# A Hierarchical Clustering Algorithm for Multivariate Stratification in Stratified Sampling *

Stephanie Zimmer † Jae-kwang Kim† Sarah Nusser †

**Abstract**

Stratification is used in sampling to create homogeneous groups. A number of methods have been proposed for stratification of populations using covariates of the variable of interest. These include Dalenius and Hodges' (1959) cumulative root frequency method, the Lavallee and Hidiroglou (1988) algorithm, and the Gunning and Horgan (2004) geometric stratification method. All of these methods assume you have one variable of interest and one correlated auxiliary variable known for the population. Many surveys have more than one important variable of interest as well as many auxiliary variables. The method we propose considers multiple variables of interest. We use a superpopulation model to create a distance metric between elements in the population that depends on multiple auxiliary variables. Using the proposed metric, a hierarchical clustering algorithm can be used to implement the optimal stratification automatically by combining elements into strata that are closest together. Our method is motivated by the NASS June Area Survey (JAS), where we have multiple auxiliary variables to stratify sample segments and want to make estimates for several crop and livestock parameters.

## 1. Motivating Problem

The National Agricultural Statistics Service (NASS) provides timely, accurate, and useful statistics in service to U.S. agriculture. One tool they use to achieve this goal is the Quarterly Agriculture survey. During one of the four surveys, the June Agriculture Survey, NASS uses an area frame, in addition to their list frame which is used in all of the four surveys. The area frame is used to estimate the incompleteness of the list frame. An area frame is expensive but has complete coverage of the population. Currently, every state has an area frame with a few states having new area frames each year [2]. The process in making each area frame is very costly because it is labor intensive and time intensive.

Two main problems in the current stratification is that unlike some other area frames, it is not a permanent frame. It must be updated because stratum definitions become out of date. Another problem is that recently constructed frames have PSUs that do not meet definitions. An example of a PSU not meeting definition would be if a PSU is in a strata defined as more than 85% cultivated land has less than 85% cultivated land.

## 2. Goals of Research

As an alternative to the current area frame stratification, we consider a list frame approach that is based on permanent frame units which partition the US land area. For instance, the Public Land Survey System (PLSS) is a division of the United States that divides the land into townships (approximately 36 square miles) and sections (approximately 1 square mile). The PLSS can be considered as a sampling frame for many states.

For current auxiliary information to define strata, we propose to use the Cropland Data Layer (CDL) to characterize the sampling units into strata. The CDL uses satellite images

*This research was funded in part by the USDA National Agricultural Statistics Service via cooperative agreement 583AEU10012.

†Iowa State University, Statistical Laboratory & Department of Statistics, Center for Survey Statistics & Methodology

to classify land into different types of land at a pixel level [1, 5]. By summarising the area of pixels classified as a type of land, we can generate an estimate of the different types of land in the sampling units from the frame. Figure 1 illustrates an example of the CDL from 2011. The next question we address is how to stratify the frame units using this information.

