Analysis of the Dr. Douglas G. Frank voter fraud theory

Phantom ballots and the case of the polynomial credit line

John Henderson (jw.hendy [at] gmail)

Compiled with LaTeX via `Org-mode`

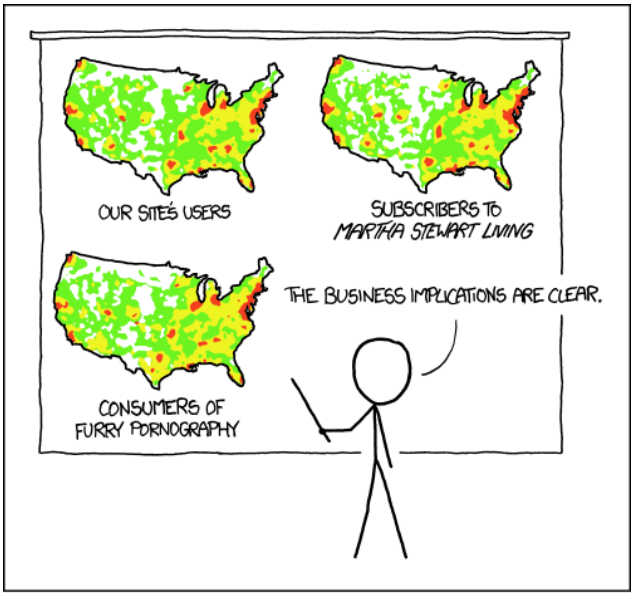
Contents

1 Introduction

This analysis presents a refutation of the voter fraud theory put forth by Dr. Douglas G. Frank, PhD, as contained in a legal brief filed by attorney Matthew DePerno in the case against Atrim County, Michigan.[1] The theory may be distilled as follows:

- bloated voter rolls (number of registered voters) in various Michigan counties feature counts at or exceeding the known 18+ population for those counties
- a 6th order polynomial (referred to as a "key") can be fit such that the registered voters by age, the mean county turnout, and this key may used to accurately predict the reported votes per age in that county
- this finding serves as evidence of a nationwide voter fraud mechanism: "phantom ballots" are injected into to the true vote count in proportion to the key, which is uniquely programmed by each state according to their demographics
- these phantom ballots are injected predominantly on behalf of young voters; the gap between registration levels (high) and turnout rates (low) for this demographic creates a "credit line" of fake votes that can be drawn from during an election

tl;dr: this theory amounts to Dr. Frank expressing surprise that the age demographics of registered voters and votes cast *both* correlate to the age demographics of the overall population. In summary, this theory is an age-based analog of this trivial phenomenon:[2]



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

[1] `collective_response_to_motions_for_protective_order_040921.pdf`, available in the DePerno Law collection of Bailey v. Antrim County documents: https://www.depernolaw.com/bailey-documents.html

[2] "Heatmap," xkcd: https://xkcd.com/1138/

2 Claims in detail

The legal brief formally summarizes the claim as follows (pgs 3-4):

> To be clear, at least four (4) of the so-called battleground states have implemented an algorithm used to regulate and shift votes in the 2020 elections. These algorithms are unique to each particular state. [... rant about the difficulty working with the Michigan Qualified Voter File structure ...] Nevertheless, after countless hours of work going through the Michigan database, Plaintiff's expert, Douglas G. Frank, PhD, has uncovered the algorithm (a sixth degree polynomial).

The following data sources are cited for Dr. Frank's analysis (pg 5):

> 1. Blue Curve. Population data extracted from the 2019 U.S census at census.gov. This is the blue curve on each chart for the 9 counties examined, which shows the census data per age group.
> 2. Black Line. The state registration database for October 2020 used in the November 3, 2020 election. This is the black line on each chart.
> 3. Red Line. The state voter database from January, 2021. This is the red line on each chart.

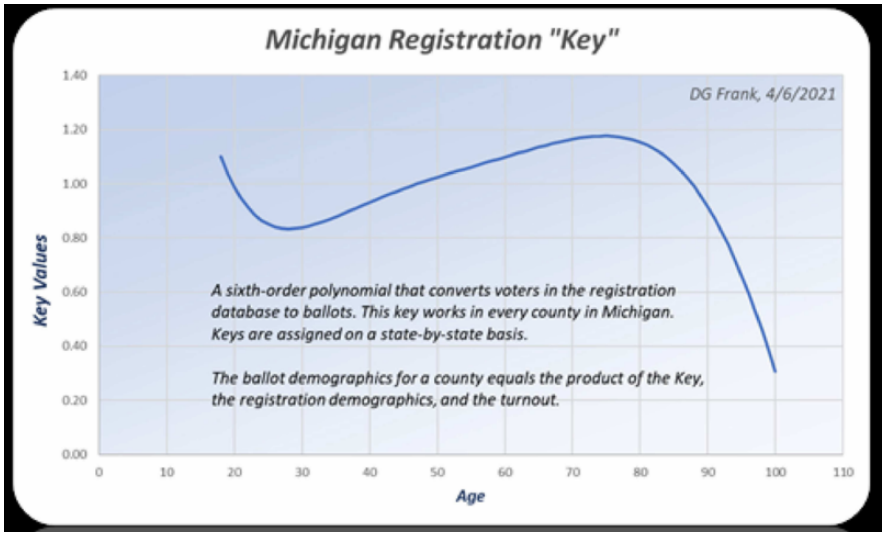
We are given the following list of core conclusions drawn from Dr. Frank's findings (pg 5):

> - Voter registration is consistently near, or exceeding county population demographics.
> - There are over 66,000 ballots recorded that are not associated with a registered voter.†
> - The ability to predict ballot demographics with such remarkable precision (average correlation coefficient of R = 0.997 demonstrates the activity of a regulating algorithm.
> - This confirms, as seen in several other states, that ballots are being harvested at the precinct level, regulated at the county level, and determined at the state level.
> - the degree of precision observed confirms that algorithms had access to voting databases and voting activity before, during, and following the November 3, 2020 election.

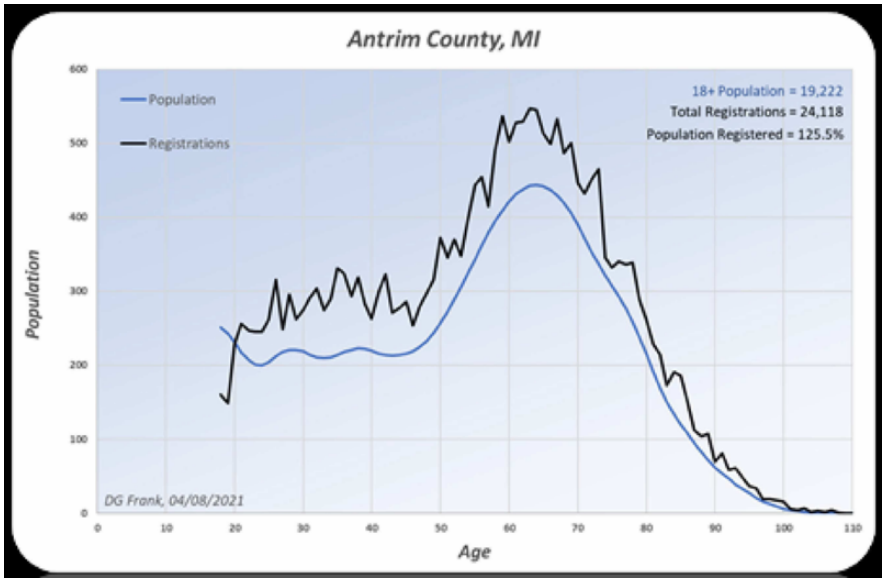
† *Note*: this is assumed to mean that the database(s) containing recorded votes and registered voters contains an ID that was used to merge the data, finding 66,000 non-matching entries. There is no way to confirm this without the raw data, this claim is not detailed in the remaining brief, and some MI officials have stated they are not aware of, nor can find this data.[3]

[3] "I'm not sure where he got the numbers from," Uzarski [Elections Director, Kent County, MI] said. "I've looked for them myself. I have not been able to find the source." https://www.politifact.com/factchecks/2021/apr/16/douglas-frank/no-evidence-michigan-used-algorithm-manipulate-ele/

