

From Election to Interpolation: Ballot Completion in Ranked Choice Voting

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Jay Lee

May 2019

Approved for the Division
(Mathematics - Statistics)

Heather Kitada Smalley

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

Introduction	1
Chapter 1: What is ranked choice voting?	3
1.1 Frequently Used Terms	4
1.2 Claims about RCV	4
1.3 History of RCV in the US (SF in particular)	6
1.4 Why, or why not, implement RCV?	6
1.4.1 Pros	6
1.4.2 Cons	10
1.5 Research into SF?	13
Chapter 2: Methods and Structure	15
2.1 Data Structure and Source	15
2.1.1 Election-specific data	15
2.1.2 Demographic data	16
2.2 Regressions	19
Chapter 3: Results	21
Conclusion	23
Appendix A: The First Appendix	25
Appendix B: The Second Appendix, for Fun	27
References	29

List of Tables

2.1	Combined Precincts - Original	18
2.2	Combined Precincts - Weighted	19

List of Figures

Abstract

One of the arguments against implementing ranked-choice voting (RCV) is that RCV is harder for voters to participate in. Two of the reasons for this are the more complicated ballot design and the extra effort that goes into forming a ordered preference of candidates. To evaluate this claim, we examine rates of ballot errors and undervoting (ranking fewer than the allowed number of candidates) in some American elections conducted with RCV. Results show that idk yet.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Welcome to the *R Markdown* thesis template. This template is based on (and in many places copied directly from) the Reed College LaTeX template, but hopefully it will provide a nicer interface for those that have never used TeX or LaTeX before. Using *R Markdown* will also allow you to easily keep track of your analyses in **R** chunks of code, with the resulting plots and output included as well. The hope is this *R Markdown* template gets you in the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build onto your results down the road.

Hopefully, you won't have much of a learning period to go through and you will reap the benefits of a nicely formatted thesis. The use of LaTeX in combination with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may have never had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities. After working with *Markdown* and **R** together for a few weeks, we are confident this will be your reporting style of choice going forward.

Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* also allows you to read in your data, to analyze it and to visualize it using **R** functions, and also to provide the documentation and commentary on the results of your project. Further, it allows for **R** results to be passed inline to the commentary of your results. You'll see more on this later.

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about the final appearance of their document should use *R Markdown*. Of particular use should be anyone in the sciences, but the user-friendly nature of *Markdown* and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should make it of great benefit to nearly anyone writing a thesis project.

Chapter 1

What is ranked choice voting?

Ranked choice voting (RCV), also known as the alternative vote (AV) or instant-runoff voting (IRV) is an alternative voting method to the first-past-the-post (FPTP) or “plurality” election system more familiar to American voters, where the candidate with the most votes wins. Each voter, instead of choosing their highest preference among a set of candidates for an office, ranks some subset of the candidates in order of preference. This system (or a close variant) is used in Australia, Maine, and some American municipalities: San Francisco, CA; Minneapolis, MN; and Cambridge, MA; among others.

The single-winner RCV tabulation algorithm generally proceeds as follows:

1. For each voter, identify their most preferred candidate that has not yet been eliminated. Count up these preferences by candidate.
2. If one candidate has a majority ($50\% + 1$) of the unexhausted votes, they are declared the winner and counting stops.
3. The candidate with the lowest number of votes is eliminated.
4. The ballots counted for that candidate are each transferred to the voter’s next choice if one exists, or if one does not exist the ballot is “exhausted” and removed from counting for further rounds.
5. Return to 1.

Most jurisdictions that use RCV have slightly different rules for edge cases and ballot errors, but this algorithm is what distinguishes RCV from other ranked voting systems (e.g. Borda, Condorcet, Contingent, etc.). A close variant of RCV is the single transferrable vote (STV) method¹, which can be used to elect multiple candidates, i.e. for a school board, instead of just one. In the US, this is used in Cambridge, MA and Minneapolis, MN to elect multi-member offices.

¹More accurately, RCV is the single-winner implementation of the STV algorithm.

1.1 Frequently Used Terms

Below are some definitions for frequently used terms later on. These are not all ubiquitous (for example, “undervote” has another meaning in most voting research), but we define them here for clarity later on.

