Handling missing observations with multiple imputation

BirdNumbers 2016, Halle, Germany

Thierry Onkelinx, Koen Devos & Paul Quataert
Research Institute for Nature and Forest, Brussels, Belgium

Flanders
State of
the Art

www.INBO.be

Introduction

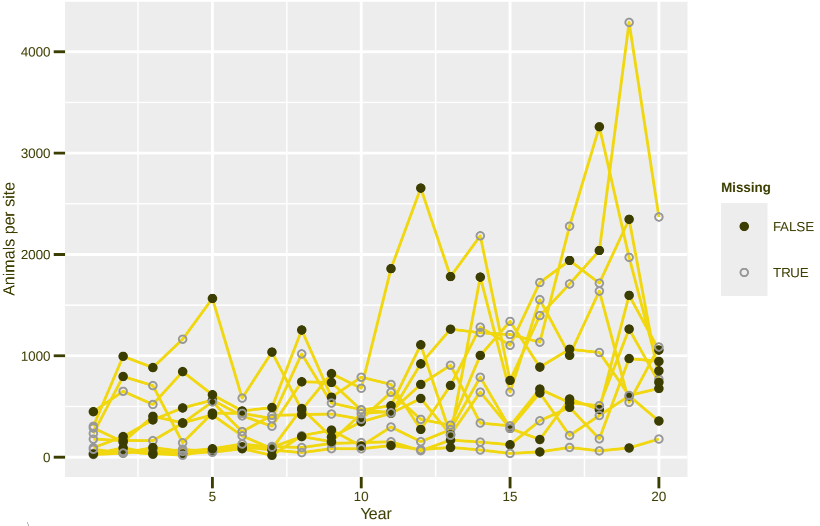
www.INBO.be

Handling missing observations

The best solution to handle missing data is to have none.
– Sir Ronald Aylmer Fisher

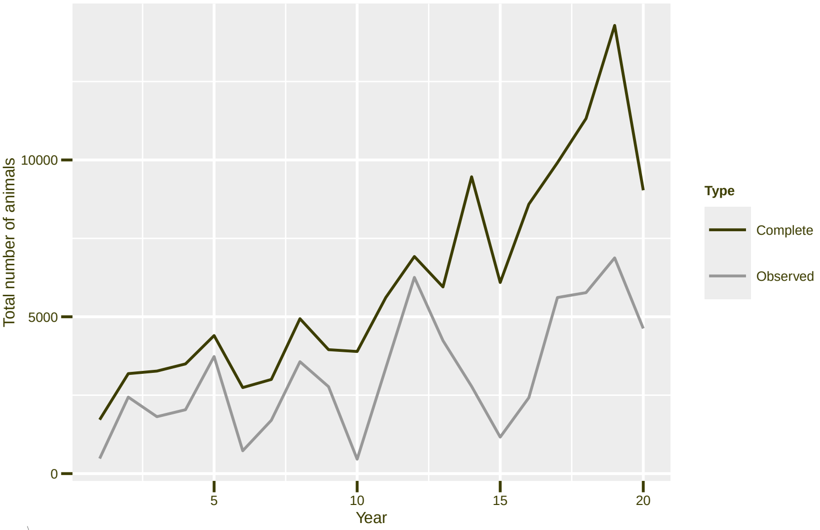
- In practice we can only try to minimise the missingness
- An increase in missingness will lead to a decrease in power
- Analysis can handle missing data (e.g. average number of animals)
  - No need for imputation
- Analysis cannot handle missing data (e.g. population totals)
  - Imputation is required

Number of animals per site



Flanders
State of the Art

Population totals



www.INBO.be

Flanders
State of the Art

Some imputation methods

- Popular in ecology for analysis of population trends
  - Underhill index, 118 citations (Underhill & Prys-Jones, 1994)
  - TRIM, 310 citations (Pannekoek & Van Strien, 2005)
  - birdSTATs, Access shell around TRIM (Meij, 2013)
  - All are **single** imputation methods
- Popular in medical and social science
  - **Multiple** imputation, 9625 citations (Rubin, 1987)
  - Only emerging in field of ecology

Flanders
State of the Art

Single imputation versus multiple imputation

The similarities

- Replace missing values with imputed values
- Imputed values are based on a model
  - The model can be very basic
    - A constant
    - The overal mean
  - The model can be elaborate
    - Use available covariates (e.g. year, season, site, climate, ...)
    - Use correlation structures (e.g. temporal, spatial, ...)
    - Use a relevant distribution (e.g. Poisson, negative binomial, ...)
    - Use zero-inflation
- Final analysis on the augmented dataset

