Imputing Missing Attitudinal Data on National Identity and Immigrant Cultures by Building a Neural Network

Melissa Lopez Reyes[1]*, Marcus Joseph L. Reyes[2]
[1] *De La Salle University, Department of Psychology*
[2] *University of the Philippines Diliman, Electrical and Electronics Engineering Institute*
**Corresponding Author: melissa.reyes@dlsu.edu.ph*

Abstract: Data sets from large-scale, multicountry surveys on social issues typically have missing data. Various imputation methods have been used to assign values to missing data in order to avoid biased parameter estimates. A common imputation approach is to predict the missing data from known relevant or associated information. In the case of a survey containing associated scales, values imputed to missing data for a scale item are predicted using the other items in the scale, as well as items in associated scales. Typical imputation methods are closed form based on a theoretical data distribution that may not correspond to the empirical data distribution. Moreover, because missing data are unknown, there are no available procedures for testing the accuracy of the imputed data. This paper demonstrates data imputation by building a neural network. Data imputations were done on items pertaining to views on national identity and immigrant cultures, using the National Identity III module of the International Social Survey Programme. The conceptual foundation of the procedure is described. Although the accuracy of imputed values against the test data was low, their correlations were from medium to large and there were more smaller than larger discrepancies. Moreover, intercorrelations of scales using imputed data follow the same pattern as those of the original data. Results indicate the challenge of data imputation with complex social phenomena.

Key Words: attitude toward immigrant cultures; back propagation; International Social Survey Programme; national identity; neural networks

1. INTRODUCTION

Occurrences of missing data have been a common problem in social science surveys and the problem is compounded when there is evidence that the missing data are not missing at random (Donders, Van Der Heijden, Stijnen, & Moons, 2006). Because discarding missing data in statistical analyses would lead to biased estimates, various methods have been designed to replace, or impute, missing data. Several methods have gone beyond the simpler methods of replacing missing values with the mean or with the last occurring value (Gelman & Hill, 2006). A general logic for imputation methods is to use information from related information, for example, by using multiple regression to predict the missing data from some reasonably chosen set of predictors (Allison, 2000), or by examining observed data patterns to finding the values that mimic these data patterns (Andridge, & Little 2010).

An added consideration is when missing data occur across several variables. The multivariate normal distribution has been used in regression models for imputation (Gelman & Hill, 2006). Another development is to use several imputed values (multiple imputation; Honaker, King, & Blackwell, 2011; Lall, 2016). The multiple imputation by chained equations (MICE; Plumpton, Morris, Hughes, & White, 2016) involves both a multivariate distribution and multiple imputation when imputing data for items in multiple scales; in this case, imputed values on a scale item is predicted using data from other items in the scale and in other scales.

1.1 Data Imputation for a Multicountry Data Set

This paper demonstrates imputation with a portion of the 34-country data set of the National Identity III module of the International Social Survey Programme (ISSP Research Group, 2013; Ganzeboon, 2017). Data imputations were done on survey items concerning citizens' notions of national identity and of immigrant cultures.

One notion of national identity is the ethnocultural notion, wherein language, religion, and ancestry are seen as defining characteristics of belonging to a nation. The ethnocultural notion is associated with antipathy toward immigrants, believed to be rooted in the dominant group's feelings of threat from the 'other', or, in some cases, in the desire to homogenize national identity. The ethnocultural notion does not lend itself easily to considerations of inclusion and openness to diversity. It is important to examine more inclusive notions of national identity, such as the legitimation notion (having citizenship, living in the country for long, being born in the country) and the civic notion (respecting institutions and laws, feeling of belonging to the nation). Citizens who subscribe to these more inclusive notions of national identity tend to have a positive attitude toward immigrant cultures. These data are useful in showing the extent to which inclusive valuing of national identity predicts valuing of immigrant cultures more so in some countries than in others. Cross-country analyses are needed to examine both the universality and specificity of sociocultural variables that promote diversity and inclusion.

1.2 Building a Neural Network for Data Imputation

To impute a value to a missing datum on an item, for example, the ancestry item of the ethnocultural notion, we built a neural network using as input data the responses to that item, to the other ethnocultural items, and to items of the other scales (i.e., legitimation and civic notions and receptivity to immigrant cultures). The output data generated from the neural network were the imputed values to replace the missing data on the ancestry item.

