Model-Based Optimal Designs for a Multipurpose Farm Survey

Jay Breidt

NORC at the University of Chicago

IASS Webinar

April 24, 2024

Joint work with Ben Reist and Ruochen Ma (NORC), and Lu Chen (NISS/NASS), with thanks to other NORC and NASS colleagues

Ongoing collaboration to improve survey methods

- Abreu, Denise
- Benecha, Habtamu
- Black, Alison
- Boim, Jason (NORC)
- Broz, Terry
- Cheng, Yang
- Dau, Andrew
- Drunasky, Lindsay
- Duan, Franklin
- Emmet, Robert
- Gerling, Michael
- Gibson, Fleming
- Herbek, Greg
- Keller, Tim
- Murphy, Tara
- Olbert, Everett
- Robinson, Tina
- Russell, Charles
- Sarkar, Bayazid
- Sartore, Luca
- Scherrer, Cathryn
- Smith, Holly
- Smith, Leslie
- Vance, Wendy
- Wing, Taylor (NORC)
- Zhang, Ruiyi

National Agricultural Statistics Service (NASS)

- One of 13 principal statistical agencies in the decentralized US federal statistical system
- NASS is the survey and estimation arm of the US Department of Agriculture
  - conducts Census of Ag every five years
  - fields hundreds of surveys each year
  - compiles extensive administrative data
  - covers nearly every aspect of US agriculture

Common problem in establishment surveys

- Frame = list of establishments: $U = \{1, 2, \ldots, k, \ldots, N\}$
  - assume complete coverage for purposes of this talk
- Characteristics of interest: $C = \{1, 2, \ldots, J\}$:

$$
\begin{aligned}
y_{jk} & \quad \text{for characteristic } j \in C \text{ on establishment } k \\
T_{yj} &= \sum_{k \in U} y_{jk}
\end{aligned}
$$

- Characteristics have different constraints: $C = C_0 \cup C_1 \cup C_2$
  - C_0: no specified constraints
  - C_1: specified precision targets
  - C_2: specified other constraints
- Frame measures of size (MOS) for $j \in C_1 \cup C_2$:

$$x_{jk} \geq 0, \quad \text{known for all } k \in U$$

- Each MOS is nonnegative, $x_{jk} \geq 0$, and often highly skewed

Common problem in NASS surveys

- Frame = list of farms in US state: $U = \{1, 2, \ldots, k, \ldots, N\}$
  - assume complete coverage for purposes of this talk
- Characteristics of interest: $C = \{\text{crop}_1, \text{crop}_2, \ldots, \text{crop}_J\}$:

$$
\begin{aligned}
y_{jk} & \quad \text{harvested acres of crop } j \text{ on farm } k \\
T_{yj} &= \sum_{k \in U} y_{jk} = \text{total harvested acres of crop } j
\end{aligned}
$$

- Characteristics have different constraints: $C = C_0 \cup C_1 \cup C_2$
  - C_0: no specified constraints for sunflowers, ...
  - C_1: specified precision targets for corn, soybeans, ..., oats
  - C_2: specified other constraints for potatoes, sugar beets
- Frame measures of size (MOS) for $j \in C_1 \cup C_2$:

$$x_{jk} \geq 0, \quad \text{historic acres of crop } j \text{ on farm } k$$

- Each MOS is nonnegative, $x_{jk} \geq 0$, and often highly skewed

Frame imperfections

- Populations are dynamic and frames are imperfect
- Farms often have multiple crops $y_{jk} > 0$, which may not align with frame acres $x_{jk} > 0$:

	Study variable, y_{jk}	
Frame variable, x_{jk}	$y_{jk} = 0$	$y_{jk} > 0$
$x_{jk} = 0$	true zero	**false zero**
$x_{jk} > 0$	**false positive**	true positive

- Perfect frame would have only true zeros and true positives

Sampling design problem

- Draw a probability sample of farms, $s \subset U$, using $\{\pi_k\}_{k \in U}$
- Estimate the population characteristics
  - Horvitz-Thompson estimators, $\widehat{T}_{yj} = \sum_{k \in s} \pi_k^{-1} y_{jk}$
  - Calibrated estimators, $\widetilde{T}_{yj} = \sum_{k \in s} \omega_k y_{jk}$, using frame totals T_{0j} as controls
- Determine first-order inclusion probabilities $\{\pi_k\}_{k \in U}$ with:
  - bounds on inclusion probabilities: $0 < \delta \leq \pi_k \leq 1$
  - budgetary constraints: $\sum_{k \in U} \pi_k$ not too big
  - no constraints for crops $\in C_0$
  - precision constraints on crops $\in C_1$
  - additional constraints (but not precision) for crops $\in C_2$

Single-MOS model-based optimal design, I

- Suppose that heteroskedastic regression through the origin is a reasonable **superpopulation model** for characteristic y_{jk} with measure of size (MOS) x_{jk}:

$$y_{jk} = \beta_j x_{jk} + \sigma_j x_{jk}^{\gamma_j} \varepsilon_{jk}, \quad \{\varepsilon_{jk}\} \text{ uncorrelated}(0, 1)$$

