Alternative Futures for the Conduct of the 2030 Census

Contact: Dan McMorrow — dmcmorrow@mitre.org

November 2016

JSR-16-Task-009

Approved for publication—distribution only by sponsor; U.S. Census Bureau, 4600 Silver Hill Road Washington, DC 20233; Dr. John Thompson.

JASON
The MITRE Corporation
7515 Colshire Drive
McLean, Virginia 22102-7508
(703) 983-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* November 2016	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE

Alternative Futures for the Conduct of the 2030 Census

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER
1316JA01

5e. TASK NUMBER
PS

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
JASON Program Office
7515 Colshire Drive
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

JSR-Task-16-009

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for publication—distribution only by sponsor.

13. SUPPLEMENTARY NOTES

The Census Bureau asked JASON to consider alternative futures for 2030 and to propose a starting point from which the Census Bureau can begin to develop a 2030 strategy. JASON's main recommendations takes advantage of the increased avail-ability of high-quality government administrative data; e.g. data collected by the Internal Revenue Service (IRS) and Social Security Administration (SSA). Research at the Census Bureau and by academics suggests that 90%or more of the U.S. population could be located in a combination of IRS and SSA records; these data contain most of the variables collected in the census short form. JASON recommends that the Census Bureau consider starting the 2030 Census with an "in-office" enumeration of the population using existing government administrative records. That would be followed by a second step using additional data and more traditional methods to find people not present in government records and to "fill in" variables that might be missing in these records.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Mr. John Thompson
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	UL		**19b. TELEPHONE NUMBER** *(include area code)* 301-763-2135

Contents

1 EXECUTIVE SUMMARY

The U.S. decennial census, codified in the U.S. Constitution, has taken place every decade since 1790, with 2020 marking the 24th census. Since 1880, the Census Bureau has relied on maps, and later actual addresses, to find locations and then enumerate the population at those locations. Since 1970 master address lists have been created prior to the enumeration and confirmed by canvassing the country. At each "address," including homeless shelters and institutions, individuals who "live and sleep here most of the time" are counted and geo-located at that address.

The Census Bureau faces pressure to efficiently and effectively conduct an accurate and precise decennial census. The public response rate in recent censuses has declined from nearly 80% to the mid-60% range requiring huge increases in door-to-door efforts and corresponding costs for the non-response follow-up (NRFU). In 2010, enumerators made over 100 million visits to housing units. The total Census cost per housing unit was approximately $94, a 34% increase from 2000 (in constant 2010 dollars).

The Census Bureau typically starts planning the new decennial census at least 8-10 years ahead of time. This, too, presents challenges as the social, economic, demographic, political, environmental, and technological factors will change in the intervening period. The Census Bureau asked JASON to consider alternative futures for 2030 and to propose a starting point from which the Census Bureau can begin to develop a 2030 strategy.

JASON's main recommendations takes advantage of the increased availability of high-quality government administrative data; e.g. data collected by the Internal Revenue Service (IRS) and Social Security Administration (SSA). Research at the Census Bureau and by academics suggests that 90% or more of the U.S. population could be located in a combination of IRS and SSA records; these data contain most of the variables collected in the census short form. JASON recommends that the Census Bureau consider starting the 2030 Census with an "in-office" enumeration of the population using existing government administrative records. That would be followed by a second step using additional data and more traditional methods to find people not present in government records and to "fill in" variables that might be missing in these records.

A transition to an "in-office" enumeration using administrative records also suggests a paradigm shift in the way the Census Bureau conceptualizes the enumeration. Historically, the Census Bureau has used the Master Address File (MAF), a list of housing units, as a frame, enumerating the people in each unit. Government administrative records are organized at the level of individuals, indexed by Social Security Numbers or Tax Identification Numbers. 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. This shift of focus aligns with the declining size of households and the current mobility estimates of less than 1/3 of the population maintaining the same residence during the decade, both of which reduce the efficiency of organizing the census around a list of housing units.

JASON's recommendations build on innovations in the Census 2020 Operational Plan, specifically the Census Bureau's proposal to make new use of administrative records. The 2020 plan calls for using records from the 2010 Census and U.S. Postal Service (USPS) to construct approximately 75% of the MAF "in-office," with the remainder of the work done "in-field." Later in the enumeration process, administrative records from the IRS and SSA will be used to optimize NRFU visits and in certain cases to impute a housing unit, response if an initial visit fails.

JASON recommends that the Census Bureau use the data collection efforts already underway for the 2020 Census to undertake research on: (i) how much of the population is covered in IRS and SSA records, (ii) how many of the census short-form variables can be collected using those records, (iii) which populations and variables are missing from the records, and (iv) alternative data sources and "in-field" methods to be used to complete an accurate enumeration. Considerable research, testing and experimentation will be needed to move to a census that starts with administrative records. JASON recommends that this process begin as soon as possible in order that the Census Bureau be in a position to make a decision in five years about whether to move forward with the approach for 2030.

Although outside of the direct tasking for this report, JASON considered several other topics. The first is the American Community Survey (ACS). The ACS replaced the census long form in 2000 and it suffers from declining response rates and costly NRFU. Many of the ACS questions have known answers in administrative records, e.g., type of housing unit, property value, age of structure, and household income. The "in-office" approach for the decadal census could establish much of this underlying data and so provide an opportunity to re-think the ACS, perhaps shifting its focus toward information not available in other data sources such as beliefs and measures of subjective well-being that require survey elicitation.

A second additional topic relates to the MAF. The Census Bureau invests considerable resources to construct the MAF but cannot share it with outside parties despite its potential value as a public good. If the Census Bureau adopts JASON's suggestion of relying mainly on administrative records for enumeration, it may not be necessary to construct a MAF at all in 2030. However, JASON believes the Census Bureau could create value by working with the USPS, Department of Transportation (DOT), local governments, and private sector firms to construct a continuously maintained National Address File that would be public information and not subject to Title 13 restrictions on data sharing. JASON also notes that a continuously maintained address file potentially could be integrated with a continuously maintained population register, forming the basis for a rolling census that could be verified every decade to satisfy constitutional requirements.

The report concludes by considering some issues relating to historical census records. The ability to link historical census records is important for understanding trends and intergenerational issues within the United States. Currently, the Census Bureau has not been able to efficiently digitize the hand-written names in the 1950-1990 census records, in part because they are protected under Title 13. The report discusses the potential for using Optical Character Recognition to parse names into individual "words" that would be exempt of Title 13, allowing the use of cheap and fast transcription.

JASON offers sixteen specific recommended actions to be taken by the Census Bureau. These recommendations are described in greater detail in the main text.

Regarding moving to an "in-office" enumeration:

1. Re-conceptualize the census by organizing it around people rather than housing units.
   (a) Identify people, then place them.
   (b) Start the count from "ninety percent" by using "in-office" enumeration.
   (c) Use field activities to fill in the gaps and validate what is known.
2. Start enumeration with administrative records from IRS, SSA and past census data to construct an "in-office" census that is as complete as possible.

3. Develop a multifaceted strategy to find people who do not appear in the "in-office" enumeration.

4. Use research and near-term experimentation to explore who will not be enumerated with this approach, what data fields will be lacking, and strategies for gap-filling.

5. Continue and expand efforts to acquire data from other agencies, which will be critical to the success of "in-office" enumeration.

Regarding trade-offs:

6. Create a set of metrics and criteria by which an "in-office" approach can be evaluated against traditional "self-response plus NRFU" approaches.

7. Examine the utility and cost of expanding the use of administrative records to be a rolling census that would provide an up-to-date population to satisfy enumeration requirements between decadal censuses.

8. Develop a list of options detailing the estimated cost of the 2030 Census as a function of the "accuracy and coverage" desired, which could be used by the Census Bureau and Congress to decide "how good is good enough."

Regarding research and testing:

9. Develop and start a set of experiments now to test the "in-office" enumeration concept.

   (a) Utilize massive administrative data linkage tests to confirm percentage of population enumerable by this strategy.

   (b) Confirm ability to identify subpopulations that will be consistently missed.

10. Explore and test alternative approaches to reach the remaining hard-to-count populations (gaps).

    (a) These could include remote and street-level sensing, crowd-sourced Citizen Enumeration, novel data such as state Medicaid records for low-income population, and partnership with HUD to count the homeless.

11. Continue to plan for and test enumeration strategies in the face of natural disasters, terrorist attacks, and temporary dislocation of large numbers of people.

    (a) Create partnerships with FEMA to locate (place) temporarily displaced individuals and trusted civil servants (e.g., firefighters).

Regarding the American Community Survey (ACS):

12. Reconsider ACS and how much data the Census Bureau elicits from different people at different times, keeping in mind replication of administrative data and high cost of asking the first question.

    (a) Use ACS to follow trends relevant to future Census Bureau data collection, such as how long people maintain their email addresses and cell phones versus their residences.

    (b) Use ACS to monitor public trust (a subjective measure).

    (c) Commission a study on what are the high-value data per dollar.

Regarding issues beyond Census 2030:

13. Create a public National Address File outside of Title 13.

14. Be alert for new public and private sector data sources that may become available.

15. Develop a pilot “rolling census” project relying mainly on administrative data.

16. Explore new methods for name recognition and OCR to digitize 1950- 1990 censuses.

2 INTRODUCTION

The decennial census, codified in the U.S Constitution, has taken place every decade since 1790. This ”actual enumeration” of all persons in each state determines the states’ representation in the House of Representatives. It is important to note that the term “persons” for the decennial enumeration is different from the term “citizens” used elsewhere in the Constitution and includes all people residing in the United States. The decennial census is also required by statute and judicial order to support the states in developing the boundaries of state legislative districts. This means that people associated with a state must be geolocated within that state. Further, the Voting Rights Act of 1965 (PL 89-110) imposes requirements for the census to collect demographic data beyond simple counts.

The year 2030 will be the 240th anniversary of the first U.S. Census and the 25th decennial census. Now is an opportune time to consider social, economic, and technical changes that may effect the creation of a 2030 enumeration of the U.S. population.

2.1 JASON Tasking

The Census Bureau faces pressure to efficiently and effectively conduct an accurate and precise decennial census. The Census Bureau typically starts planning the new decennial census at least 8-10 years ahead of time. This presents challenges as the social, economic, demographic, political, environmental, and technological factors will change in the intervening period. Two specific concerns are public perception of the census and trust in government; both issues could affect willful response to the enumeration process.

The public response rate in recent censuses has declined from nearly 80% to the mid-60% range, requiring substantial increases in door-to-door efforts and corresponding costs for the non-response follow-up (NRFU). In 2010, enumerators made over 100 million visits to housing units. The total cost per housing unit to the Census Bureau was approximately $94, a 34% increase from 2010 (in constant dollars). The American Community Survey, which replaced the census long form in 2000, also suffers from declining response rates and costly NRFU.

To support the Census Bureau's planning of the 2030 Census JASON has been asked:

> What are the possible alternative futures that could affect
> the conduct of the 2030 decennial census?

JASON was given a broad scope and asked to consider:

- leading indicators to watch - those that would tip or shape the futures,
- factors that Census can monitor, manage and plan for now, and
- key factors/approaches/concepts for testing in 2020.

This study is to propose a starting point from which the Census Bureau can begin to develop a 2030 strategy.

2.2 JASON Study Process and Study Focus

JASON was introduced to the topic through presentations by the briefers listed in Figure 2.1 . These individuals represented experts internal and external to the Census Bureau on issues ranging from the history of the census to international census enumeration models. The briefers attended the full set of presentations and participated in the accompanying discussions. Materials recommended by these individuals, together with a wide range of other publicly available materials, were reviewed and discussed by JASON. JASON gratefully acknowledges the efforts of Maryanne Chapin, Assistant Division Chief for Decennial Census Management Division (Census Bureau) for coordinating the briefings and responding to the numerous questions posed by JASON throughout the study period.

John Thompson	Director	Census Bureau
Margo Anderson	Distinguished Professor	U. of Wisconsin
John Abowd	Associate Director for Research & Methodology	Census Bureau
Deirdre Bishop	Chief of Geography Division	Census Bureau
Evan Moffet	Assistant Division Chief for Geographic Operations	Census Bureau
Amy O’Hara	Chief of Center for Administrative Records Research & Applications	Census Bureau
Nancy Bates	Senior Researcher for Survey Methodology	Census Bureau
James Whitehouse	Chief of the Census Redistric8ng & Vo8ng Rights Data Office	Census Bureau
David Johnson	Research Professor	U. of Michigan
Arona Pistiner	International Cooperative Programs Officer	Census Bureau
Tom Louis	Professor	Johns Hopkins

Figure 2.1: Briefers for the JASON 2016 study

This study focuses on the actual enumeration process versus the dissemination of the census results. The current guiding legislation for the conduct of the census is considered carefully but is not used to constrain possible new approaches to the enumeration.

The remainder of the report is organized as follows. Section 3 provides an overview of the Census Bureau’s practice, including salient legislation that governs current operations. Section 4 presents the JASON’s recommended strategy for 2030. Section 5 reviews the usefulness of administrative records. Section 6 discusses the process that could be used to implement the 2030 strategy, along with proposed research to support the process. Sections 7 and 8 look at a range of additional topics the Census Bureau might find valuable such as the implications of 2030 strategy on the American Community Survey, the creation of a National Address File, moving toward a rolling census, and a method for transcribing the 1950-1990 censuses. Section 9 provides a summary and the JASON findings and recommendation. Appendices with a brief discussion of extreme social, technical, demographic, and environmental change scenarios are also provided.

3 CURRENT PRACTICE

The goal of the census enumeration is to "count everyone once and only once and in the right place." To meet the geo-centric requirements of the census enumeration, the Census Bureau has relied on maps, and later actual addresses, to find locations and then enumerate the population at those locations. Since 1970 master address lists, i.e., the Master Address File (MAF), have been created prior to the enumeration and confirmed by canvassing the country. This canvassing has traditionally been done with field workers "walking" the country to visually confirm the addresses. The MAF is a list of all known "housing units" in the U.S., including locations for home- less shelters, institutions, and other places where individuals could "live and sleep most of the time."

No names are attached to the MAF. Rather, the MAF is used as the basis for canvassing the country a second time to count the people associated with each address on the MAF (housing unit). This second canvassing uses a variety of modalities to reach each housing unit and collect the data. This includes self-response using mail-out/mail-back forms, volunteer responses using publicly available "Be-Counted" forms, and personal contact by an enumerator. The 2020 Census is also taking advantage of the population's growing access to and use of the Internet. Internet responses will be encouraged and it is anticipated this response mode will increase the self-response rates over 2010.

The previous JASON study (JASON, 2015) advised the Census Bureau on technology approaches to respondent validation for internet responses that would not significantly change the rules for being counted, but could avoid unfortunate surprises associated with the transition to the Internet. In light of the recent cyberattacks on the Australian census (Ramzy, 2016), JASON encourages the Census Bureau to revisit the specific recommendations on fraud detection and the importance of red-teaming operations leading up to 2020.

3.1 Salient Legislation

There are several pieces of legislation, directives, and Supreme Court decisions that govern the census enumeration process. This section highlights some of the more salient pieces with respect to the innovations the Census Bureau plans for 2020 and recommendations to be made in this report for 2030.

First, the U.S. Constitution (Article 1, Section 2) requiring a decennial enumeration says

> “The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they [Congress] shall by Law direct.”

The reason to point this out is that the Constitution does not prescribe the process to be used to create the enumeration.

A second important piece of legislation is Title 13 from the U.S. Code. Title 13 directs that information collected by Census Bureau be kept confidential. This means private information will never be published, the Census Bureau will only collect information to produce statistics, and Census Bureau employees will be sworn to protect confidentiality. In addition to the data collected directly on the census forms or American Community Survey forms, the MAF is also protected under Title 13. A 1982 Supreme Court decision (U.S. Supreme Court, 1982) ruled that lists directly associated with the enumeration, even an address list without names, is considered “information as confidential.”

The last piece of legislation pertinent to this study is Title 26 from the U.S. Code, which states

> “... (1) permits the IRS [Internal Revenue Service] to share FTI [Federal Tax Information] with the Census Bureau for statistical purposes in the structuring of censuses and national economic accounts, as well as for conducting related statistical activities authorized by law.”

In addition,

> “Publication of all statistical products by the Census Bureau, including those based in whole or in part on administrative records covered by Title 26, are subject to disclosure avoidance procedures
> ….”

This means any IRS data used by the Census Bureau must be protected under Title 13.

Title 26 gives the Census Bureau authority to use IRS administrative records to support the census enumeration. By definition (OMB, 2006) an administrative record is information collected to administer a program, business, or institution. This includes information such as processing benefit applications or tracking services received.

The Census Bureau receives a monthly record update from IRS. In addition, the Census Bureau has decades of memoranda of agreements with the Social Security

Administration (SSA) allowing the use of SSA administrative data for mutual benefit. The Census Bureau currently receives quarterly data updates from the SSA. In the late 1990's the Census Bureau began receiving data from the Center for Medicare and Medicaid Services (CMS) through interagency agreements. All of these administrative records will be used to support the 2020 Census and play a principal role in the JASON recommendations for 2030.

3.2 Plans for the 2020 Census

The plans the 2020 Census are documented in the 2020 Census Operational Plan (U.S. Census Bureau, 2015). These are summarized below and depicted in Figure 3.1.

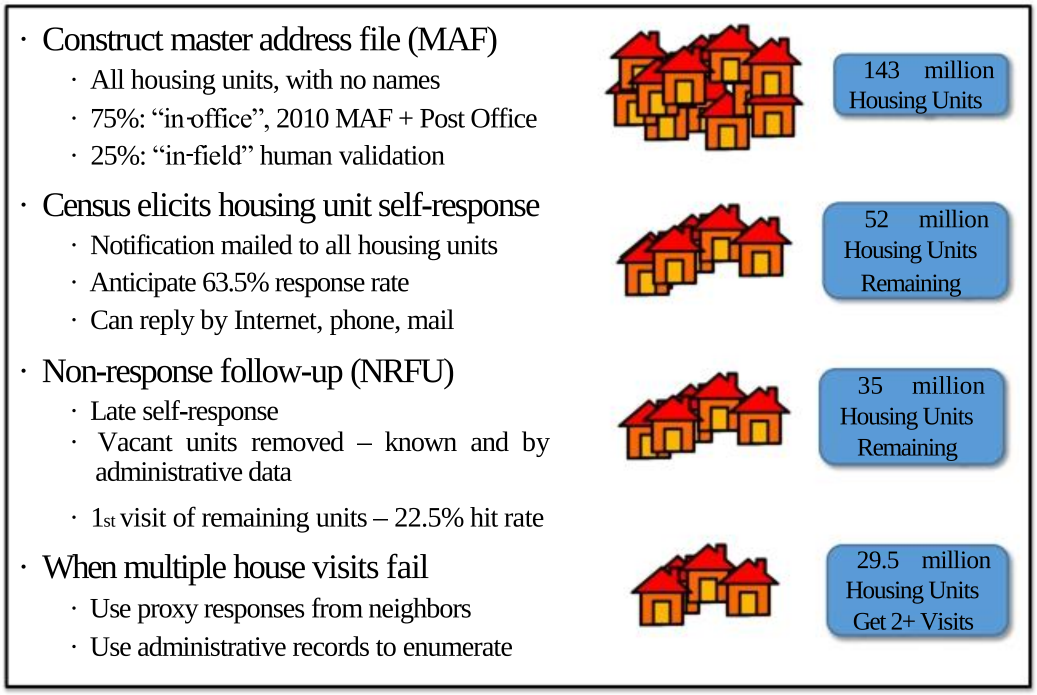
The Census Bureau estimates that MAF canvassing will result in approximately 143 million housing units. One innovation for 2020 will be to conduct an "in-office" address canvassing using the 2010 MAF, Post Office files, and other data such as Google Maps. It is expected that 75% of the MAF will be confirmed "in-office," leaving only 25% to be confirmed through field operations.

