Deep Learning for Data Imputation and Calibration Weighting

Yijun Wei[1,2], Luca Sartore[1,2], Jake Abernethy[2], Darcy Miller[2]
Kelly Toppin[2], Michael Hyman[2], Clifford Spiegelman[3]

[1]National Institute of Statistical Science (NISS)
[2]United States Department of Agriculture
National Agricultural Statistics Service (USDA NASS)
[3]Texas A&M University, College Park (TAMU)

ywei@niss.org

JSM 2018 – Leading the Estimates Towards Known Benchmarks

August 1, 2018

Disclaimer

The Findings and Conclusions in This Preliminary Presentation Have Not Been Formally Disseminated by the U.S. Department of Agriculture and Should Not Be Construed to Represent Any Agency Determination or Policy.

Acknowledgements

Denise Abreu
Valbona Bejleri
Matt Fetter
Thomas Jacob
Andrea Lamas
Johnathan Lisic
Beth Schlein
Nell Sedransk
Linda Young

Presentation outline

PART I: MOTIVATION

1. Census of Agriculture
2. A joint methodological framework

PART II: CHALLENGES

3. Data imputation
4. Dual-System Estimation
5. Integer calibrated weights

PART III: METHODOLOGY

5. Deep learning approach
6. Concluding remarks

PART I
MOTIVATION

1. Census of Agriculture
2. A joint methodological framework

Census of Agriculture

Every five years, USDA's National Agricultural Statistics Service (NASS) conducts the Census of Agriculture.

- 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.
- NASS also obtains information on most commodities from administrative sources or surveys of non-farm populations (e.g. cotton ginning data).

Census data imputation

To handle missing data in Census of Agriculture, USDA's National Agricultural Statistics Service (NASS) employs decision logic tables (DLT) for missing data detection and imputation strategy determination.

There are three imputation strategies.

1. Deterministic.
2. Previously reported data.
3. Nearest neighbor donor imputation

DSE: Dual-System Estimation

NASS uses DSE to adjust its estimates by generating weights assigned to each data-record.

- DSE requires two independent surveys to produce adjusted estimates for under-coverage, non-response and incorrect farm-classification at the national, state and county levels.
- The adjusted weights are used as starting values for the calibration process.
- The weights are calibrated to ensure that the Census estimates are consistent across all levels of aggregation and in agreement with information from other sources.

Calibration

A solution $\hat{w}$ such that $T = Aw$, where

T is a vector partitioned into y known and y^* unknown population totals,

A is the matrix of collected data from a population, and

w is a vector of unknown weights.

Calibration finds the solution of the linear system $y = \tilde{A}w$, where

$\tilde{A}$ is a sub-matrix of the collected data.

NASS publishes its estimates by using
integer weights
to avoid fractional farms.

A joint methodological framework

Toppin et al. (2017) developed a joint methodology to perform a DSE and produce integer calibrated weights. They obtained encouraging results by optimizing simultaneously a likelihood with two penalties:

- for LASSO logistic regression, and
- for the calibration benchmarks.

An extension of this research highlights the possibility of a unified approach to data imputation, DSE, and calibration by exploiting deep learning models.

PART II

CHALLENGES

3. Data imputation
4. Dual-System Estimation
5. Integer calibrated weights

Challenges with data imputation

In Donor imputation, each recipient is classified into an appropriate stratum, and the imputation is limited to donors in its stratum.

- The euclidean distance (a sum of squares of distances) between each donor in the same stratum and the recipient is calculated using normalized matching variables.
- The geographic distance is also calculated between each donor in the same stratum and the recipient.
- The geographic distance is added to the euclidean distance and the sum is adjusted by the weight of the census year.
- Donor with lowest value is selected for imputing the recipient.
- Ratio variable is used to scale or adjust the selected donor value for the recipients's value that need to be imputed.

Challenges with DSE

Young et al. (2017) developed a capture-recapture method that requires two independent samples to produce weights associated with the Census records.

These are computed through separated logistic regressions, and the resulting probabilities contribute to form the value of the DSE weights.

$$w_i^{(DSE)} = \frac{\Pr(F|F_{Census})}{\Pr(C|Farm)\Pr(R|C,F)\Pr(F_{Census}|R,C,F)},$$

where

$\Pr(C|F)$ accounts for under-coverage,

$\Pr(R|C,F)$ accounts for non-response, and

$\Pr(F|F_{Census})$ and $\Pr(F_{Census}|R,C,F)$ account for incorrect farm-classifications.

Challenges with integer weights

DSE weights are usually adjusted to be consistent accross all level of aggregation, and rounded to integers to satisfy NASS pubblication standards.

