ALASKA NATIVE AND AMERICAN INDIAN TRIBE SAMPLING FRAME CONSTRUCTION AND SAMPLE DESIGN FOR THE NATIONAL FOOD AND NUTRIENT ANALYSIS PROGRAM

Charles R. Perry, Jr., Daniel G. Beckler, Michael E. Bellow, Linda G. Gregory, USDA-NASS
Pamela R. Pehrsson, USDA-ARS
Charles R. Perry, Jr., USDA-NASS, 3251 Old Lee Hwy., Room 305, Fairfax, VA 22030

KEY WORDS: probability sample, Census of Agriculture, duplication removal, optimization

1. INTRODUCTION

In 1998, USDA's Nutrient Data Laboratory (NDL) implemented the National Food and Nutrient Analysis Program (NFNAP). The overall goal of this project is to improve estimates of nutrient measures for common foods consumed in the United States (Perry and Beckler, 2000). Additional funds were obtained that extended the program to include traditional foods consumed by Alaska natives and American Indians. Hence, a sampling plan was needed to select a statistically representative sample of members of those two groups from which traditional foods were to be obtained and analyzed. In order to get representative nutrient values, the sampling plan required that all American Indians who consume traditional foods have a chance to be selected. This paper details the construction of the list of American Indian tribes and procedures that were used to obtain a sample of native peoples, which was then used to obtain samples of traditional foods. A similar approach will be used for sampling of Alaska natives.

2. SAMPLE DESIGN APPROACH

Ideally, a complete list of American Indians who consume traditional foods would be used to select a probability sample of individuals. Since such a list was not available, another method was used to obtain a probability-based sample. The following is the four-step sample design procedure applied for the American Indian frame:

1. Obtain a complete list of tribes including counts of number of individuals in each tribe, and merge it with a list containing a measure of the concentration of agricultural production in the tribes' general locations (Agricultural Statistics Districts). The merged list constitutes the sampling frame.

2. Divide the list of tribes into as many homogeneous equal size groups as there are tribes to be sampled. Homogeneity is measured in terms of the tribes' geographic dispersion and the concentration of various present day agricultural commodities. These commodities serve as a proxy for the availability of commodities used in traditional foods.

3. Select a representative tribe or sample of tribes from each group with probability proportional to size (with population as the size measure).

4. Obtain an equal amount of each sampled food from each selected tribe or sample of tribes, to be used for nutrient analysis.

The sampling frame for Alaska natives, separate from the American Indian frame in the lower 48 states, has been developed. For the American Indian frame, separate samples were drawn for five group sizes (6, 12, 24, 36 and 48). This design excluded from selection individuals not associated with a tribe, for example those living in cities away from tribal areas. Since the NDL targeted those who eat traditional foods as part of their regular diet, this was not seen as a serious limitation to the design. American Indians living on reservations and other Indian lands are believed to consume traditional foods more regularly than those living elsewhere, and were therefore included on the list of tribes used.

3. SAMPLING FRAME CONSTRUCTION

This section describes the construction of the sampling frame of Alaska natives and American Indian tribes. A variety of data sets were needed to construct sampling frames adequate for drawing and contacting samples of tribes. Raw data came from the sources discussed below.

The *FY 1997 Labor Force Report* (LFR), available from the U.S. Department of the Interior's Bureau of Indian Affairs (BIA), provided the most recent list of Alaska natives and American Indian tribes and their population sizes. Two items included in the report are "tribal enrollment" and "total Indian resident service population". The tribal enrollment is the official population of the tribe as defined by the tribal constitution. The total Indian resident service population is defined as "the tribe's estimate of all American Indians and Alaska natives, members and non-members, who were living on or near the tribe's reservation during the

1997 calendar year and who were eligible to use BIA funded services." Tribal enrollment is believed to be the most accurate statistic in the report, since the remaining ones were estimates.

The BIA's Internet web site contains a directory of all tribal leaders for the 554 federally recognized tribes, including name, street address, telephone number, fax number and web address (if applicable). This list was used to affix a zip code to each of the tribes on the LFR. The tribal leaders' contact information was used to reach the selected tribes to solicit participation. Once the list of American Indian tribes was obtained, *The Atlas of The North American Indian* (Waldman, 2000) was used to try and obtain the tribal affiliation of all tribes listed. For those tribes with more than one affiliation, each one was recorded separately. The Environmental Systems Research Institute (ESRI), a leading marketer of geographic information systems (GIS) software and databases, provided *ESRI Data & Maps* (August, 1999) on CD ROMS. The following information was obtained using this product: zip code centroids, county area data (for each county), county population data (Census).