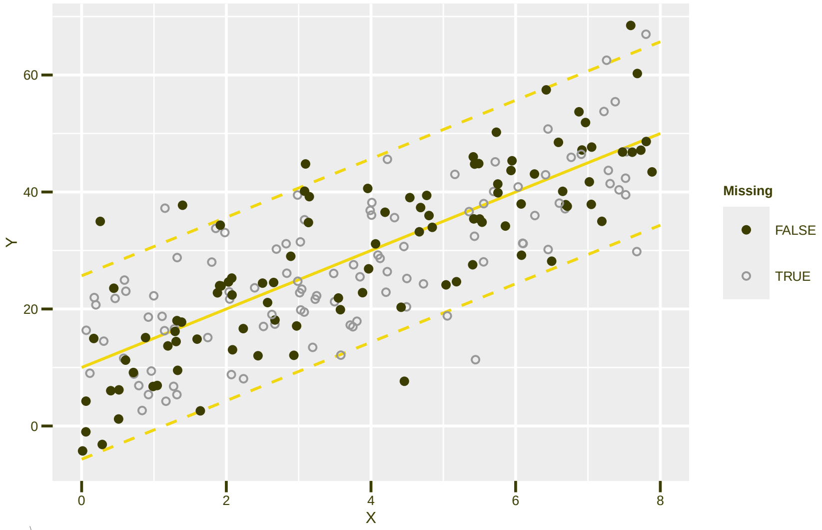
The differences

- Single imputation replaces missing values **only once**
  - It uses the best available single value: the predicted value of the model
  - Single imputation ignores *model uncertainty* and *natural variability*
- Multiple imputation replaces missing values **several times**
  - It uses each time a different random value
  - Based on
    - The distribution of predicted values of the model
    - The noise of the model
  - Multiple imputation takes both *model uncertainty* and *natural variability* into account

