DEVELOPING PRIMARY SAMPLING UNIT (PSU) FORMATION SOFTWARE

James L. Green, Sadeq Chowdhury, and Thomas Krenzke
James L. Green, Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Area sampling, geographic clustering

1. Introduction

Area probability sampling techniques are important and frequently used in survey research. The techniques can offer increased coverage and higher response rates than list or telephone sampling. The techniques are also applicable to a variety of elements including establishments, households and persons. Area probability sampling often uses primary sampling units (PSUs) at the first stage of selection. A PSU sampling frame is needed and often requires considerable professional labor to develop. For the past several years, we have invested resources in developing PSU formation software, with the objective of increasing the quality and precision offered by our PSU samples while decreasing PSU frame construction costs. We present the results of our research in this paper.

2. Background

Westat uses area probability sampling techniques quite frequently. The decennial census is the information basis for our PSU definitions, which are re-evaluated with each census release. In the 1990's, Westat maintained a master PSU frame and sample. This frame and sample were available to projects throughout Westat. As the 2000 Census data became available, we decided to review our usage of the master frame and sample. We surveyed the statistical group and found considerable use of project-specific PSU frames and samples, along with the expected master frame and sample usage. After reviewing the project-specific circumstances, we determined that the demand was sufficient to justify developing some timesaving utilities. PSU formation software was one of the results.

We developed this software under the general guidelines and procedures that we use to develop standardized software for particular processes (refer to Krenzke and Green, (2002). In this particular case, PSU formation was a well understood problem within survey research. Westat also had sufficient demand in 2000 and 2001 to justify assigning the required resources.

3. PSU Formation in General

PSU formation in general has some very specific requirements. These requirements include constraints on the PSUs themselves, objectives that the PSU formation must satisfy, and data that are required. These requirements are discussed in the following sections.

3.1 Constraints

In general, PSUs should have the following characteristics:

- Consist of individual counties or groups of counties;
- Contain only adjacent counties; and
- Contain a minimum measure of size.

Since a variety of data is available at the county level, counties are often used as the building blocks for PSUs. We often require a PSU's constituent counties to be contiguous in an attempt to control travel costs. The minimum measure of size requirement for PSUs is usually determined by solving the following equation:

$$MIN\left(M_{\alpha}\right) \geq \frac{\left(f_{o}\right)\left(\Sigma M_{\alpha}\right)}{a},$$

where

$MIN\left(M_{\alpha}\right)$ = minimum measure of size for PSUα;
f_o = largest overall sampling rate;
M_{α} = measure of size for PSUα; and
a = number of PSUs to be selected.

Other constraints will also apply. For example, PSUs may be required to respect Census region since the PSU frame will be stratified for sampling and Census region is often used as a primary stratification variable. PSUs may be restricted within states as more surveys are demanding state-level estimates. It may also be desirable to use or respect existing Metropolitan Statistical Area (MSA) definitions in the PSU formation.

3.2 Objectives

In general, PSUs are formed to satisfy one of the following objectives:

- Minimize data collection costs; and
- Minimize between PSU variance (i.e., maximize within PSU heterogeneity).

These are, of course, the classic competing objectives of cost and variance in survey sampling. The PSUs cluster the fieldwork and thus work to minimize or at least control the data collection costs. The data collection costs are often dominated by travel costs within the sampled PSUs, and these costs can be approximated by extreme end-to-end distance data within the PSU. However, since such a sample is clustered, and since usually a small number of clusters is selected, it is best to minimize the between PSU component of variance. Most manual attempts at PSU formation focus on the first objective. WesPSU allows the user to select either objective, with accommodations for the other objective via constraints.

3.3 Data

Certain data elements are required, given the constraints and objectives above. These data elements (and the sources we use for these data elements) are as follows:

- County adjacency (Census 1990 County Adjacency file, with modifications);
- County measure of size (Census SF1 variables);
- Distance data (Census Tiger mapping file latitude and longitude coordinates); and
- Between PSU variance data (Census SF1 variables).