- Further suppose that we will draw a probability sample with inclusion probabilities $\{\pi_{jk}\}$ and use a **generalized regression estimator (GREG)** to calibrate the sample to the frame control, so that

$$\widetilde{T}_{xj} = \sum_{k \in s} \omega_k x_{jk} = \sum_{k \in U} x_{jk} = T_{0j}$$

Single-MOS model-based optimal design, II

- Under the superpopulation model, **anticipated variance** (unconditional, with respect to design and model) is

$$\mathrm{AV}_j = \mathrm{E}\left[\mathrm{E}\left[\left(\widetilde{T}_{yj} - T_{yj}\right)^2 \,\middle|\, s\right]\right] \simeq \sum_{k \in U}\left(\frac{1}{\pi_{jk}} - 1\right)\sigma_j^2 x_{jk}^{2\gamma_j}$$

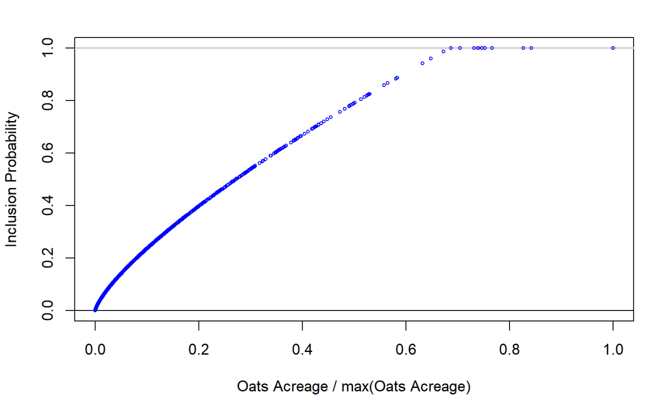
  - Cassel et al. (1976), Brewer (1979), Isaki and Fuller (1982)
- Anticipated variance is minimized by any fixed-size design with **probability proportional to size (PPS)**,

$$\pi_{jk} = \frac{n_j \sigma_j x_{jk}^{\gamma_j}}{\sum_{k \in U} \sigma_j x_{jk}^{\gamma_j}} = \frac{n_j x_{jk}^{\gamma_j}}{\sum_{k \in U} x_{jk}^{\gamma_j}}$$

  - optimal if all $\pi_{jk} \leq 1$
  - standard modification if $\pi_{jk} > 1$ is to set $\pi_{jk} = 1$, exclude unit k from frame, and recalculate with $(n_j - 1)$

Single-MOS model-based optimal design, III

- We are minimizing . . .
  - an approximation to the anticipated variance of the GREG
  - under an assumed mean model reflected by the GREG
  - under an assumed heteroskedasticity structure
- Optimal probabilities do not uniquely determine design



Single-MOS sample size determination

- Plug the optimal $\{\pi_{jk}\}$ into AV_j and divide by the squared model expectation of T_{yj} to obtain **(anticipated coefficient of variation)**^{2}:

$$\begin{aligned}\text{CV}_j^2 &= \frac{\sigma_j^2}{\beta_j^2 \left(\sum_{k \in U} x_{jk}\right)^2} \left\{ \frac{1}{n_j} \left(\sum_{k \in U} x_{jk}^{\gamma_j} \right)^2 - \sum_{k \in U} x_{jk}^{2\gamma_j} \right\} \\ &= \frac{\sigma_j^2}{\beta_j^2 T_{0j}^2} \left\{ \frac{1}{n_j} T_{\gamma j}^2 - T_{2\gamma j} \right\}\end{aligned}$$

- Plug in the target CV and solve for n_j ($j \in C_1$):

$$n_j \geq \frac{T_{\gamma j}^2}{\text{CV}_j^2 \frac{\beta_j^2 T_{0j}^2}{\sigma_j^2} + T_{2\gamma j}}$$

- Now obtain n_j using estimates of $\beta_j, \sigma_j, \gamma_j$ from past surveys

What to do with multiple measures of size?

- We have $J_1 = |C_1|$ precision targets $\{\mathrm{CV}_j\}_{j \in C_1}$, plus additional constraints from C_2
- Single-MOS approach leads to J_1 sample sizes $\{n_j\}_{j \in C_1}$ and J_1 sets of optimal inclusion probabilities:

$$\pi_{jk} = \frac{n_j x_{jk}^{\gamma_j}}{\sum_{k \in U} x_{jk}^{\gamma_j}},$$

(as usual, requires modification if $\pi_{jk} > 1$)

- But we need a single set of inclusion probabilities, not dependent on j

Options with multiple measures of size: Univariate

- Convert the multiple MOS problem to a single MOS problem and use univariate methods
- **Option 1: Give up!** Choose a single "important" MOS
- **Option 2: Compromise.** Compute a linear combination of the size measures
  - Hagood and Bernert (1945) propose first principal component
- Univariate methods are clearly suboptimal: not considered further

