HANDHELD DATA COLLECTION AND ITS EFFECTS ON MAPPING*

E. Alkire

U.S. Census Bureau, 4700 Silver Hill Road, Washington, D.C. 20233-7400 – elise.alkire@census.gov

KEY WORDS: 2010 census, automated mapping, TIGER, GPS, map spot, map scale

ABSTRACT:

This paper discusses the many challenges faced when designing and developing the automated mapping system needed to produce nearly 15 million maps to support 2010 census data collection operations. The U.S. Census Bureau used handheld technology to collect and update housing unit locations, addresses, and street features in its address canvassing operation for the 2010 census. The information collected was reflected on maps produced for subsequent field data collection operations. While the use of handheld technology substantially improved the effectiveness and efficiency of field operations, it posed a number of challenges for the development and production of maps for 6.7 million blocks, each map displaying housing unit locations that had been collected during address canvassing. Map design and development occurred before and concurrently with the uploading of housing unit location data; as a result, cartographic design, scaling, and insetting decisions had to be made based on a limited set of data from the 2008 Dress Rehearsal, and general knowledge of data in the MAF/TIGER database. Because of the short amount of time available for map production between processing of housing unit location data and the next field operation, the automated map production system had to evaluate the data to be mapped, apply scaling and insetting algorithms, determine appropriate scale and level of insetting, and identify potential display problems for review—within minutes.

* This paper reports the results of work undertaken by U.S. Census Bureau staff. This paper is released to inform interested parties of research and to encourage discussion. Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

1. INTRODUCTION

The Census Bureau's Geography Division has historically tried to employ the most current and cutting-edge technology to assist in conducting its operations (Trainor 1990). New for the 2010 census was the use of handheld computers equipped with GPS technology to collect housing unit location and feature data, which occurred during the Address Canvassing field operation. These data were subsequently displayed on maps made to support the next major field operation, beginning less than four months after the completion of Address Canvassing.

The use of handheld devices enabled efficient completion of a nation-wide data verification and update operation, and in most cases helped improve the precision and accuracy of data in the Census Bureau's MAF/TIGER database. However, as with any new technology, it also posed several challenges, particularly when it came to mapping the data that were collected. New types of data errors and anomalies, and the task of uploading and preparing for mapping such a large volume of data in time to meet tight map production schedules, presented new and unique challenges to the Geography Division.

This paper describes current and past methods of data collection, the difficulties encountered when collecting data using handheld devices, the process of integrating data into the database and preparing for use on maps, and the integral role that handheld-collected data played in allowing the Census Bureau's map production software to determine appropriate scales for the millions of highly detailed maps created to locate individual housing units across the entire nation.

2. DATA COLLECTION METHODS

While the use of handheld devices to collect housing unit and feature data was new for the 2010 census, the collection of these types of data is important in any decennial census. To understand how the use of handheld computers drastically changed this process, it is helpful to review how similar data collection processes worked in the past.

2.1 Census 2000 Block Canvassing

During the 2000 census, the Block Canvassing operation was used to verify and record housing unit locations in the field. It was limited to rural areas, and the verification of housing unit locations was done by hand, with Census field employees called "listers" marking and labeling housing units on paper maps. The data were then digitally captured at the Census National Processing Center (NPC).

This method did not allow for a great deal of precision, accuracy, or completeness. While the two-step process of marking housing unit locations by hand and then digitally capturing them helped to normalize the data, it also had the potential to introduce even more inaccuracy.

At the time, this level of detail was on par with the other data contained within the Census Bureau's spatial database, then called TIGER. In 2000, the network of features stored in TIGER (such as roads, hydrography, etc.) was not accurate enough to support the inclusion of precise GPS coordinates, so it would not have made sense to use an extremely precise method of data collection for the housing units. During the mid-2000s, improvements were made to increase the spatial accuracy of the base feature network so that it is now accurate enough to support GPS data. These improvements laid the groundwork for the use of GPS-enabled handheld devices to collect housing unit data and feature updates in preparation for the 2010 census and beyond.

2.2 2010 Census Address Canvassing

The 2010 equivalent of the Block Canvassing field operation during the 2000 census was the Address Canvassing field operation. There were several important differences between this operation and Block Canvassing, beginning with the use of handheld devices to collect data instead of the two-step process of manually recording data and then digitizing later. The handheld devices allowed data to be captured, digitized, and uploaded to the database, now called MAF/TIGER (Spahlinger 2007), by an individual field employee in one streamlined process. In addition, this more efficient method of data collection and transfer made it possible to conduct Address Canvassing nationally, rather than limiting it to certain areas.

