# Modeling Frame Deficiencies for Improved Calibrations

Eric V. Slud *

**Abstract**

One of the important reasons advanced for calibrating household sample surveys to population totals is the imperfect coverage of target frame populations by operational frame lists. The theoretical justification for design-consistency based on calibration depends on the assumption that the calibration-variable population totals are correct. When the frame list has perceptible differences from the target population, the calibrated weight-adjustment is no longer design consistent, but models of the frame may still allow consistent modified calibration estimates. Such models can be created if the periodically available data used to update frame lists are viewed as time-dependent covariates driving the rates of unit deletion from and addition to the frame as a time-varying random process. This paper proposes such a model and discusses how it might be used in the context of the Census Bureau's Master Address File.

**Key Words:** Calibration; Frame errors; Markov model; Master Address File; Survey frame; Time-dependent covariates

**Disclaimer.** *This report is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed on statistical, methodological or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.*

## 1. Introduction

The Master Address File (MAF) is the continuously maintained list of dwelling-unit addresses used by the US Census Bureau in its two largest data collection efforts, the Decennial Census and the continuously operating American Community Survey (ACS). It is the list from which working frames for the census and the ACS are constructed (as "extracts"), and in the future will serve as the basis for frame construction in several of the most important sample surveys conducted by the Census Bureau including the Current Population Survey, Survey of Income and Program Participation, and American Housing Survey. The MAF is repeatedly updated based on information periodically supplied by the US Post Office (the Delivery Sequence Files) along with the post-2000 Census Demographic Area Address Listing, surveys of new construction, and address canvassing for the decennial census. "Address canvassing" refers to the massive program, last executed in 2009, in which Census Bureau workers compare what they can see on the ground concerning all actual or potential dwelling units with the MAF's current listings.

The elements of the MAF, or MAFID's, correspond to unique addresses indexed ("geocoded") to standard Census Bureau geographic units such as the census block. The MAFID's are intended to cover all unique locations (in the US and Puerto Rico) of potentially residential structures. They are identified as group quarters or housing units or 'invalid addresses' (business structures or buildings where no one could live), but MAFID's often have a longer life than buildings, e.g., when a house is demolished but a new one is built on the same lot. Therefore, it makes sense to view map locations with MAFID's as the basic units of the MAF whose status evolves over time.

As part of the current 2020 Census research effort, Census Bureau researchers (Young et al. 2014, Tomaszewski and Boies 2014) are developing predictive models for numbers

*Census Bureau, Center for Statistical Research & Methodology, 4600 Silver Hill Road Rm 5K004, Washington DC 20233-9100, and Mathematics Dept., Univ. of Maryland, College Park, MD 20742

of adds and deletes to MAF based on data from the last round of address canvassing conducted in preparation for the 2010 Census. The intended application of such models is to the targeting of address canvassing for the 2020 Census, potentially allowing address canvassing to be done at a much-reduced cost by omitting blocks with the lowest expectation of producing MAF changes by canvassing. The predictive models of Young et al. (2014) for block-level counts of adds and deletes, which will be discussed in greater detail in Section 3 below, are based on explanatory variables extracted from several sources of data assembled as part of the 2009 address canvassing operation, including the US Postal Service Delivery Sequence Files (DSF) available every 6 months, and the Demographic Area Address Listing (DAAL) program following the 2000 Census. These models, which took the form of zero-inflated generalized linear models, made use only of predictive covariates aggregated from housing units up to the level of census blocks. They represent data from a snapshot of MAF changes related to a temporally focused round of address canvassing associated with the decennial census, although many of the data sources are and will continue to be recurrent, available periodically.

Several data streams serve the Census Bureau in its efforts to update the MAF housing inventory for ongoing intercensal surveys such as the ACS. These include the following specific sources: (i) the Postal Service DSF, generated twice each year, containing many unit-level status variables for mail delivery, occupancy, and address stability; (ii) geographically indexed new-construction estimates provided by the Census Bureau's Manufacturing and Construction Division – non-mobile homes based on the Survey of Construction, and mobile home totals based on the Survey of Mobile Home Placements; (iii) the blockgroup-level Planning Data Base (PDB), containing neighborhood demographics and census and ACS characteristics; (iv) occasional data from Address Canvassing (in preparation for decennial censuses), observed (but not without error) as described in the Johnson & Kephart (2013) Census evaluation report; and (v) other data assembled and codified by Geography Division, such as the classification of blockgroups into Urban and Rural categories based on population density, on proximity by map-coordinates and roads to urban areas, and on aerial imagery of pavement surfaces and terrain. It should be mentioned that a significant proportion of the new addresses provided by DSF require time and additional effort to be geo-coded, an operation of the Census Bureau's Geography Division that imposes time-lags on MAF updating well beyond the 6 months between successive DSF releases.

The purpose of listing these data sources for MAF updates is to emphasize that they could be used to support a regular time-evolving database through which predictive models for the MAF might be derived. Such a database could be used not only in Targeted Address Canvassing — the application envisioned in 2020 Census research — but in forecasting and estimating the likely discrepancies over time between the MAF listing and the true status on the ground of potentially residential structures in the US.