During the first six weeks of the census, the Census Bureau will elicit housing-unit self-response through mail-out/mail-back and phone interview methods, as with past censuses. A second innovation for the 2020 Census will be the option of using the Internet for self-response. (This was the topic of JASON (2015)). It is anticipated that the Census Bureau will obtain a 63.5% self-response rate, yielding 52 million housing units left for non- response follow-up (NRFU) operations.

For the NRFU operations, the Census Bureau plans to use a cost-effective strategy for contacting and counting people to ensure fair representation of every person in the United States. Once the housing units that did not respond via Internet, telephone, or mail, are known, the Census Bureau will use administrative records to identify vacant units so enumerators do not have to visit these addresses. These housing units will be removed from the MAF universe (an estimated 5.6 million housing units). Additional (late) self-responses will be collected after the first six weeks of operations (approximately 1 million units). For the remaining cases, enumerators will make a first visit to collect the information in-person. This leaves 46 million housing units that will need to be visited. During the first visit, the Census Bureau expects to resolve 22.5%, or approximately 10.3 million the housing units. This leaves 35 million housing units unresolved.

The third major innovation for the 2020 Census is the use of administrative records at this point. After providing everyone an opportunity to respond by Internet, telephone, mail or in person, and only if high quality administrative records from trusted sources exist, the Census Bureau plans to use the administrative records data as the response data for the housing unit. It is estimated that approximately 6 million of these housing units will be able to be enumerated through the use of administrative records.

Where high-quality administrative records are not available from trusted sources, the Census Bureau will continue in-person visits to reach non-responding housing units until the case is resolved. The resolution of the case could include a successful enumeration, a proxy response from neighbors, determination that the address is vacant, or that they have exhausted a maximum number of attempts to reach someone at the address. It is estimated the Census Bureau will have to visit 29.5 million housing units two or more times.



- Construct master address file (MAF)
  - All housing units, with no names
  - 75%: “in-office”, 2010 MAF + Post Office
  - 25%: “in-field” human validation
- Census elicits housing unit self-response
  - Notification mailed to all housing units
  - Anticipate 63.5% response rate
  - Can reply by Internet, phone, mail
- Non-response follow-up (NRFU)
  - Late self-response
  - Vacant units removed – known and by administrative data
  - 1st visit of remaining units – 22.5% hit rate
- When multiple house visits fail
  - Use proxy responses from neighbors
  - Use administrative records to enumerate

143 million Housing Units
52 million Housing Units Remaining
35 million Housing Units Remaining
29.5 million Housing Units Get 2+ Visits

Figure 3.1: Planned conduct of Census 2020. Figured derived from information in U.S. Census Bureau (2015).

3.3 Cost and Accuracy

The costs of the U.S. Census have been growing steadily for decades. In part, this is due to inflation and increases in population. Even after accounting for these factors, the cost per housing unit has been rising in real terms, as shown in Figure 3.2 (U.S. Census Bureau, 2012a). At the same time as per housing unit costs have increased, the accuracy of the census has been increasing in the sense that the net coverage – over or undercount – as estimated in the post-enumeration surveys has been decreasing (GAO, 2011, 2016). This is shown in Figure 3.2. However, this is only the net error. Underprivileged groups have tended to be undercounted, and whites overcounted, so that some cancellation of errors have occurred, (see Figure 3.3). Nevertheless, the general trend across all groups has been toward an increasingly accurate count.

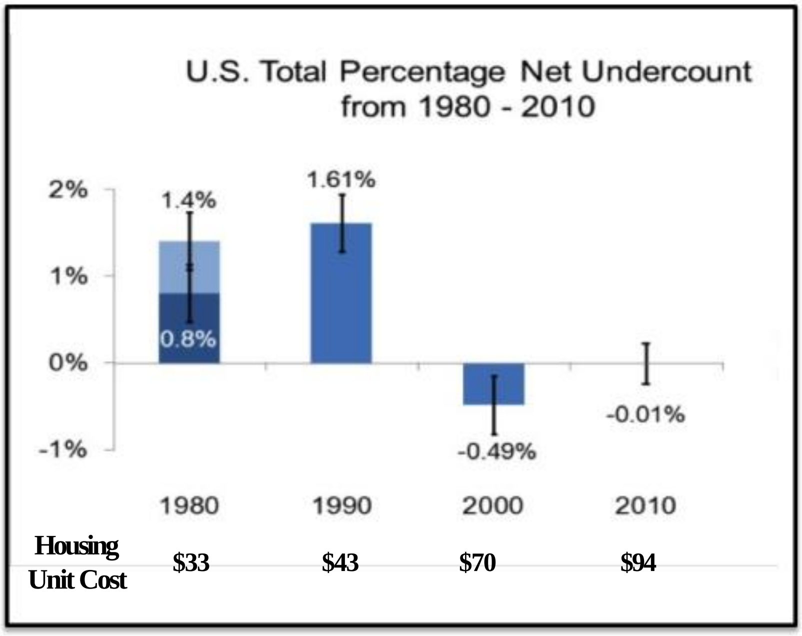
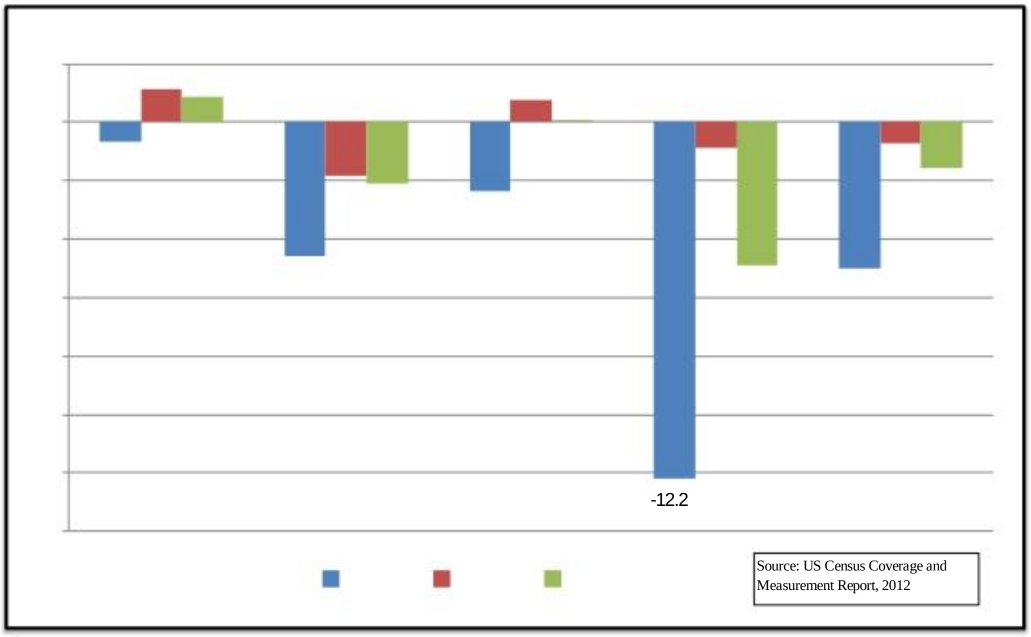


Figure 3.2: Census accuracy and cost per housing unit in constant 2010 dollars. Two net undercount estimates are available for 1980, .8% and 1.4%. Other years had single estimates. The bars are the standard errors associated with the net undercount estimates. Data from: U.S. Census Bureau (2012a), GAO (2011), and GAO (2016).

A clear understanding of the cost/benefit trade-offs between cost and coverage (accuracy) seems to be elusive across the numerous government re- ports about the census. As background for a discussion of counting strategies and priorities, it is interesting to examine how costs have been distributed in the most recent census. Table 3.1 gives the cost breakdown for field operations during the 2010 census according to the Department of Commerce's of the Inspector General (OIG, 2011).

Based on OIG (2011) description of these cost categories, field operations directed at hard-to-count populations fall under items 5 through 8 in this table. For example, items 5 and 6 involve updating non-city-style ad- dresses in rural locations, including Indian reservations, and either dropping off census forms or directly enumerating the inhabitants. Item 7 pertains to counts at campgrounds, marinas, and other places of temporary or mobile dwellings. Item 8 is the only one



Source: US Census Coverage and Measurement Report, 2012

Figure 3.3: Statistically estimated miscount by ethnicity. Source: Nancy Bates, briefing to JASON, June 2016.

that is described as an effort to count the homeless. The number is surprisingly small. It represents visits to home- less shelters, soup kitchens, and “nonsheltered outdoor locations” during the three days preceding Census Day (1 April 2010).

The data in Table 3.1 imply that the field operation costs have not been dominated by the hard-to-count populations, but rather by the need for enumerators to visit known housing units that have not self-responded. In 2010, enumerators were allowed and encouraged to visit such residences up to six times before resorting to imputation based on reports from neighbors or other evidence. The sum of lines 5-8, by contrast, is $194 million. Even item 3, which apparently involves walking every street in the country, is small by comparison to the NRFU. According to OIG (2011) the number of temporary workers involved in items 1 and 3 was in rough proportion to the costs, that is, 475,000 enumerators versus 160,000 canvassers. However, it is not possible to disentangle these numbers from the fixed

Table 3.1: Approximate costs of field operations in the 2010 Census, in mil- lions of 2010 dollars (OIG, 2011).

1.	Nonresponse Followup (NRFU)...........	$1738
2.	Field Data Collection Automation (FDCA)	$790
3.	Address Canvassing.......................	$444
4.	Vacancy Delete Check.....................	$282
5.	Update/Leave.............................	$108
6.	Update/Enumerate......................	$63
7.	Transitory Locations......................	$12.7
8.	Service-Based Enumeration	$10.8
	Total	$3,448

costs associated with the initial set-up of field offices from the operations and the actions taken in the field.

It is not clear that the total cost of the census is dominated by NRFU operation, even when taking into account all categories of Table 3.1. The total cost of the 2010 Census was nearly $13 billion (constant 2010 dollars) when summed over the decade leading up to 2010, as well as some costs of tabulation and verification in the subsequent year or two (GAO, 2016). There does not seem to be a complete report on the cost breakdown for the 2010 Census. Such reports do exist for the 2000 Census (U.S. Census Bureau (2009) and Walker et al. (2010)). These reports indicate the NRFU efforts cost $1.4 billion dollars out of a total census cost of $6 billion (in 2010 constant dollars). In addition, about 70% of the total cost of the 2000 Census was spent during the actual census year. In contrast, the Census Bureau predicts only 50% of their total 2020 costs will occur in 2020 (see Figure 3.4), likely due to some of the 2020 innovations.

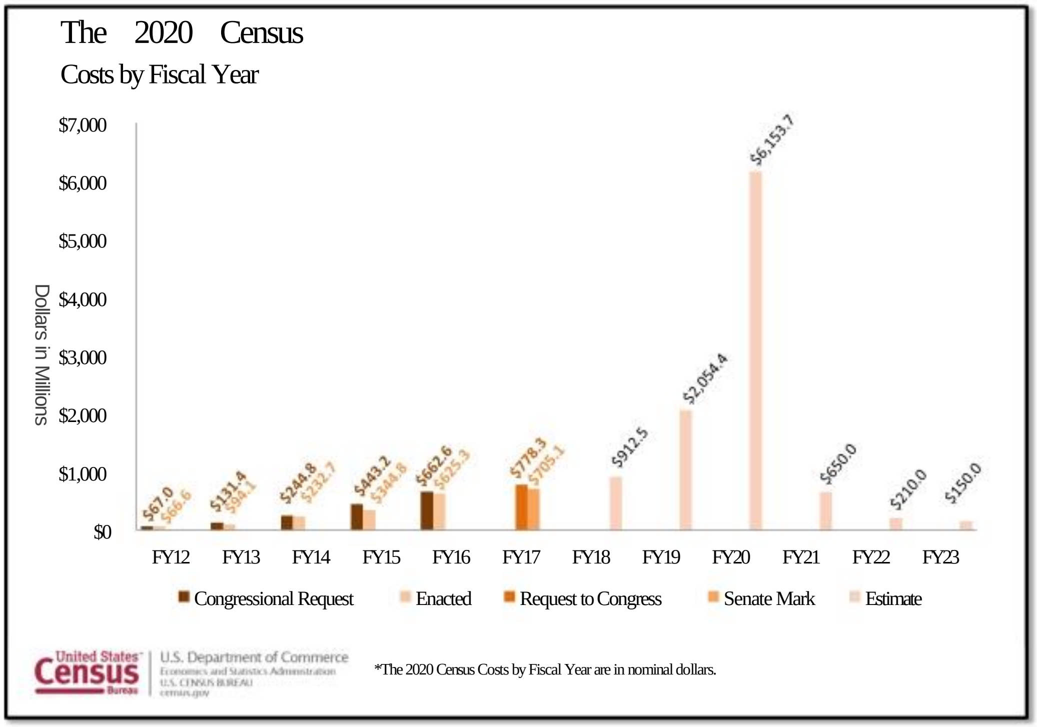


Figure 3.4: Census 2020 actual costs and future estimates. Source: Deirdre Bishop, briefing to JASON, June 2016.

What Figure 3.4 doesn't reveal is how the trade-offs between cost and coverage might be related. Figure 3.5 provides an illustrative representation of the cost and coverage based on the sequencing of the activities planned for 2020, as described in Section 3.2. The "in-office" component of the MAF Phase and the self-response and administrative data components in the Enumeration Phase are areas where the Census Bureau should be able to predict and control their fixed costs. Maximizing the "in-office" or administrative data usage to reduce costs associated with field operations and knowing this should help reduce the overall variable costs.

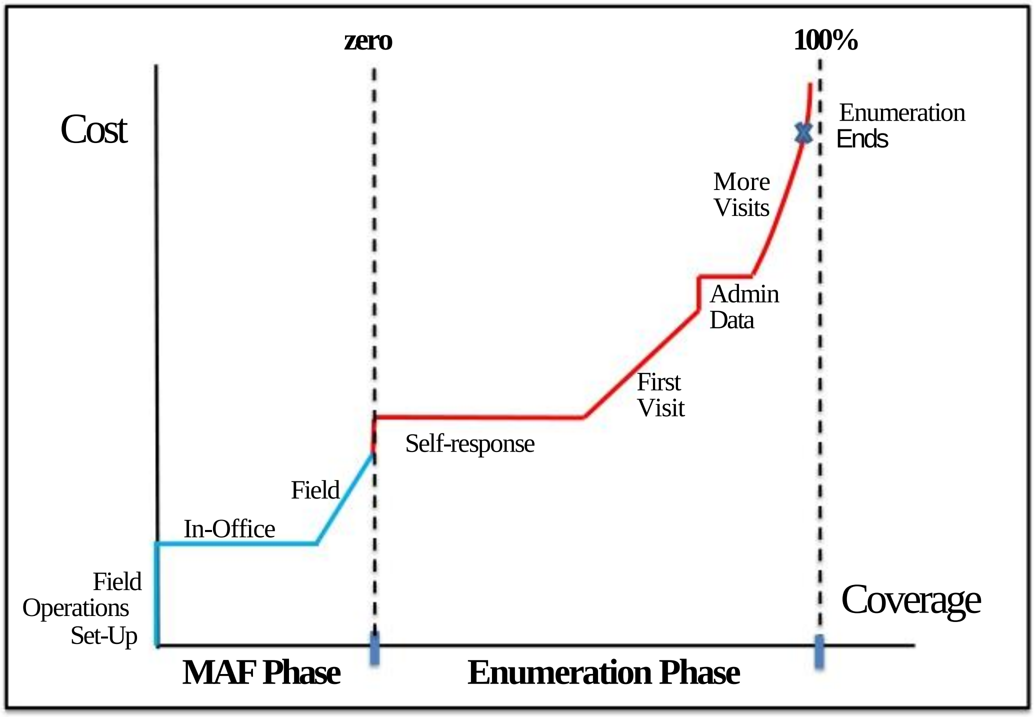


Figure 3.5: JASON illustrative depiction of cost versus coverage for 2020 Census operations.

3.4 Observations on Current Census Process

Some final observations regarding the census is that is it expensive and the costs continue to grow, although for the 2020 innovations (“in-office” MAF canvassing, Internet response, and administrative data use in NRFU) are helping to hold 2020 cost estimates to 2010 Census levels. For 2020, there will be an estimated 250,000 temporary workers (450,000 in 2010) to be employed and they will visit more than 110 million housing units. The census is imperfect, although the net error has been decreasing. However, various demographic groups continue to be disproportionately miscounted. Even with this increased accuracy, only 94.7% of the population living in physical housing units was “correctly” counted in 2010, meaning that the remaining 5.3% of the population’s data needed to be corrected or imputed (U.S. Census Bureau, 2012a).

It is unclear given the current approaches taken by the Census Bureau if or how more resources would improve coverage and correctness of the data. New approaches to the census enumeration could prove beneficial and will constitute the remainder of the discussion in this report.

4 FUTURE STRATEGY

The 2030 Census enumeration strategy proposed by JASON leverages the increased availability of high-quality government administrative data. Specifically, JASON recommends that the Census Bureau consider starting the 2030 Census with an "in-office" enumeration of the population using existing government administrative records, e.g. data collected by the Internal Revenue Service (IRS) and Social Security Administration (SSA).

A transition to an "in-office" enumeration using administrative records suggests a paradigm shift in the way the Census Bureau conceptualizes their enumeration. Figure 4.1 contrasts the 2020 Census plan with the proposed strategy for 2030. Traditionally, the Census Bureau has used the Master Address File (MAF), a list of housing units as a frame, for enumerating the people in each unit. Government administrative records are organized at the level of individuals, indexed by Social Security Numbers or Tax Identification Numbers. A census that starts with directly enumerating individuals through administrative records would involve identifying individuals first and then assigning them to their appropriate residences as opposed to the current process of identifying residences and then populating them. This section dis- cusses some of the social, demographic, and technological trends that support such a paradigm shift.