2.2.1 Canvassing Procedure: During the Address Canvassing operation, listers used handheld computers equipped with GPS technology to physically locate each building that contained individual residences, or housing units, on the ground. They then geocoded each unit to associate its address with the correct geographic location. A lister was responsible for capturing the housing units in one or more collection blocks[*] using a designated procedure. Beginning at a "convenient corner" of the block, the lister walked clockwise around it, collecting a point called a "map spot" for every address that contained one or more housing units (United States Census Bureau 2008). When the lister encountered streets or housing units on the ground that weren't on the map, or when features or housing units on the map were not in that location on the ground, he or she updated the map on the handheld device.

2.2.2 Map Spot Collection Procedure: In addition to the procedure for canvassing the blocks, listers were required to follow a specific procedure for collecting the individual map spots. As briefly mentioned earlier, a map spot refers to a point displayed on a map that represents a structure in the Census Bureau's Master Address File (MAF). The procedure for map spot collection was, for single family units, to collect a map spot at or near the main entrance. If a strong GPS signal was not available at that location, the lister was to collect it from other entrances as follows, listed in order of preference: side door, back door, garage door, or driveway. For multi-unit structures, the lister was to collect one map spot only for the first unit at the main door to the building (United States Census Bureau 2008). GPS technology vastly improved the precision of the map spot coordinates by several decimal places over the 2000 census, and improved the overall accuracy and completeness of the data. However, this accuracy was dependent on two factors that could not always be guaranteed: the strength and accuracy of the GPS signal, and a lister that correctly followed the procedures given to them. Data inaccuracies that most affected mapping were caused by failure of the lister to follow map spot collection procedures. The types of data inaccuracies that resulted from this will be discussed in the next section.

3. IMPACT OF USER ERROR ON DATA

Failure to follow the procedures outlined above was usually due to user error. It was the listers' first experience with handheld devices and they had little time to study the 500-page instruction manual that explained how to use them, so there was a high probability that mistakes would occur.

Figure 1 shows the configurations of map spot data that resulted from errors in handheld data collection. Common errors consisted of entering each unit in a multi-unit structure as a separate map spot (left), or collecting map spots for separate structures from the same ground location (right). Both resulted in groups of several of map spots with very similar coordinates.

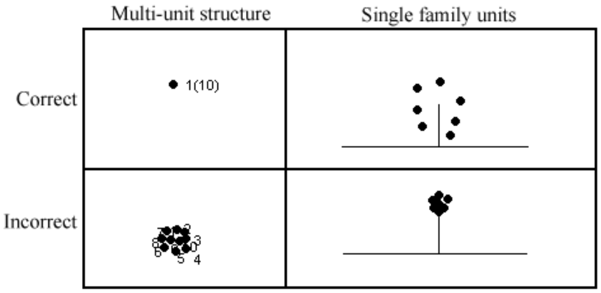

Figure 1. Correctly and incorrectly collected map spots for different types of housing

It was apparent that many of these errors were a result of misunderstanding on the part of the lister. Among urban collection blocks with several multi-unit apartment buildings, high occurrences of error were found at the lister's starting position. Further along their route, there were few or no errors, indicating that the user realized their mistake and corrected it. Figure 2 illustrates how this scenario appeared on maps.

[1*] A collection block is a geographic area that is bounded by physical features, non-visible features, and/or certain legal boundaries, and is the smallest building block that is used to support data collection during a decennial census.

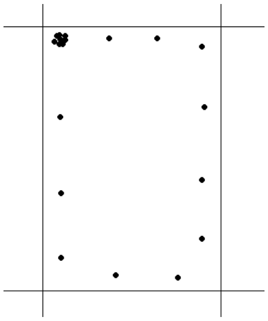

Figure 2. Bird's-eye view of a collection block. Map spots are clustered at the lister's starting position at the upper left, but are properly spaced clockwise of the starting position.

Despite occurrences of incorrectly collected map spots, it should be noted that the vast majority of map spot data collected using the handheld devices were accurate and free of user error. Nonetheless, all handheld-collected data were uploaded into the MAF/TIGER database and prepared for use in mapping. Although operations to resolve errors in the data were planned for later in the census cycle, exceedingly tight map production and field operations schedules did not allow them to take place until after the field mapping operations that immediately followed data collection. These mapping operations and the challenges faced in using the newly-collected data from the handheld devices are described in the remaining sections.