- *Overvote*: when a voter ranks multiple candidates in the same slot. This slot is typically thrown out entirely in counting, because it’s often not possible to determine which candidate was preferred.
- *Undervote*: when a voter does not rank candidates in all of the slots available to them. This is different than other definitions of “undervote”, which refer to a voter participating in one election on a ballot but not another one. This is not a problem in counting, and is explicitly allowed in the laws of most jurisdictions. A plurality election analog would be voting in high-profile races (e.g. presidential), but not down-ticket decisions (e.g. local water board).
- *Skipped vote*: when a voter ranks no candidate at slot x , but ranks a candidate at slot $y > x$. This is typically not a problem in counting, but different jurisdictions have different rules about whether a voter’s ballot is exhausted at this point or continues on to their next ranked choice. Plurality voting has no analog to this, because each race only has one “ranking” (first!).
- *Duplicated vote*: when a voter ranks the same candidate for distinct slots x and y . This is typically not a problem for counting, and the first ranking for the candidate is used. Similar to a skipped vote, plurality voting has no analog to this.
- *Ballot exhaustion*: as ballot counting progresses, some ballots will become “exhausted” when all the candidates selected are eliminated. Suppose the final count in an election is between candidates B and D, and a voter ranked candidates C-A-E. Their ballot would not be counted in this final round, as they expressed no preference for either candidate B or D. An analogous situation in a plurality election might be voting in the general election but not a runoff, that is only having a say in part of the election.

Over-, skipped, and duplicated votes are really only interpretable as “ballot mistakes”: for example, even if a voter truly prefers two candidates equally, the ballot instructions (should) make it clear that ranking them at the same slot is not allowed.

1.2 Claims about RCV

There are plenty of arguments both for and against implementing RCV in place of plurality in different jurisdictions (see the literature review), but here we’ll focus on evaluating one major argument against it - RCV is harder for voters to participate in than a plurality system. There are two major reasons cited for this:

- The physical design of an RCV ballot is usually more complicated than a plurality ballot, because there has to be a system to encode a more full preference among the candidates than just selecting one candidate
- The process of forming a multi-candidate preference inherently takes more mental energy than just choosing a favorite candidate

The first facet of this argument should be reflected in ballot errors made by voters. Compared to plurality voting, we expect more errors in an RCV ballot just because the ballot is more complicated. There are also more potentials for error in the RCV system generally. The only “errors” in a plurality ballot are incompletely marking a candidate (think incorrect Scantron bubbling, or hanging chads) or overvoting, both of which are potential pitfalls for a ranked choice ballot as well. On top of these, there are the potential errors of duplicated and skipped votes unique to ranked ballots².

The second facet should be reflected in incomplete ballots filled out by voters. Given that they understand how to encode their preferences on the ballot, there is still the non-trivial task of forming such a preference. Structurally, some of the factors that should affect this incompleteness are:

- The number of candidates running for a position
- The number of candidates voters can rank
- The number of seats elected in a given race

This first variable is at the election level (different for every election), the second is at the jurisdiction level, and the third is a mix of both. For a clear example of these differences, consider a 2016 San Francisco Board of Supervisors race (District 3) versus a 2017 Cambridge City Council race.

Factor	San Francisco 2016	Cambridge 2017
Candidates running	2	27
Candidates rankable	2 (Generally, up to 3 ³)	27 (Generally, all)
Seats elected	1	9

²These types of errors are not uniform, and some jurisdictions are more forgiving than others about rules for counting these errors. While it may be apparent that a voter who listed the same candidate 3 times (A-A-A) prefers that candidate, a candidate ranking of A-B-A is harder to extract a clear preference from. Skipped votes are where we see the most variance in jurisdiction counting rules: if a voter marks the ballot A-___-B, skipping the second slot, some jurisdictions will ignore the skip and treat B as the voter’s second choice, while others will stop counting after A is eliminated (ignoring their vote for B), and others yet will throw out the ballot entirely.

³This is changing in 2019 to be up to 10.

1.3 History of RCV in the US (SF in particular)

In the United States, there have been two major periods of RCV implementation in various jurisdictions. Between 1915 and 1950, 24 American cities chose to institute RCV as a form of local election. By 1965, however, all of these except for Cambridge, MA had eliminated the policy change. Then, in the 2000s, there was a resurgence of uptake in a different set of American cities⁴, including Minneapolis and a handful in the San Francisco Bay Area. While Cambridge has consistently used the multi-winner (STV) method to elect City Council and School Board seats, the modern resurgence of RCV almost universally deals with single-winner elections. Research argues that RCV appears in jurisdictions where there is strong multi-party support for the reform - the RCV method itself gives individual parties less power in the election process, so powerful single parties usually don't have reason to support it.

1.4 Why, or why not, implement RCV?

There are plenty of arguments on both sides of implementing RCV in jurisdictions that consider it.

