A Statistical Analysis of Possible Electronic Ballot Box Stuffing, The Case of Baldwin County Alabama Governor's Race in 2002

By

James H. Gundlach
Auburn University

Paper Presented at the Annual Meetings of
The Alabama Political Science Association
Troy, Alabama, April 11, 2003

The 2002 Governors race in Alabama erupted into controversy when Baldwin County first reported results that suggested that the democratic incumbent, Don Siegelman, had won and subsequently reported other results that gave the election to the republican challenger, Bob Riley by a margin of 3,120 votes out of 1,364,602 cast.

In this paper I demonstrate how relatively simple statistical techniques can identify apparent systematic electronic manipulation of voting results. This paper consists of four parts. The first part is an overview of the election; the second part is an analysis of county level data that suggests that both sets of results from Baldwin County are anomalous. The third part of the paper is a set of analyses of results from voting districts that identifies and describes some clear patterns in the anomalous Baldwin County final results. The final part of the paper discusses the possibilities of electronic vote manipulation and suggests mechanisms for preventing it in the future.

Some election background

I first looked at the election results by dividing the state into two components, Baldwin County and the rest of the state. I then look at the results for the 1998 and 2002 Governor's races for these two components. In the 1998 election Don Siegelman received 742,766 votes in the rest of Alabama compared to 533,772 votes for his the republican, Fob James. In Baldwin County, which is where Fob James lived, Siegelman received 17,389 votes compared to 21,004 votes for James. In 2002, Siegelman received 635,545 votes compared to 623,145 for Bob Riley in the rest of Alabama. In the first set of returns for Baldwin County Siegelman received 19,070

votes to 31,052 for Riley. In the second set of returns Siegelman was reported as receiving 12,736 votes while Riley's total did not change.

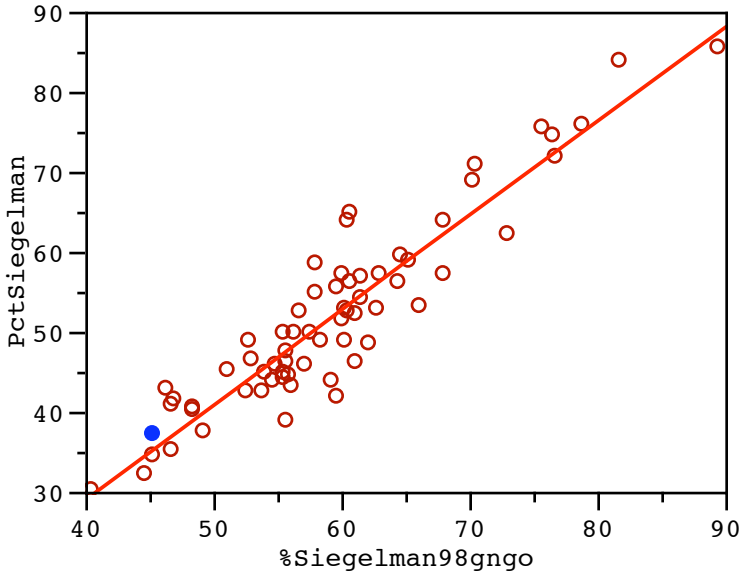
In addition to the issues reported in the press, the above review suggests three points about Baldwin County's election results that should make one suspicious.
One is the unusually large increase in the votes for the republican candidate from the 21,004 for James, the local resident in 1998 to the 31,052 for Riley in 2002. While Riley ran better across the state than James did, the only other county that showed such an increase for Riley over James from '98 to '02 was Riley's home, Clay County, which went from 2,122 to 3,176.

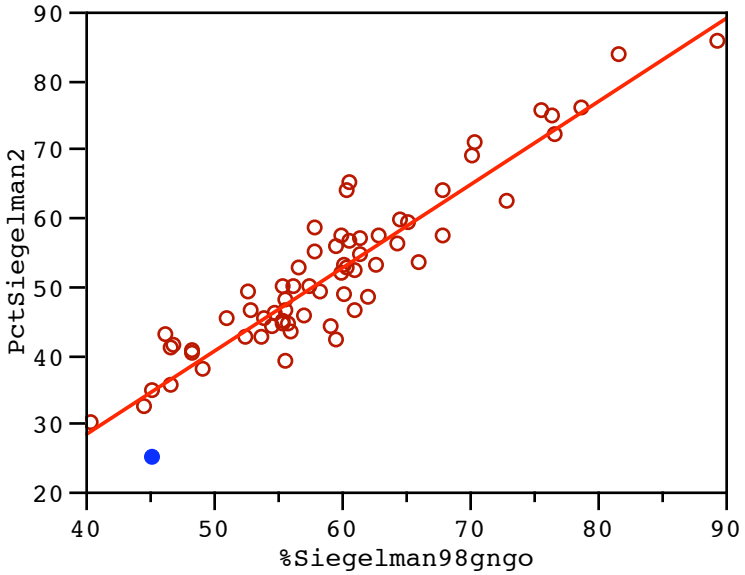
The second factor that raises suspicions is the size of the decline in Siegelman's reported vote. The difference between the two reported votes for Siegelman is a decline of almost exactly one third of the total votes finally reported for Siegelman. A one third reduction is commonly found in data that is intentionally changed but rarely the result of random errors.