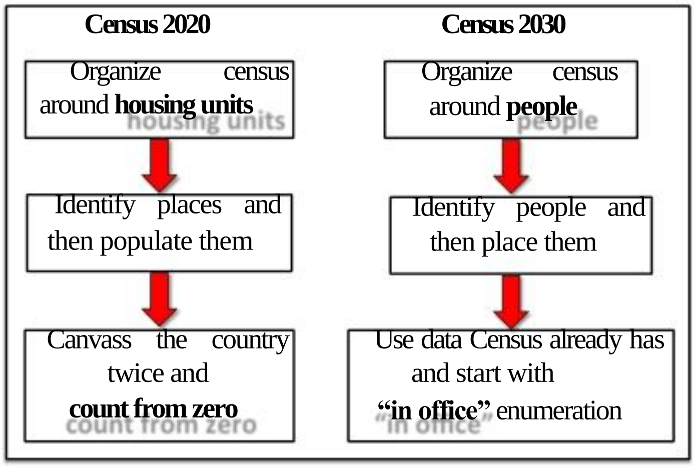

Census 2020
Census 2030
Organize census around housing units
housing units
Organize census around people
people
Identify places and then populate them
Identify people and then place them
Canvass the country twice and count from zero
count from zero
Use data Census already has and start with "in office" enumeration
"in office"

Figure 4.1: Paradigm shift in the conduct of the census for 2030.

4.1 Trends in Support of Paradigm Shift

The Census Bureau's decennial questions and the methods of asking them have slowly evolved, reflecting changes in family structure and communications. Initially, census enumerators asked the name of the head of the household along with the number of people in the household subdivided into a few categories. More recently, the names of all household members are recorded along with their age and gender and whether they are related to the householder (i.e., the person within the housing unit filling out the form). In 2010 paper forms mailed to the MAF were the primary means of collecting data. To reduce the cost of the 2020 Census, postcards with a url of an online form will be sent to MAF addresses. Non-responders will receive paper forms and if necessary phone calls and visits from enumerators.

Past changes in taking the census and continuing trends in family structure and technology suggest that changing the basis of the census from housing units to individuals should be considered. Specifically, the household structure is changing. The household size (number of individuals within a housing unit) has been and continues to decrease reducing the numerical ad- vantage of focusing on housing units. In addition, a majority of the U.S. population changes residences between decadal censuses. This situation is further complicated with a growing number of vacant or second homes that need to be validated and then removed from the MAF.

One serious concern for the Census Bureau is trust in government. Although not proven empirically, it is believed lack of trust in government has a negative effect on government data collection. All indications are that trust in government continues to decline and what this will look like in 2030 is unclear.

Communication continues to become more individualized with cell phones and email replacing U.S. Mail and landlines. There is rapid growth in the availability of data sources, both government and commercial, about individuals in the population. These can be viewed as enablers to directly enumerating individuals “in-office.”

4.2 Household Structure

The definition of household by the Census Bureau has followed the evolving structure of U.S. families.   Under Thomas Jefferson as director, the 1790 Census asked for the name of the head of the family and for numbers of other family members in five categories: free white males 16 years and older, free white males under 16, free white females, all other free persons, and slaves. The 1790 Census found 558,000 families with an average size greater than 5.1 persons.[1]

Reflecting the rising importance of wages versus family farming, in 1870 the basis for a household was changed from a common means of support to shared

[1] Some of the 1790 records have been lost, and the tabulation found had 35.8% of responses in the category of seven or more individuals (Infoplease, 2004). Assuming only seven members in this category produces a minimum of 5.1 members per family.

eating (Ruggles and Brower, 2003). In 1930 group quarters were recognized, and in 1950 the locations of college students were shifted from their family homes to their college quarters. In all census years, however, most households were defined as people related to one another and living together in a dwelling (housing unit) with shared cooking and eating facilities.

The 2010 Census asked the name, date of birth, gender, and race of Person 1, i.e., the person filling out the census form, in addition to the number of people staying in the housing unit on April 1st, whether there were additional people, and whether the unit was a house, apartment or mobile home (U.S. Census Bureau, 2010). For each additional person the form requested name, birth date, gender, age and whether the person was related to Person 1. This census found 116.7 million housing units containing an average household size of 2.59 people in a population of 308.7 million (Lofquist et al., 2012). For comparison, in 1960 there were 52.8 million housing units with an average household size of 3.29 people in a population of 180.7 million (Hobbs and Stoops, 2002). Reasons for the recent changes include more married people living to old age with the means to live alone, even after one spouse dies; later marriages and more divorces; and fewer children per family (Glick, 1984). The trend continues, with 2.54 individuals in the average 2015 household.

Ruggles and Brower (2003) criticized the evolving household definitions used by the Census Bureau and recommended changing the basis for enumeration,

> “All measures should be taken at the individual level except where there is a compelling reason to use household-level measures.”

This view is consistent with the findings and recommendations of this JASON study.

4.2.1 Changing residences

According to Hansen (1998), in 1948 20.2% of the population (28.7 million individuals aged one and over) moved every year. In 2015, the percentage decreased to 11.5%, but, owing to the increased population, the total number grew to 35.7 million. The cumulative effect is that the median duration in their present residence is only 5.2 years for individuals 15 years and older. More pertinent to the decadal census, 66.7% have been in their present residence 9 years or less (see Figure 4.2).

Figure 4.2: Duration at present residence for individuals 15 years and older. Data from Hansen (1998).

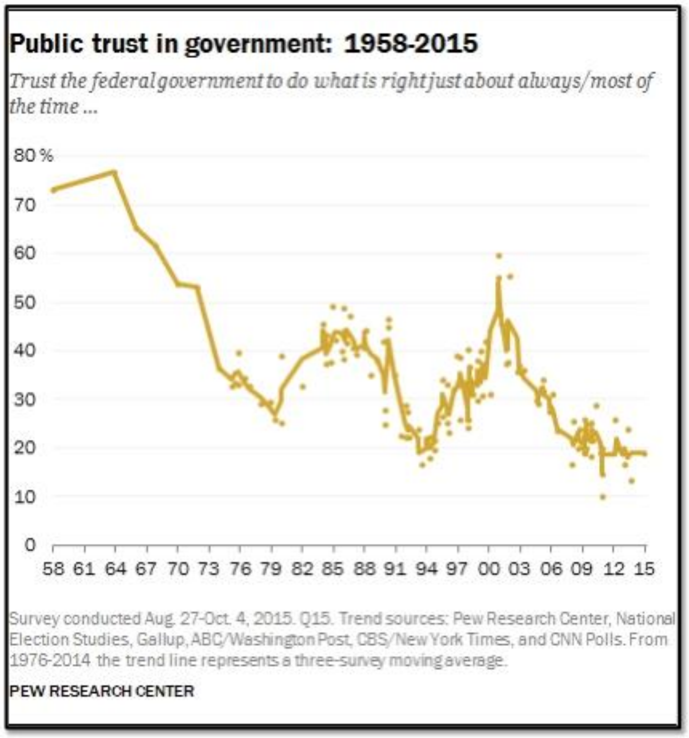
4.2.2 Second homes

The 2001 American Housing Survey (AHS) revealed that the 108.2 million households in the U.S. owned 12.9 million residences (housing units) that were either vacant or not their usual residence (HUD, 2004). The 2001 Housing and Vacancy Survey (HUD, 2004) put the number of secondary residences higher, at 14.5 million. The trend continues. The 2013 AHS estimated 132.8 million housing units with 13 million either vacant or not their usual residence. Thus, under current census procedures, the addresses of approximately 10% of the U.S. housing stock need to be identified and removed from the list of viable household addresses for enumeration (MAF) before or during the decadal census enumeration.

4.3 Trust in Government

Lack of trust in government could negatively impact an individual's willingness to respond to a government survey or census. In mid-2015, 19% of Americans indicated that they "Trust the government always / most of the time" (Pew Research Center, 2015a). The historically high level for this question was 77% in 1964. The current level of trust, 19%, was previously recorded in 1994, and there is significant variability in the measure, as illustrated in Figure 4.3. The primary reasons for this variability are the state of the economy and the levels of military and political conflict, e.g., Vietnam War controversy, terrorist attacks of September 11, 2001 terrorist attacks (Pew Research Center, 2015a). The highest level of "trust in government" since 1980 was 54% on October 25, 2001.

Going forward to 2030, this trend could stay the same, improve, or get worse. In all cases, the past history indicates that national and world events drive substantial variability. The trust concerns of the American public depend heavily on recent events. Terrorism appears to be one of the factors that may change the public's views on government monitoring. Pew Research Center (2015b) found that recent terrorist acts by agents of ISIS have resulted in more American adults feeling that the government is not doing enough to curb acts of terrorism by more electronic surveillance of individuals. The concerns over civil liberties and privacy of individuals has fallen dramatically to 28% of those surveyed in December 2015 compared to the 47% in July 2013, in the aftermath of the Edward Snowden leaks.



Public trust in government: 1958-2015

Trust the federal government to do what is right just about always/most of the time ...

Survey conducted Aug. 27-Oct. 4, 2015. Q15. Trend sources: Pew Research Center, National Election Studies, Gallup, ABC/Washington Post, CBS/New York Times, and CNN Polls. From 1976-2014 the trend line represents a three-survey moving average.

PEW RESEARCH CENTER

Figure 4.3: Graphic on trust in government from Pew Research Center (2015a).

It is difficult to predict the type of events that will occur by 2030 which may negatively impact the public's perception of government data collection as part of the decadal census. An enumeration process that has less reliance on direct self-responses could help mitigate the risk of falling response rates due to fluctuations in public trust in government.

4.4 Communication

The Census Bureau is well aware of the trends in Internet and mobile phone usage and plans to take advantage of this in 2020 by allowing Internet responses through the use of mobile apps. These trends are likely to continue as 2030 approaches.

4.4.1 Internet usage

Based on the 2013 American Community Survey of 60,000 households, File and Ryan (2014) found that 74.4% of households have Internet access, nearly all via high-speed connections. Significantly lower access was found in Black and Hispanic households, in households with limited English speaking or incomes less than $25,000, and when the householder had not graduated from high school.

Statistics from Pew Research Center (Perrin and Duggan, 2015) show 85% of U.S. adults have access to the Internet. The number of people with access has grown since the 1990s, but the growth rate has decreased in recent years. As seen in Figure 4.4, the percentage without access has been dropping steadily, but it appears to be stabilizing near 15%. As of 2015, 96% of Americans aged 18-29 have access to the Internet, but only 58% of the adults older than 65 have access. Besides age, Internet use is also correlated with income and level of education. By 2030, if the adults currently with access are assumed to retain access, then at least 81% of seniors will use the Internet. Accordingly, the total number of American adults with Internet access should increase. Given the demographics of households without access (File and Ryan, 2014), further decreases in the Internet access growth rate will likely be slowed.

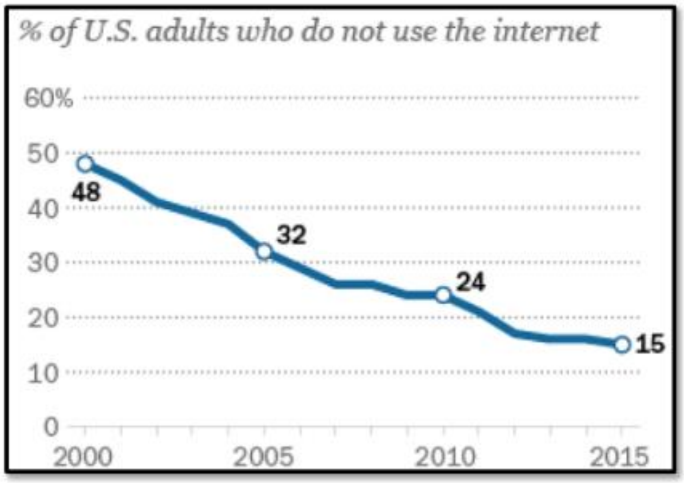


Figure 4.4: Pew Research Center surveys of U.S. adults, 2000-2015. Data from 2015 includes surveys conducted March 17-April 12, May 28-31 and June 10-July 12. Source: Anderson and Perrin (2015).

4.4.2 Land lines and mobile phones

Mobile phones are the primary means of contacting increasing fractions of the U.S. population. In 2014, over 40% used only cell-phones and fewer than half of households had landlines (Richter, 2015). Figure 4.5 shows the gap between landline and cell-phone usage closing. According the Pew Research Center (Anderson, 2015), in 2015 98% of adults had cell phones. One informal online survey (Blowoutcards.com, 2016) reported that 55% of respondents had kept the same cell phone number for at least 10 years; another (Crackberry.com, 2016) reported 28% holding the same number for a decade. In either case, retention of cell phone numbers is comparable to or greater than the 31% of the population older than 15 who maintain one residence for a decade or longer (Ruggles and Brower, 2003).

Groups with low cell phone usage include rural populations, those with no college education, households with annual incomes less than $30,000, and those older than 65 (Smith, 2015). Many are likely in the “hard-to-count” category for census enumeration.

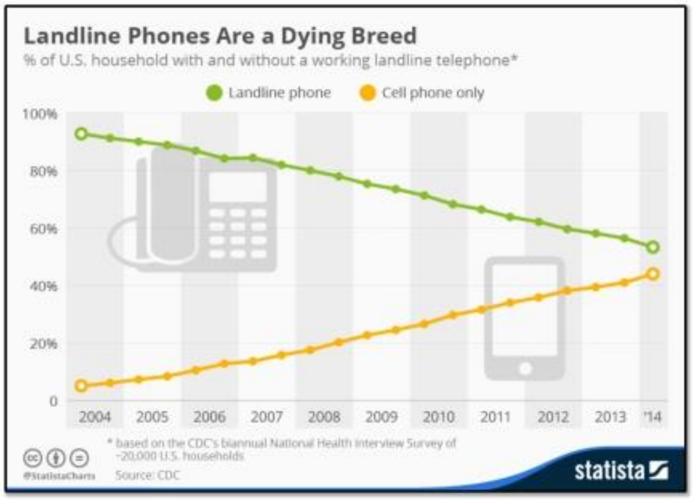


Figure 4.5: Households with landline phones (green) and with only cell phones (orange). Source: Richter (2015).

If retention of cell phone numbers continues, by the 2030 Census it might be considerably easier to build a master list of cell phone numbers than to survey all housing units.

4.5 Data Availability

The growth in available data, both government and commercial data, should serve as enablers for the 2030 Census. First, most data collected by the Census Bureau in 2020 will already exist in government administrative records. Due to Title 26 and other agreements made between the Census Bureau and other government agencies, these data are in principle available to the Census for statistical purposes, as in enumerating the population. The Census Bureau already plans to make use of administrative records in the 2020 non- response follow-up (NRFU) operations. The promise of these administrative records is discussed in detail in Section 5.

Commercially available data is simply exploding. Private sector entities are acquiring and selling data about the population. Here too, the Census Bureau has already begun to explore the utility of using external commercial and locally available public data sources to augment and enhance their statistical products (Keller et al., 2016). These data sources can range from housing and education to medical health records.

As the Census Bureau is well aware, even if trust in government were quite high, there would still be concern about sharing information because of the potential for losses by computer hacking. Recent large data breaches, such as Office of Personnel Management, Target Corporation, Citibank, Australian 2016 Census, and others, make people naturally concerned about the safety of their personal information.

According to Rainie and Duggan (2015), many Americans are willing to provide private information to companies or the government depending on incentives and tradeoffs between risk and gain. For example, 47% accept the tracking of their shopping habits by retailers for receiving discounts as part of customer-loyalty programs.   When providing personal information over the Internet, Americans weigh the risks of how the information may be used, how the information may be stolen, and how the information may be stored. Internet users believe that the government is one of the more trustworthy data collectors and place confidence in the security of the information regarding their private activities (Rainie and Duggan, 2015).

For the 2030 Census, the Census Bureau could consider devising incentives that the public might be willing to trade for what they consider to be sensitive information. For example, the Census Bureau could partner with the commercial sector, offering special incentives if consumers agree to share their data with the Census Bureau. This could become part of their commercial “license agreements.” An even stronger arrangement could be made with service providers, say the utility companies, where access to the service could require agreement to share your data with the Census Bureau under Title 13 protections. The value of the incentives will likely change depending on the public perception and trust in government.

5 ADMINISTRATIVE RECORDS

The Census Bureau has access to an increasing amount of high-quality government administrative data. Significant examples include data collected by the Internal Revenue Service (IRS), Social Security Administration (SSA), and the Centers for Medicare and Medicaid Services (CMS). This Section provides a discussion of IRS and SSA administrative records and their potential for use in the creation of an "in-office" enumeration of the population for the 2030 Census.

5.1 Administrative Data Contain Rich Data

IRS and SSA records are likely to be the richest source of administrative records information for the Census Bureau's purposes. The IRS collects a wide range of tax-related information on U.S. citizens and residents. The IRS database includes individual 1040 tax returns, W-2 forms submitted by employers, 1099-Misc forms for contract work, 1099-G forms recording unemployment insurance benefits, 1099-Int forms submitted by banks for accounts with interest payments greater than $10,1098 mortgage interest forms submitted by lenders, 1098-T forms submitted by institutions of higher education, and a wide range of additional filings.

Tax returns and IRS informational filings can be linked using Social Security Numbers (SSNs) or Individual Tax Identification Numbers (ITINs). The SSA maintains a master file of all SSNs that includes information from SSN applications. Since 1989, all US states have followed a policy of issuing SSNs at birth. For individuals who must pay taxes but who do not have an SSN, the IRS will issue an ITIN.

The combination of data fields collected by IRS and SSA includes much of the information in the Census short form. Individuals report their current mailing address on their 1040 tax return and individual addresses are also included in many informational filings, including W-2 forms and the 1099 and 1098 forms mentioned above. These tax filings do not ask for date of birth, gender, race or ethnicity. However, both SSN and ITIN applications collect date of birth and gender, and SSN applications collect a coarse version of individual race and ethnicity.

An important point to emphasize is that even individuals who do not file income tax returns are likely to appear in IRS records due to third-party information filings. For example, a retiree with only social security income will not need to file an income tax return, yet the SSA will file an SSA-1099 with the IRS. Similarly an individual with employment income may receive a 1099-G if she received unemployment benefits, and will be in the IRS database if she received employment income (and hence a W-2) or contract income (and hence a 1099-MISC) in prior years.

The U.S. tax system also provides individuals and married couples with a strong incentive to report dependents on their tax returns in order to take advantage of the dependent tax exemption or the earned income tax credit. The current version of the 1040 form includes room for the names and SSNs of four dependents, with an additional check box if a filer has more than four dependents.

While IRS and SSA records provide especially rich data for the Census Bureau's purposes, other agencies such as CMS or the Immigration and Naturalization Services also may have useful administrative records. State and local administrative records will also prove useful such as individuals registered for Medicaid, Supplemental Nutrition Assistance Program, educational State Longitudinal Data Systems, or local real estate tax assessments.

Some of the information collected by federal, state, and local agencies is clearly redundant. However, there is value for the Census Bureau in obtaining redundant administrative record information as it can be used for error correction and validation. This will be particularly helpful if an individual has moved recently or has multiple homes, two situations expected to be prevalent for the 2030 Census enumeration (see Section 4.2).

5.2 Administrative Records Are Useful

The use of administrative records for census is not a new idea (Alvey and Scheuren, 1982). Key questions for JASON's proposed 2030 Census strategy are:

- Exactly how many individuals who should be enumerated can be found in administrative records?

- How much information about these individuals can be found?

Recent research at the Census Bureau and by academics provides some initial evidence on the potential usefulness of IRS and SSA administrative records.

