Using Discriminant Analysis to Examine Recent Trends in San Francisco Elections – A Method to Analyze Multi-Candidate Races

David Latterman
Fall Line Analytics
dlatterman@flanalytics.com

Summary

Because of Ranked Choice Voting, future San Francisco elections will have multiple candidates without a runoff scenario. It is then necessary to use methods that can analyze these types of races, as well as have some predictive value. Discriminant analysis is a technique that predicts outcomes based on a set of independent variables. Here, I look at the results of the last two Mayoral general elections (2003 and 1999) and the ballot measures of that year that help explain the outcome.

Discriminant analysis can also help to compare how votes on a particular set of issues match up to how people chose their candidate. In addition, the analysis can find the issues that best reflect how voters or precincts choose their candidates. With this, it is possible to discern what kinds of issues voters use to choose candidates in a citywide race.

In these analyses, I find that good government ballot measures are an important indicator of whom voters choose, both in 2003 and 1999. In 2003, homelessness and social welfare issues also matched well with candidate choice. In 1999, it was also business and transportation issues.

By knowing what issues voters connect candidates to, we see how campaigns can tailor their message geographically to important issues to sway voters. Each campaign is different, and there are different important issues each year. By looking at several races over time, we can establish a pattern of voter trends given certain combinations of candidates and issues.

Introduction

Much of the analytical work that has been done in San Francisco from myself and others pertains to examining the results of a specific election. It is generally a post-election explanatory examination of what factors contributed to a win or loss for a candidate or an issue. Ordinary Least Square (OLS) regression models lend themselves to these types of analyses, due to the nature of the quantitative electoral data.

Rich DeLeon, with his Progressive Voter Index (PVI), moved away from this model by statistically aggregating past initiative elections into a large, internally consistent factor. PVI has both explanatory and *predictive* value, since PVI correlates quite well with most San Francisco issues in which there's any difference between the left and the right.

There are several other types of multivariate analyses that deal with quantitative and categorical data that can be used to look at San Francisco (or any) elections and trends. Ultimately, we want to find techniques that have predictive value as well as explanatory value, and look at a bigger picture. We also want to use analyses that can examine elections with more than two candidates. Now that Ranked Choice Voting (RCV) is here to stay, it's no longer prudent to look only at the vote count of one candidate. Finally, we would like to find a way to distinguish among the myriad ballot measures San Francisco voters face each year, in order to figure out those that voters used to choose a candidate.

This paper assumes that, for the most part, San Francisco voters place issues before candidates, and use their positions on various issues to make their candidate choices. Although San Francisco candidates at times assume a cult of personality, generally, people choose candidates based on their social values. It follows that we can use voters' issue choices to try to predict or at least understand their candidate choices.

Discriminant analysis

Discriminant analysis is a type of multivariate analysis in which a group of independent variables can be used to classify a group of values, based on some kind of dependent variable. Here, as with OLS analyses, the units of analysis can be poll respondents (voters) or precincts, while the independent variables are quantitative; for example, a set of demographic variables or past election results. The goal of the model is to classify as many voters or precincts correctly as possible. The more people or precincts the model classifies correctly, the better the model, and the more useful we can say the independent variables are in predicting outcomes. We want to know what independent variables best classify those values.

One advantage of discriminant analysis is that the dependent variable is categorical. This means we don't look at the percentage of a vote that someone received as the dependent variable. Instead, we look directly at the outcome. For instance, instead of the dependent variable being the percent vote candidate x received, it's whether the outcome was actually candidate x, or candidate y, or candidate z.

The dependent variable can be two or more categories. It can be yes/no on a proposition, or a choice among two or more candidates. Generally, discriminant analysis is often used when there are more than two dependent categories. Because discriminant analysis has some pretty strict mathematical assumptions, logistic regression is usually used when there are only two categories. When there are more than two categories, logistic regression output is somewhat harder to understand; moreover, discriminant analysis provides graphical output, so we can see the results a little more clearly.

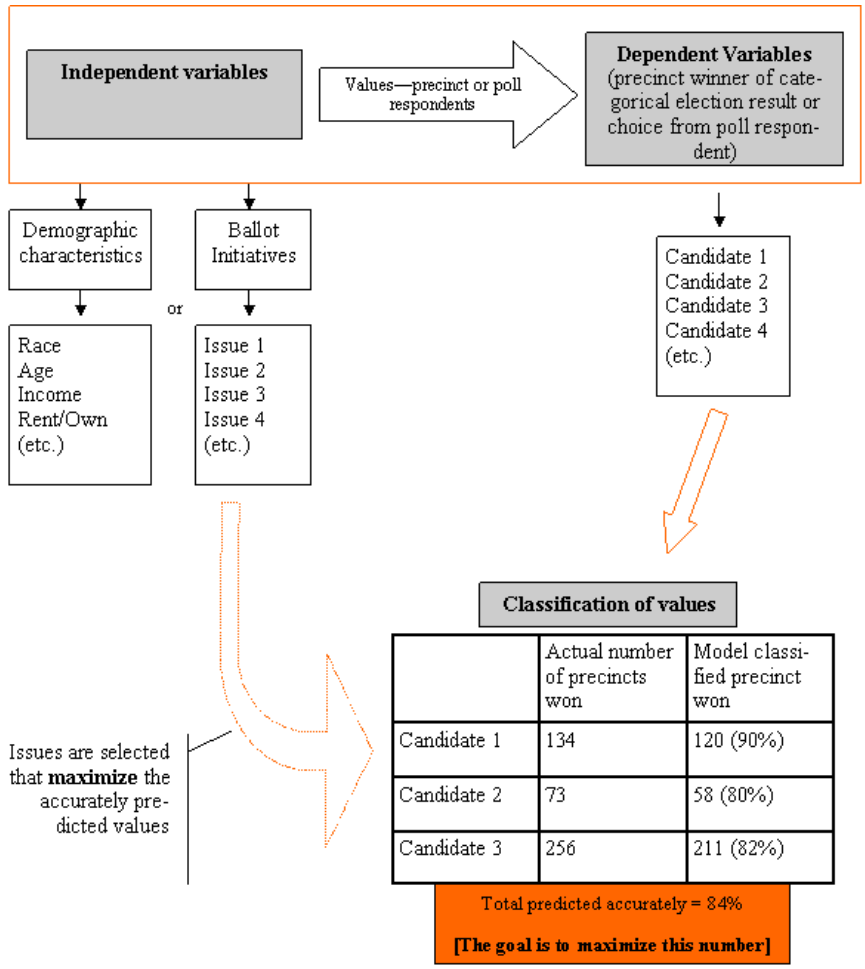
So in a political discriminant analysis, we try to "predict" whether or not the unit of interest (again: voter or precinct) will "vote" for the candidate (which is the dependent variable) based on a set of independent variables. For instance, in the 2003 Mayoral general election, there were several candidates, and a bunch of initiatives. Looking at the initiative results by precinct, we can predict what mayoral candidate the precinct *should*