The goal of this paper is to describe the way in which a time-dependent database of housing-unit and area-level covariates could be used to estimate frame errors based on the MAF. The overall perspective advanced here is that the MAF is necessarily a stochastic process, that is, a listing of units whose individual status varies nondeterministically over time, but evolves through mechanisms that can be described through regression relationships in terms of a sufficiently detailed set of time-dependent covariates. We begin, in Section 2, by distinguishing the true target population of residential housing structures, which we call the *ideal frame*, from the *working frame* of MAF entries or of extracts (for the purpose of a particular survey) from which designed survey probability samples are drawn. We briefly argue in Section 2 that in survey analyses which make use of calibration or population controls – and this is true of essentially all Census Bureau surveys – the discrepancies between ideal and working frames enter survey estimates primarily but not exclusively through er-

rors in the frame totals of calibration variables. In Section 3, we describe the data inputs and statistical form of the single-time-point models of Young et al. (2014) currently under study for Targeted Address Canvassing and their relation to latent state models which explicitly take time dynamics into account. Section 4 introduces the class of Markovian models with exogenous time-dependent covariates that we propose should be used to describe the time-evolving MAF, and discusses how to define the underlying Markov-process state-space. In Section 5, we show how Markovian models like those of Section 4 would lead to predicted frame transition rates and therefore also to predictions of numbers of transitions (adds and deletes at unit level, aggregated to suitably chosen geographic areas). Some remarks about the fitting of the models in Section 4 are also made in Section 5, but unless the database of time-dependent covariates is available at unit level, this leads to some partially unsolved technical problems. Finally, we conclude in Section 6 with a summary of our proposal for a time-evolving and regularly updated US unit-level housing database and its application to the estimation of survey frame errors, a proposal which naturally leads to suggestions for further research.

## 2. Ideal versus Working Frames: the Role of Calibration

Calibration, raking or "generalized raking" are terms used to refer to the most important ways in which statisticians analyzing the results of a survey attempt to correct for possible errors of frame coverage (Deville and Särndal 1992, Fuller 2009). The idea is that one collects survey data including not only the outcome variable $Y_i$ of interest, for sampled and responding units $i$, but also some auxiliary variables $X_i$. The auxiliary variables here are assumed to be vectors of unit-level measurements known or observed for all sampled units (whether they respond to the survey or not) for which the target population totals are also known. Since the target population $U^*$ is generally not precisely the same as the working frame list $U$ from which the probability sample is drawn, we must distinguish these two underlying frame populations.

The probability sampling design of a survey generates inverse inclusion probability (*base* or *design*) weights $d_i = 1/\pi_i$, where $\pi_i$ is the probability that unit $i$ in the working frame $U$ is included in the sample, where the sample is denoted $S \subset U$. Denote by $n$ the number of elements of $S$ (usually assumed nonrandom) and by $N$ the number of elements of $U$. Let $R_i$ denote the indicator that a unit $i$ responds (or would respond if sampled) to the survey; let $t_X^* = \sum_{i \in U^*} X_i$ be the vector total of $X_i$ over the target-population, and $t_X = \sum_{i \in U} X_i$ be the corresponding total over the working frame. Assume further, from now on, that $X_i$ has first coordinate 1 for all $i \in U$.

For notational simplicity, we denote for any attribute $Z_i$, scalar or vector, $t_Z \equiv \sum_{i \in U} Z_i$, so that for example $t_{RX} = \sum_{i \in U} R_i X_i$ and $t_{(1-R)Y} = \sum_{i \in U} Y_i(1 - R_i)$. Similarly, we denote $\hat{t}_Z \equiv \sum_{i \in S} d_i Z_i$, the Horvitz-Thompson design-weighted survey estimator.

With $t_Y^* = \sum_{i \in U^*} Y_i$ as the population parameter to be estimated, and $X_i$ as calibration variables, one version of the linear-calibration estimator (when there is no separate nonresponse adjustment) is defined by

$$\hat{t}_{Y,cal} = \sum_{i \in S} w_i Y_i \quad , \quad w_i = d_i R_i \Big\{ 1 + (t_X - \sum_{j \in S} d_j R_j X_j)' \Big( \sum_{j \in S} d_j R_j X_j X_j' \Big)^{-1} X_i \Big\}$$

Here the weights $w_i$ are equivalently defined as minimizers of the loss-function $\sum_{i \in S} R_i (w_i - d_i)^2/(2d_i)$ which quantifies the difference between the design and adjusted weights, subject to the calibration constraint $\sum_{i \in S} w_i X_i = t_X$.

The calibration estimator $\hat{t}_{Y,cal}$ uses only the working frame information including calibration-variable totals presumed to come also from the working frame. Under general

conditions, we show that $\hat{t}_{Y,cal}$ is a design-consistent estimator of $t_Y$. These conditions are slightly more complicated than those given by Deville and Särndal (1992) because unlike that paper, our version of the calibration estimator simultaneously adjusts for nonresponse. For a formulation like the one here, see Särndal and Lundström (2005, Chap. 6).

