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

Andrew M. Raim

Center for Statistical Research and Methodology
U.S. Census Bureau
Washington, DC, 20233, U.S.A

**Abstract**

The U.S. Census Bureau maintains the Master Address File (MAF) as a basis for the decennial census and household surveys it conducts throughout the decade. To prepare for the 2010 Census, the Census Bureau organized a full-scale address canvassing operation where field representatives walked most of the United States to find and correct errors on the MAF. The Census Bureau is now developing strategies to avoid the high cost of a full in-field canvassing for the 2020 Census and to reduce errors throughout the decade. One idea has been to use statistical models to study and predict coverage errors found on the MAF. Such models could potentially help to inform address listing fieldwork and other less costly alternatives being considered by the Census Bureau. Previous MAF modeling work has simply ordered point estimates to suggest regions which might contain significant coverage error and require further attention. In this work, we consider the problem in a decision theoretic framework. By choosing an appropriate loss function, we can study consequences of canvassing decisions under a selected model with data from past operations.

**Key Words:** Address Canvassing; Bayesian; Fieldwork; Loss; Optimal; Risk.

## 1. Introduction

The Master Address File (MAF) is a database maintained by the Census Bureau to record addresses of all known housing units in the United States. The MAF is used to prepare address lists for the decennial census and household surveys such as the American Community Survey (ACS). Because the MAF is critical to the mission of the Census Bureau, it is maintained through several mechanisms. The primary source of new addresses is the Delivery Sequence File (DSF) which is received from the United States Postal Service twice per year. Addresses are also contributed through Census Bureau listing programs such as Local Update of Census Addresses, Demographic Area Address Listing, and Community Address Updating System (U.S. Census Bureau, 2014).

To prepare for the 2010 Decennial Census, the Census Bureau carried out a large scale field operation known as the 2010 Address Canvassing operation (AdCan). This operation involved approximately 111,000 field representatives (FRs) walking 6 million census blocks in the U.S. and Puerto Rico and comparing an address list extracted from the MAF with addresses encountered on the ground (U.S. Census Bureau, 2012). A valid housing unit which was discovered on the ground but missing from the MAF was deemed an *undercoverage error*, and the address was added to the MAF to correct the error. This corrective action was recorded primarily in one of two ways, determined at the end of AdCan. If the address already existed on the MAF but failed to be included on the address list for FRs, it was recorded as a *matched add*. In many cases, the MAF entry was incomplete (e.g.,

*Email: andrew.raim@census.gov.

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

some location information was missing) and only became usable after it was rediscovered in the field. On the other hand, if the added housing unit was new to the MAF, a *new add* outcome was recorded in the operation. An address from the MAF that was found not to be a habitable housing unit was deemed an *overcoverage error* and the address was marked for deletion from the MAF and a *delete* outcome was recorded in the operation.[1]

The extensive 2010 Address Canvassing operation prepared the MAF for the decennial census and provided a wealth of data on coverage errors. However, the cost of nearly half a billion dollars made it the second most expensive operation in the 2010 Census, after non-response followup operations (Boies et al., 2012). Furthermore, a large proportion of blocks were seen to be stable, with little or no coverage error. After 2010, the Census Bureau began research on initiatives to avoid a large in-field canvassing operation before the 2020 Census. If stable blocks could be identified and set aside, canvassing costs could be substantially reduced without allowing a significant increase in coverage errors. One major effort has been to build statistical models from the AdCan outcomes and predict coverage error for census blocks (or other levels of geography) using covariates available before the operation. One of the first statistical modeling efforts was to use logistic regression (Boies et al., 2012). Young et al. (2016) took a count regression approach using the zero-inflated negative binomial model. Raim and Gargano (2015) also considered zero-inflated negative binomial regression, but with a more exhaustive variable selection and more candidate predictors. Heim and Raim (2016) considered several models for count regression with spatial dependence among the random effects. Current operational plans by the Census Bureau do not directly involve models, but instead focus on in-office canvassing activities such as reviewing satellite images to detect changes in housing units (U.S. Census Bureau, 2015a,b). Although models are not planned for use in 2020 Census operations, they are still of general interest from a research perpective and could be considered for use beyond the 2020 Census.

This paper will take a closer look at one particular aspect of the statistical modeling approach: how to make decisions based on an estimated model. In previous work such as Boies et al. (2012), Raim and Gargano (2015), Young et al. (2016), and Heim and Raim (2016), models were evaluated by a method similar to the following. Suppose there are $n$ census blocks, and we are allowed to canvass a limited number $k < n$. Let $y_1, \ldots, y_n$ denote the observed counts of new adds from AdCan, and let $\hat{y}_1, \ldots, \hat{y}_n$ denote predictions from a given model. Let $\mathcal{Z}_k \subset \{1, \ldots, n\}$ be the set of $k$ blocks with the largest predictions $\hat{y}_i$. We can then consider quantities such as the count $\sum_{i \in \mathcal{Z}_k} y_i$ which would have been captured through a limited canvassing guided by the model. One fundamental shortcoming of this method is that it does not use all three components of the multivariate outcome (new adds, matched adds, and deletes). Another shortcoming is that it does not weigh the cost of reduced effort (when a block is excluded from $\mathcal{Z}_k$) against the cost of missed coverage errors. Deciding not to canvass a particular block results in an immediate cost savings from the canvassing operation; however, resulting coverage error could result in reduced quality of the decennial census.