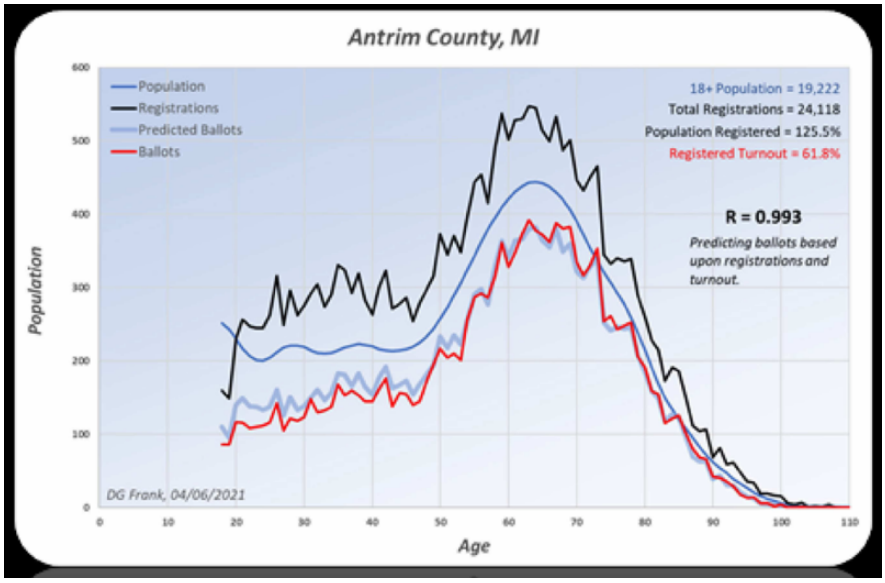
Moving on to the mechanism of this theory, we are informed that each state has a sixth degree polynomial "key" which "unlocks the door and uncovers the ability to manipulate data and results," and Michigan's key is presented as an example (pg 7):



Next, Antrim County is highlighted, and we are shown the population age demographics (blue) and registered voters from the database (black), illustrating that registered voters exceed population for almost every age (pg 9):



The key and Antrim turnout rate were subsequently applied to the registered voter curve to obtain the predicted votes (light blue), with the official reported votes shown in red (pgs 9-10):



There are additional sub-claims introduced in the document, however it will be easier and more interesting to address these in the context of the analysis. At this point, all the key points have been made, and the justification for belief in fraud can be described as, "we should not be able to accurately predict votes by age from registered voters by age." This belief will be revealed as confused in the sections that follow.

3 Tools used

Transparency and reproducibility are lauded among the scientific community, however in my experience this is not the case among fraud theorists.[4] Data is not shared, methods are only vaguely described, and I have yet to see a link to reproducible code. As a result, reproduction entails an exercise in data hunting and reverse engineering, which is what must be done here.

This analysis was conducted using `R`, a statistical programming language, with code embedded in-line in this document using Org-mode.[5, 6] Since no data was shared, screenshots of the plots in the DePerno brief were taken, and WebPlotDigitizer was used to extract raw data approximations from each curve.[7] Methods will be discussed transparently here, with full code and data available on github for reproduction (corrections welcome).[8]

[4] This will my third public writeup of sorts, with several others via email. See https://jwhendy.github.io/blog/ for examples ("Hammer, Scorecard, and NY Times json files" and "Straight ticket vs. direct votes").

[5] The R Project for Statistical Computing. https://www.r-project.org/

[6] Org-mode is an Emacs package for notes, todos, and literate programming: https://orgmode.org/

[7] A tool by Ankit Rohatgi, enabling data extraction from a plot image. https://automeris.io/WebPlotDigitizer

[8] Code, data, and the org-mode file generating this paper: https://github.com/jwhendy/dr-frank-voter-fraud

4 Analysis

4.1 Examining the population data

After listing the sources, the brief contains the following quote (pg 5):

> The blue, black, and red lines on the graphs are data. It is not speculative or calculated. It is completely 100% data.

Due to skills ingrained early on via Sesame Street, it was apparent that among Antrim County curves, "one of these things is not like the other." We are *told* that the blue curve is directly from the from the census population data, but the other curves are jagged and this one is smooth. Why?

After significant effort, I have seen no evidence that the Census Bureau reports population for individual ages, instead using total population for age *ranges*, which is what I believe was used by Dr. Frank.[9] While one could argue that this is still "100% data," it is misleading to portray it as raw and unprocessed.

In addition, the 2019 Census Bureau data are demographic *estimates*, based on the previous decennial data (2010 in this case).[10] It's the best one can obtain, but requires applying models for births, deaths, and other factors to the 2010 data for all ages... 9yrs out. This is also county level data, where the consideration of factors like moving, local jobs, college enrollment, and any number of other variables make granular estimates more challenging vs. the national or state level.

The data believed to be used by Dr. Frank was obtained in an effort to reproduce the county populations shown.[11] The data contains population by groups, e.g. `AGE5559_TOT` corresponding to total population for 55-59yr olds. The data conveniently contains both `AGE2024_TOT` *and* `AGE1824_TOT`, from which we can compute the 18-19yr old population. Age ranges are reported in 5yr increments except the computed 18-19yr old range and `AGE85PLUS_TOT`, representing 85+ yrs of age.

To reproduce Dr. Frank's plots, we must convert population totals for an *age range* into an estimates for an *individual age*. To do this, I used the following method per range:

- `x = (min_age + max_age)/2` (mean age for the range)
- `y = population / (max_age - min_age)` (total divided by number of ages represented)
- for the 85+ age group, `x=92.5` and `y=population/15` were used to distribute the population over the range of 85-100yrs.

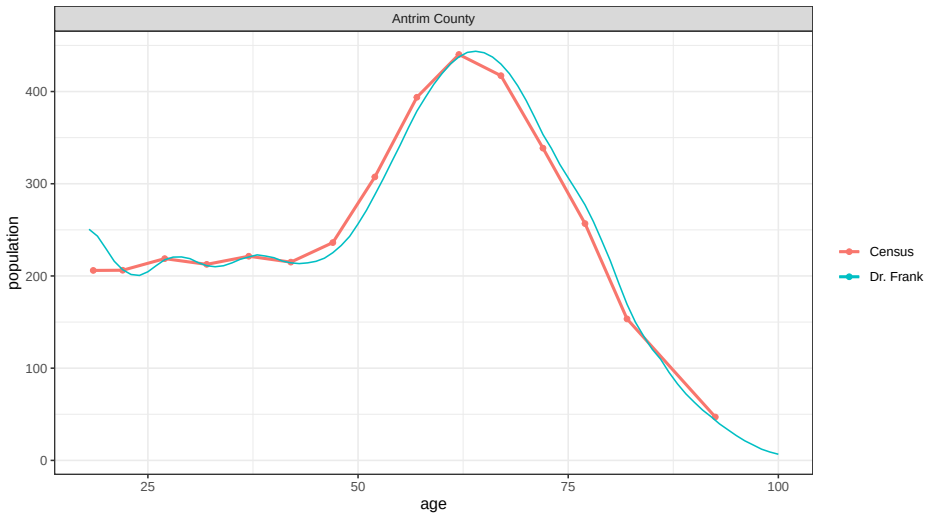
[9] This conclusion was drawn after consulting the 2010 Summary File Dataset, which appears to be the most granular data offered: https://www.census.gov/data/datasets/2010/dec/summary-file-1.html. In consulting the technical documentation, P12, "Sex by Age" data at the block level is the most likely candidate to offer this data, yet only features population in the the typical 5yr ranges.

[10] Population Estimates and Projections (2010-2019). https://www.census.gov/data/developers/data-sets/popest-popproj.html

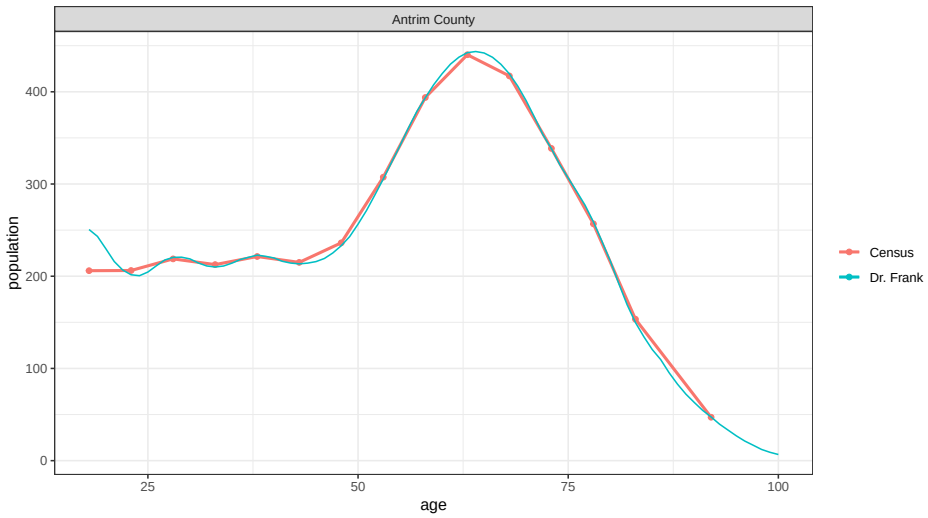
[11] "County Population by Characteristics: 2010-2019", Census Bureau. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html

Dr. Frank's plots were converted using WebPlotDigitizer, saving one file for each curve, per county. WebPlotDigitizer finds a curve matching a target color and overlays points at fixed intervals on top. As a result, the intervals do not align to ages, so values were interpolated to force alignment to fixed integer ages from 18-100.

