

2015 Administrative Records Modeling

Vincent Thomas Mule Jr., Administrative Records Modeling Team

September 3, 2014

1 Introduction

To meet the strategic goals and objectives for the 2020 Census, the Census Bureau must make fundamental changes to the design, implementation, and management of the decennial Census. These changes must build upon the successes and address the challenges of the previous Censuses while also balancing challenges of cost containment, quality, flexibility, innovation, and maintaining a disciplined and transparent acquisition decision process.

The 2020 Census is conducting a Research and Testing (R&T) Phase to obtain evidence-based decisions to address the challenges listed above. The objective of the R&T Phase is to develop preliminary designs based on solid evidence and a trade-off analysis aimed at achieving the goal of conducting the 2020 Census at a lower cost than the 2010 Census (per housing unit on an inflation-adjusted basis). Creating an evidence-based design requires making design decisions informed by the best insights that can be gathered—within schedule and budget constraints—on costs, benefits, and risks of different combinations of design options.

A primary decennial census cost driver is the collection of data from members of the public for which the Census Bureau received no reply via initially offered response options. Improving our methods for enumerating people who do not initially respond can contribute to a less costly census with high-quality results. The 2015 Census Test will allow the Census Bureau to evaluate the feasibility of utilizing the advantages of planned automation and available real-time administrative record data during the Nonresponse Followup data collection operations.

While self-response is encouraged, there will be households that do not respond; therefore, there is a need to test strategies to most effectively and efficiently collect information from those households. The 2015 Census Test will help us determine to what extent can administrative records information be utilized to eliminate or reduce the number of contacts for non-responding cases during the field workload in the NRFU operations. Administrative records can include information from federal, state or third-party sources. Examples of administrative records include Internal Revenue Service (IRS) Individual Income returns, Center for Medicare and Medicaid Services (CMS) Medicare Enrollment information or Undeliverable As Addressed (UAA) information from the United States Postal Service (USPS) related to Census Bureau 2015 Test mailings.

Section II provides a brief background on the 2010 Census Nonresponse Followup Operation. Section III provides a description of the three panels being tested during the 2015 Census Test. Two of the panels will involve using administrative records at different points. Based on available administrative records, this includes eliminating or reducing the amount of contacts,

determining the best contact time or possible usage for unresolved housing units after data collection ends. Section IV provides an overview of different approaches for using administrative records before or during NRFU to reduce contacts. Section V provides an overview of research about using administrative record information to help predict the best time to contact a case. Section VI provides a brief description about possible applications for unresolved housing units after NRFU is completed. Section VII provides a brief description about using administrative records for missing person and housing unit characteristics. Section VIII provides some questions for the committee.

II. 2010 Nonresponse Followup Operation

In the 2010 Census, the Census Bureau conducted the NRFU operation to resolve addresses after a self-response was unable to be obtained. The operation followed up on remaining units to determine if they were occupied, vacant, or non-existent. NRFU consisted of up to six visits by Census Bureau field representatives (FRs). After which, remaining unresolved units were imputed a status and household size via count imputation. The NRFU universe consisted of about fifty million addresses.

The 2010 Census Nonresponse Followup Operation occurred from mid-April until the middle of August and was composed of four field operations: Nonresponse Followup (NRFU), Nonresponse Followup Reinterview (NRFU RI), Nonresponse Followup Vacant Delete Check (VDC), and Nonresponse Followup Residual (Residual). The bulk of the 2010 Nonresponse Followup Operation was concentrated on the NRFU part. NRFU RI was designed as a quality check on NRFU enumerators' results. VDC verified vacant and delete housing units from the NRFU operation as well as a first time enumeration of some units that were not in the original NRFU universe. Residual (1) obtained population counts for units that were known to be occupied but where the count was unknown and (2) first enumerations that were not in the original NRFU or VDC universes.

