UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

2020 CENSUS PROGRAM INTERNAL MEMORANDUM SERIES: 2019.10.i

Date: April 8, 2019

MEMORANDUM FOR: The Record

From: Deborah M. Stempowski **(signed April 8, 2019)**
Chief, Decennial Census Management Division

Subject: 2020 Census Experiment: Real-Time 2020 Administrative Record Census Simulation Study Plan

Contact: Jennifer Reichert
Decennial Census Management Division
301-763-4298
jennifer.w.reichert@census.gov

This memorandum releases the final version of the 2020 Census Experiment: Real-time 2020 Administrative Record Census Simulation Study Plan, which is part of the 2020 Census Program for Evaluations and Experiments (CPEX). For specific content related questions, you may also contact the authors:

J. David Brown
Center for Economic Studies
301-763-8769
j.david.brown@census.gov

Misty Heggeness
Associate Directorate for Research and Methodology
301-763-7251
misty.l.heggeness@census.gov

United States Census 2020

2020 Census Experiment

Real-Time 2020 Administrative Record Census Simulation Study Plan

J. David Brown, Center for Economic Studies
Misty Heggeness, Associate Directorate for Research and Methodology

United States™ Census Bureau

4/08/2019
Version 5.0

Page intentionally left blank.

Table of Contents

Page intentionally left blank.

I. Introduction

This project will conduct several real-time administrative record census simulations in 2020, using all administrative records ingested by the Census Bureau as of a certain date that can inform about the Census Day population. If extended, the project will produce annual real-time administrative record census simulations after 2020 and provide predictions for where survey-style data collection will be necessary to complete an administrative record enumeration in 2030.

Our project will build on past administrative record enumeration research by including additional administrative data sources not previously available, using more accurate and comprehensive person linkage, and employing more powerful models predicting people's locations to increase coverage and person-place accuracy.[1] Unlike previous studies, it will be conducted in real time in 2020, which will show how the population statistics compare between an administrative record census and survey-style collection, in the same time frame.[2] It will also show how long it takes to execute an administrative record census and what the most time-consuming parts of the process are. The project will compare person-level, housing unit-level, and hybrid approaches to conducting an administrative record census, which will inform 2030 design decisions about whether to transition from a housing unit- to a person-based or hybrid method within the legal governance, rules, and regulations of conducting a full count census.

The project will expand upon and evaluate 2020 Nonresponse Followup (NRFU) operation administrative records innovations. We will use many more data sources, an enhanced Person Identification Validation System (PVS) process, and different methods for assigning people to locations relative to those planned for administrative record enumeration in 2020 NRFU. We will also evaluate the PVS linking methodology planned for use in 2020 operations by estimating false match and nonmatch rates and the effect of those errors on the statistics.

Models will be developed to predict where each person is located on Census Day, as well as where administrative record enumeration is most likely to diverge from survey-style enumeration. The latter can be used to target where survey-style data collection can most usefully supplement administrative record enumeration.

Comparisons of various administrative record census simulations to the 2020 Census will show the relative strengths and weaknesses of administrative record coverage and accuracy[3] by demographic characteristics, location, and level of geographic aggregation. The results from person-based, housing unit-based, and hybrid approaches to constructing an administrative record census will be examined. The effects of enhancing the record linkage infrastructure will be

[1] It may not be possible to use these improvements in the 2020 production NRFU, because the deadlines to finalize 2020 production methods are much earlier than those for this experiment.
[2] Doing the simulations in the same timeframe will demonstrate the feasibility of conducting an administrative record census in similar conditions to the actual census. Relative to a post-2020 census study, this will also reduce concern that the administrative record similations will borrow from the 2020 Census, tainting the experiment.
[3] To assess coverage, we will compare how many people are enumerated regardless of whether they are the same people in the same places; to assess accuracy, we will examine whether people are enumerated in the same places.

identified. We will study how well the models predict where administrative record enumeration and the 2020 Census agree and diverge.

If extended beyond 2020, the project will conduct annual real-time simulations and produce annual population estimates, which will be compared with other Census Bureau intercensal population estimates. This analysis will inform decisions on whether to continue researching and conducting administrative record enumeration on an annual basis after 2020. The overarching questions are whether we are able to conduct a complete administrative records census and with what level of accuracy. If so, what methods are best both in terms of infrastructure and modeling techniques for implementing a strategy that will ensure the most accurate count of the population. Our project will address several subquestions to inform these overarching questions:

- Can a record linkage methodology accurately cover people not in the Numident or Individual Taxpayer Identification Numbers (ITINs)? What is its error rate?
- What are the best models for predicting which administrative record address is the Census Day address, and how well do they perform?
- Which data sources are the most useful for coverage and accuracy and thereby pursued with the highest priority?
- How similar are the statistics produced by an administrative record census to the survey-style 2020 Census?
- How long does it take to implement an administrative record census, and what parts of the process are most time-consuming?
- Which method for conducting an administrative record census can produce statistics most similar to the 2020 Census?
- What are the cost-statistical similarity trade-offs when using different combinations of survey collection and administrative records?
- What subpopulations and geographic areas should be targeted for alternative forms of data collection?

Finally, we will refine models of person and housing unit transition rates. Repeated annual records-based censuses will facilitate the production of more powerful predictors of where survey-style data collection is needed to complete administrative record enumeration.

II. Background

To reduce costs, many countries use administrative data to assist in censuses or as a replacement to traditional censuses (Farber and Leggieri 2002, Ralphs and Tutton 2011). For several decades administrative data have been used in U.S. Census Bureau programs for population, economic, small-area income and poverty, and health insurance estimates, but they have not been used extensively in decennial census operations. For the 2020 Census, administrative records will be used to reduce NRFU fieldwork, one of the largest expenses in the decennial census.

There is pressure to continue to improve the accuracy and reduce costs of a decennial census beyond the planned administrative record use in 2020. The JASON (2016) report “suggests a

paradigm shift in the way the Census Bureau conceptualizes the enumeration… a census that starts with administrative records involves identifying individuals and assigning them to their appropriate residences as opposed to the historical process of identifying residences and then populating them." As such, the report recommended that "the Census Bureau consider starting the 2030 Census with an 'in-office' enumeration of the population using existing government administrative records."

