# USING ADMINISTRATIVE RECORDS TO PREDICT CENSUS DAY RESIDENCY

**Elizabeth A. Stuart*, Alan M. Zaslavsky, Harvard University**
**Elizabeth Stuart, Department of Statistics, 1 Oxford St., Cambridge, MA 02138**
**(stuart@stat.harvard.edu)**

**Key Words: Multiple system estimation, population, migration, hierarchical model.**

**Abstract:**

Administrative records are a promising data source for estimating Census coverage or identifying people missed in the Census. An important unsolved problem in using records is determining which of them correspond to people actually resident on Census day. We propose a hierarchical model in which one level describes the migration process, and the other describes the probabilities of observation in each of the available record systems. The observation model uses the full information in the records, including the dates associated with the records and available covariate information, and can accommodate a variety of record types, such as tax records, Medicare claims, and school enrollment lists. In addition, multiple record systems can be modeled concurrently. Posterior distributions of the in- and out-migration dates are obtained, leading to an estimate of the probability of residency in the area on Census day. This work could be useful in the context of an administrative records census, or as a way of expanding the role of administrative records in triple system estimation.

## 1. Introduction

This work utilizes administrative records to help predict census day residency. This is done using a Bayesian hierarchical model both of migration and of observation in each of the available record systems. This is useful in the context of an administrative records census, or as a way of expanding the use of administrative records in multiple system estimation.

This work has its basis in the methods of multiple system estimation. Multiple recapture estimation was originally developed as a way of estimating animal populations, but has found application in Census undercount estimation (Fienberg 1992), as well as a variety of other fields. To estimate the size of an animal population, one might capture a set of animals, mark them in some way, release them, and then make further captures at later points in time. The possible capture histories are represented by cells of a $2^k$ contingency table, where $k$ is the number of captures. One cell will be missing: the cell for individuals missed by all captures. The role of modeling is to estimate the size of this cell, thus estimating the total population size. The situation with two captures is known as dual system estimation, and similarly, that with three captures is called triple system estimation. See Darroch (1958), El-Khorazaty et al. (1977), Pollock (1991), and Seber (1982) for more information on multiple system estimation.

Early work in this field rested on a number of assumptions: the population is closed (no birth, death, or migration), the captures are independent, each individual has the same probability of capture, and individuals can be perfectly matched between captures.

Recent research has relaxed some of these assumptions. Loglinear models have been used to model the cell counts of the contingency table (Bishop et al. 1975, Fienberg 1972), allowing dependencies among captures. Unequal capture probabilities can be accommodated by calculating estimates by strata, or by a Rasch model (Fienberg et al. 1999). Bayesian methods have also been employed in this problem. In particular, George and Robert (1992) use a Gibbs sampling approach, while Smith (1991) compares Bayes, empirical Bayes, and Bayes empirical Bayes solutions.

Our approach is also related to the literature on the estimation of migration parameters for animal populations. There is a large literature on this topic, mostly as an outgrowth of the capture-recapture work. Most of these papers assume that several (usually 3-5) geographical areas have been defined and attempt simultaneously to estimate the population size and the transition probabilities among the areas. This is done by capturing animals in each location at several times, and recording where and when each animal is observed. Estimates of the total population size and the migration rates are then obtained. Much of this work involves modeling mi-

*Supported by a National Science Foundation Graduate Research Fellowship and by the US Bureau of the Census through a contract with the National Opinion Research Center and Datametrics Research Inc. Submitted for publication in the Proceedings of the 2001 Joint Statistical Meetings.

gration using Markov Chains (Brownie et al. 1993, Hestbeck et al. 1991). Dupuis (1995) provides a Bayesian approach.

In the context of the US Census, triple-system estimation has been suggested as a way to estimate the total population size. The three systems are usually taken to be the Census itself, the Post-Enumeration Survey (PES), and a series of administrative lists. There has been extensive research on the use of administrative records in the Census, for triple system estimation as well as other potential uses. Zaslavsky and Wolfgang (1993) discuss the details of using triple system estimation for the Census, and use 1988 Census dress rehearsal data to develop specific models. Larsen (1999) took a model based approach to use records to identify likely Census day residents. Logistic regression was used to develop criteria to determine likely residents, but the model did not attempt to model behavior. Zanutto and Zaslavsky (2001) use administrative records to impute for nonresponse, with both a model based and a non-model based approach. Beyond the United States, Redfern (1989) discusses the use of administrative records (and specifically, population registers) in European countries, as well as the political issues involved. Countries such as Denmark and Finland currently use elements of a register-based census, reducing both cost and respondent burden.

