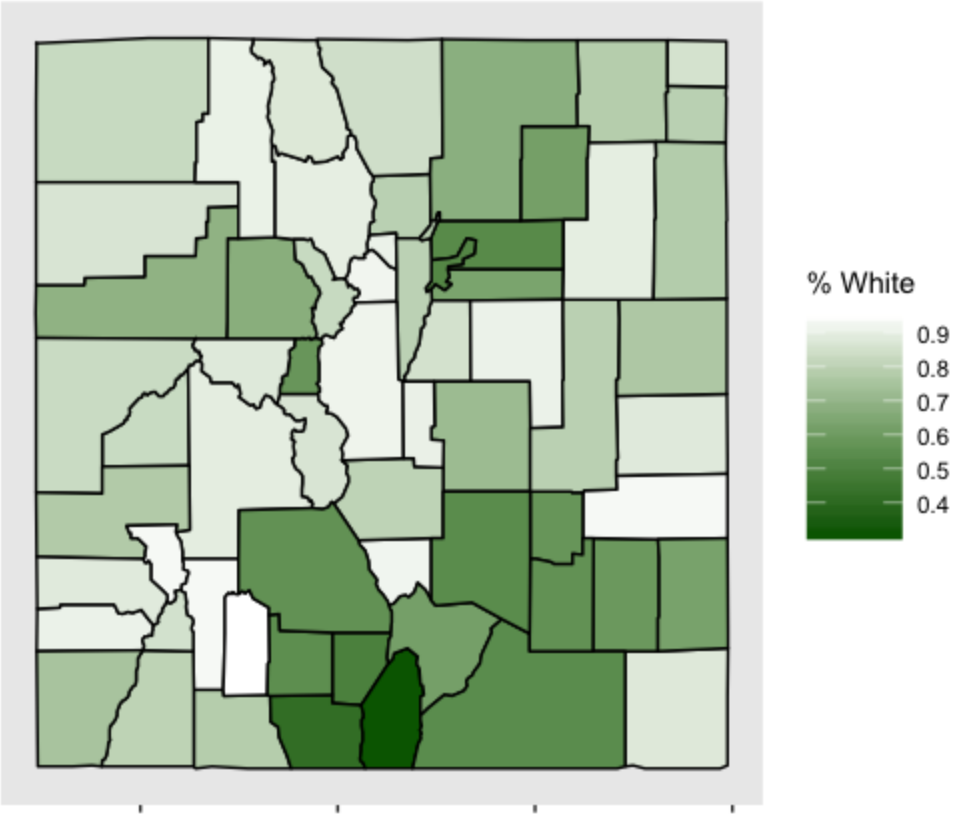

Figure 1: White voters per Colorado county

Case Selection, Data, Model Parametrization

The Centennial State and Its Voters

Demographics

Colorado, named the Centennial State due to assuming statehood on the centennial of the Union, lies in the Southwestern United States, with its Western half squarely atop the Rocky Mountains. Based on its estimated population of just over 5.5 million, Colorado is the 21st most populous state, and ranks 37th in population density. The vast majority of that population is gathered in a series of urban areas that comprise a North-to-South strip in the middle of the state, containing the Denver-Aurora-Lakewood Metro Area, Colorado Springs, Pueblo, and Fort Collins. Apart from the Western town of Grand Junction, the rest of the population resides in vast rural areas.

Colorado is landlocked, and far from any coastal town; in place of seaside resorts, Colorado attracts a substantial amount of tourists to its mountains every year. They also heavily depend on federal money and protection for national parks. Colorado has a median age of 34.3 and median household income of $65,685. Colorado's population is mostly white, with a higher minority group population density in its Southern regions, as shown in figure 3.1. [@us_census_bureau_us_2010]. The conclusion here is that Colorado is a relatively young, mostly white, and fairly well-off state that is increasingly getting more diverse, particularly in the South. These factors are important as they serve to associate Colorado with other states; such associations are useful for the replication of this study or the generalization of my results.

The State Capital is Denver; Colorado is split into 64 Counties, of which the most populous are, in no

particular order, the following: El Paso, Denver, Arapahoe, Jefferson, Adams, Larimer, Boulder, and Douglas. These counties comprise 73% of the total population of Colorado.

Table 1: Colorado population for largest counties

County	Total Population	CO Population %	Largest Metro Area
Adams	441603	0.08781	Denver-Aurora-Lakewood
Arapahoe	572003	0.1137	Denver-Aurora-Lakewood
Boulder	294567	0.05857	Boulder
Denver	600158	0.1193	Denver
Douglas	285465	0.05676	Denver-Aurora-Lakewood
El Paso	622263	0.1237	Colorado Springs
Jefferson	534543	0.1063	Denver-Aurora-Lakewood
Larimer	299630	0.05958	Fort Collins
Other	1378964	0.2742	
Colorado	5029196	100	

The Politics of Colorado

Curtis Martin (1962) notes that Colorado, due to its status as a frontier state, has always been fiercely democratic and independent. He connects this fact with Colorado's past, by pointing out that its political institutions were deeply rooted in mining culture, ordinary citizens' participation,a strong feeling of being "far away" from sources of centralized power on the coasts, and a wish for the protection and preservation of Colorado's natural environment. As such, Colorado can be described as a populist state with a strong libertarian streak, that highly values democratic processes when they serve the people or protect and fund national parks, but staunchly opposes state intervention when it is unwarranted[@martin_colorado_1962].

