Informing Maintenance to the U.S. Census Bureau's Master Address File with Statistical Decision Theory

Andrew M. Raim
Center for Statistical Research and Methodology, U.S. Census Bureau

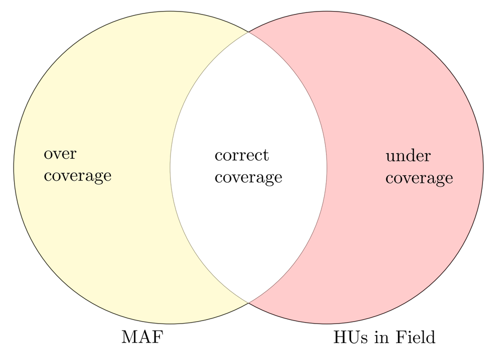
United States Census Bureau

Summary

- The Master Address File (**MAF**) is a database maintained by the Census Bureau of all known housing units.
- The MAF is used to prepare address lists for the decennial census and household surveys (e.g. American Community Survey); it is critical to the Census Bureau's business.
- MAF is regularly updated throughout the decade. Also, a large-scale block listing operation (**2010 Address Canvassing / AdCan**) was carried out before the 2010 Census.
- Census Bureau is now considering alternative strategies to prepare for 2020 Census [6]. The goal is to reduce the cost of updating the MAF without significant loss of coverage.
- Previous work with statistical models used sorted predictions to identify census blocks for closer inspection [2, 3, 7]. This may not capture decision makers' assessment of cost for wasted effort or missed coverage error.
- **Objective:** Explore use of decision theory to assist MAF maintenance. Discrete loss functions are proposed to aid intuition of decision makers.
- Decision maker quantifies: (a) severity of coverage error, and (b) loss due to possible actions under each severity level.
- Even with known "state of nature", optimal decision can vary greatly by decision maker.

Master Address File and Address Canvassing

- To prepare the MAF for the 2010 Decennial Census, the Census Bureau conducted the 2010 AdCan operation. ~111,000 field representatives (FRs) walked ~6 million census blocks in the U.S. and Puerto Rico [5].



- AdCan provided a wealth of data on MAF coverage errors.
  1. A **valid address missing** from the MAF indicated an **undercoverage** error. Address was added to the MAF and an **new add** outcome was recorded.
  2. An **invalid address present** on the MAF indicated an **overcoverage** error. Address was "deleted" from the MAF and a **delete** outcome was recorded.
  3. A **matched add** occurred where an address was already on the MAF, but could not be properly geocoded until located by AdCan.
- Census Bureau initiatives to avoid a large in-field canvassing before 2020 census:
  1. In-office canvassing using aerial imagery review.
  2. Statistical models to help inform a limited field operation.

Review of Decision Theory

- Suppose there are J possible states of nature $\Theta = \{\theta_1, \ldots, \theta_J\}$ and d possible actions $\mathcal{A} = \{a_1, \ldots, a_K\}$.
- Loss function $L(\theta, \delta)$ measures consequence of taking action δ when the state is θ.

	θ_0: No Rain Today	θ_1: It Rains Today
a_0: Leave Umbrella	0	100
a_1: Bring Umbrella	50	20

- θ usually unobservable; inferred through data $\mathcal{D}$, model $p(\mathcal{D} \mid \theta)$, and prior $p(\theta)$.
- For observed data $\mathcal{D}$, take action δ to minimize risk $r(p, \delta) = \mathrm{E}[L(\theta, \delta) \mid \mathcal{D}]$.

