STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working Paper No. 25
English only

Work Session on Geographical Information Systems
(Brighton, United Kingdom, 22-25 September 1997)

Item (iv) of the provisional agenda

THE USE OF THE TIGER/MAF SYSTEM IN CENSUS 2000 INTEGRATED COVERAGE MEASUREMENT

by

Submitted by U.S. Bureau of the Census[1]

I. THE WHAT AND WHY OF INTEGRATED COVERAGE MEASUREMENT (ICM)

[1] Prepared by Carl S. Hantman.

1. For the 2000 census, the Census Bureau is planning an Integrated Coverage Measurement (ICM) program that would use statistical methodology to improve census results for both the total population and for the various component groups of that population. Instead of taking a survey after the census to determine undercount as was done for the 1990 census, the Bureau would include the measurement of coverage accuracy in the census taking process itself. Estimates of how many people and housing units were missed or counted more than once would be available for incorporation into the final census results for a single and more accurate set of numbers.

2. The proposed ICM uses several interrelated field and processing operations designed to develop an independent estimate of persons and housing units for use in producing the final census results. The Census Bureau selects a sample of ICM block clusters (one or more adjacent blocks) for intensive address listing and household interviewing. These 25,000 clusters contain the approximately 750,000 housing units needed to achieve the desired reliability for estimation purposes. Specially trained field representatives (FR's) systematically canvass these sample block clusters using census maps and listing books to record the addresses of all living quarters there. These addresses are compared to the addresses obtained through the more universal census methods to produce an enhanced list of addresses for the sample. Field representatives then interview persons at these housing units to obtain the data that will be compared to the census data obtained for these blocks. The results of this comparison would be applied nationally to obtain the final census numbers.

3. The ICM program has multiple needs that can be effectively met by a reliable and efficient GIS. One of the principle needs is the delineation of block clusters for sample selection. Statistical and operational requirements govern the block cluster definition process. The number of housing units in a cluster is a significant concern. Special sampling procedures are used for clusters having few or no housing units. Other special procedures are applied to clusters having large numbers of living quarters. The correct identification of such clusters is thus important for accurate sampling and weighting.

4. The data collection process also places extensive requirements on block cluster formation. The degree of difficulty encountered by the field representative in accurately canvassing a cluster is determined to a great extent by the nature of the cluster. Clusters should ideally have boundaries that can be readily identified by the person in the field so that housing units will not be erroneously included or excluded from the cluster. In addition, where possible, block clusters should not have "holes" which could increase the likelihood that the FR might mistakenly include housing units which do not belong in the cluster. Multi-block clusters should consist of adjacent blocks to avoid unnecessary and costly travel between blocks which are geographically separated. Similarly, to the extent possible, clusters should not have internal physical features that make it difficult for the FR to traverse the cluster. For example, we would not want a cluster that has a limited access highway running through it, which could require the field representative to drive an excessive distance to complete the canvassing of the cluster. Clusters that cover too large an area can also require excessive time and cost to complete the listing.

5. A second principle GIS need for ICM is the production of maps for the field listing operation. When ICM staff go into the field to create the independent address list, they need maps to help them get to their assigned clusters. They also need maps to use in the canvassing operation. These cluster maps allow the FR to recognize the limits of the cluster. The FR will also mark on the cluster map the location of each living quarters found in the cluster. Subsequent matching processes use these map spots as an aid in determining housing unit coverage.

6. To provide the geographic support required by numerous programs, including the ICM, the Census Bureau has developed a comprehensive geographic information system, the TIGER/MAF System.

II. THE TIGER/MAF SYSTEM

7. For the 2000 census, as well as its other demographic data collection activities, the Census Bureau is improving and expanding its geographic support system. It is building on its 1990 census address list, the Address Control File (ACF), to create a permanent and continuously maintained Master Address File (MAF). The MAF is a national inventory of living quarters whose addresses will be continually linked to the TIGER database. The two databases, along with their associated software, form the TIGER/MAF System. The effective development and application of the TIGER/MAF system is critical to the success of the 2000 Census.

8. The TIGER/MAF system is designed to correct specific problems encountered in the 1990 census. One significant failing was the lack of consistency between the ACF and TIGER. At the beginning of the census process, the ACF reflected the geographic structure provided by TIGER. However, as census operations continued and the files were updated and modified, discrepancies developed between these two data sources. There was no system to ensure that corrections applied to one were consistently and promptly applied to the other. Not surprisingly, products such as maps and address lists generated from these files reflected these inconsistencies. A second shortcoming of the 1990 system was that once decennial census operations were completed, there was no system to update the address file and TIGER to reflect the continuously expanding inventory of housing units and the continuously changing network of streets and geographic entities.