This 1964 study of Colorado politics rings true to this day. One needs not search for long to see instances when Colorado honored this description. One example is TABOR, or the Taxpayer's Bill of Rights; a strongly libertarian, small-government, populist series of regulations that mandated a referendum for any measure that would increase state taxes, and caped government spending. TABOR was passed by referendum in 1992, and later amended in 2005 after the dot com economic crisis exposed the fact that inability to spend is very bad for a state government trying to jump start its economy. [@legislative_council_staff_tabor_2009]

Similarly, Amendment 64 passed in 2012 made Colorado one of the first states to legalize the selling, possession, and consumption of recreational marijuana; a policy advocated by progressives and libertarians alike. Colorado was also the staging ground for what has been coined the "Sagebrush Rebellion": a movement primarily consisting of ranchers in dispute with the federal government over land use laws and wildlife protection. While this "rebellion" primarily consisted of battles in local legislatures or elections in the 1970s, its echoes can be heard till today in events like the Bundy Standoff, with ranchers taking up arms against federal employees and occupying federal land [@thompson_first_2016].

Setting policy aside, this description of Colorado is also confirmed by polling data and election results. While being traditionally more conservative, inflows of immigration from the South coupled with increasing urban liberalization and tourism has led the state from leaning republican to being aggressively purple: the quintessential swing state. Colorado voted both for and later against Bill Clinton, voted for G.W. Bush twice, and has supported democratic presidential candidates since [@hamm_how_2017]. Additionally, when polled on trust of federal or local governments, Colorado residents are systematically skeptical; in a random sample poll conducted by Cronin and Loevy (2012) in 2010, 56% stated that their state officials were lazy, wasteful, and inefficient. However–again indicating a libertarian, independent streak–most Coloradoans from 1988 to today consistently believe that their state is "on the right track" [1].

Voting in Colorado

Each County individually administers local, coordinated, primary, and general elections, under the supervision of the Colorado Secretary of State. This means that each county individually handles the voters registered in that county. Unsurprisingly, the same eight most populous counties are also the counties with the majority of registered voters, as their registrants comprise 73% of total Colorado registered voters (as of November 2017). As table 3.2 shows, these eight counties have a registration rate between 60-80%, compared to a Colorado-wide rate of about 67%. Registration rates for all counties are also graphically depicted in figure 3.2. In terms of Party registration, Colorado as a whole leans democratic by a very narrow margin (figure 3.3).

Table 2: Colorado voter registration for largest counties

County	Total Registered Voters	County Registration Rate	% of Statewide Registrants
Adams	270,303	0.61	0.07
Arapahoe	410,546	0.72	0.11
Boulder	237,091	0.80	0.06
Denver	450,616	0.75	0.12
Douglas	237,659	0.83	0.06

[1] Colorado College Citizens Polls, taken from Cronin et al. [@cronin_colorado_2012]

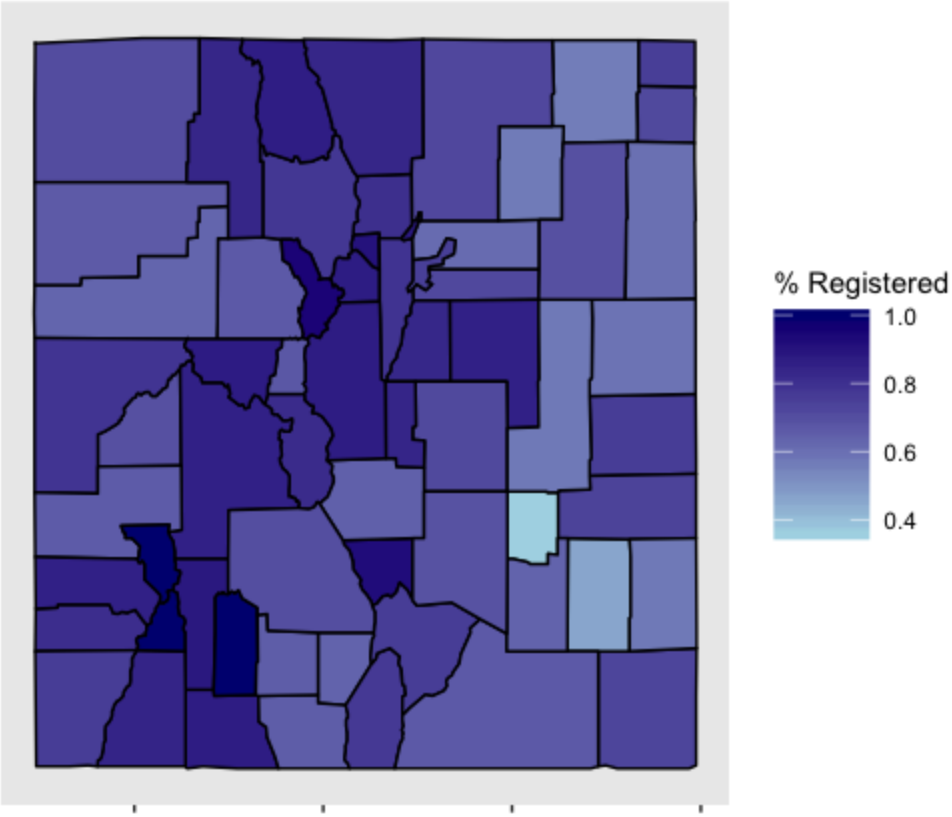

Figure 2: Registration rates per Colorado county

County	Total Registered Voters	County Registration Rate	% of Statewide Registrants
El Paso	445,708	0.71	0.12
Jefferson	422,362	0.79	0.11
Larimer	250,626	0.84	0.06
Other	1,009,392	—	0.27
Colorado	3,734,303	0.67	1.00