The final point that raises suspicions is that there should be no way to produce two different results with the computerized vote tabulation. That is, the system should not allow access to computer code or procedures that can produce different results. Computers do not accidentally produce different totals. Someone is controlling the computer to produce the different results. Once any computer produces different election results, any results produced by the same equipment operated by the same people should be considered too suspect to certify without an independently supervised recount.

A County Level Analysis: Baldwin County as an Outlier

The primary method of analysis examines changes in the level of votes for the democratic candidate between 1998 and 2002. The data were obtained from the elections page of the web site of the Secretary of State of Alabama. The Alabama Secretary of State site provides a substantial amount of data for Alabama election results in Excel format free for the downloading. The first set of analyses regresses the percent of the county vote for the democratic candidate in 2002, using both sets of Baldwin county's returns, on the percent of the county vote for the same candidate for Governor in 1998. The results of these analyses are presented below:

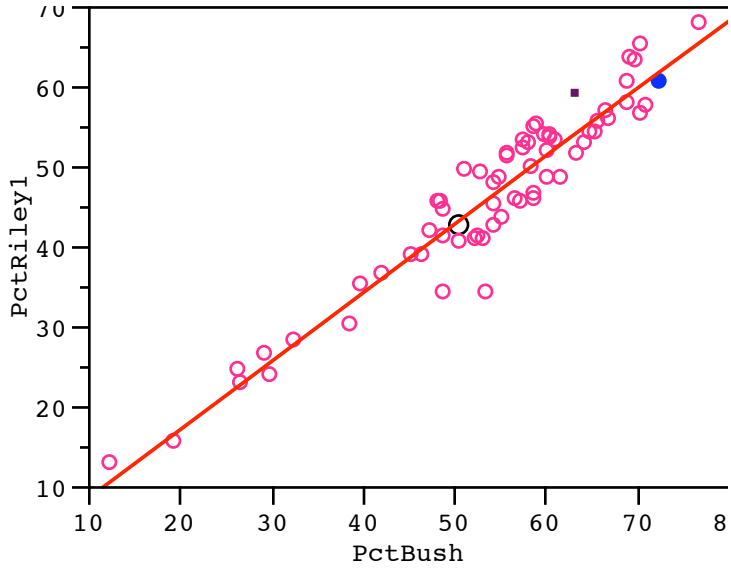


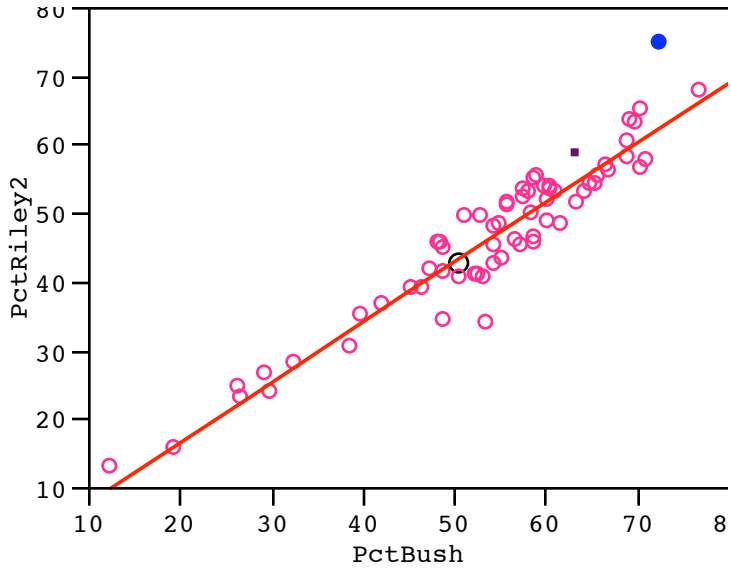


The first plot shows the relationship for the initial returns and the second shows the results for the second set of Baldwin county returns. The solid dot is Baldwin County. Note that in the initial returns, Baldwin County fits closer to the line than most of the other counties. But, in the second Baldwin County is further from the line along the vertical dimension than any other county. These results suggest that the changes made between releasing the first and second set of results made Baldwin County an outlier. This is exactly the opposite of what you would expect if the changes corrected an error in the data. That is, an error usually make the

data point deviate from expected patterns and fixing the error typically moves the data point back into the pattern. This kind of statistical irregularity deepens one's suspicions about the final Baldwin County election results.

