# Automatic stratification for an agricultural area frame using remote sensing data

Stephanie Zimmer[1,2], Jae Kwang Kim[1], and Sarah Nusser[1]

[1]Iowa State University, Ames, IA, USA
[2]Corresponding Author: Stephanie Zimmer, e-mail: sazimme2@iastate.edu

## Abstract

The National Agricultural Statistics Service (NASS) is responsible for conducting monthly and annual surveys and preparing official USDA data and estimates of production, supply, prices, and other information necessary to maintain orderly agricultural operations. One survey is the June Area Survey which utilizes an area sampling frame to collect data used to supply direct estimates of acreage and measures of sampling coverage. The population consists of all land in the USA, except Alaska, which is stratified for sampling. Currently, the stratification is labor intensive because segments on the land are hand drawn and stratified by individuals. Our goal is to make this process automatic by using permanently defined segments and auxiliary data from remote sensing, the Cropland Data Layer which classifies satellite imagery into crop types and non-agriculture categories. Using this data, we propose treating stratum identifier as a missing data value and use an expectation-maximization (EM) algorithm to assign the stratum indicator. Each segment is comprised of many pixels and segments of similar composition will be grouped together using the algorithm.

Keywords: EM algorithm, Missing data, Multivariate stratification

## 1 Introduction

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, the 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 Cotter et al. (2010). The process in making each area frame is very costly because it is labor intensive and time intensive with, on average, five full-time employees taking four months to construct one state's frame. The stratification uses many types of maps including topographic maps, Tele Atlas map, and National Agriculture Imagery Program for boundaries as well as satellite imagery and the Cropland Data Layer (CDL) to assist in the stratification Cotter et al. (2010).

There are two main problems in the current stratification. The first is that the area frame is not a permanent frame and needs updating every few years. The second is that it is a labor intensive process to manually create strata which also determine the size of the segments. Thus, our first goal will be to create a permanent sampling frame. The frame should still cover the entirety of the United States and have segment sizes that are reasonable to collect the agricultural data by enumerators in an efficient manner.

The problem of stratification is the major problem discussed in this paper. We will introduce what the goal of stratification is. We will then discuss methods that have been used in univariate stratification. Then, we will introduce two new methods we propose. One is a hierarchical clustering method and the other is an E-M algorithm.

## 2 Developing a Permanent Frame and Auxiliary Data

We propose using the Public Land Survey System (PLSS) as a permanent sampling frame which exists in most states and is comprised of approximately 1 square mile sections. For the other states, we can lay down a square mile grid on the map using GIS tools to create similar shaped sections for the entire country. For each of these sections, whether from the PLSS or created by laying down the grid, we obtain auxiliary data from the CDL.

Upon creating this frame, we will stratify the frame units using auxiliary data automatically using a stratification algorithm. The goal is that this stratification will be able to meet current goals set by NASS. One specific goal is to attain small coefficient of variation (CVs) for various measures which are specified by NASS. These measures include crop acreages for major crops. It also includes number of farms. A third measure is not-on-list cattle which are cattle that do not appear in NASS's list frame, a frame of operations and operators maintained by NASS which is separate from the area frame. The CDL uses satellite images to classify land into different types of land at a pixel level and thus we could know an estimate of the different types of land in the sampling units from the frame. The CDL uses satellite images to classify land into different types of land at a pixel level with pixels being 30m by 30m USDA-NASS-RDD (2013); Han et al. (2012). By summarizing 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. For each of the sections in the PLSS, we can overlay the CDL data and thus have a summary of CDL data for each section. Thus, the data we will have to stratify will be a list of sections and the CDL data associated with each section. The next question we address is how to stratify the frame units using this information.

## 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 estimate 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 $n_h \propto N_h S_h$. Other allocations could be used.

# 4 Current Univariate Stratification Methods

We use auxiliary data to create strata. If the auxiliary data is categorical, for example gender or crop type, strata are constructed naturally. However, if we use a continuous variable for stratification, it is not as obvious how to form strata. In their paper, Dalenius and Hodges (1959) Dalenius and Hodges Jr (1959) 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.

We focus on this specification because we consider it the most difficult problem. For our application, we already know which stratification variables exist. The choice of number of strata and sample size within strata can be investigated after we determine the way in which the population is to be stratified. We describe two univariate stratification methods found in literature. We focus on univariate methods because little was found in the literature on multivariate stratification with the exception of work done by Benedetti and Piersimoni in 2012 on multivariate stratification. Their method creates only two strata where one is completely enumerated Benedetti and Piersimoni (2012).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.