The Census Bureau has done considerable research on linking administrative data to census data, including an extensive study that links the 2010 Census to IRS and SSA records (Rastogi et al., 2012). This study first looked at the link rates for matching housing units between the administrative records and the addresses recorded in 2010 Census enumeration. The overall match rate was 92.6%. Figure 5.1 shows the variability in address match ratios by geography.

An interesting point of comparison that speaks to the increasing value of the administrative records for use in the decennial census process(s) is to compare the

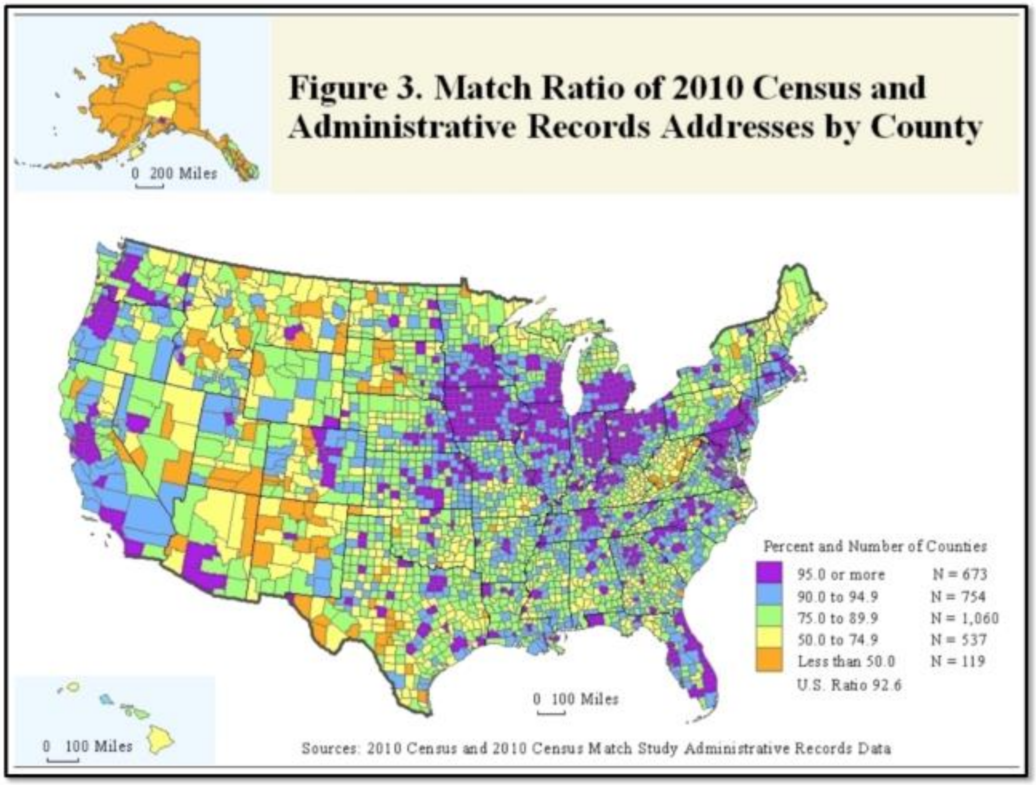


Figure 3. Match Ratio of 2010 Census and Administrative Records Addresses by County
0 200 Miles
0 100 Miles
0 100 Miles
Percent and Number of Counties
95.0 or more N = 673
90.0 to 94.9 N = 754
75.0 to 89.9 N = 1,060
50.0 to 74.9 N = 537
Less than 50.0 N = 119
U.S. Ratio 92.6
Sources: 2010 Census and 2010 Census Match Study Administrative Records Data

Figure 5.1: Match rates between administrative records and 2010 Census addresses (housing units) taken from Rastogi et al. (2012).

2010 match rate (92.6%) to the match rate from an administrative records experiment the Census Bureau conducted in 2000 (Bauder and Judson, 2003). Admittedly, it is not a direct comparison, but it is still useful. In 2000 the Census Bureau did a study that linked administrative data from IRS, Housing and Urban Development (HUD), CMS, Indian Health Services (IHS), and the Selective Service System (SSS) to the Master Address File (MAF) records for two counties in Colorado and two counties in Maryland. The match rate from this experiment was 81.0%.

The 2010 study (Rastogi et al., 2012) also considered direct matches with individuals and was able to match 98.0% of the population that had unique protected identification keys in the 2010 Census enumeration and 88.6% of the entire 2010 Census population. The match rates vary geographically and also by socio-demographic groups. Figure 5.2 shows the variability in these match ratios by geography. The 2010 study did a third match comparison looking at “person-address” pairs. The match rate for this analysis was slightly lower than the address match rates because the issues in address and person matching were compounded. Nevertheless, the results are impressive and have given the Census Bureau confidence in the use of administrative records to support the 2010 non-response follow-up (NRFU) field operations (U.S. Census Bureau, 2015; Morris et al., 2015).

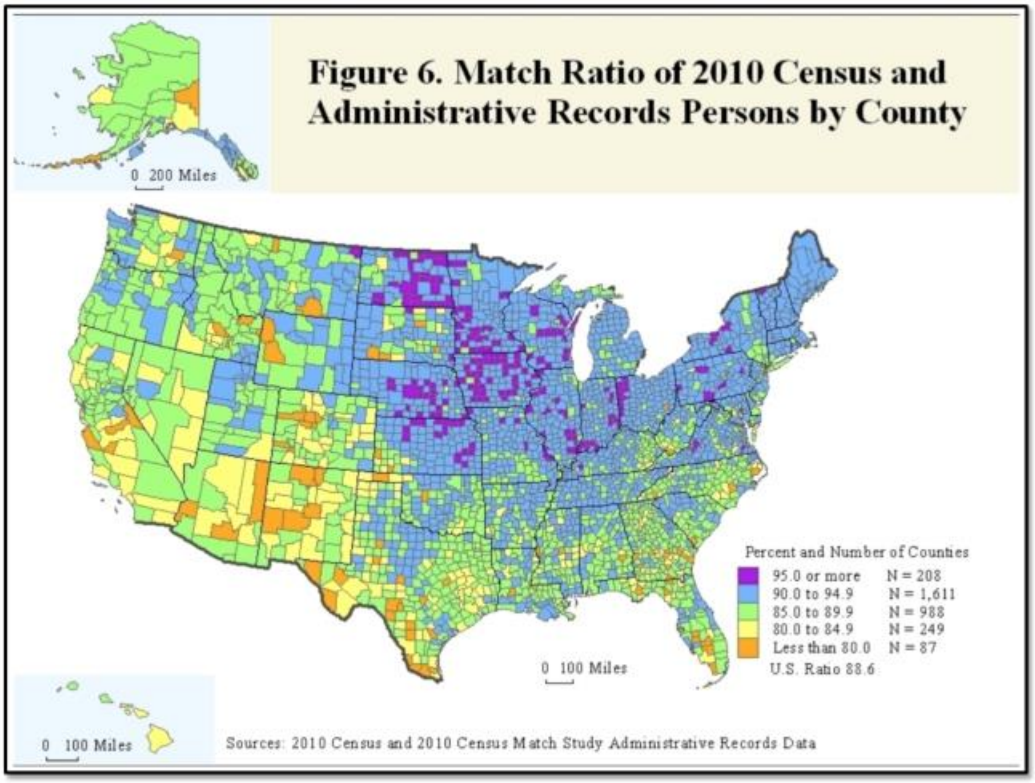


Figure 6. Match Ratio of 2010 Census and Administrative Records Persons by County
0 200 Miles
0 100 Miles
0 100 Miles
Percent and Number of Counties
95.0 or more N = 208
90.0 to 94.9 N = 1,611
85.0 to 89.9 N = 988
80.0 to 84.9 N = 249
Less than 80.0 N = 87
U.S. Ratio 88.6
Sources: 2010 Census and 2010 Census Match Study Administrative Records Data

Figure 5.2: Match rates between administrative records and 2010 Census enumeration of persons (individuals) taken from Rastogi et al. (2012).

These coverage rates and trends found in the Census Bureau studies using IRS and SSA data (Bauder and Judson, 2003; Rastogi et al., 2012) are also being found in relevant academic work. A study by Mortenson et al. (2009) used IRS data from 2003 to characterize the population of non- filers. Their study reports that 259.0 million residents could be found on an individual tax return. This can be compared to the Census Bureau’s estimate of 294.3 million residents alive at some point in time during 2003. In addition, Mortenson et al. (2009) report being able to find 29.9 million “non-filers” in IRS informational filings. They conclude that “the federal tax system contains information on 98.2% of persons who were U.S. residents during 2003.”

Finally, recent work by Chetty et al. (2014) contains some useful statistical counts on the number of individuals in different birth cohorts that can be identified in administrative data. Chetty et al. (2014) focus on counting individuals in each birth cohort who are alive, can be found in administrative records and linked to a parent. They benchmark their counts against estimates of the size of each birth cohort from the 2012 U.S. Statistical Abstract (U.S. Census Bureau, 2012b). For the 1980 to 1991 birth cohorts, their count in the tax records is at least 96%, as large as the number in the U.S. Statistical Abstract.

These recent research exercises indicate that administrative records have promising coverage of the U.S. population for the Census Bureau’s purposes. However, substantial additional research is needed on the extent of administrative records coverage, on the characteristics of those individuals who do not appear, as well as on the prevalence of missing data fields that would need to be filled in for the census enumeration.

6 CENSUS 2030 PROCESS AND PREPARATION

Adopting the paradigm shift presented in Section 4 of centering the census enumeration directly around individuals versus housing units provides a new strategy for the conduct of the 2030 Census. This strategy builds on the innovations in the Census 2020 Operational Plan (U.S. Census Bureau, 2015), specifically the Census Bureau's proposal to make new use of administrative records. Recall the 2020 plan calls for using records from the 2010 Census and US Postal Service (USPS) to construct approximately 75% of the Master Address File (MAF) "in-office," with the remainder of the work done "in-field." Later in the enumeration process, administrative records from the Internal Revenue Service (IRS) and Social Security Administration (SSA) will be used to optimize the non-response follow-up (NRFU) visits and in certain cases to impute a housing unit response.

The starting point from which the Census Bureau can build their 2030 plan is to conduct an "in-office" enumeration of the population (people) using administrative records, geographically place them, and then "fill in" the gaps of what is known. Based on the research discussed in Section 5, it is possible that nearly 90% of the population may be able to be enumerated "in-office." The gap filling will require additional data and more traditional "in-field" methods to find people and variables that might be missing or are not present in government records.

6.1 Step 1: "in-office" Enumeration

The first step in the 2030 strategy will be to conduct an "in-office" enumeration of people directly from the rich source of IRS and SSA records and past census data (e.g., 2010 and 2020 Censuses and American Community Surveys). The

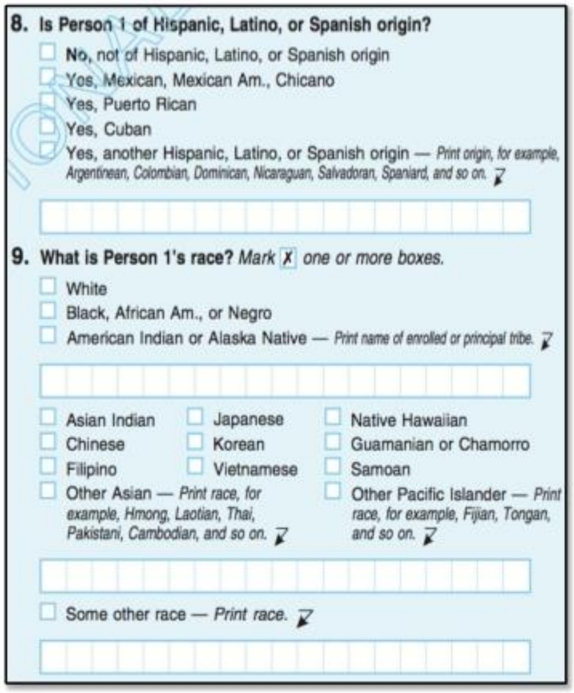
Census Bureau should not limit their use to the current year IRS data, rather they should leverage the entire IRS database going back to 1996. JASON recommends the initial administrative record focus be on IRS and SSA administration records to avoid overly diffusing the development of an "in-office" enumeration process that could be operationalized by 2030.

Research can start now to develop the "in-office" enumeration process. The research should be able to immediately identify some of the subpopulations that may be missing from the linking of this data, as well as variables such as race and ethnicity that may need to be obtained through other sources. Individuals may have multiple addresses in IRS or other administrative records. Methodology to reconcile these will need to be developed. The Census Bureau should continue to harness the growing body of research on the use of administrative records to characterize the population.

6.1.1 Challenges with race and ethnicity

JASON recognizes the Census Bureau will be presented with some challenges on gathering race and ethnicity information through administrative records. Figure 6.1 gives the current format for collecting race and ethnicity data through the census. This self-reporting is complex with many options for response. However, for most people, race and ethnicity may only need to collected once in their lifetime and can be carried forward in time through the "in-office" enumeration process(es).

There are growing numbers of government administrative records that could be used to "fill in" the race and ethnicity gaps in the enumeration. Past census and American Community Survey data will have the detailed in- formation. SSA currently collects a coarse version identifying individuals as Black, Hispanic, or Caucasian. The Affordable Care Act is requiring states to collect information on race and ethnicity, as well as other social determinants for Medicaid



8. Is Person 1 of Hispanic, Latino, or Spanish origin?

□ No, not of Hispanic, Latino, or Spanish origin
□ Yes, Mexican, Mexican Am., Chicano
□ Yes, Puerto Rican
□ Yes, Cuban
□ Yes, another Hispanic, Latino, or Spanish origin — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

9. What is Person 1's race? *Mark* ☒ *one or more boxes.*

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

□ Asian Indian
□ Chinese
□ Filipino
□ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*

□ Japanese
□ Korean
□ Vietnamese

□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*

□ Some other race — *Print race.*

Figure 6.1: Current census form questions for race and ethnicity.

reimbursements. The Census Bureau could develop a partnership with U.S. Citizenship and Immigration Services to gain some of this information. Finally, beyond federal-level data, other sources could be considered such as school records or commercial electronic health records.

6.2 Step 2: Multi-Faceted Follow-Up

The activities to follow the "in-office" enumeration will need to be developed. It is important to point out that unlike NRFU activities that typically involve more than 35% of the housing units across the country; this follow-up should involve approximately 10% of the population. These estimates may be inaccurate, but early research on the "in-office" approach should be able to determine the validity and value of this approach to the Census Bureau.

Some of the 2030 follow-up activities should be similar to the current NRFU approaches for reaching "hard-to-count" populations. The Census Bureau has decades of experience here. However, the "hard-to-count" populations that emerge from the administrative records enumeration may be different from the ones the Census Bureau has typically sought. Early research on linking the administrative records to create the "in-office" enumeration, along with some field experiments should help identify types of people, variables, and geography where potential miscounts may occur. This will help the Census Bureau adjust their "in-field" operations for NRFU in 2030.

Some natural outgrowths of the current Census Bureau's NRFU methods should also apply for 2030. These would include directly eliciting self-responses and developing (maintaining) key partnerships. The 2030 self-responses could be acquired using the most recent mail (email) or phone (mobile) information found in the IRS and SSA records for these individuals. Developing new and maintaining existing partnerships should be a priority. For example, partnerships with Housing and Urban Development (HUD) can be useful for enumerating the homeless, Federal Emergency Management Agency (FEMA) to find those temporarily displaced, U.S. Department of Agriculture (USDA) to identify geographic areas with migrant worker, or the Bureau of Indian Affairs to calibrate counts on Indian land.

Three potentially useful approaches for filling in the gaps that may be new in the context of the proposed 2030 strategy are the following:

- Citizen Enumeration: Solicit help from citizens, local organizations, crowd-sourcing, and trusted civil servants.
- Sensor Counts: Non-surveillance use of technology to collect new data and to guide follow-up field activities.

- Data Linking: State records, prison records, Medicaid records, HUD, consumer retail activity, address registries, etc.

6.2.1 Civic enumeration

The Census Bureau should look to developing field operations for 2030 that will help them identify where counts found in the “in-office” enumeration are inaccurate, both with respect to socio-demographic characteristics and geographic regions. Crowdsourcing could offer such an opportunity by enhancing the Census Bureau’s fieldwork through volunteer activities. A classic example of the use of crowdsourcing for geographic location of populations is the Audubon Christmas Bird Count (Audubon Society, 2015). The most common term for this activity is citizen science, but that is too generic (e.g., Chow (2013)). The term Citizen Enumeration will be used here.

The Census Bureau has successfully used crowdsourcing for the creation of metrics associated with mail self-response rates and hard-to-count geo- graphic areas (Erdman and Bates, 2014). This involved using Kaggle.com to host a worldwide data analysis competition for the Census Bureau (Kaggle, 2012). In contrast, JASON proposes the use of crowdsourcing in support of the 2030 enumeration that would put volunteers in the field to collect information by conducting Citizen Enumerations.

The idea is to encourage and offer limited support (conveniences) to volunteers who would describe their neighborhoods or communities of interest. This should be made easy for individuals to document locations where they are collecting information (e.g., addresses and/or GPS coordinates) and the numbers of people associated with each location, so that they can effectively serve as volunteer enumerators. It is important to be clear that JASON is not advocating this information replace the actual census enumeration data.

It could simply be used to supplement and help prioritize the “gap filling” operations following the “in-office” enumeration step.

In particular, Citizen Enumeration might be effectively applied to “hard-to-count” populations, such as Native Americans. Various non-governmental organizations (NGOs), local businesses, schools, fire or law-enforcement districts, casinos, health-care providers, and community or cultural centers, might be motivated to provide such volunteered services, either out of altruism (and the challenge involved) or if the product is viewed as serving some of their own needs. In order to be successful, the Census Bureau would have to develop means of collecting the information that is convenient and reliable both for the volunteers and for the Census Bureau, and perhaps making the crowdsourced statistical products part of the Census Bureau’s public data releases.

One approach might be to encourage texting geocoded photos taken in the field using a smartphone, perhaps with a comment on the numbers of individuals at the location of the picture (photos are often geocoded even without the user’s knowledge, or else geocoding could be specifically implemented). Though convenient this might easily be viewed as far too intrusive, and a more palatable alternative would be to simply use geocoded texts indicating a volunteer’s estimate of the number of individuals at a given location (geocoding can be added to texts from smartphones or tablets).

Clearly, many details would have to be worked out and experiments conducted to assess the value added by Citizen Enumeration to the census enumeration process(es). This basic citizen science approach has demonstrated merit for documenting populations and their changes (see reviews by Crain et al. (2014), Marshall et al. (2015), and Edgar et al. (2016)). Validation of citizen science has been an important concern, and has been addressed through one or more of i) independent observations; ii) testing of volunteers (before, during or after the tasks); and iii) mutual consistency between multiple volunteers' results (Gillett et al. (2012), Butt et al. (2013), and Vianna et al. (2014)). Results to date have been surprisingly encouraging.