In this paper, we recall the classical framework for statistical decision theory (Berger, 1993; DeGroot, 2004) and consider how it can be used to make decisions based on MAF models. We propose a simple measure of MAF error based on a multivariate outcome where the decision maker specifies cutpoints to define categories of coverage error severity. This facilitates specifying the loss function as a two-way table. From here, risks can be computed and a decision can be identified which minimizes risk; in this sense, we determine an *optimal* decision. We consider two variants of the block selection problem. In the first

[1] In practice, records are not actually deleted from the MAF, but are flagged so that they can be distinguished from habitable housing units.

**Table 1**: Example loss function for umbrella decision problem.

| | | Rain Today? | |
|---|---|---|---|
| | | No ($\omega_0$) | Yes ($\omega_1$) |
| Bring Umbrella? | No ($a_0$) | 0 | 100 |
| | Yes ($a_1$) | 50 | 20 |

variant, we may decide, for each block individually, whether or not to trigger an alert of a coverage error problem. In the second variant, we are given an allowance $k < n$ of blocks that we can select out of $n$. We find, perhaps an unsurprising conclusion, that even when the state of nature (represented by the model parameter) is fully known, the "optimal" decision can vary greatly based on the decision maker. We note that costs are expressed through loss or utility and not actual monetary costs. Recently, Thibaudeau and Morris (2016) have also explored the use of statistical decision theory in census operations, specifically for non-response followup.

The rest of the paper proceeds as follows. In Section 2, we recall the classical decision theory framework where a finite number of outcomes are possible. Section 3 formulates MAF coverage as a categorical state. Section 4 considers the block selection problem where decisions are at the individual block level, while Section 5 considers the variant where a subset of $k$ is to be selected. Finally, Section 6 concludes the paper.

## 2. Decisions with Finite Outcomes

Consider a simple scenario where there are $J$ possible states of nature $\Omega = \{\omega_1, \ldots, \omega_J\}$ and $M$ possible actions $\mathcal{A} = \{a_1, \ldots, a_M\}$. The true state of nature $\omega \in \Omega$ is not known, but information on its prior distribution $\mathrm{P}(\omega = \omega_j)$ for $j = 1, \ldots, J$ is assumed. The decision maker must select an action $\delta \in \mathcal{A}$ which leads to the best possible outcome $\langle \omega, \delta \rangle$. The consequence of taking action $\delta$ under state $\omega$ is quantified by a loss function $L(\omega, \delta)$, which must be specified by the decision maker. We will make the common assumption that $L(\omega, \delta) \geq 0$ so that $L(\omega, \delta) = 0$ indicates a best possible outcome.

For example, consider the problem of deciding to bring an umbrella to work today. Suppose there are only two possible states: no rain today ($\omega_0$) or rain today ($\omega_1$). The two possible actions are to either leave the umbrella at home ($a_0$) or bring the umbrella ($a_1$). The loss function is subjective and depends on the decision maker, but here is one possible development:

- The best outcome is $\langle \omega_0, a_0 \rangle$, with nicer weather and where we are not encumbered by the umbrella.
- The worst outcome is $\langle \omega_1, a_0 \rangle$, where we are caught in the rain without the umbrella.
- The outcome $\langle \omega_1, a_1 \rangle$ is more desirable than $\langle \omega_0, a_1 \rangle$. For the latter, we may be likely to lose the umbrella because it was not needed. In the former outcome, at least we have the satisfaction that we made a good decision.

By carefully comparing the outcomes, we have determined an ordering $\langle \omega_1, a_0 \rangle < \langle \omega_1, a_1 \rangle < \langle \omega_0, a_1 \rangle < \langle \omega_0, a_0 \rangle$, where the notation $\mathcal{O}_1 < \mathcal{O}_2$ is taken to mean that outcome $\mathcal{O}_1$ is less desirable than outcome $\mathcal{O}_2$. Next, numerical values can be assigned to the loss function; one possibility is given in Table 1.

In a Bayesian setting, the prior density $p(\omega)$ captures our uncertainty about the state $\omega$ before any data are collected. In the absense of any data, an optimal decision can be made

by selecting an action $\delta \in \mathcal{A}$ to minimize the expected loss under the prior distribution,

$$\mathrm{E}[L(\omega, \delta)] = \sum_{j=1}^{J} L(\omega_j, \delta) p(\omega_j).$$

For the rain example, suppose $p(\omega_0) = 0.5$ and $p(\omega_1) = 0.5$. Then $\mathrm{E}[L(\omega, a_0)] = 50$ and $\mathrm{E}[L(\omega, a_1)] = 35$ so that the optimal decision is to bring the umbrella.