Categories of Coverage Error

- Consider the following measure of coverage error for the ith census block,

$$Y_i^* = \frac{\text{NewAdds}_i + \text{MatchedAdds}_i + \text{Deletes}_i}{\text{HU}_i + 1}, \quad i = 1, \ldots, n.$$

- Numerator represents "units of coverage error" and denominator reduces severity for blocks with more preexisting housing units.
- Decision maker defines cutpoints $\gamma_1 < \cdots < \gamma_{J-1}$ to create meaningful categories,

$$\begin{aligned} [\gamma_0 < Y_i^* \leq \gamma_1] &\equiv \text{Least severe coverage error}, \\ &\vdots \\ [\gamma_{J-1} < Y_i^* \leq \gamma_J] &\equiv \text{Most severe coverage error}, \end{aligned}$$

where $\gamma_0 = -\infty$ and $\gamma_J = \infty$ are fixed.

- Let $(Y^*, \boldsymbol{x})$ denote random coverage measurement and fixed covariate for a given block. Let $\mathcal{D} = \{(\text{NewAdds}_i, \text{MatchedAdds}_i, \text{Deletes}_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ denote data used to train model.
- Let $\pi_j = \pi_j(\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{P}_{\boldsymbol{\theta}}(\gamma_{j-1} < Y^* \leq \gamma_j \mid \boldsymbol{x})$ be the probability of category $j = 1, \ldots, J$.
- The exact form of $\boldsymbol{\theta}$ (the "state of nature") is determined by the model. If π_j is not a tractable function of $\boldsymbol{\theta}$ and $\boldsymbol{x}$, can approximate by Monte Carlo.
- We will consider risk functions which depend on $p(\boldsymbol{\theta} \mid \mathcal{D})$ through posterior category probabilities $\mathrm{E}[\pi_j \mid \mathcal{D}]$, $j = 1, \ldots J$.

A Two Decision Problem to Aid In-Office Canvassing

- Census Bureau is considering aerial imagery and other in-office alternatives to a full scale canvassing operation.
- Using past data on MAF coverage errors, statistical models could help by triggering high-risk census blocks for closer review. With sufficiently good predictors, it may be possible to capture errors not detectable by other in-office approaches.
- Consider $J = 5$ categories of coverage error for each block,

$$\{\text{None}, \text{Lo}, \text{Med}, \text{Hi}, \text{Severe}\}, \text{ with cutpoints } \gamma_1 = 1, \gamma_2 = 4, \gamma_3 = 10, \gamma_4 = 20.$$

- For a given census block, there are two possible actions: **do trigger** the block for review (a_1) and **do not trigger** the block for review (a_0).
- Consider the linear loss function

$$L(\boldsymbol{\theta}, \delta) = \begin{cases} \boldsymbol{c}_0^T \boldsymbol{\pi} = c_{01}\pi_1 + \cdots + c_{0J}\pi_J, & \text{if } \delta = a_0 \\ \boldsymbol{c}_1^T \boldsymbol{\pi} = c_{11}\pi_1 + \cdots + c_{1J}\pi_J, & \text{if } \delta = a_1. \end{cases}$$

based on positive costs $\boldsymbol{c}_0 = (c_{01}, \ldots, c_{0J})$ and $\boldsymbol{c}_1 = (c_{11}, \ldots, c_{1J})$.

- Decision maker determines $\boldsymbol{c}_0$ and $\boldsymbol{c}_1$ before making any actual decisions.
  1. Order the $2J$ outcomes $\langle a_k, \theta_j \rangle$ from least to most desirable.
  2. Solicit loss values for ordered outcomes using algorithm from [1, Ch. 2].
  3. Let c_{kj} be the loss associated with outcome $\langle a_k, \theta_j \rangle$.
- Notice that we always want $c_{01} \leq \cdots \leq c_{0J}$ and $c_{11} \geq \cdots \geq c_{1J}$.
- Using posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$, the posterior risk is

$$r(p, \delta) = \begin{cases} \boldsymbol{c}_0^T \, \mathrm{E}[\boldsymbol{\pi} \mid \mathcal{D}], & \text{if } \delta = a_0 \\ \boldsymbol{c}_1^T \, \mathrm{E}[\boldsymbol{\pi} \mid \mathcal{D}], & \text{if } \delta = a_1. \end{cases}$$

- Optimal decision is a_0 if $(\boldsymbol{c}_0 - \boldsymbol{c}_1)^T \, \mathrm{E}[\boldsymbol{\pi} \mid \mathcal{D}] \leq 0$, and a_1 otherwise.