Scheuren (1999) gives an overview of the use of administrative records in the US Census, particularly a proposal for an administrative records census. This could reduce Census cost, provide more frequent population counts, and improve the coverage rates of populations traditionally undercounted. However, he stresses that there are many research questions still to be worked out regarding the use of administrative records. The Census Bureau's AREX 2000 experiment, and ongoing evaluations, are examining the use of administrative records as a primary source of information.

One of the drawbacks of the use of administrative records is that their coverage period does not coincide with Census day, and may extend considerably earlier. We therefore develop a model of migration that allows prediction of whether someone is still a resident on Census day, given that she appears in one or more record systems. If the administrative records are available nationally, the model can also be used to facilitate small area undercount estimation across the country.

## 2. Overview of Model

We propose a hierarchical model in which one level describes the migration process, and the other describes the probabilities of observation in each of the available record systems. The observation model uses the full information in the records, including the dates associated with the records and available covariate information, and accommodates a variety of record types, such as tax records, Medicare claims, and school enrollment lists. In addition, multiple record systems can be modeled concurrently. The posterior distributions of the in- and out-migration dates are obtained, leading to an estimate of the probability of residency in the area on Census day for each individual.

Suppose we have a series of record systems (types of administrative records, possibly a Census and/or a PES) from a geographic area. Each record is dated, providing evidence of a person being a resident in the area on that date. The total time period covered is $T_0$ to $T_1$. Define a population consisting of all people living in this area at some point during this time interval who were captured by at least one of the systems. A later version of the model will allow for individuals who were missed by all of the systems by imputing missing individuals.

We are interested in modeling the in- and out-migration times from the area: $t_{0i}$ (the time person $i$ moved in) and $t_{1i}$ (the time person $i$ moved out). The goal of the inference is to estimate the size of the population at a particular point in time, usually Census day.

Our hierarchical model has 3 levels:

Level 1 (Observational):

$P$(observation history|migration dates, covariates, parameters)

Level 2 (Migration):

$P$(migration dates|covariates, parameters)

Level 3: Priors on the parameters

Level 1 models each individual's observation in the record systems. Under the assumption of independence, the likelihoods of observation in each of the systems are multiplied together to obtain the full observation likelihood. This assumption can be modified, as discussed in Section 3.3. Level 2 describes the migration history for each individual: the in- and out-migration dates. These migration events are observed through the observation history in Level 1. Level 3 describes prior beliefs about the parameters,

either fixing them at pre-specified values, or specifying non-degenerate prior distributions.

## 3. Details of Model

In this section we present specific examples for the models at each level. More complex models can also be specified within this overall structure. In the notation that follows, $\Psi$ represents the vector consisting of all of the Level 3 parameters. Specific components of $\Psi$ are described later.

### 3.1 Migration Model

Level 2 describes the migration of the individuals, i.e. the time when the individual resided in the area. Each individual's migration history is summarized by two variables: $t_{0i}$, the time person $i$ moved into the area, and $t_{1i}$, the time person $i$ moved out of the area.

We model the population as a mixture of two types of people: never-movers, who never move in or out of the area, and movers, who migrate to or from the area (although not necessarily during the time period of observation). The in- and out-migration dates are modeled using mixture distributions to account for the two types of individuals. The parameter $r$ represents the fraction of never-movers in the area at a given point in time (considered to be constant across time). For the movers, we assume a stationary process with a constant hazard of moving ($\lambda$) that is the same for each individual.

The version of the model simulated in this paper (Section 5) assumes that $r = 0$ (there are no never-movers). From these assumptions we can deduce $q$, the proportion of the population that was in the area at the beginning of the time period of interest: $q = \frac{1}{1+\lambda(T_1 - T_0)}$. This model implies a censored exponential distribution for the length of residency and a mixture for $t_{0i}$, with a mass $q$ at $T_0$ and a uniform distribution over the remaining time, to $T_1$.

### 3.2 General Observation Model

The observation model (Level 1) describes the process of observing the individuals in the record systems. The migration history is observed through these record systems, as each person's opportunity to be observed depends on their migration history.

We first give the general framework for a single record system, and then discuss methods of combining observations from multiple systems. Sections 3.4.1 through 3.4.4 provide examples of specific record systems.

A generic approach has one indicator variable for whether an individual was in that record system type (if she filed a tax return, had a driver's license, etc.). Another variable indicates the date when she would appear. The interaction of these and the migration dates then determines whether the individual would be observed in the record system file available. The exact interpretation of these variables is specific to each record system.

Let $j$ index the type of record, and $i$ index individuals. The following variables are defined for each of the record systems ($j = 1, ..., J$):

$T_{0j}$ = Beginning of time period covered by record type $j$.