9. To meet these needs, the Census Bureau decided to build and maintain the MAF and to continually link it to the TIGER database. The MAF is being built by matching and merging two national address files, the ACF and the United States Postal Service's (USPS) Delivery Sequence File (DSF). We will keep the MAF current through cooperative agreements that the Census Bureau has entered into with the USPS and with numerous local governments in the U.S. These agreements provide for these agencies to share their address lists with the Census Bureau. The USPS will provide the DSF on a periodic basis to update the MAF. Local governments will provide their address lists to fill any holes in MAF housing unit coverage. In return, the Census Bureau will provide the relevant portions of the MAF to the local governments for their review. These partnerships should significantly improve the comprehensiveness and timeliness of the Census Bureau's housing inventory.

10. We are ensuring the consistency of TIGER and the MAF through a computerized geocoding process. Geocoding is the process of geographically locating an address. The Census Bureau has a fundamental requirement to tabulate data by individual census block, of which there are over seven million nationwide. The automated geocoding software assigns city style addresses in the Master Address File to their correct census blocks utilizing the street, address range, and USPS postal code information stored in the TIGER database. The process serves two goals. First, when a MAF address is geocoded to a census block in TIGER, the associated TIGER geographic codes are assigned to the address. These codes include not only the census block, but any other relevant census political or statistical geography associated with that block. The successful assignment of the TIGER geocode to the MAF address establishes the link between the two databases and confirms that the data in the two files is consistent. Secondly, when the software fails to establish this link, it is an indication that there is probably missing or erroneous data in TIGER. Ungeocoded addresses initiate research into resolving the problem, generally resulting in the addition of a new street or address range to

TIGER. The TIGER/MAF system will thus significantly improve the comprehensiveness of the housing inventory and the accurate assignment of geographic codes to these units.

III. USE OF TIGER/MAF FOR CLUSTER FORMATION

11. In order to use the TIGER/MAF system to establish the universe for ICM sample selection, the Census Bureau had to decide the geographic level at which sampling would occur. Two possibilities were considered; block faces and census blocks. Statistical, operational, and geographic factors were considered. Ultimately, although using block faces would reduce the number of housing units required for the sample, the operational and geographic advantages of using the census block outweighed these considerations. The geographic factors are important when considering the use of a GIS for this type of application. Most significant of these factors was the need for the sample unit to be readily identifiable both in the TIGER database and in the field. There were complex issues related to developing a definition of a block face and applying it to TIGER. The census block is the geographic atomic unit of TIGER. Each of the more than 7 million census blocks is clearly and unambiguously defined in the TIGER database. In most cases, census staff can fairly easily discern census blocks in the field as well. This was not the case for block faces. A block face could be considered one side of a street between two intersecting streets or one side of a street between any intersecting features. It could also be defined as unique combination of a census block with a street name. It might include both sides of a street or only one side. In the field, perception of a block face is fairly straight forward in a grid-like urban street pattern, but it becomes much more difficult in an area where streets are more irregular and crossed by numerous non-street features. In the TIGER database, the concept of block face does not exist. Although we could have defined a block face in terms of TIGER structure and features, implementing this definition would have required additional software development. These factors contributed to the decision to use census blocks as the sample unit.

12. The ICM block clustering software utilizes the efficiency of the TIGER topological structure and the detailed data stored in both TIGER and the MAF to automatically determine the extent of each cluster for sample selection. Starting with a county-based TIGER file partition, the clustering software collects data about each 2000 collection block in the county. From TIGER, we can determine the area of each block and the type of features that form its boundary. We know which of the bounding segments are roads (and what category of road, including primary highways, secondary roads, and local streets), rivers, railroads, power lines, or invisible features such as political boundaries. We also compute and store the length of each bounding feature. TIGER can also tell us which blocks are adjacent to each other. From the MAF, we can retrieve the most recent information we have on the number of housing units in the block.

13. The goal of the block clustering software is to form clusters that reflect, to the greatest extent possible, acceptable numbers of housing units and acceptable feature networks, area, and shape. The object is to have the clusters conform to the statistical and data collection requirements of the ICM program. To accomplish this goal, the software must analyze the data pertaining to each block and decide if the block can stand alone as a cluster or if it needs to be combined with adjacent blocks to form a cluster that meets the ICM criteria.

