OPTIMAL STRATIFICATION OF AREA FRAMES

Charles R. Perry, USDA, National Agricultural Statistics Service
and James E. Gentle, George Mason University
Charles R. Perry, USDA/NASS/Research and Development Division,
Room 305, 3251 Old Lee Highway, Fairfax, VA 22030
cperry@nass.usda.gov

ABSTRACT

The project's objective was to order the count units of an area frame so that the induced sampling strata contain contiguous blocks of count units that are homogeneous with respect to major crops. An earlier phase of this project used standard photo interpretation techniques with categorized Landsat TM and soil maps for the major agriculture producing area of Arkansas to reorder the Arkansas area frame. The reordering reduced the sampling variance as much as 50% for some major crops with no increase in variance for items not included in the reordering process. However, implementation was impractical because of the statistical expertise and time required to carry out the manual reordering process. We have now automated the procedure as an optimization problem and used simulated annealing to find near optimal solution.

This paper reports some results obtained in an application of the automated procedure to the land in intensive agricultural land-use category of the Arkansas area frame. Analysis shows the procedure to be highly effective in reducing the variance of major crops, even more effective than the manual reordering process.

Key Words: Area Frame, Landsat TM Categorized Imagery and Data, Simulated Annealing

INTRODUCTION

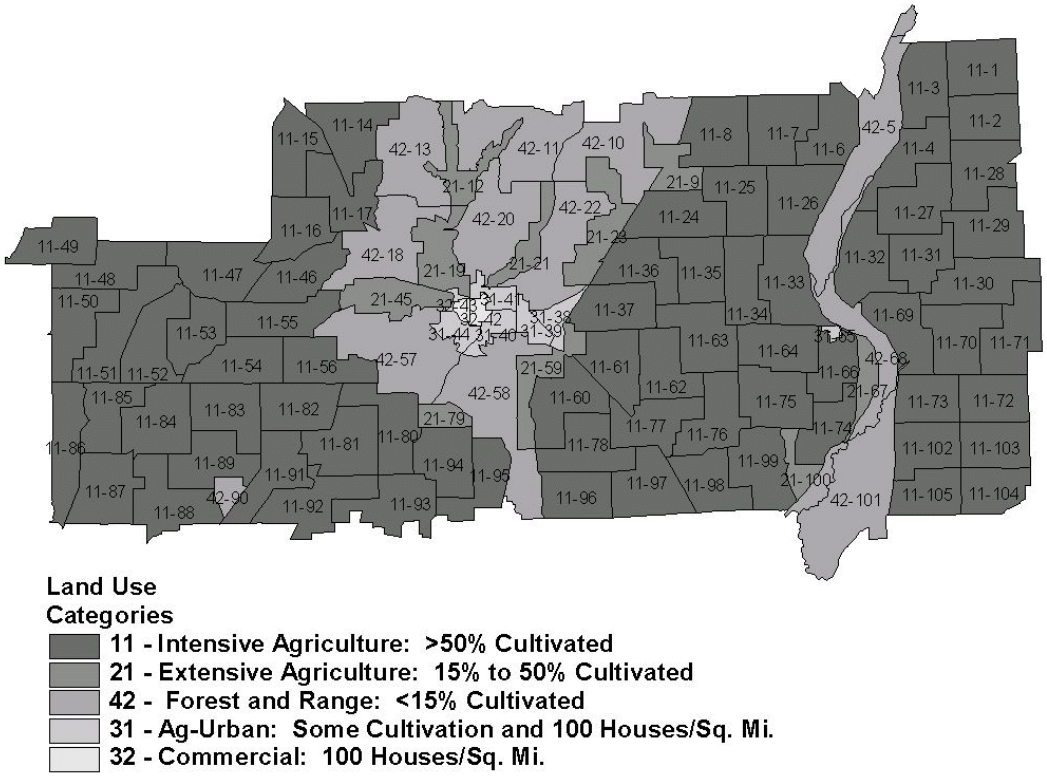
The U. S. National Agricultural Statistics Service (NASS) conducts a number of surveys of agricultural products throughout the year. The area frames for most surveys are constructed within each state by first dividing the state into various land-use categories, corresponding to the intensity of crop cultivation, and then dividing each land-use category into blocks for which permanent boundaries can be identified. These blocks are called "count units".
The total area in count units and the associated number of near equal size land units, called segments, are determined. A typical count unit contains 5-10 near equal size segments which follow permanent observable boundaries to the extent possible. The count units are then grouped into equal-sized strata from which equal-sized simple random samples of segments are drawn.

