .01, .01)	0.09

5.2 Empirical error experiments in TopDown

Next, we move to TopDown, which requires the use of input data. First, using reconstructed 2010 Texas data, we varied the relative allocation vector and the total ε, then measured the effects with an L^1 error metric included in the Census code [5]. This is a measure of block-level error: it adds the magnitudes of changes in the bins, then divides by twice the total population in the histogram.

Table 2 reports a small selection of the 100+ different scenarios explored. In general, the lowest error outcomes were observed in a few scenarios: when the budget was distributed near-equally to the levels of the hierarchy, and when half of the available budget was placed at the bottom level – beyond $\varepsilon_d = \varepsilon/2$, further bottom-weighting gave diminishing returns in block-level accuracy.

But a budget allocation that produces small block-level errors may not produce small errors for *districts*, depending on the degree of cancellation or correlation. Next, we use random district generation to understand the effects of off-spine aggregation. In particular, we employ the Markov chain sampling algorithm called *recombination* (or ReCom), which runs an elementary move that fuses two neighboring districts and re-partitions the double-district by a random balanced cut to a random spanning tree [10].

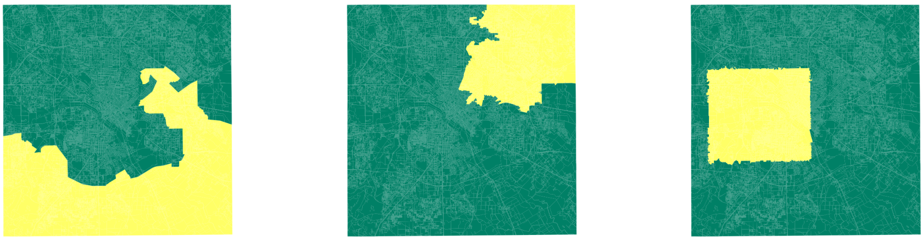


Figure 3 Three sample districts (yellow) in Dallas County, each within two percent of the ideal population for $k = 4$ districts. These are drawn by tract ReCom, block ReCom, and a square-favoring algorithm, respectively.

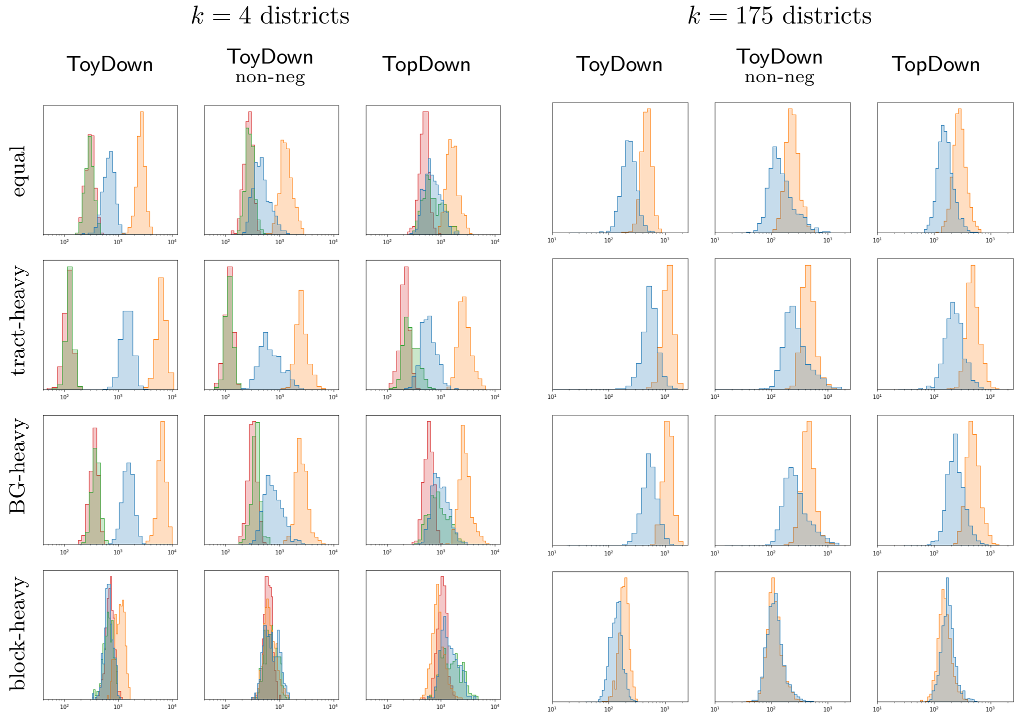
We begin with county commission districts in Dallas County, where $k = 4$. Since the 2010 population of Dallas County was roughly 2.4 million, each district will have roughly 600,000 people, making them nearly as big as congressional districts and much larger than tracts. We also include divisions of the county into $k = 175$ districts of between 13,000 and 14,000 people each for a small-district comparison. Figure 4 plots the data from our experiments on a logarithmic scale. Each histogram displays 400 values, one for each district drawn by the specified district-drawing algorithm; each value is the mean observed district-level population error magnitude over 16 executions of the specified hierarchical noising algorithm using the specified budget allocation.

First, consider two unrealistic forms of district-generation: tract Disconn (red) and block Disconn (orange), which randomly choose units of the specified type until assembling a collection with the appropriate population. These are unrealistic because they do not form connected districts; here, they are used to illustrate the effects of aggregation, neglecting spatial factors entirely. We see in Figure 4 that block-based methods generate hugely more error than tract-based methods, except if the budget allocation is concentrated at the bottom of the hierarchy. The effect is stronger for ToyDown (in keeping with Theorem 2), but is easily observed for TopDown as well.

