Deep Learning for Data Imputation and Calibration Weighting

Yijun Wei*† Luca Sartore*† Jake Abernethy† Darcy Miller†
Kelly Toppin† Michael Hyman†

Abstract

The USDA's National Agricultural Statistics Service (NASS) surveys are affected by nonresponse and by incomplete responses that may not be homogeneous across farm types and sizes. To address item nonresponse, NASS employs a variety of imputation methods such as ratio imputation, iterative sequential regression, fully conditional specification, K-nearest neighbor, carry forward of previously reported data and manual imputation to provide reliable and consistent values on NASS data. To address unit nonresponse and some other sources of error, NASS currently uses a set of generalized linear regression models to estimate the number of US farms by calibrating their corresponding weights. However, linear models cannot always capture important nonlinear features of the population. Deep learning (including artificial neural network) models are used successfully in numerous other applications in order to capture nonlinear properties efficiently. In this paper, imputation techniques and the adjustment of the survey weights are integrated. A potential unified deep learning method simultaneously adjust survey weights and impute missing values is discussed.

Key Words: Imputation, Neural network model, Calibration, Dual-system estimation, Capture-Recapture, Survey data.

1. Introduction

The US Census of Agriculture is conducted by the USDA National Agricultural Statistics Service (NASS) every five years, in years ending in 2 and 7. The Census provides a detailed picture of US farms, ranches, and the people who operate them. It is also the only source of uniform comprehensive agricultural information for every state and county in the United States. Missing information requires the use of data imputation and calibration of sample weights.

The Census is based on the NASS list frame, which contains agricultural operations that should satisfy the farm definition (O'Donoghue et al., 2009). This list frame is also used to conduct many agricultural surveys, and for this reason, the maintenance of the list frame is a crucial ongoing effort that intensifies leading up to the Census. When the list frame is "frozen" at a specific time, it becomes the Census Mailing List. This list is incomplete, and not all the agricultural operations on the list satisfy the farm definition or respond to the Census questionnaire. To address these issues, a Dual-System Estimation (DSE) methodology was developed at NASS (Young et al., 2017) in order to adjust the estimates for coverage, non-response and misclassification. NASS also obtains information on most commodities from administrative sources, which provide reliable information used to reduce the bias of the estimates through calibration.

Furthermore, missing values and outliers cannot be avoided during data collection; therefore, editing and imputation mechanisms are developed to handle these issues. These mechanisms consist of three strategies. The first is deterministic; the values to impute are determined through the evaluation of other data provided by the respondent. The second

[*] National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006, ywei@niss.org

[†] National Agricultural Statistics Service, United States Department of Agriculture, 1400 Independence Ave. SW, Washington, DC 20250

employs previously reported data, and the third is based on a nearest neighbor approach (Miller and Young, 2015).

It should be noted that the imputation is performed before DSE, which is followed by calibration. In fact, the mechanism of imputation is not considered in the calibration, which could be affected by the change of the data distribution risking additional bias in the estimates. Therefore, a unified approach to data imputation, DSE, and calibration that can exploit deep learning models is developed both with the purpose of bias reduction and learning more features (predictors).

Many approaches have been proposed and used to impute for missing or erroneously reported data. The autoregressive integrated moving-average (ARIMA) model (Nihan, 1997) and the Markov Chain Monte Carlo (MCMC) multiple imputation method (Ni et al., 2005) are two examples of imputation techniques found in the literature. With the increasing quality and quantity of data, an automatic and more efficient approach should be developed to handle massive data sets. Duan et al. (2016) proposed a denoising autoencoder (DA) deep learning model to deal with traffic data imputation, and the performance of the model was found to outperform that of ARIMA model and MCMC multiple imputation method in terms of imputation accuracy. A possible disadvantage of a deep learning strategy lies in the difficulty of explaining the model. Duan et al. (2016) addressed this by explaining the DA model by visualization.