Previous research efforts have evaluated the feasibility of a 100 percent records-based census. The Statistical Administrative Records System (StARS) was developed from select federal data sources in 1999. Decennial census research used these data to evaluate address and person counts relative to Census 2000, and for a field test (the Administrative Records Census Experiment or AREX 2000) that simulated a census in several counties that was compared with Census 2000. The research found that while address and person counts in StARS were relatively close to the counts in Census 2000 at the national level, results varied significantly by region (Farber and Leggieri 2002). The AREX 2000 research compared Census 2000 results in five counties with administrative data in StARS and found that the administrative data undercounted children, overcounted elderly populations, had difficulty identifying a correct residence of movers, and that a 15-month time gap between the administrative and census data likely contributed to the difficulties of using administrative records to enumerate the population (Bauder and Judson, 2003).

The AREX 2000 research compared a person-based approach and a hybrid person- and housing unit-based approach to constructing an administrative record census. In their person-based approach, they assigned each person to a single block. Their hybrid method assigned each person to their "best" address according to the StARS algorithm, provided that it was included in the 2000 production Master Address File (MAF). Production MAF housing units lacking people in the administrative record census were selected for follow-up survey-style data collection, and their Census 2000 population count was used in the simulation (making the results the same in the Census 2000 and the simulation for those housing units).

The 2010 Census Match Study (Rastogi and O'Hara, 2012) linked person, address, and person-address records to 2010 Census data to assess the quality and coverage of administrative data and feasibility of a records-based census. The study showed significant improvement over the AREX 2000 results in matching addresses found in administrative records to addresses in the 2010 Census (92.6 percent). It was also able to match 88.6 percent of all individuals in the 2010 Census to at least one administrative record, and 77 percent were placed at the same address. The report also evaluated quality and coverage by Hispanic origin, race, sex, and age response data in administrative records relative to the 2010 Census.

While more timely and varied sources of data were available than in the StARS data, the 2010 Census Match Study findings reaffirmed the challenges of conducting a records-based census. Some data sources such as Social Security Administration (SSA) and Medicaid data did not have addresses, and other data sources had addresses that include a post office box or other nonresidential addresses, both of which making it difficult to place all people enumerated in administrative records at a residential location. Furthermore, some individuals were associated

with multiple addresses, and the address selected often did not align with the address in the 2010 Census.

Our project will build on the AREX 2000 and 2010 Census Match Study by including additional administrative data sources, assigning unique identifiers more accurately to more people, and employing more powerful models predicting people's locations, which should result in increased coverage and person-place accuracy. Unlike previous studies, it will be conducted in real time in 2020, in parallel with the actual census. The project will compare person-level, housing unit-level, and hybrid approaches to conducting an administrative record census, which will inform 2030 design decisions about whether to transition from a housing unit- to a person-based or hybrid method. We will simulate doing a census that is 100 percent records-based, supporting that 2030 guiding principle. The simulation tabulations will produce state population counts to fulfill the constitutional mandate for apportionment and citizen voting age population by race and ethnicity at the block level to fulfill the Voting Rights Act requirement.

The project will build upon and evaluate 2020 Census innovations regarding the use of administrative records in NRFU. This project will use a greatly expanded set of data sources and person validation, as well as different methods for assigning persons to locations than those planned for administrative record enumeration in 2020 Census NRFU. We can compare our simulation results with the administrative record enumeration in NRFU for the same housing units. We can also investigate whether and the extent to which administrative record enumeration could be expanded without sacrificing quality.

The interventions with 2020 Census processes needed by the project are access to the 2020 Census production MAF, Decennial Response File (DRF), Census Unedited File (CUF), Census Edited File (CEF), and Post-Enumeration Survey (PES) as soon as they are completed.

III. Assumptions

1. The project team will obtain and maintain adequate staff resources.

2. The Internal Revenue Service (IRS) will approve the Predominant Purpose Statement (PPS) for use of Federal Tax Information (FTI) in a timely manner.

3. The project team will obtain adequate funding for computing resources to begin the project in the Integrated Research Environment (IRE).

4. The Census Bureau will fund and complete the necessary information technology (IT) requirements to move Title 26 data to the cloud.

5. The Census Bureau will continue to acquire administrative records such as Internal Revenue Service (IRS) and state program data and acquire new key data sources such as passport and visa data in a timely manner.

IV. Research Questions

1. Can a record linkage methodology accurately cover people not in the Numident or ITINs? What is its error rate?
2. What are the best models for predicting which administrative record address is the Census Day address, and how well do they perform?
3. Which data sources are the most useful for coverage and accuracy and thereby pursued with the highest priority?
4. How long does it take to implement an administrative record census, and what parts of the process are most time-consuming?
5. How similar are the statistics produced by an administrative record census to the survey-style 2020 Census?
6. What are the cost-statistical similarity trade-offs when using different combinations of survey collection and administrative records?
7. Which method for conducting an administrative record census can produce statistics most similar to the 2020 Census?
8. What subpopulations and geographic areas should be targeted for survey-style data collection?

V. Methodology

This section describes how we plan to address the research questions, followed by a discussion of the implications of the results for future testing and 2030 Census design decisions.

A. Evaluation Design

The main steps we propose to implement in the research are the following:

1. Obtain additional sources of administrative records beyond the Census Bureau's current inventory.
2. Enhance the record linkage infrastructure with additional data and methodological improvements.
3. Test the validity of assumptions incorporated in PVS, and test the application of modern entity resolution models of record linkage.
4. Process administrative records.
5. Estimate person-place models.
6. Predict the relative accuracy of administrative record enumeration by person and housing unit, including the error from linkage.
7. Conduct real-time administrative record census simulations in 2020.
8. Assess quality and coverage of the administrative record censuses in comparison to the 2020 Census.[4]

[4] We use the 2020 Census as comparison data, recognizing that they are not error free. We plan to also use the 2020 Post-Enumeration Survey (PES) as an additional comparator for statistics that the PES produces (nationally by race,

We will make a number of improvements in the administrative record infrastructure, including obtaining additional administrative data sources to reduce coverage gaps and improve model prediction, enhancing record linkage procedures to link more records, and improving model estimation techniques. Unless significant progress is made on these fronts, the discrepancies identified in the 2010 Census Match Study (Rastogi and O'Hara, 2012) compared with the 2010 Census are likely to remain too high to change that study's conclusion that the Census Bureau is not yet ready to convert to using administrative records as the primary enumeration method.