Data from the 1997 Census of Agriculture were obtained from USDA's National Agricultural Statistics Service (NASS). Groups of commodities grown in each tribal area were used to enhance the formation of homogeneous groups of tribes. The commodity groups used were the following: berries, small grains, corn, vegetables, citrus/fruits/nuts, other crops, cattle, sheep, goats and cropland. Administratively confidential estimates were obtained for each commodity group for the counties where each tribe was located. The Census data were summarized by Agricultural Statistics District (ASD), the USDA's subdivision of each state into regions of similar agricultural activity (with each county belonging to exactly one ASD). A file containing all counties in the U.S. along with their respective ASD's was obtained from NASS's Spatial Analysis Research Section.

In order for the Census of Agriculture data to be used outside of NASS, the administratively confidential status had to be removed, which was accomplished by standardizing the estimates and including only the standardized values in the final data set. The standardized estimates are not administratively confidential but are as useful for sampling purposes as the original estimates.

BIA's LFR originally contained 772 records (both Alaska native and American Indian tribes). However, the list contained much duplication, i.e., records having the same tribe name, tribal agency (a BIA administrative designation) and enrollment. The duplication was due to some tribes being listed under multiple headings. Unfortunately, due to minor spelling variations and the manner in which tribes were identified, duplication could not be removed with a simple computer process. For example, the computer was not able to identify the following two records on the report as being duplicates:

Tribe Name	Tribal Agency
Healy Lake Village	Alaska-Tanana Chiefs Conference
Healy Lake Village	Tanana Chiefs Conference

In order to eliminate all duplication, the entire report was printed and visually inspected, with a flag manually entered on records found to be duplicate. After the flagged records had been removed by computer, the list contained 585 records, each corresponding to a unique tribe. The next step was to merge tribal leader information with the 585 tribes. The original BIA Tribal Leaders Directory (TLD) contained 675 records, most corresponding to tribes but others to tribal agencies and BIA offices. Due to spelling variations and the manner in which tribes were identified on the LFR and TLD, a simple computer match-merge was not possible. Both documents were printed and visually inspected for matching tribes. Match codes corresponding to records on the TLD were manually entered on the LFR records. Later, the computer performed the actual matching (based on the match codes) and merged the TLD data with the LFR data.

There were 571 tribes in the LFR that could be matched with a tribal leader; the other 14 tribes were dropped. Most of the dropped 'tribes' were actually loose organizations of American Indians choosing to identify themselves as 'at large' instead of claiming affiliation with a specific tribe. Incidentally, the TLD contained 104 records that could not be matched with a tribe on the LFR. These records, containing leaders of BIA offices and Indian agencies, were not put on the list.

The zip code listing in the TLD was used as a proxy to identify the location of a tribe, assuming that the tribal leader lived relatively close to that location. Five-digit zip code centroids (geographic centers of zip code areas) obtained from the *ESRI Data & Maps* CD ROMS were merged onto the file containing the 571 tribes. Records were matched on the five-digit zip codes. Unfortunately, this operation did not fully succeed because several tribes had Post Office Box Only zip codes, i.e., ones used only for PO Boxes (having no actual land area associated with them). The CD ROMS did not provide centroid data for such zip codes. The United States Postal Service's (USPS) geography group confirmed that polygons and therefore centroids were not available for PO Box Only zip codes. To remedy this problem, the following two-step plan was implemented. First, the

town associated with each PO Box Only zip code was identified from the USPS web site. If another non-PO Box Only zip code was found for the town, its centroid was used (if more than one zip code was found, one was chosen at random). Second, if the town had no non-PO Box Only zip codes, the centroid of the three-digit zip code polygon, also obtained from the CD ROMS, was used. Three-digit zip code centroids were used for only 55 of the 571 tribes, mostly in Alaska.

The file containing the five-digit zip code centroids also contained state and county FIPS codes. The FIPS codes for three-digit zip codes were found manually on maps and added to the computer files, then used to append the remaining auxiliary data. Various auxiliary data sets were used to improve the ordering of tribes in order to reduce the variance of a sample of tribes.

For the American Indian frame, ASD's, county area and population data, and commodity data from the 1997 Census of Agriculture were appended to the file of 571 tribes based on matching state and county FIPS codes. ASD level data for each of these items were also appended to the file. Standard SAS procedures were used to summarize the items at the ASD level from the county level data. The resulting file contained one record for each tribe. A variety of data sets useful for sampling were included on the file (see Section 4). The stratification and sample selection procedures have been fully developed and implemented for the American Indian tribes. Similar procedures are used for the Alaska natives, with harvest regions replacing commodity data (Alaska's Department of Fish and Game has divided the state into five subsistence or harvest regions). Sections 4 and 5 refer specifically to the American Indian frame.

