Bayesian Population Interpolation and Lasso-Based Target Selection in Survey Weighting*

Devin Caughey
MIT

Mallory Wang
Medallia

November 4, 2014

Abstract

We propose solutions to two important problems that have received relatively little attention in the field of survey weighting: the construction of population targets in the face of irregularly missing data, and the optimal selection of weighting targets from the set of possible auxiliary variables. Our solution to the first problem relies on a dynamic Bayesian population-interpolation model that allows subpopulation estimates in a given year to be informed by data from other years. To address the second, we formulate the problem of target selection as one of variable subset selection, for which we propose a lasso-based solution. We demonstrate the usefulness of these techniques by using them to generate weights for quota-sampled opinion polls from the early days of survey research. Given the declining response rates, rising use of non-probability samples, and growth in potential sources of auxiliary information in modern-day polling, these methods have wide potential application in contemporary survey research as well.

*We thank Adam Berinsky, Eric Schickler, Jasjeet Sekhon, and for their help with this paper and for the use of their data. We also received valuable feedback from Danny Hidalgo, Teppei Yamamoto, Chris Warshaw, Michael Betancourt, Bob Carpenter, Luc Coffeng, and attendees at PolMeth 2014, especially James Honaker.

Contents

1 Introduction

Survey weighting—that is, the post-hoc construction of weights for use in the analysis of survey data—has become increasingly important in recent years. Weighting samples to match known population targets can reduce bias caused by unit non-response and also, if well-chosen, improve the precision of estimates (Bethlehem, 2002; Little and Vartivarian, 2005). As survey response rates have plummeted and polling organizations have increasingly abandoned probability samples in favor of internet-based opt-in surveys, weighting samples to account for differential response probabilities has become increasing crucial. Ironically, the challenges of modern survey research mirror those posed by early public opinion polls of the 1930s and 1940s, which relied on non-probability quota sampling. Constructing survey weights has been a core component of recent work by Berinsky, Schickler, and their colleagues to improve access to and analysis of these early polls (Berinsky, 2006; Berinsky et al., 2011). Finally, weighting is also central to recent advances in small-area estimation of public opinion, such as multilevel regression and poststratification (MRP; see Park, Gelman and Bafumi, 2004).

Much of the recent progress in the methodology survey weighting has come under the rubric of calibration estimation (Deville and Särndal, 1992; Särndal, 2007). Calibration estimation defines weight construction as a problem of finding the set of weights that deviate as little as possible from prior sampling weights while matching specified moments of the sample and population distributions (cf. Hainmueller, 2012, on weighting for causal inference). Calibration subsumes such well-known weighting techniques as raking and poststratification under a common framework, with different methods corresponding to different choices of target moments (joint or marginal) and distance metric.

Two problems in survey weighting, however, have received relatively little attention. The first problem concerns the accuracy and completeness of the population targets themselves. Data on the population distribution of auxiliary variables must often be compiled from multiple sources, which may contain modest inconsistencies due to sampling or measurement

error.[1] Even more commonly, population data are available only for particular points in time (e.g, census years), requiring either an assumption of time-invariant population distributions or the estimation of population proportions in years with no data. Interpolation between years with data is relatively straightforward when the structure of population data does not vary across time, but it becomes much more complex when different data are available at different points in time. These difficulties have typically led researchers to either ignore population dynamics or use only target data that are available in the same form across time.

The second problem we address in this paper is the selection of population moments to use as targets for the weights. If a smoothing method such as in MRP is used to estimate opinion, the estimates can be weighted to exactly match the multivariate population distribution. But raw survey samples themselves are typically too sparse to weight them to match the full joint population distribution. Due to sparseness in the survey data, it is often impossible to construct weights that match the joint population distribution of all available auxiliary variables. Even if it is possible, matching the full joint distribution may not be advisable given the increase in variance that it entails. Thus weighting usually requires the choice of which aspects of the population distribution to match and which not to match. Notwithstanding some useful guidance available on the subject (e.g., Särndal and Lundstrom, 2005; Bethlehem, Cobben and Schouten, 2011), the complexity of the decision problem means that it is typically addressed with simple, ad hoc methods.

This paper proposes solutions to both of these problems. For the construction of population targets, we propose a dynamic Bayesian approach in which a multinomial sampling model is used to estimate the joint population distribution based on partially observed and possibly inconsistent marginal distributions. The population distribution is allowed to evolve dynamically over time according to a Dirichlet random walk (Grunwald, Raftery and Guttorp, 1993), thus providing interpolated estimates between data points. The problem and

[1] An *auxiliary variable* is a variable observed in the sample whose population distribution is known or estimated with greater precision than the sample. Throughout this paper, we assume auxiliary variables to be categorical.

solution is similar to those of ecological inference, particularly the dynamic approach of Quinn (2004).

For the problem of selecting weighting variables, we draw on the analogy of variable subset selection for regression models. The approach we propose uses a multivariate version of the lasso (Tibshirani, 1996) to rank-order weighting specifications in terms of their ability to predict response probabilities and important outcome variables. These techniques allow us to construct much more information-rich population targets than we could otherwise, and also to optimize our choice of targets to use in the construction of weights.

The remainder of this paper is organized as follows. The next section examines the issue of population interpolation in more detail, illustrating the problem with the example of phone ownership by race and region between 1930 and 1960. It then derives and explains the interpolation model we propose as a solution. The subsequent section addresses the problem of target selection and our lasso-based approach. The penultimate section applies these two methods of population interpolation and target selection to the problem of constructing survey weights for quota-sampled polls from 1940. The final section concludes.

2 Bayesian Population Interpolation

In this section, we motivate and describe our population interpolation model. We begin with a motivating example that we will use to illustrate the problem and our proposed solution. We then briefly discuss related problems and methods before deriving our own model.

2.1 Motivating Example: Phone Ownership by Race and Region

We begin by describing a simplified version of the problem that motivated us to develop our approach to population interpolation. We were interested in constructing survey weights to mitigate biases in quota-sampled opinion polls fielded between 1936 and 1952.[2] The poll

[2]The data for these polls were cleaned and standardized under the direction of Adam Berinsky, Eric Schickler, and Jasjeet Sekhon. This work was funded by two grants from National Science Foundation, Political Science

samples exhibit class, racial, and regional discrepancies from the population. Higher-SES respondents are overrepresented in the sample, as are whites and non-Southerners. In fact, most polls entirely exclude Southern blacks (then disenfranchised). Because Southern blacks are not part of the sampling frame, we must exclude them from the target population, which we redefine to be all Americans other than Southern blacks.

