INTEGRATED REVIEW SYSTEM: PLANS FOR TRACKING U.S. CENSUS 2000 ESTIMATION PROCESSING

Randal ZuWallack, Susan L. Atha, and Philip M. Gbur, U.S. Census Bureau
Randal ZuWallack, U.S. Census Bureau, Washington, DC 20233-7600

Key Words: Verification; Enumeration

IReS Summary

The Integrated Review System (IReS) is a computer system being developed for reviewing the estimation process for Census 2000. It is a coordinated review. Each estimation procedure in the process has a review plan that is preplanned, documented, and communicated. The IReS tracks and summarizes data as it progresses through the estimation process. Results of the IReS review are available quickly and disseminated internally for further review and analysis. By providing fast, accurate results, the system is an effective means to review the estimation process and the Census 2000 population numbers and characteristics.

Background

Every ten years, the Census Bureau collects information about the population through the decennial census enumeration. Data arrives from many sources including mail returns, enumerator interviews, and follow-up procedures. Then, using many estimation procedures, the Census Bureau converts the collected data into numerous data products to be released to the public. Together, these procedures are the estimation process that transforms the raw data into complete, consistent, and confidential information available for public release. To ensure the estimation process results in quality numbers, we are developing the IReS to monitor and summarize results.

The IReS was originally developed and tested for Census 2000 Dress Rehearsal. It proved less effective than we expect for Census 2000. We found that the amount of time and resources allotted for conducting the review successfully was insufficient. Thus, the flow of data through the process overran the review. The insufficient resources also limited the amount of review that could be done and the timing of the review, resulting in an increased lag time between the procedure and the results of the review.

By beginning the system planning early and allocating sufficient staff to the project, the design for the Census 2000 IReS alleviates many of the problems that hampered the dress rehearsal review. The IReS is more extensive than in dress rehearsal, encompassing additional estimation procedures and producing results for internal review much faster, which are key factors in developing a successful review.

Role of the IReS

By monitoring and summarizing results, the IReS aids in explaining and understanding the population totals and characteristics, as well as verifies the reasonableness of the results. The IReS discovers and reports anomalies in the data so that the underlying cause can be explored and tracked throughout the processing. Additionally, the IReS is being designed to disseminate results internally for review and further analysis, rather than for documenting official results. As is done with all information that is confidential, results are restricted and have a limited distribution.

Although the IReS plays a valuable role in reviewing the results of the estimation procedures, it does not verify that the computer software developed for each procedure is correct. Additional independent reviews verify the accuracy of the computer processing for each procedure. The IReS complements these verifications.

Estimation Process Overview

The Census Bureau is releasing two sets of population numbers, one that is corrected for coverage error and a second that is not. The set of numbers not corrected for coverage is used to generate the state totals used for congressional reapportionment. Both sets of numbers are used for generating Public Law 94-171 (PL) counts for each census block, which states may use for redistricting.

We are developing the IReS to review the estimation procedures for both sets of numbers. The three procedures that the IReS is reviewing for developing the uncorrected numbers are:

- Unclassified Estimation
- Edits and Allocations
- Disclosure Avoidance

In addition to these three procedures, the IReS is reviewing the following procedures for the corrected numbers:

This paper reports the results of research and analysis undertaken by Census Bureau Staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

- SBE Estimation
- Matching and Follow-up
- Missing Data Imputation
- Dual System Estimation
- Small Area Estimation

The review also includes Housing Unit Dual System Estimation, which estimates housing unit coverage. Finally, the IReS is conducting a content review of the Hundred percent Census Unedited File (HCUF) to ensure the estimation process is beginning with a quality product.

The primary files upon which the IReS is based include the HCUF, the Hundred percent Census Edited File (HCEF), the Hundred Percent Estimated Detail File (HEDF), the Hundred Percent Detail File (HDF) and the P- and E-sample files from the Accuracy and Coverage Evaluation (A.C.E.) (Hogan, 2000).

Brief descriptions of the above mentioned procedures and files appear in the next sections. Also, the flowchart in Figure 1 displays the relationship between the estimation procedures and the various inputs and outputs.