Options with multiple measures of size: Stratification

- Multivariate stratification has a long history and is closely related
  - can approximate PPS problem as piecewise constant within strata, or otherwise adapt stratification methods
- **Option 3: Deep stratification.** Sort $\{x_{jk}\}_{k \in U}$ for each j, divide into bins, cross all bins to form multi-way strata
  - Tepping, Hurwitz, Deming (1943), Kish and Anderson (1978)
- **Option 4: Clustering.** Form homogeneous clusters using $\boldsymbol{x}_k$
  - Skinner, Holmes and Holt (1994) reference several papers
- Stratification leads to **multivariate allocation problem**
  - Friedrich, Münnich, and Rupp (2018) is an excellent review with extensions

Options with multiple measures of size: Multiple frame

- Consider J_1 frames, $U_j = \{k \in U : x_{jk} > 0\}$
- **Option 5: Multiple frame sampling.** Skinner, Holmes and Holt (1994) draw independent stratified samples and combine via multiple frame methods
  - In our setting, draw independent PPS samples with each set of $\{\pi_{jk}\}_{k \in U}$, then combine via multiple frame methods:

$$T_z^* = \sum_{j \in C_1} \sum_{k \in s_j} \frac{z_k}{\sum_{j \in C_1} \pi_{jk}} = \sum_{k \in U} z_k \frac{\sum_{j \in C_1} I_{jk}}{\sum_{j \in C_1} \pi_{jk}}$$

    is unbiased for T_z
  - Not identical to Horvitz-Thompson estimator (which requires deduplication across samples)
  - Weights $1/(\sum_{j \in C_1} \pi_{jk})$ may be less than one

Options with multiple MOS: Combining one-at-a-time

- One-at-a-time optimal probabilities for each MOS: $\{\pi_{jk}\}_{k \in U}$
  - rely on parameters of one-at-a-time models: $\beta_j, \sigma_j, \gamma_j$
- Combine in some way to address the multiple MOS problem
- **Option 6: Average optimal PPS.** Bee, Benedetti, Espa, and Piersimoni (2010) find reasonable performance with

$$\pi_{AVE,k} = \sum_{j \in C_1} \left(\frac{1}{J_1} \right) \pi_{jk}$$

Options with multiple MOS: Combining one-at-a-time

- One-at-a-time optimal probabilities for each MOS: $\{\pi_{jk}\}_{k \in U}$
- Combine in some way to address the multiple MOS problem
- **Option 7: Optimal linear combination.** Benedetti, Andreano, and Piersimoni (2019) use a custom algorithm to find $0 \leq \psi_j \leq 1$ so that

$$\pi_{BAP,k} = \sum_{j \in C_1} \psi_j \pi_{jk}$$

minimize the maximum one-at-a-time sample size n_j needed to attain precision targets

Options with multiple MOS: Combining one-at-a-time

- One-at-a-time optimal probabilities for each MOS: $\{\pi_{jk}\}_{k \in U}$
- Combine in some way to address the multiple MOS problem
- **Option 8: MPPS.** Multivariate Probability Proportional to Size sampling.

$$\pi_{MPPS,k} = \max_{j \in C_1} \pi_{jk}$$

  - Standard method for NASS surveys: Amrhein, Hicks and Kott (1996); Kott and Bailey (2000)
  - Typically, heteroskedasticity parameter is taken to be $\gamma_j \equiv 0.75$ (following a suggestion by Ken Brewer)

MPPS at NASS

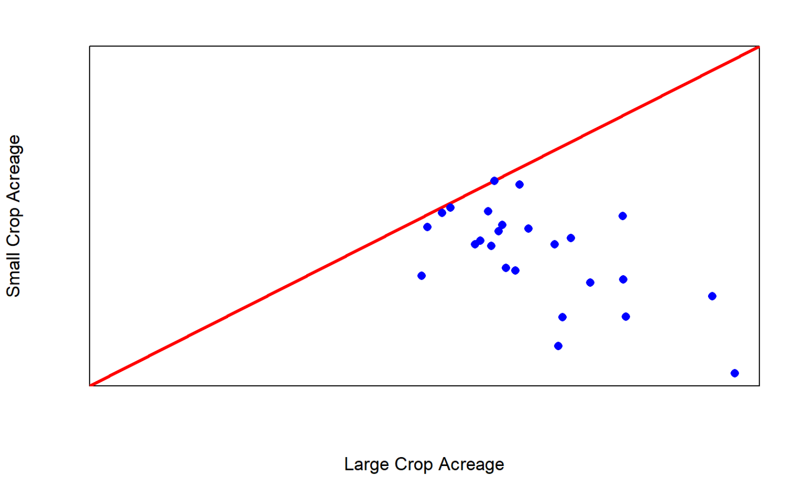
- Common in NASS multipurpose surveys, like Crops APS (Acreage, Production, and Stocks)
- Simple and fairly effective approach
- Overshoots target sample sizes for all crops:

$$
\begin{aligned}
j\text{th expected sample count} &= \sum_{k \in U} \mathbf{1}(x_{jk} > 0)\pi_{MPPS,k} \\
&= \sum_{k \in U} \mathbf{1}(x_{jk} > 0) \max_{i \in C_1} \pi_{ik} \\
&\geq \sum_{k \in U} \mathbf{1}(x_{jk} > 0)\pi_{jk} \\
&= n_j
\end{aligned}
$$