In the past 25 years, there have been a series of key changes in the way Colorado administers elections, in relation to Vote By Mail and other reforms targeted and expanding the democratic franchise. In 1992, Colorado introduced no-excuse absentee voting, allowing voters to either physically pick up a mail ballot at a Vote Center or County Office, or have a ballot mailed to them prior to election day. In 2008, this reform was expanded to a permanent Vote-By-Mail system, which gave counties the option to allow voters to be permanently put on a list of addresses that received mail ballots prior to the election. The State also entered a transitional status to full mail elections, giving counties the option to make all coordinated local elections, general elections, and primary elections exclusively VBM. In 2013, the Colorado State Legislature passed HB13-1303: The Voter Access and Modernized Elections Act, which mandated that every voter currently registered receive a mail ballot for all future elections. The Act also expanded the use of Vote Centers instead of traditional polling places, instituted same-day voter registration, and revamped the way active and inactive voter status was designated on voter rolls–more on this in future sections [@hullinghorst_voter_2013]. These changes are summarized in Table 3.3.

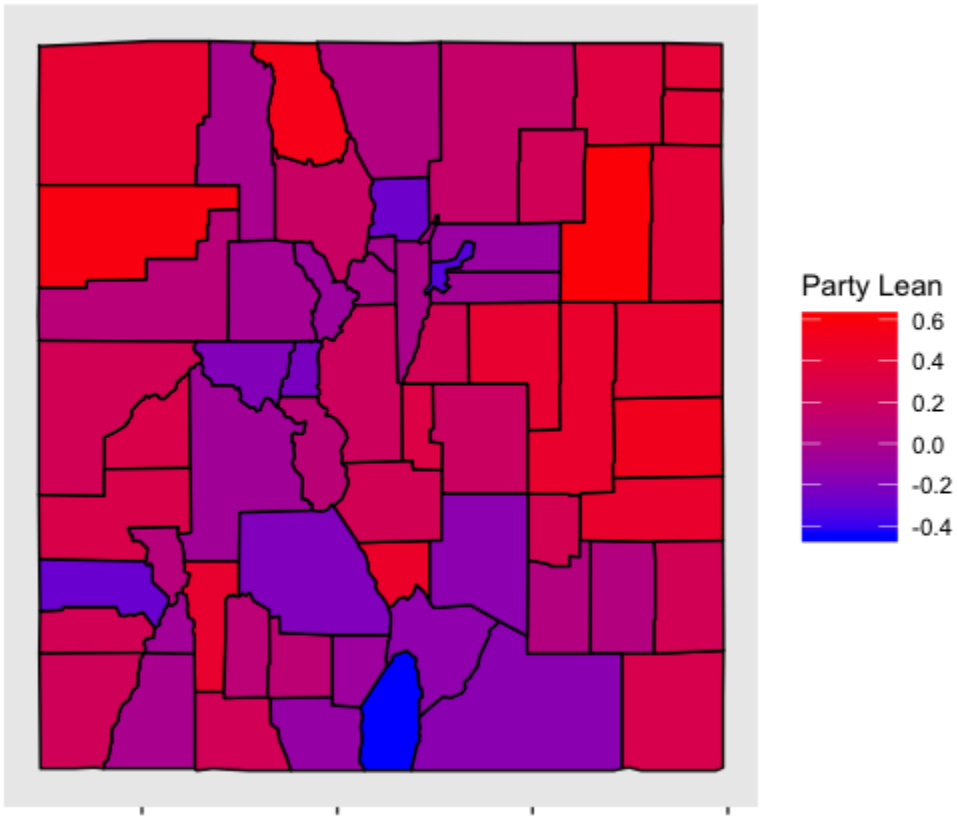

Figure 3: Democratic/Republican party lean per Colorado county

Table 3: Key changes to Colorado elections policy

Year	Key Changes
1992	No Excuse Absentee Statewide Implementation
2008	Permanent No-Excuse VBM Lists, Option of Full-VBM Elections
2013	Automatic Mail Ballot System Implemented Statewide, Established Vote Centers

Colorado as a Case for this Thesis

Colorado presents such an interesting case for research on Vote By Mail exactly because it has gone through such a long transitional process to reach its current elections system. It has steadily developed voting policy through a mixture of state mandates, county action, and outside policy motivations. Colorado's streak of independence and direct democracy is also very apparent in this shift in electoral practices, since they have been passing policies trying to expand participation for a very long time. It gives researchers access to approximately 22 years during which at least part of the state conducted elections by mail, making comparative, county- or individual-level case studies particularly alluring. Colorado's streak of independence and direct democracy is also very apparent in this shift in electoral practices, since they have been passing policies trying to expand participation for a very long time.