In practice, the strata are formed by first ordering the count units and then defining the strata systematically as sub-sequences of count units. Although not absolutely necessary, it is desirable that the strata consist of contiguous count units. Because the count units are of various sizes, a typical stratum will contain two partial count units. Figure 1 shows the standard arrangement of count units within a given county.

Count units are numbered in a simple serpentine pattern. Strata within land-use categories are formed from the sets of sequentially-numbered count units which consist of contiguous blocks, except when a disruption occurs in the land-use category, as shown in Figure 1.

OPTIMAL FORMATION OF STRATA

The method of organizing count units into strata is one of the most important steps in this process. The objective, obviously, is to form strata that are as homogeneous as possible with respect to each item in the survey. Because of the standard ordering, shown in Figure 1, to systematically form strata, some heterogeneity results within strata. However,



this heterogeneity is not as great as it would be if we formed strata in other ways; or, equivalently, if we reordered the list of count units in a random fashion.

One approach to forming better groupings of count units is to use satellite images of crops, and manually form groups of "similar" count units. Once the groups are identified, the count units within a given group can be numbered so that sequential numbers are contiguous, and then strata can be formed systematically as before.

This manual effort resulted in an increase of 20% to 100% in the relative efficiency of major crops. See Perry and Nayak (2000) which will be issued later this year for further details.

There are a number of disadvantages to this manual process. Two are obvious: the process is labor intensive, and it requires a high level of expertise. Another disadvantage of this process is the classification errors of crops using satellite data. The satellite data provides only approximately 80% correct classification. Another disadvantage is the subjectivity inherent in judging the "similarity" of count units. The objective in the manual reordering process is to order the count units such that units close in the ordering have similar amounts of the crops of interest. This process results in the strata, which are short strings of count units, being homogenous with respect to the crops of interest. In Arkansas we arranged the count units within land-use categories so that the first units in the orderings had high concentrations of one crop and ended with units with high concentrations of another crop. In addition, we attempted to maintain smooth transitions through the concentrations of other significant crops.

EXPLICIT FORMULATION OF THE OPTIMIZATION PROBLEM

The objective is to form strata that have minimal within-stratum variances, with respect to a set of target variances, for all items in the survey. It is clear that there may be no solution to this optimization problem, so the practical problem to be solved is to form strata for which a weighted sum of within-relative-variances is minimized. Thus the objective function is of the form :

$$\sum_{i\,(strata)} \sum_{j\,(crops)} w_j R_{ij} \qquad (1)$$

where w_j is the relative weight given to the j^{th} agricultural product, and R_{ij} is the variance of the j^{th} agricultural product within the i^{th} stratum, relative to a target variance for that product.

Because the count units have different sizes, and because the strata may contain some partial count units, the within-stratum variance for the j^{th} agricultural product is proportional to a weighted sum of squares:

$$R_{ij} \propto \sum_{k \in i^{th}\ stratum} p_k (x_{jk} - x_j)^2, \qquad (2)$$

where p_k is the geographical size of the k^{th} count unit, x_{jk} is the amount of the j^{th} crop in the k^{th} count unit, and x_j is the weighted mean of the amounts of the j^{th} crop over the count units in the i^{th} stratum.

In addition to the within-stratum variances of the agricultural products, it is desirable to include a component that measures spatial dispersion within the strata. There are a number of reasons for this. One is to achieve strata consisting of contiguous, or at least nearby, count units. Another reason is that the spatial variation likely encompasses the variability of a number of underlying determinants of crop suitability. This can result in two benefits. There may be other agricultural products that we may wish to include in future surveys, and small spatial variation will make it likely that there will be small variations in these other products. In addition, the spatial component may smooth out the variability of crop variances over time.

The spatial variance is measured in the same way as the crop variances in equation 2, where the *x*'s are 2-vectors containing the spatial coordinates of the centers of the count units. Instead of using the variance directly, however, we use a function of the variance that has been determined empirically to relate spatial variance to agricultural product variance. See Perry (1979).