Procedures and Files for Uncorrected Census Counts

Unclassified Estimation: Unclassified Estimation imputes for missing housing unit status (occupied, vacant, delete) or the number of persons for any occupied census housing unit without household size, using a nearest neighbor hot deck.

HCUF: The HCUF contains the 100% data items from the census enumeration short form as well as census operational variables. Observations may have incomplete or inconsistent demographic data.

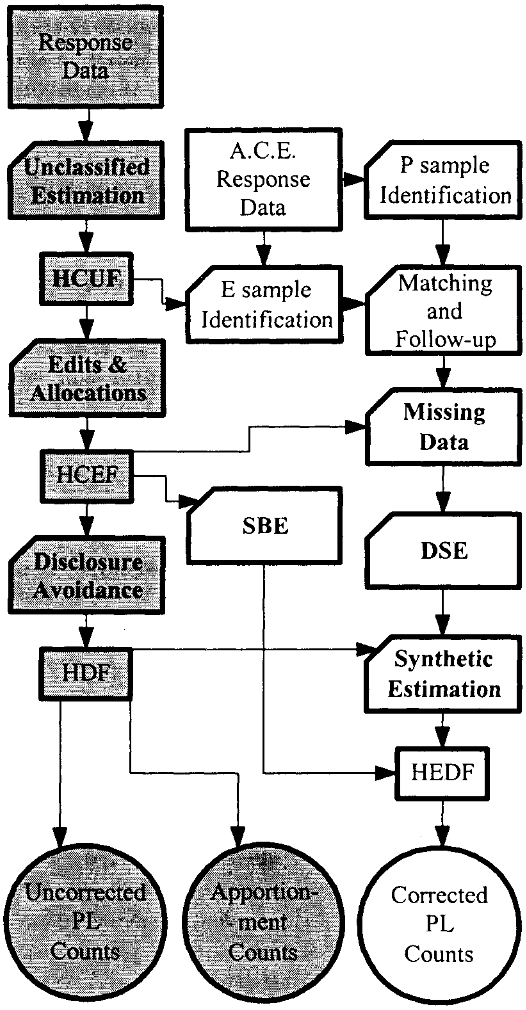
Edits and Allocations: The 100% Edits and Allocations is the process of editing inconsistent 100% data items and imputing for missing 100% data items collected during the census enumeration. Population items include relationship, sex, race, origin, age, and date of birth. Household items include householder determination and tenure.

HCEF: The HCEF contains the 100% data items from the census enumeration short form as well as census operational variables. No observations have incomplete or inconsistent demographic data.

Disclosure Avoidance: To maintain the confidentiality required by law (Title 13, United States Code), the Census Bureau assures that the published data do not disclose information about specific individuals, households, or housing units. The primary means of assuring confidentiality consists of exchanging the data for similar households. This means that pairs of household records that match on a cross tabulation of certain key variables but are in different geographic locations may be swapped across those geographic locations.

HDF: The HDF contains the 100% data items from the census enumeration short form as well as census operational variables. No observations have incomplete or inconsistent demographic data. Confidentiality of publicly released data products resulting from the HDF is assured.

Figure 1. Flow of the Census Files and Estimation Procedures (Not including all input and output files)



Procedures and Files for Corrected Census Counts

SBE Estimation: Service Based Enumeration (SBE) is the process of enumerating persons without usual residence by visiting shelters, soup kitchens, mobile food vans, and targeted non-sheltered outdoor locations.

Multiplicity estimation is used to account for the people who use service facilities, but not on the day of the enumeration (Kohn and Griffin, 1999).

Matching and Follow-up: The P-sample people are matched to the census, first with a computer and then clerically if needed. After the matching is completed, field follow-up is conducted for selected cases. Another clerical match is done after follow-up is completed.

A.C.E. Missing Data Procedures: Missing data procedures for the A.C.E. impute or adjust for missing information essential for calculating dual system estimates of the population. The missing data in the A.C.E. occurs in two forms, unit missing data resulting from noninterviews and item missing data. Item missing data is further divided into two categories, missing person characteristics (age, sex, race, Hispanic origin and tenure), and unresolved person status (match, correct enumeration, and residence).

