Using Neural Networks to Assess Uncertainty for the Imputation of an Area Survey

Sean Rhodes[1], Luca Sartore[1,2], Tara Murphy[1] , Arthur Rosales[1]

[1] National Agricultural Statistics Service, 1400 Independence Ave SW 20250 Washington D.C., USA

[2] National Institute of Statistical Sciences, P.O. Box 33762, Washington D.C., USA 20033

Background

Every year the United States (US) Department of Agriculture's National Agricultural Statistics Service (NASS) conducts the June Area Survey (JAS) based on an area frame, which has complete coverage of all land in the contiguous US. Response rates have been declining in many federal surveys, including the JAS, leading to increased item and unit nonresponse. NASS has begun the exploration of automatic imputation for the JAS using models based on artificial intelligence (AI). Previous research has found that NASS's Predictive Cropland Data Layer (PCDL) and Crop Sequence Boundaries (CSBs) have good predictive power for corn and soybeans (two major US crop commodities). This poster presents the application of customized neural network architectures to estimate the uncertainty associated with the JAS imputed acreage.

Methods

5877 JAS tracts, from the 2021 sample and each with less than 150 acres, were used for analysis. These records were from the Corn Belt States (IL, IN, IA, MI, MN, MO, OH, WI).

To estimate the standard errors (SEs) for the imputed values, let

- A_i = ground-truth Farm Service Agency (FSA) acreage of a specific commodity planted within digitized tract i,
- W_i = total size of JAS digitized tract i expressed in acres,
- $\hat{A}_i$ = weighted average between the PCDL, $\hat{A}_{i,P}$, and CSB predicted acreage, $\hat{A}_{i,C}$, computed for digitized tract i, where the weights are nonlinear functions of input variables $\boldsymbol{X}_i$:

$$\hat{A}_i = \hat{A}_{i,P}\,\omega_P(\boldsymbol{X}_i) + \hat{A}_{i,C}\omega_C(\boldsymbol{X}_i).$$

The absolute value of the relative error made when imputing the acreage of the tract i is defined as

$$Y_i = \left|\frac{\hat{A}_i - A_i}{W_i}\right|,$$

where Y_i is assumed to be distributed as an inflated beta random variable, which also accounts for zeros and ones (Ospina and Ferrari, 2012).
The maximum log-likelihood is computed using the Inflated Beta distribution, i.e.,

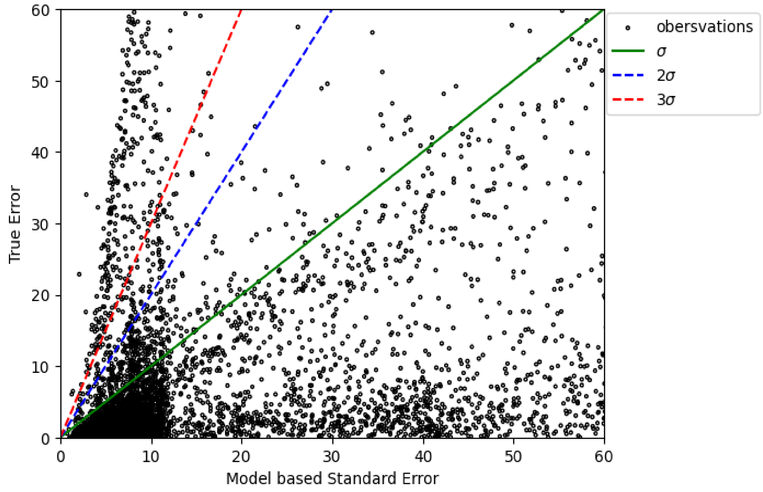
$$f_Y(y;\mu,\sigma,\nu,\tau) = (1+\tau+\nu)^{-1}\begin{cases}\nu, & \text{if } y=0,\\ f_B(y;\mu,\sigma), & \text{if } y\in(0,1),\\ \tau, & \text{if } y=1,\end{cases}$$

