Integer Programming for Calibration

Luca Sartore, PhD – National Institute of Statistical Sciences,
Kelly Toppin – National Agricultural Statistics Service,
Clifford Spiegelman, PhD – Texas A&M University

May 3, 2016

Outline

- Census of Agriculture
- Calibration
- Integer calibration
- Results
- Conclusion

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.
- CML also contains misclassified agricultural operations
- 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

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.

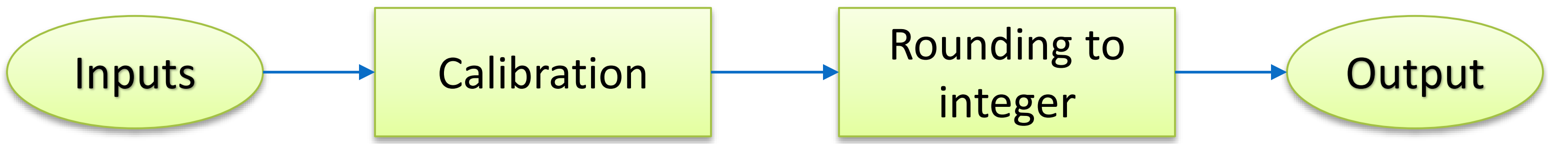
2. Rounding the weights from step 1.

Problems with old approach

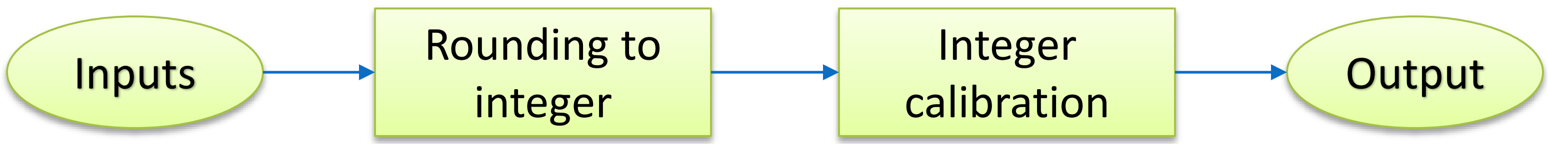
- Too many missed targets
- Final weights are very different than initial (DSE) weights
- Computational intensive and time consuming

Alternative proposal

- Old approach



- New approach



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

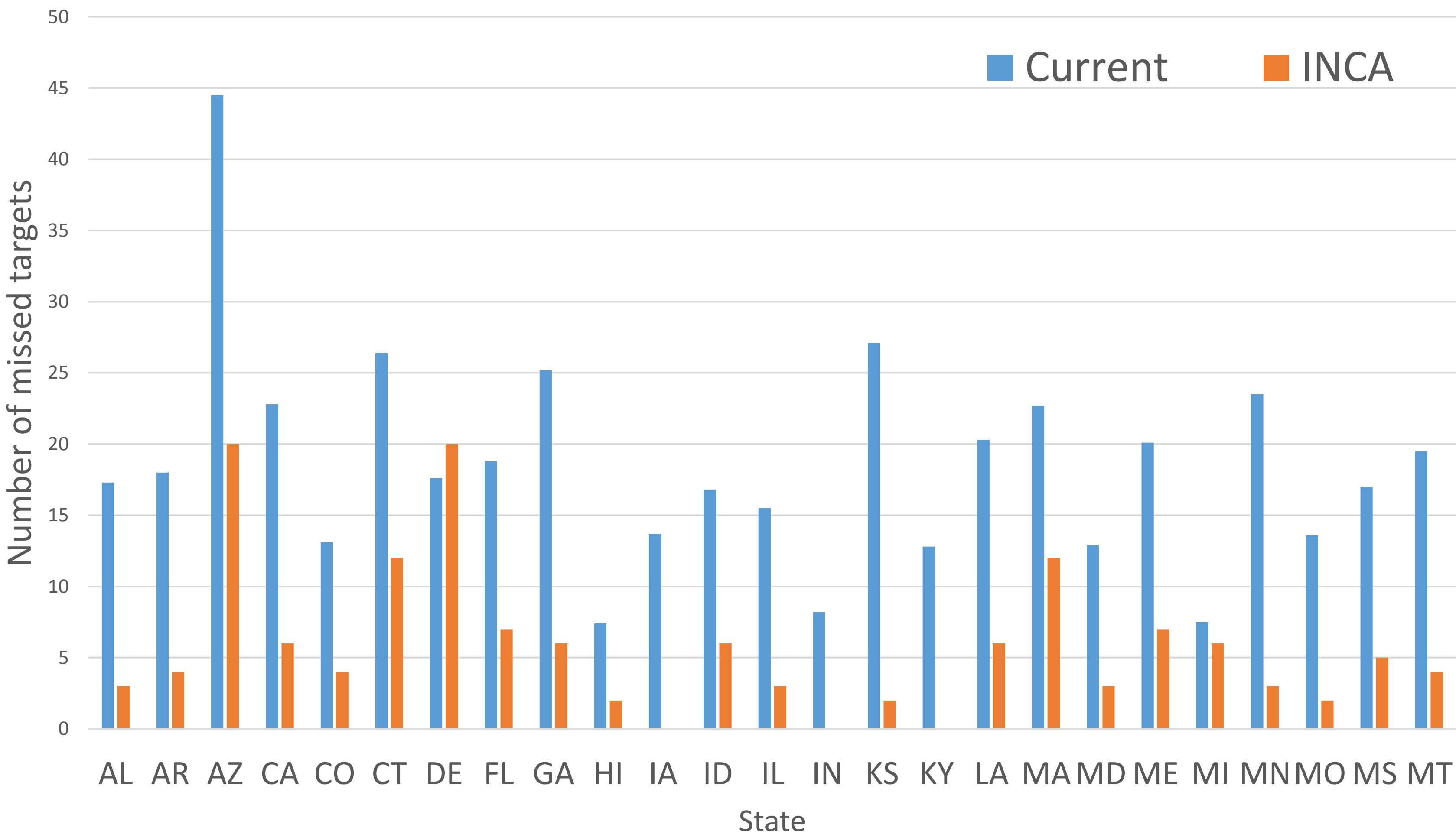
1. All unfeasible weights are truncated to their closest boundary, and in order to minimize the objective function, non-integer weights are then rounded sequentially according to an importance index based on the gradient.