Acceptance of Citizen Enumeration to support the Census Bureau operations does not come without some concerns. Johnson and Sieber (2013) discuss the reasons that governments have difficulty incorporating such volunteered information (their work refers to open, democratic societies, such as the U.S. and Canada). However, there are very successful government examples such as the National Weather Service's Cooperative Observer Program (NOAA, 2016) that has been providing essential meteorological information for more than a century.

6.2.2 Sensor counts

Today there is an explosion of sensors enabling complete and persistent coverage of people, infrastructure, and the environment. This includes video, portal monitoring (e.g., EzPass), vehicular counting, cell phone traffic, etc. Figure 6.2 is just one of many examples. The country is experiencing ubiquitous connectivity for both people and things and this trend is expected to continue, if not accelerate. Already, cellular and WiFi infrastructures are being built up and exploited in unexpected ways. Biometric technologies are improving and proliferating. This includes everything from DNA matching and finger printing to gait analysis from video images. Robotic technologies such as autonomous passenger and freight vehicles and customer-facing services may play a role in increasing population mobility or taking people off the grid completely.

All of these technologies, often integrated, are manifest in a variety of venues, including commerce, defense, finance, political campaigns, health, and security. The "Internet of Things" is a term coined to convey the concept

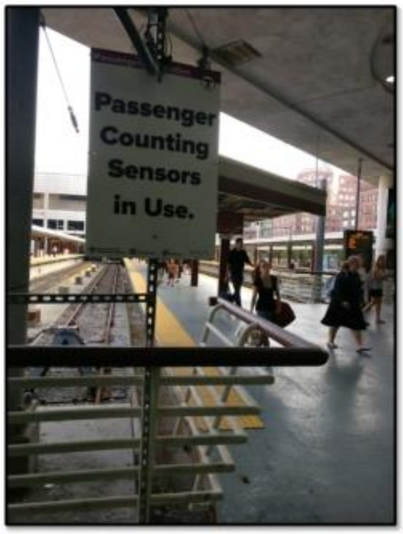


Figure 6.2: Counting people at train station in Boston.

of a global network of connected objects. Dissimilar WiFi-capable devices that connect to each other (e.g., vehicles, smartphones, household appliances, machines, security systems, wearable devices) are establishing a network of physical devices that continually exchange information. "Smart cities" are leveraging these data technologies in such applications as trac and parking management, improved energy efficiency, better delivery of municipal services, public health and safety.

Privacy concerns under current law and societal norms would emerge if the Census Bureau proposed using this these technologies for direct enumeration. However, similar to the proposed use of Citizen Enumeration, the Census Bureau could consider exploiting these technologies to help cue and prioritize follow-up activities for filling in the gaps from the "in-office" enumeration step. For example, some of these technologies could be used to help verify the aggregate count in a census block or tract. Failures to verify the aggregates would represent "outliers" in the enumeration and could be useful in efficiently allocating resources for more detailed follow-ups. Provocative examples of what could be currently done to find both individuals and aggregates are given next.

The cellular system generates, as part of its operation, geolocation data for all operating handsets. Although individual level (single handsets) cellular data are currently solely in the hands of the cellular carriers, cellular verification at the aggregate level is being commercially sold. Airsage (2016) aggregates handset geolocations from cellular carriers and sells analyzed datasets with geo-resolution to the census block level. Also, the count (and nationalities) of handsets served by any particular cell tower at any time can be obtained in real time through a commercially available software defined radio within tower range.

Aggregate count verification might also be accomplished through street- level sensors in a city. For example, pings collected by urban WiFi networks are specific to the unique MAC (Media Access Control) address on WiFi enabled devices. The analyses of pedestrian ping patterns over time allows segregation into residents (appear regularly at all hours/days), workers (appear regularly only during working hours on workdays), and visitors (all others). Figure 6.3 presents some recent data collected in lower Manhattan in New York City (Kotokosta and Johnson, 2016). The proliferation of urban WiFi networks could allow this sort of verification to be accomplished on a broad scale. A similar approach would be through gait and/or facial analysis of persistent street-level imagery.

6.2.3 Linking data

The JASON proposed strategy for the 2030 Census will require a considerable amount of linking data for a successful "in-office" enumeration. At the start of the enumeration, the "people," in essence, are "hidden" and need to be identified (imputed). Places, or geo-locations, also need to be identified and attached with high probability (⇠ 1.0) to each individual. Research in graphical and generative models is moving quickly and the Census Bureau should use these innovations to their advantage for their massive heterogeneous data integration.

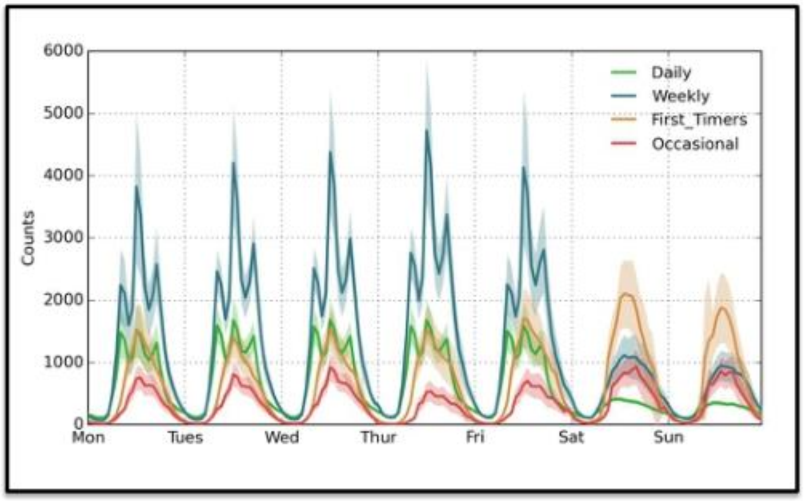


Figure 6.3: Weekly WIFI activity, by user group, for network in lower Manhattan (Kotokosta and Johnson, 2016

Below is a brief description of some of these advances.

A graphical model is a probabilistic representation of a casual related set of random variables (nodes of a graph) that are linked by conditional dependencies. The dependencies may be as general as a complete multivariate joint distribution over the input variables, or any kind of simpler model (e.g., a univariate distribution that depends only on a function of the inputs). Figure 6.4 shows a simple example with four variables. At scale, state-of-the-art graphical models today can have thousands or more variables. By the year 2030, models with millions of variables may be solvable.

What makes graphical models useful are that they can be "solved" (a process usually awkwardly termed, "have inference performed on them,") conditioned on data. That is, some nodes in a graphical model may be "visible," while others are "hidden," Then, visible nodes can be set to measured or known values, after which the probability distributions (including, if desired, joint distributions) of all the other "hidden" variables can be calculated. There is a large literature about how, computationally, to perform this inference (Koller and Friedman,

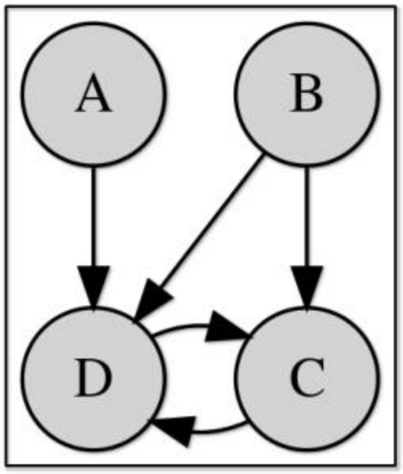


Figure 6.4: Graphical model with four random variables. A and B are independent variables, D depends on A, B, and C; and C depends on B and D. Source: Graphical model. (n.d.) (2016).

2009). Thus, the “secret sauce” for a large graphical models lies in the details of the highly developed inference algorithms, and also in finding useful approximations for very large systems. (For example, large graphs can be simplified by marginalizing over whole subgraphs.)

Generative models (Jordan, 2002) are closely related to graphical models, but attempt to “hide” the graph (and secret sauce) behind the user- friendly interface of a so-called probabilistic programming language (Probabilistic programming language. (n.d.), 2016). A generative model is a “for- ward” or hierarchical Bayes description of a process, including its Bayes priors. For the 2030 Census, a generative model might draw N, the number of people in a particular census block from a prior based on the number of housing units, the results of the 2020 Census, and other information. Now, for every person 1...N, the model generates, with specified prior probabilities, the administrative records that a person is likely to “throw off” tax returns, school enrollments, etc. The model continues down to the level of individual fields in these records (first and last names with neighborhood-specific priors, addresses with priors compatible with the street names in the census block, and so forth). A model for the misspelling or mistranscribing of names is included.

A model is sufficiently complete when it can be run in the forward ("generative") direction and, drawing from the specified priors, produce a completely synthetic census block including all the relevant fields in all expected administrative records. Then it can be run again (with different random draws) and produce another possible synthetic result for the same census block.

So what good does this do? The power of inference on graphical models is what makes them valuable for the proposed census enumeration process based on linking massive amounts of data. Given the data corresponding to actual administrative records the model can be solved probabilistically all the way up the Bayes hierarchy. The output can be, for example, the most probable list of true names of census block residents, and/or a probability distribution for the value of N, the number of residents in the block. Because the model includes priors on the forward generation of record fields (e.g., true name "William" can generate "Bill" or "Willy" or "W." on a form, or for that matter anything else close to "William" in Hamming distance), it will automatically merge conflicting records in the most probable way, without any requirement for exact matches.

So-called "probabilistic programming languages," (Probabilistic programming language. (n.d.), 2016)) allow the specification of generative models in a user-friendly computer language. Venture (MIT), Figaro (Charles River Analytics), and BLOG (Berkeley) are relatively mature languages, but there are also many others. Google and Microsoft both support platforms. It seems likely that, for use in a 2030 time frame, the Census Bureau would develop its own language, capable of interfacing to multiple inference engines (e.g., Markov Chain Monte Carlo, forward-backward propagation, etc.).

6.3 Research to Prepare for 2030

Considerable research, testing and experimentation will be needed to move to a census that starts with an “in-office” enumeration based on administrative records. Now is the time to define and begin the research program for the 2030 Census. The Census Bureau can use the data collection efforts already underway for the 2020 Census to undertake research on:

- how much of the population is covered in IRS and SSA records,
- how many of the census short-form variables can be collected using those records,
- which populations and variables are missing from the records, and
- alternative data sources and “in-field” methods to be used to complete an accurate enumeration.

JASON recommends that this process begin as soon as possible in order that the Census Bureau be in a position to make a decision in five years about whether to move forward with the “in-office” enumeration strategy proposed in this report for 2030.

An exemplar research and testing agenda for the 2030 Census should include comparing results that would have been obtained in an “in-office” enumeration based on 2010 Census data, IRS and SSA administrative records to the past and future 2020 Census experiments. This should include identifying geographic areas and subpopulations that are difficult to enumerate or missing from the linked data (i.e., administrative records and census 2010 data). The Census Bureau should conduct a mini-version of the 2030 in a limited set of geographic areas in parallel with 2020 dress rehearsal and then again with the actual 2020 Census.

The Census Bureau should develop a set of experiments to test alternative ways of reaching missing populations and filling in imperfect data (e.g. sensor counts, Citizen Enumeration). This should include conducting early research on approaches filling in data gaps for race and ethnicity enumeration to assess magnitude of this challenge.

The Census Bureau should develop a continuous and robust research program in data linkage and imputation across all types of data, from government administrative sources to emerging commercial data. It was outside the scope of this study for JASON to explore the potential limitations of the Census Bureau's computational environment to support this research. Nonetheless, JASON encourages the Census Bureau to ensure they have or can acquire the adequate hardware and software solutions to support this research and ultimately support the corresponding operations for the 2030 Census.

Finally, the 2030 strategy proposed here presents a novel way to approach the population enumeration. Past methods of determining the accuracy and coverage for the census may be challenged. The Census Bureau will need to begin research on a post-enumeration strategy consistent with this new enumeration process.

6.4 Legal and Public Perception

This section closes with some brief statements about how the proposed paradigm shift in the conduct of the census for 2030 fits with the legislative framework surrounding the Census Bureau's operations and public perception. Both topics are important to the public's future trust in the Census Bureau.

The "in-office" enumeration does not appear to violate the Census Bureau's current guiding legislation. Sampling has not been proposed. The linking of massive amounts of data, including administrative records, is a form of imputation. This is consistent with what the Census Bureau is al- ready planning to do with administrative records for 2020.

Currently the residency rules, most notably "where you live and sleep most of the time" or in some cases "where you slept on April 1st," align with the traditional census process of enumerating people within housing units. Some thought and potential restatement of these rules may be needed under the proposed 2030 strategy. The post-enumeration evaluation method may change under the new process. This will need to be vetted in the research community and with Congress.

On the topic of public perception, currently people trust the Census Bureau even without fully understanding its operations. Unless introduced to the pubic carefully, the move to an "in-office" enumeration could feel like government surveillance. Also, the Census Bureau will need to consider whether or how to give everyone an opportunity to feel they have been counted and even validate their data. This study scope did not go into technical approaches that might be implemented to meet that need.

7 NEW AND RENEWED STATISTICAL PRODUCTS

The "in-office" enumeration process proposed for 2030 opens opportunities for innovation in other census products such as the American Community Survey (ACS) and the development of a National Address File. This census enumeration approach could also provide a pathway to conducting a rolling census and even a national registry. Both options could stabilize the Census Bureau's operations and costs associated with the census enumeration over each decade.

7.1 The American Community Survey

JASON considered implications of the proposed census 2030 strategy on the American Community Survey (ACS). The ACS replaced the census long- form in 2000. It also suffers from declining response rates and costly non-response follow-up (NRFU) activities. The ACS is under constant scrutiny by Congress with questions about its utility (see Figure 7.1).

Figure 7.1: American Community Survey controversy.

What seems clear is that many of the ACS questions have known answers in administrative records, e.g., type of housing unit, property value, age of structure, and household income. The Census Bureau's Center for Administrative Records Research and Applications (CARRA) has begun re- search to validate this claim (O'Hara et al. (2016), Keller et al. (2016), and Ruggles (2015)). The proposed "in-office" approach for the decadal census has the potential to directly establish much of this underlying data. This creates the opportunity to re-think the ACS, perhaps shifting its focus toward information not available in other data sources such as beliefs and measures of subjective well-being that require survey elicitation.

7.2 National Address File

The Census Bureau invests considerable resources to construct the Master Address File (MAF). Unfortunately, the MAF is protected under Title 13 and cannot be shared with outside parties despite its potential value as a public good (Craig, 2006; U.S. Supreme Court, 1982). If the Census Bureau were to adopt JASON's suggestion of relying mainly on administrative records for the census enumeration, it may not be necessary to construct a MAF to drive the census enumeration process. This would open the opportunity for the existence of a MAF-like file, a National Address File, outside of Title 13. The Census Bureau could use the National Address File to support the "in-office" enumeration process as a linkable source of data to help geo-locate individuals. However, if it is not the central survey frame for the overall enumeration, simply one of many data sources, it could well remain exempt from Title 13 protection.

JASON suggests that the Census Bureau may be able to create public- good value by working with the U.S. Postal Service (USPS), Department of Transportation (DOT), local governments, and private-sector firms (e.g., Google, Uber, Amazon) to construct a continuously maintained National Ad- dress File that could be publicly shared. Similar to the MAF, this would be an address list only, with no person identifying information attached. JASON also notes that a continuously maintained address file could be integrated with a continuously maintained population register, forming the basis for a rolling census that would be verified every decade to satisfy constitutional requirements (see Section 7.3).

JASON did not focus on other Census Bureau activities outside of the decennial census. However, JASON is aware that the MAF forms the basis for many national household-level surveys that the Census Bureau conducts for the U.S. government. A National Address File, properly constructed, could replace the MAF as the sampling frame for these surveys. This could open some entrepreneurial opportunities for the private sector in the conduct of these national surveys.

7.3 Rolling Census - Beyond 2030

Consideration should be given to moving the U.S. toward using a rolling census, i.e., one that is updated continually and verified for the decadal census. Benefits could include greater accuracy, by using increasingly complete in- formation from commercial and governmental sources, and stabilizing costs by eliminating the need to start afresh every 10 years, as well as the required surge in funding during the census year. In considering how a rolling census might proceed, there are two options, (1) based on enumerating housing units and then populating them, along with accounting for special populations, e.g., homeless estimates; and (2) based on directly enumerating individuals.

7.3.1 Rolling census based on housing units

First, a rolling census parallel to current Census Bureau operations of creating a census through the enumeration of housing units (the MAF) and then populating them is discussed. Rather than creating and then updating a MAF starting two years before every census, it a "master dwelling file" maintained continuously by several federal agencies based mostly on state and local government administrative records could be developed. In this context, a dwelling is a MAF housing unit with the information about who lives in the unit attached. Assessors in the nation's 3,143 counties (including county equivalents in Louisiana and Alaska) have the greatest interest in locating every taxable structure in their domains. They also have the advantage of being nearby. In addition to the Census Bureau, Housing and Urban Development (HUD), the USPS, and the Federal Emergency Management Agency (FEMA) have strong interests in knowing the location and composition of every dwelling. The federal agencies could cooperate with each other and with state and county governments to maintain a master dwelling file based on county tax records and other local administrative records. Following this approach, the work required to identify vacant dwellings could be significantly reduced by using Internal Revenue System (IRS) and county tax records, some of which tax second homes differently than primary residences. Taking a census would be confined to verifying the list using an efficient but statistically valid procedure.

Methodology for counting special groups like the homeless would still need to be developed. Approaches that can be implemented on an on-going basis, versus once a decade, could be adopted. For example, every year on a single night in January HUD sponsors a count of homeless persons in emergency shelters, transitional housing, and Safe Havens. On odd-numbered years, unsheltered persons are added to the count. Most recently, Henry et al. (2015) reported 564,708 homeless one night in January 2015. Of these, 64% were individuals; the remainder were families with children. Overall, homelessness declined by 2% between 2014 and 2015 and by 11% since 2007. People conducting the survey try to determine race, gender, and age. Consistent with current Census Bureau practice, there is frequently no mention of names and an acknowledgement that some homeless do not want to be found.

7.3.2 Rolling census based directly on individuals

If the Census Bureau moved toward a strategy of directly enumerating individuals, this could lead nicely to a rolling census based directly on individuals. There are many reasons for maintaining a registry of persons in the U.S., but it has not happened owing to concerns about requirements for mandatory personal identification cards. These, however, seem to be less threatening in view of the Real ID Act passed by the U.S. Congress in 2005 (Public Law, 2005) enforcing stricter requirements for IDs, e.g., those acceptable to the Transportation Security Administration (TSA). Only five states and American Samoa have not complied and face having their driver's licenses rejected by TSA after January 22, 2018 if they do not upgrade their licenses.

Given the records maintained by IRS, states, counties, and companies, compiling a personal registry with redundant crosschecks should be feasible. Automatic updating should also be possible as the source records are modified reflecting new taxpayers, deaths, and moves. With the existence of such a list, a decadal census would involve seeking verification from each individual without building the database from scratch.