Dual System Estimation: Dual System Estimation is the procedure for measuring the degree of population coverage error observed during the census enumeration. By comparing the census enumeration results to A.C.E. results, we calculate dual system estimates (DSEs) for different post-strata, based on geography and demographic variables. For each post-strata, we then calculate coverage correction factors (CCFs) by taking the ratio of the DSE to the census count.

Synthetic Estimation: We use Synthetic Estimation to calculate population estimates below the post-strata level such as blocks, tracts, counties, congressional districts, and states. We calculate estimates by applying the CCFs to block level population counts. Then, we can tally the block level estimates to get estimates for higher levels of geography.

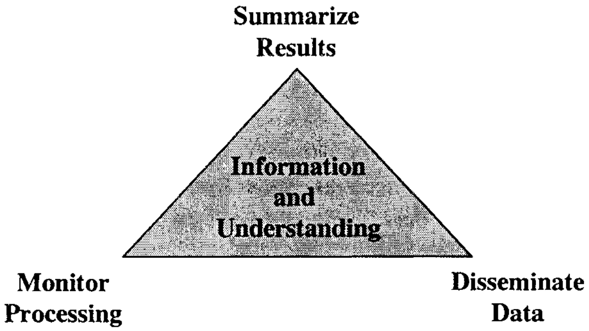
HEDF: The HEDF contains the 100% data items from the census enumeration short form as well as census operational variables. No observations have incomplete or inconsistent demographic data. The file is corrected for coverage error. Confidentiality of publicly released data products resulting from the HEDF is assured.

Housing Unit DSE: Housing Unit DSE is the procedure for measuring the degree of housing unit coverage error observed during the census enumeration. Similar to the person DSE, housing unit DSE compares census housing units to an independent list of housing units to calculate dual system estimates for different post-strata. For each post-strata, a CCF is then calculated by taking the ratio of the DSE to the census housing unit count.

System Components

As stated earlier, we are developing the IReS to monitor and summarize results of the estimation process. To do this, the system relies on three primary components: summarizing results with pre-specified tables, monitoring the data processing for unexpected data patterns and outliers, and disseminating data internally for further review. The three components, explained in the following paragraphs, all work with one another to provide an understanding and an explanation of the data processing results. Figure 2 is a graphical representation of the IReS's three components.

Figure 2. IReS Component Relationship



Summarize Results: To ensure a complete review, we identify pre-specified tables prior to each estimation procedure. These tables provide key information at the state and county level for understanding and explaining the results of the procedure and the processing itself. The tables fall into two categories, operational summary tables and standard tables. The operational summary tables are specific to each procedure and provide a summary of the processing. Standard tables are tables that are generated multiple times throughout the entire process. By comparing the standard tables before a procedure to those same tables generated after, we can gather an understanding of the impact the procedure has on population and housing numbers.

To determine what information should be provided in these tables, the IReS staff reviews procedure documents and specifications to identify key information. Then, the staff working on the procedure reviews the tables and provides input as to what they feel is necessary to sufficiently summarize and understand the procedure. Finally, we present the proposed tables to management for their input. This process ensures that several opinions with different levels of involvement in the procedure are accounted for, making the tables as complete and efficient as possible.

The IReS staff then designs, programs, and tests the tables prior to the procedure's implementation so that the results are disseminated efficiently. Within three days of the completion of the estimation procedure, we will report the results of the summary tables internally.

Monitoring Data Processing: In addition to tracking data from one procedure to the next, the IReS assesses the reasonableness of the data by monitoring the data processing at the local census office (LCO), county, and tract level. The monitoring focuses primarily on processing variables, such as imputation rates or non-response adjustments rather than population and housing unit totals. For each state, we calculate summary statistics for the processing variables such as the median, minimum, and maximum at the tract, county, and LCO level. Depending upon availability, we may use independent data sources, such as past census results and demographic analysis as a comparison. These outside sources provide information about certain characteristics of the population, which may identify unexpected data patterns when compared to Census 2000 numbers.

In addition to the state level monitoring, we identify tract, county, and LCO level outliers for processing variables. For example, county imputation rates are considered outliers if the percentage of imputations in the county exceeds the percentage of imputations in the state by a significant amount.

