United Nations ECE/CES/GE.41/2018/14

Economic and Social Council

Distr.: General
6 July 2018

Original: English

Economic Commission for Europe

Conference of European Statisticians

Group of Experts on Population and Housing Censuses

Twentieth Meeting
Geneva, 26–28 September 2018
Item 7 of the provisional agenda
Geo-spatial information

Validating the inputs used to create geo-referenced Census data

Note by the Unite States Census Bureau*

Summary

The United States Census Bureau (Census Bureau) is continuously validating the inputs used to create geo-referenced data. In preparation for the 2020 Census, the Census Bureau conducted a 100 percent review of addresses in its' Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) System. Staff members compared different vintages of aerial imagery and housing unit counts using geographic information systems (GIS). This combination of human and technological resources enabled the Census Bureau to validate the accuracy of approximately 70 percent of the addresses while working in the office. Staff will validate the remaining 30 percent during an in-field operation in 2019.

While the use of this technology helped improve efficiency and reduce costs related to the 2020 Census, the Census Bureau is already researching ways to streamline the process for the 2030 Census. This document will focus on those ideas, including automated change detection and continual in-field validation by professionally trained geographers.

* Prepared by Deirdre Dalpiaz Bishop and Michael Ratcliffe.

GE.18-11157(E)
1811157

I. Introduction

1. The Census Bureau is continuously validating the inputs used to create geo-referenced data. In preparation for the 2020 Census, the Census Bureau conducted a 100 percent review of addresses in its' Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) System. Staff members compared different vintages of aerial imagery and housing unit counts using geographic information systems (GIS). This combination of human and technological resources enabled the Census Bureau to validate the accuracy of approximately 70 percent of the addresses while working in the office. Staff will validate the remaining 30 percent during an in-field operation in 2019.

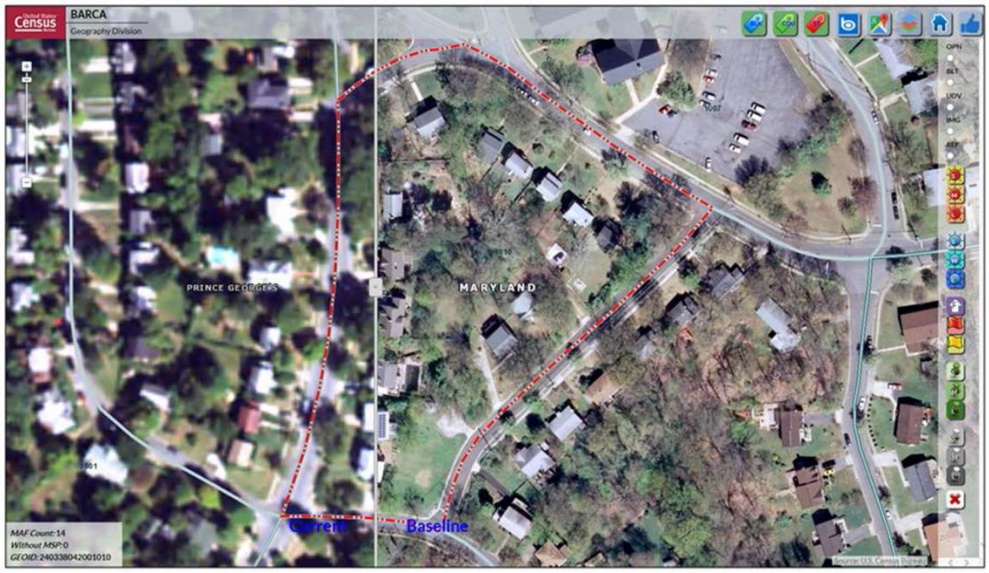
2. While the use of this technology helped improve efficiency and reduce costs related to the 2020 Census, the Census Bureau is already researching ways to streamline the process for the 2030 Census. This document will focus on those ideas, including automated change detection and continual in-field validation by professionally trained geographers.