Table 1 documents the final census resolution of cases in the 2010 Census Nonresponse Followup operation. These results include cases in any of the four Nonresponse Followup operations listed above. Of the 50 million housing units in the Nonresponse Followup Operation, 31 million or about 60 percent were determined to be occupied. The remaining 40 percent were determined to be unoccupied. Of those, 14 million were determined to have been vacant units and 5 million were determined to not meet the Census Bureau's definition of a housing unit. Some examples of non-existent units are businesses or uninhabitable units.

Table 1: Final Resolution Status of 2010 NRFU Operation Cases

Occupied	Vacant	Non-Existent Units
31 million	14 million	5 million

III. 2015 Census Test

The 2015 Census Test has the following objectives related to efficiency and effectiveness of the Nonresponse Followup Operations:

- Test methods of reducing field workload with administrative records;
- Test implementation of adaptive design fieldwork during data collection;
- Test the dynamic case assignment and optimization of case assignments against a traditional approach to NRFU;
- Test a new field management structure against a traditional approach to NRFU.

In order to test these objectives, the 2015 Census Test is implementing three panels. The first panel is a control panel. This panel will use contact strategies similar to what were used in 2010 to provide a comparison panel. While this panel is similar to 2010, it will utilize similar Research and Testing Operational Control System (RTOCS) and automated interviewing using a hand-held phone as was done in the 2014 test.

The second and third panels will utilize the Reorganized Census with Integrated Technology (ROCKIT) field management, administrative record usage to reduce contacts and adaptive design contact strategies. One main difference between the second and third panels is the usage of administrative records to eliminate or reduce the number of contacts during Nonresponse Followup.

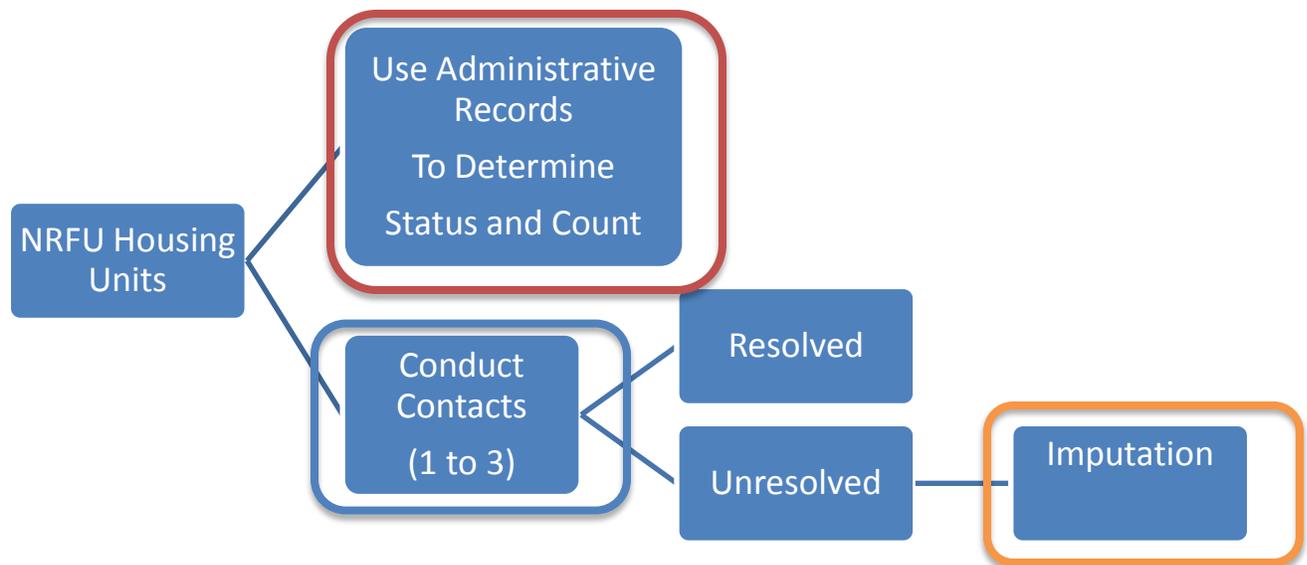
For these two panels, upcoming figures will show the flow related to administrative records usage and fieldwork. A red box indicates where administrative records are used to either eliminate or reduce the number of contacts for a unit. For these units, we have utilized administrative records to determine that the unit is either occupied with population, vacant or is a non-existent housing unit. Section IV will provide more information on this application.