Unexpected data patterns, identified either through pre-specified tables or monitoring of the processing variables, lead to further investigations of the data. Since complete exploration of the data is resource intensive, it's not feasible to allocate staff and computer time to do this for the entire country. Thus, using unexpected results or preselected states or counties based on prior information to target areas is beneficial in terms of efficiently using staff and computer resources.

As appropriate, we'll document unusual data patterns and outliers in special reports within one working day of discovery. Otherwise, we will summarize the monitoring statistics in a report within three days of the completion of the estimation procedure.

Disseminating Data: We expect the pre-specified tabulations to include much, but possibly not all of the necessary information for understanding the census processing. Thus, we are creating a data set and table generator for disseminating data that were not foreseen as being relevant to the estimation process. The generator is an interactive application that allows the user to easily provide the specifications of the desired table or data set as input to the system. The generator prompts the user for required information such as the level of geography, the variables, and whether there are any variables used to subset the data. Access to the major databases, including the HCUF, the HCEF, the HDF, the HEDF and the P- and E-sample files are available through the table and data set generator. The generator provides a tool for efficiently accessing and investigating the data to uncover the cause of unexpected results.

With the generator in place, we expect most requests to be fulfilled by the IReS staff within one working day, depending on the accessability of the major databases. Requests that require access to more obscure data files may take more time.

Resources: Required vs. Available

One of the major concerns about the effectiveness of the IReS is the available resources. The project requires considerable staff time, data storage, and processing space. In addition to staff working directly on the IReS, the project needs support from other staffs, such as those responsible for designing, implementing, or programming the various estimation procedures.

The IReS has several options based on the resources available. The current disk space allotted to the IReS is about 50 gigabytes, which we expect to be sufficient for storing SAS® programs and output, but not for storing data files permanently. The IReS is most efficient if SAS data sets of the major data files are easily accessible, though clearly the current disk space cannot support the storage of these data sets for the complete United States. Based on files for the Census 2000 Dress Rehearsal as a rough estimate, California alone would need around 8 gigabytes to store housing unit and person data in SAS data sets. So the options for the IReS are 1) request enough disk space to store the full databases as SAS data sets, 2) store compressed extracts from the major databases, 3) store files temporarily for review, or 4) have access to the production files in ASCII format.

Each of these options has associated negatives. More disk space is the most effective solution, but there is considerable cost and management involved. Options two and three both have no additional monetary cost, but each compromises on efficiency and adds to file management. The final option has no additional costs or file management, but efficiency suffers since SAS data sets won't be available.

The most likely scenario for solving the resource problem is a combination of options 2, 3, and 4 above. We can create the summary and monitoring tables on a flow basis, meaning files are rotated onto the IReS disk space for review and removed upon completion. We can store file extracts containing variables commonly requested on the IReS disk for quick access with the table and data set generator. Data requests requiring variables not on these extracts would require access to the ASCII production files. This allows for a complete, effective review of the estimation procedures, but slightly impedes the ability to fulfill data requests. In some instances, we may use random sampling and targeted sampling to limit the amount of resources needed for the review.

Designing the Computer System

The IReS is designed to be a user-friendly interactive application. Using the windowing feature in SAS, we are developing a menu based system in the OpenVMS™ environment. The menus allow the user to navigate to the procedure being reviewed and enter in the required parameters for running the review. The system prompts the user for information such as input and output directories, the level of geography, and whether it is a full review or a report from a previous review. Upon entering the required information, the system creates the pre-specified tables and the monitoring statistics specific to that procedure, or creates a table or data set. Figure 3 is the main menu for the IReS prototype being developed with test data based on the Census 2000 Dress Rehearsal.

The ease of use improves efficiency. It allows multiple analysts to access the programs, improving the speed of the review. The amount of training is minimal, reducing the amount of time needed to sufficiently learn the system.

Example: HCUF Content Review

The review plan for the HCUF Content Review, which is the first process covered under the IReS, is a good example for illustrating the setup for a typical review. The primary goal of the review is to determine whether the HCUF creation process is resulting in a quality product, which in turn will not have negative implications on the census procedures that follow.

To best review the content, we are forming a team with representatives who have diverse knowledge of variables on the HCUF. The team approach is beneficial for two main reasons, 1) the size of the HCUF is too large for an individual to sufficiently review in a timely fashion, and 2) the HCUF contains a wide array of variables and review results are better understood by team members familiar with the variables.