We compare that with the more realistic district-generation algorithm block ReCom (blue), which builds compact and connected districts out of block units. This tends to give error levels in between the extremes set by the other two. Likewise, tract ReCom (green) builds compact and connected districts from tracts. One reasonable mechanism by which ReCom has much lower error than Disconn is that ReCom districts will tend to have higher "hierarchical integrity," keeping higher-level units whole just by virtue of being connected and plump. The interior of ReCom districts will thus contain many whole block groups and tracts. Near the boundary, block groups and tracts are more fragmented, leaving the corresponding block-level errors uncancelled. These fragmentation ideas are explored more fully in Section 6 and some sample districts are depicted here.

The cancellation effect is significant: in most experiments, the error level for ReCom districts is much closer to that of tract Disconn than block Disconn (recall the data is plotted on a logarithmic scale). Overall, drawing districts out of larger pieces (e.g., using tract Disconn instead of ReCom, or ReCom instead of block Disconn) lowers error magnitude significantly in the best case and has little or no effect in the worst case.

Although tract ReCom and tract Disconn behave very similarly under ToyDown, the compact districts perform noticeably worse than their disconnected relatives once we pass to the full complexity of TopDown. At first this seems puzzling, because compact and



Green: tract ReCom, Red: tract Disconn, Blue: block ReCom, Orange: block Disconn

Figure 4 These histograms show district-level error on a log scale for various combinations of budget splits (rows), district-drawing algorithms (colors), and noising algorithms (columns). We include both large districts and small districts, dividing the county into $k = 4$ and $k = 175$ equal parts. Each histogram displays 400 values, one for each district drawn by the specified algorithm, plotting the mean observed district-level population error magnitude over 16 executions of the noising algorithm using the specified budget allocation.

connected districts are being punished by the geography-aware TopDown. But the reason for this is apparent on further reflection: *spatial autocorrelation* is causing the post-processing corrections to move nearby tracts in the same direction, impeding the cancellation that makes counts usually more accurate on larger geographies.

In the end, the story that emerges from these investigations is that, with full TopDown, the best accuracy that can be observed for large districts occurs when they are made from whole tracts and the allocation is tract-heavy; an equal split is not much worse. For districts with population around 13,000, $\varepsilon = 1$ noising creates errors in the low hundreds for compact, connected districts, with the best performance for block-heavy allocations. Again, an equal split is not much worse, suggesting that this might be a good policy choice for accuracy in districts across many scales.

6 Geometrically compact vs hierarchically greedy districts

The analysis above suggests that the district-level error E_D will depend not only on the randomness of the noising algorithms, but also on the geometry of D and the structure of H. This section studies the hypothesis that districts that disrespect the geographical hierarchy will tend to have higher error magnitude. This section defines the *fragmentation score*, relates a district's fragmentation score to its error variance under ToyDown, and compares the fragmentation of two simple district-drawing algorithms on homogeneous hierarchies and simple geographies. Ultimately, we find that the explanatory value of the fragmentation score decays as we move to more realistic deployment of TopDown. This discrepancy raises important questions for future study: Which of the many additional features of TopDown attenuates the fragmentation–variance relationship?

We define a score intended to capture the contribution to $\mathrm{Var}(E_D)$ of the shape of the district with respect to the hierarchy. Recall that $\hat{h}$ denotes the parent of node h.

▶ **Definition 5** (Fragmentation score). *For $D \subseteq H_d$, let* $\mathsf{Frag}(D) = \sum_{h \in H} (w_h - w_{\hat{h}})^2$.

Because weights are in $[0, 1]$, the score obeys $0 \le \mathsf{Frag}(D) < |H|$ for all districts, with higher scores indicating the presence of more units that are only partially included in D.

This fragmentation score is reverse-engineered from the expression for the variance of district-level population errors when using ToyDown with privacy divided equally across levels of the hierarchy (Corollary 3): namely, $\mathrm{Var}(E_D) = \frac{8d^2}{\varepsilon^2}\left(w_1^2 + \mathsf{Frag}(D)\right)$. When the district D itself is a random variable sampled from some distribution, the expected fragmentation $\mathbb{E}(\mathsf{Frag}(D))$ is similarly related to $\mathrm{Var}(E_D)$. Namely, using the law of total variation, when each level gets ε/d privacy budget:

$$\mathrm{Var}(E_D) = \mathbb{E}\left(\mathrm{Var}(E_D|D)\right) + \mathrm{Var}\left(\mathbb{E}(E_D|D)\right) = \mathbb{E}(\mathrm{Var}(E_D|D)) = \frac{8d^2}{\varepsilon^2}(\mathbb{E}(\mathsf{Frag}(D)) + \mathbb{E}(w_1^2)).$$

When ε is allocated unequally across levels, as for the other splits in Table 1, the simple analytical relationship between the fragmentation score and the error variance breaks down.

Observe that a hierarchy H does not capture all of the geometry relevant to district drawing. In particular, H does not directly encode any information about block adjacency, and therefore we can't detect from H that a district is contiguous. For algorithms to generate contiguous districts, we need to make use of the plane geometry associated to H. We restrict our attention to the simplest case: homogeneous hierarchies (where every node on level $\ell < d$ has exactly n_ℓ children) and *square tilings.* (where each unit on level ℓ is a square and has n_ℓ children that cover it with a $\sqrt{n_\ell} \times \sqrt{n_\ell}$ grid tiling).

We analyze the fragmentation score for two simple district-drawing algorithms (see Appendix C). The Greedy algorithm builds a district from the largest subtrees possible, only subdividing a subtree when necessary. It takes as input H and $k \in \mathbb{N}$ and returns a district of size $N = \lfloor |H_d|/k \rfloor$, assembled by starting with the largest available units at random and adding units that are adjacent in the labeling sequence without passing size N, then allowing one partial unit, and so on recursively at lower levels. Observe that Greedy depends only on the hierarchy H. The Square algorithm takes as input a square, homogeneous hierarchy H and $k \in \mathbb{N}$ such that the district size is a perfect square, $|D| = |H_d|/k = {s_d}^2$. It outputs a uniformly random $s_d \times s_d$ square of blocks.