A blue box indicates where field contacts are occurring. During this part of the operation, administrative records will be utilized along with other information to predict the best time to contact the unit. Section V gives a brief overview of this research.

Finally, the orange boxes identify units that are unresolved after NRFU is completed. These units may not have been able to determine if the status of the unit (occupied, vacant or non-existent) or may not have been able to determine the population count. This is another place with the potential usage of administrative records. Section VI provides more information on this research.

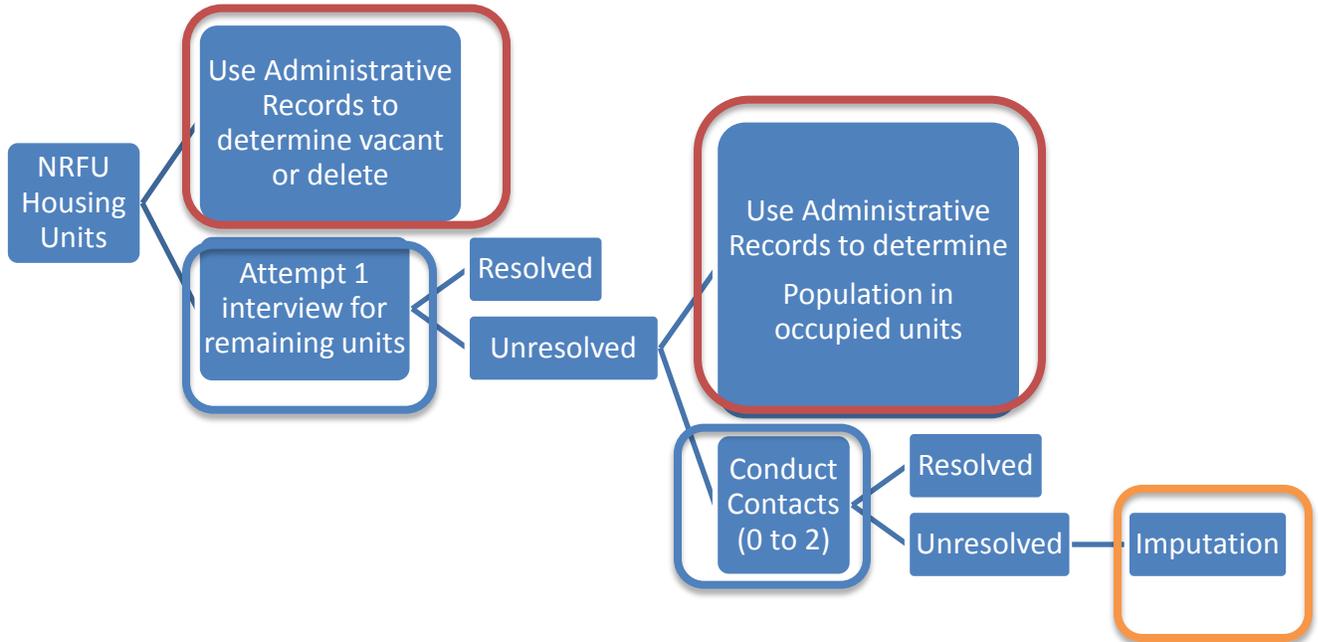
The second panel is the Full Administrative Record removal panel. In this panel, the goal is to use administrative records to determine the status (occupied, vacant or non-existent) and the count if occupied before fieldwork begins. Figure 1 shows the NRFU flow for the Full removal panel. Based on adaptive design approaches being researched, one to three contact attempts will be made to the units receiving fieldwork.

