An Automated Procedure for Forming Contiguous Sampling Units for Area Probability Samples

Bryce Johnson, Jill Montaquila, Andrew Heller, Westat
Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Abstract

In area probability samples, the primary sampling units (PSUs) are often counties or groups of contiguous counties. The secondary sampling units (SSUs) or segments are often constructed using census blocks, block groups, or tracts. Algorithms to form segments by grouping areas that are adjacent in numbering schemes (e.g., adjacent block numbers within tract, in a sorted list) sometimes result in discontiguous segments. This may be a problem if area measures (e.g., environmental data) are to be obtained or if multilevel modeling is to be used for analysis.

In this paper, we describe a study that motivated the development of an algorithm to optimize the configuration of segments based on census blocks. The algorithm is described and evaluated, and ideas for enhancements are discussed.

Keywords: TIGER, segments

1. Introduction

In many area probability samples, the primary sampling units (PSUs) are individual counties or groups of contiguous counties, and the secondary sampling units (SSUs) or segments are census blocks or groups of census blocks. When most of the individual counties are large enough to constitute PSUs (in terms of a measure of size, or MOS, appropriate for the survey) the remaining small counties are often grouped with adjacent small counties. In these cases the process of creating a sampling frame can be straightforward. However, in other cases the creation of PSUs is very challenging because many or most of the counties must be grouped. Automated techniques have been developed to make this process more efficient and can effectively handle cases where adjacent counties are joined to form PSUs (see Green, Chowdhury, and Krenzke, 2002). The formation of segments can be challenging for many of the same reasons; however, segment formation often involves a very large number of blocks (using Census 2000 block definitions, there are over 8 million blocks in the 3,141 counties in the U.S.). For the most part, census blocks are numbered in geographic sequence within census tracts, but there are a nontrivial number of exceptions. Thus, relying on numeric sorting alone to combine blocks will result in some segments that are not geographically contiguous.

For many surveys, it is not essential that the segments be geographically contiguous; segments with disjoint pieces located in relatively close proximity to each other may have a negligible or near-negligible effect on the resources required for data collection. However, for the survey that motivated the approach described in this paper, segments are required to be geographically contiguous and to have approximately uniform measure of size.

This paper describes an automated procedure developed by Westat to form contiguous segments. In Section 2, we discuss the algorithm used in this automated procedure. Section 3 presents the results of implementing the automated procedure with respect to several criteria. A summary of the algorithm's performance and plans for future enhancements are given in Section 4.

1.1 The Manual Approach

Initially segments were constructed by a manual approach. Using ArcView, a geographic information systems (GIS) software package, maps were created that delineated census blocks. Linked to these maps was a database containing the census block measure of size (the expected number of births in the block). Blocks were manually combined with the aim of achieving contiguous, uniformly sized segments. This manual process was quite labor-intensive, and it was often the case that a person had to abandon their work and restart the process because of an unlucky choice in the starting point.

In order to gain a clearer understanding of the process involved in forming contiguous segments, consider

Exhibit 1. The colors represent different segments composed of census blocks, with block measure of size displayed in the center of each block. The segments are further identified by the letters A through F and the segment MOS is identified in parentheses. The goal is to form segments composed of contiguous census blocks while minimizing variation in the segment measure of size. In this example, the target measure of size is 121.