A better version of DA based on a deep learning model is described by Gondara and Wang (2018). Unlike the previous studies, which employed the complete observations for training and dealt only with the same type of missing mechanism (e.g. missing at random), their approach and the one proposed here are designed to handle multiple missing mechanisms in the incomplete observation training dataset. Gondara and Wang's approach (2018) was also tested with different data sets that were modified by various missing mechanisms, and its performance is better than other proposed solutions.

To improve the estimates of population totals, Lemel (1976) introduced the first idea of calibration. This gained importance after Deville (1988), and it was generalized by Deville and Särndal (1992). Singh and Mohl (1996) distinguished between two types of calibration approaches:

1. those that iterate until weight restrictions are met while satisfying the calibration equations;

2. and those that iterate until the calibration equations are met while satisfying the range restrictions on the weights.

Even if both approaches are asymptotically equivalent, the interval length of the range restrictions imposed on the weights has an impact on the point estimates and their precision. In light of the DSE methodology proposed by Young et al. (2017), an alternative to these two approaches is an iterative optimization that stops only when both weight restrictions and calibration equations are met.

Imputation, DSE and calibration are sequentially performed operations and the effects of these successive algorithms on the final estimates are not obvious. This motivated Slud and Thibaudeau (2010) to develop an algorithm that simultaneously performs calibration and non-response adjustments by minimizing a multi-objective function. Subsequently, Slud et al. (2013) extended this technique to deal with soft-constraints, and Shaffer et al. (2014) studied the impact of penalties. Elkasabi et al. (2015) proposed a joint calibration estimator for a dual survey system, while Toppin et al. (2017) attempted to perform calibration and DSE simultaneously.

As the previous authors investigated the combination of DSE and calibration, a unified approach that also incorporates data imputation is proposed in this paper. A DA model is developed during the imputation stage, and the learned features are used sucessively for DSE and calibration. Further details on imputation, DSE, and calibration are discussed in the second section. The joint model will be proposed in the third section. The conclusion will be delivered in section 4.

2. Methodology review

2.1 Notation

The following notation is used throughout the paper:

$\mathbf{x}_j$	Vector of covariates in the input layer
$\mathbf{h}_j$	Vector of values stored in hidden layers
$\mathbf{z}_j$	Vector of values stored in the output layer
$\boldsymbol{\omega}$	Vector of parameters
$\sigma(\cdot)$	Activation function (often a sigmoid)
$\mathcal{U}$	A finite heterogeneous population
$\mathcal{S}$	A sample selected from $\mathcal{U}$
s_k	Statistical unit sampled from $\mathcal{U}$
α	Elastic-net factor controlling the regularization
$\mathbf{A}$	An $n \times p$ matrix of collected data
$\mathbf{a}_k$	The k-th row of matrix $\mathbf{A}$
$\mathbf{y}$	Vector of targets (known totals)
ℓ_k	Lower bound of the k-th target
u_k	Upper bound of the k-th target
$\mathbf{w}$	Vector of final calibrated weights
g_k	Upper bound of weight restrictions

2.2 Imputation

2.2.1 Missing data mechanism

There are three types of missing mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). Data that are MCAR indicate that the probability of any observation having missing data is the same. Data that are MAR imply that the probability of missingness depends on observed information. Data are MNAR if the missing depends on unobserved information, or the missing in a variable happens based on the variable itself (Leke et al., 2015).

2.2.2 Handling missingness

Gondara and Wang (2018) proposed a multiple deep denoising autoencoder imputation model to handle the missing data. Unlike previous studies, which employ the complete observations in training and only deal with the same type of missing mechanism (for example, missing at random), their proposed approach is designed to handle multiple missing mechanisms in the incomplete observation dataset. US Census of Agriculture data are imputed by 3 strategies:

1. deterministic (any value that can be determined through the evaluation of relevant responses (e.g. a missing total) is imputed);

2. previously reported data (previously-reported data assembled from a variety of NASS surveys together with previous Census of Agriculture data are used for imputation), and

3. nearest neighbor donor imputation (an imputation approach named donor imputation, which adopts nearest neighbor method and is a type of automated imputation, is used).