With both the Census Data and the extracted Antrim County population curve from Dr. Frank's plot, we can compare the two results:



This works out delightfully well, and with some guess and check it appears that Dr. Frank used `[23, 28, ..., 83, 91]` for the `x` values of each age group. Other than disagreement about the low end (perhaps he did not compute the 18-19yr olds specifically as I did) and the smoothing method he used, we have near perfect alignment on all key points.



Why does this matter? For one, it serves as a sanity check on reverse engineering what's been shown. It's also highlighting that the blue curve is *not* presenting the raw data as-is, but scaling and smoothing aggregate data while presenting it as the de facto population for individual ages in each county. This will become more important later.

4.2 Registrations > population

A core mechanism of the theory is that excess registrations, particularly among young voters, enable a a "credit line" of "shadow ballots" which can be drawn on during an election to obtain the desired result. The brief puts it like so (pg 12):

> Some people might ask why the key is different in every state. The answer is different because each state has its own demographics. Outcomes are predicted based on demographics. Anyone with access to the QVF [Qualified Voter File] can change just one number in an algorithm (at the state level presumably) and modify the sixth degree polynomial to adjust the election result. For instance, in Michigan, Defendant Benson is overemphasizing the younger people. We can see that in the disparity between the black and red line on the left side of the graph. And that becomes progressively lower as the chart moves right. The younger people are the least reliable; the algorithm tilts to the younger ages because the less reliable voters will give the most shadow ballots. Think of the gap as a "credit line" that can be drawn on at any time using the algorithm.

From the plots, excess registrations are indeed apparent across Antrim and other counties. You can confirm totals yourself via the Secretary of State website and the Census.[12, 13] To the initial shock value of this finding, let's immediately point out that these two statements are worlds apart:

- there were more registered voters than voting aged population
- there were more *votes* than the voting aged population (or registered voters)

One of these requires process improvement, the other constitutes fraud. Also, let's be clear that this is not a new problem. Using this to fuel suspicion in 2020 also implicates 2016.[14] And 2006.[15] And Texas.[16] In my experience, the anticipated response to this is embracing that all prior elections *have* been controlled; this is the only way to maintain logical consistency while scrutinizing the 2020 election (though I have seen no calls to audit 2016, nor anyone running these analyses on 2016 data). In any case, it's good to recall what's being *shown* vs. merely alluded to.[17]

[12] Antrim County registered voters listed as 21,945 (2021-04-17). https://mvic.sos.state.mi.us/VoterCount.

[13] The Census reports total minus under <18yr old population as: `23,580*(1-0.177)=19,406`. https://www.census.gov/quickfacts/antrimcountymichigan

[14] https://www.politicscentral.org/report-michigan-has-24-counties-with-more-voters-than-people/

[15] https://www.govtech.com/archive/Michigan-Makes-Strides-In-Cleaning-Up.html

[16] "...counties across Texas appear to have more registered voters... than qualified citizens of voting age." https://www.chron.com/news/houston-texas/article/Conservative-watchdog-group-questions-counties-3467513.php

[17] This doesn't stop false interpretations, either, which can be seen in replies to these plots on DePerno's twitter page. "Basically there are too many ballots. Not enough people. Does that make sense?" https://twitter.com/imtwin64/status/1381030810132877313. "...When you have more votes then you have ballots you have fraud."

4.3 The function of the key

Let us accept this "young voter credit line" at face value, reminding ourselves of how this algorithm supposedly works. The legal brief tells us, "The ballot demographics for a county equals the product of the Key, the registration demographics, and the turnout." (pg 7) We are also given this additional explanation (pg 11):

> If we want to check our theory, then we simply graph the ratio between the black and the red, which creates the polynomial. The polynomial becomes the key.

So, given some registered voters for an age, `reg_n`, the county `turnout`, and the value of the polynomial for that age, `key`, the predicted votes are: `reg_n * turnout * key`.

For this injection theory to make sense, the following additional points must be true:

- the reported votes (red) consist of some number of real votes *and* some number of fake votes injected on behalf of [mostly] young people who didn't really vote
- thus, the real votes on election night were *lower* than the red line; the true count would be what is shown in red *minus* the number of injected shadow ballots
- put one more way, reality is something lower than the red line and the algorithm shifted the result *away* from reality and *toward* the black line of inflated registrations

However, go look at the key. It quickly *decreases* to <1 for the entire range of 20-50yrs old. Let's walk through what that means:

- I have some number of registered voters for e.g. age 25
- I multiply these registered voters by 0.618, the turnout for Antrim County
- I *further* multiply that by ~0.8, the value of the key for age 25, to obtain my prediction
- if you want to predict votes for 75yr olds, you use the same county turnout value, but your final multiplier will be a key value of *~1.2*.

Recall that the key is literally synonymous with the statewide manipulation algorithm used to control the election. This key, however, does *precisely the opposite* of what is claimed:

- after multiplying registered voters by an average turnout rate of 61.8%...
- the key must *decrease* the number of younger voters in order to predict reported votes, shifting them *away* from the bloated registration numbers
- the key must *increase* the number of older voters in order to predict reported votes, shifting them *toward* the bloated registration numbers

https://twitter.com/joe62160339/status/1382438444966764546

4.4 A note on turnout

On the subject of turnout, there's something curious about Dr. Frank's values. Antrim County is listed as 61.8%. Where is this from? Antrim County total votes are listed on the Michigan State website as 16,044 with a registered voter count of 22,082 for a turnout value of 72.66%.[18] Here are all values used by Dr. Frank vs. the reported turnout among the nine counties analyzed:[19]

	Dr. Frank			**Official**	**Computed**	
County	registered	18+ pop	turnout, %	turnout	% of reg	% of pop
Antrim	24,118	19,222	61.8	16,044	66.5	83.5
Barry	48,628	48,094	71.8	36,146	74.3	75.2
Charlevoix	23,279	21,337	72.8	17,103	73.5	80.2
Grand Traverse	79,537	74,536	72.8	60,668	76.3	81.4
Kent	489,234	500,078	71.3	363,695	74.3	72.7
Livingston	157,667	152,390	78.4	127,839	81.1	83.9
Macomb	670,592	694,196	71.2	497,098	74.1	71.6
Oakland	1,011,669	999,630	74.2	775,379	76.6	77.6
Wayne	1,365,392	1,339,405	61.6	878,102	64.3	65.6

I believe the turnout value used by Dr. Frank to be the calculated value required to scale *registered voters* to predicted votes using the key. Because the key is fixed per state while the registered voters vs. population surplus varies widely across the counties, a bespoke scaling factor is required to account for this.

In practice, voter turnout is reported as a % of either Voting Aged Population (VAP, all individuals 18+ yrs old) or Voting Eligible Population (VEP, 18+ minus those ineligible to vote). For example, the reported figure of 66.7% turnout for the 2020 election is *not* the percent of registered voters, but the percent of estimated VEP.[20]

Now, Dr. Frank certainly *could* have used the population as the input for the prediction, which would have enabled using the more standard values for calculated turnout. Why didn't he? Well, predicting votes from population entails multiplying his key (smooth) by turnout (a fixed number) by his population curve (smooth), which would yield a *smooth result*. This lacks the shock value of strikingly similar shapes between the input and the result. The matching shape of the registered voters, predicted votes, and votes is what enables this mathematical optical illusion to work.