▶ **Theorem 6.** *Let $D_G \sim$ Greedy(H, k), $D_\square \sim$ Square(H, k). For $n_1 \cdot n_2 \cdots n_{d-2} \ge k \ge 2$, let $L = \arg\min\{\ell : n_1 \cdot n_2 \cdots n_\ell \ge k\}$.*

$$\mathbb{E}(\mathsf{Frag}(D_G)) \le \frac{k-1}{k^2} \sum_{\ell=1}^{L} n_\ell + \frac{1}{4} \sum_{\ell=L+1}^{d-1} n_\ell; \quad \mathbb{E}(\mathsf{Frag}(D_\square)) \ge \frac{2}{3}\left(\frac{\sqrt{n_1 \dots n_{d-1}}}{\sqrt{k}} - \frac{11}{2}\right)\sqrt{n_{d-1}}.$$

Dallas County is nearly a perfect square shape, so it gives us an opportunity to set some roughly realistic parameters to evaluate these bounds. There are 529 tracts in Dallas County, with an average of 3.2 blocks groups per tract and 26.4 blocks per block group, yielding 44,113 total blocks. We can approximate these parameters by setting $d = 4$, using $k = 4$ as for the county commission districts, and setting $(n_1, n_2, n_3) = (484, 4, 25)$ which has a reasonably similar 48,400 blocks (as a result, $L = 1$). The bounds in the theorem say that $\mathbb{E}(\mathsf{Frag}(D_G)) \leq 98$ and $\mathbb{E}(\mathsf{Frag}(D_\square)) \geq 264$. Note: for homogeneous hierarchies H with equal-population leaves, the score $\mathsf{Frag}(D_G)$ is independent of algorithm randomness and can be computed exactly; for the above parameters $\mathsf{Frag}(D_G) = 90.75$. So the bound in the theorem is fairly tight, at least in this case.

To interpret the theorem, it is helpful to think of Greedy as being hierarchically greedy and Square as being geometrically greedy. That is, the former is oriented toward using the biggest possible units and keeping them whole, so that spatial considerations are secondary; the latter is oriented towards "compact" geographies with a lot of area relative to perimeter, and unit integrity is secondary. The theorem shows that compactness alone (a function of the plane geometry) does not keep down the fragmentation score (a function of the hierarchy), and indeed the bounds get farther apart as the hierarchy gets larger and more complicated. In Appendix C, we compare these theoretical results to empirical district errors, finding that fragmentation tracks well with errors in ToyDown, but that the complexity of the TopDown model weakens the relationship, suggesting a need for more sophisticated tools.

7 Ecological regression with noise

7.1 Inference methods for Voting Rights Act enforcement

When elections are conducted by secret ballot, it is fundamentally impossible to precisely determine voting patterns by race from the reported outcomes alone. The standard methods for estimating these patterns use the cast votes at the precinct level, combined with the demographics by precinct, to infer racial polarization. Because the general aggregate-to-individual inference problem is called "ecological" (cf. ecological paradox, ecological fallacy), the leading techniques are called *ecological regression* (ER) and *ecological inference* (EI). It is rare that EI and ER do not substantively agree, and we focus on ER here because it lends itself to easily interpretable pictures.

ER is a simple linear regression, fitting a line to the data points determined by the precincts on a demographics-vs-votes plot. A high slope (positive or negative) indicates a likely strong difference in voting preferences, which is necessary to demonstrate the Gingles 2-3 tests for a VRA lawsuit.

The top row of Figure 5 shows standard ER run on the precincts of Dallas County, with each precinct plotted according to its percentage of Hispanic voting age population or HVAP (x-axis) and the share of cast votes that went to Lupe Valdez (y-axis). Strong racial polarization would show up as a fit line of high slope. This process produces a point estimate of Hispanic support for Valdez, found by intersecting the fit line with the $x = 1$ line, which represents the scenario of 100% Hispanic population. The point estimate of non-Hispanic support for Valdez is at the intersection of the fit line with $x = 0$.

7.2 Summary of Experiments

ToyDown and TopDown were both run on the full Texas reconstruction from 2010. We plotted Dallas County votes from three contests: votes for Obama for president in 2012 general election, votes for Valdez for governor in the 2018 Democratic Party primary runoff, and