## 4.1 Lavallée and Hidiroglou Algorithm

The Lavallée and Hidiroglou algorithm (1988) Lavallée and Hidiroglou (1988) was intended to stratify a univariate, skewed population. Particularly, it was proposed for establishment surveys where a few establishments are very large and important in the estimate so are sampled with certainty. When given a number of strata, $H$, the algorithm chose breakpoints such that $y_{(0)} < b_1 < b_2 < \cdots < b_{H-1} < y_{(N)}$ 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 $H-1$ stratum. Some standard notation that will be used is $H$ is the number of strata; $W_h = N_h/N$ for $h = 1,\ldots,H$ 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 $h = 1,\ldots,H$ 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 $-\infty < b_0 < b_1 < \cdots < b_{H-1} < b_H = \infty$ are the stratum boundaries. Since $n_H = N_H$, the sample size in the other stratum can be written as $(n - N_H)a_h$ where $n$ is the total sample size and $a_h$ define the allocation such that $\sum_{h=1}^{H-1} a_h = 1$ and $a_h > 0 \; \forall \; 1 \le h \le H-1$. 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 (2) for $n$ with $W_h = N_h/N$ leads to

$$n = NW_H + \frac{\sum_{h=1}^{H-1} W_h^2 S_{yh}^2 / a_h}{\text{Var}(\bar{y}_{st}) + \sum_{h=1}^{H-1} W_h S_{yh}^2 / N} \tag{3}$$

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

### 4.2 Dalenius and Hodges Algorithm

Another algorithm used to stratify continuous variables was proposed in 1959 by Dalenius and Hodges. The algorithm proposed by Dalenius and Hodges minimizes variance for a stratified sampling design under some assumptions which includes an assumption that variance does not vary much from stratum to stratum. Dalenius and Hodges show that if the root frequency of groups are equal then minimum variance is achieved.

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 $H$ denote the number of strata. The following algorithm creates strata with approximately equal frequencies in each stratum.

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: $Q = \frac{1}{H}\sum_{i=1}^{J} \sqrt{f_i}$
7. Take the upper boundaries of each stratum to be the $X$ values corresponding to
   $Q, 2Q, 3Q, \ldots, (H-1)Q, HQ$.

Other stratification methods include the method introduced by Gunning et al. (2009) which uses the idea of cumulative root frequency from Dalenius and Hodges but removes the arbitrary choice of number of classes, $J$.

## 5 Proposed Methods

Both of the proposed methods are motivated by minimizing the variance of the mean estimate. We first discuss the univariate case and then an extension to the multivariate case. To do this, we re-write the variance formula as follows where $\boldsymbol{y}$ is some auxiliary data vector correlated with the response. It can be shown that $V(\bar{y}_{st}) \approx N^{-2}\sum_{h=1}^{H} \frac{1}{n_h} \sum_{i \in U_h} \sum_{j \in U_h} (y_i - y_j)^2$.

If we use Neyman allocation, that is $n_h \propto N_h S_h$, it can be shown that
$V(\bar{y}_{st}) \approx N^{-2}\sum_{h=1}^{H} \left(\sum_{i \in U_h} \sum_{j \in U_h} (y_i - y_j)^2\right)^{1/2} \propto \sum_{h=1}^{H} \left(\sum_{i \in U_h} \sum_{j \in U_h} (y_i - y_j)^2\right)^{1/2}$
$= \sum_{h=1}^{H} \left(\sum_{i \in U_h} \sum_{j \in U_h} d_{ij}\right)^{1/2} = \sum_{h=1}^{H} Q_h = Q$ where $d_{ij} = (y_i - y_j)^2$ and $Q_h = \left(\sum_{i \in U_h} \sum_{j \in U_h} d_{ij}\right)^{1/2}$. If we are able to minimize, the function $Q$ then we will also minimize the variance estimate.

### 5.1 Hierarchical Method

One way we propose to minimize $Q$ is through a hierarchical algorithm as follows where $H^*$, the number of desired stratum is given. This starts with each observation in its own strata and then merges strata step by step. Strata are merged when they are closest to each other. The idea of closeness is in the formula in Step 2. The steps for the algorithm are as follows:

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 $H = H^*$.

The hierarchical algorithm must go through $N - H^*$ steps where $N$ may be large and $H^*$ small and thus the number of iterations will be large. This means the algorithm will run slower than the other algorithms presented.