The input and output data are part of a neural network in so far as the generation of the output from the input data is through intermediate, interconnected layers of nodes or neurons (called hidden layers) with the first hidden layer connected to the input data and the last hidden layer connected to the output data. Travelling of information from input to output happens through a pattern of activation from a preceding layer to a subsequent layer, from each node to each subsequent node. A pattern of activation is defined by a collection of weights, where each weight determines the activation of one node coming from a preceding node.

In using a neural network for data imputation, the intent is to generate a travelling pattern of activation so that information about covariations of the item needing imputation with the rest of the items will inform the imputed values for the missing data. The pattern of activation is not predetermined, say from a fixed formula or from an a priori model, but is learned by the neural network from the training data fed into it. The training data consist of entries with nonmissing data on the item needing imputation. Through a number of iterations (epochs) the patterns passing through the neural network settles down; the final or learned pattern then generates the imputed values. Some test data are set aside (with known values on the item needing imputation) and are used to assess accuracy by comparing the imputed values to the actual values.

When can we say that a neural network has learned the existing covariations of the item needing imputation with the other items? A common algorithm is the back propagation where forward flows of activation are sent back to earlier layers, thereby monitoring the adequacy of predictions (the loss function) across epochs. Subsequent changes in patterns of activation are toward minimizing the loss function.

Many imputation methods are closed form and assume a certain theoretical data distribution that may not hold true empirically. In contrast, a neural network does not operate on some *a priori* distribution or formula but, rather, builds these from patterns of covariations in existing data.

2. MATERIAL AND METHODOLOGY

2.1 Data Set

We used data on notions of national identity and attitude toward immigrant cultures from the ISSP National Identity III module. Only data from respondents who are citizens of the country were included in the analyses (n = 44,906).

We used responses to questions on how important the following are to be truly [nationality]: (1) *language* - to be able to speak the country's dominant language(s) (2) *religion* – to be of the country's dominant religion (3) *ancestry* – to be of the country's dominant ancestry (4) *born* – to have been born in the country (5) *citizen* – to have the country's citizenship (6) *lived* – to have lived in the country for most of one's life (7) *respect* – to respect the country's political institutions and laws (8) *feeling* – to feel that one is of the country's nationality. The response options followed a 4-point format (1: not important at all, 4: very important).

We also used responses to questions about how receptive the respondent is to immigrant cultures: (1) *improved* - whether immigrants improve the country's society by bringing new ideas and cultures (2) *undermined* - whether the country's culture is generally undermined by immigrants (reverse-scored). The response options followed a 5-point format (1: strongly disagree; 5: strongly agree).

2.2 Procedure for Data Imputation

We built a neural network with the Python deep learning library Keras (https://keras.io/) to impute values to missing data on each of the abovementioned items. The input data for each of the 44,906 respondents consisted of the responses to the abovementioned 10 items and the respondent's country membership. Each item was represented by a vector of size (p+1), where p is the number of response options (with entry 1 for the response and 0 elsewhere); the last entry is 0, if there is a response, and 1, if the response is missing. Each of the 34 countries was represented by a row vector of size 34 (with entry 1 for the country of citizenship and 0 elsewhere). Three hidden layers were built in the neural network, each with 512 neurons. The output data consisted of vector of size p (with entry 1 for the imputed value and 0 otherwise).

We derived the imputed values from the response patterns of respondents who answered the item needing imputation. These data were randomly apportioned into two: 35,000 were used to learn the response patterns (called training data); the remaining were used as test data, where the imputed values were compared with the actual values.

The neural network was trained to predict the response to an item using the training data. We implemented the back propagation algorithm using the categorical cross entropy loss function. Training epochs were terminated when training data prediction accuracy kept improving with no accompanying improvement in test data (i.e., overfitting).

3. RESULTS AND DISCUSSION

3.1 Accuracy of Predictions

Shown in Table 1 are each item's percent of missing data, percent accuracy of predictions for the test data, and the correlation between the actual and predicted values for test data. The percentage prediction accuracies for test data have modest values, albeit there are medium to large correlations between the actual and predicted values.

