Disaggregation of Precinct Voting Results to Census Geography

Kenneth F. McCue*

California Institute of Technology

January 3, 2008

*Research Scientist, Department of Biology, California Institute of Technology.

The method of allocating (or breaking down, or disaggregating) votes to census geography is done by assigning an individual probability to each voter who voted in the election and aggregating these probabilities to the census geography (the block, at the lowest level), to obtain a total vote for the block. The following is done to do this.

- Partition voters into groups. There are technical reasons why these groups should be "homogeneous", which is a statistical concept for how well the voting behavior fits a statistical law (in this case the multinomial distribution–see below). In general, in California, we use Democrat/Republican/Independent as the groups, with provisions for the variation of minorities and economic status (see next item). If there are P precincts, then X_{gi} is the number of voters in group g, $g = 1, \ldots, G$ in precinct i, $i = 1, \ldots, P$.

- Estimate the overall voting probabilities in the state/district. The probability of a member of group g voting for a candidate in precinct i is denoted by p_{gi}, and if v is the votes for that candidate, an equation similar to the following is used to estimated p_{gi}:

$$\text{argmin} \sum_{i=1}^{P} (v_i - \sum_g X_g p_{gi})^2,$$

  where the argmin is taken over the p_{gi} (in actual fact a more complicated optimization is used but it will produce results close to this equation).

  There are too many p_{gi} to estimate each individual p_{gi}, so an average p_{gi} is estimated instead. The average used here is of the form $p_{gi} = p_g(z_i, \tau)$, where τ is a conformable vector to a set of characteristics z_i for precinct i (which is how one would include variations in minorities or economic status). Note that for a district race, each estimation should be done for each district by itself (or suitably modified, with the use of dummy variables), whereas statewide races can usually be estimated with all precincts.

- Adjust estimated probabilities to the precinct. The method of estimation described above does not ensure that the estimated precinct totals equal the actual precinct totals on a precinct by precinct basis (though the overall estimation is done so that the estimated district/state totals do match the overall district/state totals). The standard statistical methodology for adjusting estimations is followed, where a $\tilde{p}$ is estimated such that $v_i = \sum_g X_{gi}\tilde{p}_{gi}$. This $\tilde{p}_i$ (which is a G by 1 vector) is estimated by

$$\tilde{p}_i = \mathrm{E}[\hat{p}_i|v_i] \approx p(z_i, \hat{\tau}) + \mathrm{Cov}[K_i|v_i][\mathrm{Var}[v_i]]^{-1}[v_i - \sum_g X_{gi}p(z_i, \hat{\tau})],$$

where the K_i is the response count of the groups for the i^{th} precinct, with the $(K_{1i}, \ldots, K_{Gi})$ being distributed multinomially (some distributional assumptions must be made in order to estimated the covariance of K_i with v_i).

These $\tilde{p}_i$'s allocate all of the votes for a candidate in each precinct to the individual voters in that precinct. Thus the sums of these votes by census block by all census blocks will equal the sum of the vote for the candidate. For a census block split between precincts (say precinct A and B), those voters in precinct A will have an assigned voting propensity of $\tilde{p}_A$ and those in precinct B will have an assigned voting propensity of $\tilde{p}_B$.

As an example, consider a two-precinct district as described in Table I (in this table, quantities are suppressed for units in which they do not make sense). The two precincts in the district, A and B, each have a block wholely contained in the precinct (1001 for A and 1003 for B) and share a block, 1002. There are two groups in the electorate, X_1 and X_2. Through geocoding, it is known how many voters of each group is within each block and what precinct each voter is in, which is known through the registered voter rolls. The number of votes cast for a candidate (v) is known at the precinct level.

Estimated quantities are an overall probability of each group to vote for the candidate ($\hat{p}_1$ for group X_1 and $\hat{p}_2$ for group X_2) and then the adjusted probabilities ($\tilde{p}_1$ for group X_1 and $\tilde{p}_2$ for group X_2) chosen by the method described above in "Adjust estimated probabilities to the precinct". This then gives, for each precinct/block combination, an estimated vote by group ($\hat{v}_1$ for X_1 and $\hat{v}_2$ for X_2). Finally, the sum of $\hat{v}_1$ and $\hat{v}_2$ can be calculated,

Table 1: Example of Disaggregation in a Two-precinct District

Precinct	v	$\hat{v}$	Block	X_1	$\hat{p}_1$	$\tilde{p}_1$	$\hat{v}_1$	X_2	$\hat{p}_2$	$\tilde{p}_2$	$\hat{v}_2$
A		150.0	1001	100		.2	20.0	200		.65	130.0
A		80.0	1002	50		.2	15.0	100		.65	65.0
A	230	230.0		150		.2	35.0	300		.65	195.0
B		56.5	1002	100		.34	34.0	50		.45	22.5
B		73.5	1003	150		.34	51.0	50		.45	22.5
B	130	130.0		250		.34	85.0	100		.45	45.0
District	360	360.0		400	.3		120.0	400	.6		240.0

providing a $\hat{v}$ for each precinct/block combination and a check that, indeed, the allocated (or disaggregated) vote does equal the actual vote for each precinct.

For a block split between two (or more) precincts, these estimated totals can be added up across precinct/block combinations which contain that block to obtain totals for the block. For example, in block 1002 in Table I, there are 49 votes from group 1 for the candidate, 87.5 from group 2 for the candidate, for a total of 146.5 votes for the candidate.

This note does not address the following:

- Assigning voters to census geography who could not be assigned by geocoding
- Constraints on votes between elections
- Constraints on voters across elections
- Estimating when only registered voters are available (as opposed to voters who voted)
- Estimation of standard errors of the estimates
- Pratical matters of data preparation and inclusion into estimations

- Use of census block/individual data to modify the $\tilde{p}_i$
- The relationship between these estimates and estimates of racially polarized voting