The extreme end-to-end distance for any area defined by a set of counties can be calculated based on the latitude and longitude coordinates of the constituent counties. The appropriate expression can be found in cartographic literature. Between PSU variance of one or more variables can be calculated following the approach of Kostanich et al. (1981).

4. PSU Formation Software

This section describes our PSU formation algorithm in some detail, presents applications to date and provides an evaluation. The JSM presentation concluded with a brief demonstration. The software is proprietary, however sufficient detail is provided to be informative.

4.1 Algorithm

We developed an algorithm that follows the same general steps taken when forming PSUs with labor intensive approaches. The advantages, of course, are that the program can execute the steps and do the required calculations much faster, considerably more accurately and more consistently than is possible by hand. The algorithm can also focus on minimizing between PSU variance, in either a univariate or multivariate way, and that is beyond what is possible with reasonable labor intensive approaches.

Our algorithm consists of the following seven general steps:

1. Identify primary strata, hard and soft boundaries;
2. Sort counties into sufficient, deficient and solved lists within strata and hard boundaries;
3. Identify all possible PSU formation solutions for each deficient county;
4. Select a given PSU solution from all possible PSU solutions;
5. Update the lists from #2 above and repeat steps #3 and #4;
6. Terminate the process when the deficient list is empty or contains only unsolvable counties; and
7. Adjust solutions for any unsolved counties, contingent on user approval.

First, we sort all counties into mutually exclusive and exhaustive processing streams based on primary strata and any other hard boundaries. This can reduce the size of the PSU formation problem significantly. The primary strata are the strata the statistician expects to use for stratifying the PSUs prior to sampling. For example, since we often stratify PSU samples by Census region, we would pass Census region as a primary stratification variable. The user can also specify other hard boundaries. For example, the user may want PSUs to respect state lines or an urban/rural classification scheme. The user would pass a variable reflecting this scheme as a hard boundary parameter. Soft boundary parameters are also permitted. The algorithm will only cross soft boundaries when required to meet the minimum measure of size constraint.

Second, within each processing stream, we assign each county to one of three lists. These lists are as follows:

1. The sufficient list;
2. The deficient list; and
3. The solved list.

The sufficient list contains all counties that could stand alone as PSUs since they contain the minimum measure of size. The deficient list contains all counties that require combining with other counties to reach the minimum measure of size. The solved list is initially empty, but is filled as PSUs are formed. When a PSU is formed, the constituent counties are removed from the other two lists.

Third, all possible feasible (i.e., which do not violate constraints) PSU solutions for each deficient county are identified. The best of all possible PSU solutions for each deficient county is also identified. The best solution for each deficient county is the PSU solution with the lowest value of the objective function. The objective function works to minimize the end-to-end distance for the PSU, or minimize between PSU variance. This process is repeated independently for each deficient county. Counties from either the sufficient or deficient lists are available for a solution.

Fourth, a PSU solution is selected based on the MINIMAX, MINIMIN (Lingo, 1998), or RANDOM (to be implemented) parameters. If the MINIMAX option is exercised, the county with the worst (in terms of the objective function) best-possible solution is selected. If the MINIMIN option is exercised, the county with the best best-possible solution is selected. If the RANDOM option is exercised, a user-specified number of runs will be made. Each run will use a different, randomly ordered list of the deficient counties and solve the deficient counties in that order. The run with the best value for the objective function will be the one selected.

As one would expect, exercising the MINIMIN option tends to leave a few more deficient counties in the unsolved list. This happens because sometimes one or more of the few counties available for a deficient county with few possible feasible PSU solutions can be taken away early in the process. It is this possibility that the MINIMAX option is designed to prevent.

Finally, the three lists are updated based on the selected PSU. The process of PSU formation and selection is then repeated. The process is terminated when the deficient list is empty or when it contains only unsolved counties. A final, optional procedure forms PSUs for any counties remaining in the deficient list. Since these are most likely deficient counties for which no feasible solutions existed, the resulting PSUs will violate the end-to-end distance constraint. The user must approve these PSUs before they are final.