On a more general level, Colorado is interesting exactly because it is "typical" but with a wild streak. It is typical rocky mountain country, great planes country, and liberal urban city but all *in one state*. In is libertarian yet increasingly Democratic. It heavily relies on state funding for national parks, yet rebels against federal land use laws. It is a frontier state with traditional values, that overwhelmingly supports marijuana legalization. It is also a consistent purple state, with a Democratic Governor and House, but Republican Attorney General, Secretary of State and Senate. This means that Colorado is a combination of distinct national effects, but also local effects that make it significantly different from national trends as a whole. In this environment, predicting results of policy can be difficult, but extremely salient as multiple effects can be tested against each other.

The Data

This thesis relies on county and individual level models to draw conclusions on voting behaviors, and how they are affected by voting method. As such, the data I need will optimally contain the following:

- **County and individual level demographic characteristics**: race, gender, urban population
- **County and individual level voting data**: turnout, party registration, total registrants
- **Information on individual elections**: date, ballots cast, voting methods, county, election descriptions

In the process of my research, I have acquired sufficient data to cover the second and third of these areas. I was unable to procure individual level data on demographic characteristics apart from gender, age, and party registration. Reasonable conclusions can still be drawn from county or precinct aggregates.

Sources and first glance

I used two sources of data: Colorado voter records gathered by the Colorado Secretary of State's office, and demographic data from the 2010 US Census. In the process of procuring these data I was aided by a series of other researchers and professionals with experience in the field of elections administration. Andrew Menger, Postdoctoral Fellow at the Weidenbaum Center on the Economy, Government, and Public Policy at Washington University, was kind enough to give me access to data files for Colorado for the years of 2012-2016

that he had already collected for his research[2]. I directly obtained the 2017 from the Colorado Secretary of State's office, with the help of Mr. Judd Choate, Director of Elections.

2010 US Census

The US Census is conducted country-wide every ten years, with the goal of procuring accurate data on the demographic characteristics of the population. The Census uses a combination of federal field workers conducting door-to-door canvassing and statistical methods for data aggregation. From the 2010 Census, which is publicly available online, I get total population counts, characteristics on race, and rural/urban population counts for Colorado.

I use two datasets from the Census. For both, the unit of observation is one of the 64 counties of Colorado, and both include the same total population counts. One contains racial demographic characteristics and the other contain percentages of rural and urban populations in each county. The racial demographic dataset needed some wrangling work to extract a percentage of white residents in each county. Individuals were coded as "white" when they identified as exclusively white–this doesn't include mixed-race individuals reporting white ancestry.

Colorado Voter Files

As mandated by HAVA, Colorado maintains a statewide registry of all currently registered voters. This registry is typically under the purview of the Secretary of State. Voter Registration Files are constantly updated with new information on existing voters, new voters, or with the removal of inactive or otherwise ineligible voters. Therefore, this file will be different every time it is accessed or shared. Based on when this file is accessed, only a "snapshot" of the file can be obtained. Similarly with VRFs, a Voter History File is maintained and constantly updated by the state. This file is uniquely connected to its VRF: only voters showing up as registrants will have their histories included. I have both Voter Registration and History files for the years between 2012-2017, obtained with the help of Judd Choate and Andrew Menger.

In the Voter Registration files, the unit of observation is the individual voter, and all variables are initially coded as character strings. Each voter is assigned a unique voter ID, which serves as a point of reference between the two files. Broadly speaking, data in this file can be divided between three categories: first, personal identification information like address, ZIP code, or phone number; second, demographic information like age and gender; third, information pertinent to elections administration like congressional district, local elections for which the individual should receive a ballot, voter ID, and party registration. I will further elaborate on relevant variables in the wrangling section.

In the Voter History files, the unit of observation here is a single ballot cast, and all variables are initially coded as character strings. This means that for each voter registered–and so included in the VRF–the history file should contain an observation for each time they voted. This file includes two types of data: first, identifiers for the election like county, date, description, and type; second, identifiers for the individual vote including voter ID and voting method.

Why Voter Registration Files?

Voter Registration and History files are suitable sources for my analysis because they contain all the data that is necessary for a first pass at testing my hypotheses: voting method, county, election level, active registration, registration dates, and a series of individual characteristics like party registration, age, or gender. These data, and the demographic data in particular, are also in the most basic unit of observation: the ballot level[3]. This means that I do not need to establish any process to infer individual characteristics from population-wide statistics.

[2]Doctor Menger's website with links to his research can be found at www.andrewmenger.com

[3]A good heuristic for what "ballot" level means is a specific individual at the time when they cast a specific ballot. Between ballots all characteristics may change: age, gender, party registration etc., which is why the "ballot" level is distinct from the individual level.

In statistical science, sampling is the process by which individual units are selected from a population. The sample selected should be *representative* of the whole so that it can be used to infer characteristics of the general population. Despite a vast array of techniques to ensure that sampling is representative, there is always room for error. Voter Registration Files are an excellent source of data because they do not involve any sampling whatsoever; they include *all* currently registered voters and voter histories. This characteristic helps cut down on data-related errors and on the complexity of data extraction.

An additional characteristic of these data is that they are concentrated and relatively uniform. Data transfer errors still exist, and the wrangling process is never entirely straightforward. Still, the data are almost completely uniform in how variables are encoded. Over thirty five million observations are included in my final, cumulative voter history dataset, and all of them have, for example, the same types of entry for party registration (REP, DEM, UAF etc.). Admittedly, this may just be a characteristic of the Colorado files, since they are the only ones I used for my research.