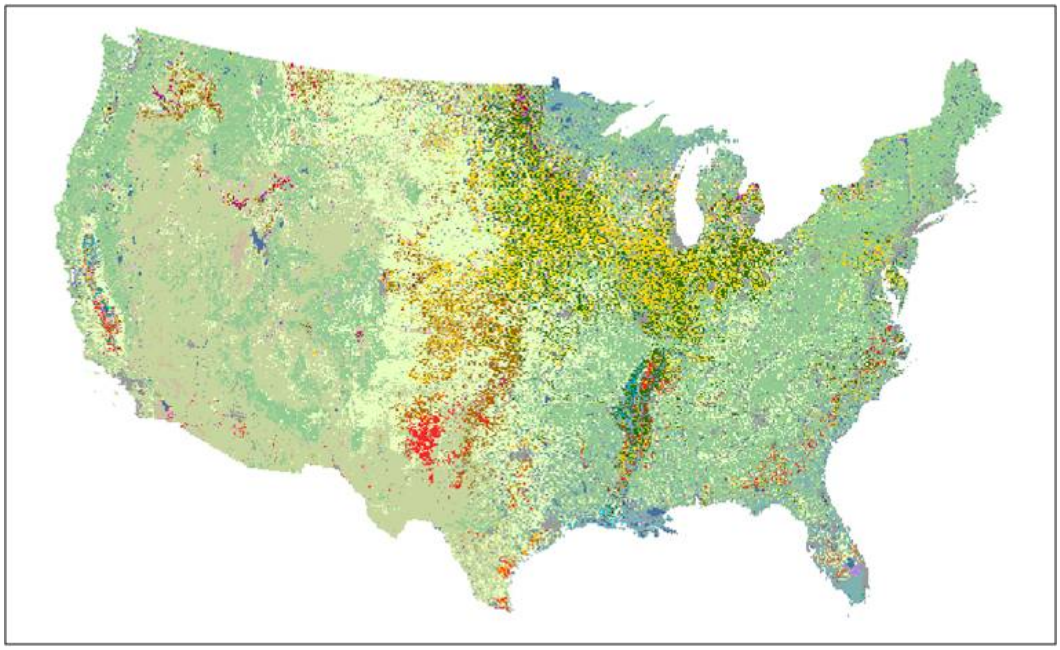

**Figure 1**: 2011 CDL where each color represents a different crop or land cover such as wetlands, water, or urban/developed.

We consider methods to stratify the units to achieve efficient estimation. We will examine three stratification algorithms, including two standard stratification algorithms and a proposed hierarchical clustering algorithm. We apply the proposed method to the June Area Survey for a few states.

## 3. Stratification

In sampling, the use of a stratified sampling partitions a population into disjoint subgroups called strata. Sampling is done within each strata, independently of the sampling in other strata. The attribute used to define the strata must be known for each unit in the frame. Sometimes the strata are natural partitions of the population. For example, gender, race, and age groups when sampling from a frame of individual people. Other times, the variable is a continuous variable without natural grouping such as income, sales, and number of employees for businesses. After stratifying the finite population, the sampling design must be chosen within each stratum, which includes the allocation of the sample. If the strata are relatively homogeneous within, then a stratified sampling design will have a lower variance of the mean than a simple random sample (SRS). If SRS is done within each strata, the variance of the estimate of the mean and the variance of that mean are as follows

$$\bar{y}_{ST} = \frac{1}{N} \sum_{h=1}^{H} N_h \bar{y}_h \tag{1}$$

$$V(\bar{y}_{ST}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \tag{2}$$

where $N$ is the population size, $N_h$ is the size of stratum $h$, $n_h$ is the sample size of stratum $h$, $\bar{y}_h$ is the sample mean of stratum $h$, and $S_h^2$ is the sample variance of stratum $h$.

For a given $n$, one must decide how to allocate the sample size to each strata. For a given response variable $y$, if one wants to have minimum variance for the estimated mean of $y$ of the population, Neyman allocation should be used where $n_h \propto N_h S_h$. If one wishes to accurately estimate the mean within each stratum, proportional allocation should be used where $n_h \propto N_h$. Additionally, power allocation can be used to balance between the variance of the estimate of the population mean and the stratum means. For power allocation, $n_h \propto N_h^p$ where $0 < p < \infty$.

## 4. Current Stratification Methods

In their paper, Dalenius and Hodges (1959) [3] state that there are four design considerations in stratification which are: the choice of stratification variables; the choice of the number of strata; the determination of the way in which the population is to be stratified; and the choice of the size $n_h$ of the sample to be taken from the $h^{th}$ stratum. The methods discussed are concerned with the third specification, the determination of the way in which the population is to be stratified. The Dalenius and Hodges algorithm and the Lavallée and Hidiroglou (1988) algorithm both choose break points for univariate strata given the number of strata desired. Other stratification methods include the method introduced by Gunning and Horgan (2009) [4] which uses the idea of cumulative root frequency from Dalenius and Hodges but removes the arbitrary choice of number of classes which will be discussed later.

### 4.1 Lavallée and Hidiroglou Algorithm

The Lavallée and Hidiroglou algorithm (1988) [6] was intended to stratify a univariate, skewed population. When given a number of strata, L, the algorithm chose breakpoints such that $y_{(0)} < b_1 < b_2 < \cdots < b_{L-1} < y_{(N)}$ where $N$ is the number of elements in the population and $y_{(h)}$ is the $h^{th}$ smallest value of the study variable. The final stratum with the largest elements is a take-all stratum while sampling is done in the other $L-1$ stratum. Some standard notation that will be used is

- $L$ is the number of strata;
- $W_h = N_h/N$ for $h = 1, \ldots, L$ is the relative weight of stratum $h$, $N_h$ is the size of stratum $h$, and $N = \sum N_h$ is the population size;
- $n_h$ for $h = 1, \ldots, L$ is the sample size in stratum $h$ and $f_h = n_h/N_h$ is the sampling fraction;
- $\bar{Y}_h$ and $\bar{y}_h$ are the population and sample means of $Y$ within stratum $h$;
- $S_{yh}$ is the population standard deviation of $Y$ within stratum $h$.

