Bayesian Decision Theory to Optimize the Use of Administrative Records in Census NRFU

Yves Thibaudeau * Darcy Steeg Morris†

Abstract

Morris, Keller and Clark (2016) and Brown (2013) propose an approach to identify reliable administrative records for enumerating the occupants of housing units in the context of the U.S. Decennial Census Non-Response Follow-up. We propose using Bayesian decision theory to extend the approach of these authors and account for costs and response propensities of field follow-ups. We elicit a loss function that emphasizes the importance of a correct enumeration for each unit. We exploit the properties of the loss function to make decisions between conducting new field follow-ups and utilizing administrative records to complete an enumeration. This leads to a general Bayesian decision theory problem. We attempt approximating the Bayes (optimal) solution of this problem through a version of "backward induction" (DeGroot 1970; Brockwell Kadane 2003). We give explicit formulas applicable to specific situations and derive possible strategies.

Key Words: Administrative Records, propensity, backward induction, NRFU, Decennial Census

Disclaimer

This paper is intended to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

1. Introduction

The non-response followup operation (NRFU) for the Decennial Population Census is a large-scale government field operation. For the 2010 Decennial Census more than 47 million personal visits and attempts at in-person enumerations -"field enumerations"- totaling about $1.6 billion (Walker et al. 2012) were carried through by professional staff to enumerate the households that could not be enumerated through a mail request or phone contact. This enormous task was carried methodically and successfully in past censuses. Nevertheless the rising costs of the logistics involved make the traditional formula prohibitively expensive (Vitrano Chapin 2011). Given this situation, the Census Bureau has sponsored several avenues of research. In particular, researchers are actively investigating the possibility of utilizing information extracted from existing administrative sources (Morris et al. 2016) to replace field enumerations. This gives rise to a case by case decision process whereby a decision whether attempting a field enumeration or accepting information from administrative records -"AR enumeration"- must be made. The critical issues are the cost of a new attempt, and the quality of the administrative information. In that respect, a successful field enumeration attempt is the gold standard, but the probability of a success may be not be high enough to justify the cost.

2. A Loss Function for Field Operation

To build our case for defining a formal decision rule it is instructive to consider simple and extreme cases of decision rules for choosing between attempting a new field enumeration

*US Census Bureau, Washington DC 20233

†US Census Bureau, Washington DC 20233

and substituting administrative information instead. One instance is the decision rule to always use the enumeration from the administrative records, regardless of field conditions. To adjudicate our decisions we propose a simple loss function

Consider NRFU, or a data collection stratum in NRFU, has N housing units, indexed by $i = 1, 2, ..., N$. Then let η_i be the true count for unit i and let d_i be the $decision$ for the count for unit i. We will consider decision rule that allows a decision d_n to be a function of the outcome of a previous decision d_m, where $m < n$. In that sense our approach is Bayesian as it allows for new decision to learn from prior information generated by previous decisions. This invites a formal application of Bayes Theorem in more complicated situations. But simple rules for updating decisions are available in basic situations and will be used in the paper instead.

Let $\lambda(d_i, \eta_i)$, the loss of making decision d_i when the true count is η_i, be defined as

$$\lambda(d_i, \eta_i) = \begin{cases} 0 & d_i = \eta_i \\ 1 & d_i \neq \eta_i \end{cases} \quad (1)$$

Therefore the loss is 1 if the decision d_i for η_i is incorrect and there is no loss if d_i is correct. We emphasize this loss function is designed to quantify the performance of the field operations. But it is not a comprehensive measurement of the accuracy of the Census count overall, in that the errors associated with the "incorrect" enumerations may not affect the error of the final count equally. The goal of the paper is to expand a theory useful for making field decisions. Measuring the final error comprehensively, whether it be undercount or overcount is beyond the scope of the paper.