4. PREPARING THE DATA FOR MAPPING

As previously mentioned, data were collected during the Address Canvassing field operation, which took place from March through July of 2009. Almost immediately after the start of data collection, the process of uploading and preparing the data for mapping began so that the maps could be completed in time for the next field operation in October. From the beginning, the Geography Division was aware of how tight the schedule was at every step of the process, so the plan for mapping took it into account.

Data from the handhelds were uploaded to MAF/TIGER in real time as they were collected. However, using the live database as the source for mapping was not optimal for product creation. The usual procedure is to take snapshots of the database at a particular point in time, called "benchmarks". About 150 benchmarks were created, one for each of the roughly 150 areas into which the Census Bureau divided the country for the purpose of managing data collection during Address Canvassing. These areas were based on population, and could range in size from smaller than a city to larger than a state. A benchmark of one of these areas could not be completed until all the data within it were uploaded. Since this required a great deal of processing time, it was expected that map production would have to begin before all benchmarks were complete.

5. AUTOMATED MAPPING WITH HANDHELD DATA

The first set of maps that used the handheld-collected data from Address Canvassing was produced to support the Group Quarters Validation (GQV) field operation, which was set to begin in mid-October, just months after the July completion of Address Canvassing. For this operation, highly detailed large-scale paper maps of each collection block in the country needed to be produced to assist enumerators in the field. A unique map for each of the 6.7 million individual collection blocks was produced using the Geography Division's in-house automated mapping software. In order to meet the early October deadline, map production had to begin in late July. In total, 7.8 million 11x17 individual map sheets were successfully produced in the allotted 2.5-month production period. Each of these maps displayed and labeled every handheld-collected map spot that was located within the map's subject block, so the scale of each map had to be large enough to show them in detail.

5.1 Census Automated Map Production System

The Geography Division uses software that was developed in-house to do the majority of its mapping, called the Census Automated Map Production System, or CAMPS. It uses Oracle database tables containing hundreds of different parameters that provide information about the map, such as content, symbolization, scaling, inset determination, text placement, and marginalia. When the software is provided a complete set of parameters, and a list of areas to map, it produces a unique map of each area in an automated fashion (Spahlinger 2007). In this way, the software is capable of producing unique maps for a large number of areas more efficiently than could any office of human cartographers creating each map interactively.

A complete set of parameters is referred to as a "project." Several projects were needed to complete production of the individual block maps for GQV. The main parameters that differed between projects were the ones that controlled scaling and insetting, the reasons for which will be expanded upon in the next few sections.

5.2 Development of Initial Map Design and Scaling Parameters

The handheld-collected map spot data were not essential for the development of all of the map design components—there were older data available with which to test and develop the basic map design. Handheld data were necessary to determine map scales because the scales needed to be based on the density and distribution of the current map spots that were to be displayed on the maps. Scaling parameters could not be finalized until these data were available for mapping.

By the time the maps were scheduled to begin production, only two small areas—Washington, DC and Hawaii—had completed the benchmarking process. As a result, these areas contained the only data available for mapping. Scaling parameters were finalized using these areas, even though they represent only a fraction of the geographic diversity found across the United States. The same set of scaling parameters developed using this limited data set had to be used to map blocks throughout the nation due to time constraints, as anticipated. It was expected that scales appropriate for urban settings, like Washington, DC, and the higher density settlement patterns in Hawaii would not always work in other areas that had lower density settlement patterns, were more rural, or otherwise different.

Had the entire nation's worth of handheld data been available for testing prior to the start of map production, spatial analysis could have been performed and blocks could have been pre-assigned to run on projects according to map spot distribution. For example, one project could have been developed for blocks with dense but even distribution of map spots, a scenario that one might expect to find in a city. Another project could have been created for blocks with less dense but uneven distributions of map spots, which might be found in more suburban areas. Since time constraints prohibited using those methods, the strategy instead was to run all the blocks on the main project initially, and the ones that failed the software's internal quality control checks were rerun on another project with different scaling parameters.

A few alternate projects were developed early to resolve common map issues, but since the software had never used GPS-collected data for scaling before, many of the problems that had to be accounted for could not be predicted. This, along with the rolling basis by which data for new areas became available for mapping, necessitated continuous development of new projects until the end of map production.

