Hitting Calibration Targets + INCA Calibration

Luca Sartore, Kelly Toppin, Andrea Lamas,, Matt Williams,
Cliff Spiegelman, Linda J. Young

National Agricultural Statistics Service

April 30, 2015

"... providing timely, accurate, and useful statistics in service to U.S. agriculture."

Census of Agriculture

- National Agricultural Statistics Service (NASS) conducts a Census of Agriculture every 5 years.
- The Census provides a detailed picture of U.S. farms, ranches and the people who operate them.
- It is the only source of uniform, comprehensive agricultural data for every state and county in the United States.

Census of Agriculture

- NASS also obtains information on most commodities from administrative sources or from NASS surveys of non-farm populations, such as
  - USDA Farm Service Agency program data,
  - Agricultural Marketing Services market orders,
  - livestock slaughter data, and
  - cotton ginning data.

Census Mail List

- Definition of farm: an agricultural operation that produced or would produced and sold agricultural products of at least $1000 during the year of the census
- Every effort is made to make the Census Mail List (CML) as complete as possible, but it does not contain all U.S. farms, resulting in list undercoverage.
- Some farms on the CML do not respond to the census, nonresponse is present.

Dual System Estimation (DSE)

- To adjust for undercoverage, nonresponse and misclassification, NASS uses capture-recapture methodology where two independent surveys are required.
- Calibration is conducted to ensure that the census estimates are consistent with the available information on commodity production.
- This DSE method produces adjusted weights that are used as the starting values for the calibration process.

Calibration

- Forces weighted estimates of calibration variables to match known totals
- Idea was introduced by Lemel and developed by Deville and Särndal.

Calibration

We want $T = Aw$, where

T is vector partitioned into y known and y^* unknown population totals,
A is the matrix of collected data from population, and
w is a vector of p unknown weights.

Find the solution of the linear system $y = A^*w$, where

y is a vector of n known point targets (benchmarks), and
A^* is a $n \times p$ submatrix of collected data.

- Often produces non-integer weights

Improving Calibration Results

- Approaches previously tested
  - Code changes
    - Stepwise variable addition (traditional) vs. all variables in approach (new)
    - New code treats soft targets as soft targets (get targets within the allowable range)
  - Allowing DSE weight input to calibration with relaxed truncation (0.9-6)
    - Traditional code truncates the DSE weights input to calibration to between 1 and 3
  - Allow R and X case records to be handled similarly to regular records
  - Allowing calibration output weights in the range of .9 to 6
    - Traditional code outputs weights in range of 1 to 6
  - Allowing use of submitted, unedited data

Integerization

- Current integerization methodology uses “linked” integerization
  - DSE weight decimal and calibration weight decimal are used to determine how calibration weight will be rounded
- Current integerization methodology cannot handle calibration weights less than 1 (some will be rounded to 0)
- Therefore, calibration research will focus on weights between 1 and 6

Michigan

Restriction	DSE Input Weights to Calibration	Limited Data Changes	Output Weights of Calibration	Targets Missed out of 175		
				After Calibration		After Integerization Avg (Min,Max)
				Total	Total Possible	
R & X-Case	Partially Adjusted (1-3)	No	(1-6)	8	2	12.6 (11,16)
R & X-Case	Fully Adjusted (1-3)	No	(1-6)	9	3	11.1 (9,13)
None	(1-6)	No	(1-6)	6	0	9.6 (7,12)
R & X-Case	(1-6)	No	(1-6)	6	0	7.5 (6,11)
R-Case & EO	(1-6)	No	(1-6)	6	0	9.6 (7,13)
None	(1-6)	Yes	(1-6)	4	1	9.8 (8,14)
R & X-Case	(1-6)	Yes	(1-6)	4	1	7.5 (6,10)
R-Case & EO	(1-6)	Yes	(1-6)	4	1	9.4 (7,12)

Note: Highlighted rows use old code (integerization conducted 10 times)
Other rows use new code (integerization conducted 100 times)

North Carolina

Restriction	DSE Input Weights to Calibration	Limited Data Changes	Output Weights of Calibration	Targets Missed out of 184	
				After Calibration	After Integerization Avg (Min,Max)
R & X-Case	Partially Adjusted (1-3)	No	(1-6)	4	6.1 (3,8)
R & X-Case	Fully Adjusted (1-3)	No	(1-6)	4	5.2 (4,6)
None	(1-6)	No	(1-6)	0	3.0 (1,5)
R & X-Case	(1-6)	No	(1-6)	3	3.9 (3,6)
R-Case & EO	(1-6)	No	(1-6)	1	2.5 (1,5)
	NO EDITS NEEDED				

