# Dynamic Ecological Inference for Time-Varying Population Distributions Based on Sparse, Irregular, and Noisy Marginal Data

## Devin Caughey[1] and Mallory Wang[2]

[1] *MIT, Political Science, 77 Massachusetts Ave., Room E53-463, Cambridge, MA 021040, USA. Email:* devin.caughey@gmail.com
[2] *Uber, 555 Market Street 4th Floor, San Francisco, CA 94108, USA. Email:* mallory.wang@gmail.com

## Abstract

Social scientists are frequently interested in how populations evolve over time. Creating poststratification weights for surveys, for example, requires information on the weighting variables' joint distribution in the target population. Typically, however, population data are sparsely available across time periods. Even when population data are observed, the content and structure of the data—which variables are observed and whether their marginal or joint distributions are known—differ across time, in ways that preclude straightforward interpolation. As a consequence, survey weights are often based only on the small subset of auxiliary variables whose joint population distribution is observed regularly over time, and thus fail to take full advantage of auxiliary information. To address this problem, we develop a dynamic Bayesian ecological inference model for estimating multivariate categorical distributions from sparse, irregular, and noisy data on their marginal (or partially joint) distributions. Our approach combines (1) a Dirichlet sampling model for the observed margins conditional on the unobserved cell proportions; (2) a set of equations encoding the logical relationships among different population quantities; and (3) a Dirichlet transition model for the period-specific proportions that pools information across time periods. We illustrate this method by estimating annual U.S. phone-ownership rates by race and region based on population data irregularly available between 1930 and 1960. This approach may be useful in a wide variety of contexts where scholars wish to make dynamic ecological inferences about interior cells from marginal data. A new R package `estsubpop` implements the method.

*Keywords:* ecological inference, survey weighting, demographic interpolation, Bayesian models, Dirichlet dynamic model

## 1 Problem: Subpopulation Estimation with Irregular Data

Social scientists are often interested in how populations evolve over time. Estimating the composition of populations is a central concern of demography as well as a frequent target of ecological inference (EI). One common use of population estimates is as targets for weights designed to make samples more representative of the population. Survey researchers, for example, often poststratify poll samples so that the joint distribution of certain variables matches the target population. Constructing such multivariate population targets is straightforward when data on variables' joint distribution can be derived from a single authoritative source, such as the Integrated Public Use Microdata Series (IPUMS) of anonymized samples of individual U.S. Census records (Ruggles *et al.* 2017). As long as the population data are observed in consistent form across time, it is also simple to make these population targets dynamic by interpolating between decennial IPUMS samples (e.g., Enns and Koch 2013).

Frequently, however, data on population distributions must be compiled from multiple sources, which may exhibit inconsistencies due to sampling or measurement error. Moreover, the structure of population datasets or the variables they include often differ across time periods. Faced with such barriers to simple interpolation, scholars typically either use time-invariant

**Political Analysis (2019)**
**vol. 27:388–396**
**DOI:** 10.1017/pan.2019.4

**Published**
11 April 2019

**Corresponding author**
Devin Caughey

**Edited by**
Justin Grimmer

Table 1. Phone ownership by race by region in 1940. Unobserved cell proportions are represented by $\pi$ and observed marginal proportions by $p$. Subscripts indicate the presence (uppercase) or absence (lowercase) of the three attributes.

| | South<br>($p_{S++} = 0.25$) | | | Non-South<br>($p_{s++} = 0.75$) | | |
|---|---|---|---|---|---|---|
| | Phone | Non-Phone | | Phone | Non-Phone | |
| Black | $\pi_{SBP}$ | $\pi_{SBp}$ | $p_{SB+} = 0.06$ | $\pi_{sBP}$ | $\pi_{sBp}$ | $p_{sB+} = 0.03$ |
| Non-Black | $\pi_{SbP}$ | $\pi_{Sbp}$ | $p_{Sb+} = 0.19$ | $\pi_{sbP}$ | $\pi_{sbp}$ | $p_{sb+} = 0.72$ |
| | $p_{S+P} = 0.05$ | $p_{S+p} = 0.21$ | | $p_{s+P} = 0.30$ | $p_{s+p} = 0.45$ | |

population targets—which is obviously problematic for studies that span periods of substantial demographic change—or confine their attention to the few variables whose joint distribution is observed in consistent form across time (cf. Leeman and Wasserfallen 2017).

As an alternative, we develop a Bayesian framework for estimating population cell proportions over time, conditional on all available population data. The data may consist of marginal or joint distributions, be observed irregularly over time, and come from multiple noisy or inconsistent datasets. Extending the dynamic EI model of Quinn (2004), our approach uses a Dirichlet random walk to allow past and future as well as contemporary data to inform the cell estimates for a given year. We motivate this model with the example of estimating phone ownership by race and region and implement it in a new R package, `estsubpop` (Caughey and Wang 2018a).