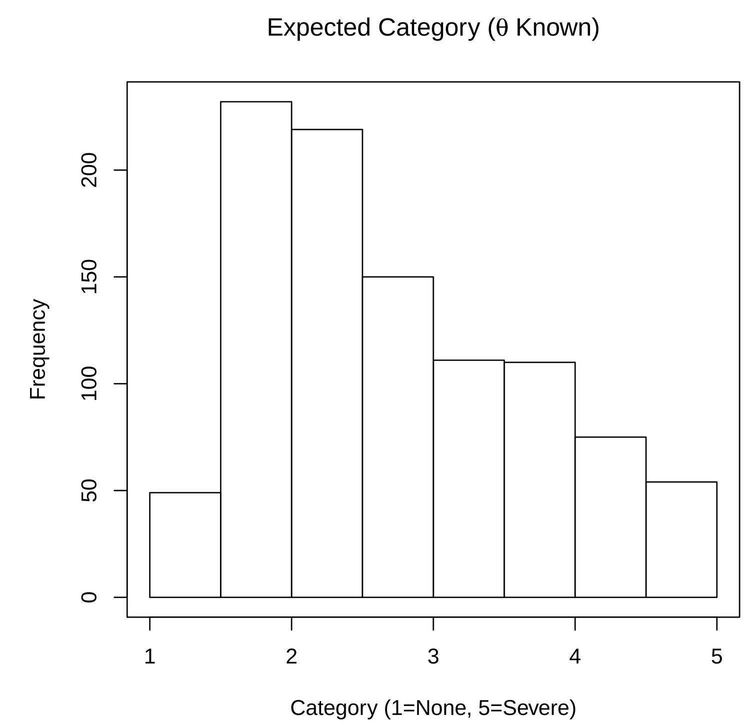
Simulation

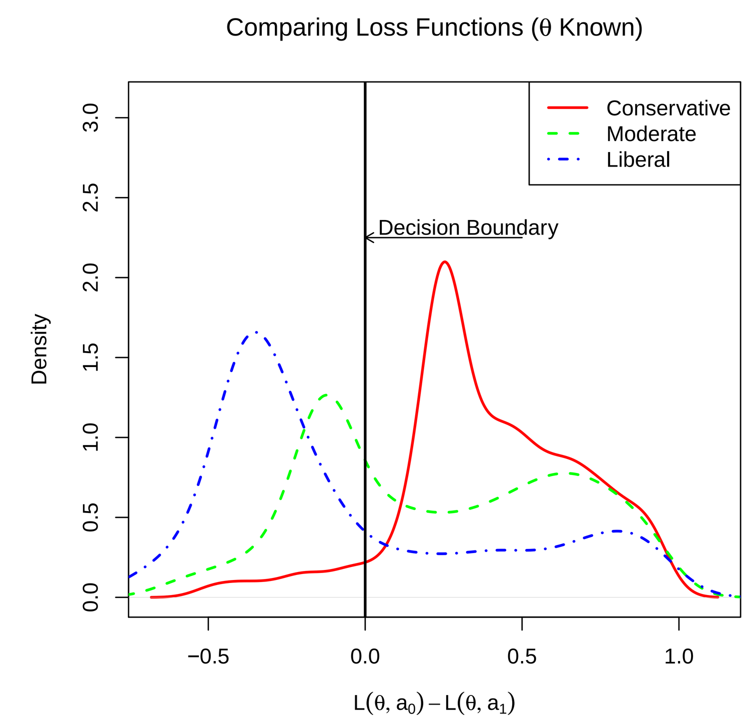
Compare several decision makers in the following scenario. Assume $\boldsymbol{\theta} = (\boldsymbol{\beta}^A, \boldsymbol{\beta}^M, \boldsymbol{\beta}^D, \tau_A^2, \tau_M^2, \tau_D^2)$ is known for now (i.e. a no-data problem).