II. Background

3. The Census Bureau began In-Office Address Canvassing (IOAC) in September 2015 as the first official 2020 Census operation. IOAC uses a combination of aerial imagery and administrative and programmatic data to identify where change is occurring in the address base on the ground. Internally developed software, called the Block Assessment, Research, and Classification Application (BARCA), is used to help detect that change.

4. As illustrated in Figure I, the BARCA displays information to assist analysts in identifying and classifying residential change. The information includes aerial imagery, address counts for each census block, points representing addresses, roads, parcel boundary layers (where available), and a variety of tools and flags to record work.

Figure I
The Block Assessment, Research, and Classification (BARCA) Application



5. During IOAC, Census Bureau analysts compare imagery from two different points in time (at the time of the last census compared to now) to determine if there is stability or change on the landscape. At the census block level, analysts work to:

(a) The information obtained is close to the situation as of 1 January 2017;

(b) Determine possible growth and/or decline of housing units;

(c) Identify possible future growth – where housing might be built;

(d) Identify missing or misaligned roads or boundaries;

(e) Determine if the census block has under-coverage or over-coverage of housing units; and

(f) Determine block status (is the block already fully developed, undevelopable, or open for future development).

6. This review takes approximately 90 seconds for each census block. To date, a nationwide review of 11 million census blocks identified 79 percent of the blocks as stable, 16 percent as requiring further research, and 5 percent as needing updated aerial imagery or administrative data. The 16 percent of blocks classified as requiring further research are flagged with colourful pins that indicate whether there appears to be under-coverage of over-coverage of addresses in the Census Bureau's address list (see Figure II). These blocks will be sent to an in-field operation for further validation.

Figure II
Flags Depicting Over-Coverage and Under-Coverage of Housing Units

7. While the use of this technology helped improve efficiency and reduce costs related to the 2020 Census, the Census Bureau is already researching ways to streamline the change detection process for the 2030 Census

III. Detecting changes to geospatial data

8. When thinking about change detection, it is critical for the Census Bureau to keep sight of our primary mission as a statistical agency and the foundational data needed to carry out that mission. What are the foundational geospatial data needed by the Census Bureau, or for that matter, any statistical agency? They are, put simply, respondent locations. These could be houses, apartments, dormitories, or any structure or object on the

landscape in which someone might live. From an economic standpoint, respondent locations could be business locations, whether traditional "brick and mortar" stores or locations from which on-line business is conducted, vendors' stalls, factories, unique divisions within a corporation's building (e.g., the accounting department, the production floor, the shipping and receiving department), or even spatially diffuse operations conducted by a company (office operations, oil wells, offshore drilling platform, refinery, pipeline). In short, data collected will be associated with particular geographic areas (i.e., geocoded), tabulated, and disseminated. This requires locational information of sufficient accuracy to collect, tabulate, and disseminate data at a level of precision necessary to meet analytical, planning, and policy-making expectations.

9. Over the past decade, the Census Bureau has seen increased expectations for precise locational information, spatially accurate features and boundaries, and related attributes and metadata. Accompanying this expectation for spatially accurate data is an expectation that data will be as up-to-date as possible and that geospatial data can be used to link multiple, sometimes disparate, datasets.

10. This drives us to a few fundamental questions regarding change detection and the management and updating of geospatial data:

(a) What changes do we need to detect?

(b) When, or how frequently, do we need to detect changes?

(c) What sources should we use?

(d) How do we know the data we have detected are accurate and authoritative?

11. Experience with IOAC, partnerships with tribal, state, and local governments, and evaluation of commercial address and spatial data files, has helped frame responses to these questions. Initial forays into automated change detection have yielded positive results, but also have raised questions related to what can be detected versus what should be updated and maintained in the Census Bureau's geospatial database.