A disk storage space versus run time dilemma is at the core of developing software to handle the general PSU formation problem. Given that, we should note that the algorithm description provided above is accurate but does not give specifics on several spacesaving and timesaving approaches that we implemented. These approaches managed the size of the problem intelligently and took advantage of particular problem characteristics to reduce run time significantly.

4.2 Applications

We developed, tested and used this software almost simultaneously. We also used the software for production immediately, completely replacing the labor-intensive approaches.

We tested and used the software on the National Center for Education Statistics' (NCES) Early Childhood Longitudinal Study-Birth cohort (ECLS-B) project. We needed to develop 2nd stage units (defined at the county level) within existing sampled PSUs in order to decrease travel time and costs within the existing PSUs. Each run was for a very specific geographic area, which restricted the size of problem considerably and allowed for detailed review and evaluation of results.

We used the software for production in the following studies:

Study (agency)	Measure of size used
Adult Literacy and Lifeskills (NCES)	Noninstitutional civilian population
Commercial Building Energy Consumption Survey (Department of Energy)	Counts of commercial building
Department of Transportation Commercial Truck Survey (DOT)	Miles of interstate/limited access highways
Early Childhood Longitudinal Study-Birth Cohort (NCES)	Annual births by occurrence
Head Start (ACYF)	Head start programs
National Assessment for Educational Progress–Writing On-line, Oral Reading (NCES)	Schools
National Assessment of Adult Literacy (NCES)	Household population
National Health and Nutrition Examination Study (NCHS)	Household population
State Assessment of Adult Literacy (NCES)	Household population
Survey of Youth in Residential Placement (DOJ)	Counts of youth facilities

4.3 Evaluation

Since we used this software for production almost immediately, comparisons between manual approaches and WesPSU had to be constructed. First, we compared a typical manual effort for two states to the same WesPSU attempt. Second, we compared the ECLS-B PSU frame, which was constructed manually before WesPSU was available, to a WesPSU run with the same constraints, objectives etc. In all cases, the software was considerably faster, cost less, and created PSUs with smaller mean end-to-end distance. Similar evaluation measures are provided in the standard WesPSU printed output.

Application	Time	Mean distance in miles (minimax/ minimin)	Standard deviation
Two states (AL, NE) Manual	5 hours	597	423
Two states (AL, NE) WesPSU	53 seconds	309/344	298/317
National (ECLS-B) Manual	200 hours	128	134
National (ECLS-B) WesPSU	60/14 hours	116/106	118/117

4.4 Demonstration

The presentation at the August 2002 Joint Statistical Meetings included a demonstration of the software. PSUs were formed within the state of Pennsylvania. The attached map shows some results, as well as the values selected for parameters.

