Applying multiple imputation on waterbird census data

T. Onkelinx[a], **K. Devos**[b] and **P. Quataert**[a]

[a]Biometrics & Quality Assurance
Research Institute for Nature and Forest (INBO)
Anderlecht, Belgium
Thierry.Onkelinx@inbo.be

[b]Species Diversity
Research Institute for Nature and Forest (INBO)
Anderlecht, Belgium

Keywords: missing data; citizen sciences; monitoring of biodiversity; survey design and analysis

Abstract: The bulk of waterbird censuses are conducted by volunteers. Shifts in availability of volunteers leads to missing data. Missing data in waterbird censuses are commonly imputed using the method described by Underhill and Prys-Jones (1994). This method models the raw counts in terms of several covariates like site, month, year,... The missing data are imputed by the predictions of this model. Finally, the augmented dataset is analyzed.

Since this method uses the predicted values to impute to missing data, the variances of the parameters of the final analysis are likely to be underestimated. Two key factors play a role in this. Firstly, the apparent number of observation in the final analysis is higher than the true number of observations. Secondly, predicted values are less variable than raw counts. Therefore the imputation adds data with much smaller variation. Furthermore, the method potentially biases the parameters estimates, depending on the starting values of the imputation.

We compare this imputation method with a model-based multiple imputation (Rubin, 1987). A missing value is imputed by random number based on the distribution of the prediction for this value rather than the predicted value itself. The variance of the augmented dataset will increase depending on the uncertainty of the imputations and is most likely higher than when a complete dataset would have been available. Whereas imputing predicted values would result in a lower variance compared to a complete dataset. Next, we apply the same analysis to the augmented dataset. The results will off course depend on the randomly imputed numbers. We accommodate this by repeating the imputation and analysis process several times. The parameters of interest are averaged over the different imputations. Their standard errors are based on the standard errors of the individual imputations and the variance among the parameter estimates.

References

Rubin, D.B. (1987) Multiple imputation for Nonresponse in Surveys. *John Wiley & Sons, New York*

Underhill, L.G. and Prys-Jones, R.P. (1994) Index Numbers for Waterbird Populations. I. Review and Methodology. *Journal of Applied Ecology*, 31:463-480.