Note: Highlighted rows use old code (integerization conducted 10 times)
Other rows use new code (integerization conducted 100 times)

Texas

Restriction	DSE Input Weights to Calibration	Limited Data Changes	Output Weights of Calibration	Targets Missed out of 346		
				After Calibration		After Integerization Avg (Min,Max)
				Total	Total Possible	
R & X-Case	Partially Adjusted (1-3)	No	(1-6)	9	0	24.7 (21,32)
R & X-Case	Fully Adjusted (1-3)	No	(1-6)	14	5	19.1 (15, 26)
None	(1-6)	No	(1-6)	Error		
R & X-Case	(1-6)	No	(1-6)	12	3	25.4 (16,32)
R-Case & EO	(1-6)	No	(1-6)	5	1	22.9 (16,35)
None	(1-6)	Yes	(1-6)	Error		
R & X-Case	(1-6)	Yes	(1-6)	11	7	25.5 (20,32)
R-Case & EO	(1-6)	Yes	(1-6)	4	3	22.0 (13,34)

Note: Highlighted rows use old code (integerization conducted 10 times)
Other rows use new code (integerization conducted 100 times)

Findings

- Most targets that cannot be hit, are unable to be hit because the data do not support the targets

Recommendations

- Targets need to be evaluated
- Integerization process needs more research
  - Do other integerization methods allow for more targets to be hit?
  - Other integerization methods allow calibration weights to be less than 1.

Outline

- Calibration
- Rounding
- Integer calibration
- Results
- Conclusion

NASS Census 2012 Calibration

- The targets used in calibration are the commodity products (commodity targets), and the 65 farm targets.
- Each target is calibrated within a pre-specified tolerance range, which is generally less than 2% of the target.

NASS Census 2012 Calibration

- NASS has a need for integer weights for its final totals in the census publication. It uses a two part process.

1. Linear truncated calibration to produce non-integer weights.

2. Rounding the weights from step 1.

Integerization

- Current integerization (KR) methodology uses "linked" integerization
  - DSE weight decimal and calibration weight decimal are used to determine how calibration weight will be rounded
- Current integerization methodology cannot handle calibration weights less than 1 (some will be rounded to 0)

Problems with old approach

- Too many missed targets
- Final weights are very different than initial (DSE) weights
- Computationally intensive and time consuming

SimCa code (first attempt)

- Get target within its interval. The old method tried to hit each target's point value instead of target's interval.
- The second feature was that targets are calibrated simultaneously instead of the sequential approach present in the old code.

Preliminary results

State	Method	Missed	After old rounding
MI	Old	9	11.1 (9,13)
	SimCa	6	7.5 (6,11)
NC	old	4	5.2 (4,6)
	SimCa	3	3.9 (3,6)
TX	old	14	19.1 (15, 26)
	SimCa	12	25.4 (16,32)

New rounding Method

- INCA (rounded)
  - Explicit gradient
  - Starts with real calibrated weights

Preliminary Results with new rounding

State	Rounding	Missed
MI	Current Rounding	7.5 (6-11)
	INCA rounded	6
NC	Current rounding	3.9 (3-6)
	INCA rounded	3
TX	Current rounding	25.4 (16-32)
	INCA rounded	9

Alternative proposal

- Old approach

- New approach

Hitting Calibration Targets
April 30, 2015

Description of the problem

- The following objective function is minimized:

$$\min_{w \in \mathcal{W} \subseteq \mathbb{N}^p} \sum_{i=1}^{n} \rho_{\ell_i, u_i}(y_i - a_i^\top w) + \lambda P(w)$$

ℓ_i is the lower bound for $a_i^\top w$,
u_i is the upper bound for $a_i^\top w$,
$\rho(\cdot)$ is a generic loss function,
λ is a non negative scalar,
$P(\cdot)$ is a distance from the original weights