## 5.2 E-M Method

We investigated how k-means performed for stratification even though it does not have the same objective function. In k-means, we are trying to estimate the group means and it is the solution to

$$E\left[\sum_{i=1}^{N}\sum_{h=1}^{H}(y_i - \boldsymbol{x}_i^T\boldsymbol{\mu})^2\boldsymbol{x}_{ih}|\boldsymbol{y}\right] \tag{4}$$

where $y_i|i \in U_h \sim N(\mu_h, \sigma^2)$ and

$$x_{ih} = \begin{cases} 1 & \text{if } i \in U_h \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Thus $\boldsymbol{x}_i$ is the indicator vector of the strata. If we knew $X$ then $\hat{\boldsymbol{\mu}} = (X'X)^{-1}X'y$ but we will first estimate $X$ and then $\boldsymbol{\mu}$ alternately where estimating $X$ is the Expectation step and estimating $\boldsymbol{\mu}$ is the Maximization step Dempster et al. (1977) which is the k-means algorithm. This is minimizing SSE which, in survey statistics terms would be $\sum N_h S_h^2$ but we wish to minimize $\sum N_h S_h \propto Q$.

We propose minimizing the similar term $\sum N_h^2 S_h^2$. If we do this, we can view the problem as $y_i|i \in U_h \sim N(\mu_h, W_h^{-1}\sigma^2)$ where $W_h \propto N_h$ and thus we will wish to minimize

$$E\left[\sum_{i=1}^{N}\sum_{h=1}^{K}N_h(y_i - \boldsymbol{x}_i^T\boldsymbol{\mu})^2\boldsymbol{x}_{ih}|\boldsymbol{y}\right]. \tag{6}$$

Under this model, the log-likelihood of $(\boldsymbol{\mu}, N_1, \ldots, N_g, x_i)$ is as follows

$$l(\boldsymbol{\mu}, N_1, \ldots, N_h, x_i) = \sum_{i=1}^{N}\sum_{h=1}^{H} x_{ih}\log\left[f_h(y_i|\mu_h, N_h)\right]. \tag{7}$$

where

$$f_h(y_i|\mu_h, N_h) = \left(\frac{2\pi\sigma^2}{N_h}\right)^{1/2}\exp\left\{-\frac{N_h}{2\sigma^2}(y_i - \mu_h)^2\right\}. \tag{8}$$

Then, we get the estimate of $\hat{x}_{ih}$ as

$$P(i \in U_h|y_i) = \hat{x}_{ih} = \frac{\hat{N}_h f_h(y_i|\hat{\mu}_h, \hat{N}_h)}{\sum_{k=1}^{H}\hat{N}_k f_k(y_i|\hat{mu}_k, \hat{N}_k)} \tag{9}$$

We use fractional imputation to estimate the parameters and the strata classification. Since an observation can only be in one strata, we can create weights for each of the $H$ possibilities Kim (2011). Thus, for each observation $i$, we have $H$ pairs of data $(\boldsymbol{x}_i^{(h)}, y_i)$ for $h = 1, \dots, H$ and a corresponding weight for each weight where the weight is

$$w_{ij}^* \propto f(y_i|\boldsymbol{x}_i^{*(j)})\pi(\boldsymbol{x}_i^{*(j)}) = f(y_i|\boldsymbol{x}_i^{*(j)}; \hat{\theta})\hat{\pi}(\boldsymbol{x}_i^{*(j)}) \tag{10}$$

with $\sum_{j=1}^H w_{ij}^* = 1$, $f$ as defined above, and $\pi(\boldsymbol{x}_i^{*(j)}) = P(i \in U_h|y_i)$. Once we have weights, we can estimate the parameters as

$$\hat{\mu}_j = \frac{\sum_{i=1}^N w_{ij}^* y_i}{\sum_{i=1}^N w_{ij}^*} \tag{11}$$

$$\hat{\pi}_j = \frac{\sum_{j=1}^N w_{ij}^*}{N} \tag{12}$$

$$\hat{\sigma}^2 = \sum_{j=1}^H \sum_{i=1}^N w_{ij}^* (y_i - \hat{\mu}_j)^2 \hat{\pi}_j \tag{13}$$

We iterate estimating $w_{ij}^*$ and the parameters until convergence and choose the strata for observation $i$ with the largest weight. Initial values for $\hat{\boldsymbol{\mu}}$ could be from k-means.

## 6 Conclusion and Future Work

We propose using automatic stratification to stratify NASS's area frame. This will be efficient in terms of labor and time. It will also be statistically efficient if the CDL is a good approximation of what is on the land. We will compare our two new methods to the existing stratification methods. We will also compare using the automatic stratification to the current stratification methodology used by NASS.

## References

Benedetti, R. and Piersimoni, F. (2012). Multivariate boundaries of a self representing stratum of large units in agricultural survey design. *Survey Research Methods*, 6(3):125–135.

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

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

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

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

Han, W., Yang, Z., Di, L., and Mueller, R. (2012). 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.

Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1):119–132.

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

USDA-NASS-RDD (2013). USDA-NASS-RDD spatial analysis research section. `http://www.nass.usda.gov/research/Cropland/SARS1a.htm`. Accessed September 18, 2012.