These steps are described below:

Obtain additional sources of administrative records: We will obtain and integrate additional administrative record sources to plug coverage gaps and improve the predictive power of the models. The 2010 Census Match Study (Rastogi and O'Hara, 2012) and this project's analysis of differences between administrative record enumeration and the 2020 Census can inform us where additional sources can be most beneficial. We will conduct analysis of the relative contributions of the current sources to coverage and predictive power, which will inform decisions on which current sources could be dropped to free up funds for new acquisitions, if necessary.[5]

New sources that could be particularly valuable include state driver's licenses; voter registration data; state-level low-income assistance program participation; and data sources provided by local governments.[6]

As additional sources are added, the person-place models will be reestimated both so that people found only in the new sources can be included and to harness this information to improve prediction.[7]

Enhance the record linkage infrastructure with additional data and methodological improvements: One of the most important reasons for administrative record coverage gaps is the fact that several million U.S. residents are either not in the SSA Numident, or the personally identifiable information (PII) in their other administrative records is different from how it appears in the Numident. We will add U.S. Citizen and Immigration Services (USCIS) legal permanent resident and naturalization data, Customs and Border Protection (CBP) visa data, and State Department passport data to the reference files,[8] covering some of the people not in the Numident.[9]

ethnicity, and age group). Below we discuss separate comparisons with 2020 Census records for which there are different probabilities of error. Note that in NRFU housing units enumerated via administrative records in 2020 Census production, comparisons with survey-style data collection will be limited to the PES.

[5] Decisions to discontinue sources will need to factor in other Census Bureau needs and uses of those sources.

[6] The availability of data sources for the project will be dependent upon maintaining and in some cases revising existing agreements with current data providers and obtaining them for new sources.

[7] Note that we will want to reestimate the models periodically anyway, because relationships between administrative record and survey-style enumeration could change over time. Our ability to do this throughout the decade will depend on continuing survey-style collection of ACS housing unit roster information.

[8] Reference files are the files used to validate the PII in a person record. The current reference files in PVS are the Numident and ITINs.

[9] Note that the intention to do this is already public knowledge, as the March 1, 2018, memo from the Census Bureau to Commerce Secretary Wilbur Ross mentioning this has been made public as part of a FOIA request.

As a way to facilitate the linkage of the remaining people who are in administrative records, but not in any of the reference files listed above, we will experiment with the BigMatch linkage procedure developed by William Winkler, as well as other entity resolution approaches (Steorts et. al. 2016). These approaches could potentially be more accurate as well. The methods are well suited to linking records across multiple files. The Steorts et al. (2016) approach is implemented by fitting models using Bayesian methods. The results of the estimation include posterior probabilities that records on incoming files are of individuals not already on the existing reference files.

Test the validity of PVS assumptions and the application of modern entity resolution models: We will evaluate the PVS and entity resolution methodologies by estimating false match and false nonmatch rates and the effect of those errors on the statistics (e.g., the population count in a geographic area or the percentage of the population with particular demographic characteristics).

Process administrative records: The Census Bureau will ingest administrative records, PVS the persons to assign Protected Identification Keys (PIKs)[10], and MAF-match the addresses to assign MAFIDs, latitude and longitude, and other geolocational codes. We will start with all administrative data that are currently available containing personally identifying information (PII) and geographic location.[11] People that the Numident or other reliable sources indicate are deceased will be dropped. We will use PIKs to have some confidence that the person exists and to be able to unduplicate the person's records to prevent multiple counting.

Estimate person-place models: We will estimate person-place models to produce a probability that a PIK is at a particular location, for each PIK-location pair (or person-address pair, where the person has been assigned unique identification number, and the address has a MAFID or other geocodes).[12] We will use ACS panels for the same years as the vintages of the administrative

[10] As mentioned above, we will test other methods for assigning unique person identifiers as well.

[11] A candidate list includes Internal Revenue Service (IRS) Individual Income Tax Returns 1040 and Information Returns 1099; Housing and Urban Development (HUD) Public and Indian Housing Information Center (PIC) and Tenant Rental Assistance Certification System (TRACS), and Computerized Homes Underwriting Management System (CHUMS); Social Security Administration Supplemental Security Record (SSR), Numident, and the Kidlink file derived from the Numident; Center for Medicare and Medicaid Services (CMS) Medicare Enrollment Database (MEDB) and Transformed Medicaid Statistical Information System (T-MSIS); Indian Health Service (IHS) Patient Registration System; U.S. Postal Service National Change of Address file; Experian; Targus/Neustar; Veteran Service Group of Illinois (VSGI); InfoGroup; Melissa Data; Health and Human Services Child Care Development Fund (CCDF); Bureau of Justice Statistics National Corrections Reporting Program (NCRP) and Post-Custody Community Supervision (PCCP); Bureau of Prisons Permanent Release Database; Veterans Affairs; Alaska Permanent Fund Dividend File; Supplemental Nutrition Assistance Program (SNAP); Temporary Assistance for Needy Families (TANF); Special Supplemental Nutrition Program for Women, Infants and Children (WIC); Homeless Management Information Systems (HMIS); Low-Income Home Energy Assistance Program; utilities records data; Corelogic; RealtyTrac; and DAR Partners.

[12] This estimation process will be repeated using different definitions of location (housing unit, block, tract, ZIP code, county, state, and latitude and longitude coordinates with different degrees of precision). By doing this we can investigate the possibility that discrepancies between administrative record and survey-style enumeration vary by geographic level. Since most moves are within small geographic areas, person-place discrepancies are likely to be much smaller at higher levels of geography.

record data to fit the models.[13] First-stage logistic regressions will be estimated separately for each administrative record source. PIKs in the ACS that have administrative records in the particular source will be included in the regression. The dependent variable equals one if the location in the administrative record is the same as the ACS location for the PIK, and zero otherwise. We will estimate separate regressions for each source, because variables that can help predict if the location is the ACS location vary across sources. For example, IRS 1040's contain variables for filing status, whether a child is living elsewhere, and the week the return was processed (measuring the vintage). Veteran Service Group of Illinois (VSGI) contains household income, owner vs. renter status, and length of residence. For sources with several years of data, such as IRS 1040s, we will construct variables for whether the person was at this location or a different one in past years to capture the person's mobility. Variables on the person's age, sex, race/ethnicity, and citizenship status (primarily from the Numident) will be included in all these regressions.