The review is tailored to variables that fall into one of five categories: categorical, numerical, geographical, time/date or name/address. We will monitor categorical variables by doing these three types of review:

- Generating a distribution of values, which is summarized in a report by listing the number of missing values, the most frequent value and the least frequent value, as well as the corresponding frequency counts for these values.
- Conducting a range check to identify values that are out of scope according to the HCUF documentation.
- Checking the variable source for variables transferred directly from the DMAF or DRF2. Variables having values different from their source will be summarized in the reports.

We will conduct a similar review for numerical variables, but the report will contain a different set of summary statistics. These statistics include the number of missing values, the mean, median, minimum, and maximum. We will also provide the frequency of the median, minimum, and maximum in the report. As with the categorical variables, we'll do a range and source check for numerical variables. For both geographical and time/date variables, we will only conduct a range check and source check. The appropriate range for each time/date variable will be identified by the analyst familiar with the timing of the operation. Values that are out of scope or do not match their source will be summarized in the report. The final type of variable on the HCUF is name/address, which is not being reviewed as part of the content review.

The team will review the variables on a sample of the HCUF files. Five hundred and fifty nine HCUF files, corresponding to each LCO, are created on a flow basis, which is driven by data availability. Following the same flow as the file processing, the content review team will evaluate the first 10 files created. For the 11th through 559th file, we will select and review a systematic sample of files. Additionally, the content review team will evaluate specific LCOs upon request from management. In all, over 10% of the HCUF files will be reviewed.

The resulting output from the review is a set of tables, one for each type of variable. Figure 4 contains a report example generated by the IReS prototype developed with test data based on the Census 2000 Dress Rehearsal.

Conclusion

The IReS is a major undertaking, but the final product is a very powerful tool for accessing information and understanding operations. One of the best qualities of the IReS is the ability to relate the estimation procedures to each other. By reviewing the entire estimation process, we can compare results at any point in the process to any other point, which provides an understanding of the interactions between and among the procedures.

A second quality worth noting is the ability to generate data sets and tables for both summarizing results and for other informational purposes. This expedites the process of disseminating the data internally for further analysis. Having data available from a centralized source enhances the coordination of the review, resulting in a complete review.

Finally, the monitoring and summary tables work in conjunction with the independent verification of the estimation procedures. These verifications focus primarily on validating computer programs, whereas, the IReS centers on understanding the procedure results and verifying the reasonableness of the outcome. Coupled together, the IReS and the independent verifications complement each other to provide a complete and effective review of the estimation process.

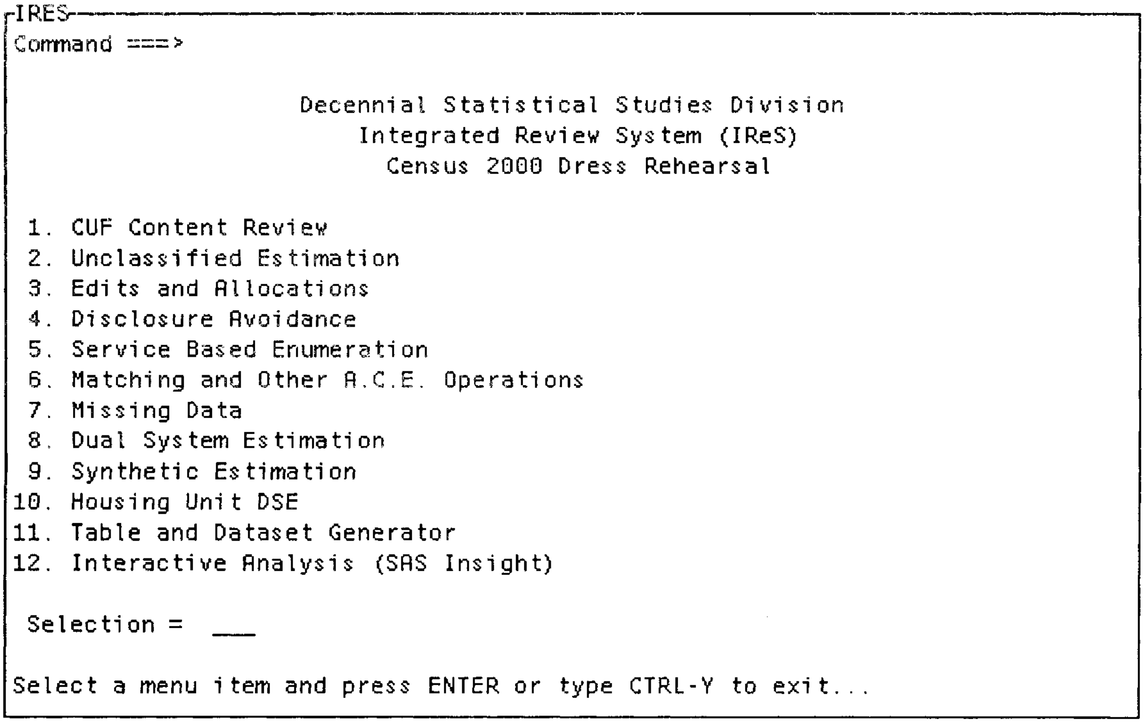
As stated previously, one of the keys to a successful review is the timely dissemination of results internally for further review and analysis. The IReS is a preplanned, preprogrammed system that provides prompt dispersion of key information. In summary, the IReS provides a planned, focused, and guided approach for reviewing the population numbers and characteristics for Census 2000.