In a superpopulation large-sample setting, under general conditions on the representativeness of the samples drawn, on the largest magnitudes of $X_i$ and $Y_i$ (as functions of $N$), and on the uniform positivity of the smallest eigenvalues of $\sum_{i\in U} R_i\, X_i\, X_i'$,

$$\hat{t}_{RX} \,=\, t_{RX} \,+\, O_P(\frac{N}{\sqrt{n}})\,,\;\; \hat{t}_{RY} \,=\, t_{RY} \,+\, O_P(\frac{N}{\sqrt{n}})\,,\;\; \hat{t}_{RYX} \,=\, t_{RYX} \,+\, O_P(\frac{N}{\sqrt{n}})$$

as both $n$ and $N$ get large and are nonrandom, and

$$\sum_{1\in S} d_i\, R_i\, X_i\, X_i' \;=\; \sum_{i\in U} R_i X_i X_i' \,+\, O_P(N/\sqrt{n})$$

Then it follows immediately from the last displayed equations that

$$\hat{\beta}_r \;\equiv\; \Big[\sum_{i\in S} d_i\, R_i\, X_i\, X_i'\Big]^{-1} \hat{t}_{RYX} \;=\; \beta_r \,+\, O_P(\frac{1}{\sqrt{n}})$$

where $\beta_r \,\equiv\, (\sum_{i\in U}\, R_i\, X_i\, X_i')^{-1}\, t_{RYX}$, and similarly that

$$\hat{t}_{Y,cal} \;=\; \hat{t}_{RY} \,+\, (t_X - \hat{t}_{RX})'\,[\sum_{i\in S} d_i\, R_i\, X_i\, X_i']^{-1}\, \hat{t}_{RYX}$$

$$=\; t_{RY} \,+\, (t_X - t_{XR})'\,[\sum_{i\in U} R_i\, X_i\, X_i']^{-1}\, t_{RYX} \,+\, O_P(\frac{N}{\sqrt{n}})$$

Now, substituting the definition of $\beta_r$ and making using of the identity $\sum_{i\in U} R_i\, X_i\, (Y_i - \beta_r' X_i) = 0$ that it implies, we find

$$\hat{t}_{Y,cal} \;=\; t_{RY} + (t_X - t_{RX})'\,\beta_r \,+\, O_P(\frac{N}{\sqrt{n}}) \;=\; \beta_r'\, t_X \,+\, O_P(\frac{N}{\sqrt{n}}) \qquad (1)$$

So far, we have not assumed anything about the mechanism of nonresponse or about the approximate equality of ideal and working frame totals. Typical assumptions about the nonresponse mechanism, with the flavor of the *pseudo-randomization model* (Fuller 2009) or the Missing at Random assumption, would restrict the dependence of the jointly distributed random variables $(R, Y, X)$ in such a way that $P(R = 1\,|\,Y,\, X) \;=\; P(R = 1\,|\,X)$ with probability 1. In the present setting — where $X_i$ and $Y_i$ are viewed as finite-population attributes that are unknown but not random, while $R_i$ are viewed either as binary random variables or constants with general dependence on $i$ — the proper form of the assumption that responders are like nonresponders in their $(X, Y)$ relationships is that for $\beta_r$ defined as above, $N^{-1}\, \sum_{i\in U}\, X_i\, (Y_i - \beta_r' X_i)$ is close to 0 as $N,\, n$ get large. Specifically, assume

$$N^{-1} \sum_{i\in U} X_i\, (Y_i - \beta_r' X_i) \;=\; o_P(1/\sqrt{n}) \qquad \text{as} \quad n,\, N \,\to\, \infty \qquad (\mathbf{R})$$

This assumption says that the working-frame least-squares regression coefficient vector $\beta$ is very close to the coefficient vector $\beta_r$ defined from the responder-subset of $U$ (i.e., from the subset of units $i$ for which $R_i \,=\, 1$). Under this assumption, combined with (1), $\hat{t}_{Y,cal} \,=\, t_Y \,+\, O_P(N/\sqrt{n})$, and the top-order $O_P(N/\sqrt{n})$ remainder term here is exactly the same as in equation (1).

The assumption (**R**) holds for example when $n = o_P(N)$ if the covariate-vectors $X_i$ are discrete with finitely many levels $X_i = x_j,\ 1 \le j \le k$, with $k$ fixed as $N$ gets large, and either if the response indicators $R_i$ are random and independently identically distributed for $i \in C_j \equiv \{a \in U :\ X_a = x_j\}$ for each $j$, or if the proportions of responders $R_i = 1$ within the cell $C_j$ tend to a limit $p_j$ as $N$ gets large, and the subset $\{i \in C_j :\ R_i = 1\}$ is chosen by simple random sampling from $C_j$.

Now we can complete our discussion of the effect of discrepancies between the ideal and working frames. The survey estimators used in practice are, like $\hat{t}_{Y,cal}$ in (1), based upon the working frame, which are design-consistent (after normalization by $N$) for the total-parameter $t'_X\,\beta_r$ over the working frame. In the real situation where $t_Y = t_X{}'\,\beta$ may differ from the target-population total $t^*_Y = t^*_X{}'\,\beta^*$, the inconsistency in the estimator manifests itself in two ways: in the difference $t_X - t^*_X$ between the calibration totals in the two frames, and also in the difference between the frame least squares coefficients $\beta = (\sum_{j\in U} X_j X'_j)^{-1} \sum_{j\in U}\ X_j Y_j$ and $\beta^* = (\sum_{j\in U^*} X_j X'_j)^{-1} \sum_{j\in U^*} X_j Y_j$. Thus, even if the calibration variable totals $t^*_X$ associated with the ideal frame $U^*$ were available from an external source to replace the totals $t_X$ in the calibration constraints, the ideal-frame design consistency of $\hat{t}_{Y,cal}$ would not be guaranteed, since the coefficient vectors $\beta$ and $\beta^*$ would not in general be the same. However, if one could reasonably assume that $\beta = \beta^*$, then the foregoing arguments based on (1) show under assumption (**R**) that a revision of the working-frame calibration totals $t_X$ to bring them into line with the ideal-frame totals $t^*_X$ would eliminate the inconsistency of $\hat{t}_{Y,cal}$.