A second-stage regression includes all administrative record PIK-location pairs for PIKs in the ACS that have administrative records in any of the administrative record sources. Once again the dependent variable equals one if the PIK-location administrative record pair is the same as the ACS PIK-location pair for each PIK, and zero otherwise. Characteristics of the administrative record location are included here, such as the housing unit type, U.S. Postal Service delivery sequence file information, and the number of other PIKs with administrative records with this location[14]. We will include indicator-indicator variables for whether each particular administrative record source lists the person at this observation's location (here) or at one or more other locations (elsewhere). These indicator-indicators are also separately interacted with the individual match propensities obtained from the first-stage regression corresponding to the indicator source for the PIK-location pair.[15] The rationale for the interactions is that the location where a source lists a person should be more likely to be the ACS location if the first-stage match propensity is high. Including indicators for each of the sources captures the degree of agreement across sources about the person's location. The coefficients from these models are used to produce person-place match probabilities for all PIK-location pairs eligible for administrative record enumeration.

We will control overfitting the models by performing k-fold cross-validation. It splits the data randomly into k partitions. For each partition it fits the model using the other k-1 groups, then uses the generated parameters to predict the dependent variable in the unused group. We will also test the models by using the parameters to predict the dependent variable in future years of ACS data and the 2020 Census.[16,17]

[13] The models implicitly assume that the ACS household roster is accurate, which may not be the case. We are unaware of a more accurate alternative, however.

[14] Large numbers of people may have the same administrative records address, e.g., people using recreational vehicle (RV) mail forwarding services. Including this variable should help address this issue - their person-place probabilities are likely to be low, making them top candidates for follow-up.

[15] In cases where the administrative record source has multiple other MAFIDs for the PIK, we will sum up the propensities for the other MAFIDs here.

[16] For example, we can apply models fit on 2017 ACS data to 2018 administrative records to produce probabilities that people in the 2018 ACS are located at various addresses. Then we will see how well those probabilities predict the actual 2018 ACS address for the person. The same thing will be done with the 2020 Census, with the added benefit of being able to do it on the full population rather than a survey sample.

[17] Note also that unlike in Rastogi and O'Hara (2012) and Brown, Childs, and O'Hara (2015), all the model coefficients will be applied to future administrative records, not to earlier administrative records used to fit the models.

Predict relative accuracy of administrative record enumeration: We will produce several different predictors, some geared toward person-based follow-up data collection and others for location-based follow-up. For each PIK, we will calculate the maximum probability of being at a particular location among all their locations in administrative records, using predictions from our person-place models fit with earlier data vintages. This predicts the probability that if counted in this location in an administrative record census, the person will not be an enumeration error (counted in the wrong place). Any person-based follow-up data collection could focus on individuals with the lowest values for this measure. For location-based follow-up, we will calculate a location-based enumeration error predictor that again uses the PIK maximum location probability described above, but now taking its average value across all PIKs at the location.

To predict where omissions are more likely, we estimate housing unit-level administrative record coverage regressions. The dependent variable when fitting the model for one version is equal to one if at least one person in the ACS household roster cannot be linked to the set of administrative records we plan to use in the simulations. A second version is a count regression for the number of people in the ACS household roster who cannot be linked. Explanatory variables include the number of un-PIKed administrative records with this location, indicators for the sources of the un-PIKed records, the number of PIKed persons with this location, demographic characteristics of the un-PIKed administrative records, and characteristics of the location, such as housing unit type and U.S. Postal Service delivery sequence file information.[18] The coefficients from these regressions will be used to produce out-of-sample predictions of incidence and number of omissions for all locations. Another measure we will use is the standard deviation of the population count in the particular location across repetitions of the simulation (explained in the next section).[19]

We can run administrative record census simulations in 2019 to generate additional measures. We will study person and location dynamics, such as the share of people who move across locations (housing unit, block, tract, ZIP code, county, state, by different levels of precision of latitude and longitude, and between the U.S. and other countries), the share that appear in one year but not the next for reasons other than being deceased, and the share that don't appear one year but appear the next for reasons other than birth. The latter two categories can be because of either emigration and immigration or administrative record coverage problems. We can model these transitions using person and housing unit characteristics, as well as past transitions as predictors. These transition probabilities could be used to supplement the measures described above for targeting survey-style data collection in future census tests.

Two additional indicators of where follow-up survey-style data collection may be most useful are housing units in the 2020 production MAF lacking anyone in administrative records[20] and addresses not in the 2020 production MAF but with people in administrative records. Locations with high concentrations of housing units of either type may be candidates for follow-up.

[18] Unlike PIKed records, we cannot obtain un-PIKed record demographic information from the Numident, but some of the other sources also contain demographic information.

[19] A higher standard deviation suggests less confidence in the administrative record count and thus greater misalignment.

[20] Bauder and Judson (2003) choose these housing units for follow-up in their simulation.

Conduct real-time simulations in 2020: How long does it take to implement an administrative records census?[21] What parts of the process are most time-consuming? Is it feasible and useful to do real-time updating of entity resolution?[22] Answers to these questions can help identify what research needs to be done in 2021-2025 to improve administrative record census execution.[23] We will process all administrative records available on a particular date (we will produce different versions with different deadline dates).[24] PIKs, MAFIDs, and other geocodes will be placed on the records. We will implement housing unit-based, person-based, and hybrid approaches and compare them.

For the housing unit-based approach, we will place people at each of their administrative record addresses that are in the 2020 Census production MAF. Person unduplication will be done within each particular housing unit, but not across housing units. Some people will be counted multiple times in different housing units, as in survey-style censuses. Individuals without administrative record addresses in the 2020 Census production MAF will not be counted.

In person-based approaches, we will create variables used in our person-place models. The person-place model coefficients will be applied to these variables to produce the probability that each PIK-location pairing is correct. After dropping PIK-location pairs for locations not included in the particular simulation (e.g., locations that can only be determined at the state level for a simulation that uses location below the state level), the remaining PIK-location probabilities will be rescaled to sum to one for each PIK. Multiple replications of the census will be constructed.[25] Each PIK is placed at one location per replication (and thus counted only once), and the location is selected randomly among the person's locations using their location probability as the weight.[26] In different variants of this approach we will change the geographic aggregation (housing unit, block, tract, ZIP code, county, state, and latitude and longitude coordinates of different degrees of precision). The locations will not be restricted to ones found in the 2020 Census production MAF. Individuals without administrative record addresses that can be geocoded to the level of geography used in the particular simulation will not be counted.