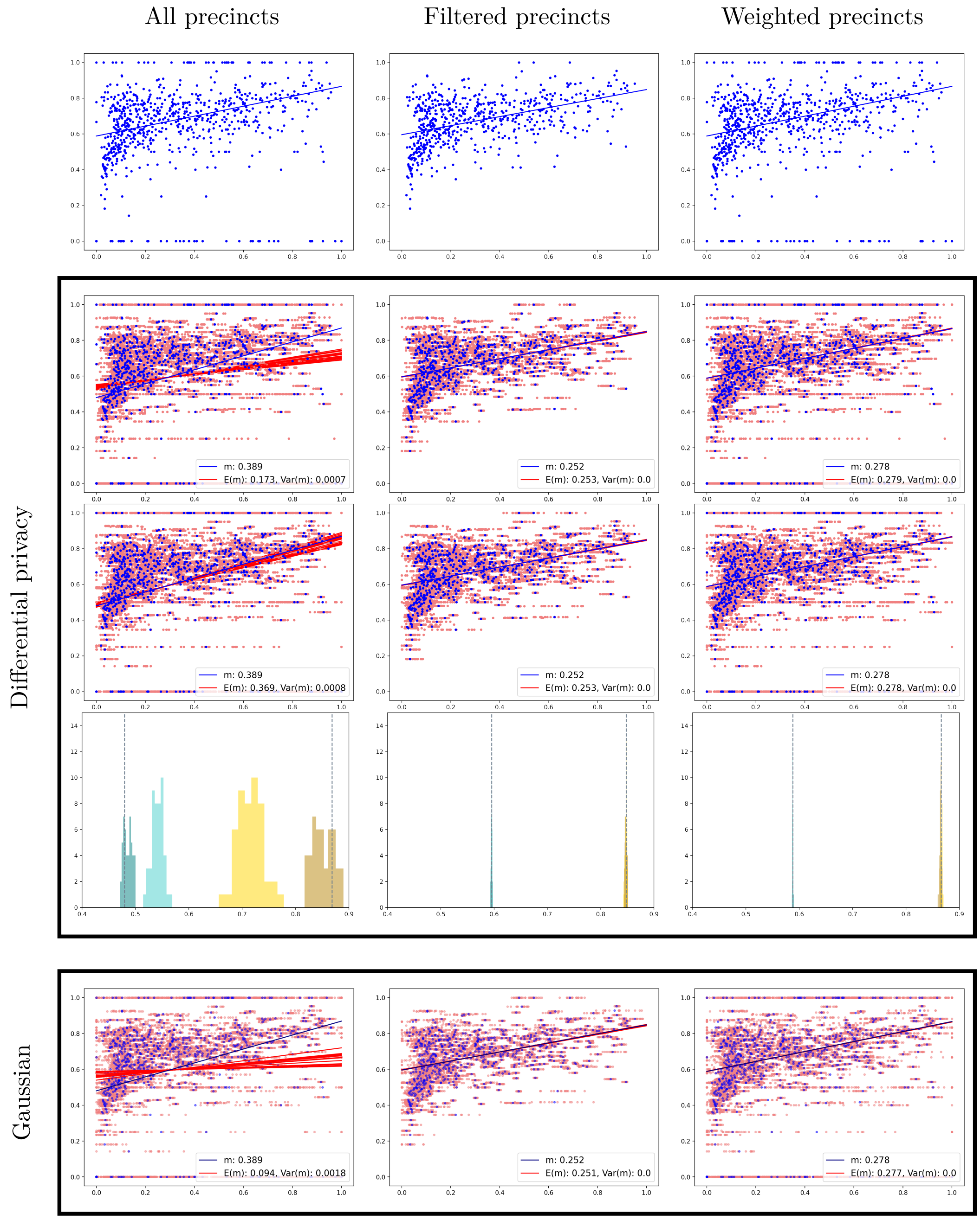


Figure 5 Comparing ecological regression on un-noised data (top row) with various styles of noising. ER is re-run on data noised by differentially private ToyDown (second row), and data noised by TopDown (third row), both with $\varepsilon = 1$, equal split. The blue dots repeat the un-noised data, the pink dots show 16 runs of noised data with pink fit lines re-computed each time. Below that, the histograms show the point estimates of Latino (gold) and non-Latino (teal) support for Valdez estimated from ER on data noised by ToyDown (lighter) and TopDown (darker). The last row contrasts the differentially private algorithms with a naive variant that adds noise to each precinct from a mean-zero Gaussian distribution, set to match the average precinct level L^1 error observed in the ToyDown runs (in this case, this is $\sigma = 20$). Across all of these experiments, the conclusion is striking: TopDown performs better than ToyDown and far better than a naive Gaussian variant, even without filtering precincts; if precincts are filtered or weighted, none of the noising alternatives threatens the ability to detect racially polarized voting.

Table 3 Point estimates from ER for Dallas County in the Valdez/White primary runoff in 2018. In the first table, estimates are made with (un-noised) VAP data from the 2010 Census. In the *filtered precincts* case, precincts with fewer than 10 cast votes are excluded from the initial set of 827 precincts. In the *weighted precincts* case, precincts are weighted by the number of cast votes. The ToyDown and TopDown estimates are made from VAP data from 16 runs with $\epsilon = 1$ and an ϵ-budget with all levels given equal weighting. Variance is the empirical variance over the repeated runs of the noising algorithm and is in units of 10^{-8}, shown to two significant digits.

	All precincts (827)		Filtered precincts (626)		Weighted precincts (827)	
Race	this group	complement	this group	complement	this group	complement
Hispanic	0.869	0.480	0.848	0.596	0.866	0.588
Black	0.917	0.518	0.851	0.620	0.835	0.595
White	0.555	0.623	0.474	0.811	0.478	0.805

			All (827)		Filtered (626)		Weighted (827)	
Race	Algorithm	statistic	group	compl.	group	compl.	group	compl.
Hispanic	ToyDown	mean	0.715	0.541	0.848	0.595	0.867	0.588
Hispanic	ToyDown	variance	36000	7000	250	43	160	19
Black	ToyDown	mean	0.798	0.543	0.851	0.62	0.835	0.595
Black	ToyDown	variance	39000	2100	89	5.9	25	2.1
White	ToyDown	mean	0.476	0.674	0.473	0.811	0.478	0.805
White	ToyDown	variance	17000	8000	64	36	33	17
Hispanic	TopDown	mean	0.853	0.485	0.848	0.595	0.865	0.587
Hispanic	TopDown	variance	45000	6700	480	100	120	16
Black	TopDown	mean	0.91	0.52	0.85	0.62	0.835	0.595
Black	TopDown	variance	30000	1200	250	23	45	2.4
White	TopDown	mean	0.582	0.607	0.472	0.81	0.47	0.804
White	TopDown	variance	10000	3400	92	37	92	10

votes for Chevalier for comptroller in the 2018 general election. We chose these contests because in each, ER finds evidence of strong racially polarized voting when using published 2010 census data. All three contests gave similar findings; we'll choose the Valdez runoff contest as our focus here.