From this point of view, the development in the paper is aimed at modeling the changes in a frame like MAF between updates in order to estimate differences $t^*_X - t_X$. If both (**R**) and $\beta = \beta^*$ are tenable assumptions, then estimates of these differences could allow calibration estimators constructed from a working frame to be brought closer to consistency.

**Remark 1** The superpopulation-based assumptions provided here under which linear calibration estimators simultaneously adjusting for nonresponse (design-consistently when the working frame is the ideal frame), can also be extended to generalized calibration or raking estimators. Such estimators arise when a different loss function $\sum_{i\in S} d_i\, H(w_i/d_i - 1)$ is used to measure differences between design and adjusted weights (based on a function $H$ such that $H(0) = H'(0) = 0$ and $H''(0) = 1$), instead of the quadratic one leading to linear calibration. The theoretical machinery for proving design consistency is given by Deville and Särndal (1992) in the absence of nonresponse. When nonresponse is taken into account, the theory of Deville and Särndal shows that a generalized calibration estimator is design-consistent for $\beta'_r\,t_X$ and is asymptotically equivalent to the linear calibration estimator $\hat{t}_{Y,cal}$ in the sense that the difference between the two estimators is $o_P(N/\sqrt{n})$. Thus the discussion above concerning the effect of ideal versus working frame discrepancies applies equally when raking (the case $H(z) = (1+z)\log(1+z) - z$) or a *pseudo-empirical likelihood* method ($H(z) = -\log(1+z) + z$) is adopted subject to calibration constraints.

## 3. Latent State Models for Frame Inclusion

We consider next some of the models which have been and might be fitted to data consisting of variables at unit and aggregated (block) levels describing frame elements before and after frame updating, together with indicators at unit level of frame additions and frame deletions. We have argued in the Introduction that unit additions and deletions could be studied statistically for the Census Bureau MAF through periodically updated unit-level descriptors and MAF status descriptors over time. The point of view advanced here is that unit-level or aggregated MAF (or other frame) status descriptors should be viewed

as a time-evolving stochastic process, with status changes (adds or deletes), new arrivals (i.e., new frame elements arising through new construction or conversion from other uses like businesses or splittings of previous frame elements), and removals (demolitions, condemned structures, zoning changes from residential to business, and purging of duplicates).

Current efforts at modeling frame changes have focused on counts of adds and deletes of frame units at block level as a result of a single round of address canvassing. Young et al. (2014) studied the counts respectively of block-level adds $N_b^+$ and deletes $N_b^-$ as a result of address canvassing in 2009-2010, in preparation for the 2010 Census, in terms of predictor variables observed before canvassing through the Postal Service Delivery Sequence Files, the post-2000-Census DAAL program, and new-construction and vacancy surveys. The predictor variables used are described in some detail by Young et al. (2014), but fall into the following broad categories:

**(1)** *Geographic characteristics of blocks*. Variables in this class include address-types (city-style, rural-route, or non-city) cross-classified with DSF coverage indicators, indicators of American Indian or Native Hawaiian Homeland status, Urban indicator, and indicator that the block was matched to the Census Bureau's TIGER integrated geographic database. Other variables of this sort that could have been included are found in the regularly updated Planning Data Base containing summary characteristics of census block-groups from the most recent ACS and decennial census.

**(2)** *Characteristics of individual housing structures*. These variables included a Trailer-as-HU indicator, indicators that a unit was included in the previous census or the MAF, had a small multi-unit address, was excluded from USPS delivery statistics, had one of a defined set of postal delivery types (business curbline mailbox, residential curbline mailbox, cluster mailbox in restricted community, etc.), is seasonally occupied, is vacant, or had one of several categories of 'valid' or 'invalid' address. Several of the variables mentioned here are characteristics of address stability as seen by the Postal Service, and the DSF files contain other such variables.

In the data explorations of Young et al. (2014), the unit-level variables were aggregated or recoded to block-level variables (e.g., indicators of at least one address of various types in the block), and interactions or other recodes combining multiple variables were not considered. Thus, some of the variables found by Young et al. to drop out of variable-selection screening might still play a role in models allowing interactions, and unit-level models of the types fitted by Young et al. (2014) with predictors selected from an expanded set of explanatory variables have yet to be tried.

Letting $X_b^\circ$ denote vectors of block-level predictor variables, and $N_b^+,\ N_b^-$ respectively denote the block-level counts of adds and deletes, the models fitted by Young et al. (2014) were of the form of *Zero-Inflated Latent State Models*, in which the counts $N_b^+$ or $N_b^-$, assumed independent across distinct block indices $b$, are represented in the form $N_b = \epsilon_b \cdot \nu_b$, where $\epsilon_b$ is a binary variable satisfying a logistic regression

$$P(\epsilon_b = 1 \,|\, X_b^\circ) \,=\, (1 \,+\, \exp(-\beta' X_b^\circ))^{-1}$$

and $\nu_b$ is a count random-variable (conditionally independent of $\epsilon_b$ given $X_b^\circ$) following a Poisson or Negative-Binomial Generalized Linear Regression Model with rate-parameter $\exp(\gamma'\, X_b^\circ)$. The essence of this model is that $\epsilon_b \in \{0, 1\}$ is a latent state describing the entire block $b$ which, when equal to 0, indicates that the count $N_b$ (of frame changes, adds or deletes) must be 0.