Strata will be constructed using a stratification variable $X$. Stratum $h$ consists of all units with an $X - value$ in the interval $(b_{h-1}, b_h]$ where $-\infty < b_0 < b_1 < \cdots < b_{L-1} < b_L = \infty$ are the stratum boundaries. If we use SRS in the strata, $\bar{y}_{st} = \sum W_h \bar{y}_h$ with variance

$$\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) S_{yh}^2 \tag{3}$$

Since $n_L = N_L$, the sample size in the other stratum can be written as $(n - N_L)a_h$ where $n$ is the total sample size and $a_h$ define the allocation such that $\sum_{h=1}^{L-1} a_h = 1$ and $a_h > 0$ $\forall\, 1 \leq h \leq L-1$. For example, one could use Neyman allocation if one assumes a uniform cost per unit and wishes to achieve minimum variance of the mean.

Solving Equation (3) for $n$ leads to

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_{yh}^2/a_h}{\mathrm{Var}(\bar{y}_{st}) + \sum_{h=1}^{L-1} W_h S_{yh}^2/N} \tag{4}$$

Then the optimal stratum boundaries are the values of $b_1, \ldots, b_{L-1}$ that minimize $n$ subject to a constraint on the precision of $\mathrm{Var}(\bar{y}_{st}) = \bar{Y}^2 c^2$ where $c = \frac{\sqrt{V(\bar{Y})}}{\bar{Y}}$ is the target coefficient of variation, CV. Alternatively, you can minimize $\mathrm{Var}(\bar{y}_{st})$ for a fixed $n$.

Since $y$ is not known, one can use a model for discrepancy between the stratification variable and the survey variable. In 2002, Rivest [7] introduce the idea of using a model between a known variable and the study variable. Let $\{x_i, i = 1, \ldots, n\}$ denote a known stratification variable that is available for all $N$ units in the population. Stratum $h$ consists of the population units in the interval $(b_{h-1}, b_h]$. To optimize the stratification, $E(Y|b_h \geq X > b_{h-1})$ and $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$ must be known. One model that could be considered is the log-linear model where $\log(Y) = \alpha + \beta_{\log} \log(X) + \epsilon$ where $\epsilon \sim N(0, \sigma^2_{\log})$. Another model is the linear model where $Y = \beta_{lin} X + \epsilon$ where $\epsilon \sim (0, \sigma^2_{lin} X^\gamma$ where $\gamma \in (0, \infty)$. The model chosen depends on the actual relation between stratification variable and study variable where these are common relationships. One uses the model to calculate $E(Y|b_h \geq X > b_{h-1})$ and $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$.

To find the optimal boundaries, it is suggested to use the Sethi algorithm [8]. The algorithm Sethi introduces is for stratification when using proportional, equal, and optimal (Neyman) allocation. The main difference between the Sethi algorithm and the Lavallée algorithm is the use of a take-all stratum. Lavallée and Hidiroglou specify how to find the optimal boundaries for the log-linear case, but it should follow similarly for other models. To begin, we should consider $W_h = \int_{b_{h-1}}^{b_h} f(x)dx$, $\phi_h = \int_{b_{h-1}}^{b_h} x^\beta f(x)dx$, and $\psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} f(x)dx$ where $\beta$ is the slope from the log-linear model. These quantities are considered because the optimal stratification is a function of $W_h$, $\phi_h$, and $\psi_h$. We can rewrite Equation (4) as follows using conditional means and variances,

$$n = NW_L + \frac{\sum_{h=1}^{L-1} W_h^2 \mathrm{Var}(Y|b_h \geq X > b_{h-1})/a_{h,X}}{\bar{Y}^2 c^2 + \sum_{h=1}^{L-1} W_h \mathrm{Var}(Y|b_h \geq X > b_{h-1})/N} \tag{5}$$

where $a_{h,X}$ denotes the allocation rule written in terms of $X$. Given a model between $Y$ and $X$, $\mathrm{Var}(Y|b_h \geq X > b_{h-1})$ and $E(Y|b_h \geq X > b_{h-1})$ can be written in terms of $W_h$, $\phi_h$, and $\psi_h$. For example, if one uses the linear model $E(Y|b_h \geq X > b_{h-1}) = \phi_h$ and $\mathrm{Var}(Y|b_h \geq X > b_{h-1}) = \psi_h$. Thus the partial derivatives on $n$ with respect to $b_h$ can be evaluated for $h < L-1$ using the chain rule as follows,

