Challenges and Methods in Creating Secondary Sampling Units for Area Probability Samples

Bryce Johnson[1] and Jill Montaquila[1]
[1]Westat, 1650 Research Blvd., Rockville, MD 20876

Abstract
In area probability samples, secondary sampling units (SSUs), or segments, are often constructed using census blocks, block groups, or tracts. There are numerous challenges in forming segments that meet statistical, operational, and study goals. Creating segments that are composed of contiguous units is often a desirable and challenging goal that can be a challenging task, whether a manual or automated solution is used. A complex approach involves forming segments which are more heterogeneous based on one or several variables. There may also be some advantage in creating more compact segments. An important operational consideration is to create segments with visible boundaries (e.g., roads, highways, rivers), rather than invisible boundaries (boundaries with no physical demarcation; these are often municipal, county, or state boundaries.). All of these challenges are explored in this paper and a multifaceted solution is presented that relies on automated techniques.

Key Words: Segment, automation, TIGER, invisible boundaries, contiguous, heterogeneity

1. Introduction

In multi-stage 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. In this paper, we explore four challenges that may occur in creating secondary sampling units (segments) in area probability samples: invisible boundaries (2.1), contiguity of sub units within segments (2.2), compactness of segments (2.3), and heterogeneity (2.4). Although the topic of heterogeneity is not exclusively an area probability challenge, we discuss it here, and provide an approach that could easily be applied in the construction of other sampling frames. In each sub-section within section 2, we describe the challenges posed by each of these issues, and then discuss approaches developed to address these challenges. In section 3, we discuss a strategy that ties these four considerations together into a single cohesive method, facilitating the creation of a segment sampling frame.

A similar approach was developed and described by Green et al. (2002). However, the Green et al. (2002) work focuses on the formation of primary sampling units (PSUs) by combining counties, with an emphasis on contiguity, minimum measure of size requirements, and heterogeneity.

2. Approaches to Address Particular Segment Formation Issues

2.1 Invisible Boundaries
Boundaries that are not defined by any clear geographic feature are described as invisible boundaries, and are a common occurrence in census geography. Some invisible boundaries are necessary; for example, they may correspond to political boundaries, such as county or municipal boundaries. In such cases, recognition of the invisible boundary might be important for the construction of a segment sampling frame in the context of area probability samples. For example, if the primary sampling unit was the county, then segments must be formed within the sampled county. However, invisible boundaries frequently serve no useful purpose in area probability samples, and cause problems in survey implementation. Such problems include listing the proper housing units and collecting data from the intended sample area. Figure 1 is a satellite image of a census block, with the census boundary (in this case, an invisible boundary) delineated in red.

Figure 1. Invisible boundary example

One problem with a scenario such as that depicted in Figure 1 is in assigning the housing units to the appropriate segments that could be formed on either side of the census boundary. Avoiding the use of invisible boundaries can be facilitated in a number of ways. It may be possible, for example, to shift these boundaries after segments have been initially formed. Since this is a complex and labor intensive process, it may be advantageous to wait until segments have been selected, as opposed to modifying the sampling frame. However, this latter approach gives rise to the possibility of biases introduced by trying to make the segments meet operational goals. In order to avoid such bias it is important to establish unambiguous rules that preclude biases introduced by a practitioner who is adjusting the segment boundaries. An example of such a bias might be reluctance on the practitioners' part to increase the size of certain selected areas that are difficult to operate within (e.g., areas in which low response rates are anticipated), but an eagerness to increase the size of areas that are operationally favourable. To avoid introducing bias, the rules should be applied consistently regardless of which side of the invisible boundary is contained in the sampled segment.

A different approach is to identify invisible boundaries before segment creation and collapse units (e.g., census blocks) that share an invisible boundary. This solution involves the processing of a great deal of data, so an automated solution is likely the only feasible approach. Invisible lines can be identified in the shape files provided by the Census Bureau, through a field called the MTFCC code[1]. The MTFCC codes that begin with the letter P denote invisible lines. Currently there are 35 different types of invisible lines identified by the Census Bureau.