The alternative strategy investigated here is to use a deep denoising autoencoder model, which is based on Artificial Neural Network (ANN).

2.2.3 Artificial neural network

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

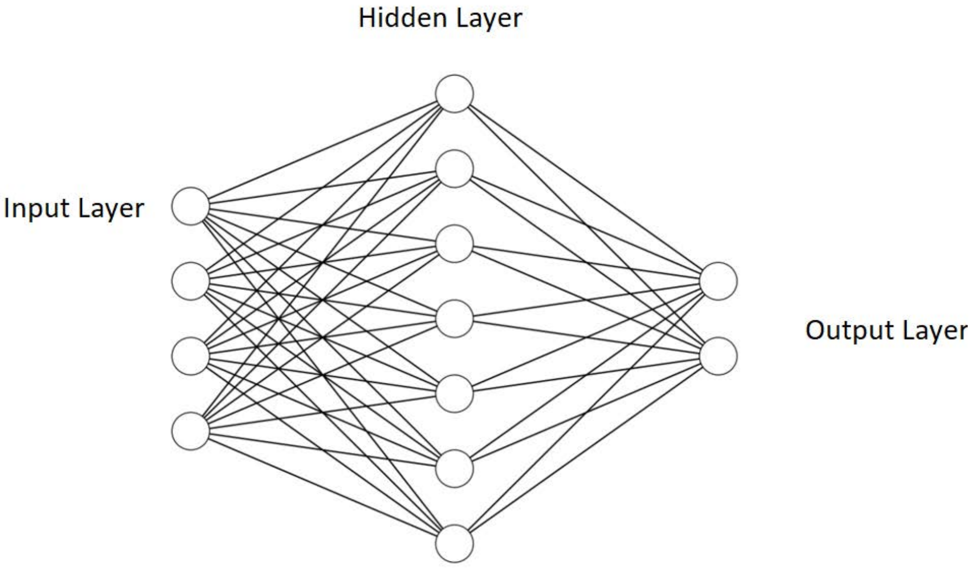

Figure 1: The structure of an artificial neural network.

In Figure 1, one hidden layer is displayed between the input and output layer. Multiple hidden layers can be in an ANN but they have the same structure.

A hidden layer consists of several neurons and these neurons are used to form output layer. Hidden neurons in a hidden layer are calculated as:

$$h_i = \sigma\left(\omega_{0i} + \sum_j \omega_{ji}\, x_j\right),$$

$$z_i = \text{softmax}\left(\omega_{0i} + \sum_j \omega_{ji}\, h_j\right),$$

where h_i represents the i-th hidden neuron, x_j represents j-th input, ω_{ij} represents parameters in i-th neuron, ω_{0i} represents bias in i-th neuron, x_j represents the j-th component, i.e. observation of input layer, and $\sigma(\cdot)$ is the activation function and can be defined as a sigmoid function, e.g. $\tanh(\cdot)$ function or the Rectified Linear Units (ReLU). All parameters will be randomly initialized and then updated via back-propagation.

2.2.4 Autoencoder and denoising autoencoder

An autoencoder is a type of ANN that is trained to attempt to copy its input to its output. Internally, it has a hidden layer that describes a code used to represent the input. The network may be viewed as consisting of two parts: an encoder represents a feature extracted process and a decoder that produces an input reconstruction. This architecture is presented in Figure 2. Three hidden layers are in Figure 2, but the size of the middle hidden layers in encoder and decoder have the same structure.

