# Optimal Stratification and Allocation for the June Agricultural Survey

Jonathan Lisic [1] Hejian Sang [2] Zhengyuan Zhu [2]
Stephanie Zimmer [2]

[1]United States Department of Agriculture National Agricultural Statistics Service, Washington, D.C. 20250, U.S.A.

[2]Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.

December 2nd, 2015

# Overview

This presentation will cover:

- Optimal stratification and allocation through simulated annealing under coefficient of variance and fixed sample size constraints.
- Application to simulated data.
- Application to the June Agricultural Survey.

# Background

The June Agricultural Survey (JAS) is an annual area survey of agriculture over the contiguous 48 states.

- Stratification is performed on a state-by-state basis.
- Characteristics of interest include major commercial crop acreages (corn, soybeans, winter wheat, etc. . . ).
- Sampling units (segments) are approximately one square mile in size (up to 268,518 segments in Texas).
- Characteristics are not necessarily correlated with each other.
- Target CVs are set for estimates, not administrative data.
- Highly correlated covariates available through remote sensing for crops.
- Fixed sample size.

# Background

The current stratification is non-optimal:

- Strata are formed through univariate bounds on cultivated acreage within segments.
- Stratification is not based on the characteristics of interest.
- Optimal allocation is performed given a stratification.

# The Problem

How do you create an optimal design under quality and sample size constraints?

# Prior Approaches

- The problem has been addressed by Dalenius and Hodges (1959) and Lavallée and Hidiroglou (1988), for the specific case of two stratum (one census and one non-census).
- Lavallée and Hidiroglou (1988) formed strata through univariate thresholding, e.g. establishments greater than 100 people.
- This work has been extended to multiple dimensions (see Benedetti and Piersimoni, 2012), but not to more strata.
- The multivariate extension initially forms boundaries through univariate thresholding each of the characteristics being sampled.
- This boundaries are relaxed through a sequence of exchanges.
- Require strong population asymmetry and the sample size cannot be fixed.

# Approach

- Use existing, computationally efficient, machine learning methods to form an initial stratification.
- Use simulated annealing to both obtain an optimal sample allocation and provide a stratification aligned with our desired objective function.
- The approach taken does not require strong population asymmetry, but requires the sample size to be fixed (potentially empty feasible region).

# Objective Function

How do you define an objective function if you have vector valued CVs, $\hat{c} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_J)$, and targets $c = (c_1, c_2, \ldots, c_J)$?

$$\hat{c}_j = \frac{\sqrt{S_j^2}}{\bar{y}_j}$$

where $y$ is the set of PSUs with fully observed administrative data indexed by $j$.

# Objective Function

We apply a penalized objective function, with penalty $\lambda$:

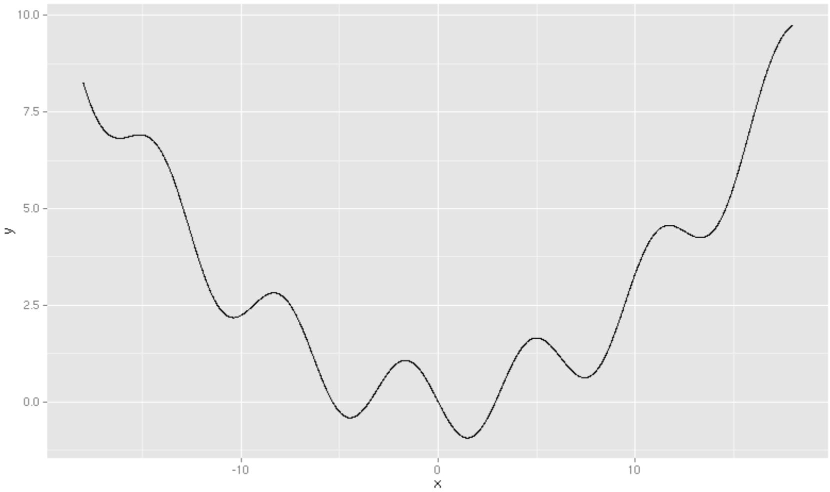
$$||\hat{c}||_2^2 + \lambda||\hat{c} - c||_{2+}^2 \tag{1}$$

- This objective function penalizes departures from the vector valued target CVs.
- The function $||x||_{2+} = \left(\sum_{j=1}^{J} x_j^2 \mathbb{I}_{x_j>0}\right)^{1/2}$.
- This approach is "soft" in that it does not have "hard" CV constraints.