$T_{1j}$ = End of time period covered by record type $j$.

$w_{ji}$ = Bernoulli variable indicating that person $i$ has a record of type $j$.

$$w_{ji}|\alpha_{ji} \sim \text{Bernoulli}(\alpha_{ji} = \alpha_j(x_i, \Psi))$$

$\alpha_{ji}$ represents the probability of individual $i$ having record type $j$, and may depend on individual covariates $x_i$ through some kind of regression model.

$y_{ji}$ = Date associated with record type $j$ for individual $i$.

$$y_{ji} \sim F_j(x_i, \Psi)$$

The distribution of $y_{ji}$ depends on the type of record, and possibly covariates $x_i$.

$z_{ji}$ = Indicator for individual $i$ being observed in file $j$.

$$z_{ji} = Z_j\left(w_{ji}, t_{0i}, t_{1i}, y_{ji}, T_{0j}, T_{1j}\right)$$

$z_{ji}$ is a function of $w_{ji}, y_{ji}$, migration dates, and the dates covered by record system $j$.

Define $T_0 = \min\{T_{0j}\}$, the beginning of the time period covered by any source, and $T_1 = \max\{T_{1j}\}$, the end of the time period covered by any source. This notation for the observation model can accommodate a variety of record systems, including administrative records files, the Census, and the PES. The framework stays the same, but the specifics of the distributions depend on the type of record system.

### 3.3 Combining observation models from multiple record systems

Under the assumption that being in a record system ($w_{ji}$) and the date associated with that ($y_{ji}$) are independent across systems, conditional on migration

dates, the probability distributions of $(w_{ji}, y_{ji})$ for each system are jointly independent. In this case, the full observation model is just the product of the individual record observation likelihoods:

$$L(z|\Psi) \propto \prod_i \left[ \prod_j P(w_{ji}|\alpha_{ji}) P(y_{ji}|w_{ji}, \alpha_{ji}) \cdot P(z_{ji}|y_{ji}, w_{ji}, t_{0i}, t_{1i}, T_{0j}, T_{1j}) \right] \cdot P(t_{0i}, t_{1i}|\lambda, q).$$

Although independence of the systems is a fairly common assumption in multiple system estimation, many studies have shown that it is not a good approximation for administrative records. To model dependence among the $w_{ji}$'s, loglinear models could be utilized. Dependence among the $y_{ji}$'s (for example, if both driver's license renewal dates and car registration files were linked to an individual's birthday) could be modeled directly. The specifics would depend on the exact records involved.

### 3.4 Specific observational models

The following are specific examples of the observation model.

#### 3.4.1 Census

In the case of the Census, $w_{Ci} \sim$ Bernoulli$(\alpha_{Ci})$ for all $i$, where $\alpha_{Ci}$ depends on each individual's characteristics, as well as the undercount rate. Since the Census records nominally cover just one day, $y_{Ci} =$ April 1 for everyone ($y_{Ci} = y_C$ for all $i$), and $T_{0C}$ and $T_{1C}$ are both April 1. The function for $z_{Ci}$ is then $z_{Ci} = 1$ if $w_{Ci} = 1, t_{0i} \leq$ April $1 \leq t_{1i}$, and $z_{Ci} = 0$ otherwise.

#### 3.4.2 Tax Returns

For tax returns, $w_{Ti} \sim$ Bernoulli$(\alpha_{Ti})$, where $\alpha_{Ti}$ represents the probability that someone with person $i$'s characteristics files a tax return. This may depend on characteristics such as age or region of the country. Since tax returns are generally filed around April 15, the distribution of $y_{Ti}$ is centered around April 15, with a distribution of early and late filers. Given the plentiful tax data, a non-parametric estimate of the distribution of filing dates $(y_{Ti})$ could be utilized. $T_{0T}$ is the beginning of the time period covered by the file, and $T_{1T}$ is the end date of the period covered by the file. The function for $z_{Ti}$ is then $z_{Ti} = 1$ if $w_{Ti} = 1, t_{0i} \leq y_{Ti} \leq t_{1i}, T_{0T} \leq y_{Ti} \leq T_{1T}$, and $z_{Ti} = 0$ otherwise.