The zero-inflated models were first introduced in Lambert (1992), in the setting of manufacturing defects, where the latent state $\epsilon_b$ was an indicator of a manufacturing process

being out of control, and the count of actual defects $\nu_b$ observed in the out-of-control state was taken to be Poisson with log-rate following a linear model in terms of covariates. At least implicitly, Lambert envisioned the underlying time-dynamics of manufacturing processes passing stochastically in and out of control. Analogously, one might consider models for the time-evolution of the latent states $\epsilon_b$ each of which indicates whether the list of frame units within block $b$ can change in a current round of updates.

Models including mechanisms of change of state over time have been studied in many forms in the statistical literature. Earliest and most prevalent are Markovian models with observable states, and such models are readily extended to allow regression-type modifications of single-time-step transitions or transition rates as a function of covariates, as in Slud and Kedem (1994). A more elaborate discussion of likelihood inference for continuous-time models with transition intensities driven by continuous-time exogenous (unmodeled) covariates is given by Slud (1992). Models of this type are often expressed in terms of transition intensities that are themselves functions of covariates that may also be time-dependent stochastic processes. Semiparametric survival-time regression models of this type have been very prominent in biostatistics for more than 20 years (Andersen et al. 1993).

Different sorts of models have been devised for the time dynamics of Markov processes in which an important part of the state is unobservable. There is a huge literature on Hidden Markov models, with widely divergent fields of application ranging from computer recognition of audio speech signal sequences to automatic recognition of genes within DNA sequences. Common to all of these papers is a discrete sequence process $(M_t, \epsilon_t)$ in which the $M_t$ sequence is observable but the latent states $\epsilon_t$ are not, and usually $\epsilon_t$ is modeled as a Markov process within a finite-dimensional parametric family while the observable sequence variables $M_t$ are assumed conditionally independent given the sequence $\{\epsilon_s\}_s$. In these models, the unobserved states $\{\epsilon_s\}_s$ are interpreted as an unobserved 'context' given which the observables follow relatively simple conditional models $f(M_t \,|\, \{\epsilon_s\}_s) = f(M_t \,|\, \epsilon_{t-j},\ j = 0, \ldots, k)$ that may themselves contain unknown parameters. These models may all be viewed as mixed-effect Markov chain models, in which the latent states $\epsilon_t$ are discrete components of heterogeneity incorporated into a Markov chain model for the observables. Among many other papers on zero-inflated count time-series, one which may be viewed as a mixed-effect time series state-space model is Wang (2010).

An important class of Hidden Markov models in social science settings is the Mover-Stayer model (Vermunt 2004). This model is assumed to govern populations of independent individuals, each of whom at each time is a member of one of two unobservable groups (latent classes), whose composition changes with time. (These might be "movers" or "stayers" in settings where social behavior related to transiency is of interest; or adherents to one of the two major political parties in models of political behavior.) These models appear also in biostatistics, where the Markov chain for transitions of the observable variables given the latent states often have a generalized linear structure in terms of individual covariates. See for example the 1999 Biometrics paper of Albert on a mover-stayer model for longitudinal disease markers.

We seek in this section to place the zero-inflated generalized linear models of Young et al. (2014), for block-level counts of MAF adds and deletes from address canvassing, into a broader context of statistical models that incorporate temporal dynamics. Many of those are latent-state hidden Markov models, generally analyzed using some variant of the EM algorithm. They are mixture models, and when they are driven by covariates through regression models, require observations of units for many successive time-points, and even then can be very hard to fit. The alternative to latent-state modeling, in a data-rich environment such as MAF, is to define states explicitly through recoded groupings of unit-level covariates, after clustering covariate-defined cells with very similar observed

behavior. That is the approach preferred and advocated here for time-dependent modeling of unit-level MAF frame transitions.

## 4. Conditional Markov Models with Time-dependent Covariates

What might a Markovian model for the units of a frame list (MAF) look like? We will assume that existing MAF units (MAFID's) $i$ are each uniquely[1] associated with block indices $b = b(i)$ and equipped with a vector of covariates $X_i(t)$ and one of a finite set of state-labels $M_i(t) \in \{D, 1, 2, \ldots, K\}$ at time $t$. Here the state $D$ stands for *Deleted* and will be treated as an absorbing state, labelling a unit as an irreversibly invalid address. View the covariates $X_i(t)$ as time-evolving but non-random frame attributes, which is to say a complicated set of unknown non-random functions of time associated with population units. In this setup, the ideal frame $U^* = U^*(t)$ itself is a stochastic process, which by convention retains all units with $M_i(t) = D$, but which is also subject to an *immigration process* according to which new units enter the frame at times of nonhomogeneous block-specific Poisson processes with rates $\mu_b(s) = \exp(\alpha' Z_b(s))$, where $Z_b(s)$ are observable block-level time-dependent covariates. Finally, the state $M_i(s)$ for unit $i \in U^*(t)$ would move independently $j \mapsto k$ for $t < s \leq t + 1$ according to a transition intensity which depends on the set $\mathbf{X}(t)$ of all covariate processes only through the unit-$i$ covariates, in the form

$$\lambda_{j,k}^{(i)}(s \,|\, \mathbf{X}(t)) \;=\; \lambda_{j,k}(s \,|\, X_i(t)) \;=\; \exp(\beta_{j,k}{}' X_i(t)) \tag{2}$$

where $\{\beta_{j,k} : j \neq k,\ j = 1, \ldots K,\ k = D, 1, \ldots, K\}$ and $\alpha$ are unknown vectors of parameters. There is no need to parameterize rates of transition from state $D$, since we assume that state is *absorbing*, so that any unit entering state $D$ remains there forever.