## 2 Motivation: Phone Ownership by Race and Region

As motivation, consider the challenge faced by Berinsky *et al.* (2011), who constructed survey weights for quota-sampled opinion polls fielded between 1936 and 1945. To ameliorate the sampling biases of the polls, the authors sought to poststratify the samples by region, race, and indicators of class status. But because poststratification (a.k.a. cell weighting) requires knowledge of the weighting variables' joint distribution in the population, these authors could not poststratify on variables for which only marginal distributions were available, such as phone ownership. They were thus forced to choose to either drop phone ownership as a weighting variable or abandon poststratification in favor of raking weights, which match variables' marginal but not joint distributions.[1]

Table 1 illustrates the structure of the problem Berinsky *et al.* faced, using data from 1940. It presents a 2 × 2 × 2 array of cells defined by the binary variables *South*, *Black*, and *Phone*. Cells' population proportions are represented by $\pi$, whose subscripts indicate the presence (uppercase) or absence (lowercase) of the three attributes. If these cell proportions were known, they could be used to create poststratification weights by dividing them by the cells' corresponding proportions in the sample. But data on the joint distribution of these three variables is not available until the 1960 IPUMS, when phone-ownership rates were much higher than in 1936–45. All that is available in the 1936–45 period is information on the marginal distributions of *Phone* and *Black* within each region.[2] These observed marginal proportions are represented by $p$.

If race and phone ownership were independent within region, the cell proportions could be estimated by multiplying the corresponding marginal proportions. Under this assumption, the phone-ownership rate in the South in 1940 would be naively estimated to be 19% for both blacks and whites. Unfortunately, this assumption is highly implausible, for phone ownership was almost certainly much more common among whites. This racial disparity is clear in the 1960 IPUMS, which reports a phone-ownership rate of 70% for Southern whites versus 39% for Southern blacks.

[1] Berinsky *et al.* (2011) ultimately decided to create raking weights for phone ownership, but discouraged use of these weights in favor of poststratification weights based on education or occupation.

[2] The 1940 IPUMS contains *Black* × *South*. *Phone* × *South* can be derived from AT&T corporate records.

The problem, then, is how to formally incorporate the 1960 information on the joint distribution of *South*, *Black*, and *Phone* into our cell estimates two decades earlier.

## 3 Related Problems and Methods

The problem illustrated in the previous section is closely related to classic problems in EI, such as the oft-studied example of voter registration by race. (In EI, the quantities of interest are often defined as conditional rates rather than cell proportions, but the latter are a function of the former; King 1997, 29–31.)[3] EI is unreliable without supplementary data or prior information, and Bayesian models offer a convenient way of incorporating additional information. Most relevantly, Quinn (2004) develops a dynamic EI model that shrinks estimates across time periods, which he notes are a powerful source of such information when data are temporally dependent. Quinn's model, however, is designed for situations where the margins of the same $2 \times 2$ table are observed in the same form in each time period. By contrast, we are interested in applications where there may be multiple (possibly inconsistent) data sources available in a given year, and the structure and content of the data may differ across years. We therefore augment Quinn's approach with ideas borrowed from demography (e.g., Bryant and Graham 2013), specifically the idea of combining (1) multiple observation models for different data sources; (2) a "demographic account" that encodes the logical relationships among different population quantities; and (3) a transition model for change in population parameters across time periods.[4]

## 4 A Bayesian Model for Dynamic Population Estimation

Estimating the joint population distribution of $V$ nonnested categorical variables is equivalent to estimating the population proportions of $C = \prod_v^V L_v$ cells, where $L_v$ indicates the number of possible values that variable $v$ can take. Let $\boldsymbol{\pi}_t = (\pi_{t1}, \ldots \pi_{tc}, \ldots, \pi_{tC})'$ denote the simplex of cell proportions in period $t \in \{1 \ldots T\}$. In each period, population data are available on the joint distribution of $M_t \geqslant 0$ subsets of the $V$ variables. Each variable subset $m \in \{1 \cdots M_t\}$ contains $V_{tm} \leqslant V$ variables, whose levels define $G_{tm} = \prod_w^{V_{tm}} L_w$ groups, each composed of $H_{tmg} \geqslant 1$ cells. The population data for each variable subset $m$ in period $t$ consist of the simplex $\boldsymbol{p}_{tm} = (p_{tm1}, \ldots p_{tmg}, \ldots, p_{tmG_{tm}})'$. Each group proportion $p_{tmg}$ is an estimate of group $g$'s true population proportion $\phi_{tmg}$, the sum of the proportions $\pi_{tc}$ of the cells that compose group $g$.