have chosen, based on our prediction. We can also do this with demographic variables, or a collection of selected past initiative or candidate results. We use precincts as the units of analysis because data are easily available - and I use them here - but this works very well with individual poll respondents, where the candidate they choose is the dependent variable, and the independent variables are various other questions.

In discriminant analysis, we can choose to either enter a bunch of independent variables into the model altogether, and see how well the outcome is predicted; or, we can enter the variables "stepwise", which means the model will select the issues that most successfully match the dependent variables to the "correct" outcome. Therefore, in a race with several candidates, and many issues, we can find the issues that best predict the candidate outcome, by what percent of people or precincts the model gets right.

Figure 1 shows the process graphically. Independent variables help classify values by predicting how they will the dependent variable. In a stepwise model, the model will select the combination of initiatives that best classify the values. Follow the chart clockwise.

Figure 1:Graphical look at the process of a discriminant analysis. Independent variables help to classify values into groups.



What discriminant analysis can tell us

In San Francisco most elections now have multiple candidates. And, with RCV firmly in play, we won't have two-candidate runoff scenarios anymore. All elections – at least those with candidates - for the foreseeable future will need to be analyzed looking at several candidates as the outcome. This process lends itself to one of the multivariate techniques used to analyze categorical results of more than two categories. It also lends itself to analyzing polls in which the respondent has many candidates from whom to choose.

We often try to link election results to various factors, and then turn around and predict the next election using similar sets of factors. Other similar kinds of predictive analytics are often used with the larger national elections, but are also applicable at the local level if the data are good enough.

Discriminant analysis can sift through dozens of variables – like ballot initiative results or demographic variables, to see which ones best predict a precinct voting trend. The issues that the model selects will best successfully classify the precinct result. We can then say these issues were likely important factors in determining what concerns the voters had in mind when they chose a certain candidate.

It is important to note that these model-selected issues aren't necessarily the most important topics of the day in the voter's mind – it's hard to know that without a direct exit poll; but, we can find the issues (and kinds of issues), that best consistently classify voting behavior.

By looking at what precincts are correctly classified by the model, and by looking at those that are misclassified, we get a brief glimpse into what voters are thinking when they vote.[1] We can examine why a voter or precinct is misclassified – perhaps there's a pattern that hinge on one issue that most analysts overlook.

Of course, this is still a post-election analysis. But by looking at the results, we can begin to predict what kinds of initiatives match best with candidate results. This process basically links issues with candidates. We see what issues best "predict" an election result.

[1] As usual, in dealing with aggregate precinct voting totals, this is only a proxy for individual voting behavior. Ideally, we'd perform this on poll data, or a reliable citywide poll, but since they aren't usually readily available, precinct data must suffice. I gladly accept poll data donations.

How discriminant analysis works – an example[2]

Imagine we conduct a survey that asks some well-traveled people to group selected cities into three categories: highly desirable, somewhat livable, and please-god-don't-make-me-live-there. We ask the respondents to rate the cities on many different characteristics, like weather, size, traffic, etc. What discriminant analysis does is create functions – from two to *n*-1 where n is the number of independent variables – that assigns probabilities to the cities that they belong in their respective groups.

Perhaps the first function (called *canonical discriminant functions*) divides the cities by weather, traffic, and cultural amenities. This function divides the cities pretty well into the highly desirable and please-god-don't-make-me-live-there groups. But a second function is needed to break out the somewhat livable group. Maybe this function uses independent variables sports facilities, parks, and size. And if this function isn't enough, the model will create more until the explanatory power of each function drops below certain statistical limits.

The canonical functions are used to see which independent variables best discriminate the groups. All independent variables will be assigned to some canonical function, but some won't be statistically significant, which means they aren't as useful in explaining the dependent variable. Then, the model creates classification functions, based on the canonical functions, to assign each value to its respective group. Each unit of analysis is given a classification score that may or may not assign it to the correct group. This is based on the geometric distance from the score from the mean of all the groups, which is called the group centroid. The better the canonical functions, the more likely the model is going to correctly assign the value to its proper group. There is the same number of classification functions as there are groups.