- Can break the (higher MOS, higher probability) link for smaller crops since the larger crops will dominate

Possible broken relationship for small crops

- Highest probabilities (due to large crop acreage) for lowest level of small crop acreage



Issues with control of the sampling design

- MPPS cannot address side conditions, C_2, except by adjusting sample sizes
- Further complication is that NASS uses MPPS probabilities in Poisson sampling
- Controlling design therefore requires
  - adjusting preliminary expected sample sizes, n_j
  - (but sample sizes are random due to Poisson sampling and targets are overshot by MPPS)
  - or setting aside C_2 cases for special consideration
- Result is lack of control of design, necessitating iteration in design and selection

Two paths to improving control of the sampling design

Improve the probabilities

- Revisit models underlying the methods, updating if necessary
- Enumerate all $C_1 \cup C_2$ sample constraints and build them into the probabilities, if possible
- Is it possible to find the **Optimal Probabilities**, which minimize the expected sample size for the given constraints?

Improve the sample selection

- Applies to either MPPS or Optimal
- Poisson sampling is “least controlled” selection strategy for a given set of probabilities
- Is it feasible to use **Balanced Sampling** as alternative for selection?
  - “Most controlled” selection strategy for a given set of probabilities

Two paths to improving control, continued

- Either path can improve sampling team's control of the design, and **neither path requires the other**
- Optimal probabilities could be used for sample selection...
  - in current Poisson sampling designs
  - or in new Balanced sampling designs
- Balanced sampling could use as its inclusion probabilities...
  - existing MPPS probabilities
  - new Optimal probabilities
- The two research questions can be pursued in parallel, with any improvements implemented either alone or together

Is it possible to find the optimal probabilities?

- **Our approach: skip the intermediate steps** of determining one-at-a-time $\{\pi_{jk}\}$
- Return to anticipated CV^2 constraint, without "optimal" π_{jk}:

$$\frac{\sigma_j^2}{\beta_j^2 T_{0j}^2}\left(\sum_{k\in U}\frac{x_{jk}^{2\gamma_j}}{\pi_k} - \sum_{k\in U} x_{jk}^{2\gamma_j}\right) \leq CV_j^2,$$

which implies

$$\sum_{k\in U}\frac{x_{jk}^{2\gamma_j}}{\pi_k} \leq \frac{\beta_j^2 T_{0j}^2}{\sigma_j^2} CV_j^2 + T_{2\gamma j}, \quad j \in C_1$$

- Minimize expected sample size, $\sum_{k\in U}\pi_k$, subject to CV constraints and

$$0 < \delta \leq \pi_k \leq 1$$

Convex optimization with CV constraints

- Solve this problem via **convex optimization**:

$$\begin{aligned} \text{minimize} \quad & \sum_{k \in U} \pi_k \\ \text{subject to} \quad & 0 < \delta \leq \pi_k \leq 1 \\ & \sum_{k \in U} \frac{x_{jk}^{2\gamma_j}}{\pi_k} \leq \frac{\beta_j^2 T_{0j}^2}{\sigma_j^2} \mathrm{CV}_j^2 + T_{2\gamma j}, \quad j \in C_1 \end{aligned}$$

- Can we solve this (large) problem directly, without custom software or special computing resources?
- We use the `R` package `CVXR` (Fu, Narasimhan, and Boyd 2020)

Convex optimization with CV constraints: CVXR

- Solve this problem via **convex optimization** using the R package CVXR (Fu, Narasimhan, and Boyd 2020)

unknowns $\{\pi_k\}_{k \in U}$	`pik <- Variable(N)`
minimize $\sum_{k \in U} \pi_k$	`objective <- Minimize(sum(pik))`
subject to	`constraints <- list(`
$\pi_k \geq \delta > 0$	`pik >= delta,`
$\pi_k \leq 1$	`pik <= 1,`
$\sum_{k \in U} x_{jk}^{2\gamma_j} / \pi_k \leq B_j$	`sum(x[, j]^(2 * gamma[j])`
	`* inv_pos(pik)) <= B[j])`

- Here, the known bounds are

$$B_j = \frac{\beta_j^2 T_{0j}^2}{\sigma_j^2} \mathrm{CV}_j^2 + T_{2\gamma j}, \quad j \in C_1$$

Notes on computation

- Our (limited) experience with problem size:
  - no problems with $N = O(10^3), J = O(10)$
  - memory troubles with $N = O(10^4)$
- Break up the problem into G feasible subproblems:

$$\begin{aligned} \text{minimize} \quad & \sum_{g=1}^{G} \sum_{k \in U_g} \pi_k \\ \text{subject to} \quad & 0 < \delta \leq \pi_k \leq 1 \end{aligned}$$

$$\sum_{g=1}^{G} \left(\sum_{k \in U_g} \frac{x_{jk}^{2\gamma_j}}{\pi_k} \right) \leq \sum_{g=1}^{G} \left(\frac{\beta_j^2 T_{0j}^2}{\sigma_j^2 T_{2\gamma j}} \mathrm{CV}_j^2 + 1 \right) T_{2\gamma j,g},$$

where

$$T_{2\gamma j} = \sum_{g=1}^{G} \left(\sum_{k \in U_g} x_{jk}^{2\gamma_j} \right) = \sum_{g=1}^{G} T_{2\gamma j,g}$$