5. Acknowledgements

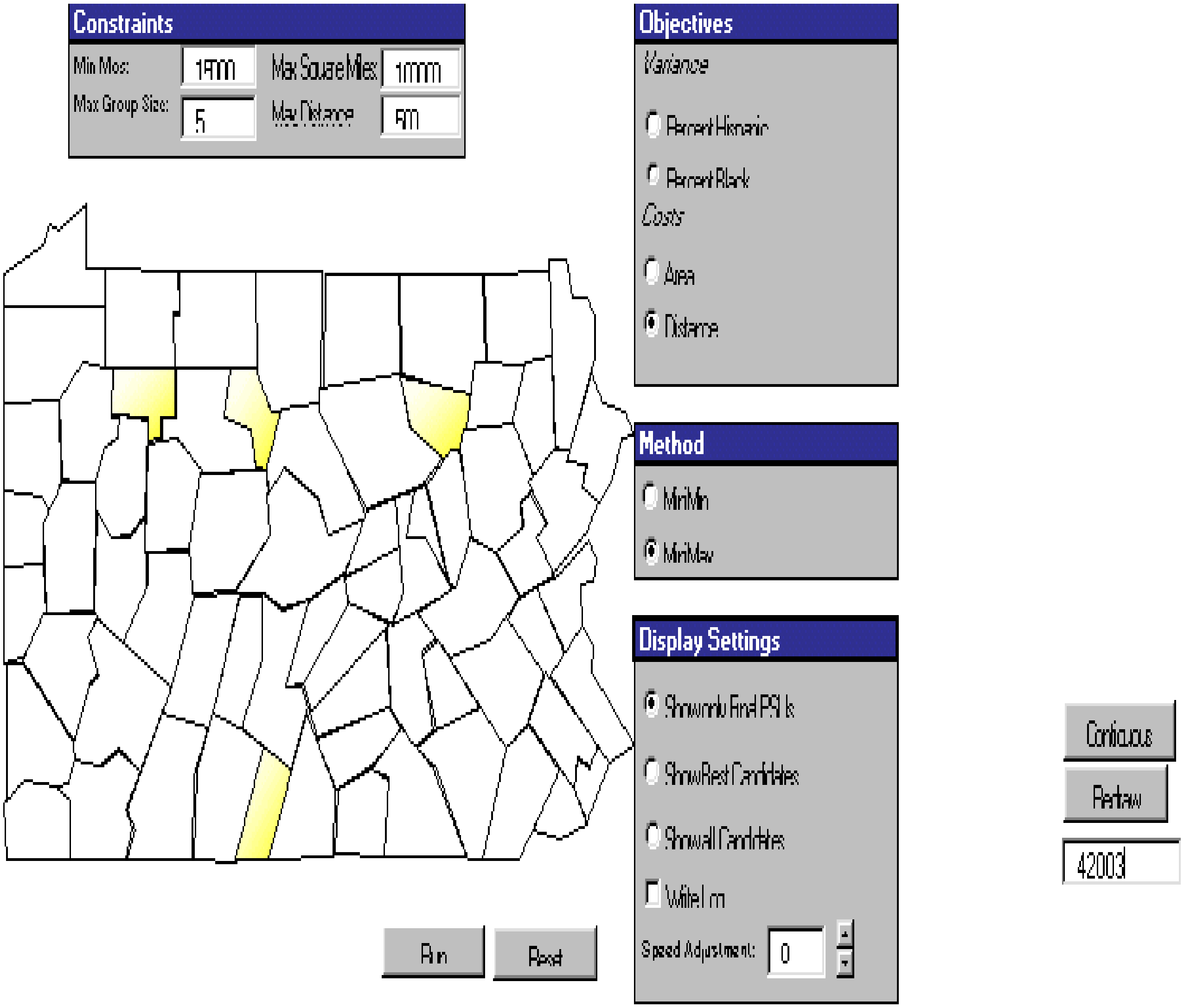
WesPSU was a collaborative effort spanning approximately 4 years from the initial ideas and casual discussions to the first production version. The development team included John Burke, Sadeq Chowdhury, Jim Green, Wen-Chau Haung, Tom Krenzke, Leyla Mohadjer, David Morganstein and Debby Vivari. We especially thank Andrew Heller for his quick and creative software design.

6. References

Kostanich, D., Judkins, D., Singh, R., and Schautz, M. (1981). Modification of Friedman-Rubin's Clustering Algorithm for Use in Stratified PPS Sampling. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 285-290.

Krenzke, T. and Green, J. (2002). When, Why and How to Develop Widely Used Standard Software for Area Sampling. *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*

Lingo. (1998). *Optimization Modeling with Lingo*. 2nd ed., *Lindo Systems*, Chicago IL.