In terms of output, we generally look at two things. First, we examine the scatterplot of each value's classification score, where the x and y axes are the values from best two canonical functions. This provides a graphical look at how well the model discriminates groups. Second, we look at the classification table, like the one in Figure 1, which tells us how accurately the values were predicted (see Figure 2 and Table 1 to see a made up output chart and classification table). Of course, San Francisco is firmly in the 'highly desirable' category.

Finally, in discriminant analysis, it is standard practice to break the sample size in half. The analysis is performed on one half of the sample, in which we know the outcome (in this case, how the respondent rated the city). Then, the functions that are created are applied to the other half, called the holdout sample, so the model can be tested without

---

[2] This is an extremely general overview. For more information, see Hair J., Anderson, R., Tatham, R., Black, W., 'Multiple Discriminant Analysis and Logistic Regression', in Multivariate Data Analysis, Prentice Hall, 1998.
Online, see: http://www.statsoft.com/textbook/stdiscan.html and http://www2.chass.ncsu.edu/garson/pa765/discrim3.htm

knowing the answers ahead of time. Usually, the results aren't quite as good for the holdout sample, but the results are more real in that the model is tested again a "new" set of values.

As with any analysis, it's important to have enough values and independent variables to be statistically significant. A good ratio is 20 values to one independent variable, but it can be less than that. If the city survey asks about 50 cities, we shouldn't use any more than 5 variables. This also affects whether or not we use a holdout sample.

Figure 2: Sample graphical output from city survey discriminant analysis. This analysis fares pretty well grouping the highly desirable cities from those that are not, as evidenced by whether or not the groupings intersect. It doesn't group the middle values quite as well, which is typical in discriminant analysis with two extreme choices and a middle choice. Notice the blue group intersects with both the red and yellow group.

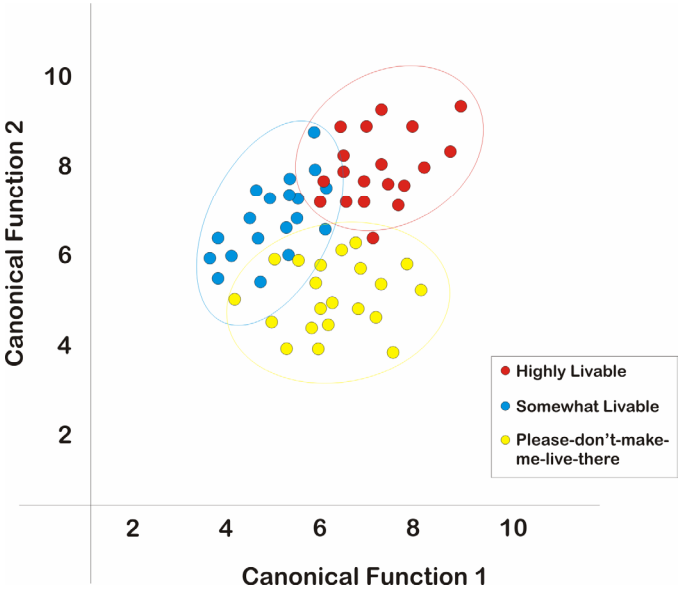


Table 1: Classification percentages of sample city survey discriminant analysis. As we saw in Figure 2, the model was accurate for the highly desirable and not so desirable categories (84% and 86%, respectively). The middle group was predicted less successfully (68%). However, with an overall success rate of 80%, this is a pretty good model, and therefore we can say that the independent variables in the model (survey questions) are good predictors of how people will rate cities.

	HL	***SL***	***PGDMMLT***
n	19	19	21
Classified correctly	16 (84%)	13 (68%)	18 (86%)

Total classified correctly: 80%

Using discriminant analysis to delineate San Francisco political trends

My goal in using discriminant analysis is to see if we can determine which issues – meaning ballot initiatives – are better predictors for how people will vote in multi-candidate elections. We look at past elections to see which initiatives best classify election results. For any election with a dozen or more ballot initiatives, perhaps there are a few of those that can accurately predict a precinct's choice of candidate. The next time around, these issues can be used as the basis of a campaign marketing effort, or in a poll to see how to push voters to the issues that they care about in relation to a candidate.

For this paper, I examined two elections: the 2003 and 1999 general Mayoral elections. I looked at several other races, including the 2003 DA race and the 2004 School Board race, but I thought these best exemplified the method and what it could do. It is interesting to see the differences between the 1999 and 2003 races and what issues prove to be important.[3]

2003 Mayoral general election

Background

For this analysis, the units of analysis are the precincts (571), and the dependent variable is which candidate the precinct selected. Table 2 shows the frequencies of the Mayoral winners. For the independent variables, I use all of the November 2003 local measures. These are the issues the voters had to face at the time they chose their Mayoral candidate.

Table 3 displays all the November 2003 measures. This is a good sample of issues given half the number of precincts (235), which I used as a holdout sample to test the model. There were 13 issues from 2003, so it's an 18:1 independent variable to sample size ratio, well within acceptable limits.

Table 2: Frequency of precinct choices from the 2003 Mayoral election

Candidate	Precincts won	Percent
Gavin Newsom	439	76.9%
Angela Alioto	35	6.1%
Tom Ammiano	3	0.5%
Matt Gonzalez	94	16.5%
Total	571	

[3] I'm going to present the summary results and interpretation of the Mayoral analyses. If anyone wants to see the entire technical workup of these analyses, please email me at dlatterman@flanalytics.com and I'll email you the various tables and/or graphs that go along with the analyses.