5.3 Determining Map Scales Using Handheld Data

5.3.1 Scaling to Point Data in CAMPS: The CAMPS software uses a critical dataset called the "analysis layer" to automatically determine the scale for each map. Because the primary purpose of the block maps was to allow enumerators to locate individual housing units, the analysis layer for these maps was the set of handheld-collected map spots. Since this was point data, scaling was based on the proximity of map spots to each other. A major requirement for the block maps was that the label for each map spot could be read by the enumerator and matched to a corresponding address list, so it was crucial that each map had a scale that placed map spots far enough apart so that the label of one did not obscure the labels of other map spots nearby.

It is important to note that while this set of data is referred to as a "layer," it is not the same as a layer in some commercial mapping software applications that can be turned on or off. Instead, it is topologically integrated and elements within the data set are "aware" of one another. This awareness is stored in the form of a nearest neighbor attribute—the distance on the ground between a map spot and the one closest to it. Without this attribute, scaling based on feature proximity would not be possible.

In addition to an analysis layer with a nearest neighbor attribute, CAMPS requires two additional inputs to scale to point data: the minimum distance allowed between points on the map, and the percentage of the total points on the map sheet that had to meet this minimum distance. The minimum distance between the points on the map is determined by the cartographer with the intention of keeping the labels legible.

For example, assume that the minimum distance apart on the map that any two map spots can be and still have their labels be legible is 1/8th of an inch. The nearest neighbor attribute is stored in ground units, in this case meters, in the database. In order to determine what scale the map needs to have in order to meet the minimum distance requirement, CAMPS needs to find the minimum distance in ground units that any two map spots in the block are apart, convert that to the specified minimum distance in page inches, and apply the same scale to the entire map area (Figure 3). In this example, if the shortest distance on the ground between any two map spots in a given collection block is 3 meters, the scale that will ensure that those map spots are no less than 1/8th of an inch apart on the map of that block is about 1:1500, a large but not uncommon scale for many of these maps.

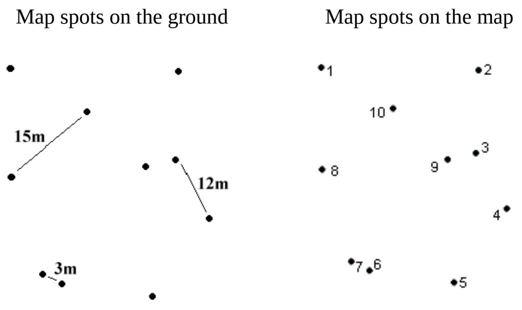

Figure 3. Map spots as they would appear on the ground and in the database are on the left, with the nearest neighbor distance in meters shown for select map spots. On the right, the same group of map spots as they would appear on a map. The distance in page units between map spots 6 and 7 would be 1/8th of an inch.

The final scaling parameter tells CAMPS what percentage of the analysis layer has to meet the minimum distance criterion explained above. Percentages less than 100 are allowed because the assumption is that the remaining elements will either be shown at larger scales on inset sheets, or will have leaders or other text placement strategies applied to their labels. This value is set separately for inset sheets; a higher value is needed because as the sheet is already an inset at a larger scale than the rest of the map, text placement strategies are the only option to improve legibility.

5.3.2 Adjustments to Scaling Parameters: The addition of the newly collected and extremely precise map spot data from the handheld devices added a new level of complexity and uncertainty to this already complicated scaling process. CAMPS makes certain assumptions when using an analysis layer for scaling, the most fundamental of which is that the data in the analysis layer are correct and accurate; that is, the data as they are stored in the database accurately represent the corresponding phenomena on the ground. As discussed earlier, this was not always the case with the map spot data collected during Address Canvassing. As a result, certain adjustments to scaling parameters were necessary to account for inaccuracies in the handheld data. Almost all of these adjustments required separate projects to be deployed with slightly different scaling parameters.

The majority of the adjustments were made to the two scaling parameters discussed above: the minimum distance allowed between map spots and the percent of map spots that had to meet this criteria. Projects with high minimum distance values were necessary to accommodate clusters of map spots (Figure 4). Adjustments to the percent of map spots evaluated were also needed because of how close together several handheld-collected map spots were (some as close as 8 centimeters apart on the ground!). Even evaluating 80 percent of the map spots on an inset may have left some labels illegible. As a result, projects that evaluated close to 100 percent of the map spots on a sheet had to be created in order for all to be labeled clearly.