For intuition, consider the problem of using the 1960 IPUMS and the 1940 data in Table 1 to estimate phone ownership by race and region between 1940 and 1960. This example involves $V = 3$ auxiliary variables (*South*, *Black*, and *Phone*), each with $L_v = 2$ levels. Of interest are the population proportions $\pi_{tc}$ of $C = 2 \times 2 \times 2 = 8$ cells in each of $T = 21$ years. In 1940 ($t = 1$), data on the joint distribution of $M_1 = 2$ variable subsets are available: $\{South, Black\}$ and $\{South, Phone\}$. For each variable subset $m$, $V_{1,m} = 2$, $G_{1,m} = 4$, and $H_{1,mg} = 2\ \forall g$. We observe $M_1 = 2$ vectors of group proportions $\boldsymbol{p}_{1,m}$, which are estimates of $\boldsymbol{\phi}_{1,m}$ and which correspond to the marginal proportions in Table 1 (with different notation). In 1960 ($t = 21$), $M_{21} = 1$ variable subset is available: $\{South, Black, Phone\}$. For this subset, $V_{21,m} = 3$, $G_{21,m} = 8$, and $H_{21,mg} = 1\ \forall g$. No data are available for years between 1940 and 1960, so $M_t = 0\ \forall t \notin \{1, 21\}$.

Our goal is to use the observed group proportions $\boldsymbol{p}_{tm}$ to make inferences about the true cell proportions $\boldsymbol{\pi}_t$. We do so with a Bayesian model that combines two submodels: an *observation model* linking $\boldsymbol{\pi}_t$ (via $\boldsymbol{\phi}_{tm}$) to the data $\boldsymbol{p}_{tm}$ and a *transition model* specifying how $\boldsymbol{\pi}_t$ evolves over

[3] For example, the phone-ownership rate among Southern blacks, $\beta_{BS} = \text{Pr(Phone|South, Black)}$, is equal to $\pi_{SBP}/(\pi_{SBP} + \pi_{SBp})$.

[4] Also related is the work of Leeman and Wasserfallen (2017), who address the problem of using multilevel regression and poststratification when the joint population distribution is unknown. Leeman and Wasserfallen's proposed solution is to combine marginal population data with survey data to create a "synthetic" joint distribution. Their method can be thought of as a static version of our approach, which can incorporate data from multiple points in time as well as from multiple sources (including surveys).

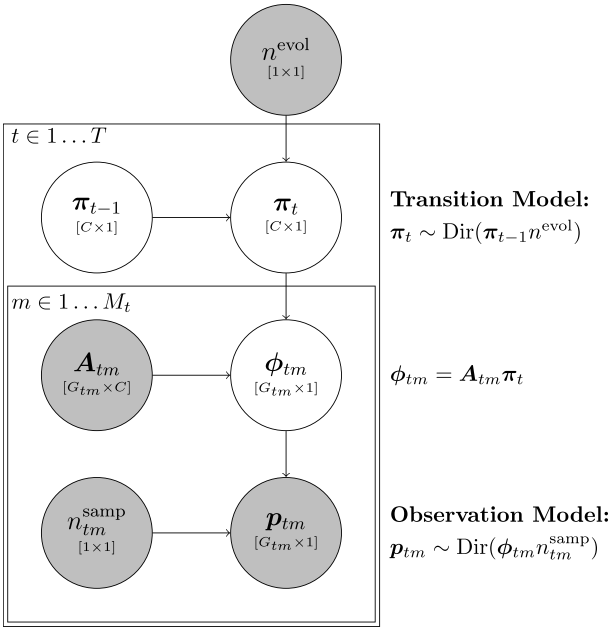

$n^{\text{evol}}$ $[1\times1]$
$t \in 1 \dots T$
$\boldsymbol{\pi}_{t-1}$ $[C\times1]$
$\boldsymbol{\pi}_t$ $[C\times1]$
Transition Model:
$\boldsymbol{\pi}_t \sim \text{Dir}(\boldsymbol{\pi}_{t-1} n^{\text{evol}})$
$m \in 1 \dots M_t$
$\boldsymbol{A}_{tm}$ $[G_{tm}\times C]$
$\boldsymbol{\phi}_{tm}$ $[G_{tm}\times1]$
$\boldsymbol{\phi}_{tm} = \boldsymbol{A}_{tm}\boldsymbol{\pi}_t$
$n_{tm}^{\text{samp}}$ $[1\times1]$
$\boldsymbol{p}_{tm}$ $[G_{tm}\times1]$
Observation Model:
$\boldsymbol{p}_{tm} \sim \text{Dir}(\boldsymbol{\phi}_{tm} n_{tm}^{\text{samp}})$