For both ToyDown and TopDown, we vary how we handle the inclusion of small precincts in the ecological regression. The options are All (every precinct is a data point in the scatterplot, all weighted equally); Filtered (only including precincts with at least 10 votes cast in that election); or Weighted (weighting the terms in the objective function in least-squares fit by number of votes cast). Filtering and weighting are done using the exact number of cast votes, not the differentially private precinct population totals, which is realistic to the use case.

For each noising run we have a block- or precinct-level matrix, $\hat{M}$ of noised counts, with height b, the number of geographic units (blocks or precincts), and width c, the number of attributes for which there are counts recorded. We also have a corresponding matrix M of un-noised counts. We can compute the L_1 error by summing over the absolute value of every entry in $M - \hat{M}$. ToyDown and TopDown were run 16 times for each configuration. Let E_{avg} be the average L_1 error across noising runs.

If we add *Gaussian* noise to each count instead, the expected L_1 error is $\sum_{i,j} E[|X_{i,j}|]$, where $X_{i,j} \sim \mathcal{N}(0, \sigma^2)$. This is the half-normal distribution, so $E[|X_{i,j}|] = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$. We rearrange to find the standard deviation $\sigma = \frac{E_{avg}\sqrt{\pi}}{bc\sqrt{2}}$ that defines the Gaussian distribution (with $\mu = 0$), so that adding a random variable drawn from it to each unit count will produce an expected L^1 error matching the average E_{avg} observed across the runs.

7.3 The role of small precincts

Practitioners who use ER have raised two questions regarding the effect of differential privacy: (1) How robust will the estimate be after the noising? (2) Will noising diminish the estimate of candidate support from a minority population? We analyzed the effects of TopDown and ToyDown on the 2018 Texas Democratic primary runoff election, where Lupe Valdez was a clear minority candidate of choice in Dallas county.[1]

We begin by observing that of the 827 precincts in Dallas County, 201 have fewer than 10 cast votes from that election day – in fact, 99 precincts recorded zero cast votes. These precincts are a big driver of instability under DP. This is not surprising; percentage swings are much higher in small numbers even if the noise injected might be low. However, downweighting these small precincts makes the estimate almost always agree with the un-noised estimate. Specifically, we assign weights to the precincts equivalent to the number of total votes in the precinct. Figure 5 shows how the estimates vary by run type and data treatment.

8 Conclusion

The central goal of this study has been to take the concerns of redistricting practitioners seriously and to investigate potential destabilizing effects of TopDown on the status quo. A second major goal is to make recommendations, both to the Disclosure Avoidance team at the Census Bureau and to the same practitioners – the attorneys, experts, and redistricting line-drawers in the field. Texas generally, and Dallas County in particular, was selected because it has been the site of several interesting Voting Rights Act cases in the last 20 years.[2]

Our top-line conclusion is that, at least for the Texas localities and election data we examined, TopDown performs far better than more naive noising in terms of preserving accuracy and signal detection for election administration and voting rights law. Perhaps more importantly, we have created an experimental apparatus to help other groups conduct independent analyses.

This work has led us to isolate several elements of common redistricting practice that lead to higher-variance outputs and more error under TopDown. The first example is the common use of a full precinct dataset, with no population weighting, in running racial polarization inference techniques. The second major example is the use of the smallest available units, census blocks, for building districts of all sizes, with no particular priority on intactness for larger units of Census geography. In both cases, we find that these were already likely sources of silent error. Filtering small precincts (or, better, weighting by population) and building districts that prioritize preserving whole the largest units that are suited to their scale are two examples of simple updates to redistricting practice. Besides being sound on first principles, these adjustments can insulate data users from DP-related distortions and help safeguard the important work of fair redistricting.

[1] We also examined the general elections for President in 2012 and Comptroller in 2018, with similar findings.

[2] This is a large county with considerable racial and ethnic diversity. Follow-up work will consider smaller and more racially homogeneous localities.

References

1 *13 U.S.C. Section 9.* URL: `https://www.law.cornell.edu/uscode/text/13/9`.

2 John Abowd, Daniel Kifer, Brett Moran, Robert Ashmead, Philip Leclerc, William Sexton, Simson Garfinkel, and Ashwin Machanavajjhala. Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge, 2019. URL: `https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf`.

3 Avery v. Midland County, 390 U.S. 474 (1968).

4 U.S. Census Bureau. *Disclosure avoidance system - End to End demonstration.* URL: `https://github.com/uscensusbureau/census2020-das-e2e`.

5 U.S. Census Bureau. *Disclosure avoidance system - End to End demonstration, L1 metric.* URL: `https://github.com/uscensusbureau/census2020-das-e2e/blob/3f2c9cf9cb3c33a4e2067bd784ff381792f7ffc0/programs/validator.py#L20`.

6 U.S. Census Bureau. *TopDown: Adding Geometric Noise to Counts.* URL: `https://github.com/uscensusbureau/census2020-das-e2e/blob/d9faabf3de987b890a5079b914f5aba597215b14/programs/engine/topdown_engine.py#L678`.

7 U.S. Census Bureau. *2010 Demonstration Data Products*, 2010. URL: `https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html`.

8 U.S. Census Bureau. *2010 Census Summary File 1*, 2012. URL: `https://www.census.gov/prod/cen2010/doc/sf1.pdf`.

9 U.S. Census Bureau. *Census P.L. 94-171 Redistricting Data*, 2017. URL: `https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html`.

10 Daryl DeFord, Moon Duchin, and Justin Solomon. Recombination: A family of markov chains for redistricting. *arXiv preprint arXiv:1911.05725*, 2019.

11 Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

12 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. *Halevi S., Rabin T. (eds) Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science*, 3876, 2006.

13 Peter Wayner JN Matthews, Bhushan Suwal. *Accompanying GitHub repository.* URL: `https://github.com/mggg/census-diff-privacy`.