Then the cumulative loss of all the decisions $\boldsymbol{d} = \{d_i\}$ for the counts $\boldsymbol{\eta} = \{\eta_i\}$ of the N NRFU units is defined by

$$L(\boldsymbol{d}, \boldsymbol{\eta}) = \sum_i \lambda(d_i, \eta_i) \quad (2)$$

$L(\boldsymbol{d}, \boldsymbol{\eta})$ is in fact the number of wrong decisions, or equivalently the number of wrong enumerations, for the N units in NRFU.

3. A Naive Decision Rule

Policy makers will determine the final "decision rule" for the field operations and to conduct NRFU. Our interest is to develop decision theoretical tools to enable policy makers to determine productive "decision rules" or "decision strategies." In that respect, it is useful to first examine over simplistic decision strategies when expanding the decision theoretical methodology.

Consider the decision rule $\boldsymbol{d}^{(0)} = \{d_i^{(0)}\}$, where the decision $d_i^{(0)}$ is to use the count extracted from the administrative records information corresponding to unit i. That is, the decision rule $\boldsymbol{d}^{(0)}$ is to always AR enumerate.

In the context of decision theory, it is possible to evaluate the "expected loss" of decision rule $\boldsymbol{d}^{(0)}$, given the loss function defined in (1) and (2). In that respect, let ρ_i by the "reliability" of the administrative information available for unit i. We define ρ_i as the probability that the administrative records yield the correct enumeration for unit i. Morris et al (2016) discuss methods to model this probability. Then the expected loss of always choosing an AR enumeration is

$$E[L\left(\boldsymbol{d}^{(0)}, \boldsymbol{\eta}\right) | \boldsymbol{\rho}\,] = \sum_i (1 - \rho_i)$$

4. A Decision Involving a Field Enumeration Attempt

The decision strategy $\boldsymbol{d}^{(0)}$ applies if no resources are available for making field enumeration attempts. This assumption is unrealistic in the case of the U.S. Decennial Census, as policy makers have consistently allocated significant funds in effort to successfully conduct as many field enumerations as possible. In our context, the goal of the field operations can be framed as an effort to conduct field enumerations while also taking advantage of administrative information to minimize the expected loss in (2) given the resources available. Before approaching the situation comprehensively, again it is useful to consider over-simplistic situations to further our understanding.

Consider the situation where resources are available for only one field enumeration attempt and we will assume the cost of such an attempt is the same for all units i. Examining this over-simplistic situation will help us understand the principles at work when deriving decision strategies for more general situations. Again we assume that a successful field enumeration attempt always leads to a correct enumeration. if the attempt is unsuccessful the decision rule reverts to the AR enumeration for unit j.

In this set-up, let $\boldsymbol{d}^{(j)} = \{d_i^{(j)}\}$ be the decision rule to attempt field enumerating unit i for $i = j$ and to use the information from administrative records for $i \neq j$. Also, let ϕ_i be the propensity associated to unit i. That is ϕ_i is the probability that the contact attempt and field enumeration are successful for unit i. Then, under the loss function defined in (1) and (2, the expected loss of $\boldsymbol{d}^{(j)}$ given the propensities $\boldsymbol{\phi} = \{\phi_i\}$ and the reliabilities $\boldsymbol{\rho}$ is

$$\begin{aligned} E[L\left(\boldsymbol{d}^{(j)}, \boldsymbol{\eta}\right)|\boldsymbol{\phi}, \boldsymbol{\rho}\,] &= E[L\left(\boldsymbol{d}^{(0)}, \boldsymbol{\eta}\right)|\boldsymbol{\rho}\,] + (1-\phi_j)(1-\rho_j) - (1-\rho_j) \\ &= E[L\left(\boldsymbol{d}^{(0)}, \boldsymbol{\eta}\right)|\boldsymbol{\rho}\,] - \phi_j(1-\rho_j) \end{aligned} \tag{3}$$

Observe the expected loss in (3) is minimized by choosing the unit j that maximizes

$$\phi_j(1-\rho_j) = \text{propensity} \times (1 - \text{reliability}) \tag{4}$$

We call $\phi_j(1-\rho_j)$ the "productivity" of attempting a field enumeration for unit j. So, holding reliability constant, productivity increases as propensity does. This reflects the good prospect for a successful field enumeration and a good investment of resources. At the same time, productivity decreases as the reliability of the administrative records increases. This is a reflection of misusing resources, in the sense that resources may not be optimally assigned when used to field-enumerate units whose count is reliably available from administrative information already.