Evidence on $\omega$ can be collected through observed data $\mathcal{D}$; by specifying a likelihood $p(\mathcal{D} \mid \omega)$ which relates the probability of $\mathcal{D}$ to $\omega$, we can use Bayes rule to obtain the posterior density $p(\omega \mid \mathcal{D})$. In a Bayesian analysis, we desire to make an optimal decision conditional on $\mathcal{D}$. This can be accomplished by selecting an action $\delta \in \mathcal{A}$ to minimize the expected loss under the posterior distribution

$$\mathrm{E}[L(\omega, \delta) \mid \mathcal{D}] = \sum_{j=1}^{J} L(\omega_j, \delta) p(\omega_j \mid \mathcal{D}). \tag{1}$$

The quantity (1) is the posterior risk, which we will notate by $r(p, \delta)$. This notation emphasizes that the risk depends on the density $p$ of $\omega$ (which may be, for example, the prior $p(\cdot)$ or the posterior $p(\cdot \mid \mathcal{D})$) and the selected action $\delta$. To illustrate a possible source of data in the umbrella problem, we could have $\mathcal{D}$ consisting of a single random variable

$$X = \begin{cases} 1 & \text{if local weather station predicts rain,} \\ 0 & \text{otherwise,} \end{cases}$$

where $X \sim \text{Bernoulli}(\omega_1)$ if it will rain and $X \sim \text{Bernoulli}(\omega_0)$ otherwise.

## 3. Categories for Coverage Error

The decision framework from Section 2 applies to outcome spaces where both the space $\Omega$ for the state of nature and the action space $\mathcal{A}$ are discrete. MAF error models have involved Bernoulli, Negative Binomial, and related regression models (Boies et al., 2012; Raim and Gargano, 2015; Young et al., 2016; Heim and Raim, 2016). In such models, the unknown state of nature $\boldsymbol{\theta}$ is represented by parameters of a likelihood. It may be difficult for a decision maker to select an action $\delta \in \mathcal{A}$ based directly on inference of $\boldsymbol{\theta}$, which may include regression coefficients, variances, dispersion parameters, etc. Instead, we will create categories of severity for MAF coverage error as functions of $\boldsymbol{\theta}$, which will facilitate decision making.

We will focus on models with count outcomes, as discussed in Raim and Gargano (2015), Young et al. (2016), and Heim and Raim (2016). Let $\text{NewAdds}_i$, $\text{MatchedAdds}_i$, and $\text{Deletes}_i$ denote (scalar) counts of new adds, matched adds, and deletes, respectively, for blocks $i = 1, \ldots, n$; these quantities are considered to be random variables. Also let $\boldsymbol{x}_i$, $i = 1, \ldots, n$, denote a vector of covariates for regression. For example, $\boldsymbol{x}_i$ may contain the area of the $i$th block, whether the block is urban or rural, and the count of housing units on the block before the new adds, matched adds, and deletes were obtained. We will denote the latter quantity as $\text{HU}_i$. The $\boldsymbol{x}_i$ are considered fixed (not random), and the goal of modeling will be to make decisions conditional on knowing these characteristics. We will denote

$$\mathcal{D} = \{(\text{NewAdds}_i, \text{MatchedAdds}_i, \text{Deletes}_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$$

as all data available to train the model. We will drop the index $i$ and use expressions such as NewAdds, MatchedAdds, and Deletes when referring to a typical block with predictor $\boldsymbol{x}$.

We would like an ordinal variable to summarize the overall coverage error for a block based on the multivariate outcome $(\text{NewAdds}_i, \text{MatchedAdds}_i, \text{Deletes}_i)$. One way to do this is to consider random variables

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

as measures of coverage error on the blocks of interest. Higher counts of new adds, matched adds, or deletes contribute to an increased coverage error, while more previously existing housing units lowers the perceived severity.[2] For example, consider two blocks with NewAdds+MatchedAdds+Deletes $= 10$; a block with HU $= 1$ would be a more alarming case of coverage error than another block with HU $= 1000$.

The decision maker can now define cutpoints $\gamma_1 < \cdots < \gamma_{J-1}$ to create meaningful categories for the range of $Y^*$,

$$[\gamma_0 < Y^* \leq \gamma_1], \quad \ldots, \quad [\gamma_{J-1} < Y^* < \gamma_J],$$

where $\gamma_0 = -\infty$ and $\gamma_J = \infty$ are fixed. The event that $Y^*$ falls into each of these categories is a (multinomial) random variable, but the associated probabilities

$$\pi_j(\boldsymbol{x}, \boldsymbol{\theta}) = \text{P}_{\boldsymbol{\theta}}(\gamma_{j-1} < Y^* \leq \gamma_j \mid \boldsymbol{x})$$

are fixed quantities given the unknown $\boldsymbol{\theta}$. To simplify the notation, we will write $\pi_j = \pi_j(\boldsymbol{x}, \boldsymbol{\theta})$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ when the context is clear. Whether $\boldsymbol{\pi}$ is a tractable function of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ depends on the the distribution of $Y^*$; if not tractable, $\boldsymbol{\pi}$ can be computed by Monte Carlo approximation.