[18] "Official Election Results November 3, 2020 2nd amended." http://www.antrimcounty.org/elections.asp

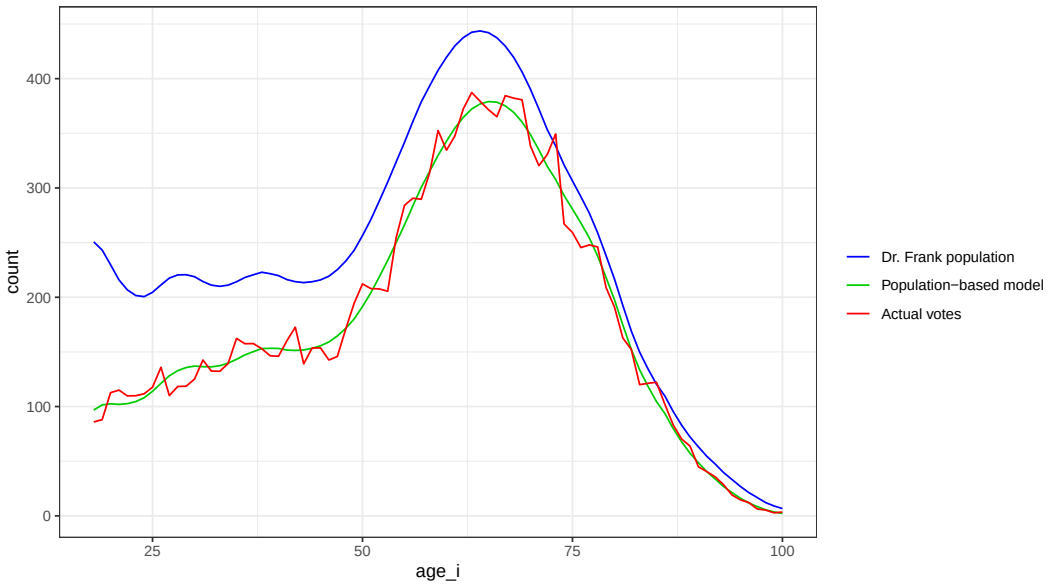
[19] https://mielections.us/election/results/2020GEN_CENR_TURNOUT.html

[20] See footnote for reported turnout on Wikipedia for an explanation of the calculation: https://en.wikipedia.org/wiki/2020_United_States_presidential_election#cite_note-4

Here I use the data extracted from Dr. Frank's Antrim County plot to create a 6th degree polynomial key for the vote predicting model: `population * VAP_turnout * key`.

```
df_key <- df %>%
  filter(CTYNAME=="Antrim County") %>%
  pivot_wider(id_cols=c(CTYNAME, age_i),
              names_from=var, values_from=val_i) %>%
  select(CTYNAME, age_i, reg, pop, pred, vote) %>%
  mutate(rat = vote/pop/0.835)

fit <- lm(rat ~ poly(age_i, 6), data=df_key)
df_key <- df_key %>%
  mutate(rat2 = predict(fit, df_key)) %>%
  mutate(pred2 = pop * 0.835 * rat2)
```

What does this prediction look like?



Compared to the original Antrim County R-value of 0.993, how does this smooth prediction fare?

```
[1] 0.993657
```

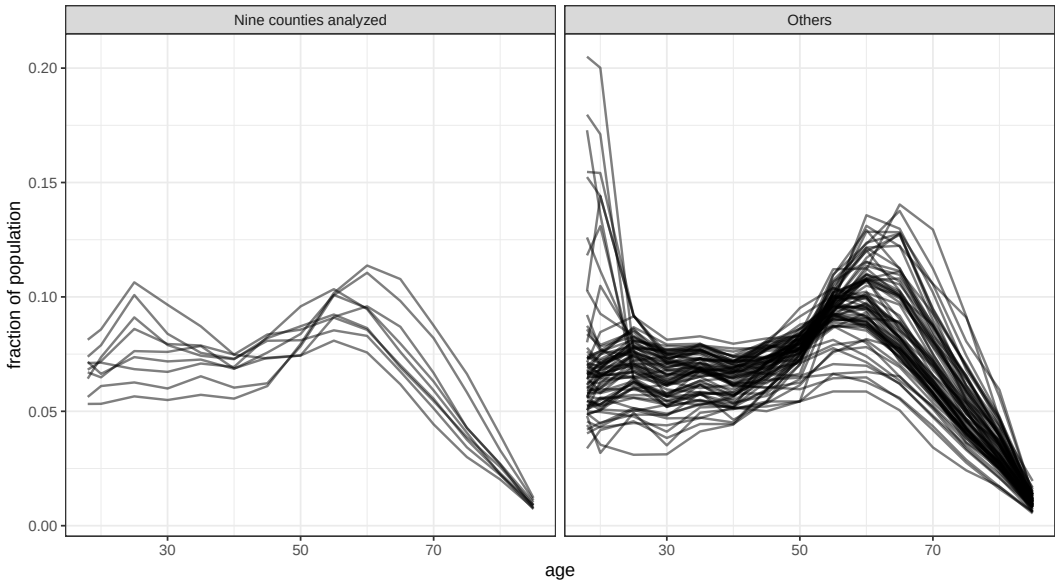
Contemplate this in light of another assertion by Dr. Frank in his full presentation of this fraud theory (which this brief is citing):[21]

> "Correlation Coefficient, R" A statistical value that indicates how well a set of values predicts a target set of data [...] *Correlations involving human behavior rarely have R values greater than 0.8*

While we're discussing the function of the key, I also want to touch on this claim (pg 7):

> We call the polynomial a "key" because it works in every county in Michigan...

But *does* it work in every county in Michigan? We don't have data for votes and registrations except for the nine counties analyzed, but we *do* have Census age data covering all MI counties. This allows us to look at the distribution of age as a percent of the total population per county.



The nine counties look rather "average" compared to some of the extremes present in the others. Ultimately, this polynomial is simply modeling turnout; as long as the turnout for each age is rather consistent across all counties, the key will work. More population of a certain age yields more registrations at that age, which is multiplied by the polynomial model for turnout, yielding more votes for that age.

That said, factors affecting turnout among an age group (e.g. a college campaign encouraging students to vote, or increasing voting accessibility for a certain age group) *would* deviate from the state mean turnout by age, which is what this polynomial is fitting. Thus, more extreme counties *may* still break the mold. For example, here are the top and bottom three counties for <30 and >65 yrs of age:

[21] `Exhibit 4.pdf`, available in the DePerno Law collection of Bailey v. Antrim County documents: https://www.depernolaw.com/bailey-documents.html

Population < 30		**Population > 65**	
Isabella County	0.481	Keweenaw County	0.42
Houghton County	0.415	Ontonagon County	0.404
Ingham County	0.401	Alcona County	0.389
Keweenaw County	0.129	Washtenaw County	0.152
Alcona County	0.121	Ingham County	0.147
Ontonagon County	0.11	Isabella County	0.131

4.5 Correlation does not equal ~~causation~~ accuracy

We all know that "correlation does not equal causation," but did you know that correlation, as in Pearson's Correlation Coefficient (R-value), does not equal *accuracy*?[22] Here are some of the claims on accuracy in the brief:

- "The ability to predict ballot demographics with such remarkable precision (average correlation coefficient of R=0.997)..." (pg. 5)
- "...we can predict the number of ballots cast in a county to 99.7% certainty without seeing the results." (pg. 6)
- "Every other county may think they are clean. They are not. Indeed, the key works in Barry County with 99.6% certainty." (pg 12)

Brief aside: the second item is patently false. We were told above, "...we simply graph the ratio between the black and the red, which creates the polynomial." With registered voters *and* votes (i.e "results") in hand, Dr. Frank found a polynomial such that `votes/registered = poly` and then proceeded to show us that, indeed, `registered * poly = votes`.

The correlation coefficient is more like asking "do these variables move together in the same patterns?" This may seem like a nuance (doesn't moving together imply that one predicts the other?), but I will demonstrate that R-value does not imply accuracy with respect to error. Using Antrim County data, I computed an R-value of 0.990 (vs. 0.993 by Dr. Frank). This was sufficient to validate the data as reasonable in light of it being extracted from a screenshot. Ages were filtered to $\leq$90 because WebPlotDigitizer was sub-optimal at the tails where curves overlapped, and the error would be proportionally larger for such small values.