Table 3: Ballot initiatives from the November 2003 election used as independent variables

Prop	Title	Result (Pass/Fail)	Percent Voting Yes	B=Bond Issue C=Charter Amendment O=Ordinance
A	School Bonds	P	70.6%	B
B	Retirement Benefits for Safety Employees	P	66.8%	C
C	City Services Auditor	P	70.6%	C
D	Small Business Commission	P	55.9%	C
E	Ethics Reform	P	61.9%	C
F	Targeted Early Retirement	P	67.6%	C
G	Rainy Day Fund	P	75.8%	C
H	Police Commission / Office of Citizen Complaints	P	51.9%	C
I	Child Care for Low Income Families	P	59.9%	O
J	Facilities for the Homeless	P	58.7%	O
K	Sales Tax for Transportation	P	74.8%	O
L	Minimum Wage	P	59.6%	O
M	Aggressive Solicitation Ban	P	59.7%	O
N	Taxi Permit Holder Disability	F	28.0%	O

Analytical methodology

Normally, discriminant analysis is preferable when the frequencies of the groupings are of relative equal size. In the 2003 election, Newsom won the most precincts by far. Still, I use Alioto's and Gonzalez's precincts as fair groupings, but I omit the Ammiano precincts because three precincts won are not significant to the overall analysis.

(This part is important) When we run a random sample of precincts, the results are slightly different each time. That is to say, the issues that create the best classification results, for that sample, change a bit due to precinct randomization for the holdout sample. After enough model runs with a random sample, the most important issues emerge, as they best classify the precincts in run after run. So, in order to be analytically fair (instead of just taking a particular run with results that I liked), I ran the stepwise model 20 times, and took the issues that emerged from the model the most frequently.

Once I had the most frequent issues, I entered those into the model altogether – not stepwise, but collectively. This is the result I then report. The important parts are the issues selected and the overall classification percentage.

The stepwise model uses Wilks' lambda as the entering criterion. F value entry is at 3.84, and removal is 2.71. Classification was performed using the already existing group sizes. This means we take into account the pre-existing distribution of precinct winners when we classify the predicted groups, instead of assigning each candidate an equal probability of winning. The reason for this is that in a San Francisco election, we know each candidate does *not* have an equal probability of winning. Certain precincts have very strong tendencies to vote in certain ways. This affects how the precincts are classified.

The holdout sample was randomly selected from all the precincts, and changed for each of the initial 20 runs. I also used a holdout sample on the final analysis published here (see above for explanation of this methodology).

One more note – in the 2003 election, many of the issues correlate with each other pretty strongly. Normally, we try to avoid this in these types of analyses. Although this is not considered a fatal flaw, it tends to upwardly bias the results a little bit.

Results

Table 4 shows the frequency of the issues that were selected from the 20 runs of the stepwise model. The five most frequent issues stand out as the issues that repeatedly best classify the precincts to the correct mayoral selections, with two issues (props I and M) being selected almost every time. Props G, C, and N were also selected over half the time.

Table 4: Frequencies of issues that 20 different random-sample discriminant analyses produced. The top five are used in the subsequent analysis, and are considered the best classification issues for the 2003 Mayoral race.

Issue	Title	Count	Freq
P_YI_03	Child Care for Low Income Families	19	19.0%
P_YM_03	Aggressive Solicitation Ban	19	19.0%
P_YG_03	Rainy Day Fund	13	13.0%
P_YC_03	City Services Auditor	12	12.0%
P_YN_03	Taxi Permit Holder Disability	11	11.0%
P_YE_03	Ethics Reform	7	7.0%
P_YD_03	Small Business Commission	5	5.0%
P_YK_03	Sales Tax for Transportation	5	5.0%
P_YH_03	Police Commission / Office of Citizen Complaints	4	4.0%
P_YL_03	Minimum Wage	4	4.0%
P_YJ_03	Facilities for the Homeless	1	1.0%
P_YA_03	School Bonds	0	0.0%
P_YB_03	Retirement Benefits for Safety Employees	0	0.0%
P_YF_03	Targeted Early Retirement	0	0.0%

These five variables, therefore, were entered into the final discriminant model. Figure 3 displays the canonical function graph, to see the physical separation among the three groups. Table 5 shows the classification table for the analysis.

Figure 3: Chart of precinct values plotted by canonical functions. Notice the strong separation between all three candidates, but especially Newsom and Gonzalez. Alioto precincts intersect more with the other candidates' precincts.

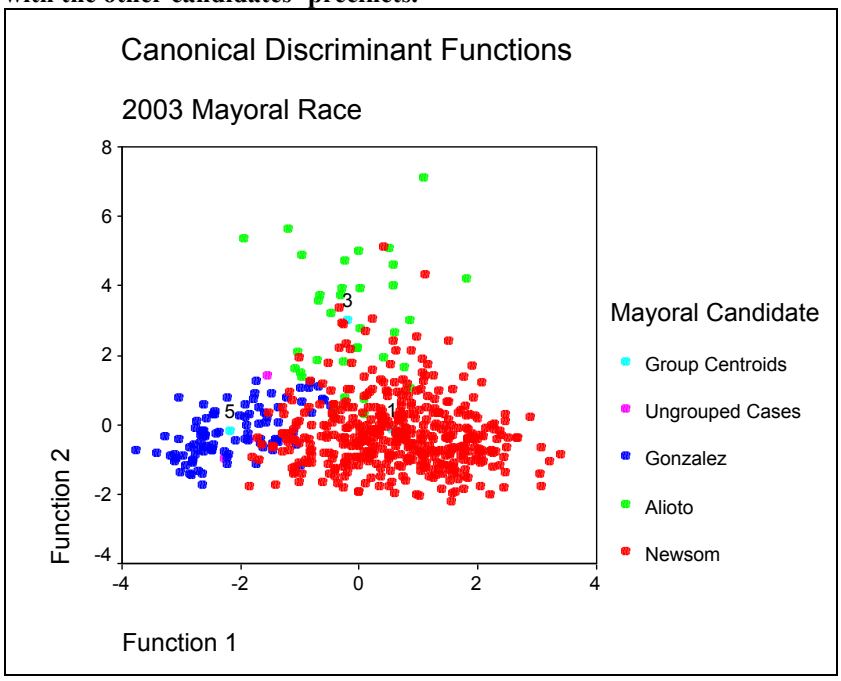


Table 5: Classification results from the 2003 Mayoral model. 92.2% of analyzed samples are predicted correctly, as are 91.9% of the holdout sample.