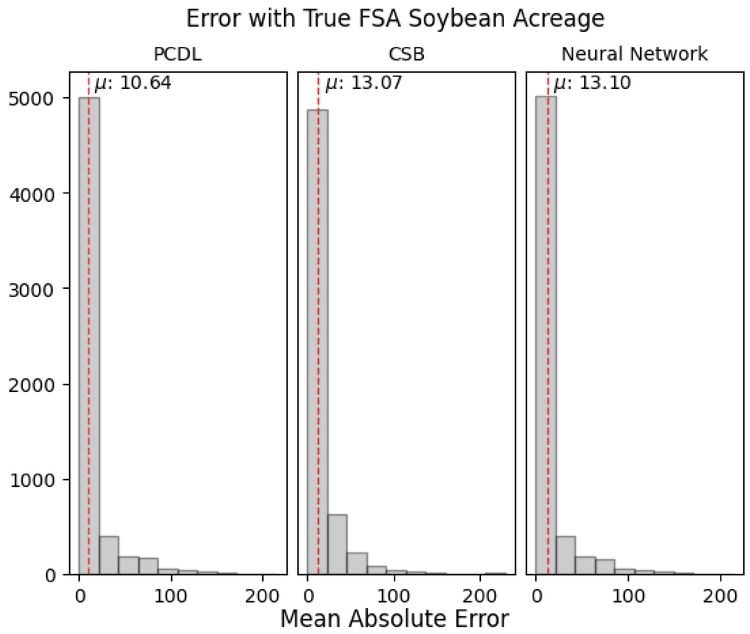
where the parameters μ, σ, ν and τ are nonlinear functions of $\boldsymbol{X}_i$. In particular, the parameters μ and σ respectively control for the location and scale of the Beta distribution, while ν and τ respectively represent the inflation probabilities for the zeros, $\Pr(Y_i = 0 \mid \boldsymbol{X}_i)$, and for the ones, $\Pr(Y_i = 1 \mid \boldsymbol{X}_i)$.
The standard error (SE) associated with the imputed value $\hat{A}_i$ is computed as

$$\mathrm{SE}[\hat{A}_i|X_i] = \sqrt{\mathrm{Var}[\hat{A}_i|X_i]} = W_i\sqrt{\mathrm{E}[Y_i^2|X_i]} = W_i\sqrt{\mathrm{Var}[Y_i|X_i] + \mathrm{E}[Y_i|X_i]^2}.$$

Three neural networks are developed and combined in a unified framework for the simultaneous computations of imputed values and their standard errors.

Model 1: Predicts the acreage using semi-parametric weighted averages.
Model 2: Assesses the uncertainty of the absolute value of the relative errors.
Model 3: Unifies the results from the models to produce both acreage and variance.





Three Neural Networks Unified for Computing Acreage and Variance Table 1.1

Layer	Neurons	Processed Information
1.Input Entropy & Diff	2	
2.Dropout (10%)	2	1
3.Dropout (10%)	2	1
4.Dense (ReLU)	12	2
5.Dense (ReLU)	12	3
6.Dropout (10%)	12	4
7.Dropout (10%)	12	5
8.Dense (ReLU)	6	6
9.Dense (ReLU)	6	7
10.Concatenate	8	(1,8)
11.Concatenate	8	(1,9)
12.Dropout (10%)	8	10
13.Dropout (10%)	8	11
14.Input	5	
15.Dense (softmax)	3	12
16.Dense (sigmoid)	2	13
17.Dense (ReLU)	10	14
18.Concatenate	5	(15,16)
19.Input Tract Acreage	1	
20.Dense (softmax)	2	17
21.Input PCDL &CSB	2	
22.Concatenate	6	(18,19)
23.Dot	1	(20,21)
24.SE (custom layer)	1	22
25.Concatenate	2	(23,24)

Table 1.1 Key

Input	
Model1	
Model2	
Intermediate Output	
Final Output (Model 3)	

Discussion

Neural networks provide flexibility that better capture the relationship between the summarized PCDL entropy and JAS imputation error for Soybean planted acres. However, entropy information alone is not sufficient to determine variability of the JAS imputation errors. A key finding from this poster is the potential to simultaneously produce imputation values and model-based SEs before the FSA data are available.

Disclaimer

The findings and conclusions in this poster are those of the authors and should not be construed to represent any official USDA, US Government, or NISS determination or policy.

References

Hunt, K., Abernethy, J., Beeson, P., Bowman, M., Wallander, S., Williams, R. (2023). Crop Sequence Boundaries (CSB): Delineated Fields Using Remotely Sensed Crop Rotations. In the Proceedings of the IX International Conference of Agricultural Statistics.
Rhodes, S., Rosales, A., Sartore, L., & Murphy, T. (2023). Uncertainty Assessment for the Imputation of an Area Survey. In Proceedings of the Joint Statistical Meeting. https://doi.org/10.5281/zenodo.8383528
Ospina, R., & Ferrari, S. L. (2012). On bias correction in a class of inflated beta regression models. International Journal of Statistics and Probability, 1(2), 1-14.
Sartore, L., Boryan, C., Dau, A., & Willis, P. (2023). An Assessment of Crop-Specific Land Cover Predictions Using High-Order Markov Chains and Deep Neural Networks. Journal of Data Science, 21(2).
US Department of Agriculture, Farm Service Agency. (2022) Common Land Units (CLUs).
Zhang, C., Di, L., Lin, L., & Guo, L. (2019). Machine-learned prediction of annual crop planting in the US Corn Belt based on historical crop planting maps. Computers and Electronics in Agriculture, 166, 104989.