A last benefit of using Voter Registration Files comes from the replicability they allow for. These files are generally public, with access to them including only data transfer and administrative costs. This makes peer-review and replication less complicated than if, say, I was using private survey data. It also allows for expansion that goes beyond the State of Colorado; my code can be adapted to fit different data, making future comparative studies more likely and less time-consuming than starting from scratch.

	VOTER_ID	COUNTY	LAST_NAME	RESIDENTIAL_CITY	VOTER_STATUS	PARTY	GENDER
1	318090	Denver	Jefferies	DENVER	Active	DEM	Male
2	1224557	La Plata	Simmons	BAYFIELD	Inactive	REP	Female
3	833164	Arapahoe	Jenkins	CENTENNIAL	Active	UAF	Female
4	654070	Boulder	Giroux	LAFAYETTE	Active	UAF	Female
5	506568	Alamosa	Provorov	ALAMOSA	Inactive	UAF	Male
6	1467732	Mesa	Hoskins	GRAND JUNCTION	Active	DEM	Female
7	1061624	Boulder	Nola	BOULDER	Active	DEM	Male
8	1896894	Denver	Arrieta	DENVER	Active	DEM	Male
9	649632	Arapahoe	Butler	LITTLETON	Inactive	DEM	Female
10	1144545	Montrose	Clement	OLATHE	Active	DEM	Male

Table 4: Sample of a Voter Registration File

	VOTER_ID	ELECTION_TYPE	ELECTION_DATE	VOTING_METHOD	PARTY	COUNTY_NAME
1	318090	General	11/03/1998	Absentee Mail	REP	Gunnison
2	1224557	Coordinated	11/05/2013	Absentee Mail	DEM	El Paso
3	833164	General	11/07/2000	Absentee Mail	UAF	Arapahoe
4	654070	General	11/08/2016	Mail Ballot	CDP	Arapahoe
5	506568	Coordinated	11/02/1999	Absentee Mail	REP	Arapahoe
6	1467732	General	11/08/2016	Mail Ballot	REP	Pueblo
7	1061624	Coordinated	11/07/1995	Absentee Carry	REP	Larimer
8	1896894	General	11/03/1998	Polling Place	DEM	Larimer
9	649632	General	11/03/1992	Polling Place	UAF	Arapahoe
10	1144545	Coordinated	11/03/2015	Mail Ballot	DEM	Jefferson

Table 5: Sample of a Voter History File

Wrangling the Data

The process of "wrangling" refers to manipulating the data into a form that can then be used for graphing, exploratory data analysis, modelling, or presentation. In this case, wrangling also included aggregating data

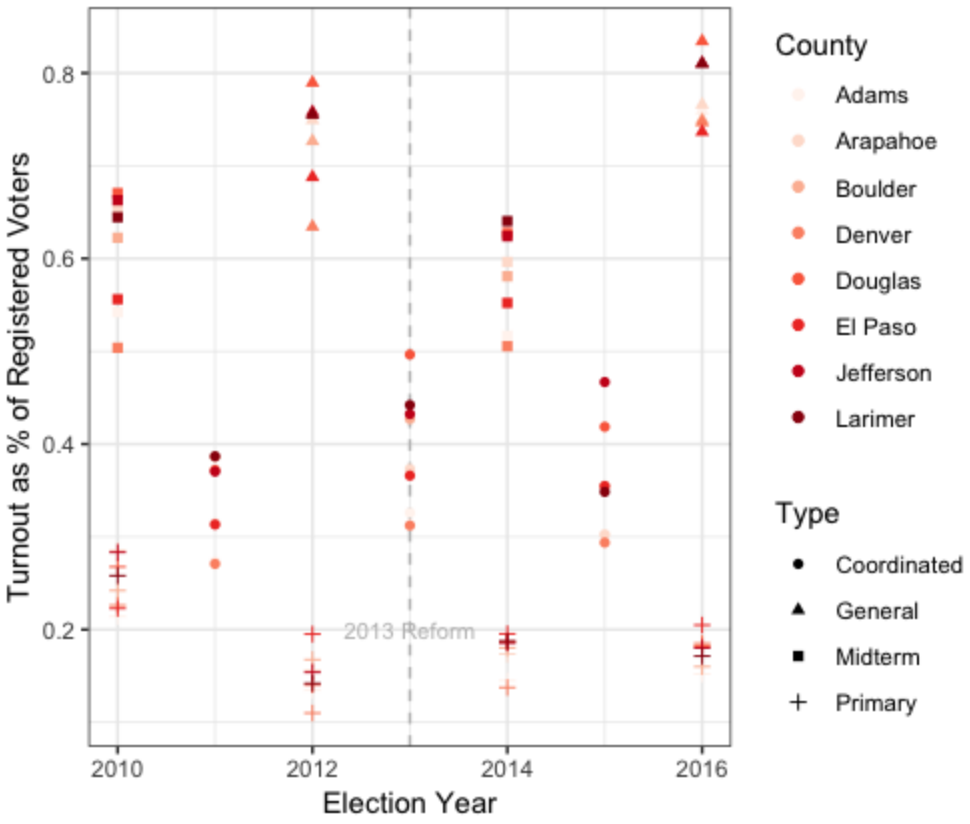

Figure 4: Turnout plot for eight largest Colorado counties, 2012-2016

across multiple sources and datasets. For this purpose, I made heavy use of the `tidyverse` *R* package, and in particular the `dplyr` package. In this section I will go through some of the key problems encountered during the wrangling of these data, and then discuss the final form each variable takes.