The concept of the productivity of a field contact and enumeration attempt is central and we will explore the implications. At the same time, it is crucial to realize it is only a theory and other elements enter the decisions, such as policy issues, which we discuss in a later section.

5. One Wave of Field Enumeration Attempts.

Building on the simplistic example of the previous section we consider a simplified NRFU with a single wave of field contact attempts, whereby every unit in NRFU (or a in a data collection stratum) is the object of a single field enumeration attempt. Given c_k is the cost

of attempting a field enumeration for unit k, total resources C and the loss function defined in (1) and (2), the expected loss is minimized by maximizing.

$$\sum_{\{k\}} \phi_k(1-\rho_k) \tag{5}$$

over all possible set of units $\{k\} \subset \{1, ..., N\}$ subject to the constraint

$$\sum_{\{k\}} c_k \leq C \tag{6}$$

Note that if costs are constant, $c_k = c$, (5) implies field enumerations are virtually attempted in decreasing order of productivity until resources are exhausted, or every unit was subject to an attempt. In practice there are other constraints, such as mobility of the field workers from one unit to the next. So this approach would need to be implemented locally, as much as it is possible. The important aspect here is our analysis provides a general guiding principle, "productivity", for optimizing resource allocation.

Also note that if the propensities and costs are constant, $\phi_i = \Phi$ and $c_i = c$ for $i = 1, ..., N$, this principle reverts to the approach considered by Morris et al. (2016). That is, initially the set of attempts $\{k\}$ is the set of all NRFU units $i = 1, ..., N$. Then units are moved from $\{k\}$ by descending order of their reliability ρ_i until the cost constraint (6) is met.

We discuss in the next section how the "productivity principle" continues to be useful as an overall guideline in more complicated situations involving multiple attempts per unit.

6. Multiple Waves of Contact Attempts

The situation when multiple attempts are allowed quickly becomes complicated from a decision theoretical perspective. A formal treatment would require a more complex notation. However for our presentation we will make simplifying assumptions allowing us to keep notation as simple as possible for this teaser presentation.

We assume again constant cost. We also assume constant propensities for all units within a same wave of field enumeration attempts and data collection stratum. For simplicity and the purpose of illustration, we consider the situation where the number of attempts is limited to two per unit. Accordingly, let Φ^{W_1} be the propensity at the first attempt and Φ^{W_2} be the propensity at the second attempt for a unit in the stratum. For now consider a very simple situation. Assume there are two units in NRFU and there are resources to attempt two field enumeration attempts. In this setup, two decision strategies are available.

Strategy A, $\boldsymbol{\delta}^{(A)} = \{\delta_1^{(A)}, \delta_2^{(A)}\}$, is to proceed as if only one attempt per unit is allowed. First attempt field enumerating the first unit. If the attempt is successful $\delta_1^{(A)}$ is the field enumeration. If the attempt fails, the unit is AR enumerated, so $\delta_1^{(A)}$ is the administrative information. Then the same process is repeated for $\delta_2^{(A)}$.

Strategy B, $\boldsymbol{\delta}^{(B)} = \{\delta_1^{(B)}, \delta_2^{(B)}\}$, differs in that if the first field enumeration attempt for the first unit fails a second attempt is made for the first unit. So $\delta_1^{(B)}$ gets two opportunities to represent a field enumeration whereas $\delta_1^{(B)}$ gets at most one opportunity, only if the first field enumeration attempt for the first unit is successful. AR information is used whenever field information could not be obtained.