#### 3.4.3 Driver's Licenses

Although driver's licenses are unlikely to be used as a record system in the Census context because of the complications of disparate state laws and data files, they are a good, intuitive example of how the model works. In this case, $w_{Di} \sim$ Bernoulli$(\alpha_{Di})$, where $\alpha_{Di}$ represents the probability that someone with person $i$'s characteristics has a driver's license and could depend on personal characteristics (in particular, age) and location in the country. Since most driver's licenses are renewed at fixed intervals of some number of years, typically on the individual's birthday, we assume that the distribution of renewal dates, $y_{Di}$, is uniform. Since we are only concerned with the most recent renewal, the right endpoint of this distribution is $T_{1D}$ (the endpoint of our observation interval), and the left endpoint is $T_{1D} - R$, where $R$ is the length of time between renewals. We then assume that anyone in the area with a driver's license would have had to renew their license at some point in this interval. The function for $z_{Di}$ is then $z_{Di} = 1$ if $w_{Di} = 1, t_{0i} \leq y_{Di} \leq t_{1i}, T_{0D} \leq y_{Di} \leq T_{1D}$, and $z_{Di} = 0$ otherwise.

#### 3.4.4 Other Types of Records

Other types of records that could be modeled in this way include the Social Security Service's Master Beneficiary Record, which is a list of anyone entitled to Social Security Benefits, updated monthly. Each individual has a probability of being a beneficiary in each month, and his observation date would be modeled as uniform through the month. The monthly files could give us fairly precise information on when individuals moved to or from the area.

Models for Medicare claims would be more complex. We can estimate the probability that an individual is a Medicare fee-for-service beneficiary. The temporal distribution of claims is more complicated since some people will have many claims in a short time period, while others may have claims very spread out.

## 4. Inference

### 4.1 Levels of Inference

The structure of the hierarchical model allows inference on each of the 3 levels: global parameters of coverage probabilities and migration, individual migration times, and individual observation and record histories. The level of inference will depend on the goal. For example, inference about the global migration parameters may be of interest to sociologists

interested in studying migration patterns. This flexibility of levels of inference enables the model to be useful for a variety of purposes.

In the Census context, we are mostly interested in inference on the second level, regarding the migration dates for individuals. It is possible to obtain posterior estimates of individual's migration dates, which lead to estimates of the probability of residency, and in turn lead to an estimate of population size on Census day. An example of this is given in Section 5.

### 4.2 Computational Methods

Draws from the joint posterior are obtained by running a Gibbs sampler, which iteratively draws from each of the full conditional posterior distributions and converges to the joint posterior (Geman and Geman 1984). The general framework is that of cycling through the three levels, drawing the parameters at each level. Here we present the specifics for the model as described in the simulation presented in Section 5. Discussion is restricted to the distributions necessary for the simulation, which includes two types of systems: the Census and driver's licenses. The priors used are $\alpha_C \sim \text{Beta}(a_{\alpha_C}, b_{\alpha_C})$ and $\alpha_D \sim \text{Beta}(a_{\alpha_D}, b_{\alpha_D})$. In addition, $q$ and $\lambda$ are considered known and so are not drawn in the scenario discussed.

Define the full parameter vector

$$\Theta = \{\{t_{0i}\}, \{t_{1i}\}, \{w_{Ci}\}, \{w_{Di}\}, \{z_{Ci}\}, \{z_{Di}\}, \{y_{Ci}\}, \{y_{Di}\}, \lambda, q, \alpha_C, \alpha_D\}.$$

The Gibbs sampler iterates through the following steps:

1. Global Parameters

   (a) $\alpha_C|\Theta_{\setminus \alpha_C} \propto$

   $$\frac{\alpha_C^{n_{11}+n_{21}+a_{\alpha_C}-1}(1-\alpha_C)^{n_{12}+b_{\alpha_C}-1}}{(\alpha_C+\alpha_D-\alpha_C\alpha_D)^n}$$

   (b) $\alpha_D|\Theta_{\setminus \alpha_D} \propto$

   $$\frac{\alpha_D^{n_{11}+n_{12}+a_{\alpha_D}-1}(1-\alpha_D)^{n_{21}+b_{\alpha_D}-1}}{(\alpha_C+\alpha_D-\alpha_C\alpha_D)^n}$$

   In this simulation, $\lambda$ and $q$ are considered known, and so draws from their posterior distributions are not necessary. The known variable $n_{11}$ represents the number of individuals caught by both systems, $n_{12}$ is the number of individuals caught by the driver's license file and missed by the Census, and $n_{21}$ is the number of individuals caught by the Census and not by the driver's license file. Finally, $n = n_{11} + n_{12} + n_{21}$ is the total number of individuals caught by any source. The posterior distributions of $\alpha_C$ and $\alpha_D$ have a form similar to that of a binomial distribution, with a modification to the denominator. The explanation for the denominator is easily seen if we consider the 2x2 table formed by the interaction of $w_{Ci}$ and $w_{Di}$. Since we only consider individuals who were in at least one of the systems, the cell $w_{Ci} = 0, w_{Di} = 0$ is not in the model and thus the sum of the probabilities of being in each of the cells is not 1. This sum is thus in the denominator. Since the posteriors of $\alpha_C$ and $\alpha_D$ are not in closed form, Metropolis-Hastings algorithms were used to obtain posterior draws from them (Gelman et al. 1995, Chapter 11). A Uniform jumping distribution was used, and acceptance rates were in the range recommended by Gelman et al. (1996).