Initial Problems with the 2017 Voter File and Solution

In my initial research I intended to only use the 2017 snapshots of the Colorado Registration and History file. The major issue I encountered, which merits discussion in its own section, comes from the fact that the records I have access to are "snapshots". What this means, is that for each person in each year of voter registration files, I have their corresponding history files for all ballots they have cast in Colorado, but not their own history of registration and migration. If, say, a voter moved from Boulder County to Summit County, I would have their votes in Boulder County show up in the voter history file, but them being registered in Summit. If you recall the turnout calculations specified earlier on, this implies an overestimation when looking back at elections that happened some time before the date of the "snapshot". Additionally, "snapshots" of current voter files do not reflect voters dropping off the rolls for whatever reason (death, moving out of the state, long term inactivity, non-confirmable personal data etc.)

After going through turnout calculations with the 2017 files, a significant majority of counties appeared to have turnout exceeding 80%, particularly for years between 2000 and 2012. This was, to put it mildly, concerning. With the aforementioned help, I was given access to similar "snapshots" for each year between 2012-2016. After similar calculations, I returned figure 3.4 for the eight most populous counties as described above, including different shapes for election type, colors for county, and a vertical line at 2013 to signify the latest major change in how Colorado administers elections.

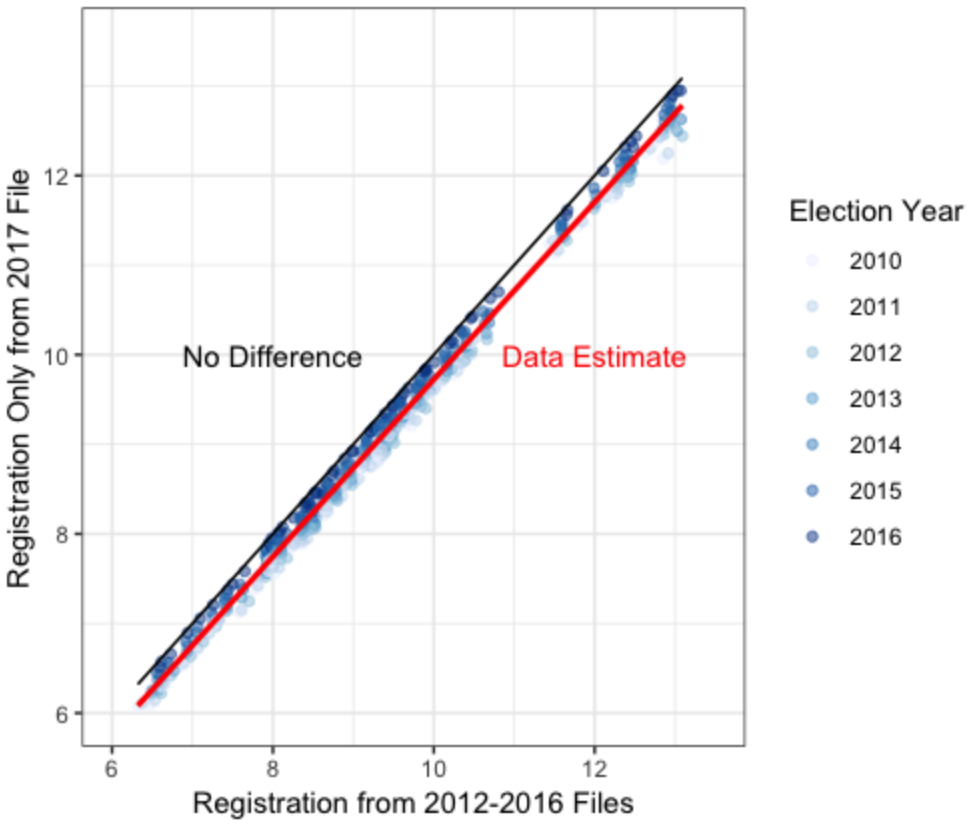

Figure 5: Comparison of registration count methods

To also further illustrate the in-county migration and dropped voter problem, I created a graph that includes logged total counts of registered voters calculated using the 2017 and the 2012-2016 files. The plot also includes a line at $y = x$. If in-Colorado migration and dropped voters are not an issue, most points on this graph should be at this line.

Two things should be clear from figure 3.5. First, there is significant deviation between the counts using just the 2017 file and all files across years. Specifically, the 2017 count consistently underestimates the total amount of registered voters–this is shown by the red linear model smoothing line. This consistent difference means that it is close to impossible to generate safe conclusions on my hypotheses using only the 2017 files and the methods I have outlined in Chapter 2. Second, counts get more accurate the closer to 2017 we get. This should be even more apparent in figure 3.6, which limits the scale to only some high registration counties, and adds a shape indicator for county.