Incorporating the spatial variation within the i^{th} stratum, D_i, into the objective function in equation 1, we have

$$f = \sum_{i\,(strata)} \left(\alpha D_i + (1 - \alpha) \sum_{j\,(crops)} w_j R_{ij} \right) \qquad (3)$$

As described above, the strata are formed systematically as subsequences of the list of count units. Thus, f in equation 3 is a function of the ordering, and the problem of optimally forming strata becomes the problem of optimally ordering the list of count units. The objective in the minimization problem therefore is to find

$$\text{MIN}[\text{of } f \text{ all possible count unit permutations}] \qquad (4)$$

Another modification to the objective function we refer to as "chaining". As mentioned above, we may at different times form strata of different sizes. It is desirable that a single ordering of the count units yields a near optimal results for strata of different sizes. We incorporate this feature in the objective function by computing the minimum of f in expression 4 over different sizes of overlapping strata.

The optimization problem in equation 4 is extremely computationally-intensive. The decision space is of size $n!$ where n is the number of count units. Because an exhaustive search is not possible, and because there are many local minima, we chose to use simulated annealing to solve the problem. Simulated annealing does not guarantee an optimal solution. Therefore, we used several control parameters to increase the likelihood of an optimal solution. Examination of the computation log of the simulated annealing process is used to determine when a near optimal solution has been obtained.

We first developed a prototype of the simulated annealing algorithm for this problem in S-Plus, which proved to be too

slow to for production work. We then developed a FORTRAN version of the program. More details of the algorithm will be available later this year in Perry and Gentle (2000).

RESULTS

We have used this procedure for the intensive agriculture land-use category (category 11) in Arkansas. Figure 2 shows the strata formed by the automated procedure when applied to the intensive agriculture category in Craighead County, Arkansas.