2. Individual Migration Parameters

   (a) $t_{0i}|\Theta_{\setminus t_{0i}} \sim \lambda e^{\lambda t_{0i}} \left(q\left(\delta(t_{0i} = T_0)\right) + (1-q)(\delta(T_0 < t_{0i} \leq T_1))\right) \cdot I\{t_{0i}^L \leq t_{0i} \leq t_{0i}^U\}$

   - $t_{0i}^L$ and $t_{0i}^U$ are bounds on $t_{0i}$, determined by the set of records observed

   (b) $t_{1i} - t_{0i}|\Theta_{\setminus t_{1i}} \sim \text{Exp}(\lambda) I\{t_{1i}^L \leq t_{1i} \leq t_{1i}^U\}$

   - $t_{1i}^L$ and $t_{1i}^U$ are bounds on $t_{1i}$, determined by the set of records observed

3. Individual Observation Parameters

   (a) $y_{Ci}|\Theta_{\setminus y_{Ci}} = T_{1C}$

   (b) $y_{Di}|\Theta_{\setminus y_{Di}} \sim \text{Uniform}(y_{Di}^L, y_{Di}^U)$

   - $y_{Di}^L$ and $y_{Di}^U$ are determined by the set of records observed

   (c) $w_{Ci}|\Theta_{\setminus w_{Ci}} \sim \text{Bernoulli}(\alpha_C)$, unless determined by observation history

   - $w_{Ci} = 1$ if $z_{Ci} = 1$
   - $w_{Ci} = 0$ if $z_{Ci} = 0, t_{0i} < T_{1C} < t_{1i}$

   (d) $w_{Di}|\Theta_{\setminus w_{Di}} \sim \text{Bernoulli}(\alpha_D)$, unless determined by observation history

   - $w_{Di} = 1$ if $z_{Di} = 1$
   - $w_{Di} = 0$ if $z_{Di} = 0, t_{0i} < y_{Di} < t_{1i}$

The ranges of possible values for $t_{0i}$, $t_{1i}$ and $y_{Di}$ are determined by the records observed for each individual and the current values of the other parameters. Depending on the dates of observation, the posterior distribution of moving dates might be diffuse across the entire observation period, or might be more limited. For example, an individual with a driver's license observed on day 25 and also observed on Census day (day $T_{1C}$) has a moving out date that is more constrained than that for someone observed only on day 25.

This leads to complications in the computations, as each individual has a different range of possible values of $t_{0i}$, $t_{1i}$ and $y_{Di}$ given the observation dates, $w_{Ci}$, $w_{Di}$, and the migration parameters. The posteriors for the moving dates and $y_{Di}$ look like the priors, but are either restricted or unrestricted due to the observation dates. Many of the steps in the Gibbs sampler thus consist of a set of cases depending on values of $z_{Ci}$, $z_{Di}$, $w_{Ci}$, and $w_{Di}$. Examples of the types of observed data and the consequences for the ranges of $t_{0i}$, $t_{1i}$, and $y_{Di}$ are given below. Some break down into cases based on the current draws of the parameters.

1. Observed in driver's license file on day $y_{Di}$ and in the Census on day $y_{Ci} = T_{1C}$.
   For this individual, we know that $t_{0i} < y_{Di}$, $t_{1i} > T_{1C}$, $w_{Ci} = 1$ and $w_{Di} = 1$.

2. Not observed in driver's license file, observed in Census on day $y_{Ci} = T_{1C}$.
   The possible ranges of $t_{0i}$ and $t_{1i}$ depend on the current value of $w_{Di}$.

   a: $w_{Di} = 0$: Since the individual does not have a driver's license, her absence from the file tells us nothing about her migration history. We thus only know that $t_{0i} < T_{1C} < t_{1i}$.

   b: $w_{Di} = 1$: The individual renewed her driver's license, but not during the time that she was in the area. For a given value of $y_{Di}$ (drawn from its posterior distribution), we know that the individual must have either moved in after $y_{Di}$ or out before $y_{Di}$. This restricts the possible values of $t_{0i}$ and $t_{1i}$.