$$\frac{\partial n}{\partial b_h} = \frac{\partial n}{\partial W_h}\frac{\partial W_h}{\partial b_h} + \frac{\partial n}{\partial \phi_h}\frac{\partial \phi_h}{\partial b_h} + \frac{\partial n}{\partial \psi_h}\frac{\partial \psi_h}{\partial b_h} + \frac{\partial n}{\partial W_{h+1}}\frac{\partial W_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \phi_{h+1}}\frac{\partial \phi_{h+1}}{\partial b_h} + \frac{\partial n}{\partial \psi_{h+1}}\frac{\partial \psi_{h+1}}{\partial b_h} \tag{6}$$

where

$$\begin{aligned}
\frac{\partial W_h}{\partial b_h} &= -\frac{\partial W_{h+1}}{\partial b_h} = f(b_h) \\
\frac{\partial \phi_h}{\partial b_h} &= -\frac{\partial \phi_{h+1}}{\partial b_h} = b_h^\beta f(b_h) \\
\frac{\partial \psi_h}{\partial b_h} &= -\frac{\partial \psi_{h+1}}{\partial b_h} = b_h^{2\beta} f(b_h)
\end{aligned}$$

so that the partial derivative of $n$ with respect to $b_h$ simplifies to

$$\frac{\partial n}{\partial b_h} = f(b_h)\left\{\left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right) + \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}}\right) b_h^{\beta} + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right) b_h^{2\beta}\right\}. \tag{7}$$

Also,

$$\frac{\partial n}{\partial b_{L-1}} = f(b_{L-1})\left\{-N + \frac{\partial n}{\partial W_{L-1}} + \frac{\partial n}{\partial \phi_{L-1}} b_{L-1}^{\beta} + \frac{\partial n}{\partial \psi_{L-1}} b_{L-1}^{2\beta}\right\}. \tag{8}$$

Note that this is a quadratic function in $b_h$ except that $W_h$, $\phi_h$, and $\psi_h$ are functions of $b_h$. However, we can iteratively find new $b_h^{\beta}$ for $h < L-1$ as follows,

$$b_h^{\beta,new} = -\left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}}\right) \Big/ \left\{2\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\right\}$$
$$+ \frac{\left\{\left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}}\right)^2 - 4\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)\left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right)\right\}^{1/2}}{2\left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)}$$

and for $h = L-1$,

$$b_{L-1}^{\beta,new} = \frac{-\frac{\partial n}{\partial \phi_{L-1}} + \left\{\left(\frac{\partial n}{\partial \phi_{L-1}} - 4\frac{\partial n}{\partial \psi_{L-1}}\left(\frac{\partial n}{\partial W_{L-1}} - N\right)\right)^2\right\}^{1/2}}{2\frac{\partial n}{\partial \psi_{L-1}}} \tag{9}$$

One can start with initial values for $b_1, \ldots, b_{L-1}$ and continue updating $b_h$ with $b_h^{new}$ until convergence.

### 4.2 Dalenius and Hodges Algorithm

First, let us assume we know the distribution of a variable. Consider a density $f(x)$ with mean $\mu = \int t f(t) dt$. The range $x_0, x_L$ of the estimation variable $x$ is cut up into $L$ parts at points $x_1 < \cdots < x_h < \cdots < x_{L-1}$ where each part corresponds to a stratum. For the $h^{th}$ stratum

$$W_h = \int_{x_{h-1}}^{x_h} f(t) dt \tag{10}$$

$$W_h \mu_h = \int_{x_{h-1}}^{x_h} t f(t) dt \tag{11}$$

$$\sigma_h^2 = \frac{\int_{x_{h-1}}^{x_h} t^2 f(t) dt}{W_h} - \mu_h^2 \tag{12}$$

A sample of $n = \sum_h n_h$ observations is taken from $f(x)$ and $\mu$ is estimated by

$$\bar{x} = \sum_h W_h \bar{x}_h \tag{13}$$

and the estimate has variance

$$\sigma_s^2(\bar{x}) = \sum_h W_h^2 \frac{\sigma_h^2}{n_h}. \tag{14}$$

Under Neyman allocation, the minimum variance is achieved and the variance equals

$$\sigma^2_{min}(\bar{x}) = \frac{1}{n}\left(\sum_h W_h\sigma_h\right)^2. \tag{15}$$