**Figure 1.** Plate diagram of the dynamic EI model. Shaded nodes indicate variables that are observed or set by the analyst.

time (see Figure 1 for an overview).[5] The observation model allows for measurement error in the observed group proportions $\boldsymbol{p}_{tm}$ relative to the true proportions $\boldsymbol{\phi}_{tm}$. We represent this stochastic relationship between $\boldsymbol{p}_{tm}$ and $\boldsymbol{\phi}_{tm}$ using the Dirichlet distribution,

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

Under this model, the expected value of $p_{tmg}$ is $\phi_{tmg}$. The precision of the sampling distribution is determined by $n_{tm}^{\text{samp}}$, which is specified by the analyst (e.g., based on the actual sample size of the data source for $\boldsymbol{p}_{tm}$).[6]

Each group proportion $\phi_{tmg}$ is the sum of the proportions of the $H_{tmg}$ cells that compose it. The relationship between $\boldsymbol{\phi}_{tm}$ and $\boldsymbol{\pi}_t$ is thus compactly described by the equation

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

where $\boldsymbol{A}_{tm}$ is a $G_{tm} \times C$ indicator matrix in which a 1 in row $g$ and column $c$ indicates that group $g$ contains cell $c$. In Table 1, for example, the $M_{1940} = 2$ vectors of observed (estimated) group proportions are

$$\begin{aligned}\boldsymbol{p}_{1940,South\times Black} &= (p_{\text{SB+}}, p_{\text{Sb+}}, p_{\text{sB+}}, p_{\text{sb+}})\\ \boldsymbol{p}_{1940,South\times Phone} &= (p_{\text{S+P}}, p_{\text{S+p}}, p_{\text{s+P}}, p_{\text{s+p}}),\end{aligned}$$

[5] This setup is similar to a state-space model, which implicitly defines a latent state $\xi$ with two equations: a measurement equation for the observed data conditional on $\xi_t$ and a transition equation for $\xi_t$ conditional on $\xi_{t-1}$ (e.g., Jackman 2009, 471).

[6] Alternatively, the observation model can be written using a multinomial sampling distribution. This requires rounding $n_{tm}^{\text{samp}}\boldsymbol{p}_{tm}$ to the nearest integer, but this minor inaccuracy maybe necessary if there are empty elements in $\boldsymbol{p}_{tm}$ because the Dirichlet cannot accommodate zero values.

and the corresponding unobserved (true) proportions are

$$\begin{aligned}\boldsymbol{\phi}_{1940,South\times Black} &= (\pi_{\text{SBP}} + \pi_{\text{SBp}}, \pi_{\text{SbP}} + \pi_{\text{Sbp}}, \pi_{\text{sBP}} + \pi_{\text{sBp}}, \pi_{\text{sbP}} + \pi_{\text{sbp}}) \\ &= \boldsymbol{A}_{1940,South\times Black}\boldsymbol{\pi}_{1940} \\ \boldsymbol{\phi}_{1940,South\times Phone} &= (\pi_{\text{SBP}} + \pi_{\text{SbP}}, \pi_{\text{SBp}} + \pi_{\text{Sbp}}, \pi_{\text{sBP}} + \pi_{\text{sbP}}, \pi_{\text{sBp}} + \pi_{\text{sbp}}) \\ &= \boldsymbol{A}_{1940,South\times Phone}\boldsymbol{\pi}_{1940}.\end{aligned}$$

The observed group proportions can be directly linked to the unobserved cell proportions by substituting (2) into (1), leading to the observation model

$$\boldsymbol{p}_{tm} \sim \text{Dir}(\boldsymbol{A}_{tm}\boldsymbol{\pi}_t n_{tm}^{\text{samp}}). \tag{3}$$

The model defined by (3) does not distinguish among cells in the same group $g$, so without further information the posterior distributions over the cell proportions will be equal within groups. If there is only $M_t = 1$ set of auxiliary variables, the posterior estimate of each cell proportion $\pi_{tc}$ in group $g$ will converge to the maximum likelihood estimate $p_{tmg}/H_{tmg}$, yielding weights identical to those that would be obtained with poststratification. If there are data on multiple sets of auxiliary variables, as in Table 1, then each of the $M_t > 1$ observation models will inform the cell estimates, yielding weights similar to those created by raking or calibration on the $M_t$ vectors of marginal proportions $\boldsymbol{p}_{tm}$.