In the following sections, we will consider problems whose loss functions are linear combinations of $\boldsymbol{\pi}$, so that the posterior risk is a linear combination of $\text{E}[\boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{\theta}) \mid \mathcal{D}]$. Here, expectation is taken with respect to the density $p(\boldsymbol{\theta} \mid \mathcal{D}) = p(\mathcal{D} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathcal{D})$, with likelihood $p(\mathcal{D} \mid \boldsymbol{\theta})$ and prior $p(\boldsymbol{\theta})$ specified by the model.

## 4. A Two Decision Problem for Individual Blocks

The Census Bureau is planning to use aerial imagery and other in-office alternatives to a full scale canvassing operation (U.S. Census Bureau, 2015b). Statistical models could augment the planned operations by triggering high-risk census blocks for closer review. A model with sufficiently good predictors might detect coverage error which is not visible through imagery. In practice, finding effective predictors for coverage error is an ongoing problem (Raim and Gargano, 2015). Here we will simply assume knowledge of an effective $\boldsymbol{x}$.

Suppose that there are $n$ blocks in the country. Focusing on the $i$th block, we would like to determine whether there is evidence of serious enough coverage error to trigger an alert. Here, our decision can be one of two choices and is local to one particular block at a time. To be concrete, let us consider $J = 5$ categories of coverage error,

$$\{\text{None}, \text{Lo}, \text{Med}, \text{Hi}, \text{Severe}\},$$

[2] We let the three types of coverage error contribute to the coverage error equally. We could also consider setting coefficients to weight new adds, matched adds, and deletes differently, and to modify the effect of $\text{HU}_i$ in the denominator. Note that $\text{HU}_i + 1$ is used in the denominator because blocks may have zero housing units before canvassing.

**Table 2**: An example ranking of the outcomes in the two-decision problem for individual blocks.

| | None | Lo | Med | Hi | Severe |
|---|---|---|---|---|---|
| $a_0$ | 9 | 6 | 3 | 2 | 1 |
| $a_1$ | 4 | 5 | 7 | 8 | 9 |

with cutpoints $\gamma_1 = 1, \gamma_2 = 4, \gamma_3 = 10, \gamma_4 = 20$. For a given census block, the two possible actions in $\mathcal{A}$ are: $(a_1)$ to trigger an alert on the block or $(a_0)$ to not trigger. We will consider a linear loss function for the $i$th block,

$$L_i(\boldsymbol{\theta}, \delta) = \begin{cases} \boldsymbol{c}_0^T \boldsymbol{\pi}(\boldsymbol{x}_i, \boldsymbol{\theta}), & \text{if } \delta = a_0, \\ \boldsymbol{c}_1^T \boldsymbol{\pi}(\boldsymbol{x}_i, \boldsymbol{\theta}), & \text{if } \delta = a_1. \end{cases}$$

The decision maker determines the nonnegative costs $\boldsymbol{c}_0 = (c_{01}, \ldots, c_{0J})$ and $\boldsymbol{c}_1 = (c_{11}, \ldots, c_{1J})$ to formulate the loss function.[3] Interpretation of this loss function is similar to the discrete loss function from Section 2; for example, if $\boldsymbol{\pi} = (1, 0, \ldots, 0)$ then $c_{01}$ represents the consequence of action $a_0$ when the category is None with probability 1. For a general $\boldsymbol{\pi}$, $L_i(\boldsymbol{\theta}, \delta)$ represents a weighted mean of costs $\boldsymbol{c}_0$ or $\boldsymbol{c}_1$.

Selection of $\boldsymbol{c}_0$ and $\boldsymbol{c}_1$ is critical for the loss function to reflect the sense of the decision maker. Meaningful values for these costs can be solicited in an intuitive way by taking the following steps:

1. Rank the $2J$ outcomes $\langle \omega_j, a_\ell \rangle$ from least to most desirable. Ties are acceptable.
2. Solicit loss values for ordered outcomes using the algorithm from Berger (1993, Ch. 2).
3. Let $c_{kj}$ be the loss associated with outcome $\langle \omega_j, a_\ell \rangle$.

In sorting the outcomes, we first note that $c_{01} \leq \cdots \leq c_{0J}$ and $c_{11} \geq \cdots \geq c_{1J}$ will be true for a rational decision maker. For example, if our action is to not trigger $(a_0)$, it is safe to assume that the best outcome for coverage severity is None, the second best outcome is Lo, and so on.

The algorithm from Berger (1993, Chapter 2) gives a principled way to solicit numerical values for the sorted outcomes; it is also described in DeGroot (2004). We will think for a moment in terms of utility rather than loss. The concepts are equivalent, with utility proportional to the negative of the loss. We will illustrate the algorithm intuitively in terms of the current problem; see Berger (1993) for a more formal description. Suppose we have ranked the outcomes according to Table 2.