Constraints

Min Mos: 19000

Max Square Miles: 10000

Max Group Size: 5

Max Distance: 90

Objectives

Variance

○ Percent Hispanic

○ Percent Black

Costs

○ Area

● Distance

Method

○ MiniMin

● MiniMax

Display Settings

● Show only Final PSUs

○ Show Best Candidates

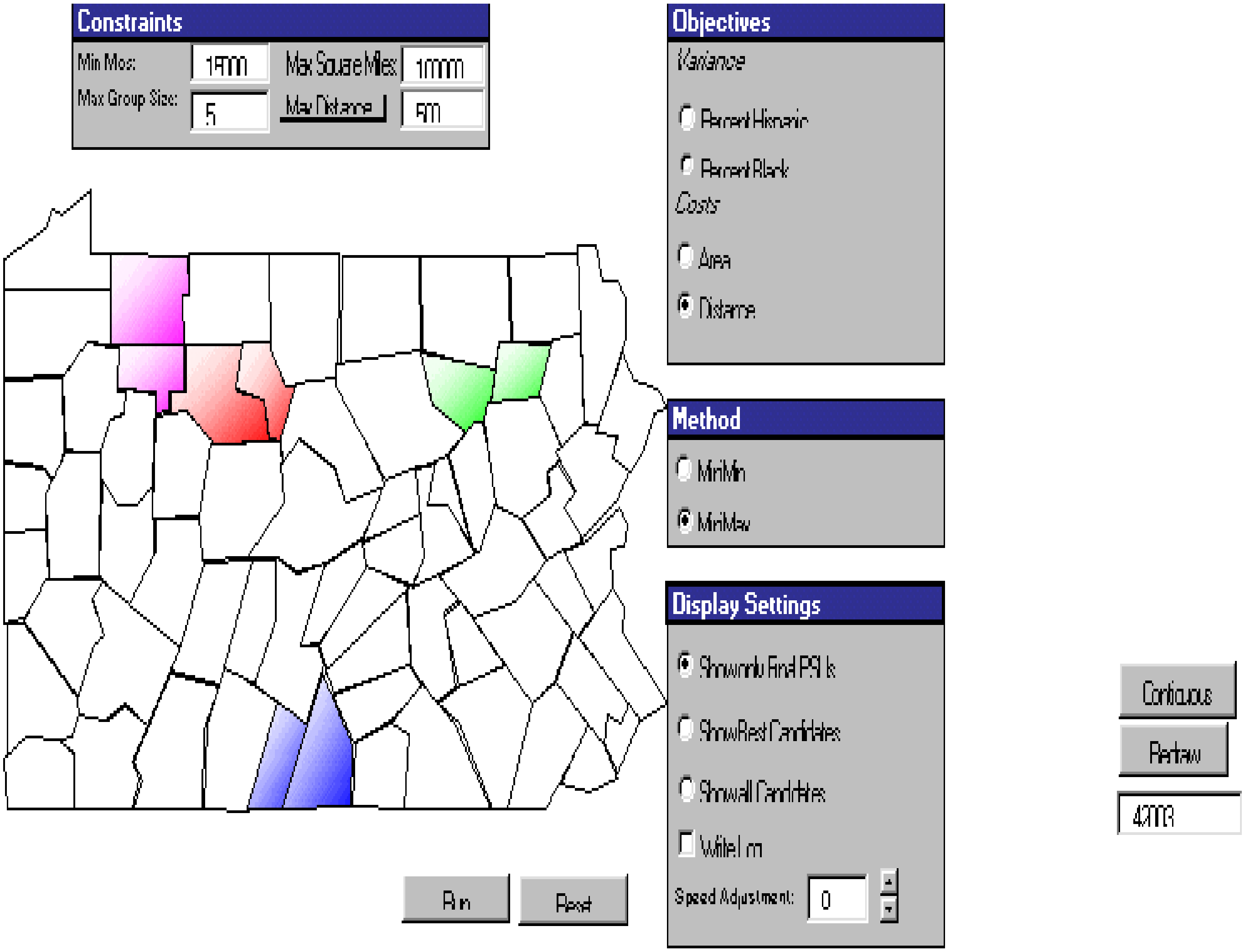
○ Show all Candidates

☐ Write Log

Speed Adjustment: 0

42003

WesPSU demonstration - deficient counties



Constraints

Min Mos: 15000 Max Square Miles 10000

Max Group Size: 5 Max Distance 500

Objectives

Variance

○ Percent Hispanic

○ Percent Black

Costs

○ Area

● Distance

Method

○ MiniMin

● MiniMax

Display Settings

● Show only Final PSUs

○ Show Best Candidates

○ Show all Candidates

□ Write Log

Speed Adjustment: 0

42003

WesPSU demonstration - formed PSUs