14 Denis Kazakov. *Census Scripts GitHub repository*, 2019. URL: `https://github.com/94kazakov/census_scripts`.

15 U.S. Census Bureau Michael Hawes. *Differential Privacy and the 2020 Decennial Census*, 2020. URL: `https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf`.

16 National Conference of State Legislatures. *2010 Redistricting Deviation Table.* URL: `https://www.ncsl.org/research/redistricting/2010-ncsl-redistricting-deviation-table.aspx`.

17 Reynolds v. Sims, 377 U.S. 533 (1964).

18 Wesberry v. Sanders, 376 U.S. 1 (1964).

A ToyDown and TopDown

ToyDown is described in Algorithm 2. It uses the *Laplace distribution* $\mathrm{Lap}(b)$ with scale parameter b, i.e., the probability distribution over $\mathbb{R}$ with mean zero and probability density function $\mathbb{P}[L] = \frac{1}{2b}e^{-|L|/b}$. It has variance $2b^2$. TopDown uses the *geometric* distribution, a discretized version of the Laplace distribution with integer support.

The inputs to TopDown are as follows. $A_{H,T} = \{a_{h,t}\}_{h\in H, t\in T}$, where $a_{h,t}$ is the number of people in h of type t; $W = (Q_1, \ldots, Q_{|W|})$ is a *workload* consisting of a collection of histograms Q; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$ is a hierarchical allocation of the privacy budget, with $\varepsilon_\ell > 0$

at each level; $B : W \to [0,1]$ with $\sum_{Q \in W} B(Q) = 1$ is a probability vector describing the relative privacy budget on each histogram in the workload; *invariants* V; and *structural inequalities* S. We write $\boldsymbol{a}_h = \{a_{h,t}\}_{t \in T}$ (and $\boldsymbol{\alpha}_h$ analogously). For a query q, we write $q(\boldsymbol{a}_h) = \sum_{t \in q} a_{h,t}$ (and $q(\boldsymbol{\alpha}_h)$ analogously).

In the first stage (lines 2-5), a geometric random variable is added to the raw counts a to produce noised counts $\hat{a}$. In the second stage (lines 6-8), the noised counts are adapted to the nearest integer values that meet a collection of equality and inequality conditions. These equalities and inequalities, over the real numbers, describe a convex polytope; therefore the post-processing can be thought of geometrically as a closest-point projection to the integer points in the convex body under L^2 distance.

The noising stages of both ToyDown and TopDown are ε-differentially private for $\varepsilon = \sum_{\ell=1}^{d} \varepsilon_\ell$. In ToyDown, this stage can be viewed as generating a single histogram at each level ℓ using budget ε_ℓ. Following the Census Bureau, we use bounded differential privacy, wherein the global sensitivity of histogram queries is 2. In TopDown, the budget at level ℓ is further divided among the $|W|$ histograms Q in the workload, each receiving $B(Q)\varepsilon_\ell$ of the budget. Because ToyDown's post-processing is data independent, ToyDown is ε-DP. TopDown's post-processing is not data independent: the invariants and structural inequalities may depend on the original data.

Algorithm 1 TopDown, based on [2].

```
1: procedure TopDown($A_{H,T}$, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d$, $W$, $B$, $V$, $S$)
2:     for $h \in H$, $Q \in W$, $q \in Q$ do
3:         $\beta \leftarrow \exp(-B(Q) \cdot \varepsilon_{\ell(h)}/2)$
4:         $G_{h,q} \leftarrow \mathrm{Geom}(\beta)$                         ▷ See [6]
5:         $\widehat{a}_{h,q} \leftarrow q(\boldsymbol{a}_h) + G_{h,q}$      ▷ Geometric mechanism with
                                                  sensitivity 2, budget $B(Q) \cdot \varepsilon_{\ell(h)}$
6:     for $\ell = 1, \ldots, d$ do
7:         Compute hierarchically-consistent       ▷ A sophisticated heuristic algorithm
           non-negative integers $\{\alpha_{h,t}\}_{h \in H_\ell, t \in T}$          out of scope for this work
           minimizing $\sum_{h \in H_\ell} \sum_{q \in W_\ell} (q(\boldsymbol{\alpha}_h) - \widehat{a}_{h,q})^2$,
           subject to the invariants: $v^*(\boldsymbol{\alpha}_h) = v^*(\boldsymbol{a}_h)$ for all $h \in H_\ell$, $v \in V$
           and structural inequalities: $s(\boldsymbol{\alpha}_h, \boldsymbol{a}_h) \leq 0$ for all $h \in H_\ell$, $s \in S$
8:     return $\{\alpha_{h,t}\}_{h \in H, t \in T}$
```

B Detailed materials and methods

B.1 Primary data sources

2010 US Census demographic data was downloaded using the Census API, and the 2010 census block, block group, and tract shapefile for Dallas County were downloaded from the US Census Bureau's TIGER/Line Shapefiles. For our VRA analysis, we obtained both statewide election results and a statewide precinct shapefile from the Texas Capitol Data Portal, which we then trimmed to the precincts within Dallas County.[3]

[3] Data comes from `data.capitol.texas.gov/topic/elections` and `data.capitol.texas.gov/topic/geography`.

