Modeling Frame Deficiencies for Improved Calibration

Eric V. Slud, U.S. Census Bureau, CSRM
Univ. of Maryland, Mathematics Dept.

JSM 2014, Monday August 4

Outline

- 1. Motivation: Census Master Address File (MAF) updating
  Calibration for Survey Nonresponse & Frame Omissions

- 2. Address Frame as a Stochastic Process
  Zero-inflated Count Models for Block-level Adds and Deletes
  Markov Modeling of transition intensities with covariates

- 3. Dynamical vs Fixed-time, Latent vs. Observed States
  Zero-inflated, Hidden-Markov, Mover-Stayer Models

- 4. Technical Challenges
  Inference from aggregated data, adjusted frame totals

Ideal Frame versus Working Frame

Master Address File (**MAF**) : continuously maintained list on which working frames for the Decennial Census and the American Community Survey (**ACS**) are based.

Ideal frame is the list of all (US & Puerto Rico) unique locations of potentially residential structures.

Updated by Postal Service Delivery Sequence File and by
post-2000 census Demographic Area Address Listing

Decennial Census updates include *address canvassing*
possibly *targeted* for Census 2020,
through modeled add & delete rates at block level

MAF Data Sources

`P.O. Delivery Sequence Files` (every 6 mos.)
include many unit-level status variables for
mail delivery, occupancy, address stability

`Blockgroup-level Planning database`
neighbhd demographics & (census/ACS) characteristics

`Geographic Data`
New-construction Records, aerial imagery of surfaces

`Address Canvassing` (in preparation for decennial census)
Observe $X_i(t), Z_i(t)$, but not without error
(Johnson & Kephart Census evaluation report, 2013)

Calibration for Nonresponse & Noncoverage

Data are: $\{(X_i, R_i, R_i \cdot Y_i) \, : \, i \in S\}$ and totals $t^*_{\mathbf{X}} = \sum_{i \in U} X_i$

$S \subset U \subset U^*$ is probability sample drawn from working frame U (within ideal U^*) with known inclusion prob's π_i

X_i predictive (unit-level) covariates, R_i unit-response indicator

Y_i unit attribute with desired (ideal-frame) total t_Y

Estimator: $\hat{t}_Y = \sum_{i \in S} w_i \, R_i \, Y_i$ using calibrated weights w_i
minimizing Loss $= \sum_{i \in S} R_i \, \pi_i \, (w_i - \pi_i^{-1})^2$ subject to
calibration constraints $\sum_{i \in S} R_i \, w_i \, X_i = t^*_X$

Design-consistent if R_i satisfies Missing-at-Random condition **and** working-frame totals t^*_X correctly reflect ideal frame.

Models for Frame Errors

Frame Deficiencies as Missing (not-yet-observed) Data
Address Canvassing as Auxiliary Data

D. Young et al. (2014, JSM), initial MAF Error Model:
block-level counts of adds or deletes
zero-inflated Poisson or negative binomial regression model
in terms of environment variables X_i

$U^* =$ true address list, $U =$ MAF

aggregated summary of stochastic transitions

$i \in U^c \mapsto i \in U$ (add) or $i \in U \mapsto i \in U^c$ (delete)

Here consider unit-level X_i-conditionally Markovian models, with Delete absorbing state **D**, some unit-splitting (garages, out-buildings), and immigration (new construction).

Markovian Unit Model with Covariates

For each dwelling unit (MAF ID) i, unit-level covariates $X_i(t)$ for neighborhood, address & postal-delivery stability, occupancy and residential status evolve over time.

Units have states $Z_i(t) \in \{D, 1, 2, \ldots, K\}$, related to covariates but not ascertained completely except just after canvassing. Think of states as clusters based on covariates.

Assume transition $j \mapsto k$ rates $\lambda_{jk}(t|\mathbf{X}) = \exp(\beta_{jk}{}' X(t))$ depend only on covariates $X(t)$ and coefficient vectors β_{jk}.

For MAF updating, Deletes relate to $\{\lambda_{jD}(t)\}$ or $\{\beta_{jD}\}$; and Adds to transitions from invalid to valid MAFIDs or new construction.

Zero-Inflated Latent State Models

Lambert (1992, Zero-inflated Poisson)   *original paper*

mentioned in/out-of control setting, latent time dynamics

Zero-inflated models (Young et al. 2014, JSM) applied to MAF for $b =$ block index, given block-level covariates X_b^* :

$$N_b = \epsilon_b \nu_b , \quad \begin{cases} P(\epsilon_b = 1 | X_b^*) = \texttt{plogis}(\beta' X_b^*) \\ \nu_b \sim \texttt{NegBin}(\texttt{exp}(\gamma' X_b^*), \kappa) \end{cases}$$

Interpret counts $N_b = \sum_{i \in b} I_{[Z_i(1)=D]}$ as block aggregates of deletes at time 1.