To apply these decision strategies (or rules) we assume the units are ordered by non-decreasing order of AR relieability, $\{\rho_i\}$. be the propensity at the second attempt, for either units. Appendix A shows, using a simple instance of Bayesian backward induction

(DeGroot 1970; Brockwell Kadane 2003), the expected loss of strategy 2 is less than the expected loss of strategy 1 if and only if:

$$\Phi^{W_1}(1-\rho_2) \leq \Phi^{W_2}(1-\rho_1) \tag{7}$$

This means strategy B is advantageous if the second attempt at contacting and field enumerating the first unit, when needed, is more productive than a first attempt at contacting and field enumerating the second unit. Since reliability is non-decreasing it implies strategy 2 is advantageous if the propensity at the second wave of attempt is equal or larger than the propensity at the first wave. That is

$$\Phi^{W_1} \leq \Phi^{W_2} \tag{8}$$

This result can be extended to more units. Meaning applying strategy B repeatedly has a lower expected total loss if equation (8) holds steady, which is the case if we assume constant propensities across wave. Note that we expect fewer units will be subject to field enumeration attempts on the whole if strategy B is followed rather than if strategy A. But the increased productivity leads to an increase in the expected correct enumerations.

7. Expected Number of Units Subject to Enumeration Attempts

Given the assumptions above, the decision strategies considered in the paper imply it is possible to estimate how many units will be the object of a field enumeration attempt and accordingly how many units can be excluded from the field work. Strategy A is straightforward. Under strategy B, for the units subject to contact attempts the expected number of attempts per unit is:

$$\Phi^{W_1} + 2\left(1-\Phi^{W_1}\right) = 2 - \Phi^{W_1} \tag{9}$$

Therefore, if there are N units and resources are available for M field enumeration attempts, we expect approximately the $N - M/\left(2-\Phi^{W_1}\right)$ units associated to the administrative information with the highest reliability will never be subject to contact attempts and can be excluded from field work at the onset.

It is important to realize that once the high reliability units have been removed from the field work, the order of the field enumeration attempts no longer matters.

Similar formula can be derived for other situations. But the general case, that is arbitrary numbers of waves and propensity patterns, is harder to track algebraically. In that case, backward induction can be used but is non-trivial.

8. Policy Issues

We have introduced a meaningful loss function and simplistic decision rules showing the usefulness and power of the concept of "productivity." According to this concept, decisions rules are productive when we minimize the expected number of incorrect enumeration which in turns implies using administrative records over field interviews if and only if the records are highly accurate.

It is important to realize policy issues may legitimately superseed the concept of "productivity" defined in (4) in the sense that additional value and utility may be assigned to field enumerations. The U.S. Census Bureau has traditionally emphasized the desirability of a "live enumeration", be that through a questionnaire returned by mail, or electronically, or an enumeration over the phone. Proxy and hot-deck imputation are very last resort means.

One possibility under consideration (U.S. Census Bureau 2020 Operational plan) is to attempt a field-enumeration for every unit once before considering doing any AR enumeration (U.S. Census Bureau, 2015). In that event, the theory presented here would become relevant only after that initial attempt.

9. Conclusion

We have introduced to concept of "productivity" in (4) as a guide for making field decisions during NRFU. We have derived algebraic formula for simple situations. More complicated patterns of propensity characterizing the waves of contact attempts may make algebraic onerous and likely require computational solutions. But some basic principles are emerging. Based on the principle of productivity, a scenario including successive waves of stable or increasing propensities through successive waves of field enumeration attempts suggests early removal of highly reliable units at the onset, and a focus on field efforts, repeated if needed, for units with low reliability.

By contrast, a scenario involving declining propensities through successive waves of field enumeration attempts suggests limiting initial removals and conducting broader rounds of enumeration attempts in the earlier waves to take advantage of the higher propensities. Under that scenario, cutoffs for AR enumerating high-reliability units could be introduced in later, lower propensity waves of attempts.