[22] See the definition of Pearson's coefficient: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. The numerator, covariance, may provide a more intuitive definition of what correlation means: https://en.wikipedia.org/wiki/Covariance

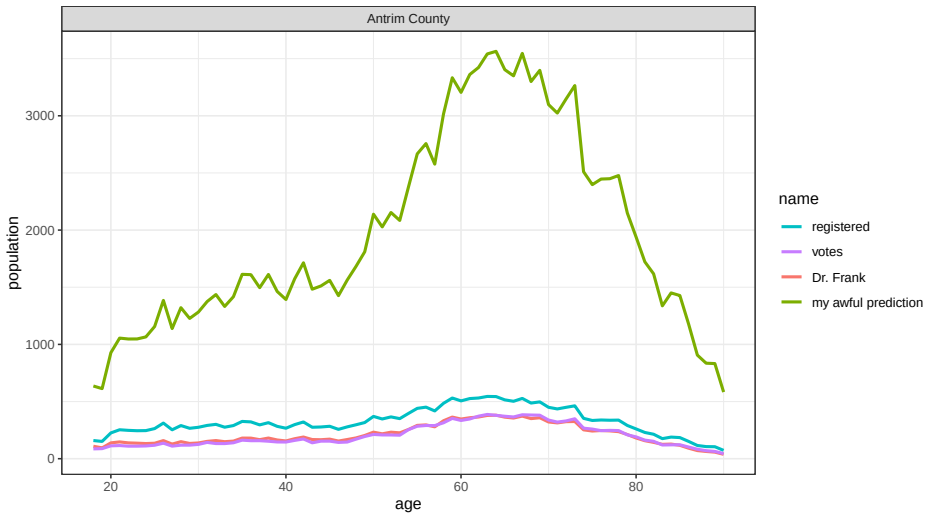
```
df_pred <- df %>%
  filter(CTYNAME=="Antrim County", var %in% c("reg", "vote", "pred")) %>%
  pivot_wider(id_cols=c(CTYNAME, age_i), names_from=var, values_from=val_i) %>%
  filter(age_i < 91)

cor(df_pred$pred, df_pred$vote)
```

[1] 0.9901213

Above, using population as a predictor resulted in a smoother curve that was still correlative (R=0.99), but the resulting prediction was still quite in line with actual votes. What if we used a different key? Here, I multiply registered voters by a uniform sequence from 4 through 8 the same length as the data:

```
df_pred <- df_pred %>%
  mutate(pred2 = reg * seq(4, 8, length=length(reg)))
```

Plotting this new prediction along with Dr. Franks data yields the following result:

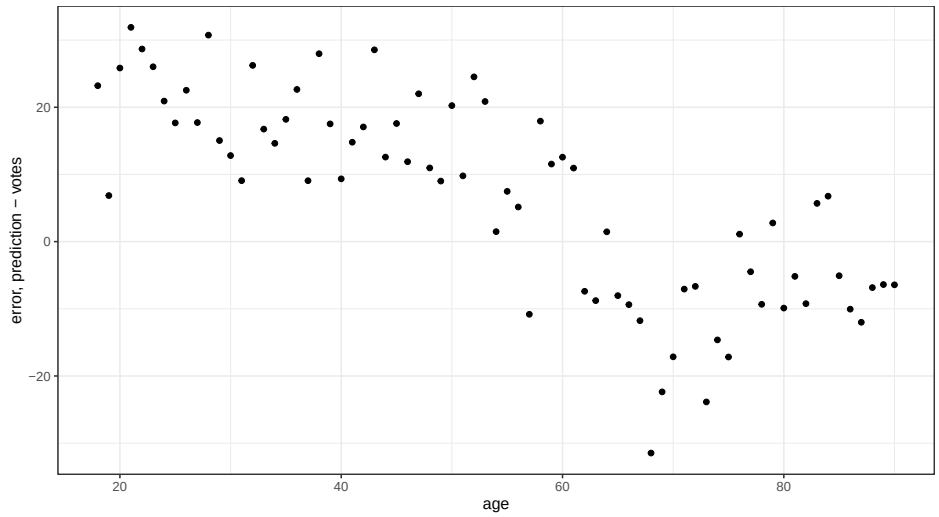


Could this prediction be described as "accurate" or "precise"? Yet what do we find?

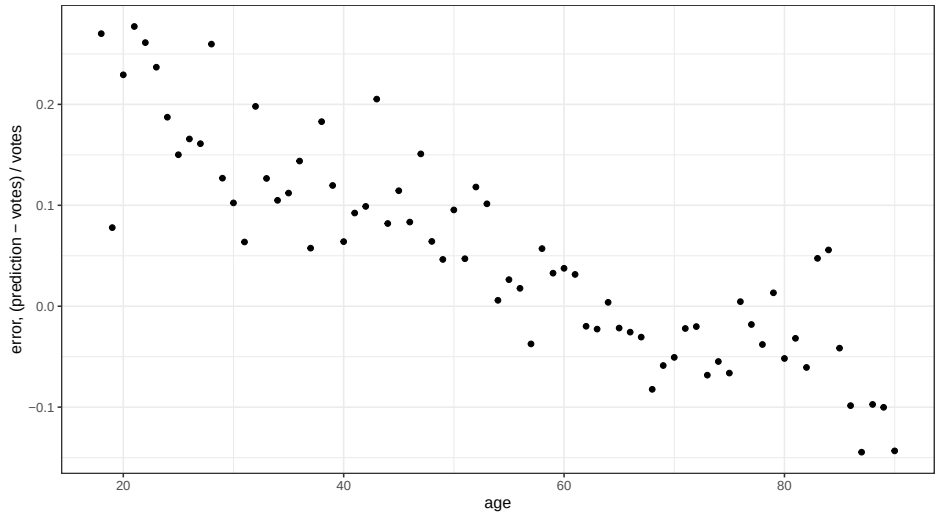
```
cor(df_pred$pred2, df_pred$vote)
```

[1] 0.9930492

How… is this possible? The correlation coefficient is useful for, well, correlating, and the data *do* correlate: "The red line is almost a direct image of the black line, but just lower on the graph." (pg 11) Indeed, my result is just higher on the graph. This toy example illustrates that correlation != accuracy. We can verify this further via a residual plot of `error = (prediction-actual)`:[23]



We're seeing errors of +30 to -20 across the range of ages, and if you didn't take note in earlier plots, Antrim County is *tiny*. Here's the prediction error as a percent of the voters at each age, highlighting that "99.3% correlation" can still manage -15% to 25% error.



[23] https://towardsdatascience.com/how-to-use-residual-plots-for-regression-model-validation-c3c70e8ab378

The residuals indicate a bias in the prediction. They should center about the line `error=0`, be randomly distributed, and have no obvious trend across the independent variable, `age`. We have a downward trend in this case, which I'll point out is in the opposite direction of what this theory proposes. The key is creating a prediction that's *too high* vs. the actual result for younger ages (+error) and *too low* for older ages (-error).

Root Mean Squared Error is a better assessment of accuracy, and is shown for all nine counties, computed using Dr. Frank's prediction (light blue) and reported votes (red) in the plots:

County	**RMSE, absolute**	**RMSE, fraction**
Antrim County	16.3	0.115
Barry County	22.9	0.068
Charlevoix County	14.9	0.093
Grand Traverse County	37.2	0.062
Kent County	135.6	0.04
Livingston County	59.6	0.052
Macomb County	145.7	0.03
Oakland County	320.2	0.042
Wayne County	367.7	0.039

4.6 Explaining the illusion

Having examined several math-based voter fraud theories, they tend to share some commonalities:

- the data, methodology, and code aren't shared
- honest scientists should poke at their work, explore counter-explanations, and discuss limitations ahead of time; these theories lack almost any rigor, yet the authors (and their attorneys) seem comfortable "publishing" to the courts to support accusations of the utmost gravity
- the theories ultimately contain some type of mathematical sleight of hand to portray something as "odd" or "weird" without ever defending this claim; we're never given a "non-weird" reference for context (cf. the birthday problem sounds "weird" despite it just being the math of probability manifesting in real life)[24]

Given all that's been said, then, what *is* the trick? Across theories I've analyzed, it often amounts to what *isn't* being shown more than what is. For starters, we're given no references to what this looks like in previous years, other states, and so on. Theories focus entirely on the swing states; what would this analysis look like in states that have zero justification to cheat? What would this analysis look like in states like Wyoming, Idaho, Utah, Oklahoma and West Virginia who all carried a ~20 point Republican margin?[25]

All we've been shown are some curves that match in shape and... that's it. Ask yourself why you think that these curves *shouldn't* match. Putting fraud aside for a moment, what would this look like with ~fraud glasses on?

- citizens of a certain age exist
- some of those *same* citizens register to vote
- some of *those* same citizens vote

What would "break" this explanation? How would it *not* be the case that these curves match? In other words, given some e.g. spike of 73yr olds who exist in the population, why wouldn't that spike in demographics show up in the number of 73yr olds who register and the number of 73yr olds who vote? Why *wouldn't* these values be in proportion to one another?