4. STRATIFICATION OF AMERICAN INDIAN FRAME

To ensure that the samples of traditional foods analyzed were representative of the foods eaten by American Indians in the lower 48 states, tribes were divided into homogeneous strata of equal size. A tribe or local collection of tribes was randomly selected from each stratum from which food samples were to be collected. The homogeneity of the tribes is measured in terms of their geographic dispersion and the concentration of various present day agricultural commodities grown in the area. The agricultural commodities were used as proxies for the traditional commodities available in the area where the tribes are located, since data on traditional commodities were not available. The sampled tribe from each stratum was selected with probability proportional to the tribe's population size. For nutrient analysis, an equal amount of the foods of interest will be collected from each sampled tribe or local collection of tribes.

Strata were formed by first ordering the tribes and then defining the strata as subsequences of tribes. The objective was to order the tribes so that the induced sampling strata would contain homogeneous groups of tribes with respect to both geographic dispersion and the amount of the commodities grown in the ASDs associated with tribes. Since the tribes are of various sizes, a typical stratum contains two partial tribes.

Statistically, the objective was to form strata having minimal within-stratum variances with respect to a set of target variances for the commodities and the spatial dispersion. Clearly, there may be no solution to this optimization problem, so the practical problem was to form strata for which a weighted sum of within-relative-variances was minimized. Formally, the objective function is of the form :

$$f = \sum_{i=1}^{L} \sum_{j=1}^{C} w_j R_{ij} \qquad (1)$$

where L is the number of strata, C is the number of agricultural commodities, w_j is the relative weight given to agricultural commodity j, and R_{ij} is the variance of commodity j within stratum i relative to a target variance for that commodity. Since the tribes have different sizes and the strata may contain some partial tribes, the within-stratum variance for commodity j is proportional to a weighted sum of squares:

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

where:

p_k = population size of tribe k
x_{jk} = amount of commodity j associated with tribe k
$\bar{x}_j$ = weighted mean of commodity j over the tribes in stratum i

In addition to the within-stratum variances of the agricultural commodities, we desire a measure of the spatial dispersion within the strata. One reason for such a measure is to achieve strata consisting of nearby tribes, while another is that the spatial variation likely captures some of the variability of a number of underlying determinants of commodity suitability of the geographic area where the tribes are located. Two benefits can result from this: 1) there may be other commodities that could be included in the stratification but for which no data are available, and 2) minimizing the spatial variation will

make it likely that there will be small variations in these other commodities. In addition, the spatial component may smooth out the variability of the commodities and other factors over time. The spatial variance is measured in the same way as the commodity variances in equation (2), where the *x*'s are 2-vectors containing the spatial coordinates of the centroids of the tribes. However, instead of using the variance directly, a function of the variance that has been determined empirically to relate spatial variance to agricultural commodity variances is used (Perry and Hallum, 1979).

Incorporating the spatial variation within the stratum *i* (D_i) into the objective function in equation (1), we have:

$$f = \sum_{i=1}^{L} \left(\alpha D_i + (1 - \alpha) \sum_{j=1}^{C} w_j R_{ij} \right) \qquad (3)$$

where α is a weighting factor. As described above, the strata are formed systematically as subsequences of the list of tribes. Thus, f in equation (3) is a function of the ordering, and the problem of optimally forming strata becomes one of optimally ordering the list of tribes. The objective of the optimization problem is to minimize f over all possible permutations of the tribes.

Another modification of the objective function is known as *chaining*. A single ordering of the tribes yielding near optimal results for strata of slightly different sizes is desired. This feature is incorporated in the objective function by minimizing f over different sizes of overlapping strata. The optimization problem is extremely computationally intensive with decision space of order $n!$, where n is the number of tribes. Because an exhaustive search is not possible and there are many local minima, the method of simulated annealing (Kirkpatrick et. al., 1983) was chosen to solve the problem. While not guaranteeing an optimal solution, simulated annealing when properly controlled can almost always produce a nearly optimal solution. Several control parameters were used to increase the likelihood of a nearly optimal solution. The computation log of the simulated annealing process was examined to determine exactly when such a solution had been obtained.