Here the structure of the data becomes clear: for each county, there are a series of almost vertically distributed points, which get closer to the $y = x$ line the closer the counts get to 2017. Through this series of tests, it became clear that using multiple years of data was necessary in order to conduct an accurate test of my hypotheses. My selection was later vindicated, when looking at comparisons between reported rates of turnout[4] and turnout calculated through my dataset for the 2014 midterm election (see fig. 3.7).

The differences are insignificant. They exist because of "noise" added on because of errors in the data, misreporting, registration records redacted due to privacy concerns, voters dropped before the "snapshot" occurred, and other similar factors.

[4]Turnout is calculated over all registered voters

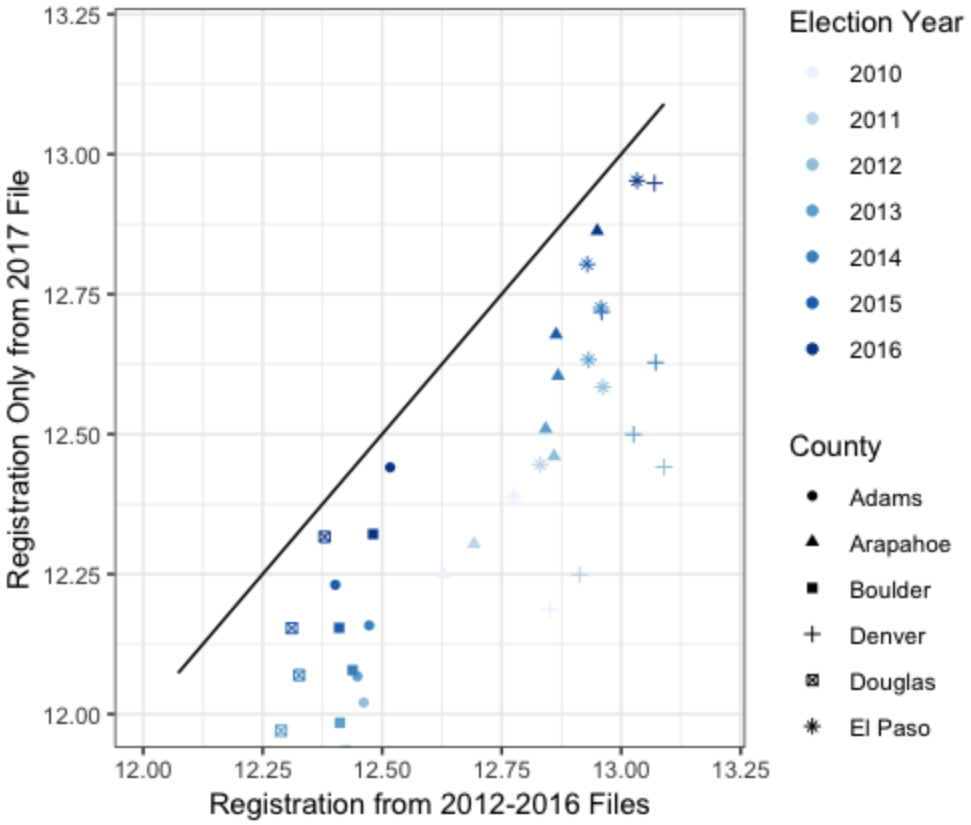

Figure 6: Comparison of registration count methods only for a few counties, 2012-2016

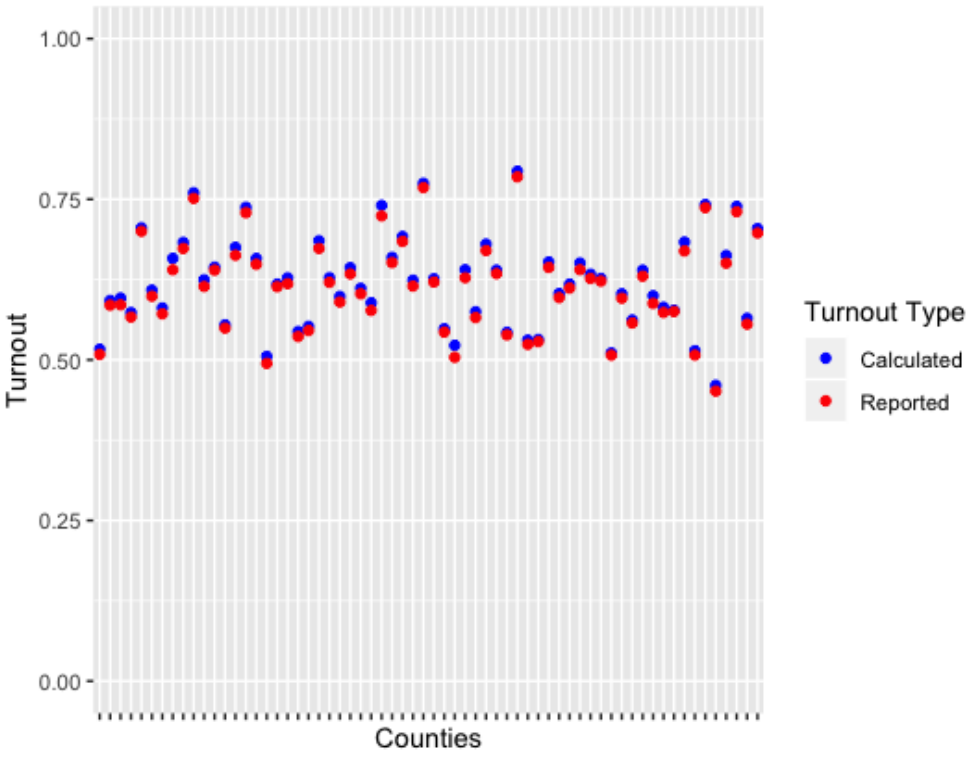

Figure 7: Comparison of reported and calculated turnout for 2014 midterms across county