Notes on computation, continued

- Now solve each of the G feasible subproblems:

$$\begin{aligned} \text{minimize} \quad & \sum_{k \in U_g} \pi_k \\ \text{subject to} \quad & 0 < \delta \leq \pi_k \leq 1 \end{aligned}$$

$$\left(\sum_{k \in U_g} \frac{x_{jk}^{2\gamma_j}}{\pi_k} \right) \leq \left(\frac{\beta_j^2 T_{0j}^2}{\sigma_j^2 T_{2\gamma j}} \mathrm{CV}_j^2 + 1 \right) \left(\sum_{k \in U_g} x_{jk}^{2\gamma_j} \right)$$

- Partition potentially constrains the solution space
  - but does not impose any additional constraint if $T_{2\gamma j,g} \neq 0$ for exactly one g
  - constraints are minimal for a random partition with G small, hence we get a good approximate solution
  - (change G or rerandomize and get a very similar solution)

Additional constraints: Domain sample size targets

- Domain sample size targets based on **observed $\boldsymbol{x_{jk}}$** for $j \in C_2$:

$$\text{expected sample count} = \sum_{k \in U} \mathbf{1}(x_{jk} > 0)\pi_k \geq m_j$$

  - see Falorisi and Righi (2015) for multi-domain problem with known domain indicators
- Domain sample size targets based on **predicted $\boldsymbol{y_{jk}}$** for $j \in C_2$:

$$\sum_{k \in U} \mathrm{E}\left[\mathbf{1}(y_{jk} > 0) \mid \boldsymbol{x}_k\right] \pi_k = \sum_{k \in U} \rho_j(\boldsymbol{x}_k)\pi_k \geq m_j$$

  - requires new propensity models $\rho_j(\boldsymbol{x}_k)$ for domain membership
- Either type of constraint is convex in $\{\pi_k\}_{k \in U}$

Optimization with additional constraints

- Solve this problem via **convex optimization** using the R package CVXR (Fu, Narasimhan, and Boyd 2020)

unknowns $\{\pi_k\}_{k \in U}$	`pik <- Variable(N)`
minimize $\sum_{k \in U} \pi_k$	`objective <- Minimize(sum(pik))`

subject to
$\pi_k \geq \delta > 0$
$\pi_k \leq 1$
$\sum_{k \in U} x_{jk}^{2\gamma_j} / \pi_k \leq B_j$
 	
$\sum_{k \in U} \mathbf{1}(x_{jk} > 0) \pi_k \geq m_j$
	`constraints <- list(`
	`pik >= delta,`
	`pik <= 1,`
	`sum(x[, j]^(2 * gamma[j])`
	`* inv_pos(pik)) <= B[j],`
	`sum((x[, j] > 0) * pik)`
	`>= m[j])`

Additional constraints: Domain area targets

- Want the sample to capture a specified proportion of a domain's total area
- Domain area targets based on **observed** $\boldsymbol{x_{jk}}$ for $j \in C_2$:

$$\frac{\text{expected sample area}}{\text{total area}} = \frac{\sum_{k \in U} x_{jk} \pi_k}{\sum_{k \in U} x_{jk}} \geq p_j$$

- Domain area targets based on **predicted** $\boldsymbol{y_{jk}}$ for $j \in C_2$:

$$\frac{\sum_{k \in U} \mathrm{E}\left[y_{jk} \mid \boldsymbol{x}_k\right] \pi_k}{\sum_{k \in U} \mathrm{E}\left[y_{jk} \mid \boldsymbol{x}_k\right]} = \frac{\sum_{k \in U} \alpha_j(\boldsymbol{x}_k) \pi_k}{\sum_{k \in U} \alpha_j(\boldsymbol{x}_k)} \geq p_j$$

  - requires new models $\alpha_j(\boldsymbol{x}_k)$ for response acreage
- Either type of constraint is convex in $\{\pi_k\}_{k \in U}$

Example of computing optimal probabilities

- Frame acres for $N = 23,528$ farms in one US state (2017 Census of Agriculture):

$$\boldsymbol{x}_k = \left[(x_{jk})_{j \in C_1}, (x_{jk})_{j \in C_2} \right]^\top$$

- Specific precision targets for $J_1 = 6$ crops:

$$C_1 = \{\text{barley, corn, dry beans, oats, soybeans, spring wheat}\}$$

- Sample size and acreage coverage targets:

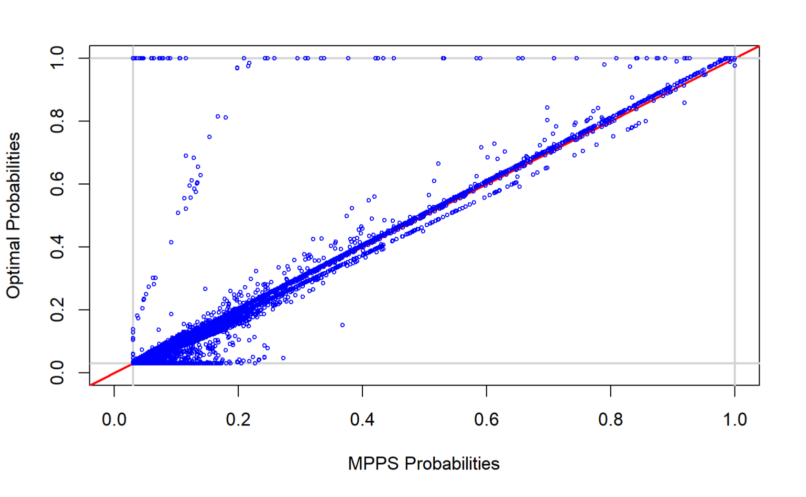
$$C_2 = \{\text{potatoes, sugar beets}\}$$

- Partition into subproblems for optimization:

$$U = \{\text{any small crop}\} \cup (\cup_g \{\text{only corn or soybeans}\})$$

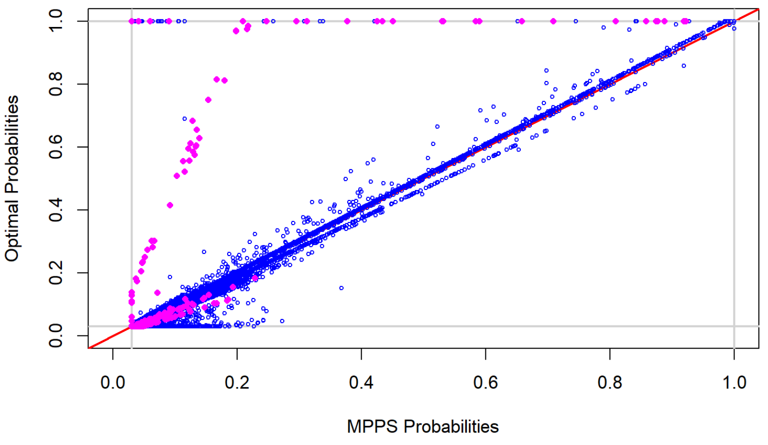
Optimal probabilities versus MPPS probabilities

- Probabilities are highly correlated but far from identical



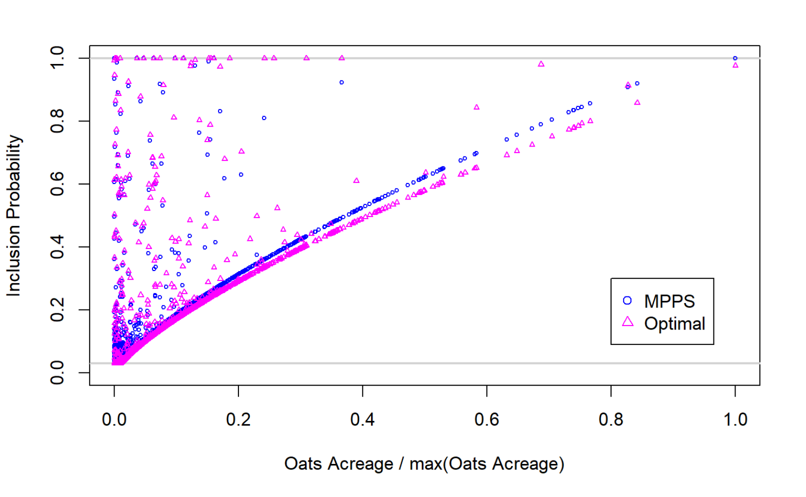
Optimal probabilities versus MPPS probabilities

- Satisfying C_2 potato sample size constraint



Optimal probabilities versus MPPS probabilities

- Large farms are less dominant in Optimal than MPPS



Simulation of a farm population

- Simulate a population starting with frame acres, $\{\boldsymbol{x}_k\}_{k \in U}$ for $N = 23,528$ farms from 2017 Census of Agriculture
- Simulation steps:
  - given number of frame crops, simulate number of crops
  - given number of crops, simulate crop types
  - given crop types, simulate crop acreages
- Iterate over time to simulate population dynamics

Given frame number of crops, simulate number of crops

- For farm k, use its number of frame crops f_k (with nonzero frame acres) to simulate its number of actual crops, c_k
- Use conditional probability distributions, $\mathrm{P}[c_k = j \mid f_k = i]$, estimated from 2019 Crops APS (Acreage, Production, and Stocks) survey data:

Number of frame crops, f_k	Number of crops, c_k: 0	1	2	3	4
0	0.823	0.150	0.025	0.001	0.001
1	0.221	0.691	0.081	0.006	0.001
⋮	⋮	⋮	⋮	⋮	⋮
5+	0.046	0.000	0.318	0.590	0.046

Given the number of crops, simulate crop types

If c_k satisfies...	then ...
$c_k = 0$	no crop types to simulate
$c_k > f_k = 0$	draw from population distribution of crop types
$f_k \geq c_k > 0$	draw from frame crop types for farm k
$c_k > f_k > 0$	any crop type is possible, frame crops more likely

- True zeros, false zeros, true positives, and false positives are all possible
- Simulation parameters are tuned to match the rates seen in real data