Consider a quota-sampled poll fielded in 1940. From the 1940 U.S. Census, we know the joint distribution of *Region* (South vs. North) and *Race* (black vs. white), but these variables' joint distribution with *Phone Ownership* (phone vs. no phone) is not available until the 1960 Census, when overall phone ownership was much more common. Using AT&T corporate records, however, we can determine the proportion of phone owners at the regional level in 1940. Table 1 represents the observed and unobserved aspects of the joint distribution of *Region*, *Race*, and *Phone Ownership* as a three-dimensional array.

	South			North		
	Phone	No Phone		Phone	No Phone	
Black	$\pi_{(111)}$	$\pi_{(112)}$	$p_{(11\bullet)}$	$\pi_{(211)}$	$\pi_{(212)}$	$p_{(21\bullet)}$
White	$\pi_{(121)}$	$\pi_{(122)}$	$p_{(12\bullet)}$	$\pi_{(221)}$	$\pi_{(222)}$	$p_{(22\bullet)}$
	$p_{(1\bullet1)}$	$p_{(1\bullet2)}$		$p_{(2\bullet1)}$	$p_{(2\bullet2)}$	

Table 1: Three-dimensional array of the population distribution of *Phone Ownership* by *Race* by *Region*. Unobserved cell proportions are represented by π and observed marginal proportions by p. Gray cells (Southern blacks) are not part of the target population.

If the target population were all Americans, we could use raking to produce weights that match the distribution of *Region* $\times$ *Race* as well as of *Region* $\times$ *Phone Ownership*. These weights would accurately incorporate all known information about the population.[3] Because the target population excludes Southern blacks, however, this approach would only be valid if *Phone Ownership* and *Race* were uncorrelated in the South, in which case the

Program: SES-0550431 (Berinsky and Schickler, "Collaborative Research: The American Mass Public in the 1930s and 1940s," 2006–2010) and SES-1155143 (Berinsky, Schickler, and Sekhon, "Collaborative Research: The American Mass Public in the Early Cold War Years," 2012–2014). For further details on the project and the data, see Berinsky (2006) and Berinsky et al. (2011).

[3]Although this is rarely noted in the weighting literature, the loss function that raking implicitly minimizes is the Kullback-Leibler cross-entropy, making raking weights optimal in an information-theoretic sense (Wittenberg, 2009).

required interior cells could be obtained by the equations $\pi_{(121)} = p_{(12\bullet)} \times p_{(1\bullet 1)}$ and $\pi_{(122)} = p_{(12\bullet)} \times p_{(1\bullet 2)}$. Unfortunately, the 1960 Census, not to mention historical intuition, indicates that Southern blacks were much less likely than Southern whites to own phones. We are seemingly stuck with two bad choices: either make an independence assumption between *Phone Ownership* and *Race* that we know to be very wrong, or do nothing to mitigate the biases caused by the upper-class bias of the poll sample.

An alternative is to find a way to represent all the information we know about the population while making the weakest assumptions we can. In the broadest sense, we want to allow the information we know from the 1960 Census—that whites are overrepresented among phone owners and blacks underrepresented—to inform our estimates of race-specific phone ownership rates in 1940. Moreover, we would like to make such estimates for all other years in the 1930–60 period, even years where we observe no data at all.

2.2 Related Problems and Methods

Since the problem described in Section 2.1 is similar to issues in demography and ecological inference, we briefly review related work in these fields before describing our own approach.

Demographers have long grappled with how best to estimate population totals and composition from census and other data. Particularly relevant for our purposes are interpolation methods for estimating subnational population counts between censuses (Swanson and Tayman, 2012). Classical methods in this area were typically deterministic, but most recent developments in demographic interpolation and forecasting have used a Bayesian approach (e.g., Raftery et al., 2012; Bryant and Graham, 2013; Wheldon et al., 2013). The models in these works are typically quite complex and incorporate demographic accounting identities, fertility information, and other application-specific knowledge into the model itself. But the models' basic structure typically involves a sampling model for the data, a transition model for the temporal evolution of the population counts or proportions, and a set of equations specifying the logical relationships among demographic quantities.

Another natural reference point for our work, this one closer to political science, is the problem of ecological inference—that is, inference about individuals from aggregate data (Freedman, 2001). Many ecological inference problems, such as the example of a 2×2 table of voter registration by race discussed by King, Rosen and Tanner (2004) and many others, have a very similar if not identical structure to the problem of estimating interior cells from marginal totals. As Wakefield (2004) notes, such inferences are unreliable in the absence of supplementary data or prior information, and scholars in this field too have found Bayesian models an effective way to incorporating such information.

Many ecological inference models bring in such supplementary information via a hierarchical model that borrows strength from observably similar cross-sectional units (e.g., King, 1997). As an alternative, Quinn (2004) proposes to borrow strength across time instead. He argues that temporal dependence in the interior cell proportions is not a "statistical nuisance" but rather "an important piece of background knowledge" (Quinn, 2004, 207). Quinn incorporates this information via a local-level dynamic linear model (DLM) for the logits of the cell proportions, which shrinks the cell estimates towards the estimates from adjacent periods.[4] He applies this to a dynamic version of the canonical ecological inference problem of estimating voter registration by race.

Quinn (2004) provides a very useful template for our own work, but it is worth noting the differences in the problems we respectively address. First, in Quinn's application the data are observed in the same form in every time period. By contrast, as illustrated by our motivating example, we focus on situations in which the available information differs by period and many periods have multiple data sources or lack data entirely. Relatedly, in some years we know certain variables' joint distribution, whereas only data on marginal proportions are available in Quinn (2004). Finally, our problem is higher-dimensional than the 2×2 tables typical of ecological inference, involving as many as 9 categorical variables and 37,000 cells in each of 30 years.

[4]Transforming the proportions to the logit scale allows the use of Gaussian DLMs that would otherwise violate the unit-interval support of the proportions (Cargnoni, Muller and West, 1997).