14. We initially define each block that has fewer than three housing units as a unique cluster. As mentioned above, these blocks with few or no housing units are sampled at a different rate than other clusters. The software then assigns each block that has more than 100 housing units to a unique cluster. These are the blocks that have more than the ideal number of housing units and will require subsampling. We then use the bounding segment data from TIGER to find any remaining blocks that do not share a common side with any other unclustered block. These blocks will also each become a unique cluster because we do not

want to form any discontiguous block clusters by combining blocks that are not adjacent to each other. We then begin the process of combining blocks into clusters according to the ICM clustering specifications, with the goal of minimizing the number of clusters that fail to meet these criteria.

15. First we attempt to combine any block that shares an invisible boundary with another block. Invisible boundaries create problems for field representatives because it is often difficult to determine whether a housing unit along these boundaries is in or out of the cluster. Following this step in the process, our primary consideration becomes the number of housing units in the block. Clusters with between four and nine housing units are undesirable for ICM and the software attempts to combine them with adjacent clusters wherever possible. While examining adjacent block clusters for possible combination, we consider numerous factors in selecting the "best" block or blocks for cluster formation. The software examines the common bounding segment(s) of each adjacent cluster. We do not cross areal water such as lakes or wide rivers, nor do we cross limited access highways. We will not combine with a block or blocks that would create a cluster with more than 100 housing units. We will not cross streams or railroads except where there are no other adjacent blocks available for clustering. The software also attempts to avoid creating clusters that are larger than 15 square miles. We attempt to make clusters that are reasonably compact by using a "shape index" derived by dividing the area of the cluster by its perimeter. Finally, we attempt to ensure that clusters do not have any gaps or holes; that is, a block cluster should not completely surround another block cluster. As explained above, such holes also create difficulties for field representatives.

16. We can efficiently and successfully apply these clustering criteria because the structure of TIGER allows use to maintain the relationship of census blocks to their component features and to other blocks. The integration of TIGER and the MAF allow us to effectively apply current housing unit data to the clustering process. The continuous maintenance and updating of these two databases ensures that the best available data will be used for cluster formation. The TIGER/MAF system can thus produce a reliable universe of block clusters that meet the standards for ICM sample selection.

IV. TIGER MAPS

17. Once the sample block clusters have been selected, ICM field representatives visit these areas to create the independent address list for the coverage measurement process. They need maps to help them get to their assigned clusters and maps to use in the canvassing operation. A total of over 50,000 unique map sheets must be produced for ICM. To meet this and many other field mapping needs, the Census Bureau has developed the Single MIM-Based Integrated Mapping System (SMIMS).

18. The SMIMS is a user-driven map production system that allows timely production and delivery of large volumes of diverse field maps. For the ICM, staff in 12 Census Bureau regional offices will use SMIMS to produce cluster and cluster locator maps for the 25,000 sample block clusters. Using a WEB-based interface, running on the Census Bureau's intranet, staff will select the types of maps they need and the areas for which they need them. The system then reads a series of cartographic definition tables that reflect the specifications for the particular map series. These specification include the types of features to be displayed on the maps and how these features will be symbolized. The mapping software then accesses the TIGER database on local UNIX workstations or servers to retrieve the definition of the block cluster and to extract the cartographic data required for the particular map type. It retrieves, in separate layers, the areal features, linear features, political boundaries, and landmarks needed for the particular map. The mapping software determines the optimum map scale and decides the best location for all text labels. The individual layers are then integrated to form the

complete map, which is produced as a Map Image Metafile (MIM), a device-independent, graphic file format. The MIMs are stored on CD-ROMs for production on laser printers at the census field office.

19. The SMIMS represents a significant improvement over previous methods of large volume map production used at the Census Bureau. For the previous decennial census, map file production was centralized at a single computer installation. Map files were stored on computer tapes and shipped to regional plotting facilities. Once plotted, the maps needed to be cut and shipped to the local census office for use in the field. The process was time consuming and created production bottlenecks on the computer and at the plotting sites. Under the new system, map production is controlled at the regional level, where each office can produce the maps needed for the operations and areas for which they are responsible. Instead of relying on expensive and specialized plotting devices, ordinary office printers can now be used to generate the paper maps in the field offices where they are needed. Distributed processing results in greater local control of the system, reducing bottlenecks and getting the maps into the hands of field staff more quickly and efficiently.