In general, however, there may be not only multiple sets of auxiliary variables available in a given year but also different sets across years. Data from other years can thus provide information distinguishing cells for which no individuating data are available in year $t$. To pool this information across time periods, we model the temporal evolution of the proportion vector $\boldsymbol{\pi}_t$ using a Dirichlet transition model (cf. Grunwald, Raftery, and Guttorp 1993),

$$\boldsymbol{\pi}_t \sim \text{Dir}(\boldsymbol{\pi}_{t-1} n^{\text{evol}}), \tag{4}$$

where $\pi_{t-1,c}$ is the expected value of $\pi_{tc}$.[7] In periods with no data, $\boldsymbol{\pi}_t$ will be interpolated with values informed directly by the immediately adjacent periods and, indirectly, by all previous and subsequent estimates.

Since the variance of the Dirichlet is inversely proportional to the sample size, the hyperparameter $n^{\text{evol}}$ governs the degree of pooling across periods. It is possible either to set $n^{\text{evol}}$ exactly (as depicted in Figure 1) or to give it a hyperprior, but either way it should be specified with care because the degree of pooling over time can substantially affect inferences. In our application, we found that with a diffuse prior, $n^{\text{evol}}$ was usually estimated to be less than 2,000, which is too low to propagate much information across our three-decade period of interest.[8] We therefore recommend that analysts use substantive judgement to select a value or informative prior for $n^{\text{evol}}$. In general, $n^{\text{evol}}$ should be set large enough to propagate information across years without data, but not so large as to outweigh data when they are observed. One way to think through particular values of $n^{\text{evol}}$ is to reason backward from a "typical" yearly change. For instance, the expectation that a cell that currently constitutes half the population will change by a percentage point between years implies a choice of $n^{\text{evol}} = 0.5 \times (1 - 0.5)/0.01^2 = 2{,}500$. For an illustration of the consequences of different choices of $n^{\text{evol}}$, see Supplementary Appendix A.1.3.

[7] In $t = 1$, $\boldsymbol{\pi}_t \sim \text{Dir}(\boldsymbol{\pi}_0 n^{\text{prior}})$, where $\boldsymbol{\pi}_0$ is a simplex of user-specified prior means and $n^{\text{prior}}$ is the prior "sample size." A natural default is a uniform prior with $n^{\text{prior}} = C$, which leads to $\boldsymbol{\pi}_t \sim \text{Dir}(\boldsymbol{1}_C)$.

[8] The implied sample size of the dynamic prior is proportional to the number of intervening time periods. Thus, if $n^{\text{evol}} = 1{,}000$, data measured 30 periods away has the same informativeness as a sample size of 33 (1000/30).

## 5 Validation

To validate our model, we examine how accurately it recovers the known cross-tabulation of *South*, *Black*, and *Phone* in the 1960, 1970, 1980, and 1990 IPUMS based on partial information about the joint distribution.[9] We compare estimates based on four data configurations:

(1) One-way marginal distributions in each year;
(2) Marginals in each year for *Black* × *South* and *Phone* × *South* but not *Black* × *Phone*;
(3) Three-way crosstab in 1960 but otherwise the same two-way marginals as 2 above;
(4) Three-way crosstabs in all years.

Case 3 closely mirrors the data structure in our substantive application except that the 1960 crosstab data are used to inform estimates for future rather than past years. For all cases we set $\pi_{0c} = 1/8\ \forall c$, $n^{\text{prior}} = 8$, and $n^{\text{evol}} = \exp(10) \approx 22{,}026$.[10] We estimate the model in the Bayesian simulation program Stan (Stan Development Team 2018), as called from R by `estsubpop`.[11]

Figure 2 plots the key comparison, case 3 (with 1960 crosstabs) versus case 2 (without), focusing the estimated population percentages of phone-owning and non-phone-owning blacks in each region. Despite sharing the same data as case 2 in 1970, 1980, and 1990, case 3's credible intervals (CIs) track the true IPUMS targets in these years much more closely. This is due to 1960's information on phone ownership by race, which is propagated forward in time. The performance of all four sets of estimates is formally compared in Figure 3. As one would hope, estimates based on the full joint distribution in every year (case 4) perform best in terms of root mean squared error (RMSE) and CI noncoverage.[12] However, the case 3 estimates (dotted line) are very accurate as well and clearly outperform cases 2 and 1. Not only are case 3's RMSEs a tenth as large as case 2's, but its CIs also do not exhibit false precision. In short, this simulation validates the usefulness of a model that can accommodate varying data structures over time, without which the crosstab data from 1960 could not be utilized.

## 6 Application

We now apply this approach to a more elaborate version of our running example, estimating the joint distribution of *South*, *Black*, and *Phone* in each year between 1930 and 1960. Table 2 details the population data we use to inform our estimates. While marginal data somewhat more frequent in this application than in the validation example, the key similarity is that data on full distribution are available only for 1960, well outside our main period of interest (1936–45).