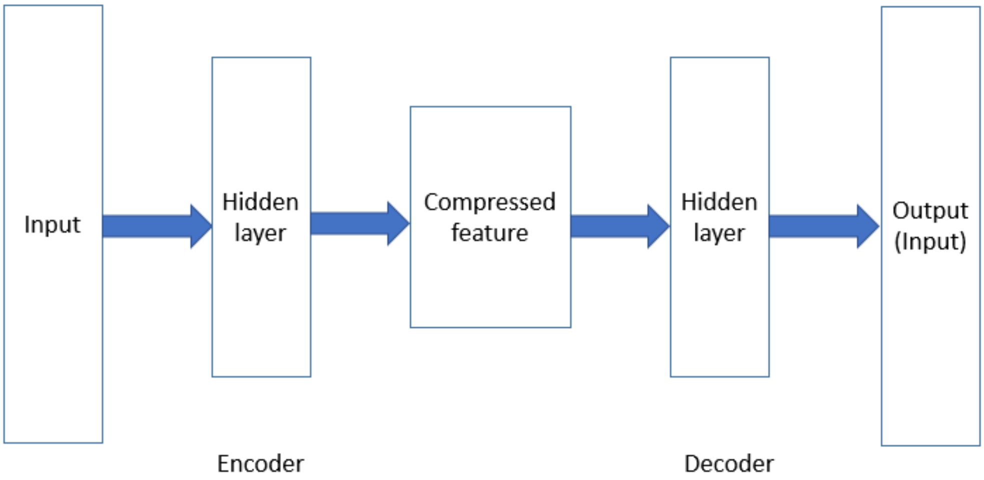

Figure 2: The structure of an autoencoder network.

Denoising autoencoder is an extension of autoencoder (Vincent et al., 2008) that seeks to learn more robust features, i.e. avoid overfitting, by corrupting the input data. Corrupting can be completed through a number of ways and one approach is randomly removing some observations.

2.3 Dual-system estimation

DSE methods are sophisticated techniques to estimate the number of unsampled units within a closed population $\mathcal{U}$. At least two independent samples are collected from the same finite population, $\mathcal{S}_1 \subset \mathcal{U}$ and $\mathcal{S}_2 \subset \mathcal{U}$, must satisfy the inequality $\mathcal{S}_1 \cap \mathcal{S}_2 \neq \emptyset$. It is possible to estimate the cardinality of the set $\mathcal{U} \backslash (\mathcal{S}_1 \cup \mathcal{S}_2)$ under the assumption that it is not empty.

Alho (1990) proposed a logistic regression model in a capture-recapture setting to estimate the size of a heterogeneous closed population. When all units in the population belong to a sample $\mathcal{S}$, it is called a census. However, a complete enumeration rarely occurs due to under-coverage, non-response and, sometimes, misclassification. Thus, most censuses can be viewed as extensive surveys that require adjustments to extend the inferential results to the entire population. This can be achieved by a capture-recapture methodology that takes into account the enumeration issues mentioned above.

Young et al. (2017) developed a DSE method for the US Census of Agriculture that performs separate logistic regressions to compute four distinct probabilities that are successively used to compute the adjusted weights. These are formulated as

$$w_k^{(\text{DSE})} = \frac{\Pr(F|F_{\text{Census}})}{\Pr(C|F)\Pr(R|C,F)\Pr(F_{\text{Census}}|R,C,F)}, \tag{1}$$

where $\Pr(F|F_{\text{Census}})$ and $\Pr(F_{\text{Census}}|R,C,F)$ respectively address farm over-count and under-count due to incorrect farm-classification, $\Pr(C|F)$ accounts for under-coverage given that the statistical unit is a farm record, and $\Pr(R|C,F)$ quantifies the propensity of a farm to respond to the Census questionnaire. This approach however does not take into account possible nonlinear relationships among the data. Also it necessitates extensive training to deal with variable selection to improve the accuracy of the results.