Dalenius demonstrated that the set of $\{x_h : 1 \leq h \leq L\}$ which corresponds to minimum variance stratification satisfies the following

$$\frac{\sigma_h^2 + (x_h - \mu_h)^2}{\sigma_h} = \frac{\sigma_{h+1}^2 + (x_h - \mu_{h+1})^2}{\sigma_{h+1}} \tag{16}$$

However, in practice, the distribution of the study variable is unknown. Only the stratification variable is known. It has been shown that an approximation to Equation (16) when $L$ is large is

$$W_h\sigma_h = \text{constant}. \tag{17}$$

A rule of thumb is introduced under the assumption that relative variance does not vary much stratum to stratum and leans to the following condition

$$W_h\mu_h' = \text{constant} \tag{18}$$

where $\mu_h'$ refers to the measure of a highly correlated variable to the study variable.

Next, Dalenius and Hodges introduce an approximation which will be useful in real problems. The first transformation introduced is

$$y(u) = \int_{-\infty}^{u} \sqrt{f(t)}dt. \tag{19}$$

Let $\lim_{u\to\infty} y(u) = H$. Then the roots $x_1', \ldots, x_h', \ldots, x_{L-1}'$ to the following equations are taken as the first approximations to the points which satisfy Equation (16)

$$y(u) = \frac{h}{L}H, \qquad h = 1, \ldots, L-1 \tag{20}$$

Since the distribution of the study variable is never known, the following algorithm is applied to an auxiliary variable to construct strata. $J$ is chosen arbitrarily, but should be much larger than the desired number of strata. Let $L$ denote the number of strata.

1. Arrange the stratification variable $X$ in ascending order
2. Group $X$ into $J$ classes
3. Determine the frequency in each class for the frame: $f_i$ $(i = 1, 2, \ldots, J)$
4. Determine the square root of the frequencies in each class
5. Cumulate the square root frequencies, $\sum_{i=1}^{J}\sqrt{f_i}$
6. Divide the sum of the square root of the frequencies by the number of strata: $Q = \frac{1}{L}\sum_{i=1}^{J}\sqrt{f_i}$
7. Take the upper boundaries of each stratum to be the $X$ values corresponding to $Q, 2Q, 3Q, \ldots, (L-1)Q, LQ$.

It is not common to use a model between the auxiliary variable and the study variable, but one idea is to find a an estimate of the study variable using a model and the auxiliary data and then apply the algorithm to the estimate of the study variable.

## 5. Proposed method

The proposed method uses a hierarchical algorithm to minimize a distance function which is the equivalent to minimizing the variance of the estimate of a mean under a stratified design where the variance is $V(\bar{y}_{st}) = N^{-2}\sum_{h=1}^{H} N_h^2 S_h^2/n_h$. Under Neyman allocation, $n_h \propto N_h S_h$ leads to $V(\bar{y}_{st}) \propto \sum_{h=1}^{H} N_h S_h$. Note that $S_h^2 = (N_h-1)^{-1}\sum_{i\in U_h}(y_i - \bar{Y}_h)^2 = N_h^{-1}(N_h-1)^{-1}\sum_{i\in U_h}\sum_{j\in U_h}(y_i-y_j)^2$. Let $d_{ij} = (y_i-y_j)^2$. Minimizing $Q = \sum_{h=1}^{H}\left(\sum_{i,j\in U_h} d_{ij}\right)^{1/2}$ is essentially the same as minimizing $\sum_{h=1}^{H} N_h S_h$. One way to minimize $Q$ is through a hierarchical algorithm as follows where $H*$, the number of desired stratum is given.

1. Set $H = N$.
2. Compute the distance between strata by $d^*_{h,h'} = \left(\sum_{(i,j)\in U_h\cup U_{h'}} d_{ij}\right)^{1/2} - \left(\sum_{(i,j)\in U_h} d_{ij}\right)^{1/2} - \left(\sum_{(i,j)\in U'_h} d_{ij}\right)^{1/2}$
3. Find the pair with the smallest value of $d^*_{h,h'}$. Merge strata $h$ and $h'$. Then we now have $H-1$ partitions because of merge. Set $H = H-1$
4. Go back to Step 2. Continue until $H = H^*$.