3. Observed in driver's license file on $y_{Di}$, not in Census file.
   Again, there are two cases, depending on the current value of $w_{Ci}$.

   a: $w_{Ci} = 0$: Since the person is not in any Census record, this gives us no information on the individual's migration history. We only know that $t_{0i} < y_{Di}$ and $t_{1i} > y_{Di}$.

   b: $w_{Ci} = 1$: This implies that the person is in the Census, but was not in the area of interest on Census day. Since our population is defined as anyone in the area at some point between $T_0$ and $T_1$, we thus know that the individual must have moved into the area before $y_{Di}$ ($t_{0i} < y_{Di}$), and out of the area after $y_{Di}$ but before Census day ($y_{Di} < t_{1i} < T_{1C}$).

## 5. Simulation

### 5.1 Simulation Parameters

We assume that two systems are available: a file of driver's license records and the Census. We assume that being in the Census file ($w_{Ci}$) is independent of being in the driver's license file ($w_{Di}$) and thus the full observation likelihood is the product of the likelihoods of being observed in each of the two systems. The observation period starts at $T_0 = 0$ and ends at $T_1 = 365$ (measured in days).

Census day is at the end of this time period, day 365. The observation model for the Census is described in Section 3.4.1. We assume that $\alpha_{Ci} = \alpha_C$ for all $i$, implying that everyone has the same probability of being in the Census. The observation model for the driver's licenses is described in Section 3.4.3. Again, we assume that $\alpha_{Di} = \alpha_D$ for all $i$. We use a renewal period of one year ($R = 365$) and have driver's license file coverage of one year, ending at Census day. The distribution of the most recent renewal date is thus approximated as Uniform$(0, 365)$. This set-up gives us more information on the migration dates and file coverage. If someone is not observed in the driver's license file, we know that it is either because she did not have a license ($w_{Di} = 0$) or because she was not in the area at the time of renewal ($y_{Di} < t_{0i}$ or $y_{Di} > t_{1i}$).

The migration model is that described in Section 3.1, with a mixture model for $t_{0i}$ and an exponential distribution for the time before moving out. To simplify this example, $q$ and $\lambda$ are assumed to be known. The values are $q = .8$ and $\lambda = \frac{1}{1825}$, which correspond to an average duration of stay of 5 years (Hansen 1998). As discussed earlier, we assume that there are no non-movers in the population.

At Level 3, the prior distributions on $\alpha_C$ and $\alpha_D$ are Beta$(1, 1)$, which are non-informative conjugate priors.

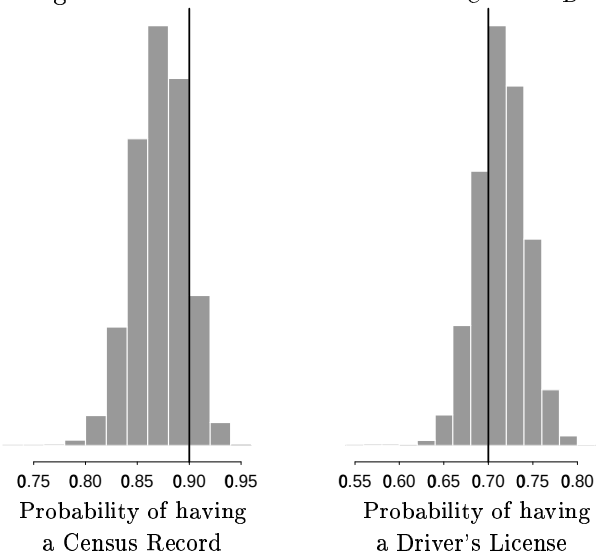
### 5.2 Results

The simulated data set consisted only of people observed in one or both of the systems, resulting in a sample size of 427. The "true" parameter values are shown in Table 1, as well as posterior estimates from the Gibbs sampler. The variable $N_C$ is the size of the population on Census day. Histograms of the posterior distributions of $\alpha_C$ and $\alpha_D$ are shown in Figure 1, and of $N_C$ in Figure 2. The vertical bar in each plot represents the "true" value in the simulation.

Table 1: Posterior Estimates of Parameter Values

| Parameter | True Value | Posterior Mean | 95% Posterior Interval |
|---|---|---|---|
| $q$ | 0.8 | NA | NA |
| $\lambda$ | $\frac{1}{1825}$ | NA | NA |
| $\alpha_C$ | 0.9 | 0.87 | (0.82, 0.92) |
| $\alpha_D$ | 0.7 | 0.71 | (0.66, 0.77) |
| $N_C$ | 408 | 407 | (397, 417) |

Figure 1: Posterior Distributions of $\alpha_C$ and $\alpha_D$




The Census file had 370 individuals observed on Census day. The addition of one record system, the driver's license file, added 37 individuals to the Census day population count. In addition, the posterior intervals for the three main parameters and the population size on Census day covered the true values. Sensitivity to starting values and priors was checked and all iterations converged to similar values.