I think the true "magic trick" in this theory was accidental: the Census Bureau doesn't have data for individual ages, and thus a smooth approximation was used. Imagine if the shape of the population plot precisely mirrored the others. For whatever reason, intuition says it's "weird" to be able to scale the curve of registered voters into a similarly shaped prediction of resultant votes... but why? They *both* have a common ancestor: the age demographics of the population itself.

[24] https://en.wikipedia.org/wiki/Birthday_problem
[25] https://worldpopulationreview.com/state-rankings/most-republican-states

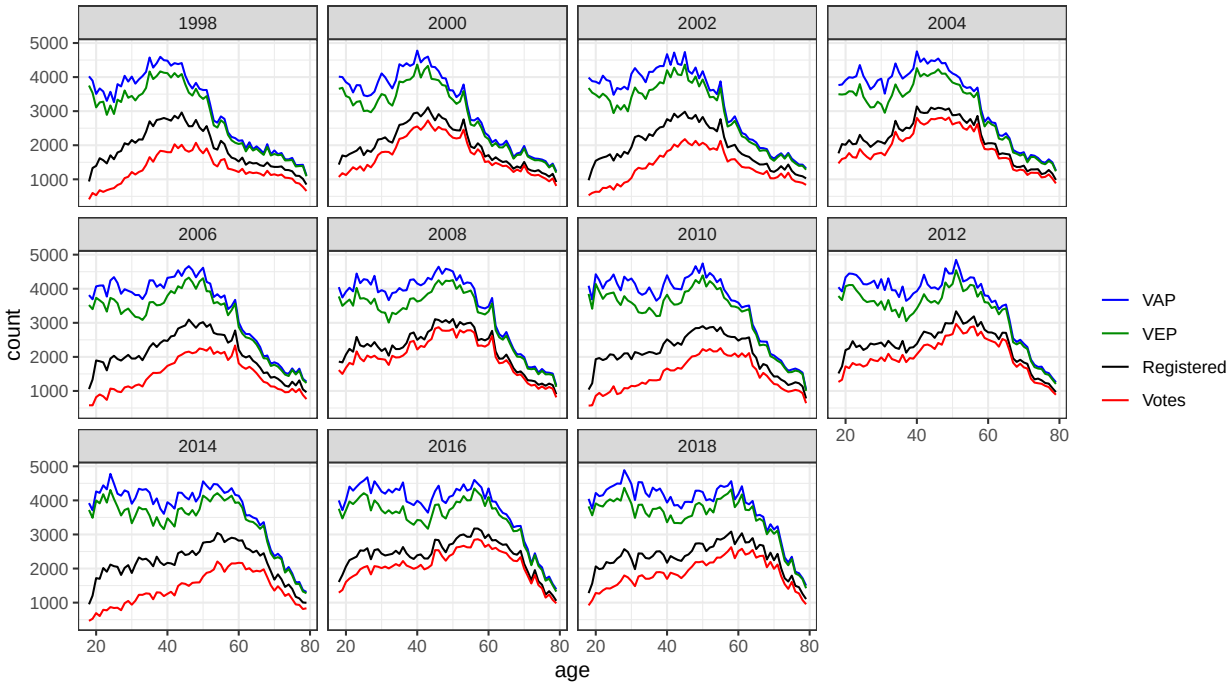
Now, there are certainly different turnout rates across the ages, but there's no "ledge" where suddenly a lower turnout among e.g. 26yr olds transitions to 100% turnout at age 27. Turnout is just an aggregate measure of contributing factors: engagement with politics, awareness, motivation, resonating issues in that election, mobility, other commitments, and so on. While we can point to characteristics among "younger" vs. "older," we cannot do the same for 26 vs. 27yr olds. Turnout differences across ages is a smooth transition, and *that* is all this polynomial is. Far from having discovered a state-controlled algorithm, the polynomial is the discovery that the same individuals of an age who exist… also register and vote.

This is also the real reason the key is <1 for younger ages: given some average turnout for a county, you have to decrease the prediction for younger voters because of their lower than average turnout rates. Given this same average turnout rate, older voters exceed it and have to be corrected upward by the ~~key~~ scaling factor to match reality.

4.7 Taking a step back

The root of this trick is that the population, as plotted, doesn't look like the other curves. How *could* we find out if both registrations and votes truly looked like the population? The Census has national election data for single years of age all the way back to 1998, containing Voting Aged Population (VAP), Voting Eligible Population (VEP), registered voters, and votes cast.[26]

Here's what that looks like:



[26]Starting in 2004 and earlier, total demographics were reported as "Total Population" (VAP) along with another column for "non-citizens." This was used to calculate VEP by subtracting the non-citizen count from VAP. From 2006 onward, the data contains both numbers for the total population (VAP) and US citizens (VEP).

Isn't it remarkable how the shape of the distributions match so closely for people who (a) exist, (b) are eligible to vote, (c) who register to vote, and (d) who actually vote? For giggles, note the spike around age 53 in the 2000 election. Follow it from election to election and you can actually see that bump in the population walk its way through time!

Contrast the reality that some ages are disproportionately high vs. their neighboring ages with how this is portrayed by Dr. Frank and DePerno (pg 10):

> Importantly, we always see at least 2 spikes on the right side of the graph that rise above the population line. This is a breadcrumb (a clue that the data is being controlled by something). In this case, the spikes on the right side actually reveal that an algorithm controls the results. This is a fact (not speculation) because they exist in every county in every state that has been tested. There is no way that every county in the US would have this same feature. **The spikes appear because every county in every state is being regulated by the census.**

There is no regulation at work other than the mere fact that the individuals who exist decide if they will register and subsequently vote. Here's an analogy:

- cars in the US have some demographics of color (red, silver, white, etc.)
- cars in the US also get in accidents
- using the distribution of cars sold by color (popularity), we could fit a polynomial curve that would accurately predict the distribution of colors among cars in accidents

Would this be evidence that the state is using an algorithm to programmatically control which cars get into accidents and at what rate? If we did this by year and saw a blip in a certain color (some shade of red was very popular one year) and this "blip" was present in *all* counties and states, would *that* be indicative of anything?

No: cars and people exist, and any reasonably uniform distribution of scaling factors (e.g. a smooth polynomial) multiplied across the population shape is going to to *retain that shape*. That "blip" in the year 2000 is now the *same* blip we're seeing around age 73, 20 years later. The blip exists in all counties because among other ages, that age is proportionally above average in existence. You'll also see a lot of white cars, no matter which county you visit.[27]

4.8 Reversing the illusion

To remove the illusion, we need to show that a truer representation of the population mimics both the registered voters and votes. This turned out to be quite difficult, as I couldn't find population by individual age *anywhere* via any Census Data. I did stumble upon an age pyramid for the US which was suitable for extraction via WebPlotDigitizer.[28] For transparency, I do not know how

[27] https://www.liveabout.com/most-popular-car-colors-4160630

[28] See image on this page: https://en.wikipedia.org/wiki/Demographics_of_the_United_States. Direct link: https://en.wikipedia.org/wiki/Demographics_of_the_United_States#/media/File:USA2020dec1.png

this data was obtained or inferred, though I have made a request for clarification.[29] Using this US individual age data, we can attempt to connect the dots between various data sets to better simulate a county population estimate:[30]