Description of the algorithm

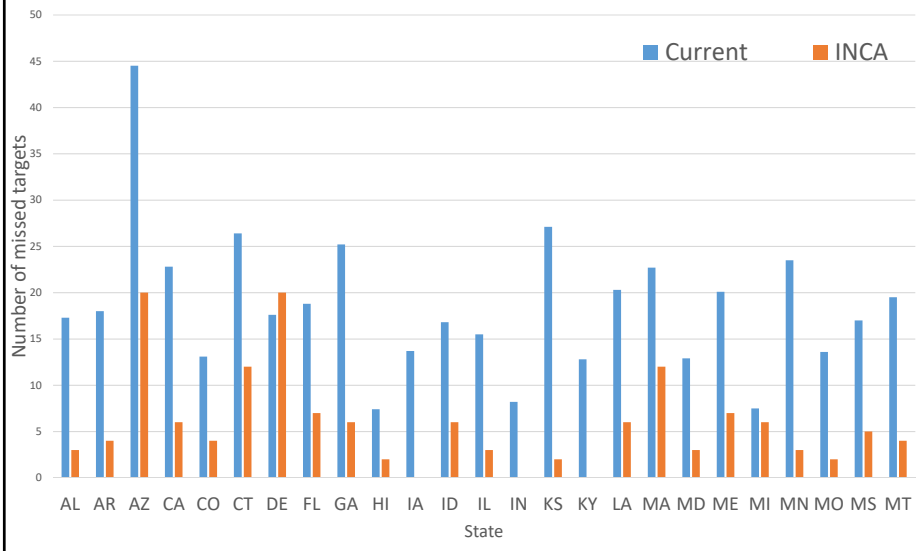
1. All unfeasible weights are truncated to their closest boundary, and in order to minimize the objective function, non-integer weights are then rounded sequentially according to an importance index based on the gradient.

2. Each weight, according to the magnitude of the gradient, is allowed to move unit-shifts which decreases the objective function.

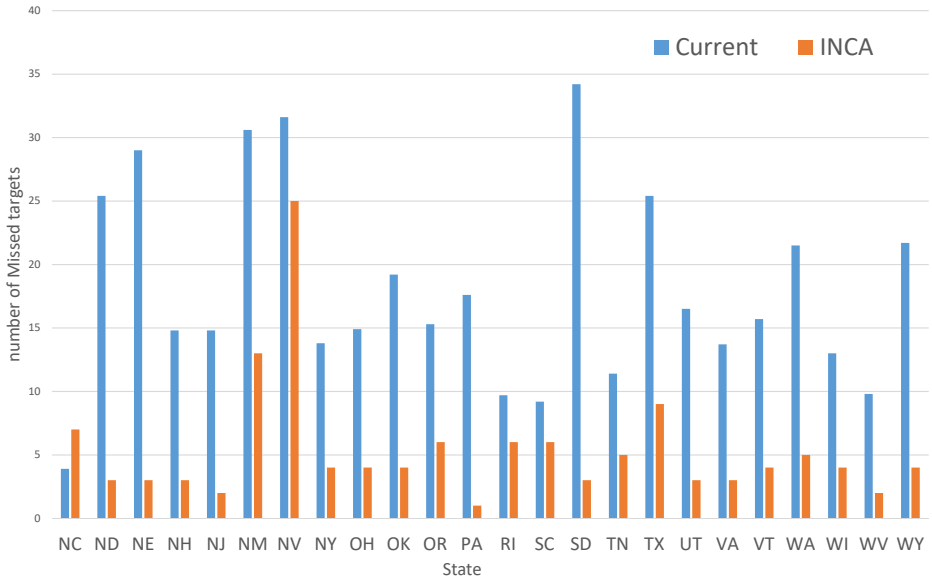
Integer Calibration (INCA)

- Based on gradient
- Using C programming languages with SAS wrapper
- Output weights are in the set {1, 2, 3, 4, 5, 6}
- Output weights are close to the input weights