View $\epsilon_b = 0, 1$ as time-1 latent state for (all) units in block b

Time-varying Latent State Models

Could regard $N_{b,t}$ as time-dependent block-counts (say of deletes) with associated latent states $\epsilon_{b,t}$ and time-dependent covariates $X_b^*(t)$.

Since $\epsilon_{b,t}$ are not observed we have a `Hidden Markov Model`; since they drive the time-dependence, a truly time-dependent model would specify their time-dynamics.

Models of this sort are given by Wang (2010), Albert (1999) in a biostat setting with analysis via EM algorithm.

Vermunt (2004) describes similar `Mover-Stayer Models` with latent binary state in social-science context.

Unit-level Delete Probabilities in Markov Model

Probabilities we care about are (for $j = 1, \ldots, K$, units i)

$$P_{j,D}(0, t|\mathbf{X}) \;=\; P(Z_i(t) = D \,|\, Z_i(0) = j,\, X_i(s),\, 0 \le s \le t)$$

If probabilities of 2 or more transitions are negligible, obtain these conditionally given the states $Z_i(0) = j$, approximately as $P_{j,D}(0, t|\mathbf{X}_i) \approx 1 - \exp(-\int_0^t \exp(\beta_{jD}{}' X_i(s))\, ds)$ (small):

block-level deletes conditionally become sums of independent Bernoulli variates with these success prob's.

If covariates are block-level (constant over $i \in b$) and $n_b(k) = \sum_{i \in b} I_{[Z_i(0)=k]}$, obtain forecasts of Delete totals

$$\sum_{i \in b} I_{[Z_i(t)=D]} \overset{\mathcal{D}}{\approx} \sum_{k=1}^{K} M_k \;, \qquad M_k \sim \text{Poi}\Big(n_b(k)\, t\, e^{\beta_{kD}{}' X_b^*(0)}\Big)$$

ZIP-type Model as Special Case

General-link ZIP model arises with $K = 2, t = 1,$ and

states $\left\{ \begin{array}{c} k = 1 \\ k = 2 \end{array} \right\}$ corresponding to $\left\{ \begin{array}{l} \text{rare Deletes, } \lambda_{1D} \approx 0 \\ \text{appreciable Delete rates} \end{array} \right.$

$$P(\sum_{i \in b} I_{Z_i(t)=D]} = m \,|\, X_b(0)) =$$

$$E\Big(\texttt{dpois}(m, n_b(2) \exp(\beta_{2D}{}' X_b(0))) \Big| X_b(0)\Big)$$

where the logistic component of ZIP is replaced by $P(n_b(2) = 0 \,|\, X_b(0))$.

Statistical Consequences of Reformulation

For estimation/forecasting of frame deficiencies:

• no need to estimate transitions among states $1, \ldots, K$ if short times between successive updates result in few changes.

• 'states' $\{1, \ldots, K\}$ represent covariate-defined clusters from which unit transition-rates to Delete or Add status are different: these should be sought even in Zero-inflated modeling efforts like those of Young et al. (2014): suggests Disaggregation.

• separate models for rates of block-level occurrence of New Construction must be found.

Summary & Further Research

• unit-level Markovian models are proposed for `Add` or `Delete` address-updates in MAF, with covariate-based address clusters as states

• statistical inference of transition parameters would most naturally be done using regularly observed update-data

• when numbers of MAF IDs initially in states cannot be observed in single rounds of address-canvassing, resulting approximate models resemble the zero-inflated models of Young et al. (2014) for block add/delete counts

• models are needed also for unit-level rates (and ultimately, errors) of `Adds` and `Deletes` under regular updating versus address-canvassing.

Thank you !

Eric.V.Slud@census.gov

evs@math.umd.edu

Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are the author's and not necessarily the Census Bureau's.

References

D. Young, N. Johnson, R. Pennington (2014 JSM paper)
Zero-inflated MAF count models

W. Fuller (2009) **Sampling Statistics** *re: calibration*

R. Gill (1988) Scand. Jour. Stat.
Markov intensities, and inference from aggregated data

N. Johnson & K. Kephart (2013) "2010 Census Evaluation of Address Frame Accuracy and Quality"

J. Vermunt (2004) Mover-Stayer Models (SAGE Encyclopedia)

P. Wang (2010) Jour. Appl. Stat.
Markov zero-inflated regression models