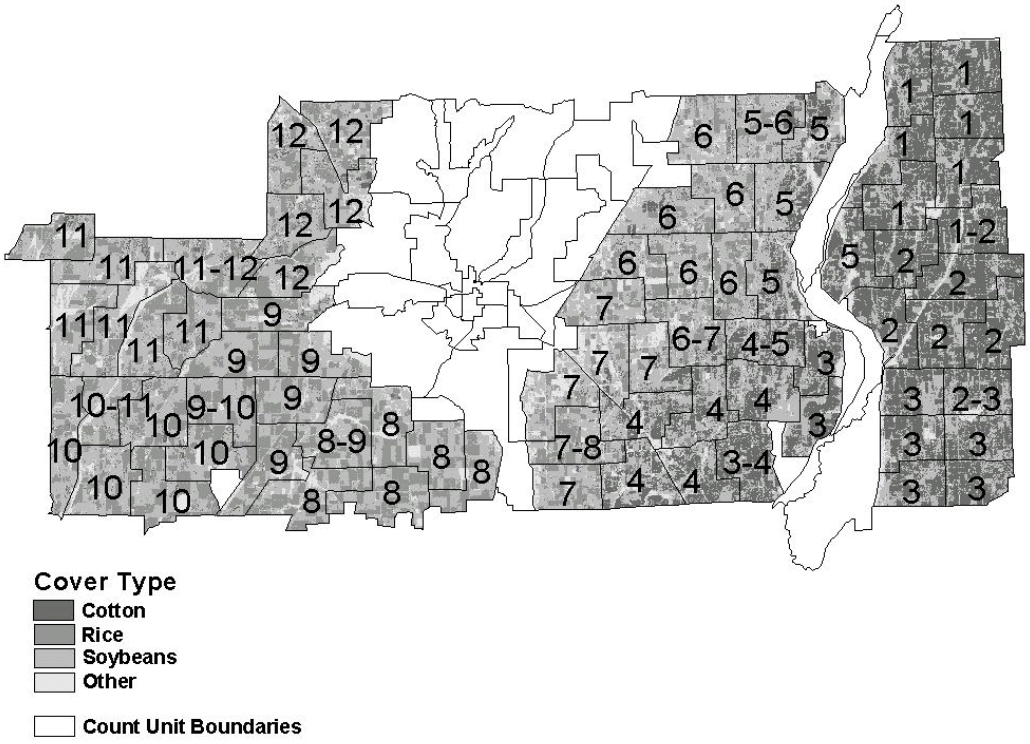


Figure 3 shows estimated relative planted acreage variances for several crops using the automated process in the intensive agriculture land-use category (category 11) of Arkansas . Figure 4 shows estimated relative planted acreage variances using the manual process. Landsat data were used in the automated process for two of the items shown, cotton and rice. However, no Landsat data were used for the third item, corn. A comparison of these graphics show that the automated process significantly reduced the relative variances for all three crops. The variance reductions were consistent with our objective, especially the major reduction in the cotton variance using the automated process.

REFERENCES

Perry, Charles R., and Cecil Hallum (1979), Sampling Unit Size Considerations in Large Area Crop Inventorying Using Satellite-Based Data, *Proceedings of the Survey Research Methods Section of the American Statistical Association Conference*, Washington, D.C.

Perry, Charles R. (2000), Improving the Efficiency of the Arkansas Area Frame Using Categorized Satellite Imagery, NASS Technical Report, U.S. Department of Agriculture.

Perry, Charles R., and James E. Gentle (2000), Optimal Stratification of Area Frames, NASS Technical Report, U.S. Department of Agriculture.