Figure 1: Full Administrative Records Removal Panel



The third panel is the Hybrid Administrative Records Removal panel. Figure 2 shows the flow of this panel. In this approach, administrative records are only utilized before NRFU if they are determined to be vacant or non-existent housing units. For the remaining units, one visit is attempted to the units. If the unit is still unresolved after one visit then at this point, we will utilize administrative records to determine those that are occupied with a population count based on administrative record information. For the remaining units after this point, up to two additional contacts can be conducted based on the adaptive design strategy. After fieldwork is completed, imputation will be utilized for the remaining unresolved cases.

Figure 2: Hybrid Administrative Records Removal Panel



IV Possible Usage of Administrative Records to Reduce Contacts

The previous section described the two experimental panels where administrative records would be utilized to reduce the number of contacts for units during the Nonresponse Followup operation. In this section, we will describe some of the approaches that we are researching. It will then briefly focus on how those approaches can be utilized to determine a) unoccupied units including vacant and non-existent and b) occupied units determined to have good household compositions. This section provides possible applications for the red boxes shown in Section III.

Rule-Based Assignment

This approach is based on rules of the people who are included in administrative record files and information that can indicate an address is vacant or non-existent. One example of a very simple rule is that if a NRFU housing unit record has people reported on an IRS 1040 return with less than 6 people then the unit will be considered occupied with that household. Another example is using United States Postal Service Undeliverable-As-Address information combined with other conditions to determine that a housing unit is vacant. We will be researching possible rule-based approaches based on analysis of 2010 Census data and 2014 Census Test results. One area where we have done modeling research to help determine rules is using decision trees. The final nodes of decision trees provide excellent sources of possible rules to implement.

Predictive Modeling

The Administrative Record Modeling Team is also researching using predictive models. Predictive models have been utilized to attempt to predict the probability of a desired outcome. The predicted probability can be utilized to make decisions about whether we should continue to conduct fieldwork or use the administrative records instead. Our team has researched several predictive modeling approaches including:

- logistic and multinomial regression,
- decision trees
- random forests

For predictive modeling, we need to identify a suitable set of training data, a desired outcome to model and a set of covariates available. Representative training data is used to determine how desired outcome is associated with possible information about the unit.

In our application, we have several possible covariates including:

- Census Bureau Master Address File (MAF) information available for the housing unit,
- Person and housing unit-level information based on administrative record sources and
- ACS estimates for block groups and tracts

With the predictive modeling approach, you need to define an outcome variable for the modeling. We have and continue to examine several different possible outcome variables to use in our predictive models.

Two outcomes that have been used are:

- a 2010 Census result:

An example of using a 2010 Census result as an outcome is using if a housing unit was determined to be vacant or not. This allows us to develop a predictive model of determining that a unit would be vacant. Decisions can be made on new incoming data based on determining those to be vacant based on if the predicted probability is greater than a specified cutoff. Morris (2013 upcoming) has done work for our team on using Receiver Operating Characteristic (ROC) curve information to help determine various possible cutoffs.

- Determine if the administrative record usage matches a 2010 census result.

Another possibility is to assess if the administrative record usage is matching a 2010 Census result for the unit. One outcome along these lines is if the administrative record housing unit count matches the census count. Similar to the vacant model in the previous bullet, a predictive model of this outcome can be developed. A housing unit could be determined to be occupied with a particular household based on having a predicted probability above a specified cutoff.

One thing that we do keep in mind when doing these predictive models is the possible reasons for the administrative records not matching the 2010 Census result. One possibility is that the administrative records include someone who should not be included in the Census at the unit. That person might have moved somewhere else. In that instance, we do not necessarily want to include that person. A second possibility is that the administrative record usage might not match the 2010 result because it could be addressing a potential error in the 2010 Census. One example is that Demographic Analysis estimated an undercount of children. Using an administrative record source that provides additional new children might end up having a lower predicted probability of the count agreeing.