Since our goal is to determine the probability of residency on Census day, we are primarily interested in the individual migration dates, $t_{0i}$ and $t_{1i}$, and their implications regarding residency on Census day. The main inference will be for individuals observed in the driver's license file and not in the Census. Figure 3 shows the predicted probability of Census day residency for an individual observed in the driver's license file at various points in time, but not in the Census. As might be expected, individuals observed later are more likely to still be in the area on Census day.

Figure 2: Posterior Distribution of Census Day Population Size

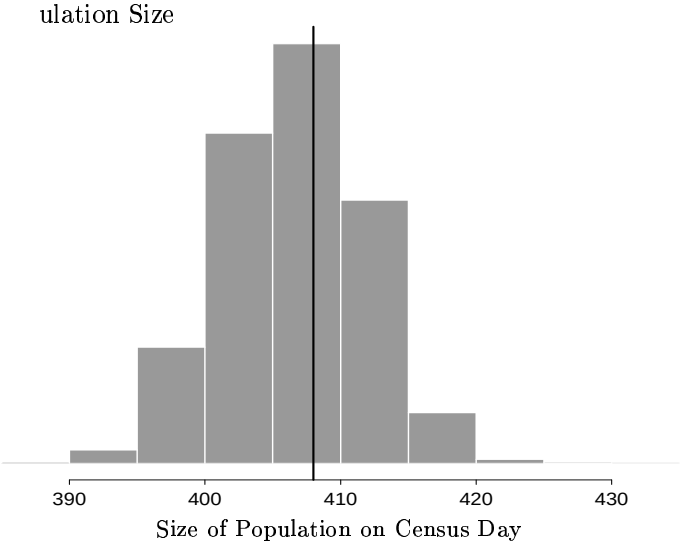
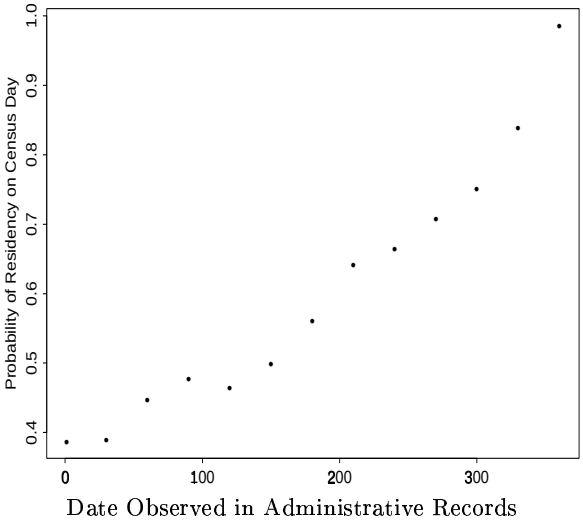



Figure 3: Probability of Residency on Census Day for those observed only in Administrative Records



## 6. Discussion

### 6.1 Extensions

The strength of this model lies in its flexibility, specifically its ability to model many types of records at the same time, either through an independence model by multiplying the likelihoods of observation of each type of record, or through a more complex model. This could be of particular use with the new Census Bureau StARS data set, which contains data from 5 different records systems.

Heterogeneity of capture probabilities associated with observable characteristics is incorporated by including covariates in the Level 1 (observation level) models, as in Alho, Mulry, Wurdeman, and Kim (1993). Unobservable heterogeneity may be modeled through common random effects affecting the probabilities of an individual being observed in each of several systems (Darroch et al. 1993). Alternatively, the joint distribution of observation in all of the record systems could be modeled directly.

Heterogeneity in the probability of moving can be incorporated by adding covariates such as demographic or area characteristics to the migration model, as can seasonality in migration.

The full model includes $N$, the total number of individuals in the population, including those unobserved. This parameter is not needed in the current version of the model as it enters only through the distribution of $\lambda$, which is considered known for this simulation. A larger model that includes $N$ will be developed. To calculate the population size on Census day, we will use the probability of those we observe of being resident on Census day, as well as weights to represent unobserved individuals.

### 6.2 Applications

Administrative records, and this model, have great potential to assist in the estimation of the undercount of the US Census. There are at least two scenarios regarding the design of a national administrative records sample. In the first scenario, the Post-Enumeration Survey and the administrative records cover the same blocks. This leads to triple system estimation (Census, Post-Enumeration Survey, and administrative records) for the blocks where the Post-Enumeration Survey and the administrative records are available. These results would then be used to adjust counts across the country.