Our hybrid approach is like our person-based approaches, but where the locations are limited to housing units in the 2020 Census production MAF. PIK-location pairs where the location cannot

[21] The record linkage process is likely to take approximately a month. The subsequent data cleaning and application of model coefficients is likely to take 2-3 weeks. Tabulation of the statistics may take a day or two.
[22] Prior to 2020, we will experiment with entity resolution methods where the algorithms are refined in real-time as new records arrive. If our tests of these methods show good results (sufficient speed and accuracy), we will implement this in the 2020 simulations.
[23] Speed is desirable, because the faster it can be completed, the later the pull date can be for administrative records used in the enumeration.
[24] Individuals that the administrative records indicate are deceased will be dropped from all simulations.
[25] The exact number will depend on how quickly a replication can be completed. The larger the number, the more informative the variance calculations will be.
[26] For PIKs with multiple locations, this means they may be placed at different locations across replications. This is similar to the idea in MITRE and Santa Fe Institute (2016) that "a person has a certain chance of being in a certain location at a certain time. The quantity of persons in a given region would then be the summation of the mass of the probability distributions."

be traced to a 2020 Census production MAFID will be dropped, and the remaining PIK-location pair probabilities will be rescaled to sum to one.[27]

We can also produce versions requiring that at least one of the administrative record sources putting a person at the address is a federal government source.[28]

The housing unit-based approach most closely mimics survey collection, so it may more closely match the 2020 Census statistics. A person-based approach has the potential to improve enumeration quality where survey collection contains errors, for example, by making greater effort to count people only once.

Demographic characteristics from the CES best race data,[29] the Master Demographics Database, and administrative records will be attached to each PIK in each of the simulations. Housing tenure information from administrative records will be attached to each MAFID for all simulations that use housing unit as the location.

The population count for a location is the number of PIKs assigned to it in the replication. In simulations restricted to MAFIDs in the 2020 production MAF, each MAFID will be classified as occupied in a replication if at least one PIK is assigned to it, and otherwise it will be classified as unoccupied. No distinction will be made between vacant and delete. Determining the number of unoccupied housing units without the use of the 2020 production MAF is out of the scope of this project.

For each simulation (except the housing unit-based approach, which will have just one replication and thus will not have a distribution),[30] we will calculate moments of the distribution (e.g., mean and standard deviation) of overall population count and by sex, age group, race, ethnicity, and citizenship by geography (MAFID, block, tract, ZIP code, county, state, and for the 50 states plus the District of Columbia). We will also calculate moments for number of occupied housing units by geography.

Compare simulations to 2020 Census: How does administrative record coverage compare with the 2020 Census (total counts, as well as and omissions in administrative records and omissions in the 2020 Census), overall and by demographics and location? To what extent do the locations of individuals common to the 2020 Census and the administrative records census agree, overall and by demographics and location? How does the degree of agreement between a simulation and the 2020 Census vary by the geographic aggregation of the counts? How does the degree of agreement

[27] All approaches using the housing unit as the location will exclude group quarters, whereas the ones using other geocodes will include them.

[28] One way to justify using an administrative record census is that people have provided information to the federal government already, and the Census Bureau is part of the federal government. This argument will be stronger if the methodology requires that a federal government-sourced administrative record puts the person at the address. With this methodology the state, local, and commercial data role would be to improve prediction, helping to choose between different addresses when federal government sources disagree with each other.

[29] The CES best race data are sourced from both Title 13 survey data and administrative records.

[30] The probability that each person is assigned to the survey location, as derived from the person-place models, can be used to produce a measure of uncertainty about the population count in each housing unit. This will be our main measure of uncertainty for the housing unit-based approach, since we will not run multiple replications for it.

vary by approach (housing unit-based vs. person-based vs. hybrid) and by the geographic aggregation of the location a person is assigned to in the simulation (placing a person in a housing unit vs. block vs. tract vs. ZIP code vs. county vs. state vs. different precision levels of latitude and longitude)?

We will compare the 2020 Census population counts with the mean counts across simulations at the national level, as well as moments of the distributions of the degree of count agreement at the MAFID, block, tract, ZIP code, county, state, national levels, and at different levels of precision for latitude and longitude coordinates. We will identify the extent to which the simulations omit people included in the 2020 Census and include people omitted from the 2020 Census. We will study person-location agreement rates (a measure of enumeration errors) between the 2020 Census and different simulations among those counted in both the 2020 Census and the simulation to which it is being compared. In addition, count and person-location agreement rate comparisons will be made by sex, age group, race, ethnicity, and citizenship.

How does the degree of similarity in the statistics depend on availability of state-level administrative records, such as SNAP and TANF? We will compare the degree of similarity with the 2020 Census in states with and without these files. In the states where we have these files, we can create additional simulations that remove these files to see how much they matter for the statsitics. This will inform how valuable the state administrative record files are for administrative record enumeration.

How does the degree of agreement between simulations and the 2020 Census differ by 2020 Census response mode and whether there is reason to doubt the housing unit's 2020 Census response accuracy? We will make separate comparisons by 2020 Census response mode, which will show the effects of substituting administrative records enumeration for each particular census operation (e.g., maybe administrative record enumeration would be a better substitute for NRFU than group quarters or update/enumerate). Comparisons will be made for housing units with no 2020 Census discrepancies, as defined by Brown, Childs, and O'Hara (2015), and housing units with at least one discrepancy.[31] This will allow us to see how the population count differences vary with survey collection difficulties,[32] and it can measure the extent to which supplementing administrative record enumeration with survey collection can improve accuracy. It could illuminate where administrative record enumeration might improve accuracy relative to survey-style data collection.[33]

[31] Discrepancies include counting a person who isn't alive on Census Day, counting the same person at another location, count imputation, proxy response with occupied status, at least one person without a PIK, different housing unit status or count across responses, move in or move out dates in the National Change of Address file suggests the person wasn't living at that location on Census Day, the count is not equal to the number of listed persons, the undercount question is answered affirmatively, and the overcount question is answered affirmatively.

[32] Differences between administrative record and survey-style enumeration results could reflect errors in survey-style collection rather than administrative records when the survey-style collection suffers from discrepancies.