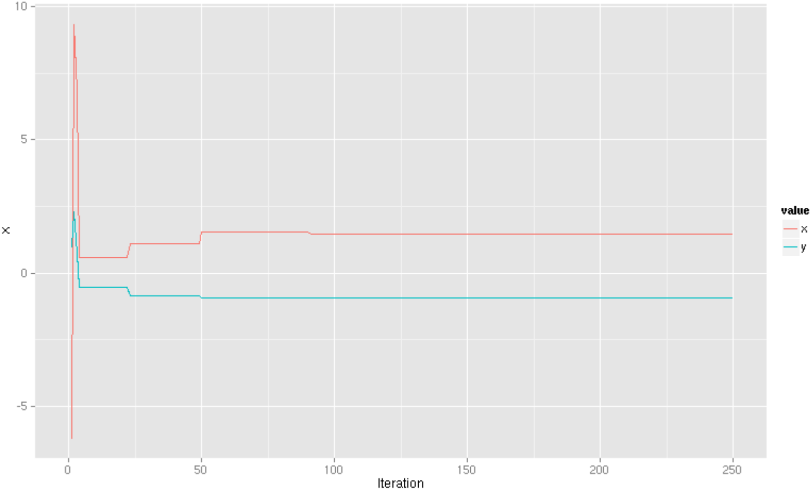
# Simulated Annealing

- Simulated annealing is a stochastic optimization process that minimizes an objective function (possibly with constraints).
- Avoids the pitfalls of ending up in a local maxima by admitting non-optimal states.
- The general form of an algorithm to perform this stochastic process is:
  1. Start with initial state $X_0$;
  2. Randomly generate a candidate state $Y_l$, $l \geq 1$;
  3. If $Y_l$ has a lower objective function than $X_{l-1}$, set $X_l = Y_l$;
  4. Else accept $Y_l$ with probability $\rho = exp\{\Delta h_l / t(l)\}$ otherwise $X_l = X_{l-1}$ $(\Delta h_l = X_{l-1} - X_l)$ ;
  5. Go back to Step 2. until a threshold of iterations has been met.

# Simulated Annealing (Example 1)



# Simulated Annealing (Example 1)



# Simulated Annealing

1. Start with initial stratification $\mathcal{I}^{(0)}$ and allocation $\eta^0$;
2. Randomly generate a candidate state $\mathcal{I}_*^{(l)}$, $l \geq 1$;
3. Randomly generate a candidate state $\eta_*^{(l)}$ (possibly the same as the prior state);
4. If $\left(\mathcal{I}_*^{(l)}, \eta_*^{(l)}\right)$ has a lower objective function than $\left(\mathcal{I}^{(l-1)}, \eta^{(l-1)}\right)$, set $\left(\mathcal{I}^{(l)}, \eta^{(l)}\right) = \left(\mathcal{I}_*^{(l)} \eta_*^{(l)}\right)$;
5. Else accept $\mathcal{I}_*^{(l)}$ with probability $\rho = exp\{\Delta h_l / t(l)\}$ otherwise $\mathcal{I}^{(l)} = \mathcal{I}^{(l-1)}$;
6. Go back to Step 2 until a threshold of iterations has been met.

In this application $t(l) = \alpha(l+1)^{-1}$ where $\alpha$ is a tuning parameter.

# Simulated Annealing (Example 2, Iteration 0)

| Index | Strata | x | y |
| --- | --- | --- | --- |
| 1 | 1 | 2.3 | 72 |
| 2 | 1 | 2.5 | 55 |
| 3 | 1 | 2.1 | 42 |
| 4 | 1 | 2.8 | 61 |
| 5 | 1 | 2.9 | 68 |
| 6 | 2 | 4.9 | 58 |
| 7 | 2 | 5.1 | 44 |
| 8 | 2 | 4.2 | 51 |
| 9 | 2 | 2.8 | 48 |
| 10 | 2 | 4.3 | 52 |

For sample size $n = (3, 3)$, $\lambda = 100$, $\alpha = 1$,
$\hat{c} = (0.082, 0.068)$, $c = (0.050, 0.100)$,
objective function $= 3.307$.

# Simulated Annealing (Example 2, Iteration 1)

| Index | Strata | x | y |
|---|---|---|---|
| 1 | 1 | 2.3 | 72 |
| 2 | 1 | 2.5 | 55 |
| **3** | $\mathbf{1} \rightarrow \mathbf{2}$ | 2.1 | 42 |
| 4 | 1 | 2.8 | 61 |
| 5 | 1 | 2.9 | 68 |
| 6 | 2 | 4.9 | 58 |
| 7 | 2 | 5.1 | 44 |
| 8 | 2 | 4.2 | 51 |
| 9 | 2 | 2.8 | 48 |
| 10 | 2 | 4.3 | 52 |

For sample size $n = (\mathbf{2}, \mathbf{4})$,
$\hat{c} = (0.108, 0.050)$, $c = (0.050, 0.100)$,
objective function $= 5.919$, and $\rho = 0.271$.