			Predicted Group Membership			Total
		Actual Mayor	Newsom	Alioto	Gonzalez	
Cases Selected	**Count**	Newsom	215	6	5	226
		Alioto	6	7	0	13
		Gonzalez	6	0	50	56
		Ungrouped cases	0	0	2	2
	%	Newsom	95.1	2.7	2.2	100.0
		Alioto	46.2	53.8	0	100.0
		Gonzalez	10.7	0	89.3	100.0
		Ungrouped cases	0	0	100.0	100.0
Cases Not Selected	**Count**	Newsom	207	2	4	213
(Holdout)		Alioto	9	13	0	22
		Gonzalez	7	0	31	38
		Ungrouped cases	0	0	1	1
	%	Newsom	97.2	.9	1.9	100.0
		Alioto	40.9	59.1	0	100.0
		Gonzalez	18.4	0	81.6	100.0
		Ungrouped cases	0	0	100.0	100.0

a 92.2% of selected original grouped cases correctly classified.
b 91.9% of unselected original grouped cases correctly classified.

Discussion

This model did a very good job discriminating among the three mayoral groups, especially between precincts choosing between Newsom and Gonzalez. For Newsom, 95% (97% in the holdout sample) of the precincts were classified correctly, while 89% (82% in the holdout sample) of the Gonzalez precincts were classified correctly. The model didn't quite fare as well with Alioto precincts (54% in the analyzed sample, 59% in the holdout sample), but that's not uncommon given politically she was in between the moderate Newsom and the progressive Gonzalez. Note all three ungrouped cases, which are precincts Ammiano won, are all placed into the Gonzalez category. Incidentally, when I ran the model the original 20 times, all of the combinations of issues selected classified about 90% of precincts correctly, give or take a couple percentage points.

The issues used to create this classification are telling, especially in their order of significance. The top two issues, by far, were the Child Care Initiative and the Anti-Panhandling Initiative. The Anti-Panhandling Initiative was strongly associated with Newsom, while the Child Care measure was a social welfare issue. The Rainy Day fund and the City Auditor are both good government measures. I think Prop N – Taxi Permits – is more of an artifact of the data. It's also the only issue that's classified by the second canonical function (see earlier example). Because it was so roundly defeated, most people who voted, voted against it. I ran the model without Prop N, and the results didn't really change.

These don't necessarily mean that these were the only important issues in the election, but these were those that collectively matched the election outcome – and could successfully predict the precinct voting patterns of the holdout sample. It gives a rough indication of what the voters were thinking when they chose a candidate. Homelessness, social welfare issues, and good government help us to predict whom the voters will choose. This also gives strong evidence that it was indeed the anti-panhandling issue (and by proxy homelessness) that helped Newsom win in 2003, and voters strongly associated Newsom with this issue.

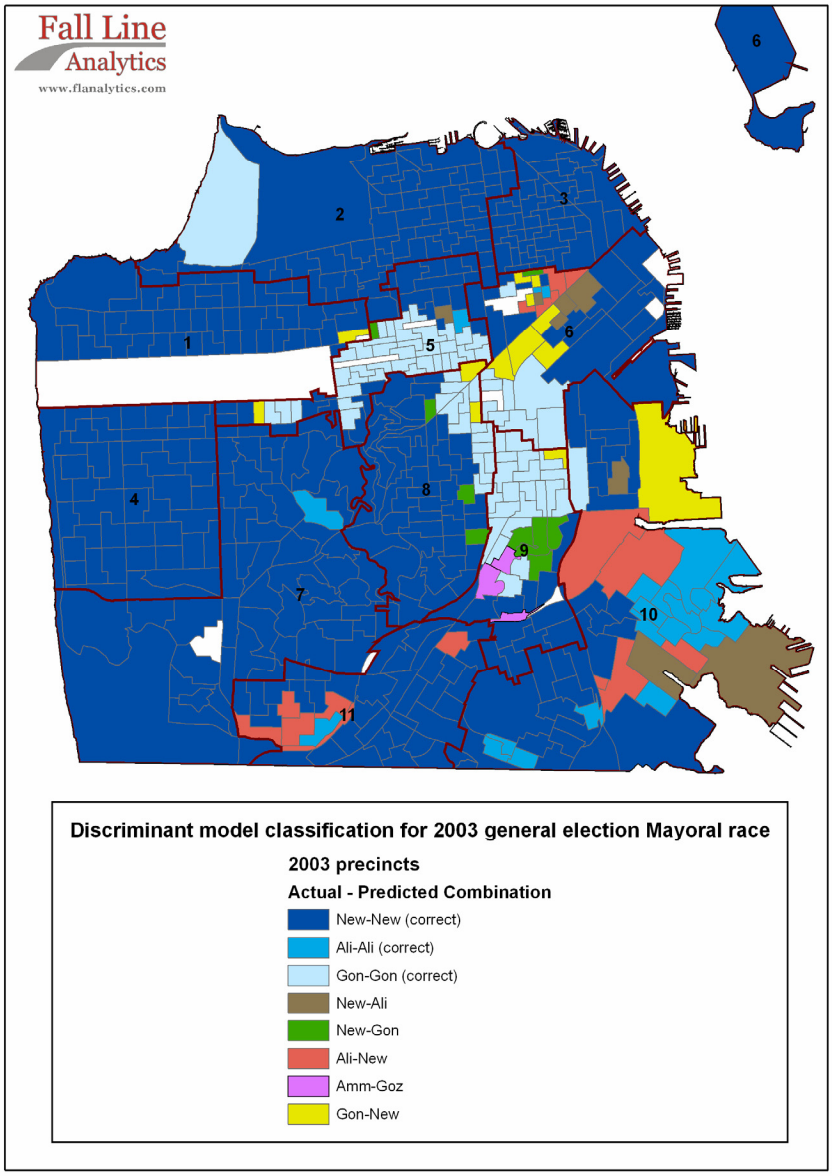
If we look at the issues that did not factor into the analysis, there are two civil service issues at the bottom. Again, it doesn't mean these aren't important, but we really can't use them to predict candidate choices in a Mayoral election.[4] More loosely, this means this issue is not necessarily coupled with a Mayor. Conversely, it's unlikely a Mayor can push this issue too far in his own direction, as opposed to an issue which was associated strongly with the candidate in his election.