Other Wrangling Issues

Wrangling the data was the majority of the work that went into this thesis. As will become clear in this section, apart from accurately processing, diagnosing, and merging the data, the process of wrangling includes several non-trivial decisions about how to treat missing values and variable specification. Including a full account would probably read like the world's most cliche crime novel: a series of elusive final datasets, a plucky yet occasionally naive young detective, two wisened mentors, clues, dead ends, frustration, compromise, and. . . spreadsheets. I will spare the reader the whole story, but I will include a non-comprehensive list of some of the difficulties associated with wrangling voter files, as it was a crucial part of the learning process I underwent while doing my research.

Missing Values: The decision on how to deal with missing values–or NAs–in a dataset is a lot more important than it may initially seem. A first, intuitive reaction might be to just disregard them; however this works under the assumption that there is no structure inherent to why these data are missing. To give just two examples, in the data I have collected the PARTY value for the 2015 voter registration file is missing. If I excluded all observations with missing PARTY values, I would be excluding a fifth of my data. Missing values were also present in the VOTING_METHOD variable of the voter history files. While this may have seemed troubling, after closer examination it was revealed that the vast majority of such missing values was concentrated in Jefferson County, and in elections prior to 2002. Therefore, these observations could be ignored, since they played no role in my final dataset. The conclusion should be that choices made on exclusion, inclusion, or estimation of missing data are very important, and should be taken with much care and consideration for the underlying structure of the data.

Data Input Errors: Is "QATAR" a political party in Colorado? State records say not. However, "QATAR" did show up as a value for the Party variable for my 2016 voter registration file snapshot. This may occur for a number of reasons, the most likely of which is the introduction of errors when transferring these data. The data have been read and written by multiple operating systems (iOS and Windows) and programming platforms

(STATA, R); they have also been uploaded, downloaded, and written unto CDs, as well as transferred between County and Colorado Secretary of State's Office when they were created. Characters that would be normally read into one platform as line or value delimiters may have been misinterpreted by another platform, with no operator error involved. In my analysis I treated all values that seemed more likely than not to be errors as NAs. There were not many of these–less than .001% of my data–but they were a hassle to find, analyze, and then recode into some useful value.

Data Size: Nothing to write home about here, just an observation that multiple voter registration files can be *huge*, which puts considerable strain on a computer's processing power. This means that wrangling has to comprise of a series of careful, deliberate moves. Brute force should be discouraged, as a dead end means several hours of melodic computer fan panic.

Joining, Merging, Spreading, and the Multiplicity of Levels: For the data to end up in any functional shape, it eventually becomes necessary to start joining datasets. Thankfully, a clear division of modelling tasks between county and individual level models means that joining on COUNTY or VOTER_ID is ideal, and fairly straightforward. As will become clear in later sections, I also had to consider the variety of different units of observation, specifically: county, individual, ballot, election, county-by-election.

Final Variable Specifications

After the conclusion of the wrangling process, the resulting datasets included a series of discrete and continuous variables. I will briefly outline them here, along with their range and values.

- VOTER_ID: Discrete variable, unique value given to each individual voter. Useful for merging.
- COUNTY: Discrete variable, the 64 counties of Colorado.
- REGISTRATION_DATE: Discrete variable, date of registration for each registrant. Useful to get total registrants on election day.
- TURNOUT: Continuous variable, in the range [0,1]. The response variable for my county-level models.
- ELECTION_TYPE: Discrete variable, the four types of elections: Primary, Coordinated, Midterm, Presidential.
- ELECTION_DATE: Discrete variable, self-explanatory.
- VBM_PCT: Continuous variable, in the range [0,1]. This is the focus of my analysis, as it counts the percentage of total ballots that were mail ballots.
- PCT_WHITE: Continuous variable, in the range [0,1]. Percentage of white residents per county.
- PCT_URBAN: Continuous variable, in the range [0,1]. Percentage of urban residents per county.
- PARTY: Discrete variable. For each voter, the party they are registered with. Can be: Republican, Democrat, Other, or Unaffiliated.
- GENDER: Discrete binary variable, Male or Female.
- AGE: The age of the individual registrant.
- VOTING_METHOD: The method used by an individual voter to cast their ballot. Is coded as either VBM or In Person, according to Table 3.4.

Table 6: Voting methods coding table

Voting Method	Description of Method	Designation
Absentee Carry	Voters who carried an absentee ballot with them from an early voting location or county office	VBM
Absentee Mail	Voters who were sent an absentee ballot, and mailed it in	VBM
Early Voting	Voters who physically went to an Early Voting location and voted	In Person
In Person	Voters who physically went to a polling place and voted on paper	In Person

Voting Method	Description of Method	Designation
Mail Ballot	Vote By Mail	VBM
Polling Place	Traditional polling place voting, discontinued in 2013	In Person
Vote Center	Voters who cast their ballots at Vote Centers	In Person