2.3 A Bayesian Interpolation Model

In this section, we derive a Bayesian model that generates year-specific subpopulation estimates informed by all the auxiliary data we have at our disposal. We begin with notation. Let $v \in \{1, \ldots, V\}$ index auxiliary variables, each of which is assumed to be categorical with L_v levels indexed by l. Let $c \in \{1, \ldots, C\}$ index the subpopulation cells defined by the auxiliary variables, where $C = \prod_v^V L_v$ if the variables are non-nested.[5] The estimands of interest are the cell population proportions $\boldsymbol{\pi}_t = (\pi_{t1}, \cdots \pi_{tc}, \cdots, \pi_{tC})'$ in each time period $t \in \{1, \ldots, T\}$ for which survey weights are required.

In each period t, population data are available on the joint distribution of M_t subsets of auxiliary variables, where $M_t = 0$ if no data are available. Each variable subset $m \in \{1, \ldots, M_t\}$ contains $V_{tm} \leq V$ variables, whose levels define $G_{tm} = \prod_w^{V_{tm}} L_w$ groups, each composed of $H_{tmg} \geq 1$ cells. The population data for each variable set m in period t consist of the G_{tm}-simplex of proportions $\boldsymbol{p}_{tm} = (p_{tm1}, \cdots p_{tmg}, \cdots, p_{tmG_{tm}})'$, which may contain measurement and/or sampling error (cf. Deville, 2000; West and Little, 2013). Each group proportion p_{tmg} is thus considered a noisy estimate of group g's true population proportion ϕ_{tmg}, which is the sum of the proportions π_{tc} of the cells that compose group g.

For further intuition, consider the example described in Section 2.1. This example involves $V = 3$ auxiliary variables (*Region*, *Race*, and *Phone Ownership*), each with $L_v = 2$ levels. Of interest are the population proportions π_{tc} of $C = 2 \times 2 \times 2 = 8$ cells in each of $T = 31$ years (1930–60). In 1940 ($t = 11$), data on the joint distribution of $M_{11} = 2$ variable subsets are available, $\{$*Region*, *Race*$\}$ and $\{$*Region*, *Phone Ownership*$\}$, each with $V_{11,m} = 2$, $H_{11,mg} = 2$, and $G_{11,m} = 4$. We thus observe $M_{11} = 2$ vectors of group proportions $\boldsymbol{p}_{11,m}$, which are estimates of $\boldsymbol{\phi}_{tm}$ and which correspond to the marginal proportions in Table 1 (though with slightly different notation). In 1960 ($t = 31$), only $M_{31} = 1$ variable subset is available: $\{$*Region*, *Race*, *Phone Ownership*$\}$. For the sake of this example, all other years

[5] An example of nested variables would be *State* and *Region*. We use the term *cells* to refer to the subpopulations whose proportions are of ultimate interest, and we use *groups* to refer to aggregations of one or more cells. Only data on group population proportions are observed.

lack data, so $M_t = 0 \; \forall t \notin \{11,\ 31\}$.

Returning to the model exposition, recall that our goal is to use the information contained in the observed group proportions $\boldsymbol{p}_{tm}$ to make inferences about the true cell proportions π_{tc}. The Bayesian approach we pursue has two primary components: an observation model linking the parameters π_{tc} and $\boldsymbol{\phi}_{tm}$ to the data $\boldsymbol{p}_{tm}$, and a transition model specifying how the π_{tc} evolve over time. The addition of prior distributions over the parameters yields a complete Bayesian model.

As noted above, the observed group proportions $\boldsymbol{p}_{tm}$ are likely to deviate from the true proportions $\boldsymbol{\phi}_{tm}$ as a result of measurement and sampling error.[6] Moreover, it is possible for the observed proportions to contain redundant information, which may not be perfectly consistent if they come from different data sources. We represent the stochastic relationship between the observed and latent proportions using a multinomial observation model, which requires that $\boldsymbol{p}_{tm}$ be converted to counts. We do this by multiplying $\boldsymbol{p}_{tm}$ by the "sample size" n_{samp}, leading to

$$n_{samp}\boldsymbol{p}_{tm} \sim \text{Multinomial}(\boldsymbol{\phi}_{tm},\ n_{samp}). \tag{1}$$

The expected value of p_{tmg} is ϕ_{tmg}, with the sample size n_{samp} (specified by the analyst) determining the precision of the sampling distribution.

Recall from above that each group proportion ϕ_{tmg} is the sum of the proportions of the H_{tmg} cells that compose it. Let $\boldsymbol{A}_{tm}$ be an $G_{tm} \times C$ indicator matrix of zeros and ones, where a 1 in row g and column c indicates that group g contains cell c. The relationship between $\boldsymbol{\phi}_{tm}$ and $\boldsymbol{\pi}_t$ is compactly described with the equation

$$\boldsymbol{\phi}_{tm} = \boldsymbol{A}_{tm}\boldsymbol{\pi}_t \tag{2}$$

because pre-multiplying $\boldsymbol{\pi}_t$ by $\boldsymbol{A}_{tm}$ sums the cells within each group. For instance, in our

[6]For instance, in our running example, the data on *Region* and *Race* are from 1% samples from the complete census and thus contain sampling error. Note also that information on *Region* appears twice, once from the census samples and AT&T records, and that these different data sources need not necessarily be perfectly consistent.

running example, if the proportions $\boldsymbol{\pi}_t$ are ordered lexicographically,[7] then the indicator matrix for *Region* $\times$ *Race* is

$$\boldsymbol{A}_{tm} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \tag{3}$$

and $\boldsymbol{A}_{tm}\boldsymbol{\pi}_t = (\pi_{t(111)} + \pi_{t(112)},\ \pi_{t(121)} + \pi_{t(122)},\ \pi_{t(211)} + \pi_{t(212)},\ \pi_{t(221)} + \pi_{t(222)})'$. $\boldsymbol{A}_{tm}\boldsymbol{\pi}_t$ is thus essentially a vector of group-specific intercepts.

Substituting (2) into (1) leads to the following observation model:

$$n_{samp}\boldsymbol{p}_{tm} \sim \text{Multinomial}(\boldsymbol{A}_{tm}\boldsymbol{\pi}_t,\ n_{samp}). \tag{4}$$

The model defined by (4) does not distinguish among cells in the same group, so without further information the posterior distributions of their distributions will be equal. Thus, given uninformative priors, a single time period, and one set of auxiliary variables, the posterior mean for each element of $\boldsymbol{\pi}_t$ will be the same as if we had weighted $\boldsymbol{\pi}_t$ with poststratification weights derived from $\boldsymbol{p}_{tm}$.[8] The only difference would be the uncertainty around the estimates, which is determined by n_{samp}.