1.4.1 Pros

No secondary elections

There are two major types of “secondary elections” used in American voting: primary elections and runoff elections. Primaries are used by political parties to select their nominee for a general election, so the voters of any one party aren't split between different candidates. Runoffs are most often used when no candidate in the general election surpasses 50% of the vote total. Typically the top two candidates from the general election⁵ advance to a later runoff. These secondary elections face two main challenges: low turnout and high cost.

Secondary elections as a whole face low turnout (Wright, 1989; Ranney, 1972). Reasons: Research shows that people don't actually like voting that much - the more frequently elections are held, the lower turnout will be for all of them generally (Boyd, 1986). Secondary elections increase the number of elections in a period, so this is one possible reason why they generally have low turnout. Further research indicates that holding elections concurrently with a presidential election “increase[s] the likelihood that citizens will vote” (Boyd, 1986). This is seen in off-year Congressional elections, where turnout drops from presidential years. Typically general elections are held concurrently with presidential elections (second Tuesday in November, super high media coverage, lots of voter outreach, yadda yadda), so secondary elections cannot be held at the same time as a presidential election and they should thus suffer in

⁴Mostly in the American West: there are 9 cities west of the Mississippi River currently using RCV and only 4 east of it

⁵Or primary election - Seven Southern states require primary winners to obtain 50% of the vote to get on the general election ballot, and some other states have a requirement of 40%. (WaPo article)

turnout. This low turnout has consequences for representation in the system. The same research (Ranney, 1972) finds that while primary voters are not ideologically unrepresentative of general election voters, they are both demographically unrepresentative and unrepresentative on some major issues. Traditional knowledge holds that primary voters are more committed partisans than general election voters, leading the eventual candidates in a general election to be polarized away from the “center” of political ideas (double check this but I’m pretty sure the cite is Hill’s *Instant Runoff Voting*).

The higher costs associated with secondary elections are a little more intuitive than turnout issues - it takes money to hold elections. Pollworkers have to be paid, facilities have to be reserved, and candidates have to do more campaigning. A 2011 City Council runoff in Plano, TX cost the city an extra \$73,000 (Plano Star Courier, 2011). A 2012 Alabama runoff for multiple seats cost the state about \$3 million.

Since RCV eliminates the need for primary and runoff elections while still ensuring majority rule (which is the main reason for these elections), it should avoid the problems of lower turnout and higher costs associated with secondary elections⁶. Summed up, RCV decreases costs and “boosts turnout via elimination of low-turnout elections” (Morales, 2018).

Ensures majority rule

In jurisdictions without rules for 50% minimums, a common phenomenon is a candidate winning an election with less than 50% of the vote (a plurality, rather than a majority). The major conceptual issue with this is that more people preferred a candidate other than the one who was elected⁷. RCV requires that a winning candidate receive at least 50% of the votes remaining⁸, ensuring that a majority of voters prefer the elected candidate to other candidates.

This is particularly important in jurisdictions (like Maine) with strong third-party support and more than two viable candidates. Former Maine Governor Paul LePage, a Republican, won his first election in 2010 with 38.1% of the vote, compared to Independent Eliot Cutler’s 36.7%⁹.

Reduces strategic voting - spoiler candidates and third parties

Strategic voting¹⁰ is a scenario where a voter does not reflect their true preferences on their cast ballot in order to affect the outcome of the election. For example, a third party supporter may cast their vote for one of the two major party candidates because otherwise they feel like “it won’t count”.

The “spoiler effect” is when a third party candidate draws votes away from the ideologically closest major party candidate, thus contributing to the election of the

⁶Or at least some of it - costs overall are indeterminate (Rhode conference paper)

⁷The ‘ideal’ for electoral systems is the Condorcet condition: the candidate elected should beat all other candidates in one-on-one contests.

⁸See below for issues with this ‘remaining’ concept.

⁹a margin of about 7,500 votes. Democrat Libby Mitchell received 19%.

¹⁰Also known as “tactical” or “insincere” voting.

other major party candidate. The most recent large-scale accusation of this was in the 2000 election. Green Party candidate Ralph Nader drew about 3% of the national vote, more than the margin of victory for George W. Bush over Al Gore¹¹. In the especially consequential state of Florida, Nader took 1.6% of the vote: almost 200 times greater than the margin between the two major party candidates of less than .01 percentage points¹² (source from FEC). Many believed that Nader, generally seen as more liberal than the Democrat Gore, drew votes from the Democratic base that would have helped Gore win the election otherwise. While research into third-party voters casts some doubt on this theory's applicability in 2000 (Herron, Lewis)¹³, public opinion still rests on the idea that Nader cost Gore the presidency¹⁴. In fact, a pro-Republican PAC aired campaign ads promoting Nader in Democratic states in an attempt to pull votes from Gore (Meckler, 2000).