Selective Service System (SSS) registration is a process with goals similar to those of census that should be included in the records accessed by the Census Bureau and which may have some lessons applicable to updating a rolling census of individuals or registry of persons. Males in the U.S. between the ages of 18 and 26 are required to register with SSS and to notify SSS within 10 days of address changes. The only exceptions are immigrants on student visas and diplomats. Most recently, compliance for the cohort was estimated at 88%. Records are updated daily and include men suspected to be in violation of the Selective Service Act. Past records for men born before 1960 are kept by the National Archives and Records Administration. Finally, note that that the Census Bureau used SSS records in their 2000 experiment with administrative records (Bauder and Judson (2003) and see Section 5.2).

The high compliance rate is attributed to: (1) requirements for SSS registration to obtain a driver's license or identification card in 40 states; (2) availability of online registration, (3) volunteer SS registrars; (4) availability of registration forms in post offices; (5) outreach initiatives, e.g., via the Harlem Globetrotters, minor league baseball teams, and the U.S. Hispanic Leadership Institute (Director of the Selective Service System, 2015). In addition, Alaska requires SS registration to qualify for payouts from the Alaska Permanent Fund established to share oil revenues. Data is also obtained from the USPS, prisons, high schools, the Immigration Services, and jobs programs. Of all 2014 SS electronic registrations, 43% were from driver's li- cense registrations, 24% came from applications to Department of Education Pell Grants, and 19% were from the online site.

A rolling census and population registry are not new ideas. Other countries have considered similar approaches. This is discussed in the next sections.

7.4 The Evolving Census in Europe

Facing pressures similar to those in the U.S., some European countries are modifying how they enumerate their populations. After 2000 seven of them adopted different approaches for their 2010 Census. In 2010 a traditional census was retained by 20 countries, mostly in eastern Europe plus the United Kingdom (U.K.), Ireland, and Portugal (see Figure 7.2). The Scandinavian countries and Austria use population registers, France conducts a rolling census, and the others combine registers with other sources. Some of the northern countries, (e.g., Denmark and Finland), do not conduct field surveys and base their censuses on statistical information extracted from registers. Once the register is established, individuals are not bothered, and costs are minimal, but it depends on strong public support for the system. Some countries combine registers with other surveys that provide verification of the data. In 2011 the Netherlands and Slovenia followed this approach to conduct a “virtual census”.

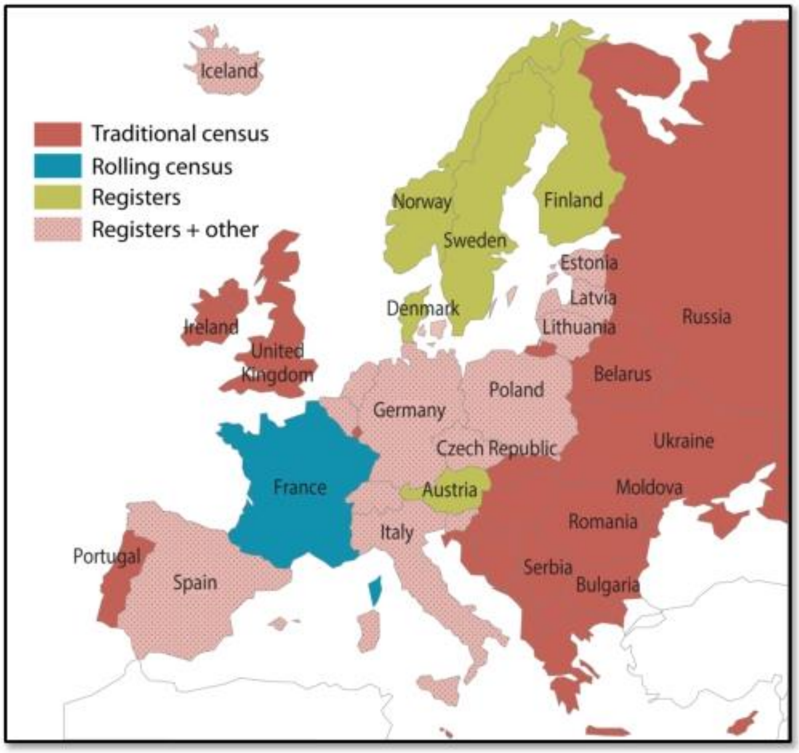


Figure 7.2: Map of census methods in European countries. Source: Valente (2010).

This Section concludes with a short review of the progress being made by France and the U.K..

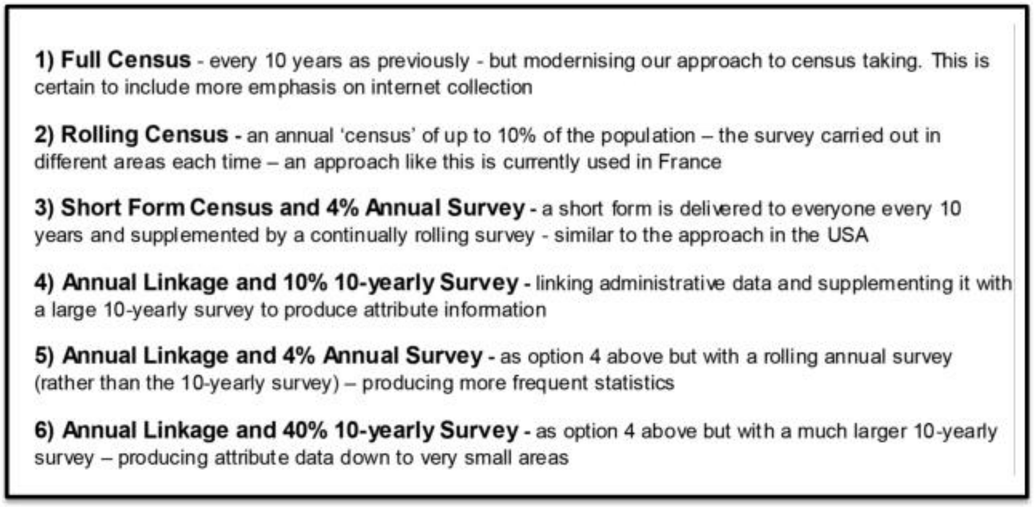
7.4.1 France

To decrease latency of the data and to remove the burden of annual counts, in 2004 France introduced a rolling census. France is divided into 36,680 municipalities, only 980 of which have 10,000 or more inhabitants. Municipalities with fewer than 10,000 inhabitants were divided into five groups with a full census in one group every year. These counts are conducted by the municipalities because they can do it more cheaply and are better positioned to find the "hard-to-count". All municipalities with more than 10,000 in- habitants enumerate 8% of their households every year. Overall, 70% of the population is covered every five years, and the census results are based on five-year moving averages, updated yearly. The new approach is believed to reduce the overall errors in the traditional census in part because the results from a tradition approach tend to be out-of-date when they are published with an error that continues to grow until the next census.

Examining France's experience after the first decade of the rolling census, Durr and Clanch´e (2013) note that costs have stabilized across the years but have increased proportionally to the population as has the data processing work load. France is now considering options that could actually result in cost reductions, even in light of population increases, such as responding via the internet.

7.5 United Kingdom

Also driven by rising costs and the difficulty of surveying a mobile population, the U.K. has been analyzing approaches in other countries as part of a program to update their census procedures. Recently, they created a list with six census enumeration options (see Figure 7.3), the third of which is similar to the U.S.'s introduction of the American Community Survey (ACS) in 2000. One option, #2, is a rolling census. Three of the options use administrative records to estimate population size and differ in how often to use a long form to determine attributes of the population. The three choices are: 10% of the population every 10 years; a lesser percentage sampling different subsets every year, like the ACS in the U.S., and sampling 40% of the population every 10 years.



1) Full Census - every 10 years as previously - but modernising our approach to census taking. This is certain to include more emphasis on internet collection

2) Rolling Census - an annual 'census' of up to 10% of the population – the survey carried out in different areas each time – an approach like this is currently used in France

3) Short Form Census and 4% Annual Survey - a short form is delivered to everyone every 10 years and supplemented by a continually rolling survey - similar to the approach in the USA

4) Annual Linkage and 10% 10-yearly Survey - linking administrative data and supplementing it with a large 10-yearly survey to produce attribute information

5) Annual Linkage and 4% Annual Survey - as option 4 above but with a rolling annual survey (rather than the 10-yearly survey) – producing more frequent statistics

6) Annual Linkage and 40% 10-yearly Survey - as option 4 above but with a much larger 10-yearly survey – producing attribute data down to very small areas

Figure 7.3: Six options considered by the U.K. for future censuses. Source: Brown (2013).

For the options in Figure 7.3, administrative records would be used in conjunction with a national address registry to estimate population size and location. Once the data are vetted on the secure side, individual identities would be removed and the data sent to the U.K. Office for National Statistics (ONS), where they would be combined and crosschecked with anonymized data from the address register and other household surveys. These data would then be the basis for enumeration and estimation of the U.K.'s population characteristics.

8 THE TRANSCRIPTION DILEMMA

U.S. decennial census records are confidential for 72 years to protect respondents' privacy. Thus, the most recent publicly available census records are from the 1940 Census, released April 2, 2012. From these records, person- level data linkages with other stored records yield information of wide interest and utility. For the 2000 Census and later, linkage keys are available, making all census data on and after 2000 quite useful through person-level links while still remaining anonymous.

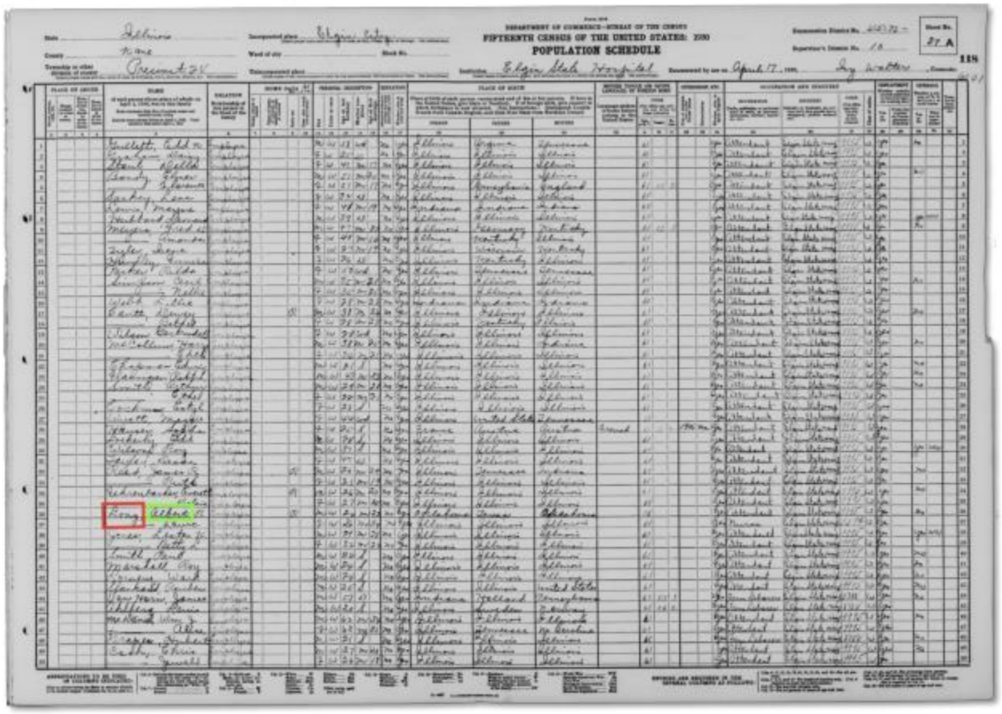
The Census Bureau would now like to make this person-level connection available for the five censuses from 1950 through 1990. Maintaining anonymity is seen as the biggest obstacle because the names are all hand- written, and need to be read and linked to names in other data without the actual name connected with the census data being revealed in the process. The most efficient and accurate way to do this would be to use human readers as with pre-1950 censuses. However, the Census Bureau feels this is not possible due to the need to protect the names.

There is a census pilot program, the 1990 Name Recovery Pilot, to demonstrate how to solve this problem, first with the 1990 Census, and then working backwards through the four remaining decades (Cronkite and Alexander, 2016). The method proposed and currently being tested in the pilot program is to employ computer driven Optical Character Recognition (OCR) for the hard part, the handwritten names, using quarantined machines and hard drives that will later be destroyed.

It is well known, however, that crowdsourcing humans to perform character recognition can process images quickly and with higher accuracy for recognizing the images than with computer-driven OCR. JASON proposes a scheme for the Census Bureau whereby it should be possible to "maintain confidentiality" and still use humans to recognize the characters, in this case the written names. This will now be described.

8.1 Scanning and Text Recognition

The Census was able to cooperate with Ancestry.com to have census and other hand-written and publically available records from before 1940 sent to India to have them manually transcribed into digital format. Due to the 72-year rule on confidentiality, the direct application of that approach was not possible for transcribing the 1950 through 1990 censuses. JASON proposes a modification to the approach that will allow manual recognition and transcription of these records with no loss of privacy, and at the same time allow increasing degrees of automatic recognition and transcription.



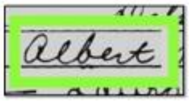
DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
FIFTEENTH CENSUS OF THE UNITED STATES: 1930
POPULATION SCHEDULE

State Illinois
County Kane
Incorporated place Elgin City
Institution Elgin State Hospital
Enumerated by me on April 17, 1930

Figure 8.1: A page of the 1930 Census, with two of hundreds of shards marked.

The key insight is that even hand-written manuscripts can be segmented into “ shards,” each of which contains a single word, using current state of the art image processing algorithms. These shards can be uniquely numbered for later reassembly and placed into a large pool. Humans can be presented with a shard and asked to enter the text that it represents.

Consider an image such as the page from the 1930 Census as shown in Figure 8.1. The image would be segmented into shards and the shards placed into a pool; the larger the pool the higher the degree of confidentiality. The system would operate by randomly selecting a shard from the pool and then presenting it to a human for recognition and transcription. For example, human presented with a shard,



Albert

might be asked, “What does this shard say?” or a confirmation, “Does this shard say ‘Albert’?” The process would repeat until a satisfactory quality level of the transcription is achieved. This would be a tunable parameter, q = 1, 2,.. such that after enough humans agree on the transcription it would be accepted as correct. Note, that if q = 1 then the cost is the same as having the words transcribed directly.

A simple process with two phases is proposed. Phase 1 assigns meaning to untranscribed shards through optical recognition. Phase 2 confirms groups of transcribed shards and rejects erroneous transcriptions. Additional confirmation phases could be imagined.

In Phase 1, there is an initial pool, called pool$_1$, of unknown shards that are created by segmenting algorithms from scanned documents. The algorithm repeats until the pool is empty, or contains only shards that have failed to be recognized after several attempts. For each human recognizer, the system randomly draws a shard from the pool and asks the recognizer to enter the text represented by the shard. This shard is then moved to a pool$_2$ to be processed in the second phase. The time to complete this task is approximately n, where n is the number of shards in the pool. It may increase slightly since Phase 2 may return difficult or erroneously transcribed shards to pool$_1$.

In Phase 2, there is the opportunity to perform quality control and also speed the process. Assume a parameter k, which is the number of shards that one is willing to present for confirmation to the human recognizer. Once k shards with the same proposed text have been accumulated in pool$_2$, present them to the human recognizer, for example, "Check all of the boxes for shards that say 'Albert'." shards that are checked are considered confirmed and are removed from pool$_2$, those that are left unchecked are returned to pool$_1$ with their proposed text cleared but with a count of the number attempts incremented. Once a shard has failed some tunable number of times, it will be considered untranscribable and will be handled manually. Since the number of unique words win, there will be many groups of size k, and so the time for Phase 2 will be approximately n/k.

In 2005, Amazon.com launched a service called Mechanical Turk—Artificial Intelligence, that provides a crowdsourcing means for humans to work on small tasks for pay. Such a system could be employed by the Census Bureau in order to transcribe the shards. This work could be done by people in the U.S., or by people in other countries, it does not matter—the system provides confidentiality inherently.

A number of augmentations could make this process more efficient. For example, one could collect like shards and ask to select those that conform to a given transcription; one could then present 10 shards and ask the human, "Select the shards that say ‘Albert’.” If machine recognition advances sufficiently, then suggestions for confirmation could be presented by machine learning algorithms.

Since every shard is randomly chosen from a large pool, there should not be concern about revealing any information using this process, as a simple analysis shows. Suppose for sake of argument that full names are composed of three elements, given, middle and family, and that there are records for $n = 3 \times 10^8$ individuals in the pool; if every name has an equal chance of being chosen, then the chance of a recognizer seeing all three names in sequence is

$$\frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} \approx 3.7037 \times 10^{-26},$$

which for all practical purposes is zero.

8.2 Automatic Transcription

The automatic recognition of hand-written script has been an active area of research for decades. It is in use by the U.S. Postal Service (USPS), and by postal services in many countries. The real question is one of accuracy. It should also be noted that the style of orthography has changed over generations as it has gone from a skill taught in school, and where a person took pride in their orthography to in many cases a skill only used in writing checks and addressing envelopes.

The Postal Services in the U.S., Australia and the United Kingdom use the HandWritten Address Interpretation (HWAI) system originally developed at the State University of New York, Buffalo. First deployed in 1996, it reads the handwritten ZIP code in order to sort the mail. It is currently used at all postal sorting centers with 95% of mail being sorted automatically. The area of automatic recognition of handwritten text, in particular digits, remains an active area of research. The National Institute of Standards and

Technologies (NIST) and the USPS both maintain training and test data sets of written digits, and many other standard sets are available as well. Convolutional neural networks have demonstrated an error rate of 0.23% (Cire, san et al., 2012). In recently years, significant effort has gone into recognition of scripts based on non-Roman alphabets (Rehman and Saba, 2014).

The interest in understanding hand-written script has increased with the use of tablets and other hand-held devices, referred to as Interactive Machine Recognition (IMR). It is common to find translation from hand- written to machine readable text on these devices, either as an intrinsic feature or as an available application. The algorithm class of choice seems to be neural networks, though other approaches have been described in the research Graves et al. (2009) and commercial Guzik et al. (2000) literature.

JASON will not attempt to say which approach is best, since it has been and remains an active area of research, but it is important to emphasize that it is a technology that is in use. When coupled with human verification, such as in the approach described in the previous section it can lead to accurate results.

Returning to the JASON scheme, even less than perfect machine recognition can be of great benefit. For example, given a set of confirmed shards from Phase 2, a machine could scan through $pool_1$ and do transcriptions and place candidates into $pool_2$. If the machine recognition is correct, they will be confirmed in Phase 2, if they are incorrect then the shards will be returned to pool. A near-perfect machine recognizer would obviate the need for humans to participate in Phase 1.

9 FINDINGS AND RECOMMENDATIONS

This report provides a starting point from which the Census Bureau can build their 2030 Census strategy. It includes a heavy emphasis on the use of government administrative records. The process is a re-conceptualization of the census enumeration process whereby a direct enumeration of people and their associated geolocations would be conducted versus an initial census of housing units followed by counting the individuals "living and sleeping most of the time" within the housing units. Figure 9.1 summarizes the 2030 vision.