We first assign utility 0 to the worst outcome(s) and utility 1 to the best outcome(s),

$$\begin{aligned} U_i(\text{Severe}, a_0) &= 0 \\ U_i(\text{None}, a_0) &= 1 \\ U_i(\text{Severe}, a_1) &= 1 \end{aligned}$$

Now identify outcome(s) whose rank is halfway between worst and best; this is $\langle \text{Lo}, a_1 \rangle$ in our example, which has rank 5. The decision maker compares two experiments: (1) having $\langle \text{Lo}, a_1 \rangle$ occur with certainty, and (2) taking the result of a gamble between a worst outcome and a best outcome, say

$$\text{P}(\langle \text{Severe}, a_0 \rangle) = 1 - \alpha, \quad \text{P}(\langle \text{None}, a_0 \rangle) = \alpha.$$

[3] The decision itself is only to select $\delta \in \mathcal{A}$; $\boldsymbol{c}_0$ and $\boldsymbol{c}_1$ remain fixed.

The decision maker must select an $\alpha$ to make the gamble exactly as appealing as having $\langle \text{Lo}, a_1 \rangle$ occur with certainty. Doing so yields

$$\begin{aligned} U_i(\text{Lo}, a_1) &= (1-\alpha)U_i(\text{Severe}, a_0) + \alpha U_i(\text{None}, a_0) \\ &= (1-\alpha)\cdot 0 + \alpha \cdot 1 \\ &= \alpha. \end{aligned}$$

To help guide the selection of $\alpha$, we can think of choosing $\alpha$ closer to 1 if the appeal of $\langle \text{Lo}, a_1 \rangle$ is closer to the appeal of the best outcome, or choosing it closer to 0 if the appeal is closer to the appeal of the worst outcome. Here, $\langle \text{Lo}, a_1 \rangle$ represents triggering an alert that might be considered unnecessary, but does not seem close to the severity of $\langle \text{Severe}, a_0 \rangle$; therefore, $\alpha > 1/2$ seems justified.

We now have utility values for outcomes with ranks 1, 5, and 9. We next solicit utilities for ranks 3 and 7. The outcome with rank 3 is $\langle \text{Med}, a_0 \rangle$; we obtain its utility by comparing the two experiments: (1) it occuring with certainty, and (2) taking the result of a gamble between the rank 1 outcome (with probability $1-\alpha$) and the rank 5 outcome (with probability $\alpha$). Again, $\alpha$ must be selected to make the two experiments equally appealing. Similarly, the rank 7 outcome is assigned utility by comparing to the rank 5 and rank 9 outcomes. The remaining outcomes can be assigned utility in a similar way, comparing to surrounding outcomes whose utility value has already been determined. Berger (1993) recommends a formal consistency check of the the completed utility function to ensure the decision maker has thought through all comparisons. The assigned utility values are in the interval $[0, 1]$ and the loss function is obtained by $L_i(\omega_j, a_\ell) = 1 - U_i(\omega_j, a_\ell)$, for each $k = 1, \ldots, M$ and $j = 1, \ldots, J$.

After formulating the loss function and observing data $\mathcal{D}$, the posterior risk for the $i$th block is

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

where $\text{E}[\cdot \mid \mathcal{D}]$ represents the expectation with respect to the posterior distribution $p(\boldsymbol{\theta} \mid \mathcal{D})$. The optimal decision is $a_0$ if $(\boldsymbol{c}_0 - \boldsymbol{c}_1)^T \, \text{E}[\boldsymbol{\pi} \mid \mathcal{D}] \leq 0$, and $a_1$ otherwise.

The following simulation illustrates the degree that the optimal solution can change with respect to changes in the loss function. Assume the model

$$\begin{aligned} &\text{NewAdds}_i \overset{\text{ind}}{\sim} \text{Poisson}\left(\exp(\beta_0^A + \beta_1^A w_i + \beta_2^A \log(\text{HU}_i + 1) + e_i^A)\right), \\ &\text{MatchedAdds}_i \overset{\text{ind}}{\sim} \text{Poisson}\left(\exp(\beta_0^M + \beta_1^M w_i + \beta_2^M \log(\text{HU}_i + 1) + e_i^M)\right), \\ &\text{Deletes}_i \overset{\text{ind}}{\sim} \text{Binomial}\left(\text{HU}_i, \text{logit}^{-1}(\beta_0^D + \beta_1^D w_i + e_i^D)\right), \\ &e_i^A \overset{\text{iid}}{\sim} \text{N}(0, \tau_A^2), \quad e_i^R \overset{\text{iid}}{\sim} \text{N}(0, \tau_M^2), \quad e_i^D \overset{\text{iid}}{\sim} \text{N}(0, \tau_D^2), \end{aligned}$$

for $i = 1, \ldots, n = 1000$. This scenario is a simplified version of regression models used in past MAF error modeling work. The predictors $\boldsymbol{x}_i = (w_i, \text{HU}_i)$, $i = 1, \ldots, n$, are generated according to

$$\begin{aligned} w_i &\overset{\text{iid}}{\sim} \text{Uniform}(0, 8), \\ \text{HU}_i &\overset{\text{iid}}{\sim} \text{NegBin}(\mu, \kappa). \end{aligned}$$