In practice, the study variable $y_i$ is unknown for the population. Assume we have $x_i$ which is closely related to the variable $y_i$. Also consider that both $x_i$ and $y_i$ may be vectors of length $p$. Let $\Sigma = \mathrm{Cov}(\boldsymbol{x})$. Now, we can define $d_{ij} = (x_i - x_j)'\Sigma^{-1}(x_i - x_j)$. Then the algorithm follows as above. Note, that if $x_i$ is univariate that this is equivalent to using squared difference which is the motivation. $\Sigma^{-1}$ is used to weight distances by the inverse of their variance which is the square of the Mahalanobis distance. Other distances will be considered in the future. For example, $d_{ij} = \left(\sum_{k=1}^{p}(x_{ik}-x_{jk})^4\right)^{1/4}$ or $d_{ij} = \max\{d_{ij1}/m_1, \ldots, d_{ijp}/m_p\}$ where $d_{ijk} = (x_{ik}-x_{jk})^2$ and $m_k = \max\{d_{ijk} : 1 \le i \le n, 1 \le j \le n\}$.

## 6. Numerical Studies

We first compare the three algorithms discussed on simulated data so that we know the truth. Then we used CDL data to stratify a few states for an agricultural survey. We use the NASS's current methodology as well as the proposed method to stratify the segments. We then compare the sample sizes necessary to achieve target CVs of estimates. sadgsag

### 6.1 Comparing Stratification Algorithms using Generated Data

Since our segments are bounded, we created a population of segments that are each 100 square units. Let $y_1$ and $y_2$ be the respective areas for Crop 1 and Crop 2 and $y_3$ be the area of the rest of the segment. Then $z = y_1 + y_2$ is the total crop acreage. We will construct the population as follows

$$x_1 \sim \chi^2_1$$
$$x_2 \sim \chi^2_1$$
$$x_3 \sim \chi^2_1$$

where $x_i \perp x_j$. Then let

$$y_i = 100\frac{x_i}{\sum_{j=1}^{3} x_j}$$

The sample size is $N = 5,000$. Our goal is to stratify the population using different algorithms and compare the variance of the estimate of the total of the three variables $y_1$, $y_2$, and $y_3$. The population was stratified univariately by each of $y_1$, $y_2$, and $z$ using three different methods - Lavallée and Hidiroglou, Dalenius-Hodges, and the new method. Additionally, we stratified the population using the new multivariate method with $p = 0.5$. In one of the multivariate stratifications, we used $\Sigma = I$ and the other we used the true variance. We used Mahalanobis distance for the unit level distance. For each of the stratifications, ten strata were used.

To compare the stratifications, we used design effect which is the ratio of the variance under the stratified design to the variance under a simple random sample. In Figure 2, the design effects are illustrated using a bar graph. The top-left panel illustrates the three algorithms that used only $y_1$ to stratify. Since $y_1$ was used in the stratification, the design effect for $y_1$ is very low. However, $y_2$ and $z$ have a design effect of greater than 1 for each of the algorithms which indicates that a simple random sample has a lower variance for these values than a design using solely $y_1$. Similar results are seen when using $y_2$ and $z$ solely in stratification in the top-right and bottom-left panel, respectively.

When using the multivariate stratification, which has results illustrated in the right-left panel of Figure 2, the design effect for all three variables is less than 0.2 for all variables using $\Sigma = I$ and less than 0.1 for all variables when using the true variance. Since, we want to be able to estimate both the total of $y_1$ and $y_2$ with similar precision, the multivariate stratification is a good compromise in reduction of the design effect.

### 6.2 Stratifying PLSS Sections using Current Method and the Proposed Method

For five study states, we will be comparing the NASS's current method of stratification to the proposed method. We will use sample size as a comparison between the stratifications where a smaller sample size is desired. The Public Land Survey System (PLSS), which is managed by the Bureau of Land Management, will serve as a sampling frame. We will use its one-square mile sections as the Primary Sampling Unit (PSU). One of the study states, Pennsylvania, is not covered by the PLSS so we lay down a grid on that state with square mile segments to mimic the PLSS elsewhere in the country. For each of these sections, we have CDL information. The CDL has categories such as corn, soy, durum wheat, urban, forest, and water. It is an estimate but, at the national level, is 85%-95% accurate for major crops. The CDL data in this study is from 2011.