In order to eliminate, or reduce, the number of invisible boundaries that are external segment boundaries, census blocks (or other geographies) may be collapsed, and these collapsed units, along with blocks that have not been collapsed, can then be used as the basic building blocks to form segments. One challenging aspect of this approach is treating the scenario in which there are multiple adjacent blocks that need to be collapsed into a single unit.

2.2 Contiguity

When forming segments it may be useful for all of the sub units that constitute the segment to be geographically contiguous. Contiguous segments may be marginally more operationally efficient than discontiguous segments. Additionally, if the survey involves environmental sampling, or if neighborhood measures are of interest, it may be advantageous to form contiguous segments.

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. Figure 2 contains an example of a census block group. If blocks 2005-2010 were combined to form a segment, that segment would be discontiguous.

1 Information regarding census shape files can be found at http://www.census.gov/geo/www/tiger/.

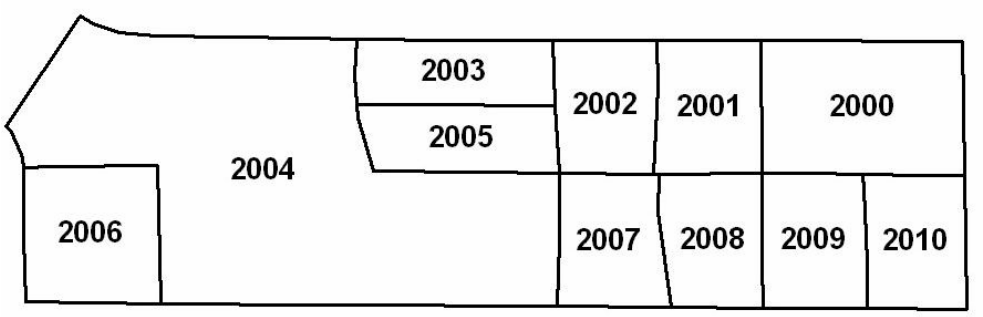

Figure 2. Sequential numbering of census blocks that could result in a discontiguous segment

If geographic contiguity is an objective, making use of geographic information systems (GIS) software is a necessity. The most basic approach involves working with the GIS software to manually combine blocks, one by one, to form segments.

We have adopted the use of an automated algorithm, which was developed at Westat (Johnson, Montaquila, and Heller (2007)). One feature of our approach is the ability to control various parameters that affect the result. However, one of the greatest advantages of this methodology is the ability to create multiple solutions, i.e., multiple possible segment sampling frames. These segment sampling frame options are then evaluated, and the final segment sampling frame is the single segment frame option that is selected as the "best" among the set.

2.3 Compactness

In forming segments, one objective may be to form compact segments. That is, given various alternatives for the shape of a segment, it may be preferable to choose the option that is most compact. Figures 3 and 4 show two different configurations of two segments created from the same set of census blocks. The desirability of creating compact segments is particularly apparent in studies in which there is environmental data collection. If environmental samples are to be obtained that are to be representative of the area, a geographically compact segment (e.g. Figure 3) may be preferable to a sprawling segment (e.g. Figure 4).