The two step application of the optimization procedure used to obtain a nearly optimal stratification of the American Indian frame for each sample size of interest (6, 12, 24, 36 and 48) will now be described. This procedure was required because it was necessary to set the target variances of the commodities after the minimum spatial variance had been determined. In this application, optimization plays a somewhat different role than in the usual sample survey where the target variances are normally given a priori for the commodities. The idea behind the use of the proxy commodities was to improve upon the homogeneity of the spatial groups with respect to the crops grown in the area without significantly increasing the spatial variance within the groups.

The first step was to use the optimization procedure to group the tribes into the desired number of strata, ignoring the commodity information. Since only the spatial components were involved at this stage, the spatial target variance could be chosen arbitrarily. Since the simulated annealing algorithm used in the optimization procedure is only guaranteed to produce a *nearly* optimal solution, the procedure was performed several times with varying control parameters to ensure that such a solution had been obtained. Once it was, the next objective was to reduce the within-stratum commodity variances as much as possible without significantly increasing the within-stratum spatial variance. That is, the goal was to create strata more homogeneous with respect to the commodity without significantly increasing the geographical dispersion of the tribes contained within each stratum.

The second step required two substeps, starting with the nearly optimal spatial stratification obtained in the first step. First, the target variances for the spatial and commodity variances were set to a fraction (e.g., 0.1) of their values at the optimal spatial stratification obtained. The optimization procedure was then rerun with a different value of the weighting factor α, which controls the relative importance of the spatial variance and commodity variances, until a solution was found that did not increase the spatial component of the objective function by more than five percent. In effect, the within-stratum commodity variances were reduced to the extent possible at the expense of a small increase in the within stratum spatial variance. The result was to make the strata more homogeneous with respect to the commodities with a minimum increase in the geographical dispersion of the tribes within the individual strata. Figure 1 is a map showing the stratification associated with the sample size 24. On this map, a location labeled '1-2', for example, contains tribes from strata 1 and 2.

5. SAMPLE SELECTION PROCEDURE FOR AMERICAN INDIAN FRAME

Chromy's algorithm (Chromy, 1971), a sequential, probability minimum replacement sampling scheme, was used to select a stratified sample of tribes from the American Indian frame. A sequential sampling scheme considers a frame's sampling units in numerical order, with a decision made as to how many times each unit will

be included in the sample. Probability minimum replacement (PMR) sample designs are probability proportional to size (pps) and allow certain sampling units to be selected more than once:

n(i) = number of times unit *i* is selected in sample
n = sample size
S(i) = size measure for sample unit *i*
S(+) = sum of size measures for all units in frame
q(i) = E[n(i)] = nS(i)/S(+)

The Chromy procedure divides the frame into *n* zones of size *S(+)/n*. One sampling unit is selected from each zone with probability proportional to size. Associated with each unit *i* is a line segment of length *q(i)*, which either falls entirely within one sampling zone or overlaps two or more zones. Figure 2 illustrates the procedure for a hypothetical case where a sample of size five is to be drawn from eight available sampling units. If *q(i)* exceeds one, then sampling unit *i* covers one or more zones completely and is known as a self-representing unit (e.g., unit 4 in Figure 2). Such units are guaranteed to appear in the sample at least once. If a unit is in part of two adjoining sampling zones but is not self-representing (units 3 and 6 in Figure 2), then it can be selected in one of the two zones but not both. By ensuring that a single unit is selected from each zone, the sample is implicitly stratified by the frame ordering. The variance is reduced as long as units in close proximity are more homogeneous than those in the population at large, which can be accomplished if units sufficiently far apart are in different selection zones. The frame is ordered by using control variables highly correlated to the quantity being measured so that neighboring units are similar.

In sampling from the American Indian frame, the zones correspond to the strata discussed earlier, with one tribe picked from each stratum.

6. ADDITIONAL APPLICATIONS

The sampling approach described in this paper lends itself to a number of applications. If sampling within a targeted tribe or area of tribes related by a designated variable is appropriate, a specific frame for that tribe or area can be extracted from the larger data set. This condition would be true for indigenous plants and animals known to exist only in a specific region, tribe specific ceremonial foods, etc. In addition, the grouping variable list (i.e., crops) can be augmented with information about specific areas, lending more precision to the grouping process. For example, specific information from published research about harvest regions in Alaska would yield more precision in the selection of groups to be sampled among Alaska natives.

REFERENCES

Chromy, J.R. (1971), "Sequential Sample Selection Methods", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 401-406.

Kirkpatrick, S., Gelatt Jr., C.D., and Vecchi, M.P. (1983), "Optimization by Simulated Annealing", *Science*, vol. 220, pp. 671-680.