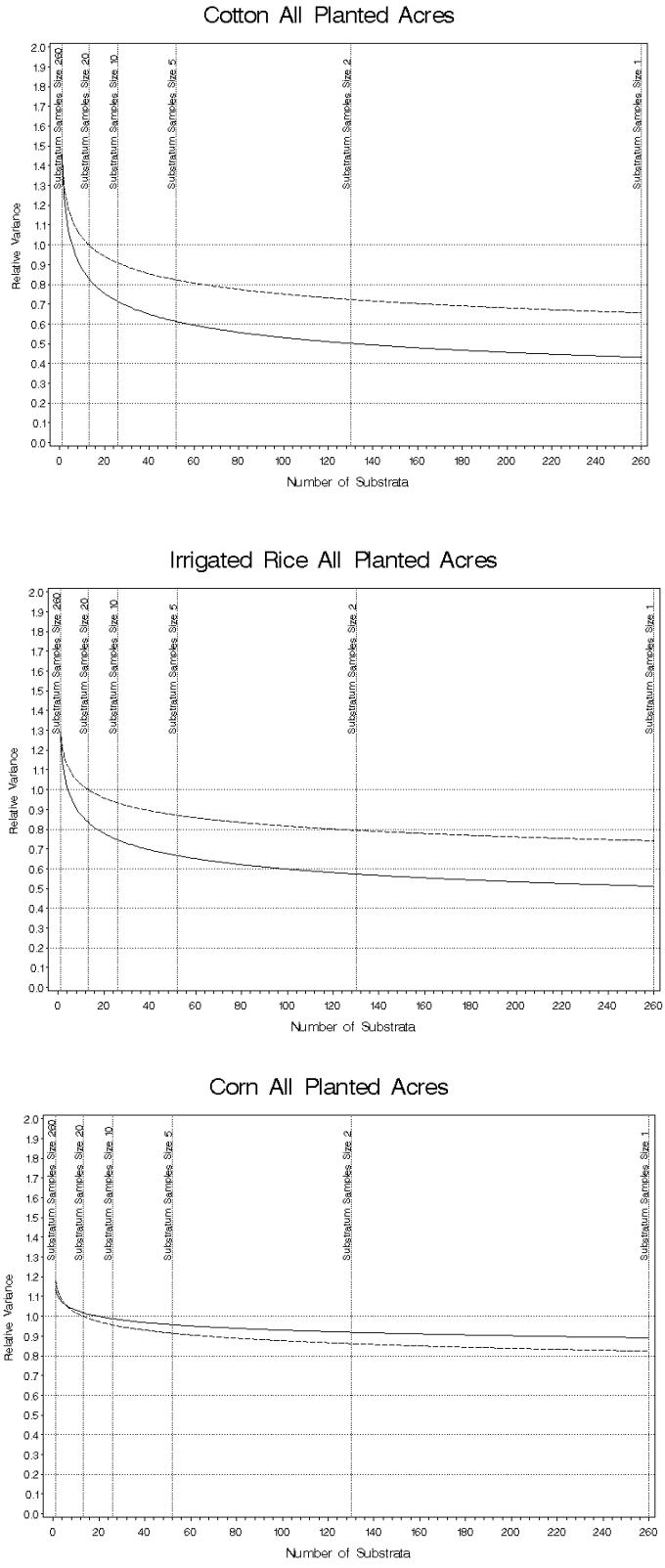


Figure 3: Comparison of crop acreage relative variances for the automated process and the current NASS ordering using 1 to 260 strata in the intensive agriculture land-use category in Arkansas.

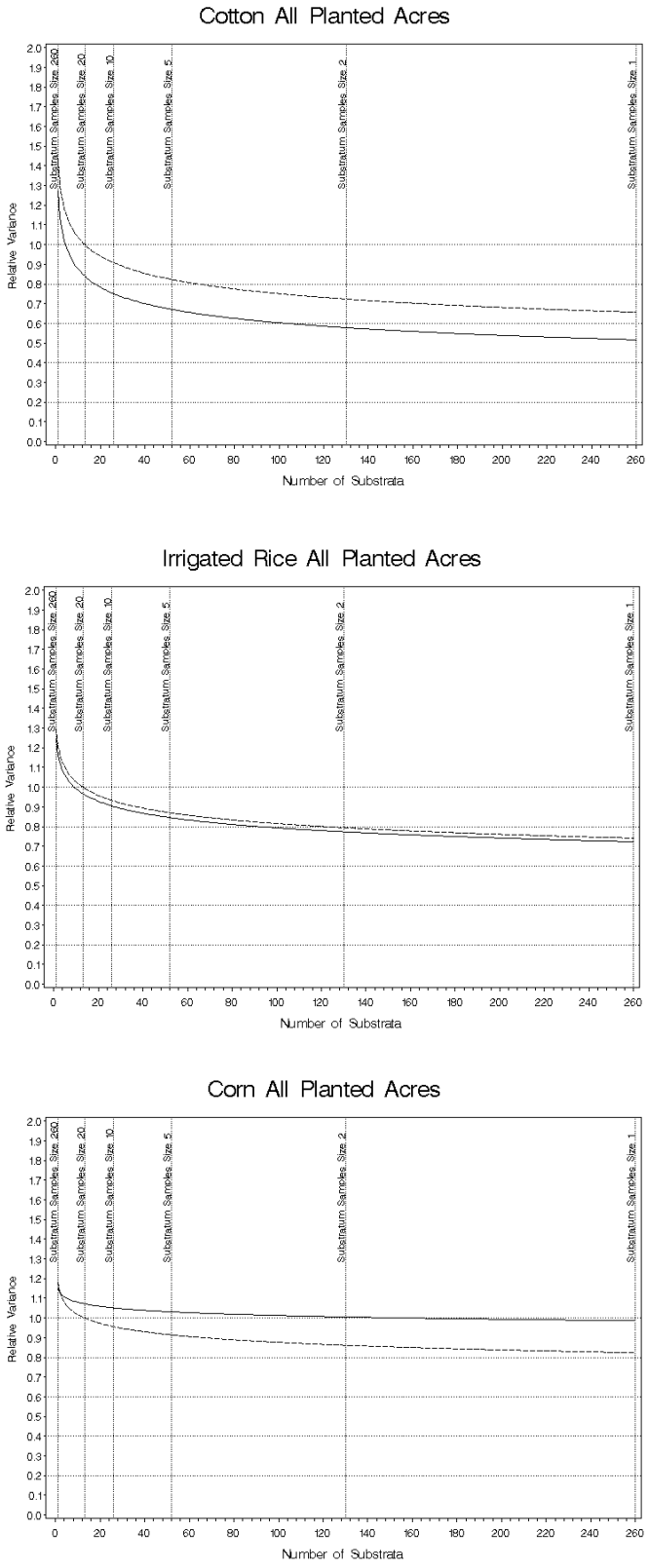


Figure 4: Comparison of crop acreage relative variances for the manual process and the current NASS ordering using 1 to 260 strata in the intensive agriculture land-use category in Arkansas.