Interestingly, the Alioto precincts were mostly split between Alioto and Newsom; none were misclassified to Gonzalez in the analysis sample, and only two in the holdout sample. This would indicate that the Alioto precinct would shade more conservative, and indeed, in the runoff, we did see those voters go primarily to Newsom.

[4] I don't show it here, but when I look at the 2003 DA race, the OCC initiative (Prop H), not important here, places near the top.

Looking at a graphical display of the results, Figure 4 is a map that shows precincts from the 2003 election. The colors represent the match between the actual precinct winner of the election and how they are classified by the model. This gives some indication of where, geographically, this type of discriminant analysis works well and where it doesn't. All the blue shades represent the various "correct" classifications. The model classified most parts of the City correctly, but it didn't do as well in parts of D6, the Outer Mission, D10, and OMI. There areas were on the fringes of support centers for the candidates. In a sense, they're the intersections of ovals drawn around the candidate clusters in the canonical function graphs (see Figure 2). They also indicate battleground areas of the City for a particular election. In the case of the 2003 general election, it was mostly among the progressive candidates, although D10 showed a battle between Newsom and Alioto.

Figure 4: Map of precincts from 2003 Mayoral general election. This shows the actual precinct winner and the predicted precinct winner. All the "correct" classifications are in blue shades.



1999 Mayoral election

Background and analysis

To have something to which to compare the 2003 Mayoral race, I looked at the 1999 Mayoral general election with its fifteen candidates. Table 6 shows the frequency of the precinct winners of that election, and Table 7 displays the 1999 ballot measures I use as independent variables. Note precinct lines and numbers have changed since 1999. In this race, there were three precinct winners: Brown, Ammiano, and Jordan. Brown won the majority of the precincts, but Ammiano also won a sizable share.

For the analysis, I use the same technique as before. 21 stepwise model runs establish the issues that most often successfully classify the groupings. These issues – here, three - were used for the final model.

Table 6: Frequency of precinct choices from the 1999 Mayoral election

Candidate	Precincts won	Precent
Willie Brown	440	64.0%
Tom Ammiano	195	28.4%
Frank Jordan	52	7.6%
Total	687	

Table 7: Ballot initiatives from the November 1999 election used as independent variables

Prop	Title	Result (Pass/Fail)	Percent Voting Yes	B=Bond Issue C=Charter Amendment O=Ordinance
A	Laguna Honda Project	P	73.2%	B
B	Firefighter/Police Retirement Benefits	P	71.4%	C
C	Supervisorial District Boundaries	P	71.8%	C
D	Sick Leave/Vacation Credit Transfers	P	75.8%	C
E	Municipal Transportation Agency	P	61.0%	C
F	ATM Fees	P	66.4%	O
G	Sunshine Ordinance Amendment	P	58.4%	O
H	Downtown Caltrain Station	P	69.3%	O
I	Octavia Boulevard Plan	P	54.3%	O
J	Central Freeway Replacement	F	47.3%	O
K	Campaign Expenditure Limit	P	79.8%	O

Results

Table 8 shows the variables that the model selected in the analysis. Three issues emerge as the strongest classifying issues: Props J, G, and F. Prop J appeared in every model, and Props G and F in nearly all of them. These three issues provided a clear cutoff as to the most useful issues in predicting the 1999 race.

Table 8: Frequencies of issues that 21 different random-sample discriminant analyses produced. The top three issues are used in the subsequent analysis, and are considered the best classification issues for the 1999 Mayoral race.

Issue	Title	Count	Freq
99_p_j	Central Freeway Replacement	21	24.7%
99_p_g	Sunshine Ordinance Amendment	20	23.5%
99_p_f	ATM Fees	18	21.2%
99_p_e	Municipal Transportation Agency	8	9.4%
99_p_a	Laguna Honda Project	4	4.7%
99_p_b	Firefighter/Police Retirement Benefits	3	3.5%
99_p_c	Supervisorial District Boundaries	3	3.5%
99_p_h	Downtown Caltrain Station	3	3.5%
99_p_i	Octavia Boulevard Plan	2	2.4%
99_p_k	Campaign Expenditure Limit	2	2.4%
99_p_d	Sick Leave/Vacation Credit Transfers	1	1.2%

These three issues are plugged into a separate model. Figure 5 displays the canonical function graph, to see the physical separation among the three groups. Table 9 displays the final classification table.