The current stratification method used by the NASS has strata that are defined by percent agriculture and urban. An example of the stratification for Pennsylvania is in Table 1. Since we are going to consider the PLSS as our frame, we will not have variable segment size depending on strata. Using the CDL data, each PLSS section's composition of cultivation and urban is determined and the section is put into a strata of class 10,20,30, 40 or 50.

For the new method, in each of the states, percent corn, percent wheat, percent cotton (when applicable), and percent cultivated are derived for each PLSS section. Then, we used previous JAS data to model JAS responses on CDL data. These estimated models were used to then estimate the acreage in each section for the quantities mentioned before. We used these as inputs into the hierarchical stratification algorithm with number of strata varying from 2 to 20.

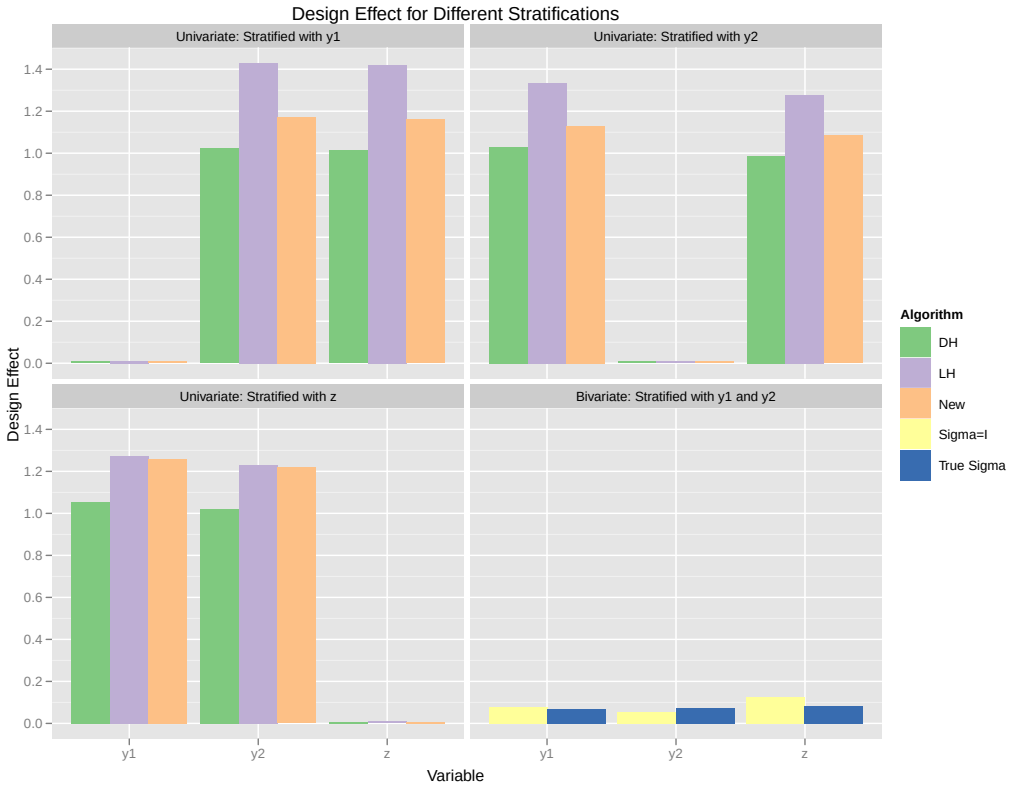


**Figure 2**: Design effect on each of the variables under each stratification.

| Stratum | Target Segment Size | Stratum Definition |
|---|---|---|
| 13 | 1.00 | 50%+ Cultivated |
| 20 | 1.00 | 15-50% Cultivated |
| 31 | 0.25 | Agri-urban: 100+ Homes per Sq Mi |
| 32 | 0.10 | Commercial: 100+ Homes per Sq Mi |
| 40 | 2.00 | 0-15% Cultivated |
| 50 | pps | Non-ag |

**Table 1**: Example of strata definitions for Pennsylvania currently used by the NASS.

Sample sizes were allocated to strata using the same method after the two types of stratifications were complete. The NASS has target CVs for estimates at a national level. However, we are only looking at a small set of states. To overcome this, we calculated the CVs that were achieved in 2010 on the state level for each of these states and made these the target CVs for this study. The allocation of the sample is a problem in convex programming. An iterative, nonlinear programming algorithm is used to provide the sample allocations. The algorithm used is guaranteed to converge [2].