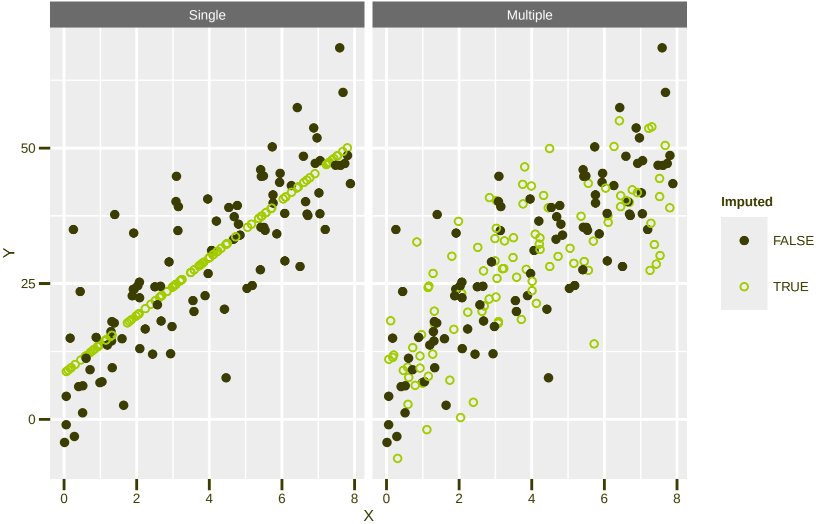
www.INBO.be

Flanders
State of the Art

Example dataset



Example of one imputation set



www.INBO.be

Flanders
State of the Art

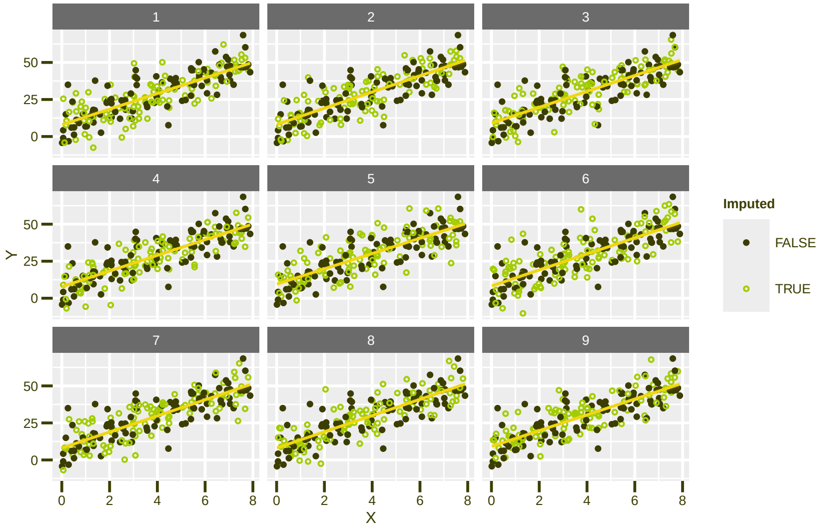
How to handle the randomness in multiple imputation?

- Since the imputed values are random, every imputation set will have different values
- Hence the results of the analysis *after* imputation will be different among imputation sets
- Solution:
  1. Create L imputation sets
  2. Run the analysis on each imputation set
  3. Average the parameter of interest B and its standard error σ_B among imputation sets using the formulas below

$$\overline{B} = \frac{1}{L}\sum_{l=1}^{L}\hat{B}_i$$

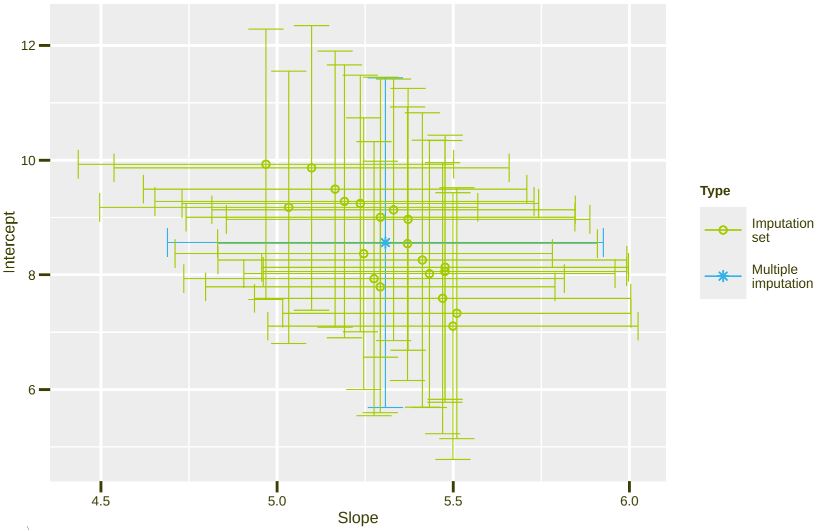
$$\overline{\sigma}_B^2 = \frac{1}{L}\sum_{l=1}^{L}\hat{\sigma}_{B_i}^2 + \frac{L+1}{L}\sum_{l=1}^{L}\frac{\hat{B}_i - \overline{B}}{L-1}$$