Ideally, DSE weights would be produced so that benchmark equations are satisfied, and rounded with the function:

$$r(x) = \begin{cases} 1, & \text{if } x < 1, \\ \lfloor x \rfloor, & \text{if } 1 \leq x \leq 6, \\ 6, & \text{if } x > 6. \end{cases}$$

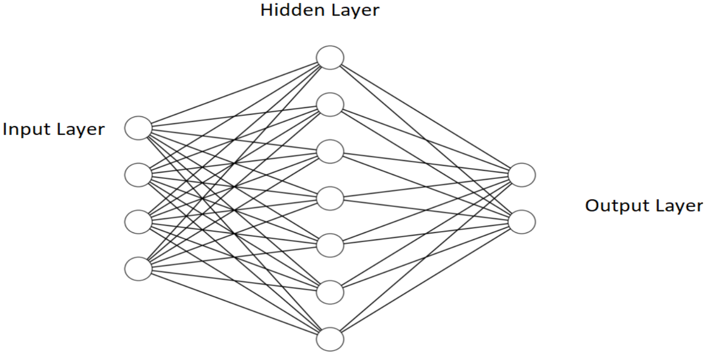
Sartore and Toppin (2016) developed a rounding algorithm that produces optimal integer weights.

PART III
METHODOLOGY

5. Deep learning approach
6. Concluding remarks

Artificial Neural Network (ANN)

- A simple ANN consists of 3 components, i.e. input layer, hidden layer, and output layer.



Artificial Neural Network - Continued

$$\text{hidden}_i = \sigma(\sum_j \omega_{ji}\, \text{input}_j + b_i),$$

where

ω_{ji} represents the j-th parameter in the i-th hidden neuron,

b_i denotes a constant (or “bias”) in i-th neuron, and

input_j is the j-th covariate (or inpunt component).‘

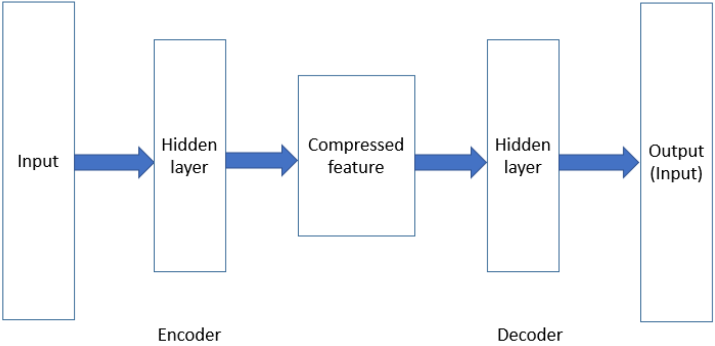
$$\text{output}_i = \text{softmax}(\sum_j \omega_{ji}\, \text{hidden}_j + b_i),$$

where

ω_{ji} and b_i are randomly initialized and updated through backpropagation.

Autoencoder

- Autoencoder is a type of ANN that is trained to attempt to copy its input to its output.



A joint deep learning framework

- An antoencoder network is constructed for imputing missing values.
- After imputing the missing values, decoder of the network is removed and the encoder is kept.
- The joint objective function is minimized by tuning the parameters of the encoder network.
- The final encoder network is used to generate the parameters assigned to each data record.

A joint optimization settings

The idea is to optimize the parameters of two networks:

1. The encoder ANN that generates the missing values.
2. The ANN used to estimate the probabilities that form the DSE weights.

Mathematically, we can formulate the following optimization problem as

$$\min_{\omega \in \mathbb{R}^p} L_{\mathrm{Imp}}(\omega) + L_{\mathrm{DSE}}(\omega) + L_{\mathrm{Cal}}(\omega),$$

where

L_{Imp} denotes the loss function used for imputation,

L_{DSE} represents the loss function used for DSE, and

L_{Cal} measures the distance from the calibration benchmarks.

Concluding remarks

- The proposed methodology is going to automate several processes within NASS, and will decrease computational time and estimation efforts.

- This approach has the potential to take into account all sources of variation and simplify the computation of standard errors.

- This method will allow to integrate all the information available to a unique framework where imputation, DSE, and calibration will be jointly preformed.

- Future research will focus on the formalization of a multifunctional network, and on a broad simulation study to identify the best estimation strategies and loss functions.

Selected References

Sartore, L. and Toppin, K. (2016). *inca: Integer Calibration*. R package version 0.0.2.

Toppin, K., Sartore, L., and Spiegelman, C. (2017). Design weights and calibration. In *Proceedings of JSM 2017, Government Statistics Section*, pages 2318–2322, Alexandria, VA. American Statistical Association.

Young, L. J., Lamas, A. C., and Abreu, D. A. (2017). The 2012 Census of Agriculture: a capture–recapture analysis. *Journal of Agricultural, Biological and Environmental Statistics*, 22(4):523–539.

Thank you!

Questions?

Yijun Wei	ywei@niss.org
Luca Sartore, PhD	lsartore@niss.org
Jake Abernethy	jake.abernethy@nass.usda.gov
Darcy Miller, PhD	darcy.miller@nass.usda.gov
Kelly Toppin, PhD	kelly.toppin@nass.usda.gov
Michael Hyman, PhD	michael.hyman@nass.usda.gov
Clifford Spiegelman, PhD	cliff@stat.tamu.edu