To simplify matters, we take $\boldsymbol{\theta} = (\boldsymbol{\beta}^A, \boldsymbol{\beta}^M, \boldsymbol{\beta}^D, \tau_A^2, \tau_M^2, \tau_D^2)$ to be known, with

$$\begin{aligned} &\boldsymbol{\beta}^A = (0.5, 0.25, 0.1), \quad \boldsymbol{\beta}^R = (0.5, 0.25, 0.1), \quad \boldsymbol{\beta}^M = (0, 0.5), \\ &\tau_A = \tau_M = \tau_D = 0.1, \quad \mu = 5, \quad \kappa = 0.25, \end{aligned}$$

**Table 3**: Utility functions for three decision makers. The utility value for each outcome is shown, with its rank in parentheses.

| (a) "Conservative" | None | Lo | Med | Hi | Severe |
|---|---|---|---|---|---|
| $a_0$ | 1 (9) | 3/8 (4) | 2/8 (3) | 1/8 (2) | 0 (1) |
| $a_1$ | 4/8 (5) | 5/8 (6) | 6/8 (7) | 7/8 (8) | 1 (9) |
| (b) "Moderate" | None | Lo | Med | Hi | Severe |
| $a_0$ | 1 (9) | 5/8 (6) | 2/8 (3) | 1/8 (2) | 0 (1) |
| $a_1$ | 3/8 (4) | 4/8 (5) | 6/8 (7) | 7/8 (8) | 1 (9) |
| (c) "Liberal" | None | Lo | Med | Hi | Severe |
| $a_0$ | 1 (9) | 6/8 (7) | 5/8 (6) | 1/8 (2) | 0 (1) |
| $a_1$ | 2/8 (3) | 3/8 (4) | 4/8 (5) | 7/8 (8) | 1 (9) |

so that the observed data $\text{NewAdds}_i$, $\text{MatchedAdds}_i$, and $\text{Deletes}_i$ are not needed for decision making.[4] [5]

Table 3 formulates utility functions for three fictional decision makers. The table entry for each outcome contains its utility value along with its rank (from least to most desirable) among all outcomes for the decision maker. All three decision makers consider $\langle \text{None}, a_0 \rangle$ and $\langle \text{Severe}, a_1 \rangle$ to be the best outcomes, both equally good because the correct decision was made. Similarly, all consider $\langle \text{Severe}, a_0 \rangle$ to be the worst outcome. Of the remaining outcomes, the "conservative" decision maker favors $a_1$ if the severity is anything except None, and therefore detection of any coverage error is considered to be important. The "moderate" decision maker has a slightly different prioritization, treating $\langle \text{Lo}, a_0 \rangle$ as a more desirable outcome. The "liberal" decision maker takes this one step further, also treating $\langle \text{Med}, a_0 \rangle$ as a more desirable outcome.

Figure 1 shows a histogram of expected categories $\sum_{j=1}^{J} j \cdot \pi_j(\boldsymbol{x}_i, \boldsymbol{\theta})$, for $i = 1, \ldots, n$, of the simulated blocks. This gives an objective summary of the blocks which does not depend on a particular loss function. We would likely decide $a_1$ for the blocks with expected category near 5, and $a_0$ for the blocks with expected category close to 1, but the reader must decide where he or she places the threshold to trigger a block. Figure 2 compares the optimal decisions by our three decision makers. The empirical density of the quantities

$$d_i = L_i(\boldsymbol{\theta}, a_0) - L_i(\boldsymbol{\theta}, a_1) = (\boldsymbol{c}_0 - \boldsymbol{c}_1)^T \boldsymbol{\pi}(\boldsymbol{x}_i, \boldsymbol{\theta}), \text{for } i = 1, \ldots, n \tag{2}$$

is shown for each utility function. Recall that (2) is the quantity used in the optimal decision rule when no data $\mathcal{D}$ is observed. The $d_i$ which fall to the left of the "decision boundary", where the horizontal axis is 0, should not be triggered according to this rule. In all, 927 blocks are triggered under the conservative loss, 609 are triggered under the moderate loss, and 322 are triggered under the liberal loss. This result emphasizes the large degree that an optimal result can vary based on the beliefs of the decision maker.

## 5. A Block Selection Problem

Consider the problem of block selection before the decennial census. Suppose it is determined that we may canvass at most $k$ of the $n$ blocks using representatives in the field. This is a greatly simplified version of the problem because blocks can vary in size and effort to canvass. Also, the cost of canvassing depends on the spatial configuration of the selected

[4] This is an example of a "no-data" decision problem.