A. What changes do we need to detect?

12. Returning to our foundational data, we need to detect changes to respondent locations. More specifically, we need to detect changes to the inventory of respondent locations and information about those locations. We typically think of change detection as identifying differences in the inventory of things—in our case, housing units (for the decennial census and demographic surveys) or businesses (for the economic census)—on the ground as determined by comparing two vintages of imagery. As noted, this is what we have been doing as part of our IOAC operation. But, as illustrated by the comparison of numbers of housing units visible in imagery to the number of addresses for the same area in the MAF, change detection can also detect differences between two data sources, with imagery as one data source and the MAF as the other. In some instances, review of two vintages of imagery reveals a lack of change over time, but comparison to the MAF indicates under-coverage or over-coverage. In this case, what has been detected is not change on the ground, but the need for change in a database. All of this is to state what is, perhaps, obvious—change detection from the standpoint of managing geospatial data involves both the detection of changes on the ground and the detection of differences between datasets, indicating the need for acquisition of new data in order to effect a change in one dataset or the other.

B. When, or how frequently, do we need to detect changes?

13. Two items must be considered when answering this question. The first is that the frequency of change detection must align with operational needs for data collection, tabulation, and dissemination. If these operations occur only once every ten years, then only decadal change is needed. If data are collected, tabulated, and disseminated on an annual basis, then annual change detection is needed. But, one can easily make the case that ongoing change detection and annual updates is more efficient than mounting a once-a-decade operation. This was part of the premise behind IOAC—conducting an ongoing process to review and detect changes to the housing landscape and assess the accuracy and completeness of the MAF. The general approach taken, then, is to conduct nationwide change detection processes along with accompanying geospatial data acquisition programs, generally carried out over the course of several years.

14. The other response to this question focuses on likelihood for change. Evaluation of results from the nationwide in-field canvassing operation for the 2010 Census revealed that changes to the MAF were limited to a small percentage of census blocks. Subsequent research leading to IOAC supported this finding, as have the results of IOAC itself. This means that a large proportion of the housing landscape will not change. Presumably, then, change detection operations could be conducted less frequently in areas of known stability and focus efforts, instead, on areas experiencing more rapid change.

15. The frequency with which we conduct change detection operations depends on the frequency of the data collection, tabulation, and dissemination cycle as well as the likelihood for change in any given area. It also depends on the level of quality, accuracy, and completeness of our geospatial data in comparison to other sources of the same data.

C. What sources should we use?

16. Our general approach to updating and maintaining geospatial data at the Census Bureau has been to rely upon datasets from authoritative sources, with “authoritative” referring primarily to the organizations responsible for creating the data. Generally, this means tribal, state, and local government partners, but in the case of the MAF, it also includes the U.S. Postal Service (USPS). Although local governments are responsible for creating and issuing addresses, because the USPS has the responsibility of delivering mail, the USPS Delivery Sequence File (DSF) also provides a reliable, authoritative source of addresses and attributes.

17. Our experience working with and evaluating various datasets demonstrates that each has particular limitations. A source may be authoritative (using the definition above) and, overall, may have a high degree of accuracy and completeness, but that does not mean it is without some level of error, inaccuracy, or incompleteness. This, coupled with temporal differences between datasets, means that a highly accurate authoritative dataset might not be sufficient at all times in all geographic areas, and that sources that have low overall quality and completeness may contain information that is usable in specific contexts. In other words, all sources are valuable for change detection as long as we understand and know their properties and characteristics, their specific utility, their limitations, and, importantly, how all of this can be harnessed for update and maintenance of the Census Bureau’s geospatial data.