In line with the approach described for imputation, artificial neural networks can introduce nonlinear features in the model and allow for a simultaneous process of the information required to fit the four probabilities. The adjusted weights are then produced as a combination of the four probabilities stored in the four neurons of the output layer, i.e.

$$z_i = \text{logit}^{-1}\left(\omega_{0i} + \sum_j \omega_{ji}\, h_j\right).$$

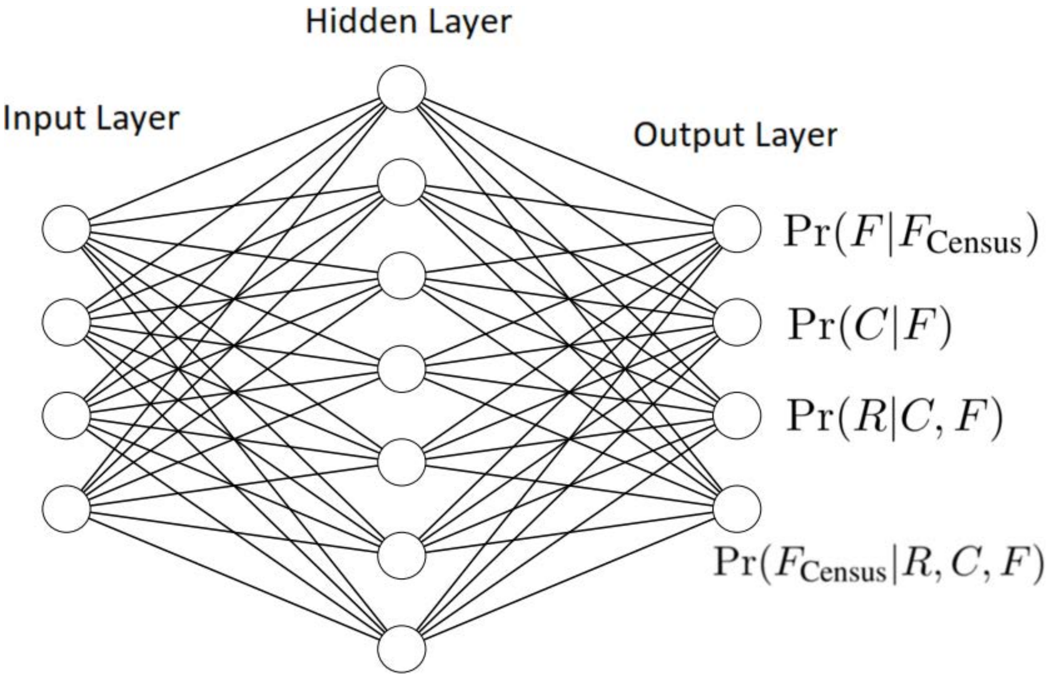

Figure 3: The structure of the DSE neural network with four neurons in output.

2.4 Calibration

The DSE weights are further adjusted to produce consistent estimates across all levels of aggregation. NASS employs integer calibrated weights to calculate the final values of its Census estimates and avoid fractional farms. Ideally, the benchmark equations should be already satisfied just by rounding the DSE weights instead of processing them by rounding algorithms developed *ad-hoc* (Sartore et al., 2018).

Calibration is generally performed to find an optimal vector of weights that satisfy a set of linear equations. At NASS, the calibration equations are evaluated as a loss function rather than considered as constraints. The reduction of the relative errors between benchmarks and estimates is performed while satisfying the range restrictions of the weights.

Given the formulation of the adjusted weights as in (1), the range restrictions are not considered as constraints but as an additional term in the objective function that penalizes those solutions that lie outside the assigned intervals. As it was pointed out by Toppin et al. (2017), the DSE weights before rounding should lie between 1 and the upper bound $g_k \in \mathbb{N}$, which is defined for any Census record, i.e. for any $s_k \in \mathcal{S}_1$. The approach proposed below is consistent with the application of artificial neural networks to estimate the probabilities involved in the computation of the DSE weights.