[5] Taking $\boldsymbol{\theta}$ to be known is equivalent to taking the prior $p(\boldsymbol{\theta})$ as a point mass at the true data-generating $\boldsymbol{\theta}$.

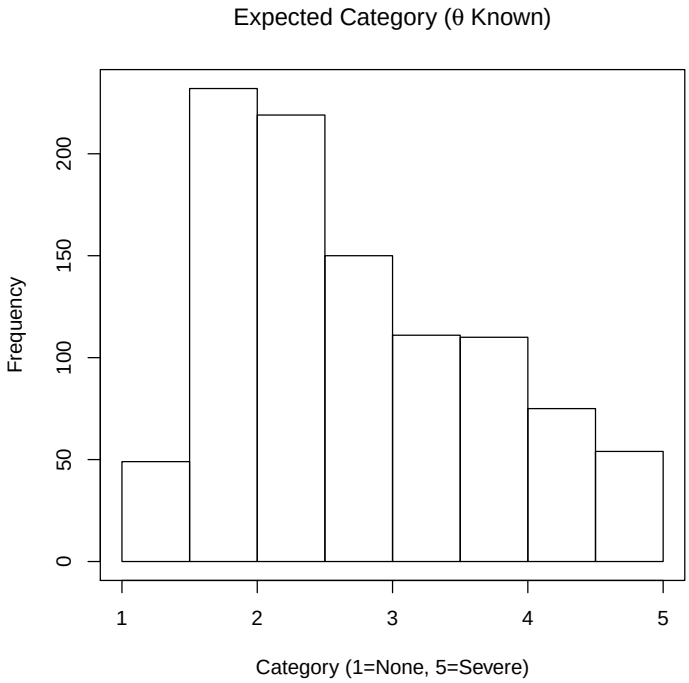


**Figure 1**: Histogram of expected categories for simulated blocks.

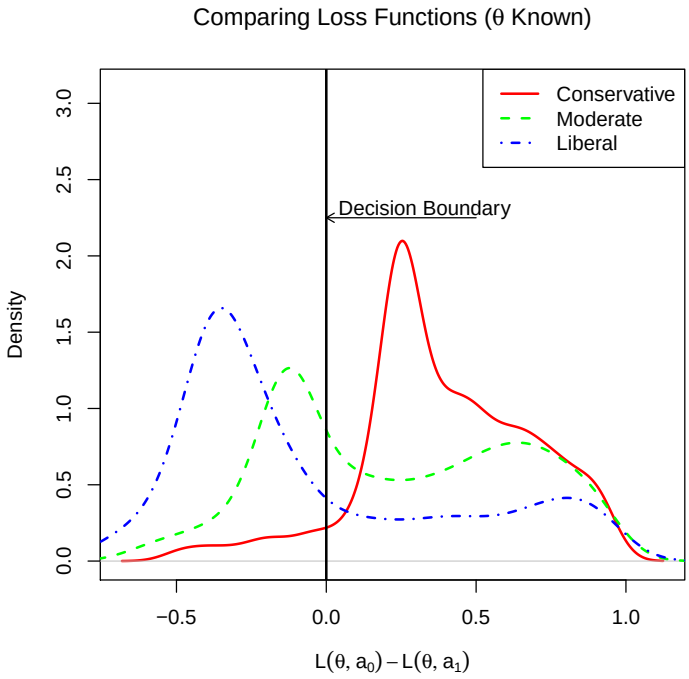


**Figure 2**: Empirical density of $d_i = (\boldsymbol{c}_0 - \boldsymbol{c}_1)^T \boldsymbol{\pi}(\boldsymbol{x}_i, \boldsymbol{\theta})$ for $i = 1, \ldots, n$, under each of the three loss functions.

blocks; for example, a selection spaced uniformly across the country will be more costly than a selection with relatively fewer clusters of nearby blocks.

Let $\mathcal{K}$ be the loss for exceeding the allowance; we will assume that it is very large so that a reasonable solution will never exceed $k$ blocks. Denote $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ as our vector of decisions, with

$$\delta_i = \begin{cases} a_1 & \text{if } i\text{th block is selected,} \\ a_0 & \text{otherwise} \end{cases}$$

Consider the loss

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} L_i(\boldsymbol{\theta}, \delta_i) + \mathcal{K} \cdot I\left\{\sum_{i=1}^{n} I(\delta_i = a_1) > k\right\},$$

where $\sum_{i=1}^{n} L_i(\boldsymbol{\theta}, \delta_i)$ represents the loss due to decision making and $\mathcal{K} \cdot I\left\{\sum_{i=1}^{n} I(\delta_i = a_1) > k\right\}$ represents the loss due to exceeding the allowance. The assumption of the decision loss being additive is rather strong—the loss due to many poor decisions might actually be larger than the sum of the individual losses—but ensures that the problem is tractable. The posterior risk given observed data $\mathcal{D}$ becomes

$$r(p, \boldsymbol{\delta}) = \sum_{i=1}^{n} r_i(p, \delta_i) + \mathcal{K} \cdot I\left\{\sum_{i=1}^{n} I(\delta_i = a_1) > k\right\}.$$