One of the advantages of a Bayesian approach, however, is that it makes it easy to integrate multiple sources of information. One potential source of additional information is the availability of population data on multiple sets of auxiliary variables in the same time period. If data on multiple variables are available, $M_t > 1$ observation models can be specified, and as with raking on multiple sets of marginal distributions, the posterior

[7] That is, $\pi_{(111)},\ \pi_{(112)},\ \ldots, \pi_{(221)},\ \pi_{(222)}$, where the parenthetical subscripts correspond to the notation used in Table 1.

[8] The poststratification weight of each cell c in group g would be $\boldsymbol{p}_{tmg}/H_{tmg}$. In general, poststratification weights are equal within groups as long as the pre-adjustment "reference weights" are equal within groups, which is the case here.

distribution of $\boldsymbol{\pi}_t$ will be informed by all of the data. Note that because the observation model allows for stochastic discrepancies between $\boldsymbol{p}_{tm}$ and $\boldsymbol{\phi}_{tm}$, different data sources need not be perfectly consistent with one another (e.g., if they were derived from separate samples from the same population).

A second potential source of information regarding $\boldsymbol{\pi}_t$ is data from time periods other than t. If cell proportions can be regarded as constant over time, the time index can be dropped and information from different periods incorporated in the same manner as data on different variables in the same period. A more realistic approach, however, is to specify a dynamic model for the evolution of $\boldsymbol{\pi}_t$ over time. Since population proportions lack a long-term mean or trend, the most plausible dynamic model in this context is a simple local-level DLM, where the value in each period serves as the prior expected value for the subsequent period. The conventional Gaussian DLM is inappropriate in this case, however, because it does not respect the constraints on the support of the proportion vector $\boldsymbol{\pi}_t$. Cargnoni, Muller and West (1997) address this problem by applying a logistic or similar transformation to the proportions so that the support of the transformed values is unbounded. We instead take an approach similar to Grunwald, Raftery and Guttorp (1993) and model the evolution of the proportions directly, using a Dirichlet distribution.

The Dirichlet would typically be parameterized in terms of a C-vector $\boldsymbol{\alpha}_t$ of "prior counts" for the C cells. In this application, however, it is convenient instead to write $\boldsymbol{\alpha}_t$ as the product of the prior expected values ($\boldsymbol{\pi}_{t-1}$) and the "prior sample size" $n_{transition} = \sum_c^C \alpha_{tc}$:

$$\boldsymbol{\pi}_t \sim \text{Dirichlet}(\boldsymbol{\pi}_{t-1} n_{transition}). \tag{5}$$

As above, the prior sample size $n_{transition}$ is set by the analyst and determines the innovation precision. In periods with no data, $\boldsymbol{\pi}_t$ will be interpolated with values informed directly by the immediately adjacent periods (Quinn, 2004, 210) and, indirectly, by all previous and

Year	Available Data
1930	*Race* × *Region*
	Phone Ownership
1935	*Phone Ownership*
1937	*Phone Ownership* × *Region*
1940	*Race* × *Region*
	Phone Ownership × *Region*
1945	*Phone Ownership* × *Region*
1950	*Race* × *Region*
1960	*Race* × *Region* × *Phone Ownership*

Table 2: Population Data for Illustrative Example

subsequent estimates. We also specify the following prior for the first period:

$$\boldsymbol{\pi}_1 \sim \text{Dirichlet}(\boldsymbol{\pi}_0 n_0). \tag{6}$$

The values for π_0 may be selected by raking or postratifying a C-vector of 1's to match a subset of available population targets. To facilitate estimation of the model, it may be advisable to choose a value of n_0 that implies a diffuse but proper prior for $\boldsymbol{\pi}_1$.

2.4 Interpolation of Phone Ownership by Race and Region

To illustrate this model, we first apply it to our running example of phone ownership by race and region between 1930 and 1960. Recall that our dilemma is that we would like to weight the polls to match the rate of phone ownership in the target population (U.S. population minus Southern blacks), but to do so we need phone ownership rates by race and region. Even if the target population included Southern blacks, it would still be better to incorporate what we know about racial differentials in phone ownership. This is exactly what our interpolation model enables us to do: incorporate all available information, including information outside the time period of real interest (1936–52), into our estimates of the target population.

Table 2 details the data on the U.S. population available in each year. We set $n_{transition} =$

$1,000,000$ and $n_{samp} = 100,000,000$. The large value of $n_{transition}$ implies a belief that the yearly fluctuations in the population proportions are relatively small, and the even larger value of n_{samp} implies a belief that measurement error in the targets is smaller than year-to-year variation. We estimate the model using the Bayesian simulation Stan, as called from R (Stan Development Team, 2013; R Core Team, 2014).[9]

Estimating the model with these data generates $C = 8$ estimated proportions in each of $T = 30$ years. Figure 1 plots the implied phone-ownership percentages by race and region.[10] As the figure shows, blacks were substantially less likely than whites to own a phone, especially in the South. Even though the racial phone-ownership disparities are only contained in the 1960 data, the model propagates this information backwards in time. The model estimates imply that in 1940, for example, 8% of blacks in the South owned a phone, compared to 23% of whites. These figures are consistent with estimates derived from external data, which suggest that 5–9% of Southern blacks owned phone in 1940. It is worth emphasizing that without the 1960 data, there would be no information to distinguish blacks and whites, whose phone-ownership rates would thus be estimated to be equal within region.

From the estimated population proportions $\hat{\pi}_{tc}$, we can also generate weighting targets for the U.S. population minus Southern blacks. Under this definition of the target population, the estimated 1940 phone-ownership rate in the South is 23%, compared to 19% when Southern blacks are included. Including the 1960 data thus makes a substantial difference in the targets as well as in any survey weights calculated from them.

The foregoing example is in several respects typical of the applications for which our model is intended. First, population data are not available in most years for which weights are desired. Second, even when population data are available, they differ in the aspects of the population distribution they measure. Sometimes marginal distributions are available

[9]Stan is a C++ library that implements the No-U-Turn sampler (Hoffman and Gelman, Forthcoming), a variant of Hamiltonian Monte Carlo. We computed 1,000 samples from the posterior distribution, discarding the first 500. We use the mean of the posterior distribution as our point estimator.

[10]The general decline in phone ownership between 1930 and 1935 is almost certainly a consequence of the Great Depression.