To compare the two stratifications, we can look at the total sample size required to achieve the target CVs. In Figure 3, the sample sizes required to reach the target CVs are plotted where the sample size is $\sum_{h=1}^{H} n_h$. In all states except for Indiana, we can achieve the target CVs with a smaller sample size with the hierarchical method than the current method used by the NASS. We hypothesize the reason for Indiana being different is that its agriculture is fairly homogeneous and thus a multivariate stratification considering multiple crops is similar to a univariate stratification only considering percent cultivation

and percent urban.

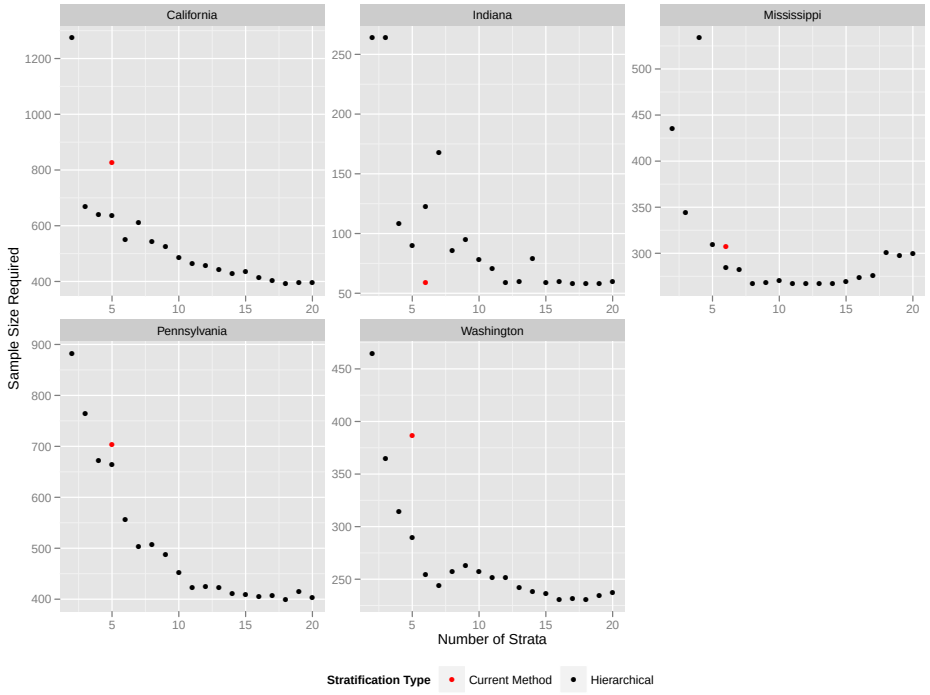


**Figure 3**: Sample sizes required to achieve target CVs using current method and hierarchical method of stratification with number of strata varying for the hierarchical method.

## 7. Conclusion and Future Work

Multivariate stratification is promising for multi-purpose surveys. We plan to examine the effect of different unit-level distance functions, the $d_{ij}$ as well as the allocation rule, which comes down to the choice of your $p$. As with all allocations, it depends on whether you would like to accurately estimate within strata or across strata.

For the NASS survey, we also intend on using the 2012 CDL, which none of the stratification depended on, to act as a proxy and estimate variances of estimation for quantities that the NASS is interested in. We will also examine how to incorporate geography into the stratification and possibly using geographic distance in the distance function for the hierarchical algorithm.

## References

[1] USDA-NASS-RDD spatial analysis research section. `http://www.nass.usda.gov/research/Cropland/SARS1a.htm`. Accessed: 09/18/2012.

[2] J. Cotter, C. Davies, J. Nealon, and R. Roberts. *Agricultural survey methods*, chapter Area frame design for agricultural surveys, pages 169–192. Wiley, 2010.

[3] T. Dalenius and J.L. Hodges Jr. Minimum variance stratification. *Journal of the American Statistical Association*, 54(285):88–101, 1959.

[4] P. Gunning, J. Horgan, and W. Yancey. Geometric stratification of accounting data. *Contaduría y Administración*, (214), 2009.

[5] W. Han, Z. Yang, L. Di, and R. Mueller. Cropscape: A web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computer and Electronics in Agriculture*, 84:111–123, 2012.

[6] P. Lavallée and M. Hidiroglou. On the stratification of skewed populations. *Survey Methodology*, 14(1):33–43, 1988.

[7] L.P. Rivest. A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28(2):191–198, 2002.

[8] VK Sethi. A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5(1):20–33, 1963.