The calibration equations are incorporated into a loss function that accounts for benchmarks defined both by numbers (hard targets) and intervals (soft targets).

$$\sum_{k\in\mathcal{H}}\left|\frac{\mathbf{a}_k\mathbf{w}-y_k}{y_k}\right|+\sum_{k\in\mathcal{I}}\begin{cases}(y_k-\mathbf{a}_k\mathbf{w})/(y_k-u_k), & \text{if } \mathbf{a}_k\mathbf{w}>u_k,\\ (y_k-\mathbf{a}_k\mathbf{w})/(y_k-\ell_k), & \text{if } \mathbf{a}_k\mathbf{w}<\ell_k,\\ 0, & \text{otherwise.}\end{cases}$$

where $\mathcal{H}$ represents the set of indexes of the known totals that need to be matched exactly, $\mathcal{I}$ represents the set of indexes of the totals that has to lie within confidence boundaries, $\mathbf{a}_k$ corresponds to the k-th row of $\mathbf{A}$, which is an $n\times p$ matrix of collected data (some of which are imputed), y_k denotes the k-th known totals, while ℓ_k and u_k respectively represent the lower and upper bound of the confidence interval of the k-th target. The vector of final calibrated weights $\mathbf{w}$ is computed such that its components satisfy the following equation:

$$w_k^{(\text{Cal})}=\lfloor w_k^{(\text{DSE})}\rceil.$$

Since the rounding $w_k^{(\text{DSE})}$ does not take into account of the range restrictions on the weights, it is necessary to add a penalty defined as

$$\begin{cases}1-w_k, & \text{if } w_k<1,\\ w_k-g_k, & \text{if } w_k>g_k,\\ 0, & \text{otherwise.}\end{cases}$$

2.5 Penalty function

Usually, regression models are fitted and evaluated to select the best predictors according to specific criteria. Among the most common variable selection techniques, the elastic-net penalty developed by Zou and Hastie (2005) is a function of the parameter to estimate and combines the LASSO and ridge penalty (Zou and Hastie, 2005; Friedman et al., 2010). It is formulated as

$$(1-\alpha)\frac{1}{2}\|\omega\|_2^2+\alpha\|\omega\|_1,$$

where the notation $\|\cdot\|_1$ represents the L^1-norm used to perform the LASSO regularization, and $\|\cdot\|_2$ denotes the L^2-norm for the ridge regularization. The factor α is used as a trade-off between the LASSO ($\alpha=1$) and ridge regularization ($\alpha=0$). The elastic-net can be a valid solution to reduce the number of edges between neurons of two different layers in the networks described in the previous sections. Regression models are fitted using all the predictors; therefore, adding this penalty to the objective function is useful for dealing with many correlated predictor variables and hidden neurons.

3. Joint optimization

The proposed approach simultaneously optimizes two neural networks. The first network, the autoencoder, addresses the problem of missing values by providing a reliable value to be used successively by the other network. The second network removes the decoder part of the first network and employs features extracted from the imputation process to quantify the DSE probabilities that are used to compute the adjusted weights. Mathematically, the adaptation of these networks can be formulated as a single objective function that can be expressed as the sum of the loss functions employed for imputation, DSE and calibration.

Without loss of generality, the objective function is formulated as

$$\min_{\omega} L_{\text{Imp}}(\omega) + L_{\text{DSE}}(\omega) + L_{\text{Cal}}(\omega) + L_{\text{Pen}}(\omega) + L_{\text{Res}}(\omega),$$

where L_{Imp} denotes the loss function used for imputation, L_{DSE} represents the loss function used for DSE, L_{Cal} measures the distance from the calibration benchmarks, L_{Pen} quantifies the elastic-net penalty, and L_{Res} accounts for the range restrictions on the calibrated weights.