However, there could be factors operating on the republican side that may account for this. Baldwin County voters, like most in Alabama, voted with strong majorities for President Bush in the last election. And Bush campaigned for Riley in Alabama. It can be argued that this change in the vote simply reflects the effect of Bush on the Baldwin County voters. Analyses using the percent of vote for Bush in 2000 to predict the percent of vote for Riley in 2002 are presented below. The first scatter plot shows the Baldwin County percentage based on the first reported results. Again the first results, those giving the election to Don Siegelman, fit the pattern shown by the other counties. The second results has Baldwin County an extreme outlier showing that Riley is receiving about fifteen percent more of the vote than predicted by the vote for Bush in 2000. The small square dot is Clay County, Riley's home. It is worth noting that home county advantage is only about a third as large as the second count gave Riley in Baldwin County. Thus, the analysis of the republican side of the election also increases the suspicion that Baldwin County results are manipulated.





An analysis of Voting District Data

Given that the county level analysis deepens rather than reduces suspicions that the Baldwin County results are manipulated to make Riley the winner, I then extended the analysis to a set of voting districts to see if the suspicious pattern was a significant deviation from expected results. For this analysis I produced a data set for comparable voting districts in Baldwin, Montgomery, and Shelby Counties for the Governor's race in 1998 and 2002. It should be noted that determining which voting districts are comparable over two elections that are four years apart is difficult and tedious because several voting districts were changed either in name or boundaries or both between the two elections. I attempted to add the voting districts for Jefferson County but was unable to verify sufficient compatibility between the 1998 and 2002 results to include them in the analysis. Out of Montgomery and Shelby Counties' 122 voting districts I was able to verify sufficient consistency in names and boundaries to use 70. Of Baldwin County's 65 reported voting districts I was able to produce 39 comparable voting districts for both 1998 and 2002. Most of this reduction in numbers was due to aggregating boxes or beats to insure comparability. For example Fairhope Civic Center has boxes for four voting districts that shifted their boundaries between them from 1998 to 2002. By aggregating the four boxes into one geographic unit for both 1998 and 2002 I

was able to create one comparable voting district for both years. I conducted regression analyses similar to the county level analyses above for two sets of voting districts, those outside of Baldwin County and those in Baldwin County. If the first reported results were accurate for Siegelman but inflated for Riley, you would expect the slope for Baldwin County to be about a third lower than the slope for the other voting districts. The results are shown below:

Montgomery and Shelby County Results

Term		Estimate	Std Error	t Ratio	Prob>|t|	r
Intercept		7.8636457	23.97778	0.33	0.7440	0
Siegelman98		0.853541	0.029685	28.75	<.0001	0.961 25

Baldwin County Results

Term		Estimate	Std Error	t Ratio	Prob>|t|	r
Intercept		-9.771309	11.75185	-0.83	0.4110	0
Siegelman98		0.6979152	0.021321	32.73	<.0001	0.983 16 9

There are two important points to note in comparing these results. First, the correlations, r=. 96 and .98 are quite strong. That means that the 1998 Siegelman vote is an adequate predictor of the Siegelman 2002 vote. Second, is the difference in the estimates. The estimate, or slope, for the voting districts outside of Baldwin County is .85. This means that for each vote that Siegelman got in 1998 he got .85 votes in 2002. The fact that this matches the pattern in the state outside Baldwin County suggests that the selected voting districts adequately represents the rest of the state. Remember that in the first part of the paper, I showed that Siegelman got 85% of the vote in 2002 that he got in 1998. The regression for the Baldwin County voting districts shows a slope of .697. This means that results for Baldwin County are substantially

different from the voting districts outside of Baldwin County. A significance test for the difference between the slopes shows that the two slopes are significantly different from each other, $t=6.19$, $p<.0001$.