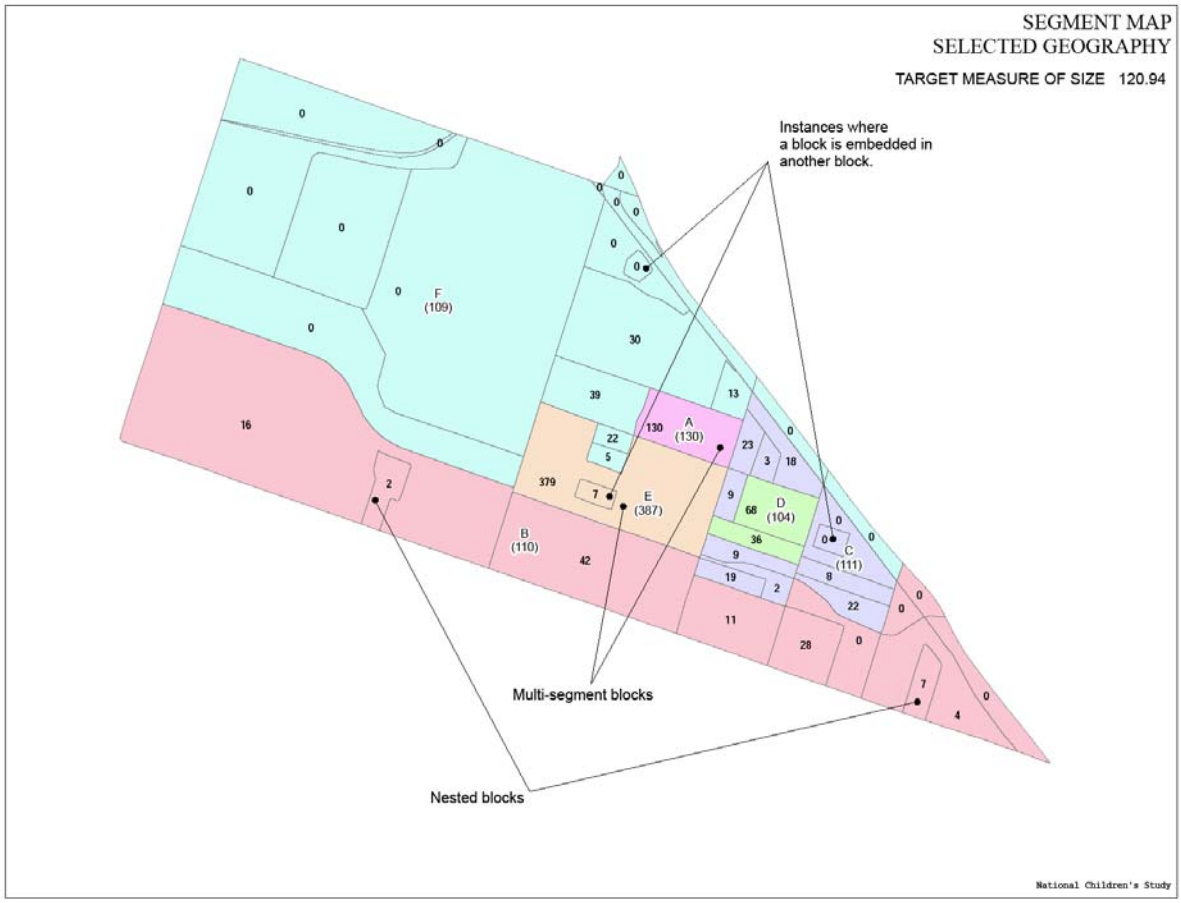

Exhibit 1. Segments in a selected geography

This example presents numerous challenges for both manual segment formation and an automated approach. Consider the variation in the geographic size of the blocks, their irregular shape, and variation in the block measure of size. Additional challenges include cases where a block is completely embedded within another block (e.g., Segment E, blocks with measure of size 379 and 7) and cases where a single block exceeds the target measure of size for the segment (Segment A), which we call a multi-segment block. The latter situation may cause drastic variation in the measure of size between segments. The constraint of contiguity in conjunction with the complexities of the census blocks, present a very challenging problem in minimizing variation in measure of size across segments.

2. An Automated Procedure for Forming Contiguous Segments and Comparison

An automated procedure was developed with a number of important goals in mind: creating segments composed of contiguous blocks, achieving minimal (or near minimal) variability in the segment measure of size, controlling the number of segments formed, and creating compact segments. The complexity involved in attempting to meet all of these goals simultaneously necessitated a computer-driven solution. An iterative process was used in order to create numerous solutions from which the best solution is selected.

The automated procedure begins with a number of input files and control parameters. An essential requirement of the automated procedure is a census block adjacency file, which specifies adjacencies between blocks. Other files and parameters specify the block measures of size, the target measure of size for the segments, the number of segments to be formed, and the number of iterations. Some blocks have a measure of size that exceeds the target measure of size (multi-segment blocks). These blocks are temporarily removed from the process and the number of segments to be formed in the remaining geographic area is determined.

One of the key ingredients to this process is to integrate a component of randomness, creating variability in the solutions among iterations. The vehicle for this random component is the selection of blocks that will serve as starting points, or seeds, from which segments will be formed. Once identified, blocks adjacent to the segment seeds (candidate blocks) may be added to the segment. In this way, each segment "grows" as new blocks are incorporated. However, the order in which the segments are "grown" is controlled, so as to achieve full coverage of the geographic area.

The candidate blocks are ranked with respect to three criteria, so as to select the block yielding the optimal solution. These criteria include number of adjacencies, distance, and block measure of size. Since the compactness of the segment is important to this process, having more adjacencies with a growing segment is given priority. We examine the number of shared borders of a candidate block with the segment that is being grown. The distance between the segment that is being grown and a candidate blocks is also evaluated and employed. Lastly, the measure of size of a candidate block is taken into account so as to reduce the variability in the segment measure of size.

Once all of the blocks have been assigned to an initial segment, a new process begins which transfers blocks between segments. The motivation for this step is to shuffle the blocks on the segment borders in order to decrease the variability in segment measure of size. This process continues until transfers no longer result in decreased variation.

The overall process yields a solution which is a function of the random seeds. We can control the number of solutions with an input parameter that determines the number of process iterations. Each iteration is associated with a new set of randomly selected seeds. The segments formed from the iteration yielding the lowest variation in the segment measure of size is selected as the best solution.