According to this formulation, units in the frame population follow independent Markov processes driven by covariates governed by (2), and new units enter the frame according to block-specific inhomogenous Poisson processes with rates that may change as block-level covariates $Z_b(t)$ do. The observable unit-level covariates $X_i(t)$ might also contain some block-level covariates, and all covariates are assumed to be observable at integer time-unit intervals as new DSF and construction and vacancy survey data come online. However, although the states $M_i(s)$ might be partially updated at regularly spaced integer times $t$, states are not ascertained completely except just after address canvassing. As mentioned in the Introduction, full and correct information about frame inclusion is not truly available even from address canvassing, but we make the oversimplifying assumption that covariates and states are observed at integer times $t$, e.g. at the 6-month intervals when DSF updates have been received and processed. The frame size is then unknown only between integer-time updates, so that at time $s \in (t, t + 1]$, intermediate frame changes between $t$ and $t + 1$ are missing data. If properly specified and fitted, the model might be used — as described in the next section — to forecast the frame changes up to the next update-time $t + 1$.

The states $M_i(t)$ themselves have not yet been defined clearly, since it is proposed to define them after data explorations as a result of clustering of transition intensities in unit-level models fitted from a database of regularly updated frame-unit covariates summarizing neighborhood characteristics, address and postal-delivery stability, along with unit occupancy and residential status. The interpretations of the states — essentially as labels for classes of units defined from covariates with roughly similar rates of transition to the Delete-state $D$ — will be discussed more fully in Sec. 4.1 below.

In the notations of this section, the data used in fitting the zero-inflated GLM models of Young et al. (2014) are obtained from a single interval $(t, t + 1]$ during which

[1]This is a somewhat over-optimistic assumption of no geo-coding errors and no duplicates.

an address-canvassing operation takes place, and the block-level counts of adds and deletes are aggregated from unit-level data as

$$N_b^+ \;=\; \sum_{i \in U^*(t+1)} I_{[b(i)=b,\; M_i(t+1) \neq D]} \;-\; \sum_{i \in U^*(t)} I_{[b(i)=b,\; M_i(t) \neq D]} \tag{3}$$

and

$$N_b^- \;=\; \sum_{i \in U^*(t)} I_{[b(i)=b,\; M_i(t+1) = D]} \tag{4}$$

### 4.1 Discussion of States in the MAF Setting

Several special definitions and interpretations regarding the handling of MAFID's should be taken into account in any model of the type described in Section 4.

- The units being indexed, the MAFID's, should in principle consist of all map-spots with potentially residential dwelling units.
- Two types of frame changes might be considered 'adds': first are those representing residential conversions of MAFID's currently identified as 'invalid addresses', like businesses or some other type of structure on MAF; and second are the genuinely new units, which might either be completely new construction or else splitting of structures like garages or sheds or subdivision of houses or apartments to create new dwelling units which would receive new MAFID's. Note that a MAF add based on splitting would generate a new MAFID so that the corresponding rates in a proper model would involve splitting from individual units with unit covariates, while other adds would not be associated with any single existing unit, and the rates of immigration of these would be based only on block-indexed covariates.
- To account for states related to deletes, some changes of MAFID valid addresses to an invalid status should, if potentially reversible, not be treated as removal of the unit or entry into the absorbing state $D$, but rather as entry into a class $K$ of dormant quasi-deleted units from which (rare) returns to valid-address status can occur. Deletes associated with demolished structures do sometimes retain their MAFID's, and new construction on that same site in the future would constitute a special type of add transition linked to a specific address rather than to the block as a whole.

## 5. Estimates and Predictions of Frame Transitions

Because of the necessary technicalities described in Section 4.1 requiring some adds in MAF to be treated as Markov-process 'immigration' and others as unit splittings and still others as state-transitions, we consider in this section only deletes in discussing how model parameters would be estimated and used to forecast frame changes at the next future update-period. The response data available up to discrete update-time $t$ for use in model fitting would be the indicators $I_{[M_i(u)=D]}$ for units $i \in U^*(u-1)$ for $u = 1, 2, \ldots, t$, with the corresponding covariates $X_i(u-1)$. These observed delete-indicators would, according to the model, have expectations conditional on covariates and $M_i(u-1) = j$ equal to time-homogeneous Markov-process transition probabilities

$$P_{j,D}(u-1, u \,|\, X_i(u-1)) \;=\; P_{j,D}(0, 1 \,|\, X_i(u-1)) \tag{5}$$

where we explicitly disallow the possibility that adds within $(u-1, u]$ could also become deletes within the same update-period, and we note that under the model (2) the Markov

transitions are time-homogeneous because the transition regression coefficients $\beta_{j,k}$ are not time-dependent. The probability of transition from $j$ to $D$ in the time-interval $(u-1, u]$ must allow for all trajectories within the time-interval in which unit $i$ starting at state $j$ hits intermediate states before landing in $D$; but all of the intermediate transition-steps share the same covariates $X_i(u-1)$ since, according to our formulation in Section 4, the covariates do not change between successive integer update times.

If we consider the situation as of time $t$, delete-indicators $I_{[M_i(u)=D]}$ for units $i \in U^*(u-1)$ for $u \leq t$ are the data, with conditional expectations of observables at time $u-1$ expressed as (5). The forecast probabilities for time $t+1$ deletes would be $P_{k,D}(0, 1 \,|\, X_i(t))$ for $i \in U^*(t),\ M_i(t) = k$.