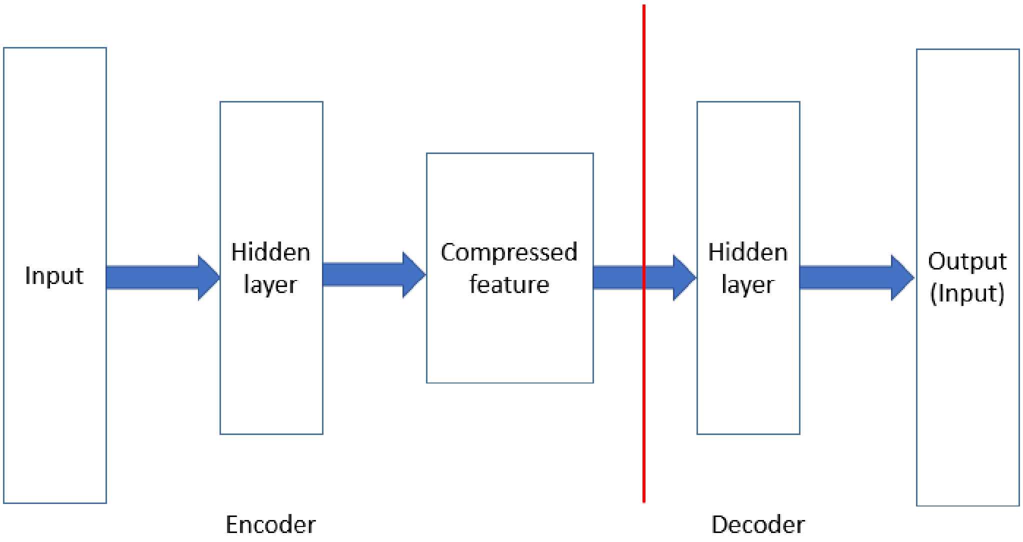

Figure 4: The first network used for imputation.

Figure 4 displays the first network, which is used for imputation; mean square error and cross entropy are the most common loss functions for numerical and categorical variables respectively. These two loss functions will be minimized simultaneously through backpropagation. A grid search is conducted to look for the number of layers and neurons in terms of the minimized loss. The model can be trained for 500 epochs using an adaptive learning rate with a time decay of 0.99 and Nesterov's accelerated gradient (Nesterov, 1983, 2012, 2013). The dropout ratio is set to 0.5 and $\tanh$ is used as the activation function because $\tanh$ performs better than other activation functions for small or moderate size datasets. Since a DA model requires complete data at the initialization, the median for numerical variables and the most frequent class for categorical variables are used initially to replace the missing values within each variable. These missing values will be imputed later when the network is constructed (Gondara and Wang, 2018). In practice, when the DA model is constructed, the decoder part is removed and the second network is formed as indicated in Figure 5. The weights of hidden layers of this network are fixed before the addition of new hidden layers. Then the weights in the newly added hidden layers will be trained through back-propagation to minimize the objective function. Finally, the probabilities are calculated and the calibrated weights can be calculated by combining those probabilities and truncating to the lowest integer.

The values generated by the encoder are used by the other loss functions; consequently their impact can be measured on both the DSE probabilities and the calibrated totals.

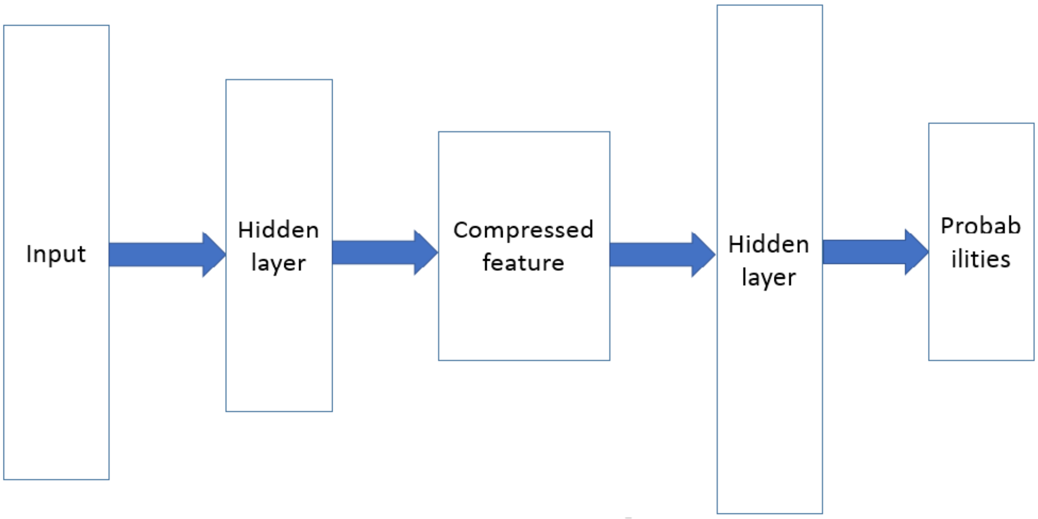

Figure 5: The second network used for DSE.