Perry, C.R., and Beckler, D.G. (2000), "A National Sampling Plan for Obtaining Food Products for Nutrient Analysis", *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

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

Perry, C.R. and Gentle, J.E. (2000), "Optimal Stratification of Area Frames", *Proceedings of the Second International Conference on Establishment Surveys*.

Waldman, Carl (2000), *The Atlas of The North American Indian*. New York, NY: Checkmark Books.

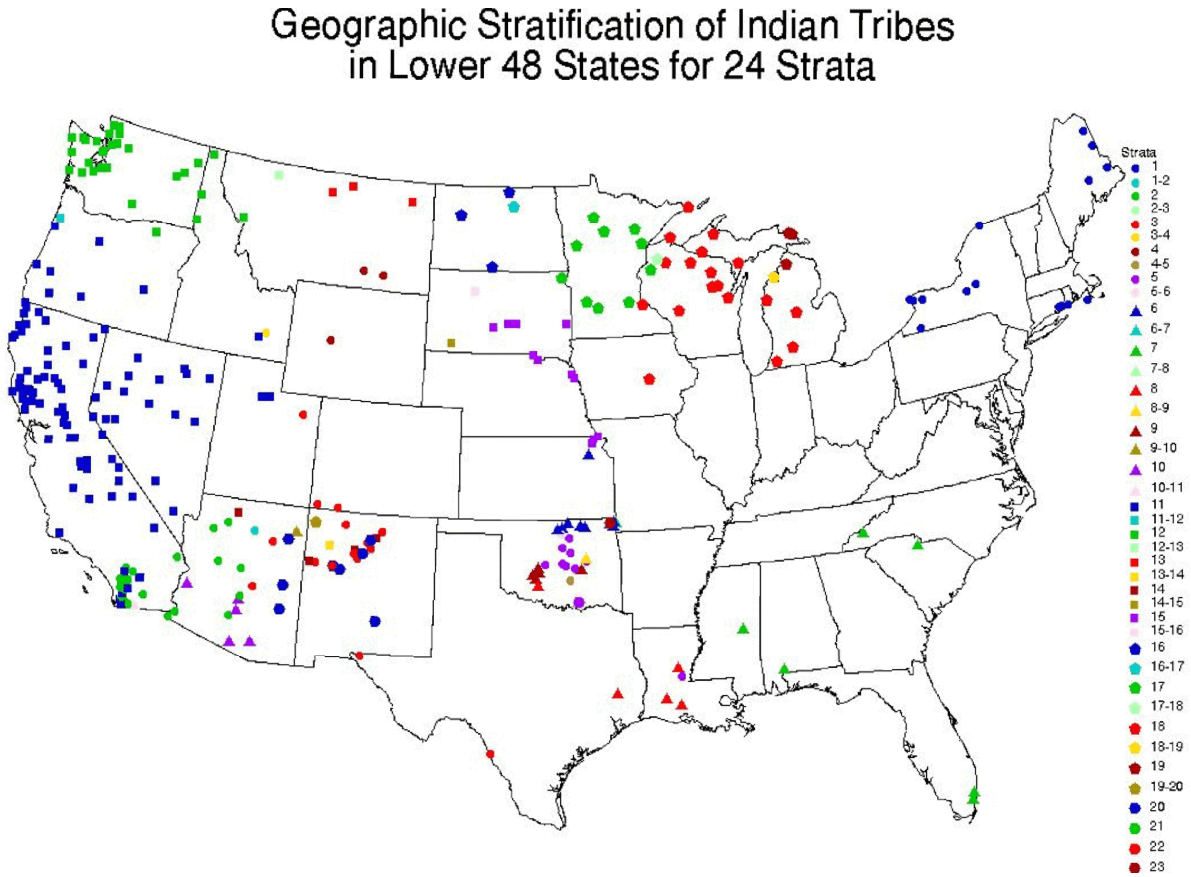

Figure 1: Stratification Example

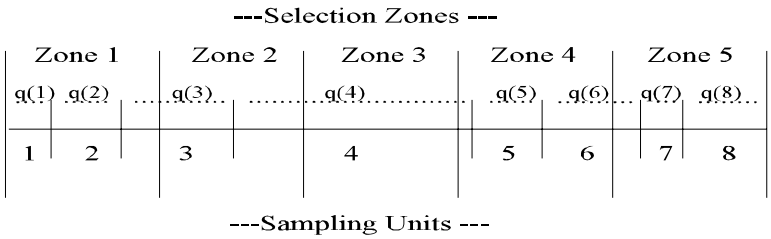

Figure 2: Illustration of Chromy's Algorithm