Table 1. Prediction Accuracies

Item	% of missing data	% prediction accuracy for test data	Correlation between actual and predicted for test data*
Language	1.6	71.1	.58
Religion	3.5	57.7	.63
Ancestry	2.4	62.2	.68
Born	1.7	67.3	.69
Citizen	1.6	70.9	.61
Lived	2.1	64.5	.63
Respect	2.5	64.9	.48
Feeling	2.1	70.0	.54
Improved	4.8	47.0	.42
Undermined	5.3	44.7	.43

* Spearman rank correlation coefficient

We considered training and test data that yielded inaccurate predictions. For these data, there were more smaller than larger discrepancies (see Table 2).

Table 2. Absolute Discrepancies between Actual and Inaccurately Predicted Training and Test Data

Item	% of data with such absolute discrepancy*			
	.5 to 1	1.5 & 2	3	4
Language	82.2	13.8	4.0	
Religion	66.7	22.92	10.4	
Ancestry	79.4	16.5	4.0	
Born	81.3	15.0	3.7	
Citizen	85.5	11.4	3.1	
Lived	84.8	12.9	2.3	
Respect	81.5	13.6	3.9	
Feeling	83.8	12.4	3.8	
Improved	68.9	23.7	5.4	1.9
Undermined	67.7	25.2	5.7	2.3

*Discrepancies for notions of identity take values of .5, 1, 1.5, 2, and 3. Discrepancies for immigrants' culture take values of 1, 2, 3, and 4

3.2 Correlations Among Items Within Scales

Items on national identity can be grouped into scales. As earlier mentioned, the ethnocultural notion includes the *language*, *religion*, and *ancestry* items. The legitimation notion includes the *born*, *citizen*, and *lived* items. The civic notion includes the *respect* and *feeling* items. The *improved* and *undermined* items are attitudes toward immigrant cultures and form one scale. As shown in Table 3, item intercorrelations within a scale given the data with imputed values are similar in magnitudes given the data with missing values deleted (see Table 3).

Table 3. Correlations among Items within Each Cluster

Item	1	2
Ethnocultural notion		
1 Language		
2 Religion	.20[a] / .09[b]	
3 Ancestry	.26 / .13	.52 / .51
Legitimation notion		
1 Born		
2 Citizen	.55 / .41	
3 Lived	.55 / .49	.50 / .52
Civic notion		
1 Respect		
2 Feeling	.34 / .48	
Immigrant cultures		
1 Improved		
2 Undermined	.31 / .25	

[a] computed from data with imputed values

[b] computed from data with at least one item missing

4. CONCLUSIONS

Many imputation methods are closed form and assume a certain theoretical data distribution that may not hold true empirically. In contrast, a neural network does not operate on a fixed formula or on some *a priori* distribution, but arrives at imputed values based on patterns of covariations obtained from existing data.

This paper shows the generation of imputed data by building a neural network. Data imputations were done on items from the National Identity III module of the International Social Survey Programme that pertain to respondents' valuing of national identity and of immigrant cultures.

Although the accuracy of imputed values against the test data was low, their correlation were from medium to large and there were more smaller than larger discrepancies. Moreover, intercorrelations of scales using imputed data follow the same pattern as those of the original data. Results indicate the challenge of data imputation with complex social phenomena.

Subsequent analyses of the use of neural works in imputing multicountry data may include examining whether patterns specific to some countries are picked up by the network, thus, leading to differentiated activation patterns across countries.

5. REFERENCES

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78* (1), 40-64.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59* (10), 1087-1091.

Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, *56* (10), 968-976.

Ganzeboom, H. (2017). *International Social Survey Programme: National Identity III - ISSP 2013 (Netherlands)*. Cologne, Germany: GESIS Data Archive, ZA5517 data file version 1.0.0. doi: 10.4232/1.12921

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45* (7), 1-47.

ISSP Research Group (2013). *International Social Survey Programme: National Identity III – ISSP 2013*. Cologne, Germany: GESIS Data Archive. ZA5950 data file version 2.0.0. doi: 10.4232/1.12312

Keras: The Python deep learning library. Retrieved from https://keras.io/

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, *24* (4), 414-433.

Plumpton, C. O., Morris, T., Hughes, D. A., & White, I. R. (2016). Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC Research Notes*, *9* (1), 45.