Algorithm 2 ToyDown.

```
procedure ToyDown($A_H = \{a_h\}_{h \in H}, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d$)    ▷ (Single attribute)
    for $h \in H$ do
        $L_h \sim \mathrm{Lap}(2/\varepsilon_{\ell(h)})$
        $\widehat{a}_h \leftarrow a_h + L_h$    ▷ Laplace mechanism with sensitivity 2, budget $\varepsilon_{\ell(h)}$
    for $\ell = 1, \ldots, d$ do
        Compute hierarchically consistent $\{\alpha_h\}_{h \in H_\ell}$
        minimizing $\sum_{h \in H_\ell} (\alpha_h - \widehat{a}_h)^2$
    return $\{\alpha_h\}_{h \in H}$

procedure MultiAttrToyDown($A_{H,T} = \{a_{h,t}\}_{h \in H, t \in T}, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_d$)
    for $h \in H$, $t \in T$ do
        $L_{h,t} \sim \mathrm{Lap}(2/\varepsilon_{\ell(h)})$
        $\widehat{a}_{h,t} \leftarrow a_{h,t} + L_{h,t}$    ▷ Laplace mechanism with sensitivity 2, budget $\varepsilon_{\ell(h)}$
    for $\ell = 1, \ldots, d$ do
        Compute hierarchically consistent
        (optionally, non-negative) $\{\alpha_{h,t}\}_{h \in H_\ell, t \in T}$
        minimizing $\sum_{h \in H_\ell, t \in T} (\alpha_{h,t} - \widehat{a}_{h,t})^2$
    return $\{\alpha_{h,t}\}_{h \in H, t \in T}$
```

We use a person-level dataset obtained by applying a reconstruction technique to the block-level data from Texas from the 2010 Census.[4] The reconstructed microdata records contain block-level sex, age, ethnicity, and race information consistent with a collection of tables from 2010 Census Summary File 1. We note that this reconstruction follows the same strategy used by the Census Bureau itself as the first step of its reidentification experiment [15], based on [11].

The reconstructed data is far from perfect. Unlike the Bureau, we do not have access to the ground truth data needed to quantify the errors. The Bureau's own reconstruction experiment reconstructed 46% of entries exactly, plus an additional 25% within ±1 year error in age [15]. We note that our reconstructed data contains no household information, because this was not present in the tables used in the constraint system. This is significant because the TopDown configurations for the US Census Bureau's 2010 Demonstration Data Products [7] include household-based workload queries and invariants.

B.2 TopDown configuration

The exact configuration files and code for all the runs are available in this paper's accompanying repository [13]. The TopDown code used for this paper was modified from the publicly available demonstration release of the US Census Bureau's Disclosure Avoidance System 2018 End-to-End test release [4]. The input data fed to the algorithm was obtained by restructuring the reconstructed 2010 block-level Texas microdata into the 1940s IPUMs data format. Most importantly, the reconstructions allowed for 63 distinct combination of races whereas the End-to-End release only allows for 6 races, so all multi-racial entries were re-categorized as Other in our TopDown runs.

[4] A team led by data scientist and journalist Mark Hansen at Columbia, including Denis Kazakov, Timothy Donald Jones, and William Reed Palmer, designed an algorithm to solve for the detailed data, which we describe in this section. Code is available upon request [14].

Because TopDown's post-processing is done level by level, the noisy counts in Dallas County do not depend on the noisy counts at the tract-level or below in counties other than Dallas. We modified the census reconstructed data to focus on Dallas county and minimize the computation time spent processing the other 253 counties in Texas. Specifically, for every non-Dallas county, we placed all of the population into a single block.

We do not enforce certain household invariants that the Census Bureau is planning to enforce, and our workload omits household queries that are used in Census's demonstration data products. Our choice to omit household queries and invariants is result of our use of reconstructed 2010 census microdata which does not include household information. We did perform additional runs with household invariants and queries using crude synthetic household data, the results of which are available in the data repository accompanying this paper [13]. In those runs, the population in each block was grouped into households of size 5 with at most one group smaller than 5. Ultimately, we focused on the experiments that did not require synthetic household data.

The TopDown runs without the household workload or invariants use a workload consisting of two histograms: $Q_{detailed}$ and $Q_{va,eth,race}$ with 10% and 90% of the budget respectively. (The additional runs with households includes an additional households and group quarters histogram in the workload assigned 22.5% of the budget, leaving 10% and 67.5% for $Q_{detailed}$ and $Q_{va,eth,race}$ respectively.) The End-to-End TopDown code reports a differentially private estimate of the L^1 error with $\varepsilon = 0.0001$ not included in privacy budget specified elsewhere in the configuration file and discussed elsewhere in this paper.

C District fragmentation

Algorithm 3 Greedy.

```
procedure GREEDY($H, k$)
    if $k = 1$ then
        Return $H$
    $N \leftarrow \lfloor |H_d|/k \rfloor$, $D \leftarrow \emptyset$, $h^* \leftarrow h_1$
    while $N > 0$ do
        For $h^*$ and $D$, let $S(h^*, D)$ be the set of
        children $h$ of $h^*$ that are disjoint from $D$.
        while $\exists h \in S(h^*, D) : |h| \leq N$ do
            $D \leftarrow D \cup h$            ▷ Associating $h$ with the blocks descendent from it
            $N \leftarrow N - |h|$
        Pick $h^* \in S(h^*, D)$
    return $D$
```

Algorithm 4 Square.

```
procedure SQUARE($H, k$)
    $s_d \leftarrow \sqrt{|H_d|/k}$                    ▷ Side length in blocks of the district
    $S_d \leftarrow \sqrt{n_1 \cdot n_2 \cdots n_{d-1}}$          ▷ Side length in blocks of the region
    Sample $i, j \in \{1, \dots, S_d - s_d + 1\}$ uniformly at random
    return $D_{i,j}$, the square district with top left corner at $(i, j)$
```

In Section 6, we defined the fragmentation score and its relationship to error variance for ToyDown, and analyzed the expected fragmentation score of districts produced by different district drawing algorithms. Now we apply TopDown to examine the relationship between a district's population error and geometry, as captured by the fragmentation score.