[33] For example, suppose a housing unit has multiple 2020 Census production responses with discrepant counts, and the administrative records for the housing units are associated with high predicted probabilities of being at that address (in other words, they appear to be of high quality). In such a case it is likely that the administrative records would provide a more accurate enumeration than the production responses.

What are the effects of enhancing the record linkage infrastructure on coverage and person-place agreement? We will distinguish PIKs by which reference file was used to validate them, then study coverage changes if PIKs from particular reference files are dropped. We will compare person-place agreement rates with the 2020 Census by PIK-reference file groups. We can also study how these rates vary by PVS score, which measures the degree of confidence in the record's validation.

When testing the entity resolution linkage methods, we can measure confidence in PIK assignments by posterior probabilities, and those probabilities can be propagated through the model to give posterior distributions (and hence measures of uncertainty) of the totals. We can assess the validity of the linkage method by how well these uncertainties relate to the actual person-place agreement rates.

How well do the different methods of predicting the degree of agreement between administrative records simulations and the 2020 Census at the person and location levels perform? We will show how quickly the 2020 Census and each simulation's results converge as more people, housing units, or higher-level locations are assigned to follow-up survey-style data collection.[34] For the person-location probability measure, we will start with no follow-up, then add people to follow-up beginning with those with no location in the particular simulation (their administrative record address could not be geocoded to the level used in the simulation), then add people based on their person-location probabilities, ranked from low to high, until all are assigned to follow-up. Similarly, for the housing unit- and higher-level measures, we will start with housing units with no one assigned to them, then add housing units based on their probabilities, ranked from more anticipated differences to fewer. This will inform the extent to which survey-style data collection can be targeted at particular individuals, housing units, or geographic areas where administrative record enumeration is most different from survey-style collection. The better the predictions, the smaller the amount of survey collection that is needed to achieve a certain quality level.

Using these housing unit rankings, we will calculate the cost of enumerating different shares of them by survey methods vs. using administrative records. This will inform trade-offs between cost savings and statistical differences with survey collection.

How does the degree of agreement between the 2020 Census and the simulations compare to agreement between the 2020 Census and the Census Bureau intercensal estimates products such as demographic analysis (DA) and ACS estimates? Making such comparisons can inform whether administrative records have the potential to improve upon other intercensal population estimates.

Informed by this analysis, a decision will be made on whether to continue researching and conducting administrative record enumeration on an annual basis after 2020. The analysis will also shed light on which approaches are most promising, which additional data sources are the highest priorities, and what record linkage and modeling improvements are needed. If the decision is to continue this line of research, then the next two steps will be taken.

[34] Following the 2000 AREX methodology (Bauder and Judson, 2003), we could replace the simulation result with the 2020 Census result for people or locations targeted for follow-up survey-style data collection. Note, however, that this exercise will be less informative for housing units where administrative record enumeration or vacancy determination is applied in 2020 NRFU, since survey-style data collection is not done for those cases.

Table 1 provides a summary of the comparisons.

Table 1. Summary of Comparisons

Measure	How Tabulated
Population Count in 2020 Census, PES, and Each Type of Simulation	National, State, Race/Ethnicity, Citizenship, Age Groups, Sex, Census Enumeration Method, Census Discrepancy Type, Reference File Source for Person Linkage, Administrative Record Source
Omissions in 2020 Census and Each Type of Simulation	National, Race/Ethnicity, Citizenship, Age Groups, Sex, Census Enumeration Method, Census Discrepancy Type
Person-Place Agreement Rate between Administrative Record Simulations and 2020 Census	National, Race/Ethnicity, Citizenship, Age Groups, Sex, Census Enumeration Method, Person-Place Probability Groups, Census Discrepancy Type
Population Count in 2020 Census Alone vs. Different Combinations of Preferred Simulation and 2020 Census	Speed of convergence, using different survey-style targeting measures
Population Count in 2020 Census, Preferred Simulation, DA, and ACS	National Overall Count (not for ACS), Sex, Age Groups, Race/Ethnicity

B. Interventions with the 2020 Census

This project will not intervene with the 2020 Census.

C. Implications for 2030 Census Design Decisions and Future Research and Testing

The results of this study will inform decisions about the extent to which future censuses should rely on administrative records to enumerate populations. This study could also lead to further results in intercensal years:

1. Conduct real-time simulations in 2021 and future years.
2. Produce annual population estimates.

Conduct real-time simulations in future years: We will follow the same steps as in the 2020 simulations, but focusing on approaches that produce the best results based on comparisons with the 2020 Census, including enhancements (additional data sources to address coverage gaps, further enhanced record linkage, and improved models using the 2020 Census in the estimation) to address weaknesses in the 2020 simulations.

Produce annual population estimates: The annual simulations could be used to produce annual population estimates at different levels of geography, if so desired. Adjustment factors based on

the comparison between the 2020 Census and the 2020 simulations could be applied to the annual administrative record counts. Person and housing unit transition rates across years will also be calculated. As the number of points in time increases, the accuracy of transition prediction models should improve.

VI. Data Requirements

Data File/Report	Source	Purpose	Expected Delivery Date
IRS Form 1040	IRS	enumeration, prediction	available
IRS 1099	IRS	enumeration, prediction	available
IRS 1099-R	IRS	enumeration, prediction	available
IRS W-2	IRS	enumeration, prediction	available
2000 Decennial PIK Crosswalk	Census Bureau	record linkage	available
Census 2000	Census Bureau	prediction	available
2000 Hundred Percent Detail File	Census Bureau	prediction	available
2000 BOC PIK Crosswalk	Census Bureau	record linkage	available
2000 Census Unedited File (CUF)	Census Bureau	prediction	available
2010 Census Unedited Files	Census Bureau	prediction	available
2010 Census Edited Files	Census Bureau	prediction	available
2010 Census PIK Crosswalk	Census Bureau	record linkage	available
2010 Census Undeliverable-As-Addressed	Census Bureau	prediction	available
2018 End-to-End Test	Census Bureau	prediction	07/01/2019
2020 Census DRF	Census Bureau	Census cost and quality assessment	09/01/2020
2020 Census CUF	Census Bureau	comparison	10/01/2020
2020 Census CEF	Census Bureau	CVAP production	12/01/2020
2020 Census PES	Census Bureau	comparison	02/01/2021
2000-2019 ACS	Census Bureau	prediction, demographic characteristics	available
ACS PIK Crosswalks	Census Bureau	record linkage	available
Current Population Survey Annual Social and Economic Supplement (CPS ASEC)	Census Bureau	prediction	available
Current Population Survey Basic Monthly Files	Census Bureau	prediction	available
Survey of Income and Program Participation (SIPP)	Census Bureau	prediction	available
SIPP Crosswalk Files	Census Bureau	record linkage	available
CPS PIK Crosswalk Files	Census Bureau	record linkage	available
Census Kidlink	Census Bureau	record linkage	available
Master Address File Extracts	Census Bureau	housing frame, prediction	available
Master Address File Auxiliary Reference File	Census Bureau	address processing	available