The combined findings of a strong relationship between the 1998 and 2002 votes in Baldwin County as well as outside, and the different slope strongly suggests a systematic manipulation of the voting results. In addition, a comparison of the slopes provides a way to estimate the apparent nature of the manipulation of the results. By dividing the Baldwin County slope, .697, by the slope for the other voting districts, .854, and subtracting the results from 1.00 you get an estimate of the proportion of the Siegelman vote in each voting district that that apparently disappeared from the official Baldwin County results. This yields a result of .18, which is about half as much as predicted from the hypothesis based on the first reported results. This raised the question about how could a process of moving X number of votes from one candidate to the other results in the mysterious production of erroneous results that were 2X above the final reported results?

The answer is surprisingly simple. This is a common error pattern to appear in programming spreadsheet calculations. My hypothesis is that someone was moving a little more than 3,000 Baldwin County votes from Siegelman to Riley by calculating a fifth of Siegelman's votes in each voting district, rounding it to a whole number, adding the resulting value to Riley's votes in that district and then subtracting that number from Siegelman's vote. However instead of subtracting the calculated number they added it to the vote for Siegelman. This is a common error created by using copy and paste to produce the invisible formulas for cells of spreadsheets. The result was a first report of county vote totals that had percentage distributions close to what was expected but a total vote that much higher than expected. Once they went back and fixed the procedure so that it performed as they desired, a reasonable total vote and Riley winning the election, the difference between the first and second reporting of Siegelman's vote was twice the number of electronically shifted votes. If what I hypothesized happened, then the total votes for Baldwin County was 27,866 votes for Riley and 15,283 votes for Siegelman. This would have produced state totals of 669,039 votes for Riley and 671,652 votes for Siegelman. The only way we will know for sure is if the paper ballots for Baldwin County are recounted.

How Baldwin County Results Could Have Been Manipulated:

When Baldwin County reported two sets of results, it was clear to me that someone had manipulated the results. There is simply no way that electronic vote counting can produce two sets of results without someone using computer programs in ways that were not intended. In other words, the fact that two sets of results were reported is sufficient evidence in and of itself that the vote tabulation process was compromised. I will next describe how the system employed in Baldwin County is supposed to work and suggest four ways in which the results could have been manipulated.

The system employed in Baldwin County works much like a digital camera that stores the pictures on a computer chip that is physically removed from the camera and inserted into a reader that is attached to a personal computer that then transfers the images to the hard drive of the personal computer to be edited and used. The voting machine reads a paper ballot and writes the information on a cartridge that saves the results of all the ballots cast on that machine. After the polls close, the cartridge is transported to the county courthouse where it and all the other cartridges from all the county voting machines are inserted into a reader attached to the tabulating computer and the files are transferred to the cartridge to the hard drive of the tabulating computer. Once all the files have been transferred to the hard drive of the tabulating computer, a previously written program is ran to read the files from the individual voting machines and produce summary tabulations. The code in the tabulating machine is not supposed to be changed. This system provides several points at which the data can be tampered with.

First, and perhaps the least likely, is altering the recorded information on the cartridges between the polling place and the county courthouse. This is especially difficult to do if the results of all the voting boxes are to be changed. And, it would require using a computer that could emulate the output from the voting machine.

A second way, would be to install something like a computer worm or virus on the tabulating computer that would intercept the data stream when the cartridges were being read, modify the data in a desired way, and send the modified data to the hard drive. This would require someone with quite high level of computer programming skills and would be fairly labor intensive. It would also

be difficult to specify the amount that the results should be fudged when needed. But it would be a modification that once created and put in place could well modify the results in every county that uses this system.

The third approach would simply require access to a program to could edit the data files once they are stored on the computer hard drive using the keyboard and monitor attached to the tabulating computer. This would require a relatively long period of unobserved access to the tabulating computer between reading the cartridges and tabulating the final results. News reports suggest the opportunity for this kind of manipulation was available in this case.

The fourth approach, and the one I would take if I were to do it, would be to install a 802.11 card on the tabulating computer, along with enabling software, and use a similarly equipped laptop in a nearby room to modify the data files immediately after they were read from the cartridges. This would simply require access to the tabulating computer at some time before the election to install the card and after the election to remove the card.

Conclusion:

In this paper I show how some relatively simple statistical analysis techniques can be used to identify probable electronic manipulation of voting results. The Baldwin County results attracted attention because two results were reported. This was probably due to mistakes made in the data manipulation procedures. If this kind of electronic ballot stuffing is done in the future, voters and candidates cannot count on similar errors to serve as flags to bring the process under review. With a little work before an election building necessary data sets and some more work entering returns on election night and over the next few days, statistical analysis can point to probable ballot stuffing, electronic or otherwise. This could lead to more honest vote counting as well as greater trust in the electoral process and government in general.