We fix the a total budget and an equal allocation across levels: $0.2 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = \varepsilon_5 = \varepsilon_6$, as in Table 1. (We do not need to noise the nation because we are focusing on Texas; we do need to noise Texas even though its total population is invariant, because its population by race is allowed to vary.) We apply ReCom to build districts out of tracts, block groups, and blocks – all of which are part of the census hierarchy – and add a realistic variant that builds from whole *precincts*. These are about the same size as block groups and are more commonly used in redistricting.

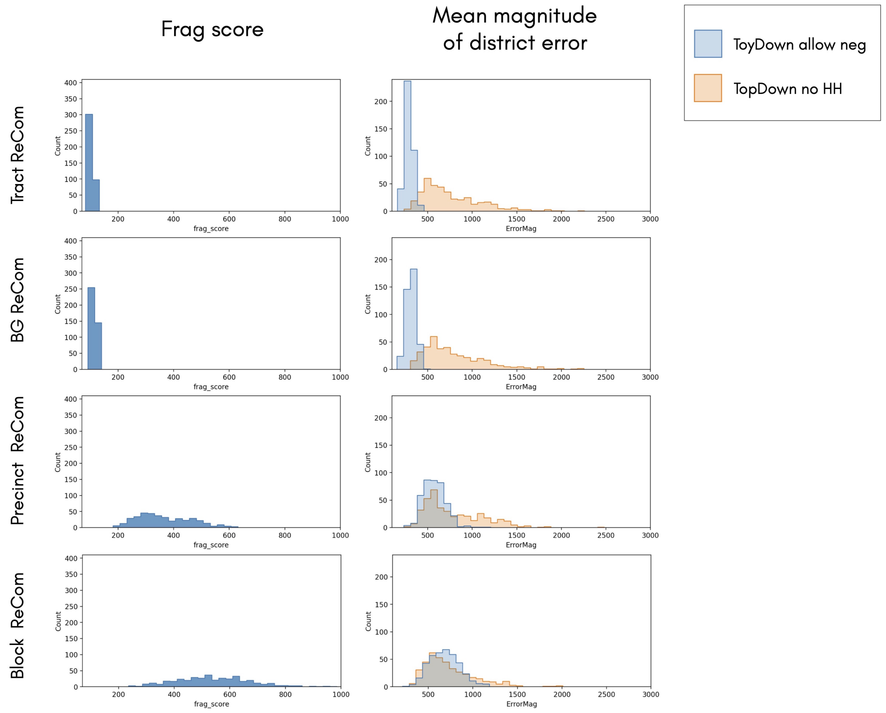


Figure 6 Do the building-block units of districts matter? Histograms of fragmentation score (left column) and mean error magnitude (right column) are shown across four district-drawing algorithms that prioritize compactness. (Dallas County, $k = 4$.) We see that using larger units leads to significantly lower fragmentation and correspondingly low district-level error in ToyDown, but the advantage erodes when we pass to TopDown.

Figure 6 plots the data from our experiments. Each of the 12 histograms displays 400 values, one for each district drawn by the specified district-drawing algorithm. The histograms on the left plot the fragmentation score of each district; the histograms on the right plot the mean observed district-level population error magnitude over 16 executions of the specified hierarchical noising algorithm.

The size of the constituent units is observed to have a controlling effect on the fragmentation score, as expected. As we would expect, this carries over to the simplest ToyDown (allowing negativity). (Note that since the error has zero mean, higher variance drives up the mean magnitude of error.) But the choice of base units makes far less difference by the time we move to full TopDown. These observations are consistent, again, with a strong similarity across spatially nearby units. All four kinds of ReCom will tend to produce compact, squat districts whose units are more closely geographically proximal than would be observed with disconnected or elongated shapes. Random noise is uncorrelated, but the post-processing effects can be highly spatially correlated because of spatial relationships in the underlying counts by race, ethnicity, and voting age.

D Robustness of noisy ER

Figure 7 extends the findings from Figure 5 with more splits and allocations, showing that as long as small precincts are filtered out, ecological regression for RPV analysis in Dallas County is robust to changes in the allocation of the privacy budget across the levels of the hierarchy and the total privacy budget for TopDown. The corresponding plots for ToyDown are essentially indistinguishable. (ER with precincts weighted by population is similarly robust.)

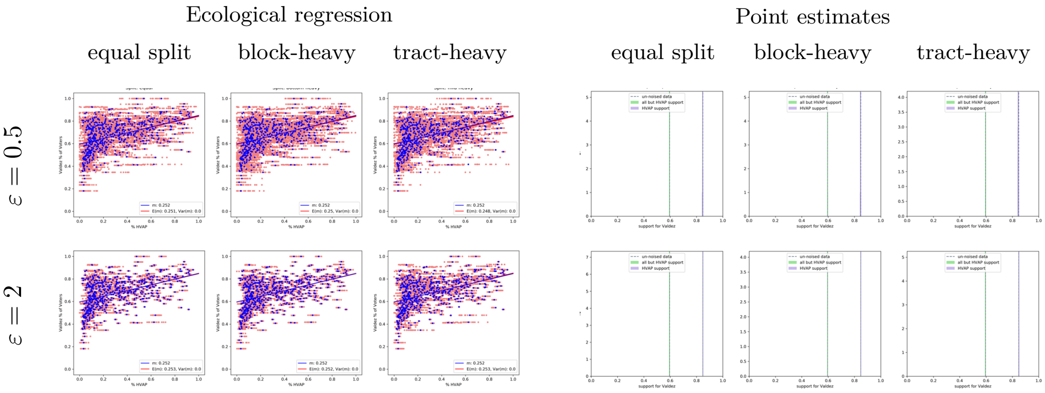


Figure 7 Ecological regression for the Valdez-White runoff election with $\varepsilon = .5$ and $\varepsilon = 2$ and three different budget allocations, together with corresponding point estimates for Latino and non-Latino support for Valdez, with small precincts filtered out as in Figure 5. Findings stay remarkably stable.