Figure 5: Chart of precinct values plotted by canonical functions. Notice the strong separation between Ammiano and Jordan, while both groups have some intersection with Brown precincts (especially Jordan).

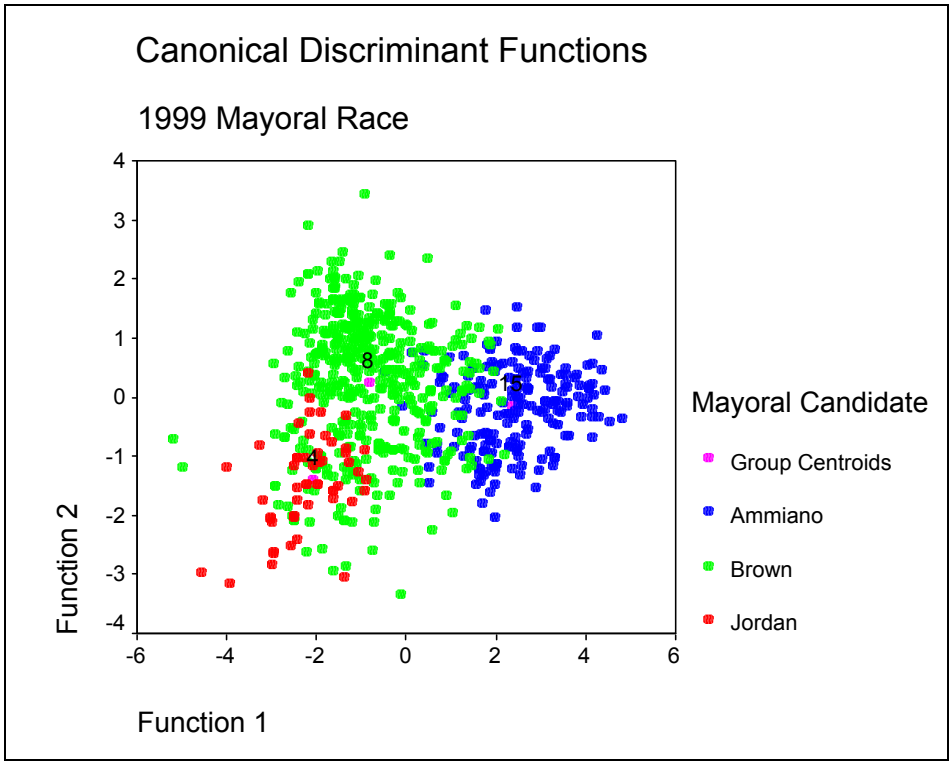


Table 9: Classification results from the 1999 Mayoral model. 85.4% of analyzed samples are predicted correctly, as is 84.7% of the holdout sample. Note no Ammiano precincts were misclassified into Jordan precincts and vice versa.

			Predicted Group Membership			Total
		Actual Mayor	Jordan	Brown	Ammiano	
Cases Selected	**Count**	Jordan	9	12	0	21
		Brown	11	182	14	207
		Ammiano	0	9	77	86
	%	Jordan	42.9	57.1	.0	100.0
		Brown	5.3	87.9	6.8	100.0
		Ammiano	0	10.5	89.5	100.0
Cases Not Selected	**Count**	Jordan	9	22	0	31
(Holdout)		Brown	10	208	15	233
		Ammiano	0	10	99	109
	%	Jordan	29.0	71.0	0	100.0
		Brown	4.3	89.3	6.4	100.0
		Ammiano	0	9.2	90.8	100.0

85.4% of selected original grouped cases correctly classified.
84.7% of unselected original grouped cases correctly classified.

Discussion

This model also did a good job classifying the precincts into their Mayoral choices. Overall, the model predicted 85% of precincts correct for the analyzed sample and also 85% for the holdout sample. 88% of the Brown precincts were classified correctly in the analysis sample, and 89% in the holdout sample. 90% of the Ammiano precincts were predicted correctly in the analyzed sample and 91% in the holdout sample. These are very good values. Again, the third candidate, Frank Jordan, didn't do as well in the model. 43% of his precincts were predicted correctly while only 29% were predicted correctly in the holdout sample.