Household Population for Occupied Units

For occupied units based on administrative records, one of the challenges is to determine the household population for the unit. One way of determining this is based on rule-based processing. Rules are developed to determine which people associated with the unit to use. The selection will be based on the possible sources being considered. Rules can be developed to associate a person with a specific housing unit if they are associated with multiple units. The selected people would be used for the unit. An alternative to determine the population count is to use an imputation allocation of the population count and household. Instead of using people directly associated with the unit, this alternative would draw a population count based on a distribution of counts. This alternative could address if there are any possible concerns about directly using administrative records associated with a unit.

Possible Utilization To Reduce Contact Examples

Table 2 documents a simple example of two possible utilizations for vacant housing units. A very simple rule is given. The 2014 had additional rules that included multiple Undeliverable-As-Addressed reasons from the mailings. Table 3 documents a simple example for discussion purposes of three possible utilizations for occupied units. These three possible utilizations have different combinations of rule-based, predictive models or imputation.

Table 2: Possible Utilizations of Vacant Housing Units

Option 1: Rule-Based Approach	Option 2: Predictive Model
<p>Example:</p> <p>If</p> <p>USPS Undeliverable-As-Address indicates Vacant and</p> <p>there is no person records associated with the unit from administrative record sources</p> <p>Then determine the unit to be Vacant</p>	<p>Example 1:</p> <p>If the predicted probability of the unit being vacant is greater than or equal to a cutoff</p> <p>Then determine unit to be Vacant</p>

Table 3: Some Possible Utilizations for Occupied Housing Units with Population (Before or During NRFU)

	Possible Option 1	Possible Option 2	Possible Option 3
Occupied Determination	<p>Rule-Based Approach</p> <p>Example:</p> <p>If an IRS return filed by April 30th or Medicare Beneficiary associated with unit</p> <p>And No USPS UAA information</p> <p>Then considered occupied</p>	<p>Predictive Model</p> <p>Examples</p> <ol style="list-style-type: none"> 1. Predicted probability of the unit being occupied \geq cutoff <p>Or</p> <ol style="list-style-type: none"> 2. Predicted probability of the administrative record count matching the census count \geq cutoff 	<p>Predictive Model</p> <p>Examples</p> <ol style="list-style-type: none"> 1. Predicted probability of the unit being occupied \geq cutoff <p>Or</p> <ol style="list-style-type: none"> 2. Predicted probability of the administrative record count matching the census count \geq cutoff
Household Composition Given Occupied	<p>Rule-Based Approaches</p> <ul style="list-style-type: none"> • Rules to determine which people associated with the unit to use from several possible sources 	<p>Rule-Based Approaches</p> <ul style="list-style-type: none"> • Rules to determine which people associated with the unit to use from several possible sources 	<p>Imputation</p> <ul style="list-style-type: none"> • Impute a population count and household based on similar housing unit characteristics

Note: Need to determine both occupied and have a good household composition to use in each column.

Coverage Improvement Using Administrative Record Sources

The Census Bureau is researching to determine the administrative record sources to utilize. One question is what sources are needed either before or during NRFU to make our necessary status and population determinations so we can stop fieldwork. One possibility is that a smaller number of sources could be utilized during the NRFU portion. Additional sources could be utilized after NRFU similar to a coverage improvement operation to produce the final population for the unit. By using a coverage improvement approach, there is the possibility that federal, state or other administrative record providers could better provide us with information related to the April 1st status of the unit. This might be a more acceptable alternative as compared to asking them to provide that same information either before or during the NRFU operation. For the 2015 test, NRFU fieldwork will start on May 15th but no determination has been made for when NRFU fieldwork will start in 2020.

Evaluation of Usages

As part of testing different potential usages during our research, our team evaluates the results at both macro- and micro-levels. Macro-level comparisons involve calculating population, housing unit and demographic totals based an implementation of the potential usage. This allows us to compare these alternative totals to the 2010 Census results. We can also do comparisons to 2010 Census Coverage Measurement results. This allows us to assess if potential utilizations might be addressing coverage differences like the underestimation of young children for example. Population and housing unit results can be generated at national, state, county, tract and block level. We implement approaches and summary measures that allow us to assess numerical and distributional differences.