3. Performance Results

In the process of developing segments we created an initial set of contiguous segments for some of the PSUs using a manual approach. In some cases we are able to make comparisons between the manual approach and the automated approach with respect to the standard deviation of the segment MOS. The following discussion focuses on standard deviation as a key performance measure. Contiguity was also an important motivator in developing the automated procedure and was achieved consistently.

Table 1 gives the results for the three most rural PSUs among the seven PSUs for which data was available. Included in the table are the target MOS, the number of blocks (to illustrate the magnitude of this optimization problem), and the performance results as measured by the standard deviation of the segment MOS. The automated procedure performed extremely well, as evidenced by the small standard deviation relative to the target MOS. In the case of Rural_2, we were able to make a comparison to the manual approach. For this PSU the automated procedure outperformed the manual approach with respect to standard deviation.

Table 1. Performance results of the automated and manual approaches to segment formation in rural areas

	Large geography			
			Standard deviation of MOS	
PSU Name	**Target MOS**	**Number of blocks**	**Manual**	**Automated**
Rural_1	304	5,285	Not available	0.32
Rural_2	218	2,150	7.08	0.33
Rural_3	102	6,663	Not available	5.86

Each of the remaining PSUs was partitioned into geographic strata. Table 2 shows results for two of the geographic strata from each PSU. These strata were selected for presentation in this table to exhibit both good results and poor results (shaded). For the most urban PSUs we have results for the automated procedure with full coverage of the PSU (large geography) and a subset geographical area (small geography). In these three urban PSUs we used a two-stage sampling approach in which smaller geographies were created and then sampled. The small geographies were selected in a random process with probability proportionate to size. We did not examine the performance in the small geography setting for Urban_1 because we did not use a two-stage sampling approach in this PSU.

Table 2. Performance results of the automated and manual approaches to segment formation in urban areas

PSU Name	Stratum	Target MOS	Large geography			Small geography	
			Number of blocks	Standard deviation of MOS		Number of blocks	Standard deviation of MOS
				Manual	Automated		Automated
	10	186	1,377	3.92	0.81	*	*
Urban_1	6	169	952	3.42	8.07	*	*
	9	122	2,631	7.5	11.7	316	0.2
Urban_2	11	122	2,103	6.8	18.8	63	45.5
	4	83	721	3.8	9.8	275	0.4
Urban_3	14	86	1,049	2.9	4.3	29	24.9
	5	112	1,281	4.8	10.2	218	0.7
Urban_4	2	111	757	5.7	8.2	130	24.2

* Small geography statistics were unavailable because a single-stage sampling approach was used.

The reason for examining both large and small geographies was to contrast the performance of the automated procedure in different settings. We wanted to determine whether the algorithm performed differently in smaller sub-areas than in the area as a whole. In all cases the standard deviation reported excludes multi-segment blocks, as this helps in making meaningful comparisons.

In many cases the automated approach performed well with respect to the standard deviation of segment MOS for smaller geographies. However, in each of the three PSUs there was one stratum where this was not the case. Although only two strata are exhibited here per PSU, the remaining strata had good results in the small geography setting. For the three cases where the automated procedure performed poorly we investigated whether a manual approach would have yielded better results. We concluded that a manual approach could not improve on the automated solution. The automated procedure, like the manual approach, is limited by numerous constraints. Difficulties associated with the shape, size, placement, and number of blocks (as illustrated in Exhibit 1) are compounded in some small geographic areas.

4. Conclusion and Future Development

The automated procedure was successful in consistently creating segments composed of contiguous blocks. The effectiveness of the automated procedure in minimizing the variance of segment MOS was studied in several settings. The automated procedure performed well in large rural geographies, relatively poorly in large urban geographies, and well in small urban geographies. The examination of small urban geographies revealed that some geographic areas are more conducive to optimal performance results than others, and that an area with fewer blocks can result in poor results regardless of whether an automated or manual approach is used.

Several enhancements to the automated procedure are planned. We intend to develop new methodologies to address performance in large urban geographies which will give the automated procedure broader application. In many instances census blocks are not defined by visible boundaries (e.g., streets, bodies of water, etc.) and are instead defined by invisible boundaries (e.g., political boundaries, property lines). Methodologies are being developed that will attempt to avoid the use of invisible boundaries as segment

boundaries. In many studies it is useful either to maximize or minimize heterogeneity within segments. We will investigate methodologies which will utilize population characteristics in segment formation to achieve this goal.

Acknowledgments

The authors would like to thank Leyla Mohadjer and Andrea Piesse for their careful review and input.

References

Green, J., Chowdhury, S., and Krenzke, T. (2002). Developing Primary Sampling Unit (PSU) Formation Software. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 1239-1242.

Montaquila, J., Brick, M., and Curtin L. (2007). Methods for Sampling Households to Identify 100,000 Births. *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*

Software and Data Resources

ArcGIS is a GIS software product of ESRI.

Topologically Integrated Geographic Encoding and Referencing system (TIGER) is a product of the U.S. Census Bureau.