Geocoded Address Extract File	Census Bureau	address processing	available
Master Demographics File	Census Bureau	demographic characteristics	available
2010 Census Coverage Measurement Estimate and Results files	Auxiliary Reference File	prediction	available
Title 13 Race and Ethnicity File	Auxiliary Reference File	demographic characteristics	available
CES Best Race File	Auxiliary Reference File	demographic characteristics	available
LEHD Employment History File (LEHD-EHF)	Auxiliary Reference File	prediction	available
LEHD Employer Characteristics File (LEHD-ECF)	Auxiliary Reference File	prediction	available
LEHD Individual Characteristics File (LEHD-ICF)	Auxiliary Reference File	prediction	available
LEHD Unit to Worker Impute	Auxiliary Reference File	prediction	available
Master Beneficiary Records (MBR)	Social Security Administration	prediction	2015 available, MOU in progress for future years
Supplemental Security Records (SSR)	Social Security Administration	prediction	available
Disability Application File (831)	Social Security Administration	prediction	available
Social Security Numident File	Social Security Administration	record linkage, demographic characteristics	available
HHS Child Care and Development Fund (CCDF)	Health and Human Services (HHS)	enumeration, prediction	available
HHS Temporary Assistance for Needy Families (TANF)	Health and Human Services (HHS)	enumeration, prediction	available
HHS Indian Health Service (IHS)	Health and Human Services (HHS)	enumeration, prediction	available
CMS Medicare Enrollment Database	Health and Human Services (HHS)	enumeration, prediction	data in-house, MOU in progress
CMS Medicaid and CHIP Information System (MSIS and T-MSIS)	Health and Human Services (HHS)	enumeration, prediction	data in-house, MOU in progress
Comp Homes Underwriting Management System	Housing and Urban Development (HUD)	enumeration, prediction	available
Multi-Family Tenant Characteristics System	Housing and Urban Development (HUD)	enumeration, prediction	available
TRACS data	Housing and Urban Development (HUD)	enumeration, prediction	available
PIC data	Housing and Urban Development (HUD)	enumeration, prediction	available

Office of Personnel Management Files (OPM)	Office of Personnel Management	prediction	available
Veteran's Administration Records (VA)	Veteran's Adminis	enumeration, prediction	available
Selective Service System	Selective Service System	enumeration, prediction	available
National Change of Address Files (USPS)	United States Postal Service	enumeration, prediction	available
Army Service and Post Service Data (DOD)	Department of Defense	enumeration, prediction	available
Department of Defense Records (DOD)	Department of Defense	enumeration, prediction	available
Bureau of Prisons Permanent Release Database	Bureau of Prisons	enumeration, prediction	available
Federal Housing Authority Loan data	Federal Housing Authority	enumeration, prediction	available
U.S. Citizenship and Immigration Services visa and naturalizations data	Department of Homeland Security	record linkage, demographic characteristics	MOU in progress
Immigration and Customs Enforcement Student Exchange and Visitor Program (SEVIS)	Department of Homeland Security	record linkage, demographic characteristics	MOU in progress
U.S. Marshals Service incarceration data, with DHS citizenship status	Bureau of Prisons and Department of Homeland Security	record linkage, demographic characteristics	MOU in progress
U.S. Customs and Border Protection arrival/departure data	Department of Homeland Security	record linkage, demographic characteristics	MOU in progress
U.S. State Department Passport Services passport data	Department of State	record linkage, demographic characteristics	MOU in progress
U.S. State Department Worldwide Refugee and Asylum Processing System (WRAPS)	Department of State	record linkage, demographic characteristics	MOU in progress
Supplemental Nutrition Assistance Program (SNAP)	State agencies	enumeration, prediction	some states available
Supplemental Nutrition Program for Women, Infants, and Children (WIC)	State agencies	enumeration, prediction	some states available
Temporary Assistance to Needy Families (TANF)	State agencies	enumeration, prediction	some states available
Low Income Home Energy Assistance Program (LIHEAP)	State agencies	enumeration, prediction	available
Alaska Permanent Fund Dividend File	State agency	enumeration, prediction	available
Homeless Management Information Systems (HMIS)	County agencies	enumeration, prediction	available
Utilities Records Data Veteran Service Group of Illinois (VSGI)	VSGI, Inc	enumeration, prediction	available
Corelogic	Corelogic	prediction	available
DAR Partners	DAR Partners	prediction	available
Experian	Experian	prediction	available
InfoGroup	InfoGroup	prediction	available
Melissa Data	Melissa Data	prediction	available

RealtyTrac	RealtyTrac	prediction	available
Targus/Neustar	Targus	prediction	available

VII. Risks

1. If the Census Bureau does not provide full funding for staff or provide staff with the needed skills, then the project scope will have to be narrowed.

2. If the Census Bureau does not maintain and in some cases revise agreements with current data providers, then some subpopulations will be poorly covered in the simulations.

3. If the Census Bureau's DMS approval process for provisioning data to researchers is not streamlined, the project may not be able to produce results in a timely manner. This is a particular concern for this project, since it involves so many datasets and is under time pressure due to the real-time aspect.

4. If the Census Bureau does not acquire additional data sources such as State Department passport and visa data, then some subpopulations will be poorly covered by the simulations.

5. If the public and/or stakeholder groups are concerned by the alternative population estimates produced by the simulations, then legal challenges may occur.

VIII. Limitations

1. The applicability of the simulations to conducting an administrative record census in the future depends on the availability of the same data sources in the future, which may not be the case. Some additional data sources may become available in the future, while others may no longer be available.