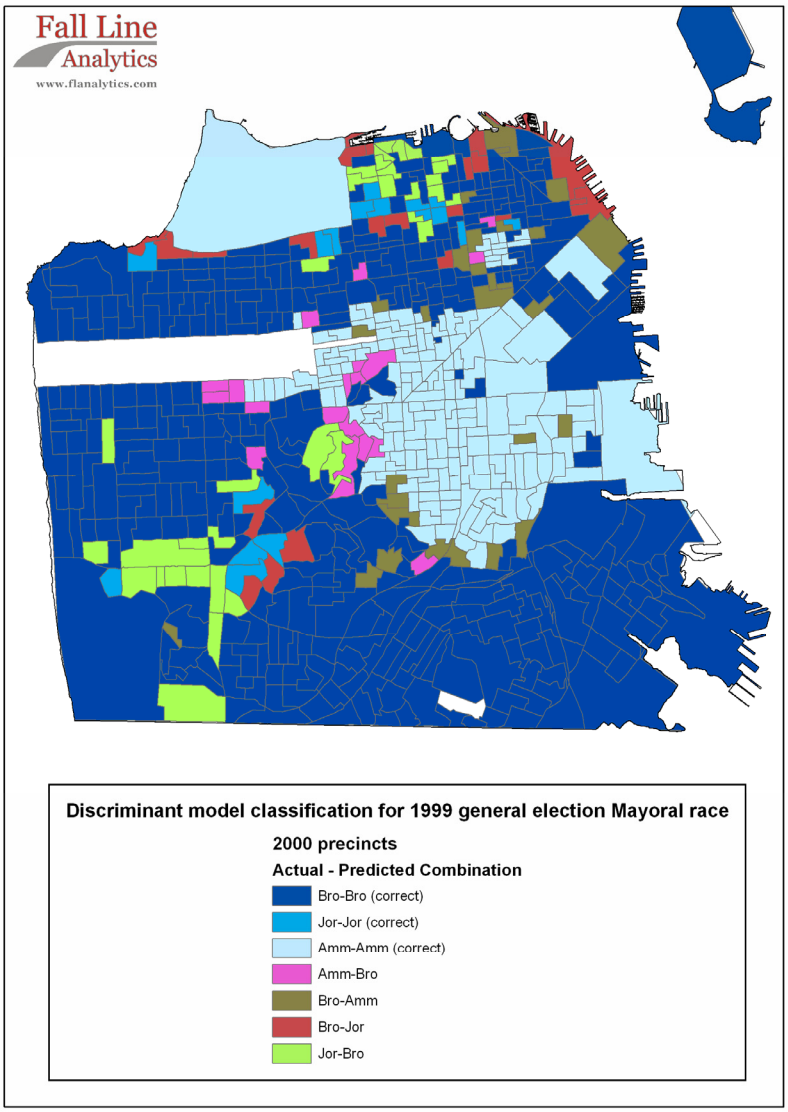
Unlike the 2003 race, the left and the 'middle' candidates (if we can call Brown that) have the best discrimination efforts. Jordan, who ran to the right in this race, has his precincts classified between him and Brown. Of note, no Jordan precincts were classified as Ammiano precincts, in either the analyzed sample or the holdout sample. As with Newsom and Gonzalez, this indicates the two candidates were politically well differentiated.

Examining the issues that classified the precincts, three issues really stood out as the most important. We see first the Central Freeway Replacement – a contentious construction issue, and the Sunshine Ordinance, another good government measure. Then, we have ATM fees. This is a bit harder to interpret, but it could mean government oversight in business matters, which is always relevant to San Francisco voters. Ammiano was pretty strongly identified with the Sunshine Ordinance, and also the effort to stop ATM fees. These patterns don't necessarily show the same pattern as in the 2003 election, but many of the issues were different. Government oversight is prevalent in both races, and is a strong, useful predictor of how people will vote.

Figure 6 shows the map – using 2000 precincts – of the model's classification patterns. It's different than the 2003 race, because the misclassification came mostly from Jordan and Brown. Thus, in a race with a credible conservative candidate, these are the areas that may be battleground areas. We can expect to see the same pattern if someone runs to the right of Newsom in 2007.

All of the blue shades represent the "correct" classification patterns. The most misclassification occurred in parts of the Marina, Chinatown, the newly renamed Barbary Coast, West of Twin Peaks, and Merced again. These areas are those that may have voted against conventional wisdom between the three issues highlighted here and the candidates best associated with those positions. Further research, by looking at correlations between the individual candidates and the vote results of the issues, could bear this out better.

Figure 6: Map of precincts from 1999 Mayoral general election. This shows the actual precinct winner and the predicted precinct winner. All the "correct" classifications are in blue shades.



So what does all this mean?

Discriminant analysis is another way to examine historical election trends, while helping to frame the issues important to future elections. However, unlike regular OLS models this one allows us to classify precincts into groups (for whom they voted). We predict the groupings based on a set of selected issues or other independent variables.

This method doesn't have any definitive answers, but what it does allow us to do best is create linkages between ballot measures and candidates. Many issues are important to voters in a given election, but what we see here are the issues that best match how the precincts choose candidates in a *multi-candidate* race. It also allows us to see what precincts (and voters) defy "conventional wisdom", by choosing different candidates than would be expected given their vote on the selected issues each model produces.

We see that in 2003, a set of good government and social welfare measures best predict precinct voting patterns, in a race primarily of one moderate and many more liberal candidates. In 1999, a good government, transportation, and a business issue best predict that race – which was mainly between a left, middle, and more conservative candidate. These political patterns repeat themselves in San Francisco elections. So, in future races, if we see a government oversight issue or a social welfare issue, we will know to use them to predict whom voters will choose.

Once it is known what issues match candidate voting trends in certain races, candidates and their teams know how to move voters to "their" side of those respective issues. A subsequent analysis to this one is to look at how the precincts voted on the selected issues in the discriminant analysis, but in a sense it doesn't really matter, since all the candidates in a race will try to sell the voters on their vision of the issues that seem to be important. For instance, candidates know "good government" is important in Mayoral (and other citywide races), so it's up to them to sell themselves on this aspect of their platform, tailoring their message – whatever it may be – to the specific audience. Likely, this will be helped by polling on such issues to see where the electorate, depending on demographics, stands. This kind of discriminant analysis is an obvious precursor to push polling a well, once key issues are identified.

Discriminant analysis is just one way we can look at multi-candidate races, but since with RCV we'll no longer have a one or the other scenario, it helps to filter issues that are important a complicated race. I also looked at the 2004 school board race, and discriminant analysis did a pretty good job, classifying about 60% of the precincts correctly, based on five precinct winners.[5] One big problem was that the model could not discriminate between Eric Mar and Norman Yee, because of the powerful Asian identity vote in the Richmond and Sunset, even among more conservative voters. For distinguishing among Hiles, Wynns, and Sanchez, the model did quite well. The fact that the model itself had so much trouble determining Mar vs. Yee precincts speaks to the power of the Asian identity vote, a topic of another paper.

[5] There were seven winners, but two candidates only won a couple precincts.