# Simulated Annealing (Example 2, Iteration 2)

| Index | Strata | x | y |
|---|---|---|---|
| 1 | 1 | 2.3 | 72 |
| 2 | 1 | 2.5 | 55 |
| 3 | 2 | 2.1 | 42 |
| 4 | 1 | 2.8 | 61 |
| 5 | 1 | 2.9 | 68 |
| 6 | 2 | 4.9 | 58 |
| 7 | 2 | 5.1 | 44 |
| 8 | 2 | 4.2 | 51 |
| **9** | $\mathbf{2} \rightarrow \mathbf{1}$ | 2.8 | 48 |
| 10 | 2 | 4.3 | 52 |

For sample size $n = (\mathbf{2}, \mathbf{4})$,
$\hat{c} = (0.092, 0.069),\ c = (0.050, 0.100)$,
and objective function $= 4.315 < 5.919$.

# Simulated Annealing

Why would this work?

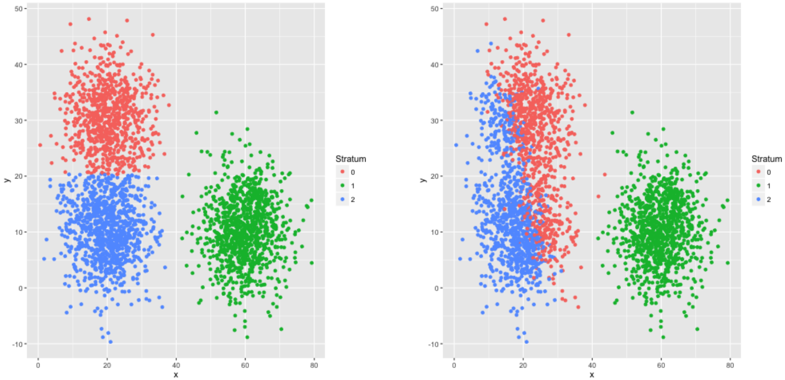
- Each move is reversible, ensuring that for an infinitely long run time with exact precision the method will converge to the global minima.
- For large populations with small sample sizes, there is little change needed to retain optimal allocation for single PSU exchanges.
- Furthermore, if a large change in optimal allocation needs to occur after a single PSU exchange, that PSU probably shouldn't be moved.

# Simulation

A simulation was performed with two population sizes, N=2,800 and N=280,000, both with sample size 60.

- $\frac{N_1}{N} = \frac{8}{28}$, $\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$,
  $\mu_1 = (60, 10)$, and $\Sigma_1 = \begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix}$.
- $\frac{N_2}{N} = \frac{10}{28}$, $\mathbf{x}_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$,
  $\mu_2 = (20, 10)$, and $\Sigma_2 = \begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix}$.
- $\frac{N_3}{N} = \frac{10}{28}$, $\mathbf{x}_3 \sim \mathcal{N}(\mu_3, \Sigma_3)$,
  $\mu_3 = (20, 30)$, and $\Sigma_3 = \begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix}$.
- $\lambda = 10,000$.
- $c = (0.020, 0.070)$.

# Simulation (N=2,800)

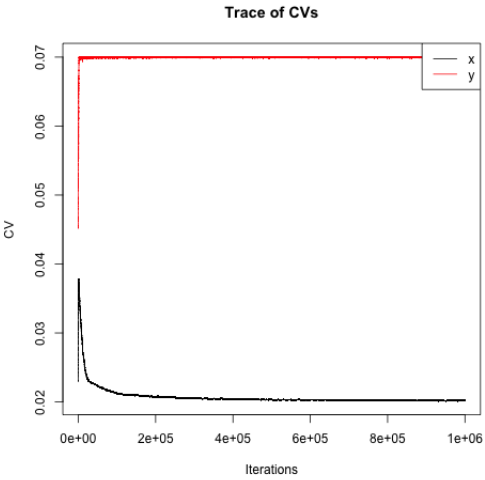


# Simulation (N=2,800)

| Target | Univariate Optimal $X$ | K-means Optimal Alloc. | Simulated Annealing |
|---|---|---|---|
| 0.020 | 0.017 | 0.023 | 0.020 |
| 0.070 | 0.084 | 0.050 | 0.070 |