18. To illustrate this, we turn again to our IOAC experience this decade. After completing a first review of all 11 million census blocks in the nation, the Census Bureau began a process through which to detect change in specific blocks, triggering them for re-review. A trigger, simply put, is anything that suggests there might be a change on the

ground or a change to the MAF or both. Processing of an address data source, such as the DSF, resulting in the geocoding of new addresses to a block, is a change detection process. In this case, the comparison of tallies of addresses in a block from one vintage of the DSF to another constitutes the change detection process, triggering the block for a new round of imagery review to validate the change relative to what is visible on the ground. As with the initial review, the result of re-review could be validation that the MAF has kept up with change on the ground or an indication that despite new addresses from the DSF, sources have not kept pace with change on the ground and additional data are needed (perhaps from a different source). Re-review could also identify temporal differences between datasets, particularly if the address data source contains more housing units in the block than are visible in imagery. Again, understanding the properties and characteristics of each source is critical to knowing how to react to the results of a change detection process.

D. Understanding accuracy and completeness

19. In planning a future of geospatial data update and maintenance that incorporates a robust automated change detection process, it is not enough to develop processes to detect change – the Census Bureau must first have a clear measure and understanding of the accuracy and completeness of the data – both in its geospatial database and in the data sources used to detect change and make updates. The mere presence of additional items in a source dataset is not necessarily an indicator of deficiency or under-coverage in the Census Bureau's geospatial assets. Likewise, lack of information in a source dataset might not be an indicator of inaccuracy but could be related to the particular use and purpose of the dataset.

20. For example, an address dataset may contain addresses for houses that have not yet been built. A roads dataset intended for routing and navigation may not include roads that are not yet drivable. Or, it might include attributes indicating the road is planned or even constructed, but not yet navigable. This does not necessarily mean the quality of the dataset is suspect; rather, the Census Bureau needs to understand the dataset's primary purpose and how that purpose relates to our own, and then make appropriate decisions about the changes we detect.

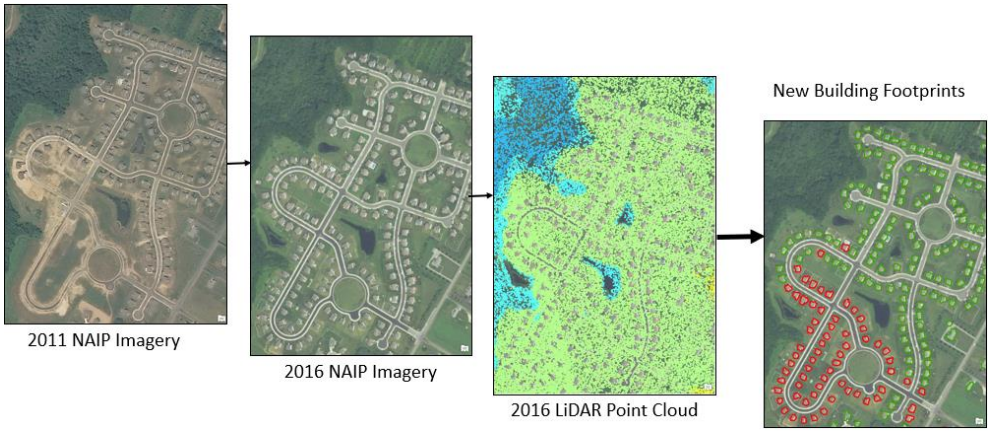
IV. Validating the inputs used to create geo-referenced census data – the future

21. Although interactive imagery review has been the primary means for validating whether the MAF is complete in each individual block, our IOAC process has grown to include other methods for validating the MAF, such as comparison of numbers of address records present in other address lists. Processing of tribal, state, and local address files through the Geographic Support System (GSS) Program yielded high match rates to the MAF and relatively low numbers of new addresses. Considering that the USPS draws from these same sources to update the DSF, which we then process every six months, files obtained through the GSS Program validated the overall quality and completeness of the DSF as well as the MAF. Evaluation and processing of five commercial address lists, yielded similar results. In-office MAF update operations utilized on-line address, parcel, and GIS data from local governments to update and validate address and feature data—in effect, a virtual canvass.