It is important to realize that efforts to optimize NRFU field strategy are coupled with intense efforts to optimize "self-response." (Blummerman, Dalpiaz Bishop, Dinwiddie 2016). Self response remains the preferred situation for retrieving information for the Census. After efforts to generate self-responses are exhausted, other efforts will be made to obtain the requested information directly through field enumeration from U.S. households, as discussed in this paper. The information obtained through self-response and field enumeration will be archived and will be available for review for policy makers and future generations. As such this information is extremely valuable. For Census 2020, policy makers may decide to use administrative records to supplement information obtained from self-response and field work because of financial or logistic constraints. In that eventuality our research could be helpful. Other research on imputation -a last resort method- based on administrative records is also being conducted (Keller 2016).

10. Appendix

Using a simple one-stage backward induction (DeGroot 1970; Brockwell Kadane 2003), under strategy A the expected loss is

$$E[L\left(\boldsymbol{\delta}^{(A)},\boldsymbol{\eta}\right)\Big|\Phi^{W1},\Phi^{W2},\boldsymbol{\rho}\,] \;=\; \left(1-\Phi^{W1}\right)(1-\rho_1)+\left(1-\Phi^{W1}\right)(1-\rho_2) \tag{10}$$

Under strategy B the expected loss is:

$$\begin{aligned} E[L\left(\boldsymbol{\delta}^{(B)},\boldsymbol{\eta}\right)\Big|\Phi^{W1},\Phi^{W2},\boldsymbol{\rho}\,] \;=\;& \Phi^{W1}\left(1-\Phi^{W1}\right)(1-\rho_2) \\ &+\left(1-\Phi^{W1}\right)\left(\left(1-\Phi^{W2}\right)(1-\rho_1)+(1-\rho_2)\right) \end{aligned} \tag{11}$$

We have

$$E[L\left(\boldsymbol{\delta}^{(A)}, \boldsymbol{\eta}\right) \Big| \Phi^{W1}, \Phi^{W2}, \boldsymbol{\rho}\,] \quad < \quad E[L\left(\boldsymbol{\delta}^{(B)}, \boldsymbol{\eta}\right) \Big| \Phi^{W1}, \Phi^{W2}, \boldsymbol{\rho}\,]$$

If and only if

$$\Phi^{W_1}\left(1-\rho_2\right) > \Phi^{W_2}\left(1-\rho_1\right) \tag{12}$$

REFERENCES

Blumerman, L. M., Dalpiaz Bishop, D., and Dinwiddie, J. (2016), " Plans and Innovations for the 2020 Decennial Census of the United States," *Statistical Journal of the IAOS*, 32(2): pp. 159-166.

Brockwell, A., and Kadane, J. (2003), "A Gridding Method for Bayesian Sequential Decision Problems," *Journal of Computational and Graphical Statsitics*, 12(3): 566-584.

Brown, J. D., Childs, J.H., and OHara, A. (2015), "Using the Census to Evaluate Administrative Records and Vice Versa", Federal Committee on Statistical Methodology (FCSM) Research Conference, Washington, DC.

Degroot, M. (1970) (Republished 2004). *Optimal Statistical Decisions*, Wiley.

Keller, A. (2016), "Imputation Research for the 2020 Census", *Statistical Journal of the IAOS*, 32(2): 189-197.

Konicki, S., and Adams, T.. (2016), "Adaptive Design Research for the 2020 Census" *Statistical Journal of the IAOS*, 32(2): pp. 167-176.

Morris, D., Keller, A., and Clark, B. (2016), "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census", *Statistical Journal of the IAOS*, 32(2): pp. 177-188.

U.S. Census Bureau. (2015), *2020 Census Operational Plan: A New Design for the 21st Century.*

Vitrano, F., and Chapin, M. (2011), "Possible 2020 Census Designs and the Use of Administrative Records: What is the impact on cost and quality?" Federal Commitee on Satistical Methodology (FCSM) Research Conference. Washington, DC.

Walker, S., Winder, S., Jackson, G., and Heimel., S. (2012), "2010 Census Nonresponse Followup Operations Assessment", 2010 Census Planning Memoranda Series, No. 190.