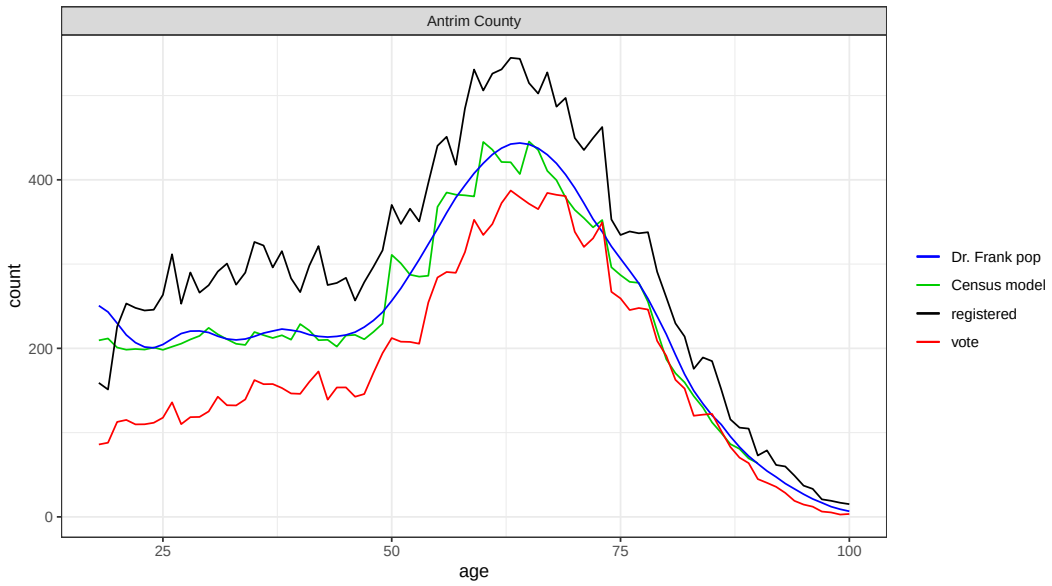
- US and county populations by age group, represented as `us_age_group` and `cty_age_group`
- the population for individual ages extracted from the age pyramid, `us_age_pop`
- the US and total county populations, `us_pop_total`, `cty_pop_total`
- relative frequencies for each age group: `us_age_freq = us_age_group/sum(us_age_group)` and `cty_age_freq = cty_age_group/sum(cty_age_group)`
- a `scale_factor` for each age group, `cty_age_freq/us_age_freq`

Now we can take the US population by age and (a) scale to match the total county population, and (b) adjust each age range according to the scaling factor. This latter adjustment preserves the relative "shape" of the US distribution (e.g. 73yr olds as more populous than 72 or 74yr olds) while accounting for a county a having higher or lower relative proportion of individuals in the 70-74yr old age group.

With the values above, we now use the age pyramid data for a specific age, `us_age_pop`, to predict the county population for that same age, `cty_age_pop`, in the following manner:

```
cty_age_pop = cty_pop_total/us_pop_total * us_age_pop * scale_factor
```

How does this look in practice for Antrim County?



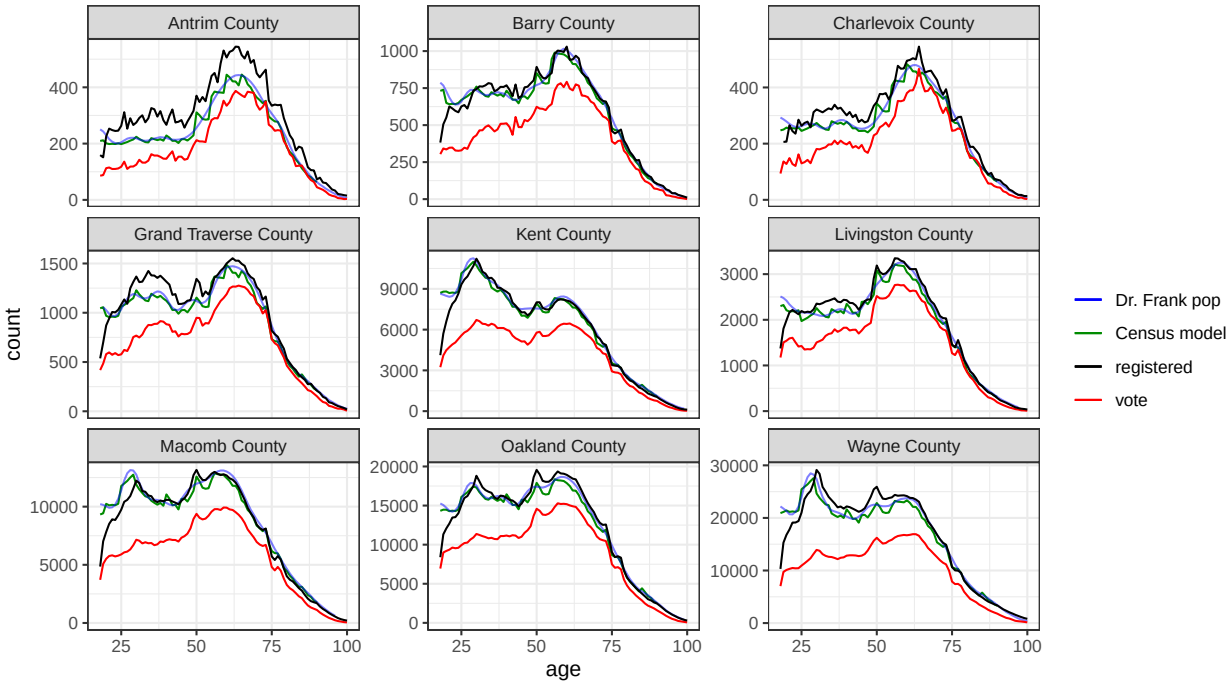
[29] See Talk page for this image: https://commons.wikimedia.org/wiki/File_talk:USA2020dec1.png

[30] Population estimates by age and sex, 2019. https://www.census.gov/data/tables/2019/demo/age-and-sex/2019-age-sex-composition.html

This is not perfect, nor can it be given what we have to work with. This is a crude attempt to scale *total US population demographics* to a small county, adjusting via per-county weightings (by groups of 5 years), using forecasts 9yrs out from the raw data they rely on. Still, this is a better approximation than Dr. Frank's, and it highlights why claims such as this are flawed:

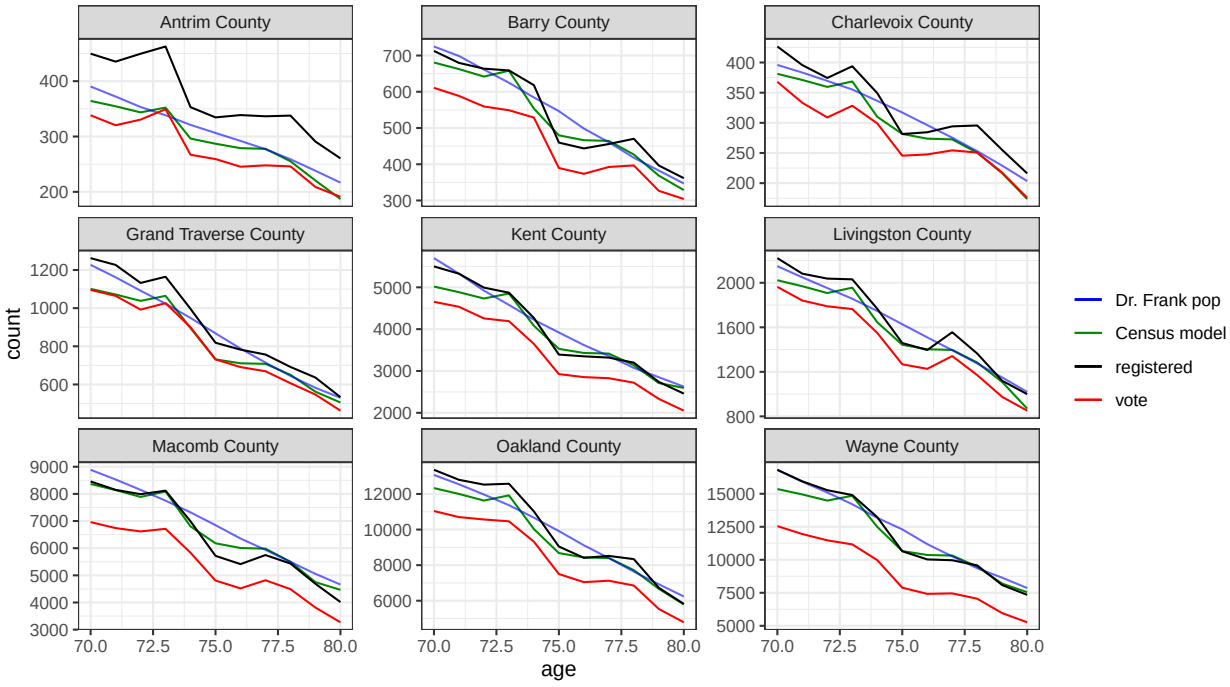
> Importantly, we always see at least 2 spikes on the right side of the graph that rise above the population line.

For starters, the statement is patently false: it's only *apparently* true in Antrim, Charlevoix, and Grand Traverse, which happen to be among the smallest counties (more sensitive to error). Even so, we can see that this derived population removed the more prominent of these occurrences in Antrim. The same is true when we replicate across all counties in Dr. Frank's data set. Note the superior fit to both registrations and votes vs. the smoothed population curve.