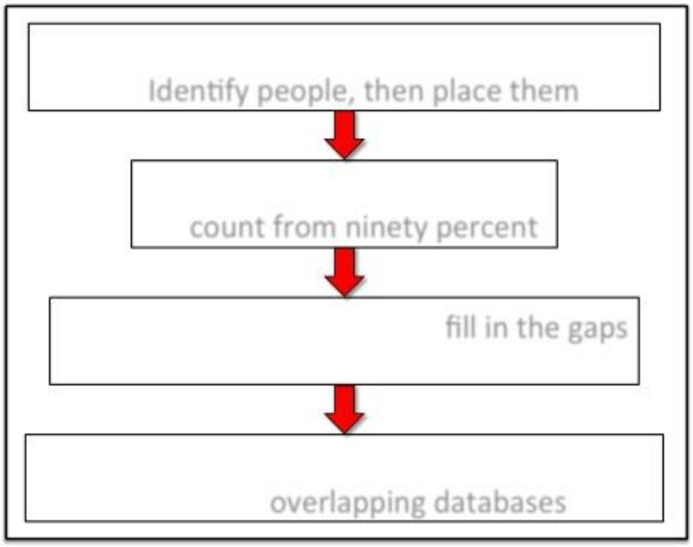


Figure 9.1: Proposed strategy for the 2030 Census process.

This study started by reviewing the Census Bureau's plans for the 2020 Census, with some focus on cost and coverage (accuracy). JASON's proposed 2030 process may provide the opportunity for the Census Bureau to predict, control, and even reduce their fixed costs. This would include maximizing the "in-office" enumeration, or administrative data usage, and optimizing additional operations needed to "fill in" the gaps. These costs versus coverage trade-offs may be knowable a-prior, thus helping to reduce the overall variable costs. Figure 9.2 illustratively compares the 2020 situation (see Figure 3-5) with the 2030 Census proposal.

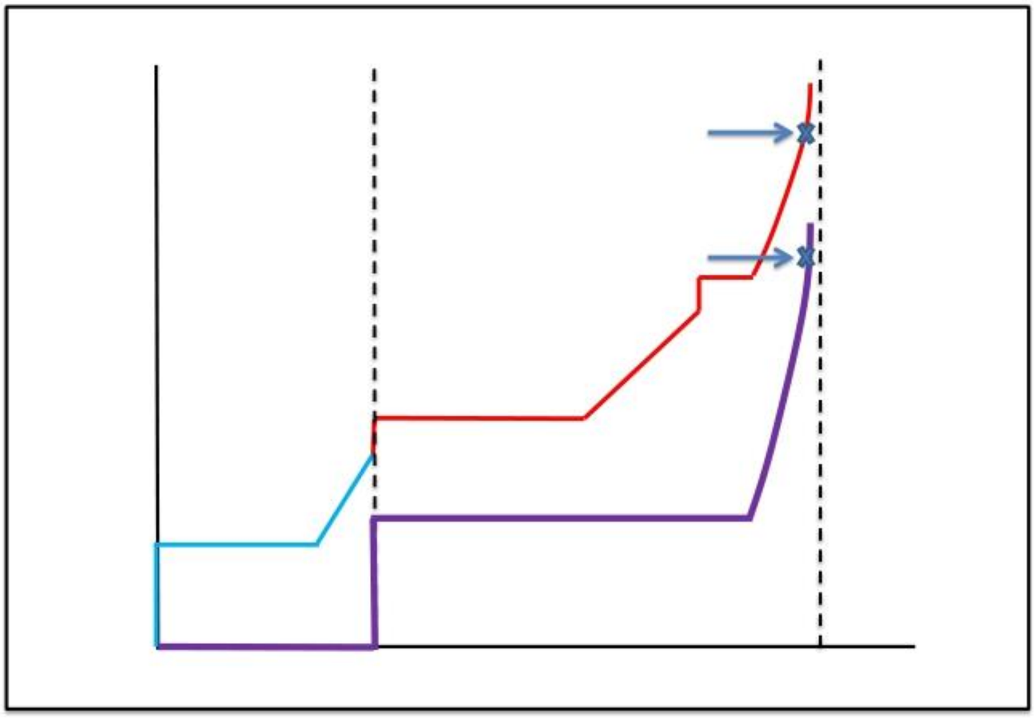


Figure 9.2: Illustrative depiction of cost versus coverage for 2020 Census operations versus proposed 2030 Census operations.

9.1 Findings

JASON offers four key findings to support the value in reformulating the census enumeration process for 2030.

1. The U.S. decennial census data enable the study of American society and its evolution.
2. The U.S. decennial census process is expensive. Significant cost drivers are:

(a) the rising costs per housing unit (approximately $94 per housing unit in 2010 constant dollars);

(b) the low self-response rates (<65%); and

(c) the labor-intensive Master Address File (MAF) generation and Non-Response Follow-Up (NRFU) activities.

3. Census process is imperfect.

4. A large majority of the information collected for the census already exists in government administrative records.

JASON offers three findings that support focusing the enumeration around directly counting individuals versus housing units.

1. The place of residence may not be the most persistent means of contact in the future or the most effective way to find the people.

2. The population enumeration coverage through government administrative records is growing, as well as availability and usefulness of third- party data sources.

3. The data and technologies available today for mapping addresses create new opportunities for building a public National Address File.

9.2 Recommendations

JASON offers sixteen specific recommended actions to be taken by the Census Bureau.

Regarding moving to an “in-office” enumeration:

1. Re-conceptualize the census by organizing it around people rather than housing units.

   (a) Identify people, and then place them.

   (b) Start the count from “ninety percent” by using “in-office” enumeration.

   (c) Use field activities to fill in the gaps and validate what is known.

2. Start enumeration with administrative records from IRS, SSA and past Census data to construct an “in-office” census that is as complete as possible.

3. Develop a multifaceted strategy to find people who do not appear in the “in-office” enumeration.

4. Use research and near-term experimentation to explore who will not be enumerated with this approach, what data fields will be lacking, and strategies for gap filling.

5. Continue and expand efforts to acquire data from other agencies, which will be critical to the success of “in-office” enumeration.

Regarding trade-offs:

6. Create a set of metrics and criteria by which an “in-office” approach can be evaluated against traditional “self-response plus NRFU” approaches.

7. Examine the utility and cost of expanding the use of administrative records to be a rolling census that would provide an up-to-date population to satisfy enumeration requirements between decadal censuses.

8. Develop a list of options detailing the estimated cost of the 2030 Census as a function of the “accuracy and coverage” desired, which could be used by the Census Bureau and Congress to decide “how good is good enough.”

Regarding research and testing:

9. Develop and start a set of experiments now to test the “in-office” enumeration concept.

   (a) Utilize massive administrative data linkage tests to confirm percentage of population enumerable by this strategy.

   (b) Confirm ability to identify subpopulations that will be consistently missed.

10. Explore and test alternative approaches to reach the remaining hard-to-reach populations (gaps).

   (a) These could include remote and street-level sensing, crowd-sourced citizen Enumeration, novel data such as state Medicaid records for low-income population, and partnership with HUD to count the homeless.

11. Continue to plan for and test enumeration strategies in the face of natural disasters, terrorist attacks, and temporary dislocation of large numbers of people.

   (a) Create partnerships with FEMA to locate (place) temporarily displaced individuals and trusted civil servants (e.g., firefighters).

Regarding the American Community Survey (ACS):

12. Reconsider ACS and how much data the Census Bureau elicits from different people at different times, keeping in mind replication of administrative data and high cost of asking the first question.

    (a) Use ACS to follow trends relevant to future Census Bureau data collection, such as how long people maintain their email addresses and cell phones versus their residences.

    (b) Use ACS to monitor public trust (a subjective measure).

    (c) Commission a study on what are the high-value data per dollar.

Regarding issues beyond census 2030:

13. Create a public National Address File outside of Title 13.

14. Be alert for new public and private sector data sources that may become available.

15. Develop a pilot "rolling census" project relying mainly on administrative data.

16. Explore new methods for name recognition and OCR to digitize 1950- 1990 censuses.

References

Airsage (2016). airsage. http://www.airsage.com/.

Alvey, W. and F. Scheuren (1982). Background for an administrative record census. In Proceedings of the Social Statistics Section. American Statistical Association.

Anderson, M. (2015). Technology device ownership: 2015. http://www.pewinternet.org/2015/10/29/technology-device-ownership-2015/.

Anderson, M. and A. Perrin (2015). 15% of Americans don't use the internet. Who are they? http://www.pewresearch.org/fact-tank/2015/07/28/15-of-americans-dont-use-the-internet-who-are-they/.

Audubon Society (2015). Christmas bird count. http://www.audubon.org/conservation/science/christmas-bird-count.

Bauder, M. and D. Judson (2003). Administrative records experiment in 2000 (AREX 2000) household level analysis. Technical report, U.S. Census Bureau.

Blowoutcards.com (2016). How long have you had the same cell phone number? http://www.blowoutcards.com/forums/off-topic/885281-how-long-have-you-had-same-cell-number.html.

Brown, J. (2013). Beyond 2011: Options explained 2. Technical report, Office of National Statistics, United Kingdom.

Butt, N., E. Slade, J. Thompson, Y. Malhi, and T. Riutta (2013). Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecological Applications 23 (4), 936–943.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014). Where is the land of opportunity? The geography of intergenerational mobility in the united states. Technical report, National Bureau of Economic Research.

Chow, T. E. (2013). We know who you are and we know where you live: A research agenda for web demographics. In Crowdsourcing Geographic Knowledge, pp. 265–285. Springer.

Cire, san, D., U. Meier, and J. Schmidhuber (2012). Multi-column deep neural networks for image classification. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3642–3649. IEEE.

Crackberry.com (2016). How long have you had the same cell phone number? http://forums.crackberry.com/general-blackberry-discussion-f2/how-long-have-you-had-your-cellphone-number-465097/.

Craig, W. (2006). Master address file for state and local government. In Proceedings of the URISA 2006 Annual Conference. Vancouver, British Columbia, Canada.

Crain, R., C. Cooper, and J. L. Dickinson (2014). Citizen science: a tool for integrating studies of human and natural systems. Annual Review of Environment and Resources 39 (1), 641.

Cronkite, D. and T. Alexander (2016, May 4). 1990 census names recovery project. https://www.census.gov/fedcasic/fc2016/ppt/2_8_Recover.pdf.

Director of the Selective Service System (2015). Annual report to the congress of the United States – fiscal year 2015. Technical report, Selective Service System.

Durr, J.-M. and F. Clanch´e (2013). The French rolling census: a decade of experience. In 59th ISI World Statistics Congress. International Statistical Institute.

Edgar, G. J., A. E. Bates, T. J. Bird, A. H. Jones, S. Kininmonth, R. D. Stuart-Smith and T. J. Webb (2016). New approaches to marine conservation through the scaling up of ecological data. Marine Science 8.

Erdman, C. and N. Bates (2014). The US Census Bureau mail return rate challenge: Crowdsourcing to develop a hard-to-count score. In Statistics, #2014-08. U.S. Census Bureau.

File, T. and C. Ryan (2014). Computer and internet use in the United States: 2013. In American Community Survey Reports, ACS-28. U.S. Census Bureau.

GAO (2011, April). 2010 CENSUS Preliminary Lessons Learned Highlight the Need for Fundamental Reforms. Technical Report GAO-11-496T, United States Government Accountability Office.

GAO (2016, June). 2020 Census: Census Bureau needs to improve its life-cycle cost estimating process. Technical Report GAO-16-628, United States Government Accountability Office.

Gillett, D. J., D. J. Pondella II, J. Freiwald, K. C. Schiff, J. E. Caselle, C. Shuman, and S. B. Weisberg (2012). Comparing volunteer and professionally collected monitoring data from the rocky subtidal reefs of southern California, USA. Environmental monitoring and assessment 184 (5), 3239– 3257.

Glick, P. (1984). American household structure in transition. Family planning perspectives 16, 205–211.

Graphical model. (n.d.) (2016). In Wikipedia retrieved August 18th 2016. https://en.wikipedia.org/wiki/Graphical_model.

Graves, A., M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidt- Huber (2009). A novel connectionist system for unconstrained handwriting recognition.

IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5), 855–868.

Guzik, K. J., A. P. Huff, and R. Nag (2000, April 25). Handwriting recognition method and apparatus having multiple selectable dictionaries. US Patent 6,055,333.

Hansen, K. A. (1998). Seasonality of moves and duration of residence. Number 66. U.S. Department of Commerce, Economics and Statistics Ad- ministration, Census Bureau.

Henry, M., A. Shivji, T. deSousa, R. Cohen, and Abt Associates Inc. (2015). The 2015 Annual Homeless Assessment Report (AHAR) to Congress. Technical report, U.S. Department of Housing and Urban Development.

Hobbs, F. and N. Stoops (2002, November). Demographic trends in the 20th century. In Census 2000 Special Reports, CENSR-4. U.S. Census Bureau.

HUD (2004). How many second homes are there? https://www.huduser.gov/periodicals/ushmc/spring2004/article_USHMC-04Q1.pdf.

Infoplease (2004). U.S. households by size, 17902006. http://www.infoplease.com/ipa/A0884238.html.

JASON (2015). Respondent validation for non-id processing in the 2020 decennial census. hhttps://fas.org/irp/agency/dod/jason/census. pdf.

Johnson, P. A. and R. E. Sieber (2013). Situating the adoption of vgi by government. In Crowdsourcing geographic knowledge, pp. 65–81. Springer.

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems 14, 841.

Kaggle (2012). U.S. Census return rate challenge. https://www.kaggle.com/c/us-census-challenge.

Keller, S., S. Shipp, M. Orr, D. Higdon, G. Korkmaz, A. Schroeder, E. Molfino, B. Pires, K. Ziemer, and D. Weinberg (2016). Leveraging External Data Sources to Enhance Official Statistics and Products. Technical report, Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech.

Koller, D. and N. Friedman (2009). Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Kotokosta, C. and N. Johnson (2016). Urban phenology: Toward a real-time census of the city. Technical report, Paper presented at the Association for Public Policy Analysis and Management Research Conference, Washington, DC.

Lofquist, D., T. Lugaila, M. O'Connell, and S. Feliz (2012, April). House- holds and families: 2010. In 2010 Census Briefs. U.S. Census Bureau.

Marshall, P. J., C. J. Lintott, and L. N. Fletcher (2015). Ideas for citizen science in astronomy. Annual Review of Astronomy and Astrophysics 53, 247–278.

Morris, D. S., A. Keller, and B. Clark (2015). An approach for using administrative records to reduce contacts in the 2020 census. In Proceedings of the Government Statistics Section.

Mortenson, J. A., J. Cilke, M. Udell, and J. Zytnick (2009). Attaching the left tail: A new profile of income for persons who do not appear on federal income tax returns. In National Tax Association Proceedings.

NOAA (2016). NOAA national weather service cooperative observer pro- gram. http://www.nws.noaa.gov/om/coop/.

O'Hara, A., A. Bee, and J. Mitchell (2016, March). Preliminary research for replacing or supplementing the income question on the American Community Survey with administrative records. In 2015, American Community Survey Research and Evaluation Report Memorandum Series #ACS16-RER-6. U.S. Census Bureau.

OIG (2011, June). Census 2010: Final report to congress. Technical Report OIG-11-030-I, Department of Commerce, Office of the Inspector General.

OMB (2006, February 14). Guidance for Providing and Using Administrative Data for Statistical Purposes. OMB Memorandum M-14-06.

Perrin, A. and M. Duggan (2015). Americans' internet access: 2000-2015. http://www.pewinternet.org/2015/06/26/americans-internet-access-2000-2015/.

Pew Research Center (2015a). Beyond distrust: How Americans view their government. Technical report, Pew Research Center.

Pew Research Center (2015b). Views of government's handling of terror- ism fall to post-9/11 low. http://www.people-press.org/2015/12/15/views-of-governments-handling-of-terrorism-fall-to-post-911-low/.

Probabilistic programming language. (n.d.) (2016). In Wikipedia retrieved august 18th 2016. https://en.wikipedia.org/wiki/Probabilistic_programming_language.

Public Law (2005). Emergency Supplemental Appropriations Act for Defense, the Global War on Terror, and Tsunami Relief, 2005. In Public Law 109-13. U.S. Congress.

Rainie, L. and M. Duggan (2015). Americans' internet access: 2000-2015. hhttp://www.pewinternet.org/files/2016/01/PI_2016.01.14_Privacy-and-Info-Sharing_FINAL.pdf.

Ramzy, A. (2016, April 11). Australia Stops Online Collection of Census Data After Cyberattacks. The New York Times, New York Edition, p. A6.

Rastogi, S., A. OHara, J. Noon, E. A. Zapata, C. Espinoza, L. B. Marshall, T. A. Schellhamer, and J. D. Brown (2012). 2010 Census match study. In 2010 Census Planning Memoranda Series No. 247. U.S. Census Bureau.

Rehman, A. and T. Saba (2014). Neural networks for document image preprocessing: state of the art. Artificial Intelligence Review 42 (2), 253–273.

Richter, F. (2015). Landline-phones are a dying breed. https://www.statista.com/chart/2072/landline-phones-in-the-united-states/.

Ruggles, P. (2015, January). Review of administrative data sources relevant to the american community survey. In ACS Information Memoranda Series Memorandum No.: 2015-03. U.S. Census Bureau.

Ruggles, S. and S. Brower (2003). Measurement of household and family composition in the United States, 1850-2000. Population and development review 29, 73–101.

Smith, A. (2015). U.S. smartphone use in 2015. http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/.

U.S. Census Bureau (2009, December). 2000 Census of population and housing. In History, PCH-R-V1. U.S. Census Bureau.

U.S. Census Bureau (2010). 2010 Census form. https://www.census.gov/schools/pdf/2010form_info.pdf.

U.S. Census Bureau (2012a). Census Bureau releases estimates of undercount and overcount in the 2010 Census. hhttps://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html.

U.S. Census Bureau (2012b). Statistical Abstract of the United States: 2012.

U.S. Census Bureau (2015, November). 2020 Census operational plan: A new design for the 21st century. Prepared by the Decennial Census Management Division, U.S. Census Bureau.

U.S. Supreme Court (1982). Baldrige v. Shapiro, 455 U.S. 345, No. 80- 1436, Argued December 2, 198, Decided February 24. https://supreme. justia.com/cases/federal/us/455/345/.

Valente, P. (2010). Census taking in Europe: how are populations counted in 2010? Population & Societies (467), 1.

Vianna, G. M., M. G. Meekan, T. H. Bornovski, and J. J. Meeuwig (2014). Acoustic telemetry validates a citizen science approach for monitoring sharks on coral reefs. PloS one 9 (4), e95565.

Walker, S., S. Susanna Winder, G. Jackson, and S. Heimel (2010). 2010 Census nonresponse followup operations assessment. In 2010 Census Planning Memoranda Series, No. 190. U.S. Census Bureau.

A APPENDIX: Extreme Scenarios

In addition to future trends discussed in the main report, JASON considered the possibility of more extreme events that might be consequential for the Census Bureau. This section briefly outlines a few such scenarios, with implications for 2030 planning.

Scenarios can provide an alternative way for thinking about the future of complex sociotechnical systems. Scenario development is commonly used in commerce areas, for example, energy system projections, to explore alternative futures. Scenarios are stories of “possible” extreme sociotechnical circumstances. They can be useful platforms for exploring how exogenous drivers might interact and for testing the robustness of future plans.