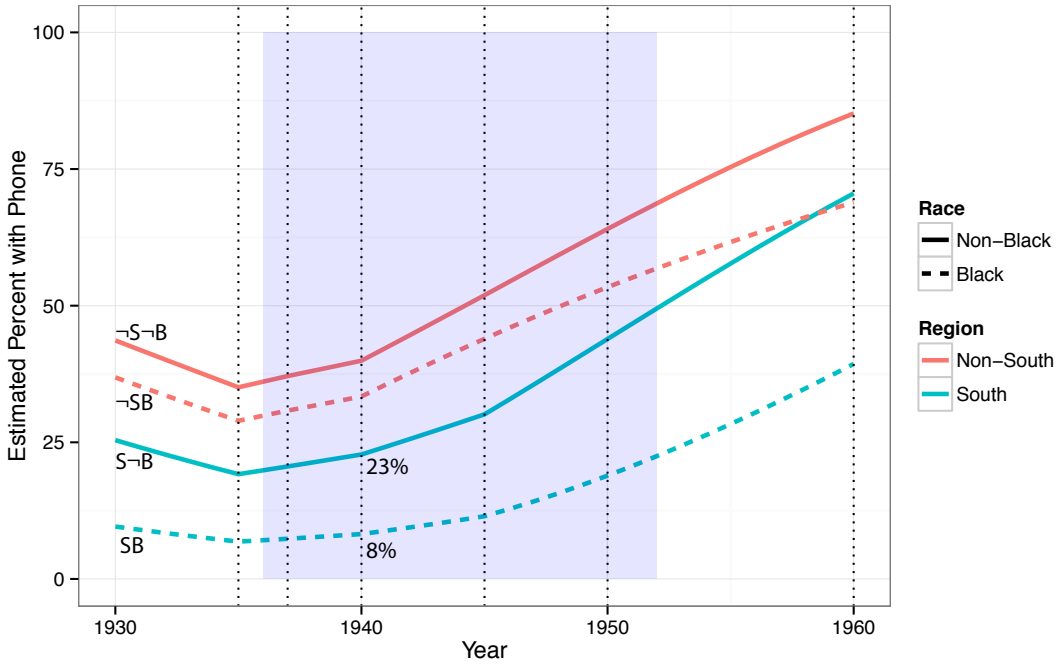

Figure 1: Estimated Phone Ownership by Race and Region, 1930–60. The shaded area denotes the years for which we require population targets. Vertical dotted lines indicate years for which population data are available.

and sometimes partial joint distributions; only once, in 1960, is the full joint distribution observed. Actual applications will share the same basic structure, but they are likely to involve more auxiliary variables and thus many more population moments to potentially match. It is this problem, that of selecting population targets, that the following section addresses.

3 Target Selection Using the Lasso

3.1 The Problem of Target Selection

In applications like Section 2.4, with only $V = 3$ variables and $C = 8$ cells, it is generally possible and desirable to poststratify poll samples so as to exactly match the joint population

distribution of the auxiliary variables.[11] Assuming the auxiliary variables are categorical, the poststratification weight of each sample unit i in cell c for survey s at time t is $w_{i[tc]} = n_s\hat{\pi}_{tc}/n_{cs}$, where n_s is the survey sample size and n_{cs} is the number of sample units in cell c. As C increases, however, it becomes more likely that $n_{cs} = 0$ for at least one cell, rendering $w_{i[tc]}$ undefined for that cell. Even if no cell is empty, the weights may become very large, resulting in unacceptably high-variance estimates. The problem of $C > n_s$ is likely to increase in the future as more sources of auxiliary data become available for weighting (e.g., consumer surveys; see West and Little, 2013).

One potential solution to this sparseness problem is to use a hierarchical model to generate shrinkage estimates for every cell, even empty ones; this is the basic motivation for MRP (see Gelman and Little, 1997). An alternative approach is to use only a subset of population moments as targets for the weights. Classical raking weights, for example, match only the marginal distributions of the auxiliary variables. More generally, under the calibration framework it is possible to use any set of population moments, joint or marginal, as weighting targets. The question then becomes how to choose the subset of moments to use as targets.

As is well known, weighting reduces nonresponse bias to the extent that the adjustment cells defined by values of weights are homogenous with respect to the response probabilities and with respect to the outcome variable (e.g., Little, 1986). This is because nonresponse bias is approximately equal to the population covariance between the response probabilities (ρ_i) and the outcome variable (y_i), divided by the average response probability. As a consequence, one should select auxiliary variables that predict the probability of response or the outcome variable of interest, preferably both. Beyond this, much of the specific advice on choosing auxiliary variables is essentially heuristic. Bethlehem, Cobben and Schouten (2011, chapter 9), for instance, suggests pre-selecting certain auxiliary variables based on substantive and theoretical knowledge, then selecting the rest based on their relationships with nonresponse and the main survey variables.

[11] The exception would be if some population cells are excluded from the sample by design, which of course is the case with the polls of interest in Section 2.4.

More formal selection criteria have also been proposed (for a review of indicators of nonresponse bias, see Wagner, 2012). Some, such as Särndal and Lundstrom (2008), recommend R^2-like statistics that capture variability in response probabilities. The advantage of focusing only on response probabilities is that the choice of weights does not depend on the outcome of interest, of which there may be many in a given poll. The downside is that weights increase the variance of outcome estimates unless they predict the outcome well. This has lead other works to suggest indicators that take into account the outcome variable as well as the response probabilities (e.g., Särndal and Lundstrom, 2005, 121–2).[12]

Whether derived from response probabilities only or from the outcome variable as well, the statistics referenced above provide an indicator with which to rank different sets of auxiliary variables by their estimated reduction in nonresponse bias. This requires calculating the nonresponse statistic for every possible variable subset, which quickly becomes computationally burdensome, especially when considering interactions (that is, the joint distribution of two or more auxiliary variables).[13]

3.2 Target Selection as Variable Selection

We build on this growing literature but take a different approach that conceptualizes the problem as equivalent to that of variable subset selection in model specification. As Särndal and Lundstrom (2005, chapter 10) observe, weights reduce nonresponse bias to the extent that the response influence $\omega_i = 1/\rho_i$ or the study variable y_i are well predicted by a linear combination of the auxiliary variables. Thus choosing the optimal set of auxiliary variables reduces to choosing the regression specification that best predicts ω_i and y_i, subject to any constraints on the resulting weights.