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

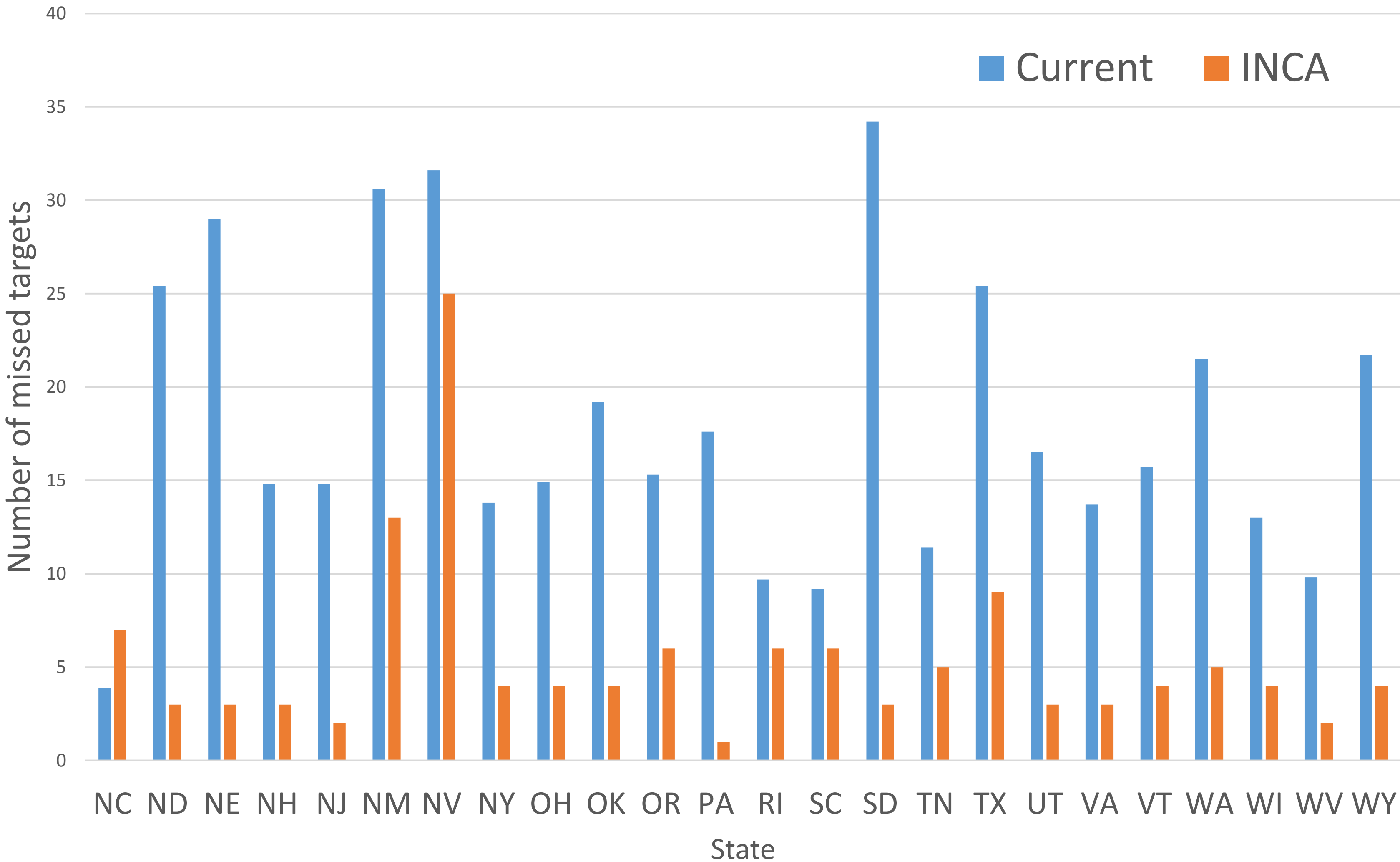
Integer Calibration (INCA)