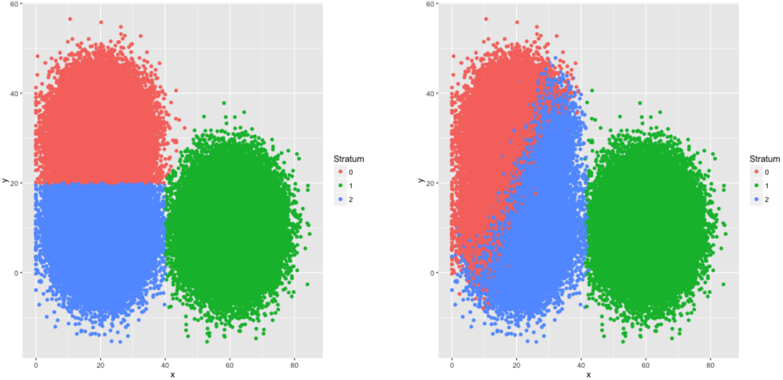
Table: Attained CVs for simulated population size of 2,800.

# Simulation (N=2,800)




Run Time = 7 seconds for 1,000,000 iterations

# Simulation (N=280,000)

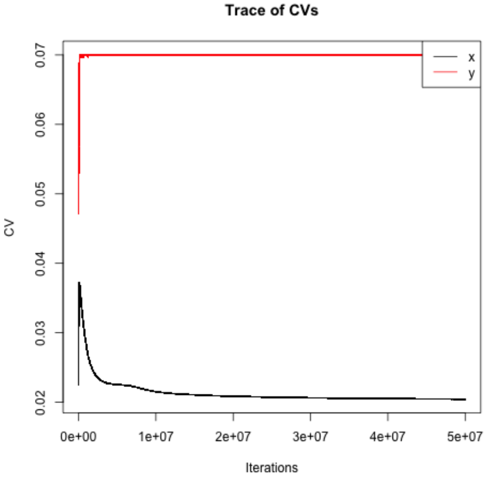


# Simulation (N=280,000)

| Target | Univariate Optimal $X$ | K-means Optimal Alloc. | Simulated Annealing |
|---|---|---|---|
| 0.020 | 0.017 | 0.022 | 0.020 |
| 0.070 | 0.089 | 0.047 | 0.071 |

Table: Attained CVs for simulated population size of 280,000.

# Simulation (N=280,000)




Run Time = 3.0 hours for 50,000,000 iterations

# Speed and Stability

That's a lot of iterations!

- Computational Speed:
  - Variances are saved and only updated on accepted exchanges.
  - Variances and updated using online methods.
- Computational Stability:
  - After a fixed number of iterations variances are recalculated from current strata assignments.

# More Speed

Can we make this faster?

- Most successful exchanges occur near the initial boundaries between stratum from the applied machine-learning methods.
- Weighting can be applied to increase the number of exchanges near the boundaries relative to other locations.

# June Agricultural Survey

This method was tested on South Dakota.

- Target crops included cultivated acreage, corn, soybeans, winter wheat and spring wheat.
- Survey using covariate data from 2013-2014.
- Each year-by-administrative variable pair is treated as a distinct administrative variable.
- 2015-2019 response is simulated using the 2008-2012 Cropland Data Layer(CDL) (see Boryan et al., 2011).
- The algorithm was run for 5,000,000 iterations.

# June Agricultural Survey

Results:

| | Cultivated | Corn | Soybeans | Winter Wht. | Spring Wht. |
|---|---|---|---|---|---|
| Target | 0.01 | 0.05 | 0.05 | 0.19 | 0.16 |
| 2013 | 0.01 | 0.03 | 0.04 | 0.09 | 0.09 |
| 2014 | 0.01 | 0.02 | 0.04 | 0.07 | 0.07 |
| *2015 | 0.02 | 0.04 | 0.05 | 0.10 | 0.10 |
| *2016 | 0.02 | 0.04 | 0.05 | 0.10 | 0.10 |
| *2017 | 0.02 | 0.04 | 0.05 | 0.10 | 0.12 |
| *2018 | 0.02 | 0.04 | 0.04 | 0.10 | 0.11 |
| *2019 | 0.02 | 0.04 | 0.04 | 0.12 | 0.12 |

*Using CDL data from prior years.

# Open and Reproducible Research

R package available at https://github.com/jlisic/saAlloc.

# Future Work

- Consider moving to more efficient methods such as differential evolution (see Day, 2009).
- Investigate adaptive methods for weighting.
- Consider alternatives to moving a single PSU, maybe hyperplanes?
- For JAS, understand the relationship between the administrative data CVs and the estimate CVs.
- For JAS, consider ways to predict future land cover.

# Thank You

# References I

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

Claire Boryan, Zhengwei Yang, Rick Mueller, and Mike Craig. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.

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

# References II

Charles D Day. Evolutionary algorithms for optimal sample design. In *A paper presented at the 2009 Federal Committee on Statistical Methodology Conference, Washington, DC*, 2009.

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