The statistical learning literature refers to this problem as variable subset selection. In theory, the solution is to compare all possible variable subsets on some measure of fit and

[12] See also Wagner (2010) and Andridge and Little (2011), though neither of these works uses its proposed nonresponse indicator to select auxiliary variables.

[13] For this reason, Särndal and Lundstrom's (2008) proposed stepwise selection procedure for the "best possible" auxiliary vector considers only main effects.

select the best subset. Such comprehensive comparisons become computationally infeasible, however, once the number of variables exceeds 30 or 40. Approximate stepwise shortcuts do exist, but they share with best-subset selection an undesirably high level of variance (Hastie, Tibshirani and Friedman, 2009, 57–61).

As its name suggests, the lasso ("least absolute shrinkage and selection operator") is a shrinkage estimator that also acts as a variable selector (Tibshirani, 1996). The lasso regression estimator can be written as

$$\hat{\beta}^{\text{lasso}} = \text{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{J} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{J} |\beta_j| \right\}, \tag{7}$$

where λ is a tuning paramater regulating how much the coefficients are shrunk away from the least-squares estimates and towards zero. In other words, λ determines the simplicity of the model specification. Like ridge regression and other shrinkage estimators, the lasso coefficient estimates are lower-variance and less prone to overfitting than least squares. Unlike ridge regression, the lasso's L_1 penalty $\sum_{j=1}^{J} |\beta_j|$ causes some coefficients to be shrunk all the way to zero—that is, to be dropped from the regression specification (Hastie, Tibshirani and Friedman, 2009, 68–9).

The lasso thus performs a function similar to best-subset and stepwise selection but is more computationally efficient than the former and less variable than the latter. In addition, like least-squares regression, the lasso can be generalized to multivariate $\boldsymbol{y}_i$, in which case the coefficient penalties are grouped across dependent variables. The variable-specific penalties can also be modified so as to require the inclusion of certain variables in every variable subset.

3.3 A Procedure for Lasso-Based Target Selection

Our goal is to generate weights that most effectively reduce the correlation between the response probabilities and relevant outcome variables, while also satisfying any user-defined

constraints on the weights or the auxiliary variables. For example, users may want to place bounds on the weights, to require that they match certain population targets, or use a consistent set of weights across multiple surveys. The following procedure is designed to achieve this goal:

1. Construct a training dataset for use in selecting auxiliary variables. This dataset may be the same poll for which weights are desired or a separate dataset with the same variables.

2. Identify the pool of potential auxiliary variables $\boldsymbol{x}_i$ as well as the outcome variables of interest $\boldsymbol{y}_i$.

3. Estimate the response influence ω_i of each unit in the training dataset with its weight $\hat{\omega}_i$, calculated using the joint or marginal population distributions of the auxiliary variables as targets and ignoring empty cells. If the $\hat{\omega}_i$ are highly skewed, it may be desirable to apply a logarithmic or other transformation.

4. For each of a range of λ values, apply the lasso using $\hat{\omega}_i$ and $\boldsymbol{y}_i$ as multivariate responses and $\boldsymbol{x}_i$ as predictors. Set to zero the penalty for any coefficient that, based on theoretical or substantive knowledge, must be included in the auxiliary vector. Save the set of non-zero coefficients.

5. Using the actual poll data, start with the largest value of λ (i.e., the simplest model) and calculate calibration weights with the corresponding lasso-selected variable subset as moment constraints.[14] Continue likewise for successively smaller values of λ until it proves impossible to construct weights (e.g., due to empty cells in the sample).

6. Select the most "complex" set of weights that satisfy desired restrictions (e.g., bounds).

This procedure should result in the selection of weights that predict $\boldsymbol{y}_i$ and ω_i about as well as possible. It is important that the variables in $\boldsymbol{y}_i$ be chosen with care and be

[14]The two primary choices of calibration method are multiplicative weighting (i.e., raking) and linear weighting; see Kalton and Flores-Cervantes (2003) for a comparison.

limited in number. If they are too numerous, the lasso will give prediction of $\hat{\omega}_i$ low priority in the selection of variable subsets. This is problematic insofar as analytic interest lies in outcome variables not in $\boldsymbol{y}_i$, for which nonresponse bias may be reduced much less than for $\boldsymbol{y}_i$. Indeed, if many outcome variables are of interest, it may be desirable to either reduce the dimensionality of $\boldsymbol{y}_i$ with factor analysis or another method, or else to predict only $\hat{\omega}_i$, along the lines suggested by Särndal and Lundstrom (2008).

4 Application to Quota-Sampled Polls

This section applies the two techniques introduced in this paper—Bayesian population interpolation and lasso-based target selection—to the problem of constructing weights for a quota-sampled opinion poll. We begin by using the model in Section 2.3 to estimate the joint population distribution of seven auxiliary variables in each year between 1930 and 1960. Next, we use the multivariate lasso to select population targets for their predictiveness of the response probabilities, voter turnout, and presidential partisanship. Finally, on the basis of the targets selected by the lasso as well as substantive knowledge, we calculate weights for two quota-sampled polls fielded in 1940 and evaluate the weights on several external metrics.

In this example, we use the following seven auxiliary variables, all binary unless otherwise indicated: *Black*, *Farm*, *Female*, *Phone Ownership*, *Professional*, *Region* (Midwest, Northeast, South, and West), and *Urban*.[15]. Table 3 lists the population data available in each year. Note that in no year is the full joint distribution available. Also, as in the illustrative example from Section 2.1, we exclude Southern blacks from our target population.

To create as large a training dataset as possible, we combine together several hundred polls fielded in the 1936–52 period, resulting in a combined dataset with over 625,000 valid

[15]In addition to these seven auxiliary variables, three more are available: *Age* (three categories), *Education* (four categories) and *State* (48 categories). Cross-classifying on the original seven variables plus *Age*, *Education*, and *State* results in over 37,000 cells per year. Our example uses only seven auxiliary variables so as to make the interpolation model less computationally demanding.