As in Section 5, we set $\pi_{0c} = 1/8\ \forall c$, $n^{\text{prior}} = 8$, and $n^{\text{evol}} = \exp(10) \approx 22{,}026$ (see Supplementary Appendix A.1.3 for a sensitivity analysis).[13] Estimating the model with these data generates $C = 8$ estimated proportions in each of the $T = 31$ years. We transform these cell proportions into implied phone-ownership percentages by race and region and plot their posterior medians and 50% CIs in Figure 4. Notwithstanding the general growth in phone ownership after 1935, the ownership rate among blacks is always estimated to be lower than non-blacks in the same region. Note that because only a small percentage of non-Southerners in this period were African American, the ownership rate among non-Southern blacks is estimated much less precisely than that of the other three groups.

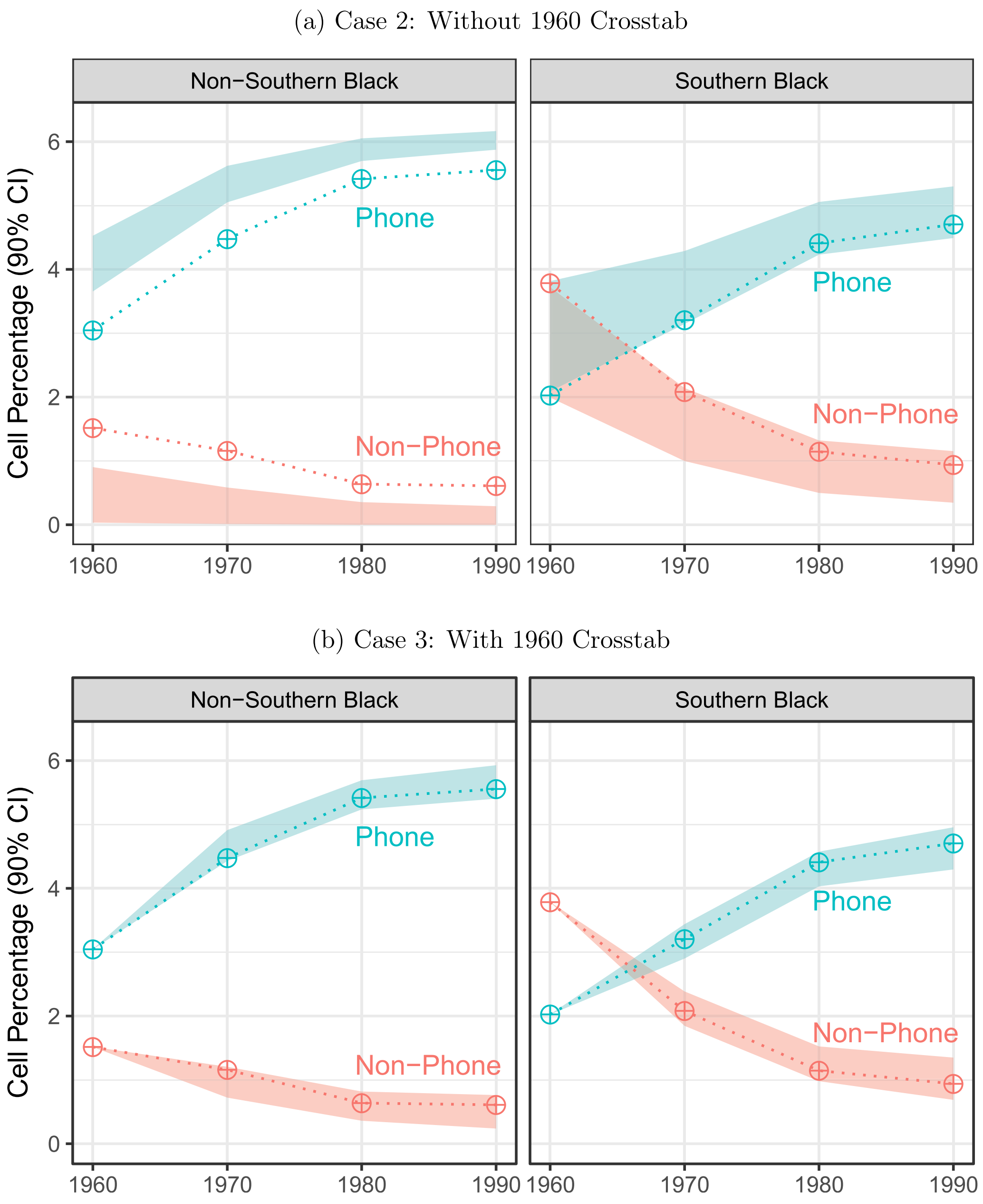
[9] Replication materials for this article can be downloaded from Caughey and Wang (2018b).