One of the reasons for major party voters to support RCV is that it avoids this spoiler issue. Under a plurality system, voters are discouraged from voting for third-party candidates because it could help elect their least-preferred of the two major candidates - "the greater of two evils", so to speak. Under RCV, however, since voters can be heard throughout multiple rounds across separate candidates, third-party voters can vote for their most preferred candidate, then still have their vote count for a better major-party candidate if their first option is eliminated.

Conversely, one of the reasons for third-party voters to support RCV is that it helps third party candidates get elected. Voters can ignore this aforementioned facet of strategic voting¹⁵ and select their truly preferred candidate. Third-party supporters who were worried about the spoiler effect, then, can vote for their true preference of a third party and not inadvertently help a less-preferred major candidate get elected. As people abandon this strategy, third parties will receive more votes from people no longer worried about the spoiler effect, and this could get third party candidates elected.

Disincentivizes negative campaigning

Ranked choice voting should incentivize candidates to avoid negative campaigning. In a plurality election, since candidates don't care about voters who are committed to their competitors, a well-thought out negative campaign will only ostracize voters who were never going to support another candidate in the first place, and perhaps bring

¹¹Admittedly, the margins are less clear-cut than this at the state level, where the margins actually matter for the Electoral College.

¹²Not to point fingers at Nader alone in this case - while he was the most popular third party candidate by far, all 8 official third-party candidates received more votes than the major-candidate margin of only 537 votes

¹³In short - while Nader's Florida voters potentially would have broken enough for Gore to put him over the top, this was more a factor of the unusually close margin between the two major candidates than anything that Nader aided in particular.

¹⁴One of the sections of Nader's Wikipedia page is entitled "Spoiler controversy" in regards to this election.

¹⁵Or any strategic voting - while not impossible, it's infeasible to vote strategically under RCV (Bartholdi and Orlin, 1990).

more swing voters to their side. Under RCV, however, alienating another candidate's voters could backfire in the event that candidate is eliminated and these voters decide to support your opponent in the next round, causing your defeat. Research supports this - a 2016 study showed that voters in cities using RCV are more satisfied with campaigns than in cities who use plurality methods, and consider the campaign to have a less negative tone overall (Donovan et al.). In San Francisco's first RCV election, there were joint fundraisers between candidates, and one district even saw regular "Candidates Collaborative" public meetings between many candidates to discuss issues affecting the district, where "the setting [was] decidedly congenial" (Murphy, 2004).

An interesting case study of this phenomena is in the 2018 San Francisco mayoral election. There were three frontrunners heading into election day, all incumbent members of the city's Board of Supervisors: London Breed, Jane Kim, and Mark Leno. As polls showed Breed ahead about a month before the election, Kim and Leno held a joint press conference to endorse the other as voters' second choices. By drawing second-choice votes from the other candidate, the remaining candidate hoped to overcome the gap between them and Breed. In the actual election, the standing when it came time to proceed to the final round of counting was 102,767 for Breed, 68,707 for Leno, and 66,043 for Kim. While a significant proportion of Kim's voters transferred to Leno after her elimination, in the final round Breed surpassed Leno by about 2,000 votes.

Though it's outside the scope of this research to tell if this cross-endorsement was effective¹⁶, there is some evidence in favor of this theory. Leno received almost 70% of the votes previously counted for Kim compared to Breed's 20%, bringing Breed's final margin of victory down to only 1 percentage point. In previous rounds of the election, no single candidate ever received more than 35% of the transferred votes from an eliminated candidate¹⁷, so this is at least an unusual observation.

Minority representation

While much of the American literature on minority representation under voting systems has focused on gerrymandering and single- versus multi-member districts, there is some study of representation under RCV independent of the number of seats up for election.

John, Smith, and Zack (2018) find an increase under RCV in the number of racial/ethnic minority candidates who run in the Bay Area, and an increase in the chance that women and minority women win their election. Their theory for this result is that RCV lowers the barrier to entry in an election, making it more feasible for minority candidates to campaign, and that women (particularly women of color) are better suited in general to the less negative campaigning and coalition-building that RCV promotes in candidates.

Because minority voter turnout in secondary elections is significantly worse than

¹⁶Other confounding factors could exist: maybe Kim and Leno had similar enough positions that this scenario would have happened without the endorsement, maybe this number is only significant because in the final rounds there were only 2 candidates for second choice votes to flow to, etc.