Example of 20 imputation sets



Flanders
State of the Art

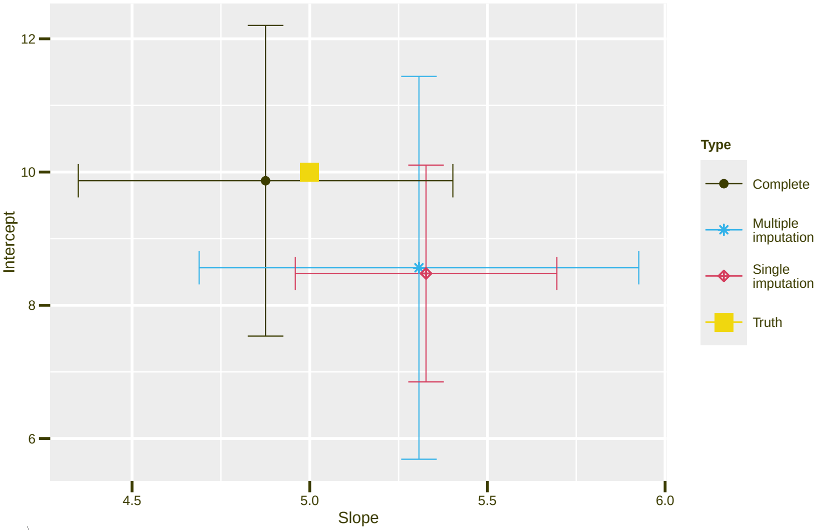
Analysis of 20 imputation sets



Flanders
State of the Art

www.INBO.be

Comparison of results



Flanders
State of the Art

www.INBO.be

Flanders
State of
the Art

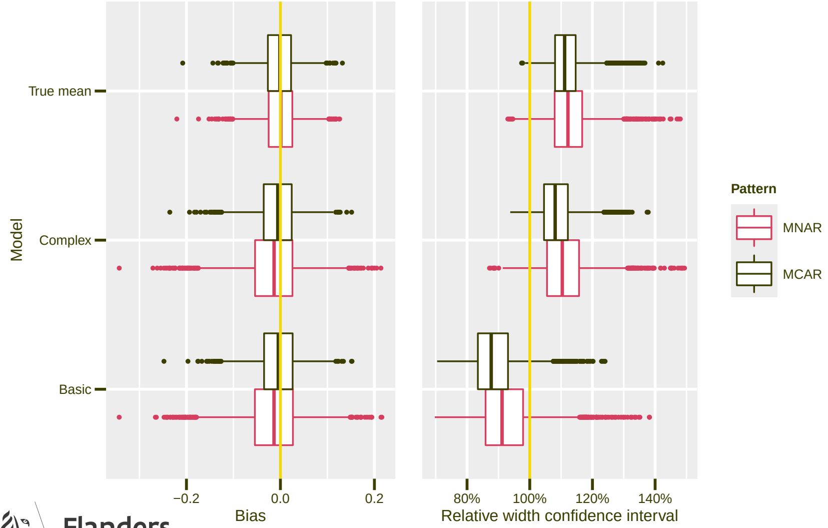
Advice on imputation

www.INBO.be

General recommendations

- Forget single imputation
  - Use **multiple** imputation
- Use a reasonable complex model
  - Too simple: model will smooth too much
  - Too complex: unstable or unreliable model
  - Use the relevant distribution!
- Number of imputations (Graham *et al.*, 2007)
  - Aim for $L = 100$ when computational effort is reasonable
  - $L = 3$ can be sufficient (<10% missing and <5% power falloff)
- Proportion of missingness
  - Multiple imputation is robust, even with 50% to 75% missing data
- Type of missingness
  - Missing not at random (MNAR) can introduce biased results

Effect of imputation model and type of missingness (Onkelinx *et al.*, in press)



www.INBO.be

Flanders
State of the Art

Available R packages

- R (R Core Team, 2013) is free and open source software for statistical computing
- Some packages for multiple imputation

Package	Counts	Mixed model	GUI	Missing covariate	Reference
multimput	X	X			Onkelinx *et al.* (2016)
Amelia			X	X	Honaker *et al.* (2011)
mice	X			X	van Buuren & Groothuis-Oudshoorn (2011)

References I

Graham J.W., Olchowski A.E., & Gilreath T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science 8(3): 206–213 doi: 10.1007/s11121-007-0070-9

Honaker J., King G., & Blackwell M. (2011). Amelia II: A program for missing data. Journal of Statistical Software 45(7): 1–47 doi: 10.18637/jss.v045.i07

Meij T. van der (2013). birdSTATs. Species Trends Analysis Tool (STAT) for European bird data

Onkelinx T., Devos K., & Quataert P. (2016). Multimput: Using multiple imputation to address missing data URL: `https://github.com/inbo/multimput` doi: 10.5281/zenodo.48423

Onkelinx T., Devos K., & Quataert P. (in press). Working with population totals in the presence of missing data. Comparing imputation methods in

References II

terms of bias and precision. Journal of Ornithology

Pannekoek J., & Van Strien A. (2005). TRIM 3 Manual (TRends & Indices for Monitoring data)

R Core Team (2013). R: A language and environment for statistical computing. Version 3.0.1 URL: `http://www.r-project.org/`

Rubin D.B. (1987). Multiple imputation for nonresponse in surveys. John Wiley; Sons, Ltd.: New York, NY.

Underhill L.G., & Prys-Jones R.P. (1994). Index numbers for waterbird populations. I. Review and methodology. Journal of Applied Ecology 31(3): 463–480 doi: 10.2307/2404443

van Buuren S., & Groothuis-Oudshoorn K. (2011). mice: Multivariate imputation by chained equations in r. Journal of Statistical Software 45(3): 1–67 doi: 10.18637/jss.v045.i03