Current VS INCA



Current VS INCA

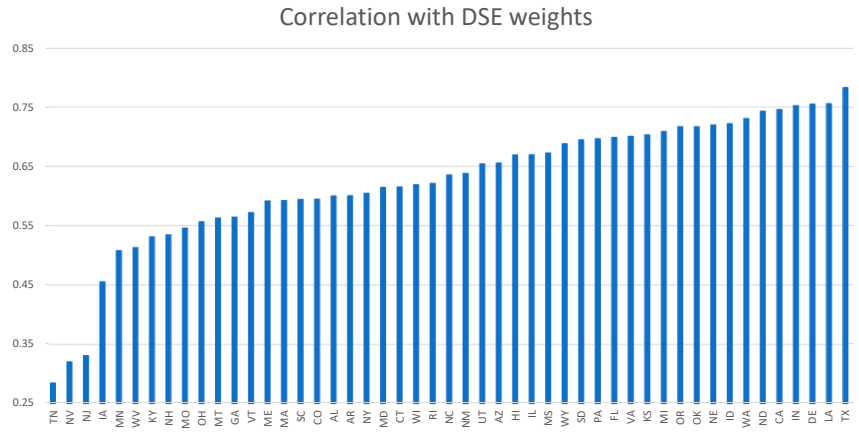


States

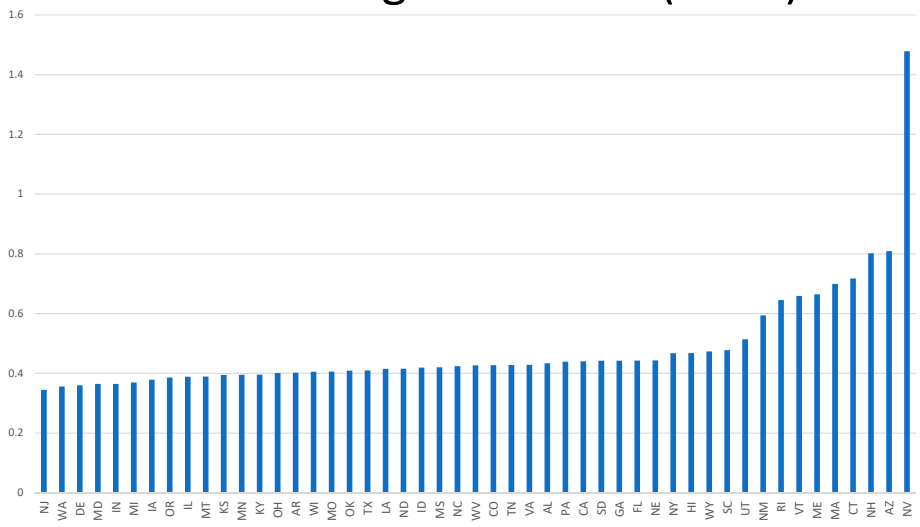
States with all possible targets attained	States with 1 - 5 missed targets			States with 5 - 10 missed targets	States with > 10 missed targets
IN, NY	MN	ID	RI	MA	NV
KY, WA	IL	HI	ME	FL	DE
IA	OR	MT	UT	NM	
KS	SC	LA		WI	
SD	MD	NE		NC	
WV	AR	OH		CT	
VA	AL	MO		AZ	
KS	GA	OK			
PA	CA	NH			
NJ	CO	ND			
TX	MS	WY			
MI	TN	VT			

INCA Missed Targets

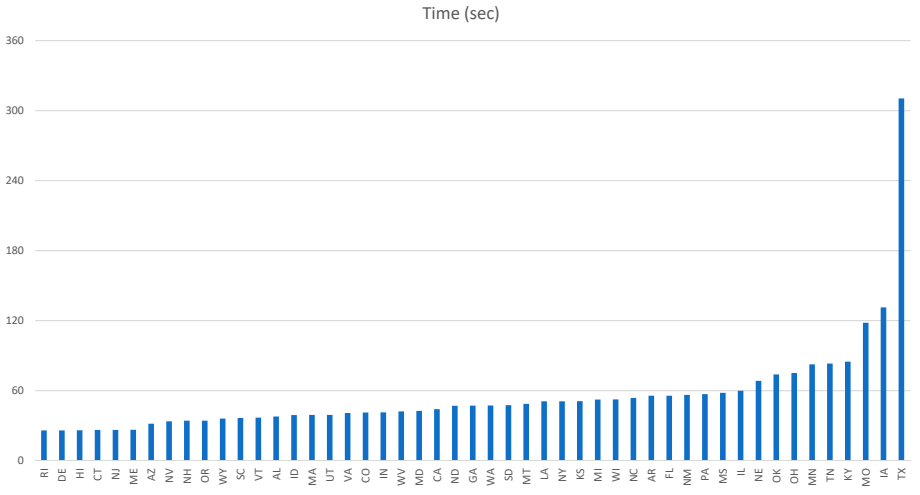
INCA DSE Correlation



INCA
Mean Average deviation (MAD)



INCA Computational Speed



- Average time per state using old code is 30 mins

Findings

- Integer Calibration decreases the number of missed targets in 48 of the 49 states
- Integer Calibration decreases calibration time

Recommendations

Move to incorporate the INCA program into 2017 Census of Agriculture

Thank you!

Luca Sartore, PhD - Luca.Sartore@nass.usda.gov
Kelly Toppin, PhD - Kelly.Toppin@nass.usda.gov
Clifford Spiegelman, PhD - Cliff@stat.tamu.edu