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

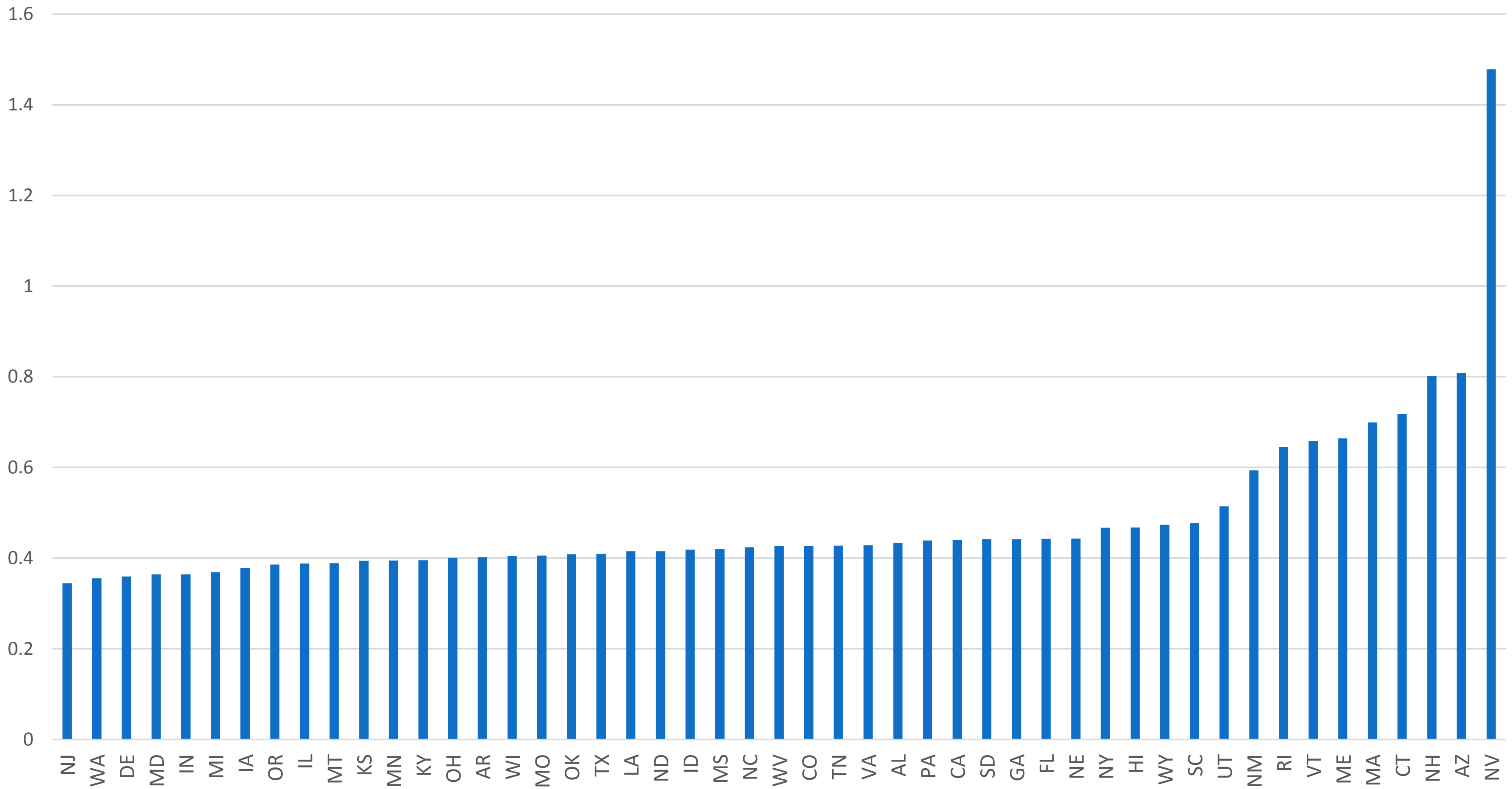
Current VS INCA



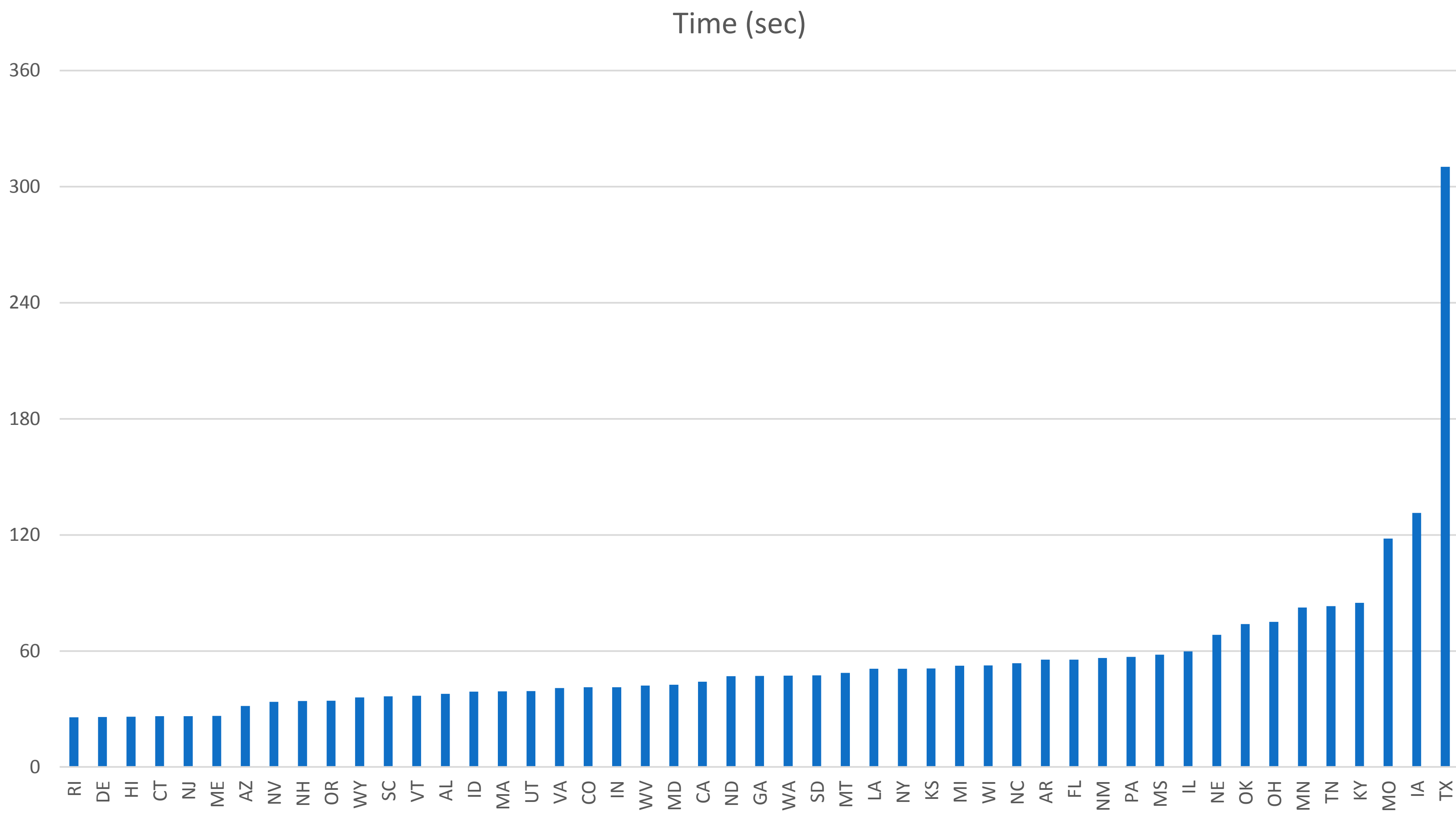
Current VS INCA



INCA
Mean Absolute Deviation from DSE



INCA Computational Speed



- Average time per state using old code is 30 mins

Findings

- Integer Calibration decreases the number of missed targets in 47 of the 49 states
- Integer Calibration decreases calibration time

Status

Moving to incorporate the INCA program into 2017 Census of Agriculture

Thank you!

Luca Sartore, PhD - Luca.Sartore@nass.usda.gov

Kelly Toppin - Kelly.Toppin@nass.usda.gov

Clifford Spiegelman, PhD - Cliff@stat.tamu.edu