These probabilities, both for model-fitting and forecasting, are simple enough to be useful only when the probabilities of two or more transitions within any single update-interval are negligible. This is an assertion about the small length of update-intervals compared to typical occupation times for units in a fixed state (other than $D$). For the remainder of this section, we assume this to be true. In that case, standard continuous-time Markov chain theory tells that for $i \in U^*(u-1)$ and $M_i(u-1) = j$,

$$P_{j,D}(0, 1 \,|\, X_i(u-1)) \;\approx\; 1 \,-\, \exp\Big( -\int_{u-1}^{u} \exp(\beta_{j,D}{}' \, X_i(s))\, ds \Big) \tag{6}$$

and that these probabilities are small for each such $i$. Using conditional independence of the indicators $I_{[M_i(u)=D]}$ across $i$, given data up to time $u-1$, it can be argued that the block-level aggregate counts $\sum_{i \in U^*(u-1)} I_{[b(i)=b,\, M_i(u)=D]}$ of time-$u$ deletes will be approximately Poisson distributed with parameter

$$\sum_{i \in U^*(u-1)} I_{[b(i)=b]} \, \exp(\beta_{M_i(u-1),D}{}' \, X_i(u-1))$$

The corresponding forecast for the count of deletes in block $b$ at time $t+1$, based on parameters $\{\beta_{j,D}\}$ and covariates as of time $t$, is obtained from

$$\sum_{i \in U^*(t):\, b(i)=b} I_{[M_i(t+1)=D]} \;\overset{\mathcal{D}}{\approx}\; \text{Poisson}\Big( \sum_{i \in U^*(t):\, b(i)=b} \exp(\beta_{M_i(t),D}{}' \, X_i(t)) \Big) \tag{7}$$

Indeed, the expression for Poisson parameter on the right-hand side of (7), with parameter estimates substituted for parameters $\beta_{jD}$ would be the forecast for the count of deletes in block $b$ at time $t+1$. Such forecasts (augmented by predictions not given here for corresponding counts of block-level adds) would be used to modify the control totals $t_Z$ for various attributes $Z$, to make them closer to the desired totals $t_Z^*$ before using them as calibration constraints.

In estimating or forecasting frame deficiencies, e.g., the number of deletes between successive updates, there is no need to estimate transitions $\lambda_{j,k}$ among states $j, k = 1, \ldots, K$ if the shortness of time-intervals between successive updates implies that the probabilities of multiple transition steps within those short intervals can be ignored.

### 5.1 Zero-Inflated Type Model as Special Case

In the setting we are now considering, where the probabilities of two or more state-transitions for a unit $i$ within a single update interval $(u-1, u]$ are sufficiently small to be ignored by comparison with the single-transition probabilities on the right-hand side of (6), a zero-inflated model for counts can arise as a special case. Suppose that data are available only from the single update time-interval $(0, 1]$, and that the covariates $X_i(0)$ entering into (2)

consist only of $X_b^\circ$ at the block level (i.e., are identical for all $i$ with $b(i) = b$). Suppose further that the units have only two non-absorbing states, so that $K = 2$, where states 1 and 2 respectively correspond to 'stable' addresses with very rare transitions to the Delete state ($\lambda_{1D} \approx 0$) and to less stable addresses for which the transition intensities $\lambda_{2D}$ modeled by (2) are not so small.

Then if

$$n_{bk} \;=\; \sum_{i \in U^*(0),\, b(i)=b} I_{[M_i(0)=k]} \qquad \text{for} \quad k = 1, 2$$

denotes the number of block-$b$ units initially in the frame in state $k$, the rarity of $1 \mapsto D$ transitions implies

$$N_b^- \;\equiv \sum_{i \in U^*(0),\, b(i)=b} I_{[M_i(1)=D]} \;\approx \sum_{i \in U^*(0),\, b(i)=b} I_{[M_i(0)=2]} \cdot I_{[M_i(1)=D]}$$

Then equation (7) at $t = 0$ implies that

$$N_b^- \;\overset{\mathcal{D}}{\approx}\; \text{Poisson}\Big( n_{b2}\, \exp(\beta_{2D}' X_b^\circ) \Big) \tag{8}$$

where for blocks with $n_{b2} = 0$, the approximate distribution of $N_b^-$ should be interpreted as degenerate at 0.

However, if the available data at time 0 includes only $X_b^\circ$ and not $\{n_{b1},\, n_{b2}\}_b$, then the Markov-transition dataset has random but unobservable initial states, and $N_b^-$ has an approximate mixture distribution, with

$$P(N_b^- = m \,|\, X_b^\circ) \;\approx\; E\Big( \texttt{dpois}(m\,,\, n_{b2}\, \exp(\beta_{2D}' X_b^\circ)) \,|\, X_b^\circ \Big) \quad \text{for} \quad m \geq 1 \tag{9}$$

where $\texttt{dpois}(m, \lambda) \;\equiv\; e^{-\lambda}\, \lambda^m / m!$ denotes the Poisson probability mass function for nonnegative integers $m$, and the expectation on the right-hand side of (9) is taken with respect to the conditional probability mass function of $n_{b2}$ given $X_b^\circ$. The model given in equations (8) and (9) is a zero-inflated type of model for block-level deletes. The zero-inflation portion of the model is precisely the conditional probability $P(n_{b2} = 0 \,|\, X_b^\circ)$, and might be approximated by a logistic or other generalized-linear regression as a function of $X_b^\circ$. The count $N_b^-$, conditionally given $n_{b2} > 0$, is then a mixture of Poisson regression distributions. Thus, the count distribution may no longer follow an exponential-family generalized linear model.