Given the crop types, simulate survey acres for each crop

- On the frame, we have total cropland acres A_k for farm k
- We have now selected c_k random crops, where crop selection probabilities are
  - linearly related to crop-specific frame acres (if non-zero)
  - or linearly related to total frame acres (if crop-specific frame acres are zero)
- **Idea:** Break A_k at random into $c_k + 1$ pieces
  - if $c_k = 0$, then all cropland acres are assigned to "remainder" = uninteresting non-crop uses
  - if $c_k > 0$, assign a small, random fraction of A_k to remainder and distribute the rest in proportion to crop selection probabilities

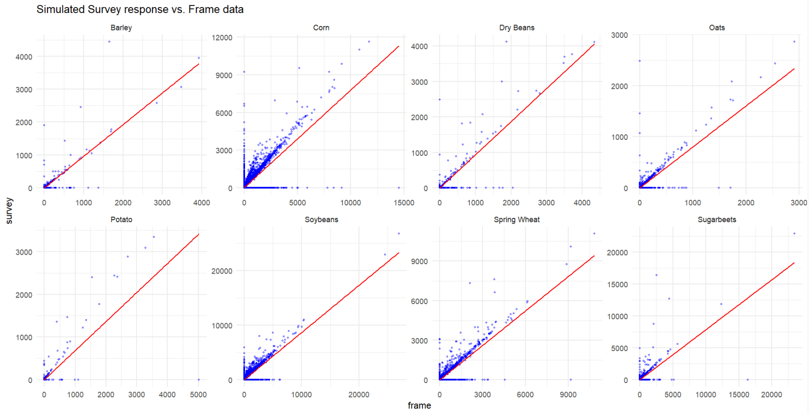
Features of the simulated population

- Works for any number/types of crops: not specific to the crops in the selected state
- Realistic variation in crop numbers and crop types
- Realistic rates of true zeros, false zeros, true positives, and false positives
- Steps can be iterated to simulate population dynamics:

$$\begin{array}{rcl}
\textbf{Frame Variables} & & \textbf{Study Variables} \\
\boldsymbol{x}_k^{(0)} & \longrightarrow & \boldsymbol{y}_k^{(0)} \\
\text{set } \boldsymbol{y}_k^{(0)} = \boldsymbol{x}_k^{(1)} & \longrightarrow & \boldsymbol{y}_k^{(1)} \\
\text{set } \boldsymbol{y}_k^{(1)} = \boldsymbol{x}_k^{(2)} & \longrightarrow & \boldsymbol{y}_k^{(2)} \\
\vdots & & \vdots
\end{array}$$

Simulated population for census +2 years

- Realistic heteroskedastic linear relationships with frame acres, *without ever introducing heteroskedastic linear models*



Monte Carlo experiment

- Simulate three years of (frame, population) data
  - fit models to year 0 "census" data
  - draw repeated samples from (fixed) year 2 population
- **Inclusion probabilities:** $\{\pi_{MPPS,k}\}_{k \in U}$ or $\{\pi_{OPT,k}\}_{k \in U}$
  - real C_1 precision constraints
  - realistic C_2 additional constraints
- **Sample selection:** Poisson sampling or Balanced sampling
- **Estimation method:** Uncalibrated or Calibrated
  - raking via `calibrate` function from `R survey` (Lumley 2004)
- For each combination of experimental factors, draw 1,000 replicate samples from fixed population
  - compute estimates for each frame crop and each survey crop
  - evaluate bias and variance of each strategy

Monte Carlo experiment: Balancing details

- Balanced sampling via cube algorithm
  - (Deville and Tillé, 2004)
  - using `samplecube` method from `sampling` package (Tillé and Matei, 2021)
- Both MPPS and Optimal are balanced on C_1 conditions:

$$\sum_{k \in s} \frac{1}{\pi_k} x_{jk} \simeq \sum_{k \in U} x_{jk}$$

- Optimal could be (but isn't) balanced on C_2 conditions:

$$\begin{aligned} \sum_{k \in s} \frac{1}{\pi_{OPT,k}} \{\mathbf{1}(x_{jk} > 0)\pi_{OPT,k}\} &\simeq \sum_{k \in U} \{\mathbf{1}(x_{jk} > 0)\pi_{OPT,k}\} = m_j \\ \sum_{k \in s} \frac{1}{\pi_{OPT,k}} (x_{jk}\pi_{OPT,k}) &\simeq \sum_{k \in U} (x_{jk}\pi_{OPT,k}) = p_j \sum_{k \in U} x_{jk} \end{aligned}$$

Monte Carlo experiment: Sample size details

- We used the following C_2 conditions:

$$\text{potatoes:} \quad \sum_{k \in U} \mathbf{1}(x_{jk} > 0)\pi_k \geq 80$$

$$\text{sugar beets:} \quad \sum_{k \in U} x_{jk}\pi_k \geq 0.5 \sum_{k \in U} x_{jk}$$

- These lead to higher expected sample sizes for Optimal than MPPS (which cannot incorporate C_2)
- To make comparisons easier, we increased MPPS expected sample size to more closely match Optimal sample size:

$$\sum_{k \in U} \pi_{MPPS,k} = 2345 \quad > \quad \sum_{k \in U} \pi_{OPT,k} = 2333$$