Under a second scenario, the administrative records are available across the entire country, not just in the Post-Enumeration Survey blocks. The records would be matched to each other and to the Census and then used to provide small area population estimates across the country. The Post-Enumeration Survey would then be used in estimation of general parameters. Although this design requires assembling much larger files, the cost would not be proportionally more than that required to obtain administrative records files for just the Post-Enumeration Survey blocks since the same systems must be accessed.

There are three main advantages to using administrative records as a second national source of information on individuals. The records could be used to add (or subtract) people for whom we have direct evidence that they were (or were not) in the area on Census day. The PES can do this as well, but provides small-area detail only for sample blocks. Hence, estimates would rely less on synthetic estimates of the undercount; fewer assumptions of homogeneity across areas would be necessary and local undercount estimates could be obtained more reliably. Finally, this would be a major step forward in the use of administrative records in the Census. The StARS database currently under development and results of modeling with it, as well as the corresponding AREX experiment in the 2000 Census, should give some indication of the potential for this method.

The migration model described here could reduce some of the problems associated with the use of administrative records. In particular, it could help reduce the amount of field follow-up needed, as it could identify the people that were more or less likely to still be in the area on Census day. Finally, the model may be useful to deal with movers in the PES. In that case, we would observe an individual on a date after Census day, and use the model "backwards" to predict residency on Census day.

In addition, there is potential for the use of this model in fields such as demography and sociology, where human migration is a major research area. The model can be extended to describe a list of events for individuals, jointly with their movement patterns. It also might be used to help identify the determinants of migration.

## References

Alho, J.M., Mulry, M.H., Wurdeman, K., and Kim, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association* 88: 1130–1136.

Bishop, Y.M.M, Fienberg, S.E., and Holland, P.H.

(1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.

Brownie, C., Hines, J.E., Nichols, J.D., Pollock, K.H., and Hestbeck, J.B. (1993). Capture-recapture studies for multiple strata including non-Markovian transitions. *Biometrics* 49: 1173–1187.

Darroch, J.N. (1958). The multiple-recapture census: I. Estimation of a closed population. *Biometrika* 45: 343–359.

Darroch, J.N., Fienberg, S.E., Glonek, G.F.V., and Junker, B.W. (1993). A three-sample multiple-recapture approach to Census population estimation With heterogeneous catchability. *Journal of the American Statistical Association* 88: 1137–1148.

Dupuis, J.A. (1995). Bayesian estimation of movement and survival probabilities from capture-recapture data. *Biometrika* 82: 761–772.

El-Khorazaty, M.N., Imrey, P.B., Koch, G.G., and Wells, H.B. (1977). Estimating the total number of events with data from multiple-record systems: A Review of methodological strategies. *International Statistical Review* 45: 129–157.

Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete $2^k$ contingency tables. *Biometrika* 59: 591–603.

Fienberg, S.E. (1992). Bibliography on capture-recapture modeling with application to Census undercount adjustment. *Survey Methodology* 18: 143–154.

Fienberg, S.E., Johnson, M.S., and Junker, B.W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society A* 162: 383–405.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* New York: Chapman and Hall.

Gelman, A., Roberts, G.O., and Gilks, W.R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics 5*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, 599–607. London: Oxford University Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.

George, E.I. and Robert, C.P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika* 79: 677–683.

Hansen, K.A. (1998). Seasonality of moves and duration of residence. US Census Bureau Current Population Reports, Household Economic Studies. P70-66.

Hestbeck, J.B., Nichols, J.D., and Malecki, R.A. (1991). Estimates of movement and site fidelity using mark-resight data of wintering Canada geese. *Ecology* 72:523–533.

Larsen, M. (1999). Predicting the residency status for administrative records that do not match Census records. US Census Bureau Administrative Records Research Memorandum Series #20.

Pollock, K.H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future. *Journal of the American Statistical Association* 86:225–238.

Redfern, P. (1989). European experience of using administrative data for censuses of population: The policy issues that must be addressed. *Survey Methodology* 15: 83–99.

Scheuren, F. (1999). Administrative records and census taking. *Survey Methodology* 25: 151–160.

Seber, G.A.F. (1982). *The estimation of animal abundance and related parameters, 2nd edition.* New York: MacMillan Publishing Co., Inc.

Smith, P.J. (1991). Bayesian analyses for a multiple capture-recapture model. *Biometrika* 78: 399-407.

Zanutto, E. and Zaslavsky, A.M. (2001). Using administrative records to impute for nonresponse. In *Survey Nonresponse*, ed. R. Groves, D. Dillman, J. Eltinge, and R. Little. New York: John Wiley and Sons. In press.

Zaslavsky, A.M, and Wolfgang, G.S. (1993). Triple-system modeling of Census, Post-Enumeration Survey, and administrative-list data. *Journal of Business & Economic Statistics* 11: 279–288.