A.1 Shifts in the Relationship of Government and Citizens

Public attitudes toward government monitoring and general trust in government can shift sharply due to external events. In recent memory, the 9/11 terrorist attacks dramatically shaped the public debate about when and how the U.S. government should monitor citizens and their communications. A striking historical example comes from the United Kingdom (U.K.).

In the late 1930, the U.K. was struggling to emerge from the Great Depression, the mood of the public was sour, and a large fraction of the population was deeply distrustful of government. It would have been difficult to obtain any adequate level of cooperation in any centrally organized government program that required al l citizens to participate.

Into this dismal situation came the gas masks. Somebody in the government, perhaps Neville Chamberlain himself, had the idea and executed it brilliantly. Within a few months, fifty million gas masks were manufactured and distributed as free gifts to the population. Grown-ups got them in grey boxes, children in bright colors. To oversee the distribution, a local network of volunteer Air-raid Wardens was established. The action had two primary purposes, to send a message to Adolf Hitler that Britain was serious, and to send a message to the British population that the government was serious. The psychological effect on the British population was immediate. Everyone knew that they were riding in the same boat, and the gas masks could save their lives. At the same time, the population in France could see that their government was not serious, because they had no gas masks.

In Britain the effect of the gas masks was to create an informal civilian organization for Air-Raid Precautions with close to a hundred percent participation. A year later, when the war began and food was rationed, a similar organization was established to oversee the fair distribution of food, and every member of the population was given an identity card. The gas masks and the identity cards were symbols of trust and cooperation between population and government. As an incidental benefit, the identity cards provided a cheap and accurate basis of information for a national census.

In the U.S. over the next decade, it is possible to envision a range of sudden shifts in public trust or attitudes toward government monitoring. While the Census Bureau's plan should be flexible enough to account for extreme events, the potential for them does not obviously favor the traditional census enumeration approach or an administrative records approach. For example, a sudden fall or rise in trust could radically reduce or increase self-response rates in the traditional census enumeration method. Similarly, a sharp change in attitudes toward government use of citizen data could alter the political landscape for using administrative records.

A.2. Breakdowns in Social or Political Order

Major breakdowns in social or political order have the potential to significantly disrupt a census enumeration. The traditional approach to census enumeration requires human enumerators to visit a large fraction of U.S. residences. Extreme scenarios resulting from localized terrorist attacks, riots, or other breakdowns in social or political order could make it unsafe for enumerators to carry out their jobs. These types of events arguably might have less consequence for an administrative records enumeration, except to the extent that they would create overall risks to the population or dislocations of many residents.

One can imagine other radical scenarios that would dramatically affect administrative records as well as a traditional enumeration, such as a concerted cyber-attack on government computer systems or a political movement that exhorted citizens not to comply with government tax or reporting obligations. Obviously, the implications of such scenarios go far beyond planning for the 2030 Census. However, considering extreme disruptions may be useful as a mechanism to test of the resiliency of the 2030 Census plan.

A.3. Population Dislocations

Natural disasters, environmental disasters and human events all can result in major population dislocations. There are many historical examples. Hurricane Katrina displaced 400,000 Gulf Coast residents, but larger disasters are conceivable. Seismologists frequently note that the Pacific Northwest is due for a large rupture on the Cascadia subduction zone. The 1700 Cascadia earthquake was a magnitude M8.9 event. Such an event in modern times conceivably could displace millions of people for months.

Other extreme weather events such as hurricanes, or flooding of coastal zones from a rapid melting of ice sheets, similarly could lead to large dis- placements. If such a displacement occurred in a time window before the decennial census count, it could seriously disrupt a traditional enumeration. A rolling census based on administrative records arguably would be more robust, although records might not accurately capture individual locations during the period of dislocation.

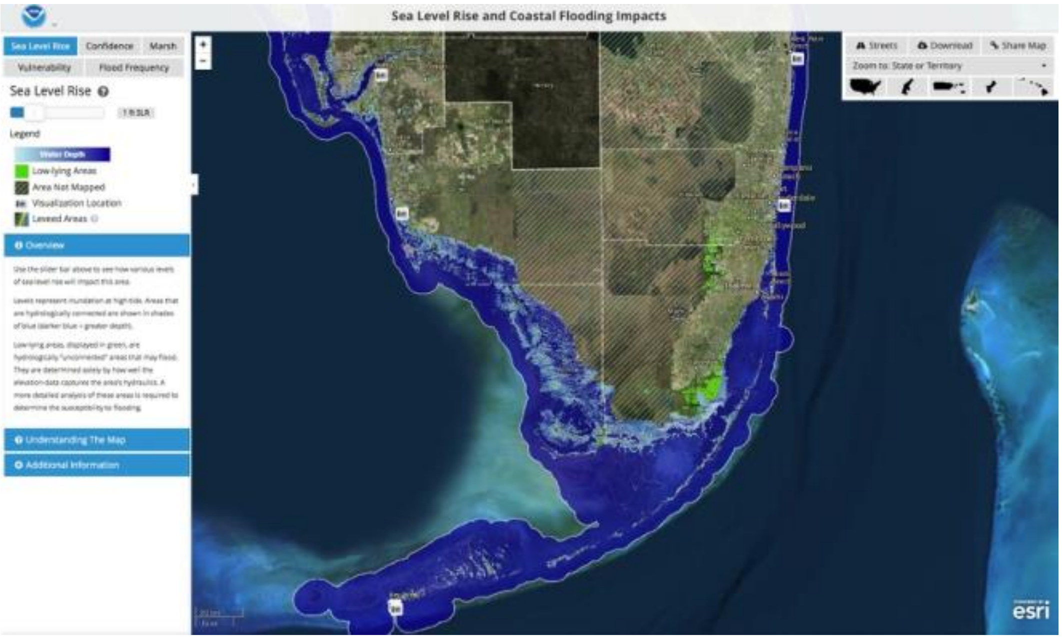
B Appendix: Environmental Changes

B.1 Long-range Projections

Long-term trends of warming temperatures, increasing drought frequency and sea level rise are components of the ongoing global climate change (Walsh et al., 2014; IPCC , 2013). These trends are expected to continue through 2030 irrespective of any changes in human emissions of greenhouse gases. The implications of these environmental changes on planning for Census 2030 are briefly summarized below, particularly for disaster contingency.

Continuation of long-term observed trends of warming, increases in drought and wildfire, and sea level rise through 2030 are expected (Walsh et al., 2014). Coastal communities may change due to infrastructure damage from storm surge, undermining of storm water drainage systems, saltwater intrusion into coastal aquifers, and negative impacts on tourism, agriculture, fisheries and energy extraction (Walsh et al., 2014; Carter et al., 2014; Garfin et al., 2014). The migration driven by these local and regional environmental changes should range from an imperceptibly small component of other demographic shifts to a more notable effect that might become apparent in other surveys prior to Census 2030.

There are detectable multi-decadal climate trends superimposed on a background of significant shorter-term variability. Hurricane intensity and rainfall rates may be increasing (Walsh et al., 2014), and rising sea level leads to increased storm surges. Dramatic ice sheet disintegration leading to rapid sea level rise has a low, but still non-zero, probability of occurring by 2030 (Walsh, et al.,



Use the slider bar above to see how various levels of sea level rise will impact this area.

Levels represent inundation at high tide. Areas that are hydrologically connected are shown in shades of blue (darker blue = greater depth).

Low-lying areas, displayed in green, are hydrologically "unconnected" areas that may flood. They are determined solely by how well the elevation data captures the area's hydraulics. A more detailed analysis of these areas is required to determine the susceptibility to flooding.

Figure B.1: South Florida, 1-foot sea level rise; darker blue indicates greater water depth. Source: https://coast.noaa.gov/slr/ .

2014; Church et al., 2013). The U.S. National Climate Assessment recommends planning for global sea level rise by 2100 of 1 to 4 feet (0.3-1.2 m), and to use a high of 6.6 feet (2 m) in situations with low risk tolerance (Walsh et al., 2014; Moser et al., 2014). The most pessimistic projections give a sea level rise of one foot by 2030. Figure B.1 shows the impact on South Florida.

Given the movement of the population toward the coastal zone (Moser et al., 2014), the above trends indicate an increase in the number of people vulnerable to a catastrophic displacement of the type that occurred in 2005 with Hurricane Katrina. Another example is Superstorm Sandy, in which tens of thousands of people in New Jersey and New York in October 2012 were displaced. Such an event would have been problematic had it occurred in March of a census year. For the 2010 Census, disaster contingency was explicitly included in the risk management plan (US Census Bureau, 2012). Such plans are not evident in the current 2020 Operational Plan (US Census Bureau, 2015), but should be developed and coordinated with the Department of Homeland Security. Disadvantaged communities tend to live in less protected locations; thus, environmental disasters can be expected to disproportionally impact people who are already “hard to count”.

B.2 Further Discussion of Historical Data

The above discussion focuses on long-range projections of environmental changes relevant to conduct of Census 2030. A different perspective is pro- vided by an assessment of the following historical data on environmental changes, beginning with coastal inundation, which is driven by a combination of sea-level rise and hurricanes. Tide gauges at most coastal locations in the U.S. show steady rates of sea level rise consistent with a global rate of 3 cm (1.2 inches) per decade; some of these records extend back for more than a century [http://tidesandcurrents.noaa.gov/sltrends/sltrends.html]. Isolated U.S. locations show much more rapidly rising (or even falling) sea levels associated with local subsidence or uplift. The historical data for U.S. hurricane numbers are widely discussed in the expert community (e.g., (Blake et al., 2011). The 60-year cyclical behavior during the 20th century has been attributed to a coherent mode of the climate system, the Atlantic Meridional Oscillation (AMO). Other measures of hurricane activity show a similar pat- tern from the mid-20th century (Figure B.2) while a more complex measure of hurricane activity over a season, shows an upward trend in the North Atlantic (Walsh et al., 2014). However, the upward trend is documented only for the past 25 years, too short compared to the AMO period to draw any firm conclusions. Thus, considering these historical trends in both sea level rise and hurricanes an argument can be made that the probabilities of disruptive coastal inundation in 2030 will not be very different than they are today.

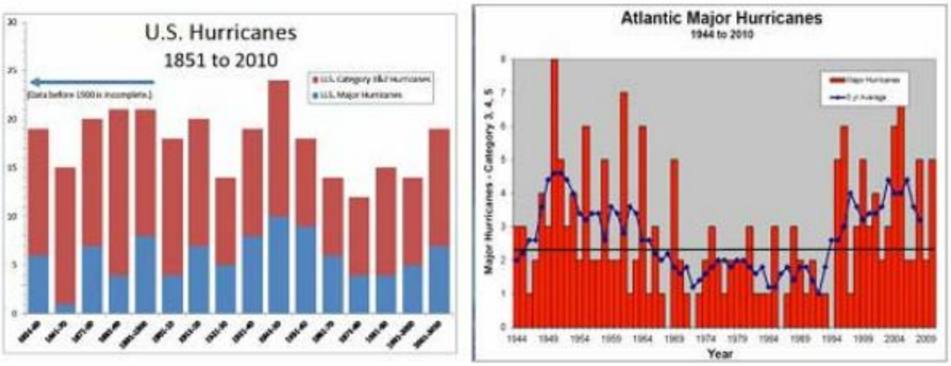


Figure B.2: Historical data for U.S. hurricanes. Source: http://www.nhc.noaa.gov/pdf/nws-nhc-6.pdf and http://www.aoml.noaa.gov/hrd/Landsea/gw hurricanes.

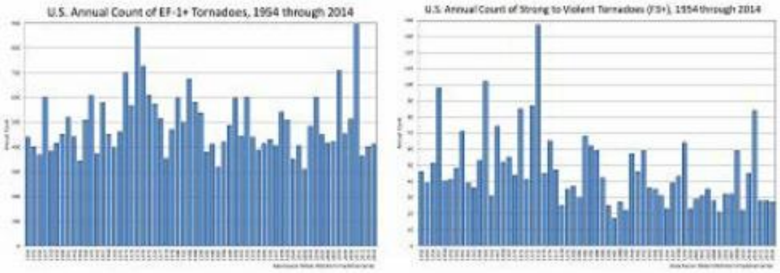


Figure B.3: Historical record of U.S. tornados. Source: http://www.ncdc.noaa.gov/climate-information/extreme-events/us-tornado-climatology/trends.

A historical record of U.S. tornados has been compiled by NOAA (e.g. Blake et al. 2011) only tornados above a minimum strength can be usefully analyzed due to expected under-reporting of weak tornados before widespread Doppler radar coverage. The data in Figure B.3 indicate there has been little trend in the frequency of the stronger tornadoes over the past 55 years. We can therefore expect that the probability of a strong tornado disrupting the 2030 Census will be about the same as it is today.

Considering drought and flood, the fraction of the U.S. under extreme climate stress

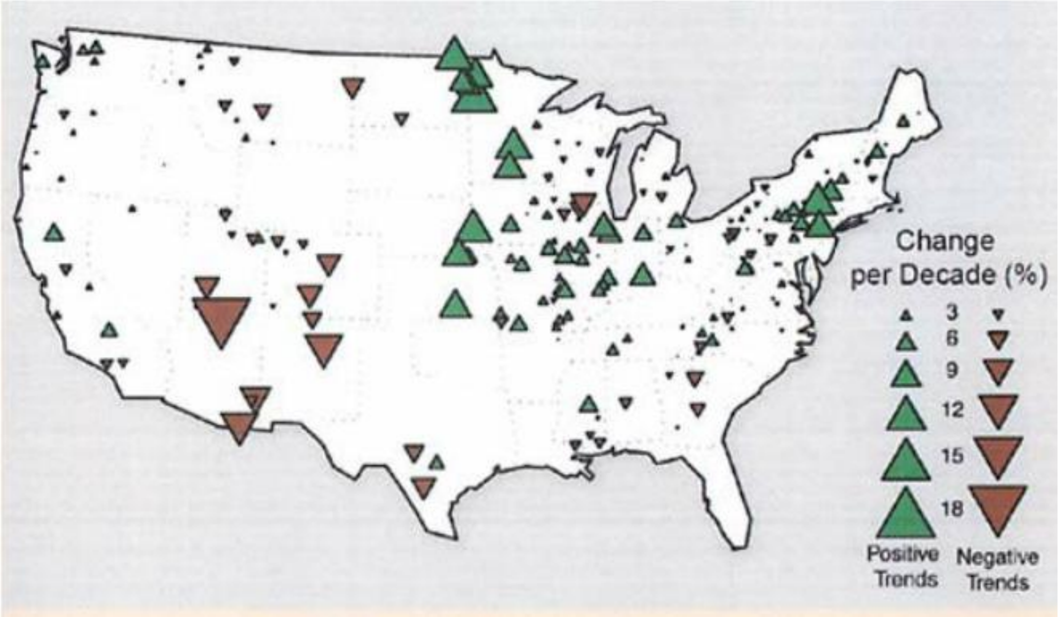


Figure B.4: Trend magnitude (triangle size) and direction (green = increasing trend, brown=decreasing trend) of annual flood magnitude from the 1920s through 2008. Local areas can be affected by land-use change (such as dams). Most significant are the increasing trend for floods in the Midwest and North- east and the decreasing trend in the Southwest. Source: Walsh et al. (2014).

(top and bottom deciles) has been relatively stable, with perhaps a slight increase in recent years. However, the past century has seen changes in the patterns of floods and droughts across the country, with more flooding in the upper Midwest and Northeast, and less in the West, as shown in Figure B.4. Finally, data on U.S. wildfires is available from the EPA [https://www3.epa.gov/climatechange/science/indicators/ecosystems/widfires.html]. The number has been fairly constant, if not decreasing, for the past 30 years, while the area burned increased by about a factor of two in 2000, and has been steady since.

B.3 Assessment

In conclusion, we emphasize that projections for 2030 based on current multi- decadal trends is complicated by considerable shorter-term variability. Considering the historical trends and metrics of diverse environmental variables, we assess that the probabilities of disruptive environmental events during the 2030 Census will not be significantly different from what they are today, with the exception that the threat of flood in the upper Midwest and Northeast might be greater. A variety of storm and climate models do predict increased frequency of extreme events over the coming decades, but uncertainties are large and definitive validation is lacking. Nevertheless, contingency planning should proceed with the present assessment revisited in 2021 and 2026 as the U.S. climate evolves and our understanding of its changes improves. Thinking beyond 2030, relocation of coastal communities can be expected to increase through 2100. A rolling census would be more accurate in the face of such relocations than the current approach of counting on April 1 of every tenth year.

References

Carter, L. M., J. W. Jones, L. Berry, V. Burkett, J. F. Murley, J. Obeysekera, P. J. Schramm, and D. Wear (2014), Chapter 17: The Southeast and the Caribbean, in Climate Change Impacts in the United States: The Third National Climate Assessment, edited by J. M. Melillo, T. C. Richmond, and G. W. Yohe, pp. 396–417, U.S. Global Change Research Program, doi:10.7930/J0N-P22CB.

Church, A. J., P. U. Clark, A. Cazenave, J. M. Gregory, S. Jevrejeva, A. Levermann, M. A. M. eld, G. A. Milne, R. S. Nerem, P. D. Nunn, A. J. Payne, W. T. Pfeffer, D. Stammer, and A. S. Unnikrishnan (2013), Fifth Assessment Report Technical Assessment Chapter 13: Sea Level, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change edited by Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley], pp. 1–124, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Garfin, G., G. Franco, H. Blanco, A. Comrie, P. Gonzalez, T. Piechota, R. Smyth, and R. Waskom (2014), Chapter 20: Southwest, in Climate Change Impacts in the United States: The Third National Climate Assessment, edited by M. Melillo, T. C. Richmond, and G. W. Yohe, pp. 462–486, U.S. Global Change Research Program, doi:10.7930/J0MS3QNW.

IPCC (2013), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, [TF, Stocker and Qin, D and Plattner, G K and Tignor, M and Allen, S K and Boschung, J and Nauels, A and Xia, Y and Bex, V and Midgley, P M (eds.)], Cambridge University Press, Cambridge, UK and New York NY, USA.

Moser, S. C., M. A. Davidson, P. Kirshen, P. Mulvaney, J. F. Murley, J. E. Neumann, L. Petes, and D. Reed (2014), Chapter 24: Coastal zone development and ecosystems, pp. 579–618, U.S. Global Change Research Program, doi:10.7930/J08G8HMN.

US Census Bureau (2012), 2010 Census Risk Management Process Operational Assessments Report, US Census Bureau.

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

Walsh, J. E., D. Wuebbles, K. Hayhoe, J. Kossin, K. Kunkel, G. Stephens, P. Thorne, R. Vose, M. Wehner, J. Willis, D. Anderson, S. Doney, R. Feely, P. Hennon, V. Kharin, T. Knutson, F. Landerer, T. Lenton, J. Kennedy, and R. Somerville (2014), Chapter 2: Our Changing Climate, in Climate Change Impacts in the United States: The Third National Climate Assessment, edited by J. M. Melillo, T. T. C. Richmond, and G. W. Yohe, pp. 19–67, US Global Change Research Program, doi:10.7930/J0KW5CXT.