Year	Available Data
1930	*Black* × *Farm* × *Female* × *Professional* × *Region* × *Urban*
1937	*Phone Ownership* × *Region*
1940	*Black* × *Farm* × *Female* × *Professional* × *Region* × *Urban*
	Phone Ownership × *Region*
1945	*Phone Ownership* × *Region*
1950	*Black* × *Farm* × *Female* × *Professional* × *Region*
1960	*Black* × *Farm* × *Female* × *Phone Ownership* × *Professional* × *Region*

Table 3: Population Data for Quota-Sampled Polls

respondents. One of the few political variables common to all the component polls is retrospective presidential vote, which we recode to three dummy variables: *Voted Democratic*, *Voted Republican*, and *Did Not Vote*.[16] Together, these variables provide a good proxy for respondents' political engagement and orientation. In addition, we poststratify every respondent using the auxiliary variable set defined above, treating the resulting weights $\hat{\omega}_i$ as estimates respondents' response influences.[17] The three vote variables and the logarithm of the poststratification weights form the multivariate response surface for the lasso.[18]

We convert the seven auxiliary variables to nine non-collinear dummy variables.[19] We include all nine in the lasso variable set along with their (non-collinear) two-way, three-way, and four-way interactions, for a total of 182 potential variables. Based on substantive knowledge of the sampling scheme (for details, see Berinsky, 2006), we require that any variable subset selected include the main effects of all the auxiliary variables as well as an indicator for Southern blacks. We accomplish this by multiplying the corresponding coefficients in the penalty $\lambda \sum_j |\beta_j|$ by 0.[20]

[16] A small residual category, those who voted for minor-party candidates, was excluded. As in nearly all other surveys, voting was substantially over-reported in our data.

[17] We poststratified the data separately by year and normalized the weights to have a mean of 1 in each year. We ignored population cells that did not appear in the sample.

[18] We take the natural logarithm of the weights to make their distribution more symmetric. In addition, this has the additional benefit of matching the log-linear functional form of the model underlying raking weights.

[19] *Region* was decomposed into the indicators *Northeast*, *South*, and *West*, with Midwest as the excluded category.

[20] We also require the inclusion of fixed effects for presidential election cycle.

We specify a grid of 50 values for the shrinkage parameter λ ranging from 10^{-4} to 10^{-1}, equally spaced on the $\log_{10}$ scale, and select a variable subset for each one.[21] The simplest (most regularized) variable subset selected by the lasso includes only the required variables. The number of selected variables increases log-linearly as λ decreases, reaching a maximum value of almost 150 variables at $\lambda = 10^{-4}$.

Based on this ranking of variable subsets, we generate weights for two Gallup polls respectively fielded in February and October 1940.[22] We choose the first because it asked respondents whether they own a car, providing a useful indicator of the class bias in the sample. The second poll is the last one Gallup conducted before the 1940 presidential election, enabling us to compare election predictions under various weighting schemes to the actual result. Each of these poll samples contained close to 3,000 respondents. Neither includes Southern blacks, so we again drop them from the target population.

We attempted to generate raking weights using each lasso variable subset as a set of moment targets.[23] We started with the simplest subset and tried each successive one until weighting proved impossible. Weighting for the February poll failed on the eighth variable subset, and the October poll failed on the seventh.[24] Under the most complex weighting specifications, the largest weights were 4 times larger than the average weight, somewhat greater than is considered ideal (e.g., Deville, Särndal and Sautory, 1993, 1018). The results reported below varied little across lasso-selected weighting specifications, suggesting that extreme weights did not cause much of a problem.

In the February 1940 poll, weighting is quite effective at reducing class bias, at least for the top two-thirds of the SES spectrum, which drives the variation in car ownership. Based on a 1948 probability-sampled consumer survey and yearly numbers on automobile

[21] To implement the lasso itself, we used the function `glmnet` (Friedman, Hastie and Tibshirani, 2010) in the R computing environment (R Core Team, 2014).

[22] The polls were AIPO #183 (February 2–7, 1940) and AIPO #219 (October 26–31, 1940).

[23] We used the R function `rake` from the `survey` package (Lumley, 2012).

[24] The February poll failed because the eighth variable subset included *Farm* × *Professional*, but this cell was empty in the poll sample. The October poll similarly failed due to an empty *Black* × *Phone* × *Urban* cell.

registrations and population (from 1930–50), we calculate a rough population target of 49–52% car ownership in 1940.[25] The unweighted percentage of car owners in the Gallup poll is much higher at 60%, with a standard error of 1%. But weighting the poll to match the most complex (feasible) set of targets yields an estimate of 51%, right in the range of our out-of-sample estimates. The inclusion of *Phone Ownership* as a weighting variable—which was made possible by the interpolation method described in Section 2—itself accounts for about half of the bias reduction.[26]

The October poll, fielded immediately before the 1940 presidential election, contains a question on presidential vote preference. Given its high correlation with the retrospective vote variable used to select targets, we should expect weights to substantially improve estimates of prospective presidential vote. The results are consistent with this expectation. In the unweighted data, 48% of respondents who expressed a preference said they would vote for the Democratic Franklin Roosevelt—almost certainly an underestimate of the population proportion, given the sample's overrepresentation of Republican-leaning higher-SES respondents. Under any of the six lasso-selected weight sets, the estimate rises to 54%, within a percentage point of FDR's ultimate share of 54.7%.[27]

In summary, the combination of model-based population interpolation and lasso-based target selection produces weights that seem to perform well. Given that the lasso-selected weight sets generate estimates very similar to those of the baseline set derived from substantive knowledge, the target-selection procedure appears to be less important in this application than population interpolation, which permits the use of *Phone Ownership* as an auxiliary variable. The lasso may be less useful in this instance because, first, our substantive knowledge of the determinants of nonresponse is usually rich, and second, empty cells in the sample

[25] Part of the uncertainty stems from having to extrapolate backwards from 1948, and part stems from the removal of Southern blacks from the target population.

[26] The car-ownership estimate using weights derived from all marginals except *Phone Ownership* is 55%.

[27] The weighted estimates of FDR's share are higher in all regions, especially outside the South, where the base percentage is higher and the class bias in presidential partisanship less pronounced. Note that the exclusion of Southern blacks from the target population does little to bias the estimates because Southern blacks were excluded from the voting population as well.

prevented the calculation of weights based on more complex targets. In the future iterations, it might be beneficial to modify target selection so as to avoid such empty cells.

5 Conclusion

This paper has introduced two advances in survey weighting methodology: a Bayesian approach to population interpolation and a lasso-based approach to auxiliary variable selection. Though we have only begun to explore these techniques and validate their performance, we are encouraged by the results of their application to quota-sampled opinion polls. Further, we believe that these techniques are of potentially wide applicability beyond the domain for which they were designed. In light of the decline of response rates, the return of non-probability sampling, and the growth in potential sources of auxiliary information, survey researchers need new tools to produce quality inferences, and we hope that our work proves useful in this regard.