$$\begin{aligned} &x_i \overset{\text{iid}}{\sim} \mathrm{U}(0, 8), \quad \text{HU}_i \overset{\text{iid}}{\sim} \text{NegBin}(\mu, \kappa), \quad \text{for } i = 1, \ldots, n = 1000, \\ &\text{NewAdds}_i \overset{\text{ind}}{\sim} \text{Poisson}\left(\exp(\beta_0^A + \beta_1^A x_i + \beta_2^A \log(\text{HU}_i + 1) + e_i^A)\right), \\ &\text{MatchedAdds}_i \overset{\text{ind}}{\sim} \text{Poisson}\left(\exp(\beta_0^M + \beta_1^M x_i + \beta_2^M \log(\text{HU}_i + 1) + e_i^M)\right), \\ &\text{Deletes}_i \overset{\text{ind}}{\sim} \text{Binomial}\left(\text{HU}_i, \text{logit}^{-1}(\beta_0^D + \beta_1^D x_i + e_i^D)\right), \\ &e_i^A \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_A^2), \quad e_i^R \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_M^2), \quad e_i^D \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau_D^2). \end{aligned}$$

Utility functions (rank for each outcome is shown in parentheses).

"Conservative"	None	Low	Med	Hi	Severe
a_0	1 (9)	0.375 (4)	0.25 (3)	0.125 (2)	0 (1)
a_1	0.5 (5)	0.625 (6)	0.75 (7)	0.875 (8)	1 (9)
"Moderate"	None	Low	Med	Hi	Severe
a_0	1 (9)	0.625 (6)	0.25 (3)	0.125 (2)	0 (1)
a_1	0.375 (4)	0.5 (5)	0.75 (7)	0.875 (8)	1 (9)
"Liberal"	None	Low	Med	Hi	Severe
a_0	1 (9)	0.75 (7)	0.875 (8)	0.125 (2)	0 (1)
a_1	0.25 (3)	0.375 (4)	0.5 (5)	0.625 (6)	1 (9)





Number of blocks triggered for review: Conservative (927), Moderate (609), Liberal (255).

Conclusions and Next Steps

- Even when θ is fully known, "optimal" result may vary greatly by decision maker.
- Investigate integrating decision framework with actual data, models, and operations.
- Incorporate model uncertainty into loss functions, to avoid too many triggers.
- Problem of selecting blocks for a limited in-field canvasssing before 2020 census.

References

[1] James O. Berger. *Statistical decision theory and Bayesian analysis.* Springer, 2nd edition, 1993.

[2] John L. Boies, Kevin M. Shaw, and Jonathan P. Holland. 2010 census program for evaluations and experiments address canvassing targeting and cost reduction evaluation report. In *2010 Census Planning Memoranda Series.* 2012.

[3] Andrew M. Raim and Marissa N. Gargano. Selection of predictors to model coverage errors in the Master Address File. Research Report Series: Statistics #2015-04, Center for Statistical Research and Methodology, U.S. Census Bureau, 2015.

[4] Yves Thibaudeau and Darcy S. Morris. Bayesian decision theory for further optimizing the use of administrative records in the census NRFU. (In Progress).

[5] U.S. Census Bureau. 2010 census address canvassing operational assessment. In *2010 Census Planning Memoranda Series: 2010 Census Program for Evaluations and Experiments.* U.S. Census Bureau, 2012.

[6] U.S. Census Bureau. 2020 census detailed operational plan for the address canvassing operation, 2015.

[7] Derek S. Young, Andrew M. Raim, and Nancy R. Johnson. Zero-inflated modelling for characterizing coverage errors of extracts from the US Census Bureau's Master Address File. *Journal of the Royal Statistical Society: Series A*, 2016.