[10] A version of the simulations with $n^{\text{evol}}$ =100,000 yielded very similar point estimates but overly narrow credible intervals. Giving $n^{\text{evol}}$ a vague prior resulted in an estimated $n^{\text{evol}}$ between 1,000 and 2,000. Cell CIs from models with $n^{\text{evol}}$ of this magnitude tended to be overly conservative.

[11] Since validation targets are available only in census years, we estimate cell proportions only in those years. We adjust for this by dividing $n^{\text{evol}}$ by the number of skipped years (10) between estimates.

[12] By *noncoverage* we mean that the CI for a parameter does not include the true value of that parameter.

[13] We estimated 4 chains, each with 10,000 iterations (half devoted to warmup). Standard diagnostics indicated convergence. Computation was done on a MacPro with 32 GB of RAM and a 3.5 GHz processor. Three chains finished in less than an hour, but one chain took 80 hours. In our experience, such variability in computation time across chains is not uncommon when sampling from these models.


(a) Case 2: Without 1960 Crosstab

(b) Case 3: With 1960 Crosstab

**Figure 2.** Cell estimates for Southern and non-Southern blacks, based on data with (bottom) and without (top) the full crosstab in 1960. Crosshairs and dotted lines indicate IPUMS targets. Shaded regions indicate 90% credible intervals.

In addition to their substantive plausibility, the estimates are roughly consistent with other sources from this period. For example, a 1935–36 government study of consumer purchases found that non-black Southerners were 3.3 times more likely to own a phone than black Southerners, which is not far from our estimated ratio of around 2.8.[14] This convergence, combined with the analogous model's accuracy in Section 5's validation analysis, bolsters our confidence that our subpopulation estimates are much more accurate than if phone ownership and race were assumed to be independent.

[14] This survey was high quality for the time but still overestimated the regional phone-ownership rate by about 10 points relative to the AT&T data we use. See ICPSR Study #8909 (doi:10.3886/ICPSR08908).

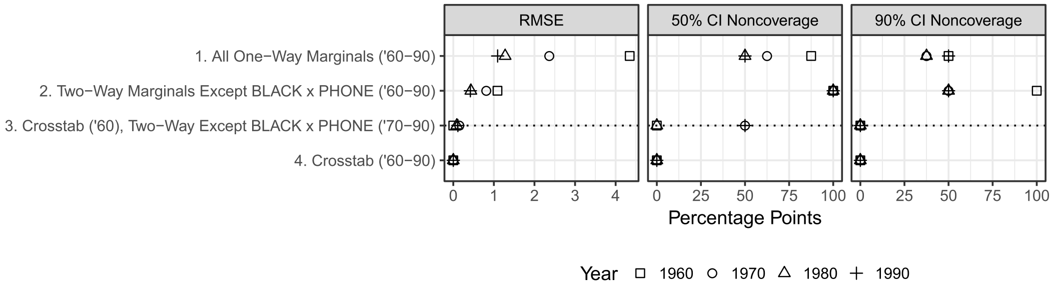


**Figure 3.** Accuracy of estimates based on different data configurations. The middle and right panels respectively report the proportion of true cell proportions not covered by the 50% and 90% CIs.

**Table 2.** Population data for example application.

| Year | Available data | Data source |
|---|---|---|
| 1930 | *Black* × *South* | IPUMS |
| | *Phone* | AT&T |
| 1935 | *Phone* | AT&T |
| 1937 | *Phone* × *South* | AT&T |
| 1940 | *Black* × *South* | IPUMS |
| | *Phone* × *South* | AT&T |
| 1945 | *Phone* × *South* | AT&T |
| 1950 | *Black* × *South* | IPUMS |
| 1960 | *Black* × *Phone* × *South* | IPUMS |