2. The 2020 Census may differ in coverage relative to past censuses due to sensitivity to and the possible addition of a question on citizenship status. The citizenship status question could potentially be discontinued after 2020.[35] Thus, the comparisons between the 2020 Census and the administrative record simulations could thus be different than they would be in the future (minus the citizenship question) for this reason.

3. Any errors in ACS household rosters will negatively affect the accuracy of the person-place models. For example, persons in the roster failing PVS will not be included in the models. The relationship between their survey and administrative records addresses could vary systematically with whether the person is successfully PVSed in the ACS, leading to less accurate predictions for such people.

[35] This was mentioned in the March 1, 2018 memo from the Census Bureau to Commerce Secretary Wilbur Ross, which has been publicly released in a Freedom of Information Act (FOIA) request.

4. When there are discrepancies between the 2020 Census, PES, and administrative records, it is impossible to know for certain which is correct in the absence of an error-free source.

IX. Issues That Need to be Resolved

1. The MOUs for some data sources have not yet been completed.

X. Division Responsibilities

Division or Office	Responsibilities
ERD	• Data sharing agreements • Data acquisition and processing
CED, CES, CODS, CSRM, ERD	• Enhance record linkage infrastructure • Evaluate PVS and entity resolution
CES	• Data aggregation • Supplementary coverage and characteristics analyses
CES, CED	• Model development and estimation • Population estimates

XI. Milestone Schedule

Evaluation Milestone	Date
Obtain additional administrative record sources	10/18 – 09/19
Enhance record linkage infrastructure	10/18 – 09/21
Evaluate PVS and entity resolution linkage processes	10/18 – 09/21
Develop person-place models using ACS data	03/19 – 06/19
Process administrative records available on July 1, 2019 for use in 2019 administrative record census	07/19-09/19
Construct administrative record census simulations for 2019	10/19
Compare 2019 simulations to March-April 2019 ACS for housing units and persons in common	11/19
Make predictions for where survey-style data collection most useful in 2020	12/19 – 02/20
Process administrative records available on July 1, 2020 for use in 2020 administrative record census	07/20-08/20
Construct administrative record census simulations for 2020	09/20
Produce 2020 administrative record census statistics	10/20
Compare 2020 administrative record census simulations to 2020 Census	11/20-06/21
Test and revise person-place models using 2020 Census	06/21 – 08/21

Evaluation Milestone	Date
Process administrative records available on this date for use in 2021 administrative record census	07/21-08/21
Write report on 2020 administrative record census simulation and 2020 Census comparisons	07/21-09/21
Process administrative records available on this date for use in 2021 administrative record census	07/21-08/21
Construct administrative record census simulations for 2021	09/21
Produce 2021 administrative record census statistics	09/21
Distribute Initial Draft Real-Time 2020 Administrative Record Census Simulation Report to the Decennial Research Objectives and Methods (DROM) Working Group for Pre-Briefing Review	09/30/2021
Decennial Census Communications Office (DCCO) Staff Formally Release the FINAL Real-Time 2020 Administrative Record Census Simulation Report in the 2020 Memorandum Series	03/01/2022

XII. Review/Approval Table

Role	Approval Date
Primary Author's Division Chief (or designee) Lucia Foster	08/13/2018
Decennial Census Management Division (DCMD) ADC for Nonresponse, Evaluations, and Experiments	02/19/2019
Decennial Research Objectives and Methods (DROM) Working Group	02/19/2019
Decennial Census Communications Office (DCCO)	mm/dd/yyyy

XIII. Document Revision and Version Control History

Version/Editor	Date	Revision Description
1.0	08/29/2018	Initial draft
2.0	02/05/2019	Incorporated comments from September 2018 DROM
3.0	02/20/2019	Incorporated comments from February 2019 Quality Process Review
4.0	03/06/2019	Incorporated comments from February 2019 DROM

XIV. Glossary of Acronyms

Acronym	Definition
ACS	American Community Survey
ADC	Assistant Division Chief
AREX	Administrative Records Census Experiment
CBP	U.S. Customs and Border Protection
CEF	Census Edited File
CUF	Census Unedited File
DA	Demographic Analysis
DCCO	Decennial Census Communications Office
DRF	Decennial Response File
DROM	Decennial Research Objectives and Methods Working Group
DSSD	Decennial Statistical Studies Division
EXC	Evaluations & Experiments Coordination Branch
FTI	Federal Tax Information
IPT	Integrated Project Team
IRE	Integrated Research Environment
IRS	Internal Revenue Service
ITIN	Individual Taxpayer Identification Number
MAF	Master Address File
MAFID	Master Address File Identification Number
NRFU	Nonresponse Followup
PES	Post Enumeration Survey
PIK	Protected Identification Key
PPS	Predominant Purpose Statement
PVS	Person Identification Validation System
R&M	Research & Methodology Directorate
SSA	Social Security Administration
StARS	Statistical Administrative Records System
USCIS	U.S. Citizenship and Immigration Services
VSGI	Veterans Service Group of Illinois

XV. References

Bauder, Mark, and Dean H. Judson, 2003, "Administrative Records Experiment in 2000 (AREX 2000) Household Level Analysis," 2000 Census Experiment Report, U.S. Census Bureau.

Brown, J. David, Jennifer H. Childs, and Amy O'Hara, 2015, "Using the Census to Evaluate Administrative Records and Vice Versa," Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference.

Farber, James, and Charlene Leggieri, 2002, "Building and Validating a National Administrative Records Database for the United States," New Zealand Conference on Database Integration.

JASON, 2016, "JSR-16-Task-009, Alternative Futures for the Conduct of the 2030 Census."

MITRE Corporation and Santa Fe Institute, 2016, "U.S. Census 2030 Challenge FY 2016 DRAFT Report".

Ralphs, Martin, and Paul Tutton, 2011, "Beyond 2011: International Models for Census Taking: Current Processes and Future Developments," Beyond 2011 Project, Office for National Statistics, Version 1.0.

Rastogi, Sonya, and Amy O'Hara, 2012, "2010 Census Match Study," 2010 Census Planning Memoranda Series No. 247.

Steorts, Rebecca C., Rob Hall, and Stephen E. Fienberg, 2016, "A Bayesian Approach to Graphical Record Linkage and Deduplication," *Journal of the American Statistical Association*, Vol. 111, pp. 1660-1672.