Figure 3. Compact segment

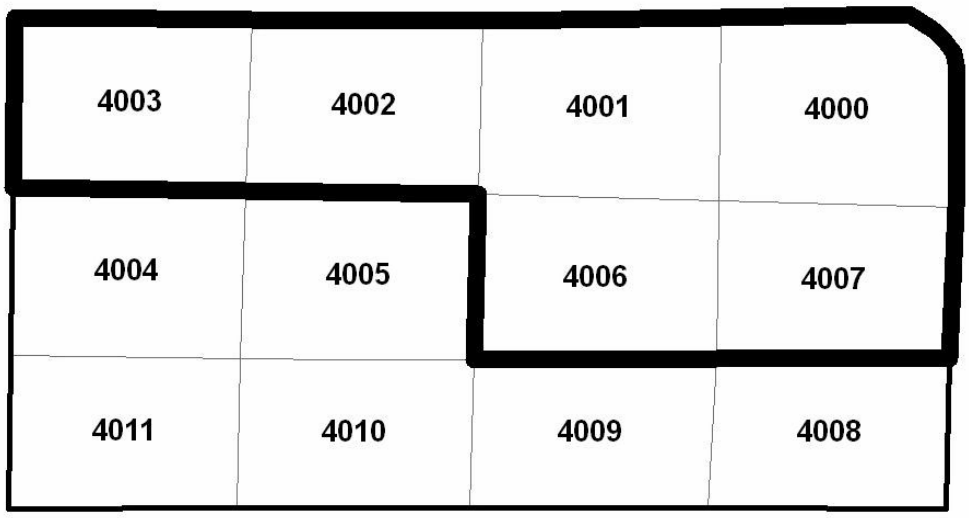

Figure 4. Sprawling segment

When designing procedures to include compactness as a criterion in segment formation, there are various alternative measures of compactness that might be considered. One simple but crude measure is to consider the ratio of interior blocks to exterior blocks. A second measure is the maximum distance between any two points in the segment. A third set of measures involves distances between centroids—e.g., the maximum distance between the centroids of any two blocks in the segment, or the maximum distance from the dynamic segment centroid to any other point in the segment. A fourth measure is the ratio of the perimeter of the segment to the area of the segment.

These measures vary in the complexity of calculation, as well as their effectiveness in characterizing the compactness of the segment. In general, the measures that are simple to calculate are less effective in characterizing the segment's compactness. For this reason, we would propose a two-stage approach to filtering segments, based on compactness. At the first stage, a simple measure (such as the maximum distance between the block centroids) could be used as a filter. At the second stage, a more complicated measure (such as the ratio of perimeter to area) could be calculated for a smaller set of candidate segment options, and a filter based on this measure could be applied.

2.4 Heterogeneity

Frequently an important segment formation objective is to maximize the heterogeneity of segments with respect to several variables. In principle, it would be useful to maximize the heterogeneity of segments with respect to a variable, or set of variables, that are highly correlated with the key measures from the particular survey. Below we discuss a statistic called the heterogeneity score, which offers a systematic approach to determining which segment frame maximizes heterogeneity within segments in the context of this metric.

Since there are often numerous important measures that are obtained from a sample survey, it may not be feasible to identify and use a set of variables that maximize segment heterogeneity with respect to all of the measures at one time. Additionally, it may be difficult to predict any sort of relationship between variables and the survey measures if the focus of the study is entirely new. One must also consider that there are many estimates that may be obtained from a survey that are unplanned. Another important challenge in addressing the topic of heterogeneity is that variables may be correlated (e.g. education and income). We propose a systematic approach that maximizes heterogeneity within segments based on a set of indices. We rely on the judgment of subject area experts to choose a set of variables that they believe will have a strong correlation with key survey estimates.

Once the set of variables has been determined we utilize principal components analysis (PCA) to reduce this full set of variables to a smaller set of principal components that will serve as our indices. There are numerous advantages to utilizing PCA, including the fact that it is widely available in statistical software and is a well understood procedure. However, one of the biggest design advantages to utilizing PCA is that the principal components are completely uncorrelated. The challenges associated with multicollinearity of the variables can thus be easily avoided.

Each record *i* (e.g. census block) is associated with a factor loading, $PRIN_{ij}$, for each principal component *j* and eigenvalue, λ_j associated with the *j*th principal component. One must decide how many of the principal components to use, however this can easily be accomplished by using the associated eigenvalue to determine the proportion of variance explained by each additional principal component.