Zooming in on the suspect region of 70-80yrs, we see alignment between the spikes in registrations, votes, *and* our Census-derived population estimate. This is an observation worth pausing for. Dr. Frank's analysis found a connection between data from the same source and about the same general "thing" (registrations and votes both being tied to voting). This population data is from a source completely separate from the election data and it *still* matches. Independent data about the ages of people who are alive (though somewhat crudely simulated) matches *separate* data about the ages of people who registered and voted.

With more accurate data I'm confident we would see that in no cases did "votes exceed the population" as is claimed. Moreover, by using this population model we can see that these "spikes" and "notches" are just people who exist in the population, who register to vote, and who vote.



4.9 The student becomes the teacher

With this hypothesis in mind, can we recreate this effect out of whole cloth? I think this illusion is so effective because of the granularity of the per-age data. Because we've likely never studied age demographics in detail, seeing the shape of these curves mirror each other is hard to write off as "just how it's supposed to be." Humans are pattern-seekers, and seeing these patterns plotted and graphed for the first time jolts the brain's recognition software.

The primary tactic used by fraud theories is what I call "argument by inception:" namely, this "phenomenon" never existed in your mind before you were told to think of it. When have you *ever* paid attention to age demographic data before you were told that voter registrations matching the shape of votes collected is "weird"? When have you *ever* wondered about the number of unique last names in Pennsylvania until someone checked?[31]

Here is another statistic you have never checked: the united states algorithmically controls the issuance of drivers' licenses across age groups according to the proportion of female smokers.[32, 33] I used data from 2009 on smokers by age, and data from 2010 on drivers by age to discover that a 6th degree polynomial is being used to regulate the issuance of licenses in the United States with 99.3% ~~correlation~~ absolute precision.

[31] This was tweeted to the President, despite its rebuttal being one google search away. https://thebl.com/politics/voter-fraud-inconsistencies-revealed-with-last-names-of-registered-pennsylvania-voters.html. The Census has long known that 62% of last names belong to only one person: https://www.census.gov/library/stories/2017/08/what-is-in-a-name.html

[32] Gallup survey of smokers. https://news.gallup.com/poll/128183/smoking-age-baby-boomer-bulge.aspx

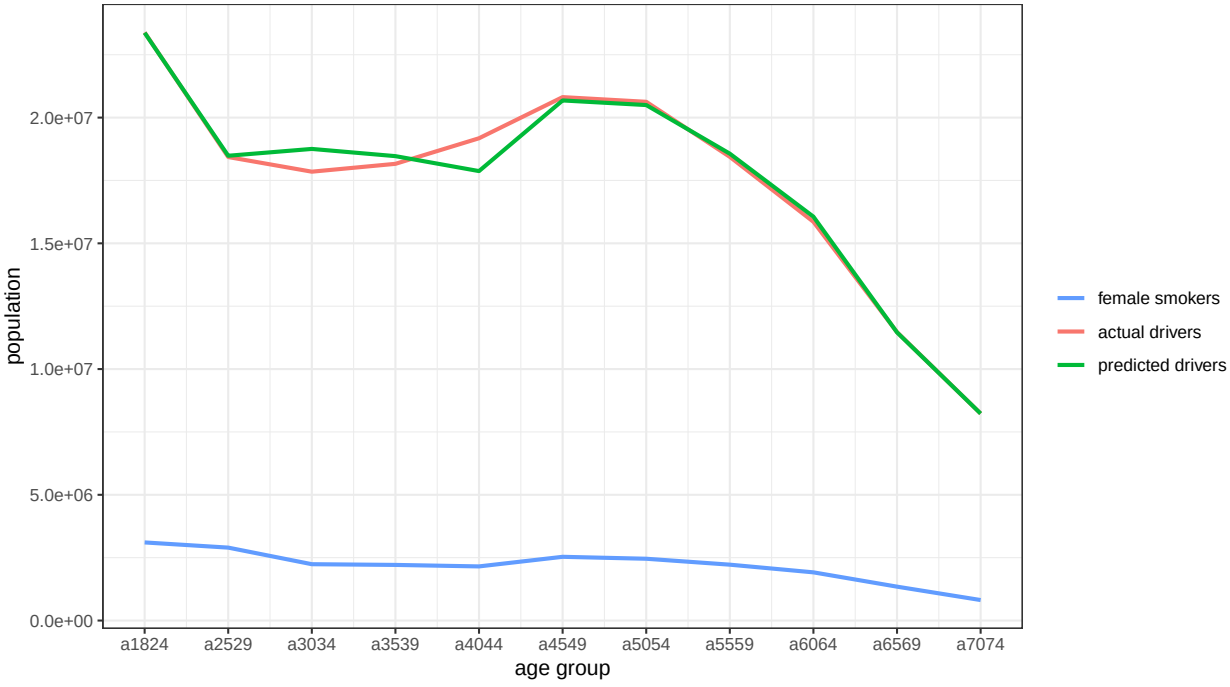
[33] Office of Highway Policy Information on drivers by age, 2010. https://www.fhwa.dot.gov/policyinformation/statistics/2010/dl20.cfm

After reading in the raw data (smoking data extracted from the plot using WebPlotDigitizer), I aligned age groups, merged the data, and applied a 6th degree polynomial model to find the following correlation value:

```
smk_drv <- smk %>% select(age_grp, smk) %>%
  merge(drv %>% select(age_grp, drv),
        by="age_grp")

fit <- lm(drv ~ poly(smk, 6), data=smk_drv)
smk_drv <- smk_drv %>%
  mutate(pred = predict(fit, smk_drv))

cor(smk_drv$pred, smk_drv$drv)
```

```
[1] 0.9926732
```

How does our prediction look (root mean squared error = 2.7%, btw)?



5 Conclusions

Hopefully this analysis has been successful in painting this proposed theory in a different light from many angles. Here is a summary of all core points made in light of the original claims:

Claim	**Response**
The blue, black, and red curves are "100% data"	The blue population curve was scaled and smoothed through 5yr age range estimates; it is not raw data
Registered voters are near or exceed populations in various counties	Agreed; Michigan should clean up its rolls, but we care about *votes* vs population, not *registered voters* vs population.
There are 66,000 votes that do not match IDs in the database	I was unable to verify this, it is not explained, and it is not apparent where to find this data at the source cited (vote.michigan.gov/VoterCounts/Index)
"...we always see at least 2 spikes" in which votes exceed county population	This only appeared to be the case in 3 of 9 counties; this was the result of using smoothed, aggregate population data, further compounded by the data being an estimate based on the 2010 Census
It is surprising that the registered voter and vote curves are so similar in shape	VAP, VEP, registered voters, and votes were verified to be "of the same shape" back to 1998; this is due to registrations and votes *deriving* from the population itself
The ability to predict votes from registered voters using a 6th order polynomial is indicative of a manipulative algorithm	Two variables that both correlate with the same upstream variable will also correlate
Obtaining high R-values showed that that the predictions were extremely accurate	RMSE is a more accurate gauge of error, and ranged from 3-11% for the nine counties; in addition, a preposterous, invented prediction of votes still achieved R=0.993
R-values > 0.8 will rarely occur when human behavior is at play	We "predicted" drivers from female smokers by age group; R=0.993 and RSME=2.7%
The "spikes" and "notches" across county data are surprising	The unique shape of age demographics (peaks, valleys, etc.) will show through in the result of uniform sampling or any relative smooth transformation applied (e.g. multiplying by a polynomial)

This theory amounts to being surprised that population age demographics manifest in other data across age demographics. To falsify this counter-hypothesis, Dr. Frank only needs to find *one* of any number of occurrences that do not fit this explanation. Some are proposed here, with others left as an exercise to the theorist and his lawyer:

- show that a state with a high +R or high +D margin in the 2020 election (e.g. WY, ID, UT, CA, MA, VT) where there would be no incentive for ballot manipulation does *not* have a strong correlation between population, registered voters, and votes across age demographics
- if specific software and/or voting machines are required for this fraud to be successful, show that these correlations do not exist in a state that does not use them
- show that these trends were not present in older data, when the use of machines and other fraud-enabling gadgetry did not yet exist (assuming 1998 is not sufficiently historic)
- if this US is viewed as uniquely suspect, find *any* example from outside the US in which a democratic country employing a free and trusted process for registrations and voting does *not* show these same trends

Merely saying something is "weird" does not make it so. When surprises emerge, it is a better practice to thoroughly scrutinize one's own mind instead of so easily blaming reality.[34]

[34] https://www.lesswrong.com/posts/tWLFWAndSZSYN6rPB/think-like-reality