This expectation is easily computed because the decision loss is additive, while the allowance loss does not depend on $\boldsymbol{\theta}$. Denoting $d_i = r_i(p, a_0) - r_i(p, a_1)$ as the reduction in risk obtained by selecting the $i$th block, and $z_i = I(\delta_i = a_1)$ we may write

$$r(p, \boldsymbol{\delta}) = \sum_{i=1}^{n} r_i(p, a_0) - \sum_{i=1}^{n} d_i z_i \quad \text{subject to} \sum_{i=1}^{n} z_i \leq k. \tag{3}$$

From (3), it can be seen that minimizing $r(p, \boldsymbol{\delta})$ is equivalent to maximizing $\sum_{i=1}^{n} d_i z_i$ subject to the constraint. The following greedy algorithm determines an optimal selection:

1. Let $\mathcal{Z}_k$ be the indices corresponding to the $k$ largest $d_i$.
2. Make decision $a_1$ for $i \in \mathcal{Z}_k$, and $a_0$ for the remaining $i \notin \mathcal{Z}_k$.

There are many possible extensions to this problem, but care must be taken to avoid intractibility. For example, consider the weighted loss functions

$$L^*(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} w_i L_i(\boldsymbol{\theta}, \delta_i) + \mathcal{K} \cdot I\left\{\sum_{i=1}^{n} I(\delta_i = a_1) > k\right\} \quad \text{and}$$

$$L^{**}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{i=1}^{n} L_i(\boldsymbol{\theta}, \delta_i) + \mathcal{K} \cdot I\left\{\sum_{i=1}^{n} w_i I(\delta_i = a_1) > k\right\},$$

where weights $w_1, \ldots, w_n$ are known to the decision maker. Loss $L^*$ leads to a similar greedy algorithm as $L$. However, minimizing risk based on loss $L^{**}$ is similar to the 0-1 knapsack problem in the computer science literature. A greedy strategy does not solve the 0-1 knapsack problem optimally, but a more clever approach such as dynamic programming is needed (Cormen et al., 2009, Chapter 16).

## 6. Conclusions

Statistical decision theory provides a formal way to make informed decisions using data. In this paper, we have considered two problems related to address canvassing and MAF maintenance at the Census Bureau. These were toy problems which ignored many complexities, but perhaps are a step toward thinking about real operational problems. There are several steps in formulating the loss function which could be subject to scrutiny: our quantity to measure coverage error, the need to specify cutpoints to determine categories of severity, and the specification of the loss function itself. We have attempted to make the process intuitive, but much responsibility is placed upon the decision maker.

Many interesting extensions to this work are possible. In the block selection problem, spatially remote blocks are more expensive to include; however, this was not factored into our simple version of the problem. Traditionally, in the statistical literature, the loss function is used to derive an optimal procedure rather than to make an actual *decision*. It may therefore be of interest to incorporate both the statistical procedure and the decision into a single loss function. Finally, we note in Section 4 that the optimal rule is based on a linear function of $\mathrm{E}(\boldsymbol{\pi} \mid \mathcal{D})$, which is a single summary of the posterior distribution. The posterior could be used further to assess variability associated with the decision; for example, credible intervals for $(\boldsymbol{c}_0 - \boldsymbol{c}_1)^T \boldsymbol{\pi}(\boldsymbol{x}_i, \boldsymbol{\theta})$ could be compared to the decision boundary at zero.

## Acknowledgements

Thanks to Drs. Scott Holan (University of Missouri), Krista Heim (U.S. Census Bureau), and Dan Weinberg (U.S. Census Bureau) for valuable suggestions which helped to shape this manuscript. Thanks also to a number of other colleagues at the U.S. Census Bureau's Center for Statistical Research and Methodology for useful discussions on decision theory and its application to MAF maintenance.

## References

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

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.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 3rd edition, 2009.

Morris H. DeGroot. *Optimal Statistical Decisions*. Wiley-Interscience, 2004.

Krista Heim and Andrew Raim. Predicting coverage error on the Master Address File using spatial modeling methods at the block level. In *JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association*. (To appear, 2016).

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.

Yves Thibaudeau and Darcy S. Morris. Bayesian decision theory for further optimizing the use of administrative records in the Census NRFU. In *JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association*. (To appear, 2016).

U.S. Census Bureau. 2010 Census Address Canvassing Operational Assessment. In *2010 Census Planning Memoranda Series: 2010 Census Program for Evaluations and Experiments*. 2012.

U.S. Census Bureau. 2020 Census Research and Testing: 2015 Address Validation Test. 2015a.

U.S. Census Bureau. 2020 Census Detailed Operational Plan for the Address Canvassing Operation. 2015b.

U.S. Census Bureau. *American Community Survey Design and Methodology*, chapter 3. January 2014.

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*. (To appear, 2016).