Monte Carlo results

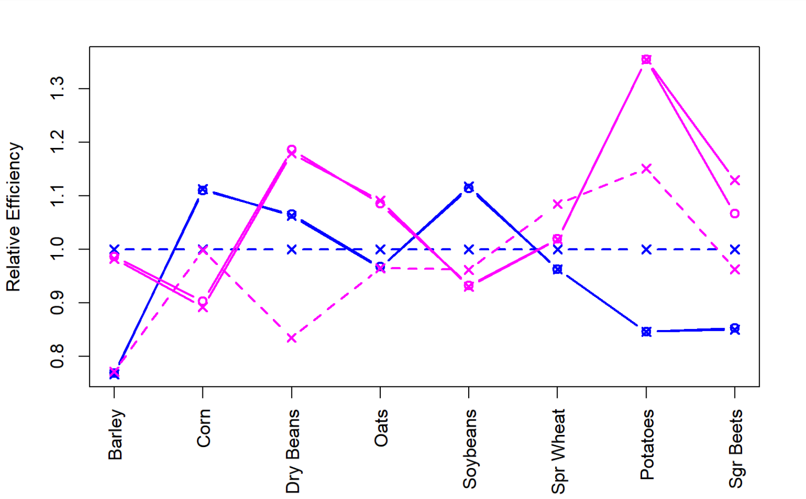
- Estimators are unbiased for population targets
- Balancing works to greatly reduce variation of sample size
- Balancing and/or calibrating works as expected for frame variables $\boldsymbol{x}$
- Results vary for study variables, depending on quality of model relating $\boldsymbol{y}$ to $\boldsymbol{x}$
- Evaluate via Monte Carlo **relative efficiency**

$$\text{(relative efficiency)} = \frac{\text{Var(MPPS Poisson Raked)}}{\text{Var(Competitor)}},$$

with values greater than one favoring the competitor

Var(MPPS Poisson Raked) / Var(Competitor)

- MPPS (blue) and Optimal (pink), Poisson (dashed line) or Balanced (solid line), Unraked (○) or Raked (×)



Discussion

- Feasible to solve for optimal probabilities
  - in a problem with realistic size and constraints
  - without custom software
  - without specialized computing resources
- Optimal design with balance dominates existing NASS methodology in limited simulation experiments
  - fair comparison is tricky: without C_2 conditions, Optimal has lower expected sample size than MPPS
- Other models can be considered
- Other features (costs, response propensities, etc.) can be incorporated in constraints
- **Thank you for your attention!**

Selected references, I

- Amrhein, J., Hicks, S., & Kott, P. (1996). Methods to control selection when sampling from multiple list frames. In ASA Proceedings of the Section on Survey Research Methods.
- Bee, M., Benedetti, R., Espa, G., & Piersimoni, F. (2010). On the use of auxiliary variables in agricultural survey design. Agricultural Survey Methods, 107–132.
- Benedetti, R., Andreano, M. S., & Piersimoni, F. (2019). Sample selection when a multivariate set of size measures is available. Statistical Methods & Applications, 28(1), 1–25.
- Brewer, K. (1979). A class of robust sampling designs for large-scale surveys, Journal of the American Statistical Association, 74, 911–915.
- Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 63(3), 615–620.
- Deville, J.C. & Tillé, Y. (2004). Efficient balanced sampling: the cube method. Biometrika, 91(4), pp.893–912.
- Falorsi, P. D., & Righi, P. (2015). Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys. Survey Methodology, 41(1), 215–236.

Selected references, II

- Friedrich, U., Münnich, R., & Rupp, M. (2018). Multivariate optimal allocation with box-constraints. Austrian Journal of Statistics, 47(2), 33–52.
- Fu, A., Narasimhan, B., and Boyd, S. (2020). CVXR: An R Package for Disciplined Convex Optimization. Journal of Statistical Software 94 (14): 1–34.
- Hagood, M. J., & Bernert, E. H. (1945). Component indexes as a basis for stratification in sampling. Journal of the American Statistical Association, 40(231), 330–341.
- Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. Journal of the American Statistical Association, 77(377), 89-96.
- Kish, L., & Anderson, D. W. (1978). Multivariate and multipurpose stratification. Journal of the American statistical Association, 73(361), 24–34.
- Kott, P.S. & Bailey J.T. (2000). The theory and practice of maximal Brewer selection with Poisson PRN sampling. In: Proceedings of the Second International Conference on Establishment Surveys, Invited papers, 269–278.
- Lumley, T. (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1–19.

Selected references, III

- Skinner, C. J., Holmes, D. J., & Holt, D. (1994). Multiple frame sampling for multivariate stratification. International Statistical Review/Revue Internationale de Statistique, 333–347.
- Tepping, B. J., Hurwitz, W. N., & Deming, W. E. (1943). On the efficiency of deep stratification in block sampling. Journal of the American Statistical Association, 38(221), 93–100.
- Tillé, Y. & and Matei, A. (2021). `sampling`: Survey Sampling. R package version 2.9. `https://CRAN.R-project.org/package=sampling`