22. These efforts revealed that validation of updates to MAF/TIGER can occur using a variety of datasets and on-line sources, many of which are amenable to automated change detection, data discovery, processing, and linkage. This leads us to envision the following process:

(a) Automated imagery analysis to identify, at a minimum, areas that have remained stable as well as areas containing change to the built environment. This process includes an automated comparison of the number of units detected in imagery for the specific geographic unit of analysis (block, grid cell, etc.) and the number of addresses in the MAF for that same unit. Our initial work in this area has shown that automated change detection takes approximately the same amount of time as IOAC—on average, approximately 80 to 90 seconds per block. Our automated methodology, however, has not yet reached the stage at which the application can distinguish between residential structures and those that have different uses. But, it can provide an initial assessment of change or no change, providing a list of census blocks or other geographic units for further investigation. Ideally, automated change detection would be able to discern which new structures are residential and which are not. This is a more complex issue that likely will require machine learning techniques that train the application to recognize residential landscape contexts based on size and spacing of structures in conjunction with street and other feature patterns that are more typical of residential development than commercial development. Figure III illustrates the use of aerial imagery from the U.S. Department of Agriculture's National Agricultural Imagery Program (NAIP) and LiDAR (Light Detection and Ranging) point clouds to detect changes to the built environment, leading to the identification of building locations and building footprints. Such data will be critical to meeting future needs for highly accurate and precise geospatial data, facilitating a wide range of locational and contextual analysis;

Figure III
Use of Imagery and LiDAR to detect change and define building footprints



(b) Some changes to the housing landscape, such as conversions of whole buildings or within structure units from non-residential to residential uses, are not detectable through imagery. For this reason, automated change detection methods must also incorporate techniques that search for and detect changes within a variety of address-related datasets. These could include, for example, administrative records, parcel data, real estate listings, and structures and addresses detected through street-level imagery collection. Sources could be located through a web crawler and once files are located, either downloaded or "scraped" for appropriate information providing indicators of new housing units and, ideally, their addresses and other attributes. This process could apply to search and acquisition of other geospatial data;

(c) Acquired geospatial data could then be validated through an automated comparison of data across multiple sources. This could be a comparison of the number of structures detected in an automated imagery review process with the number of addresses in the DSF, tribal, state, and local government address lists, or other address data sources. It could be a comparison of names attached to road features from one source to another.

Successful implementation of this approach requires spatial comparability between files and confidence and assurance that the geospatial data in each dataset used in the comparison and validation process are in their correct geographic locations. It is not enough to simply match lists of addresses—validation of data in MAF/TIGER ultimately requires assuring that data are complete and the correct locations;

(d) Our experience this decade with both IOAC and the GSS Program has demonstrated that, while much of the work to update and validate data in the MAF/TIGER database can be accomplished in the office, there are parts of the nation in which geospatial data sources are not available or are updated with sufficient frequency to meet the Census Bureau's needs. In these areas, fieldwork is still necessary to detect change, acquire updates, and validate the accuracy and completeness of geospatial data in the MAF/TIGER database. These areas often contain housing situations that may not be tracked or covered by the data sources used to update the MAF, such as houses subdivided into multiple residential units, garage apartments, informal and "self-help" housing in jurisdictions that do not issue building permits, residential structures that do not receive direct mail delivery and, therefore, are not present on the DSF, and illegal conversions of spaces for residential purposes. The atypical and complex nature of such residential settings requires a more professionalized field workforce than the Census Bureau traditionally has employed for canvassing and collection of geospatial data in the field, one trained specifically in landscape observation, use of imagery and other geospatial data while in the field, and, ideally, possessing deep knowledge of and familiar with the local landscape gained through continuous work in the field.

V. Conclusion

23. Given the inherent limitations of geospatial data sources no matter how accurate and authoritative a source may be, change detection processes must be accompanied by validation processes backed up by measures and indicators of quality and completeness, not only of the data within the MAF/TIGER database, but also the data sources used for change detection and data update.