4. Conclusion

NASS currently handles imputation, DSE, and calibration in separate steps. Toppin et al. (2017) improved the variable selection and computational costs, and designed a model that combines DSE and calibration. However, the mechanism of imputation was not considered as part of that joint approach. This may introduce some bias in the final estimates. A unified approach integrating all three processes jointly was derived in this study.

The proposed methodology has the potential 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. 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.

Acknowledgments

The authors would like to thank Nathan Cruze, Nell Sedransk, Kay Turner, and Linda Young for reviewing the earlier versions of this paper and providing useful comments. This research was supported by the intramural research program of the U.S. Department of Agriculture, NASS. The findings and conclusions in this preliminary publication have not been formally disseminated by the U.S. Department of Agriculture and should not be construed to represent any agency determination or policy.

References

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.

Deville, J.-C. (1988). Estimation linéaire et redressement sur information auxiliaire d'enquêtes par sondage.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Duan, Y., Lv, Y., Liu, Y.-L., and Wang, F.-Y. (2016). An efficient realization of deep learning for traffic data imputation. *Transportation research part C: emerging technologies*, 72:168–181.

Elkasabi, M. A., Heeringa, S. G., and Lepkowski, J. M. (2015). Joint calibration estimator for dual frame surveys. *Statistics in Transition new series*, 16(1).

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Gondara, L. and Wang, K. (2018). Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 260–272. Springer.

Leke, C., Marwala, T., and Paul, S. (2015). Proposition of a theoretical model for missing data imputation using deep learning and evolutionary algorithms. *arXiv preprint arXiv:1512.01362*.

Lemel, Y. (1976). Une généralisation de la méthode du quotient pour le redressement des enquêtes par sondage. *Annales de l'INSÉÉ*, (22/23):273–282.

Miller, D. and Young, L. (2015). Imputation at the National Agricultural Statistics Service. Number 14 in Conference for European Statisticians, Work Session on Statistical Data Editing, Budapest, Hungary. United Nations Statistical Commission and Economic Commission for Europe.

Nesterov, Y. (2012). Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate O(1/kˆ 2). In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547.

Ni, D., Leonard, J. D., Guin, A., and Feng, C. (2005). Multiple imputation scheme for overcoming the missing values and variability issues in its data. *Journal of transportation engineering*, 131(12):931–938.

Nihan, N. L. (1997). Aid to determining freeway metering rates and detecting loop errors. *Journal of Transportation Engineering*, 123(6):454–458.

O'Donoghue, E., Hoppe, R. A., Banker, D., and Korb, P. (2009). Exploring alternative farm definitions: implications for agricultural statistics and program eligibility. *Economic Information Bulletin-USDA Economic Research Service*, 49.

Sartore, L., Toppin, K., Young, L., and Spiegelman, C. (2018). Developing integer calibration weights for Census of Agriculture. *Journal of Agricultural, Biological and Environmental Statistics*, Accepted.

Shaffer, B., Cheng, Y., and Slud, E. (2014). Single-stage generalized raking application in the american housing survey.

Singh, A. and Mohl, C. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22(2):107–115.

Slud, E., Grieves, C., and Rottach, R. (2013). Single stage generalized raking weight adjustment in the current population survey. In *Proceedings of JSM 2013*, Alexandria, VA. American Statistical Association.

Slud, E. V. and Thibaudeau, Y. (2010). Simultaneous calibration and nonresponse adjustment. *Statistics*, page 03.

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.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

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.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.