In addition to macro-level comparisons, we do calculate and summarize micro-level comparisons. These can include comparisons of how often the population count agrees with the 2010 Census for example. We can also do comparisons of status determinations of units being occupied, vacant or deletes. Aggregating these allow us to summarize comparisons at the housing unit level.

V. Usage of Administrative Records Modeling During NRFU for Contact Time Modeling

Section III gave an overview of two experimental panels in the test. For these two panels, the Census Bureau is testing a new field reengineering and control system as part of the Field Reengineering system. Figure 1 and Figure 2 have blue circles around the boxes corresponding to fieldwork. In determining daily workloads for enumerators, this control system is utilizing what is the best-predicted time to attempt to contact the unit.

The Administrative Records Modeling team is working with the ROCKIT team to develop a predictive model for possible contact times. This work is utilizing contact attempt and resolution information from the American Community Survey Computer Assisted Personal Interviews. Our outcome variable in our initial research is groupings of days and times during the week. One example is weekday 9 to 3, weekday 3 to 6, weekday 6 to 10, weekend 9am to 12, weekend 12 to 3pm and weekend 3pm to 10pm. We are researching if we can build a predictive model to indicate that it might be better to contact the unit during one of those times. As the research progresses, we can see if we can expand to potentially more groupings.

For building these predictive models, we are taking advantage of potential covariates based on administrative records available for the unit. Some hypotheses are that you could attempt to contact people with young children during the day because they may be more likely to be home as compared to single person households. A hypothesis for single person households is that contacting them during the evening might produce a better chance for a successful contact. In this work, we are utilizing available administrative records to make covariates related to the household composition observed in those sources. We are combining these administrative record information with information that we have about the unit to see if that can help improve our predictions of contact times.

VI. Potential Usages of Administrative Records for Unresolved Units After NRFU

There may be some units where we were not able to determine the occupancy status and population. For these instances, our team is researching how administrative records can be utilized for these units after NRFU is completed. This section address using administrative records for the orange boxes laid out in Section III.

One possibility is to utilize administrative records that might have become available after the NRFU period finished. One example might be obtaining person records associated with applications to a federal or state program that were submitted near April 1st. The information may not be able to be delivered to the Census Bureau in time for use during our NRFU operation but may be able to be made available to us later in the calendar year. One option under this situation is to utilize this information for unresolved housing units. One example is a rule-based approach to make a housing unit occupied with the corresponding population composition based on this post-NRFU information. Predictive modeling approaches could also be utilized.

A second possible utilization of administrative record information is during the count imputation process. The count imputation process assigns a final status and corresponding population count to unresolved units. Our team is researching making modifications to the 2010 approaches. One of the modifications is to use administrative record information as covariates during the imputation. One example is the presence of person records at the unit from administrative record

sources like an IRS tax return. This has shown to be a good covariate to help discriminate occupied units.

VII. Potential Usage of Administrative Records for Missing Person and Housing Unit Characteristics

The Administrative Records Modeling and Fitness for Use teams are researching ways that demographic characteristics can be obtained from already provided information to federal, state or third-party sources. Some administrative record sources may not have race or Hispanic origin information. For these instances, we are researching if that information can be obtained from either Past Census responses or other administrative record sources. We are researching if age or sex information can be obtained from sources like the Social Security Administration Numident file. These are some examples of possible usages along these lines.

VIII. Discussion Questions

What is your reaction to how we could possibly use administrative records to eliminate or reduce the number of contacts during NRFU?

What is your reaction to how we are using administrative records to help determine the best time to contact units during NRFU?

What is your reaction to how we could possibly use administrative records for unresolved units after NRFU is completed or for missing characteristics?