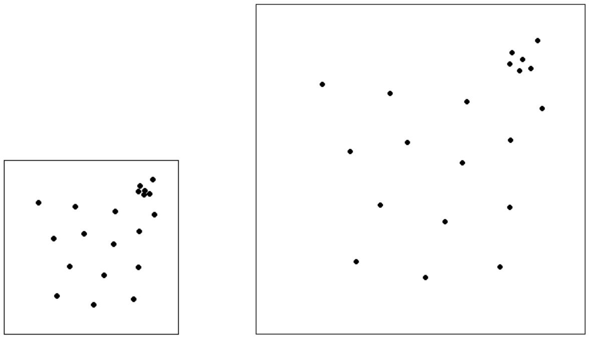

Figure 4. On the left, a group of map spots with the original scaling parameters—a low minimum distance requirement and less than 100 percent of the map spots evaluated. On the right, the same map spots with adjusted scaling parameters—a greater minimum distance value and close to 100 percent of map spots evaluated. (Map spot labels left off for clarity.)

A question that might come to mind here is why the main set of scaling parameters was not adjusted instead of making various separate parameter sets. There were two reasons why this would not have been an ideal solution. The first was that over 99 percent of the 6.7 million unique maps that were produced using the main project passed the internal quality control checks in CAMPS, indicating that those scaling parameters were appropriate for the majority of blocks mapped. The second reason that scaling parameters for a given project, especially the main project, were kept the same once it was deployed was that changing any of the scaling parameters mid-production could have had a cascading effect on all subsequent maps that could not be predicted. Because these parameters would be applied to millions of diverse geographic areas, a change in scaling that improved the appearance of one map risked making another map worse. Due to the sheer volume of maps being produced during a period of only a few months, it would have been impossible to visually inspect even a small percentage of the maps to ensure that any change did not have an adverse effect.

6. CONCLUSIONS

The main challenge that the use of handheld data posed to mapping involved errors in map spot collection that resulted in either too many points collected for a structure, or points for separate structures collected from the same ground location. Because some map spots were collected so close to each other on the ground, when mapped their labels overlapped even at "normal" inset scales, which for the block maps were already very large. In order to make the labels legible, projects had to be developed that allowed unthinkably large inset scales, against the better judgment of the cartographers designing the maps. Scales as large as 1:50, and in some cases as large as 1:7, were often required on inset sheets to clearly display badly clustered map spot data. With the correct parameters set, the CAMPS software is capable of producing maps at almost any scale, but setting it up to do so forced the cartographers who designed the projects to consider the question, at what scale does a map stop being useful? The result was that some maps were produced that were known not to be ideal as navigational tools, but met the legibility requirements nonetheless.

Despite the challenges of using handheld-collected data, the GPS-enabled devices allowed the Census Bureau to get data for the entire nation very quickly, considering the scope and volume of the data that had to be collected. They also facilitated, for the first time in Census history, nationwide data collected in one decennial operation to be used in the very next one. Even though mapping it was difficult at times, the speed at which handheld data could be collected and transferred allowed enough time for their inclusion on the millions of maps produced for the next Census operation. The use of handheld devices in data collection, along with versatile software and innovative methods of data processing, allowed record numbers of usable maps to be created in a very short amount of time.

ACKNOWLEDGEMENTS

Staff in the U.S. Census Bureau's Geography and Field Divisions supplied a great deal of the information contained within this paper. In particular, Linda Orsini in the Geography Division's Cartographic Products Branch provided knowledge and feedback throughout the research process. Staff from the Cartographic Products Branch provided feedback on the final paper as well. In addition, resources were contributed by the Geographic Areas Branch in the Geography Division and the Geographic Support Branch in the Field Division.

REFERENCES

Spahlinger, S., 2007. Proceedings from the XXIII International Cartographic Conference "The U.S. Census Bureau's Cartographic System for Supporting the 2010 Decennial Census", Moscow, Russia. http://icaci.org/documents/ICC_proceedings/ICC2007/html/Proceedings.htm (accessed 8 Sep. 2010)

Trainor, T., 1990. Fully automated cartography: A major transition at the Census Bureau. *Cartography and Geographic Information Systems*, 17(1), pp. 27-38.

United States Census Bureau, Field Division. 2008. Address Canvassing Lister Manual. Washington, DC, US