References

Andridge, Rebecca R. and Roderick J. A. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27(2):153–180.

Berinsky, Adam J. 2006. "American Public Opinion in the 1930s and 1940s: The Analysis of Quota-Controlled Sample Survey Data." *Public Opinion Quarterly* 70(4):499–529.

Berinsky, Adam J., Eleanor Neff Powell, Eric Schickler and Ian Brett Yohai. 2011. "Revisiting Public Opinion in the 1930s and 1940s." *PS: Political Science & Politics* 44(3):515–520.

Bethlehem, Jalke G. 2002. Weighting Nonresponse Adjustments Based on Auxiliary Information. In *Survey Nonresponse*, ed. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J. A. Little. New York: Wiley chapter 18, pp. 275–287.

Bethlehem, Jelke, Fannie Cobben and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: Wiley.

Bryant, John R. and Patrick J. Graham. 2013. "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources." *Bayesian Analysis* 8(2):1–32.

Cargnoni, Claudia, Peter Muller and Mike West. 1997. "Bayesian Forecasting of Multinomial Time Series Through Conditionally Gaussian Dynamic Models." *Journal of the American Statistical Association* 92(438):640–647.

Deville, Jean-Claude. 2000. "Simultaneous Calibration of Several Surveys." *Proceedings of Statistics Canada Symposium 99: Combining Data from Different Sources*, Publication No. 11-522-XCB, Statistics Canada, Ottawa, September 2000, pp. 225–230.

Deville, Jean-Claude and Carl-Erik Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87(418):376–382.

Deville, Jean-Claude, Carl-Erik Särndal and Olivier Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88(423):1013–1020.

Freedman, David A. 2001. Ecological Inference and the Ecological Fallacy. In *International Encyclopaedia of the Social and Behavioural Sciences*, ed. N. J. Smelser and P. B. Baltes. Vol. 6 New York: Elsevier p. 40274030.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1):1–22.

Gelman, Andrew and Thomas C. Little. 1997. "Poststratification Into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23(2):127–135.

Grunwald, Gary K., Adrian E. Raftery and Peter Guttorp. 1993. "Time Series of Continuous Proportions." *Journal of the Royal Statistical Society. Series B (Methodological)* 55(1):103–116.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Hoffman, Matthew D. and Andrew Gelman. Forthcoming. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* .

Kalton, Graham and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19(2):81–97.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

King, Gary, Ori Rosen and Martin A. Tanner. 2004. Information in Ecological Inference: An Introduction. In *Ecological Inference: New Methodological Strategies*, ed. Gary King, Ori Rosen and Martin A. Tanner. New York: Cambridge University Press chapter 1, pp. 1–12.

Little, Roderick J. A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54(2):139–157.

Little, Roderick J. and Sonya Vartivarian. 2005. "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31(2):161–168.

Lumley, Thomas S. 2012. "`survey`: Analysis of Complex Survey Samples." R package version 3.28-2.

Park, David K., Andrew Gelman and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12(4):375–385.

Quinn, Kevin M. 2004. Ecological Inference in the Presence of Temporal Dependence. In *Ecological Inference: New Methodological Strategies*, ed. Gary King, Ori Rosen and Martin A. Tanner. New York: Cambridge University Press chapter 9, pp. 207–233.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org/.

Raftery, Adrian E., Nan Li, Hana Ševčíková, Patrick Gerland and Gerhard K. Heilig. 2012. "Bayesian Probabilistic Population Projections for All Countries." *Proceedings of the National Academy of Sciences* 109(35):13915–13921.

Särndal, Carl-Erik. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33(2):99–119.

Särndal, Carl-Erik and Sixten Lundstrom. 2005. *Estimation in Surveys with Nonresponse*. Hoboken, NJ: Wiley.

Särndal, Carl-Erik and Sixten Lundstrom. 2008. "Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator." *Journal of Official Statistics* 24(2):167–191.

Stan Development Team. 2013. "Stan: A C++ Library for Probability and Sampling, Version 1.3." http://mc-stan.org/.

Swanson, David A. and Jeff Tayman. 2012. *Subnational Population Estimates*. New York: Springer (PDF ebook).

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Wagner, James. 2010. "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data." *Public Opinion Quarterly* 74:223–43.

Wagner, James. 2012. "A Comparison of Alternative Indicators for the Risk of Nonresponse Bias." *Public Opinion Quarterly* 76(3):555–575.

Wakefield, Jon. 2004. "Ecological Inference for 2×2 Tables." *Journal of the Royal Statistical Society. Series A (General)* 167(3):385–445.

West, Brady T. and Roderick J. A. Little. 2013. "Non-Response Adjustment of Survey Estimates based on Auxiliary Variables Subject to Error." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 62(2):213–231.

Wheldon, Mark C., Adrian E. Raftery, Samuel J. Clark and Patrick Gerland. 2013. "Reconstructing Past Populations With Uncertainty From Fragmentary Data." *Journal of the American Statistical Association* 108(501):96–110.

Wittenberg, Martin. 2009. "Sample Survey Calibration: An Information-Theoretic Perspective." Southern Africa Labour and Development Research Unit, University of Cape Town. SALDRU Working Paper No. 41. http://ideas.repec.org/p/ldr/wpaper/41.html.

A Stan Code

```
data {
  int<lower=1> T; // number of time periods
  int<lower=1> N; // number of cells
  int<lower=1> M; // maximum number of margins
  real<lower=1> priorN;
  real<lower=1> sampleN;
  real<lower=1> transitN;
  real Tgaps[T]; // periods between observed data
  simplex[N] props0; // priors
  int<lower=0> countsT1M1[144]; // observed counts
  matrix<lower=0,upper=1>[144, N] iotaT1M1; // matrix of 1s and 0s
  int<lower=0> countsT1M2[2];
  matrix<lower=0,upper=1>[2, N] iotaT1M2;
  // etc. //
}
parameters {
  simplex[N] props[T]; // period-specific cell probabilities
}
model {
  props[1] ~ dirichlet(props0 * priorN);
  for (t in 2:T) {
    props[t] ~ dirichlet(props[t - 1] * transitN / Tgaps[t]);
  }
  countsT1M1 ~ multinomial(iotaT1M1 * props[1]);
  countsT1M2 ~ multinomial(iotaT1M2 * props[1]);
  // etc. //
}
```