¹⁷Except for round 2, where all 3 votes for the same write-in candidate transferred to Breed.

white turnout (more so than in a general election), the elimination of these secondary elections should increase the relative say of minority voters in elections overall (Callaghan, 2017).

Turnout improvements

Voting system reform advocates claim that these improvements to the voting process will boost turnout by generally increasing public trust in the effectiveness of elections. While there are many different strategies employed by campaigns and advocacy groups to boost turnout, these methods don't have as much effect as changing methods to ranked voting. Phone and direct mail get-out-the-vote campaigns "typically [yield] less than 1 percentage point boosts in voter turnout", and while in-person efforts are better they are less efficient at contacting voters.

Another argument is a little more complicated to reason through - "[IRV] boosts turnout via elimination of low-turnout elections". While we typically think about turnout in terms of the number of people voting across similar elections (e.g. change in turnout from 2010 to 2014 in Congressional races), RCV improves turnout by (allegedly) increasing the number of people who get to participate in an election overall compared to the number of people who participate in typically low-turnout primary and general elections. (all Morales, 2018)

1.4.2 Cons

What is a majority, anyway?

Under RCV, the problem of ballot elimination can result in "majorities" that actually aren't. Since it's not always possible to voters to rank every candidate¹⁸, there will almost certainly be some number of voters who did not list a ranking for any of the candidates still remaining in the final round of the election. Thus the majority that the winner has collected is only a majority of the unexhausted ballots, which may make it less than 50% of the total counted ballots (Petrangelo, 2013). This does introduce a less expansive form of the spoiler problem - while voters aren't "punished" for putting a less electable candidate for their first choice, they are punished for filling all three available slots with minor candidates. In a 2015 study of four RCV elections, Burnett and Kogan found that all four had enough eliminated ballots to only give the winner an overall plurality, not the majority sought after. This is a problem that requires a substantial fix, as "even individuals who mark three distinct choices often face the prospect of exhaustion, so education alone will not fix the problem" (Burnett and Kogan 2015, p. 49).

A research question that might address this issue of majorities is whether the pluralities generated by RCV are generally "bigger" (closer to the ideal of a true majority) in some way than under FPTP, and how this is affected by technical rules

¹⁸This may be disallowed, as SF only allows you 3 rankings (due to technical limitations), or it may just be infeasible - Cambridge often has 20+ candidates on the ballot.

like the number of candidates that voters are allowed to rank. Some research¹⁹ shows that RCV does perform better than a two-round runoff in this case - as a percentage of the voters in the first round, RCV consistently has more voters counted in the final round compared to top-two runoffs (Richie, Brown, 2017).

Legal challenges

(All sourced from Maine Legislature) In addition to the political hurdle of passing RCV legislation²⁰, there are legal challenges as well. In the aftermath of Maine's recent ballot initiative instituting RCV, legal questions were brought to the state Supreme Court by the State Senate, due to conflict between the initiated bill's language about a "majority" versus the state constitution's language requiring a "plurality". There was a ruling from the court that resulted in an amended law instituting RCV for only primary elections, but a "People's Veto" later occurred through initiative that removed the law, followed by a later suit by the state Republican Party to block RV. Maine has so far used RCV in 2018 for both the primary and general elections to elect certain offices.

This anecdote exemplifies the issue - while public support may be behind RCV in jurisdictions that enact it, legal challenges abound. A (failed) legal challenge was brought against RCV in San Francisco by a defeated candidate in 2011 (Hawkins, 2011). Another challenge (also failed) was brought by Bruce Poliquin against Maine after his loss to Jared Golden in their 2018 Senate race (Mistler, 2018). In Pierce County, WA (home of Tacoma), RCV was repealed by voters after the US Supreme Court sided with a challenge to reinstate Washington's top two primary-runoff system. With the re-implementation of the top two primary, there was no need to have an RCV election (between two candidates) and the measure was repealed (Eberhard, 2017).

More complicated tabulation

By its nature, the tabulation of RCV elections is more complicated than counting results in plurality races. There are multiple rounds, so counting takes longer. Typically ballot counts are done at the precinct or county level, and then numbers are sent to a state elections office for full tabulation and verification of the results. Under RCV, however, the process typically requires all ballots to be sent (physically or electronically) to the state office for the multiple rounds of counting, further increasing the time needed to count. Additionally, if the jurisdiction is unable to obtain software to count ballots, hand counting of the ballots is necessary. This increased time in counting is an issue because "The elapsed time between Election Day and providing some results is one of the critical factors to maintaining voter engagement and trust in the process" (League of Women Voters of Maine, Comments on Proposed Rules for RCV).