We have seen in this special setting that zero-inflated count mixture models arise naturally in the unit-level Markovian model of Section 4, when data are available only in block-level aggregates and the numbers of units in different non-Delete states at the beginning of a single update interval are unobservable. More generally, statistical inference of parameters in the Markov model presents technical challenges when only block-level aggregated data are available, a problem closely related to that studied by Gill (1988).

## 6. Conclusions and Further Research

This paper has advocated the point of view that the address frame for a repeatedly conducted survey is a stochastic process, and that the unit- and area-level information used to refresh the frame should be viewed as data. The Census Bureau collects such data as part of the regular Delivery Sequence Files that it receives from the U.S. Postal Service, from the national surveys on New Construction, Housing Vacancy, and Mobile Homes that it conducts, from geographic databases that it maintains, and from the occasional address canvassing operations it undertakes as part of decennial census preparations.

Master Address File updates have in the past not been regarded as a data resource from which forecasts and estimates of frame errors might be made in terms of the dynamics of adds and deletes, but they could be developed into such a resource. This is an especially timely possibility because of the Census Bureau's current initiative to develop models of counts of block-level MAF adds and deletes, from past address canvassing, and to apply them to Targeted Address Canvassing in the 2020 Census.

A Markovian model of unit-level MAF status-changes has been proposed here, with covariate-based groupings of addresses as states. The definition of such states has been discussed in terms of underlying mechanisms of change and instability for individual residential housing units. However, the definition of such states must ultimately depend on analysis of data and tentative description of the rates of occurrence of adds and deletes in terms of block and unit level covariates. A specific conclusion from the proposals made here is that this kind of unit-level transitional analysis, from MAF updating data made as detailed as possible, should be high on the Census Bureau's research agenda. Such research would simultaneously support future targeting operations for MAF updating, and also future application to estimates and forecasts of MAF errors as experienced in the major national surveys that the Bureau conducts. Estimation and forecasting could be accomplished using regularly observed update-data through models of the sort proposed here.

The considerations of this paper suggest several specific directions of MAF update data analysis and modeling research:

- Targeting in address canvassing and specification of states for Markov models of unit address status both require an effective disaggregation of addresses into groups or clusters within which the rates of occurrence of adds or deletes are very similar, and across which these rates are as different as possible.

- It is very important that future modeling efforts concerning the rates of local adds and deletes in terms of covariates should be more detailed than has been tried so far, incorporating unit-level address covariates and block-group level neighborhood data available through the Planning Data Base, as well as interactions of these.

- As far as possible, models are needed also for unit-level rates of adds and deletes in terms of covariates available from regularly updated data sources, and not just from decennial address canvassing.

- Separate models for rates of block-level occurrence of New Construction should be developed.

- The process of developing models for unit-level address-status changes will clarify whether models for adds and deletes must take account of a few different mechanisms for MAF additions and deletions, in particular of the distinctions discussed in Section 4.1 between those additions and deletions logically related to area-level variables — such as the ones from new construction or re-zoning — and those resulting from resident behaviors at the unit level such as the splitting or merging of separate dwelling units at a single address.

## REFERENCES


Albert, P. (1999), "A mover-stayer Model for longitudinal marker data", *Biometrics*, **55**, 1252–1257.

Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993) *Statistical Models Based on Counting Processes.* Springer: New York.

Deville, J. and Särndal, C. (1992), "Calibration estimators in survey sampling" *Jour. Amer. Statist. Assoc.*, **87**, 376–382.

Fuller, W. (2009) *Sampling Statistics*, Wiley: Hoboken, NJ.

Gill, R. (1988), "On estimating transition intensities of a Markov process with aggregate data of a certain type: 'Occurrences but no exposures'", *Scand. Jour. Stat.*, **13**, 113–134.

Johnson, N. and Kephart, K. (2013) "2010 Census Evaluation of Address Frame Accuracy and Quality", 2010 Census Planning Series Memoranda No. 252.

Lambert, D. (1992) "Zero-inflated Poisson regression, with an application to defects in manufacturing", *Technometrics*, **34**, 1–14.

Särndal, C. and Lundström, S. (2005), *Estimation in Surveys with Nonresponse.* Wiley: Chichester.

Slud, E. (1992), "Partial likelihood for continuous-time stochastic processes", *Scand. Jour. Stat.*, **19**, 97–109.

Slud, E. and Kedem, B.(1994), "Partial likelihood analysis of logistic regression and autoregression", *Statistica Sinica*, **4**, 89–106.

Tomaszewski, C. and Boies, J. (2014), "Recent advancements in statistical modeling to identify Address Updating Areas for the 2020 Census", JSM 2014 Proceedings paper.

Vermunt, J. (2004), "Mover-Stayer Models", article in: The SAGE Encyclopedia of Social Science Research Methods, eds. M.S. Lewis-Beck, A. Bryman and T. Liao, http://srmo.sagepub.com/

Wang, P. (2010), "Markov zero-inflated regression models for a time series of counts with excess zeros", *Jour. Appl. Statist.*, **28**, 623–632.

Young, D., Johnson, N, and Pennington, R. (2014), "Zero-inflated modeling for characterizing coverage errors of extracts from the US Census Bureau's Master Address File". Internal Census Bureau preprint.