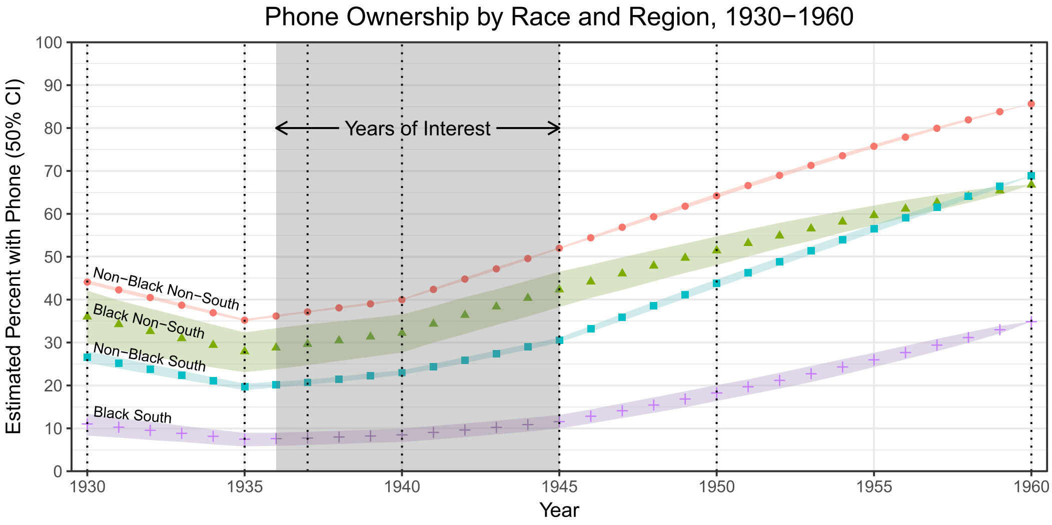


**Figure 4.** Phone ownership by race and region, 1930–60. Vertical dotted lines indicate years for which population data are available.

## 7 Conclusion

The basic approach outlined above can be applied in a variety of empirical settings. Our original motivation was creating dynamic population targets for creating survey weights, and we have used the model to create dynamic population targets for as many as 2,302 demographic types defined by the cross-classification of six variables over three decades (Caughey and Wang 2014). In addition to survey weights, such population targets may be used for multilevel regression and poststratification (Park, Gelman, and Bafumi 2004) or for generalizing causal effect estimates

(Hartman *et al.* 2015). Moreover, as Section 6 showed, the same model can be used to estimate conditional rates—the traditional focus of EI—in settings where the data structure is too irregular for conventional EI models.

The model itself could be extended and improved in various ways. One area for improvement is computational efficiency, which can become a problem as the number of cells grows. For example, in order to obtain satisfactory estimates for over 2,000 cells, we had to let Stan run for several weeks. It is possible that this problem could be overcome with approximate inference such as variational Bayes, or perhaps by reparameterizing the model so as to ease the computational difficulty of estimating proportions very close to 0. One natural alternative parameterization would be the logistic-normal distribution in place of the Dirichlet (Cargnoni, Muller, and West 1997). In addition to possible computational benefits, the logistic normal would allow for more flexible patterns of dependence across cells than the Dirichlet, whose assumption of independence across components may be undesirable in some applications.

## Supplementary material

For supplementary material accompanying this paper, please visit
https://doi.org/10.1017/pan.2019.4.

## References

Berinsky, A. J., E. N. Powell, E. Schickler, and I. B. Yohai. 2011. "Revisiting Public Opinion in the 1930s and 1940s." *PS: Political Science & Politics* 44(3):515–520.

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

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

Caughey, D., and M. Wang. 2014. "Bayesian Population Interpolation and Lasso-Based Target Selection in Survey Weighting." Paper presented at the Annual Meeting of the Society for Political Methodology, University of Georgia, Athens, GA, July 24.

Caughey, D., and M. Wang. 2018a. "`estsubpop`: Dynamic Subpopulation Estimation." R package version 1.0.

Caughey, D., and M. Wang. 2018b. "Replication Data for: Dynamic Ecological Inference for Time-Varying Population Distributions Based on Sparse, Irregular, and Noisy Marginal Data." doi:10.7910/DVN/YPMVMH, Harvard Dataverse.

Enns, P. K., and J. Koch. 2013. "Public Opinion in the U.S. States: 1956 to 2010." *State Politics & Policy Quarterly* 13(3):349–372.

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

Hartman, E., R. Grieve, R. Ramsahai, and J. S. Sekhon. 2015. "From Sample Average Treatment Effect to Population Average Treatment Effect on the Treated: Combining Experimental with Observational Studies to Estimate Population Treatment Effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(3):757–778.

Jackman, S. 2009. *Bayesian Analysis for the Social Sciences*. Chichester: Wiley.

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

Leeman, L., and F. Wasserfallen. 2017. "Extending the Use and Prediction Precision of Subnational Public Opinion Estimation." *American Journal of Political Science* 61(4):1003–1022.

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

Quinn, K. M. 2004. "Ecological Inference in the Presence of Temporal Dependence." In *Ecological Inference: New Methodological Strategies*, edited by G. King, O. Rosen, and M. A. Tanner, Chapter 9, 207–233. New York: Cambridge University Press.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek. 2017. "*Integrated Public Use Microdata Series: Version 7.0 [dataset]*." Minneapolis: University of Minnesota.

Stan Development Team. 2018. "`RStan`: the R interface to Stan." http://mc-stan.org/, R package version 2.17.3.