¹⁹from the pro-RCV advocacy group FairVote, admittedly

²⁰Neither major party is particularly interested in pushing the issue, and Republicans in particular are often opposed to it (Woodard, 2018).

In jurisdictions with some form of electronic voting, the cost of transitioning to RCV can be a disincentive as well. Not all jurisdictions have hardware capable of conducting an RCV election, and of those that do the vendor (typically jurisdictions contract to election system vendors) may not currently have software that can tabulate the results of the election. Getting around these problems is significant, as it requires either a technology upgrade²¹ or a switch to some combination of paper ballots and hand-counting. (Morales and Eberhard, 2017)

Less intuitive rules

The plurality idea, while clearly not the only voting rule out there, is one of the most simple. One of the problems with RCV is that it doesn't always satisfy this idea simply - RCV does not always agree with the plurality method on choosing a winner. In the Senate race in Maine between Bruce Poliquin and Jared Golden, while Poliquin was on top at the end of the first round of counting, neither had 50% and Golden took the lead (and the election) after other candidates were eliminated and their votes transferred to voters' second choices (Portland Press-Herald). The more complicated process and this seemingly unfair (at first glance) rejection of plurality means that voters in general have a harder time understanding and trusting the results of an RCV election. This complication leads to an increase in overvotes (Kimball), which causes more ballots to be fully or partially rejected in the counting process. RCV also sees increased costs (particularly for the first election conducted) in voter education - In Minneapolis' first RCV election, 30% of the estimated \$365,000 additional cost²² of RCV was spent on voter education (Kimball, 2010).

Three types of errors - fatigue (intentionally opting out, or accidentally skipping a portion of the ballot), confusion, and lack of information to form a preference (Neely and Cook, 2008).

Lack of true adoption by voters, only listing first choice

While RCV provides the opportunity for voters to rank multiple candidates, not all voters will do so²³. This is entirely reasonable - forming a complex preference between multiple candidates is naturally more difficult than selecting a favorite from the same set. This phenomenon of undervoting, however, means that some ballots become exhausted in the course of tabulating the election and some voters don't have a say in the final round of the election. If a voter only lists their first choice and skips the second and third choices, they lose out on two extra rounds of tabulating where their vote still counts towards the decision.

²¹This is not a problem unique to RCV - election technology across the country is aging and in need of replacement (NCSL, Funding Elections Technology).

²²About a third of this total was one-time costs for implementing the new system.

²³Which is fully legal - nobody is forced to vote.

1.5 Research into SF?

The San Francisco Bay Area in particular has been an RCV hotspot in the modern American resurgence, so some good literature exists specifically studying the impact of RCV there.

Dennis (2004) finds higher rates of overvotes in SF's first RCV election (November 2004) than were expected, somewhat positively correlated with the number of candidates on the ballot in each district. Almost a third of the ballots contained an undervote. Looking at racial and ethnic groups, while Hispanic voters were more likely to find the RCV method easy to use than Asian voters, the former were also more likely to have completed their ballot incorrectly. One possible explanation of this phenomenon is that advocacy groups "succeeded in heightening awareness about the new voting system amongst the Asian community" prior to the election, increasing their wariness of it as well.

McNeely and Cook (2008) extend this basic analysis of ballot errors to explore some theories related to racial and ethnic divisions (and others). There was a significant decrease in undervoting among racial and ethnic minority groups across 4 years of elections, and a significant increase in undervotes among women as well. Districts with more candidates²⁴ and more campaign spending were less likely to have undervotes. The likelihood of ranking all 3 available candidates was affected by racial and ethnic categories, but not as much as factors such as prior exposure to RCV and available information (campaign spending, number of candidates).

SF Examiner article (from Elections employees, fairly) about how good it's going - lower costs, more minority candidates/officeholders, high turnout, low over/undervoting rate, low number of exhausted ballots

²⁴These also had more overvotes, in keeping with Dennis' findings.

Chapter 2

Methods and Structure

Inevitably, the data format that would be ideal to conduct this research doesn't exist publicly (as it shouldn't - voter anonymity and such). The ideal model might be a logistic or other classification model that predicts whether a given voter has a ballot error or fills out the ballot incompletely, given their demographic qualities.

2.1 Data Structure and Source

2.1.1 Election-specific data

The data used comes from two main sources. From the San Francisco Elections Office , we have a cast ballot record of the election in question . This is presented by the city as two text files: - The ballot image, a 45-character fixed-width file with fields corresponding to election, candidate, unique voter number (anonymized and disconnected from any voter registration ID), ranking, precinct, and other information. Each of these is encoded numerically, so each line of the file appears as a 45-digit number. - The lookup table, an 83-character fixed-width file that defines the encodings used to create the numerical values in the ballot image.