Arriving at the heterogeneity score is a two-step procedure. The first step uses analysis of variance (ANOVA) on the factor loadings for each unit (e.g., census block), and the second step calculates the heterogeneity score based on results from the ANOVA and the use of eigenvalues as weights. We develop a model for each factor loading $PRIN_{ij}$ and segment frame $SEGID_k$ combination.

$$PRIN_{ij} = \beta SEGID_{ik} + \varepsilon_{ijk}$$

Each segment frame and principal component combination results in an F statistic, F_{jk}, from their respective model.

The factor loadings associated with each record remain static while we calculate different F statistics for each principal component/segment frame combination. Usually the application of the F-statistic is to determine whether a class variable can explain significant variation between observations; a high F-statistic indicates higher levels of homogeneity within the class variable and greater heterogeneity between levels of the class variable. However, in our application we want the segments to be as heterogeneous as possible; thus, identifying segment frames with low F-statistics is the goal of this procedure.

In the second step we calculate a heterogeneity score H_k for each segment frame *k*.

$$H_k = \sum_{j=1}^{n} \lambda_j \times F_{jk}$$

where *n* is the number of principal components (and associated eigenvalues) utilized in developing the heterogeneity score. The number of principal components used in this procedure is determined by the user; each additional principal component offers marginally less useful information in forming heterogeneous segments. The heterogeneity score is calculated for segment frame *k* and uses information from each of the *n* principal components. In the calculation of the heterogeneity score, the weight given to each principal component is the associated eigenvalue.

Because we are focused on maximizing heterogeneity within the segments, we call the score derived from the weighted combination of F-statistics a heterogeneity score but note that low levels of the score are associated with high levels of heterogeneity within the segments.

In the third step, the heterogeneity scores for each segment frame are ranked. The segment frames with the lowest heterogeneity scores are those that maximize heterogeneity within the segments and minimize heterogeneity between segments. Depending on what other criteria may be used in selecting a segment frame, one or more segment frame(s) is (are) selected from the ranked list of heterogeneity scores.

3. Composite Methodology

Many of the considerations and approaches described above can be used together to create a segment sampling frame. It may be advantageous to implement these methodologies in a certain order. For example, if one objective is to limit the number of invisible boundaries, and an automated approach is used, then collapsing units across invisible boundaries should be the first step.

In order to also consider compactness and heterogeneity, it may be necessary to create multiple segment frames, and evaluate each with respect to these measures post hoc. An illustration of the sequence of approaches is given in Figure 5.

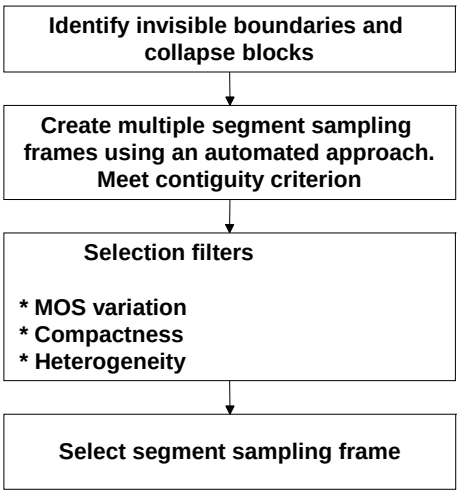

Figure 5. Sequence of approaches in forming segment sampling frame

We have used our automated algorithm to form multiple segment frames. One advantage of this algorithm is its ability to select the segment frames which have the least variation with respect to segment measure of size. The ability to control variation in the measure of size might also be used as selection criterion. Once multiple segment frames have been identified, we employ the approaches described above to evaluate compactness and heterogeneity. The order in which the criteria are considered, and the relative importance attributed to each of the criteria, depends on the objectives for the particular survey.

Acknowledgments

The authors would like to thank Daniel Levine and David Morganstein 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.

Johnson, B., Montaquila, J. (2007). An Automated Procedure for Forming Contiguous Sampling Units for Area Probability Samples. Proceedings of the Section on Survey Research Methods of the American Statistical Association.