References

Hogan, H. (2000), "The Accuracy and Coverage Evaluation: Theory and Application" *Proceedings of Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, to appear.

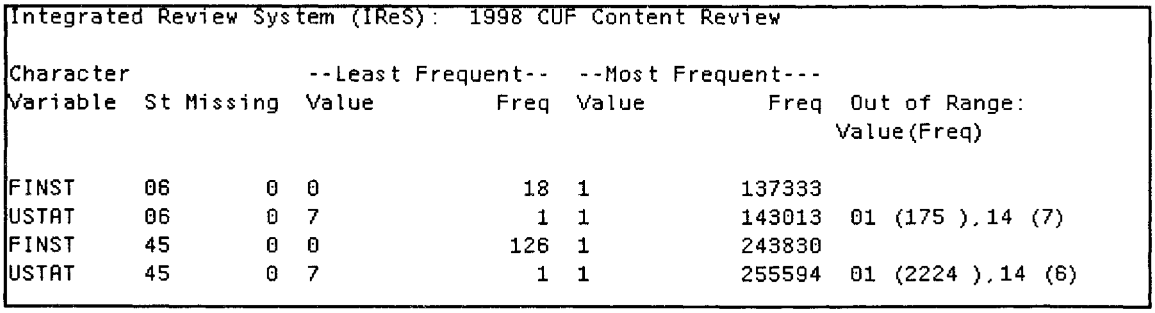
Kohn, F., and Griffin, R. (1999), "Service Based Enumeration Estimation" *Proceedings of Survey Research Methods Section, American Statistical Association*, Alexandria, VA, American Statistical Association, pp. 519-522.

Figure 3. Main Menu For the IReS Prototype



```
IRES
Command ===>

                    Decennial Statistical Studies Division
                        Integrated Review System (IReS)
                         Census 2000 Dress Rehearsal

 1. CUF Content Review
 2. Unclassified Estimation
 3. Edits and Allocations
 4. Disclosure Avoidance
 5. Service Based Enumeration
 6. Matching and Other A.C.E. Operations
 7. Missing Data
 8. Dual System Estimation
 9. Synthetic Estimation
10. Housing Unit DSE
11. Table and Dataset Generator
12. Interactive Analysis (SAS Insight)

 Selection = ___

Select a menu item and press ENTER or type CTRL-Y to exit...
```

Figure 4. Report Examples Generated by the IReS Prototype



```
Integrated Review System (IReS):  1998 CUF Content Review

Character              --Least Frequent--  --Most Frequent---
Variable  St Missing   Value         Freq  Value          Freq  Out of Range:
                                                               Value(Freq)

FINST     06       0   0               18  1            137333
USTAT     06       0   7                1  1            143013  01 (175 ),14 (7)
FINST     45       0   0              126  1            243830
USTAT     45       0   7                1  1            255594  01 (2224 ),14 (6)
```