In the ballot file, each voter is spread across three rows of data (one for each ranking). We use the `clean_ballot()` function from our `rcv` R package to read in the data for easier manipulation. From this data we have the full ranked preference set of candidates (up to 3) from each voter.

Since the voter-level data is fully anonymized, we have no demographic information at the individual level. For any given voter, since the most identifying piece of information we have about them is their precinct, it is impossible to know any one voter's gender or ethnicity. To gain insight into these demographic trends, we instead aggregate up to the precinct level.

At the precinct level, we can now only study the rates of these ballot phenomena. For example, in Precinct 1703, we may see 14% of voters undervote. Rather than building a classification model (error / no error), we can now instead build regression models with a numerical dependent variable - the rate of ballot error. Moving up this level of abstraction does remove some granularity from the model (inference and prediction at the precinct level is less specific than at the individual level), but this is

the least we can do and still be able to access demographic information.

2.1.2 Demographic data

Now that we have precinct-level rates, we need to obtain precinct-level demographic information in order to build a model. Our source of demographic data is the 2010 U.S. Census (citation). While this data is slightly out of date (about 8 years), there is more certainty about the accuracy of the measurements collected. While a more temporally accurate data set like the yearly American Community Survey (ACS) could have been used, the error margins of this data set proved too large / complicated to be useful for this study.

One consideration with this is that the voting population is not always representative of the general population. It might be more informative to obtain demographics about the voting population specifically, rather than the entire population of each precinct. However, the only information available about voters is their age (from the county voter registration file), and the discontinuity introduced by using two different data sets for demographic information is more of an “error” to me than using a less accurate (but universal) census data set.

We now encounter a problem. Since census regions (block groups, tracts) are set by the federal government, and election precincts are set by San Francisco County¹, the regions don’t line up nicely². Given this mismatch, how do we obtain demographic estimates for our precincts?

Stated more generally: if we have a division of a geographic region and some set of properties on the divisions, how do we estimate measures of these properties for other possible divisions?

Areal Interpolation

One field of GIS (geographic information system) research that looks into this is called *areal interpolation*. The goal of areal interpolation is to take a variable distributed over a set of “source zones” and estimate its distribution over a set of “target zones”. These two zoning systems must overlap at least partially to get any estimate for the target zones (e.g. information about Oregon doesn’t directly tell us anything about information in Washington, under any zoning system), but are incompatible in some way³.

Types of AI:

- Areal Weighting (what I’m doing)
- Density weighting (doesn’t really apply I think)

¹The city and county government are unified in this case, because the county comprises entirely of the city of San Francisco.

²There’s no inherent reason that they should, it’s just unfortunate for this study.

³A set of “compatible” zoning systems would be something like the US Census’ hierarchical systems: multiple blocks are combined to make a block group, multiple block groups are combined into a census tract, etc. Since the areas overlap neatly, you can directly add together certain numbers (like population) from block groups to get a very accurate estimate for the census tract.

Talk about intensive versus extensive data and how that differs.

Probably drop this paragraph. There are a few possible ways of accomplishing this. The way we use assumes a uniform spatial distribution of all non-constant properties: any sub-region has proportion of the full region's property equal to the proportion of the area of the full region that the sub-region occupies. For example, if we split a region with 100 people into two sub-regions of equal area, we assume that there are 50 people in each sub-region. To approach this method, we first spatially intersect the two division sets of the city: census tracts and election precincts. This creates a new division, where every new sub-region is a unique combination of tract and precinct. We can directly group these sub-regions together in one combination into the full set of precincts, and group them in another combination into the full set of census tracts. Using the uniform spatial assumption, we obtain demographic estimates for each of these sub-regions, then add them together in the precinct grouping to obtain demographic estimates for our precincts.

This assumption of uniformity, like any assumption, may be unfounded. There are methods, some involving satellite imagery and street location, to get more accurate information about population distribution. Quite honestly, these were just too complicated for the scope of this research. It is, however, more likely in an urban environment like San Francisco⁴ that this uniformity assumption will not drastically change our estimates.

Shit to add

R package used, and why Validation - general choropleth matching Talk about percentages versus counts, once you figure out what's going on there

Precinct Consolidation

One peculiarity in San Francisco is the combination of certain precincts during elections. California state law allows for counties to “consolidate” precincts with low numbers of registered voters during an election. This eases administration by not requiring counties to set up and staff a full polling place in a precinct with few voters, while still giving voters a physical polling location in their approximate area⁵. While San Francisco does this less often than other counties, there are still some precincts that are consolidated each election.

The areal interpolation process produces demographic estimates for areas in the map, however, which are non-consolidated. This causes a mismatch: we have demographic data for individual precincts, but election data for the consolidated precincts. To address this issue, we have split the election data (a count of overvotes and undervotes) in the consolidated precincts into their two component precincts. This

⁴as compared to the entire state of California, which has urban pockets within suburban regions within a rural overall condition

⁵This method of precinct consolidation may change with the 2018 California Voter's Choice Act, which lets counties (except for Los Angeles County) move fully to vote-by-mail in combination with voting centers (California Senate Bill 450).

split is weighted by the population of each precinct in 2010⁶. For example, suppose Precinct X/Y had 100 undervotes, the population of Precinct X in 2010 was 325, and the population of Precinct Y in 2010 was 175. Then the “population split” between X and Y is 65%-35%, and we adjust the undervotes accordingly: Precinct X should have 65 of the 100 undervotes, and Precinct Y should have the remaining 35.

Examples of this in the data are below. The first table is the initial data, with full consolidated counts doubled in the precincts (and thus doubled across rows), and the second table is the data with proper weights applied.

```
original <- sf_unadjusted %>%
  as.data.frame() %>%
  select(1,6,2:5) %>%
  filter(str_detect(precinct, '/')) %>%
  slice(1:6)

knitr::kable(original,
  col.names = c("Election Pct.", "Pct. Number", "Overvotes", "Not over", "Undervotes", "Not under"),
  caption = "Consolidated Precincts - Original Data",
  caption.short = "Combined Precincts - Original",
  longtable = TRUE,
  booktabs = TRUE)
```

Table 2.1: Consolidated Precincts - Original Data

Election Pct.	Pct. Number	Overvotes	Not over	Undervotes	Not under
Pct 1104/1105	1104	1	459	101	359
Pct 1104/1105	1105	1	459	101	359
Pct 7509/7511	7511	1	491	132	360
Pct 7509/7511	7509	1	491	132	360
Pct 7527/7528	7527	5	396	98	303
Pct 7527/7528	7528	5	396	98	303

```
adjusted <- double_cases %>%
  as.data.frame() %>%
  select(1,6, weight, 2:5) %>%
  mutate_at(vars(ends_with('count')), function(x) x * .$weight) %>%
  slice(1:6)

knitr::kable(adjusted,
```

⁶The most simple weight would be 50-50 for each split, but this is inaccurate for many of these precincts: see the gap in the weights for Pct 7527/7528. The flaw in this method was discovered after calculating turnout of over 1000% for Precinct 7527...

```
col.names = c("Election Pct.", "Pct. Number", "Weight", "Overvotes", "Not ov
caption = "Consolidated Precincts - Weighted Data",
caption.short = "Combined Precincts - Weighted",
longtable = TRUE,
booktabs = TRUE)
```

Table 2.2: Consolidated Precincts - Weighted Data

Election Pct.	Pct. Number	Weight	Overvotes	Not over	Undervotes	Not under
Pct 1104/1105	1104	0.4646860	0.4646860	213.290865	46.933284	166.822267
Pct 1104/1105	1105	0.5353140	0.5353140	245.709135	54.066716	192.177733
Pct 7509/7511	7511	0.7169455	0.7169455	352.020218	94.636800	258.100364
Pct 7509/7511	7509	0.2830545	0.2830545	138.979782	37.363200	101.899636
Pct 7527/7528	7527	0.0167966	0.0839828	6.651434	1.646062	5.089355
Pct 7527/7528	7528	0.9832034	4.9160172	389.348566	96.353938	297.910645

2.2 Regressions

Methods for regression:

- Linear ?
 - Pretty simple honestly
- Logistic with the “binomial-style” input data
 - This is more useful because our output (rate) is 0-1 limited.
 - It also overweights precincts with more people - is this good?
- For whatever goes in, I need to pick a better model selection method.

Chapter 3

Results

I have some results and they are here.

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readability and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is  
# installed and loaded. This thesisdown package includes  
# the template files for the thesis.  
if(!require(devtools))  
  install.packages("devtools", repos = "http://cran.rstudio.com")  
if(!require(thesisdown))  
  devtools::install_github("ismayc/thesisdown")  
library(thesisdown)
```

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

- Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl*. Boston, MA: Addison Wesley Longman.
- Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime*. Boston, MA: Wesley Addison Longman.
- Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.