Applying multiple imputation on waterbird census data

Comparing two imputation methods

ISEC, Montpellier, 1 july 2014

Thierry Onkelinx, Koen Devos & Paul Quataert

Thierry.Onkelinx@inbo.be

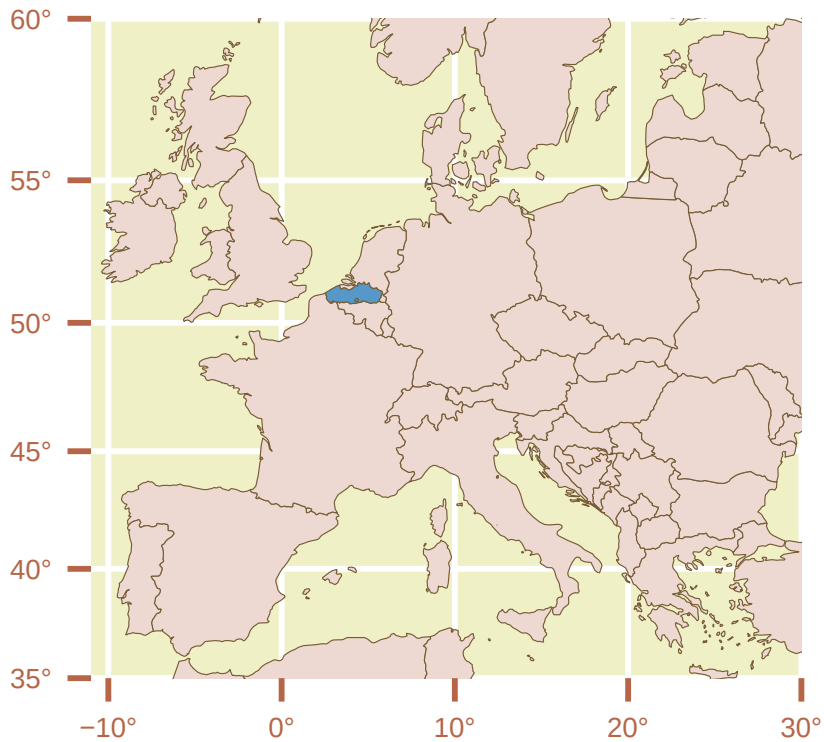
inbo

Research Institute for Nature and Forest

Contents

Waterbird census in Flanders (Belgium)

- Aim: monitor wintering birds
  - Total number
  - Average over months per winter
- Data collected by volunteers
  - 1200 sites
  - 23 winters
  - 6 months per winter
- Missing data on average 26% (9% - 42%)
  - Impution required



Underhill-index

Described by Underhill and Prys-Jones (1994)
Algorithm

1. Replace all missings with starting value
2. Fit model to imputed dataset
3. Predict missing data
4. Replace imputed value with rounded prediction when prediction is larger
5. Re-iterate from 2. until imputations are stable

Negative binomial GLM with global effects for winter, month and site

Underhill-index

Described by Underhill and Prys-Jones (1994)
Algorithm

1. Replace all missings with starting value
2. Fit model to imputed dataset
3. Predict missing data
4. Replace imputed value with rounded prediction when prediction is larger
5. Re-iterate from 2. until imputations are stable

Negative binomial GLM with global effects for winter, month and site

Potential problems

- Imputed values can never decrease: risk for **bias**
- Imputing with model predictions: risk for **reduced standard errors**

Test setup

- Generate dataset (40 sites, 24 winters, 6 months)
- Remove 25% data completely at random
- Impute missing data
- Calculate total per winter and month over all sites
- Model totals
  - Poisson regression with overdisperion
  - Estimate per winter

Test setup

- Generate dataset (40 sites, 24 winters, 6 months)
- Remove 25% data completely at random
- Impute missing data
- Calculate total per winter and month over all sites
- Model totals
  - Poisson regression with overdisperion
  - Estimate per winter
- Relative bias: $\frac{\exp(\beta_{Underhill})}{\exp(\beta_{complete})}$
- Relative SE: $\frac{\sigma_{\beta_{Underhill}}}{\sigma_{\beta_{complete}}}$

Test setup

- Generate dataset (40 sites, 24 winters, 6 months)
- Remove 25% data completely at random
- Impute missing data
- Calculate total per winter and month over all sites
- Model totals
  - Poisson regression with overdisperion
  - Estimate per winter
- Relative bias: $\frac{\exp(\beta_{Underhill})}{\exp(\beta_{complete})}$
- Relative SE: $\frac{\sigma_{\beta_{Underhill}}}{\sigma_{\beta_{complete}}}$

1. Original Underhill-index, starting value = zero
2. Original Underhill-index, starting value = geometric mean
3. Altered Underhill-index, starting value = zero
4. Altered Underhill-index, starting value = geometric mean

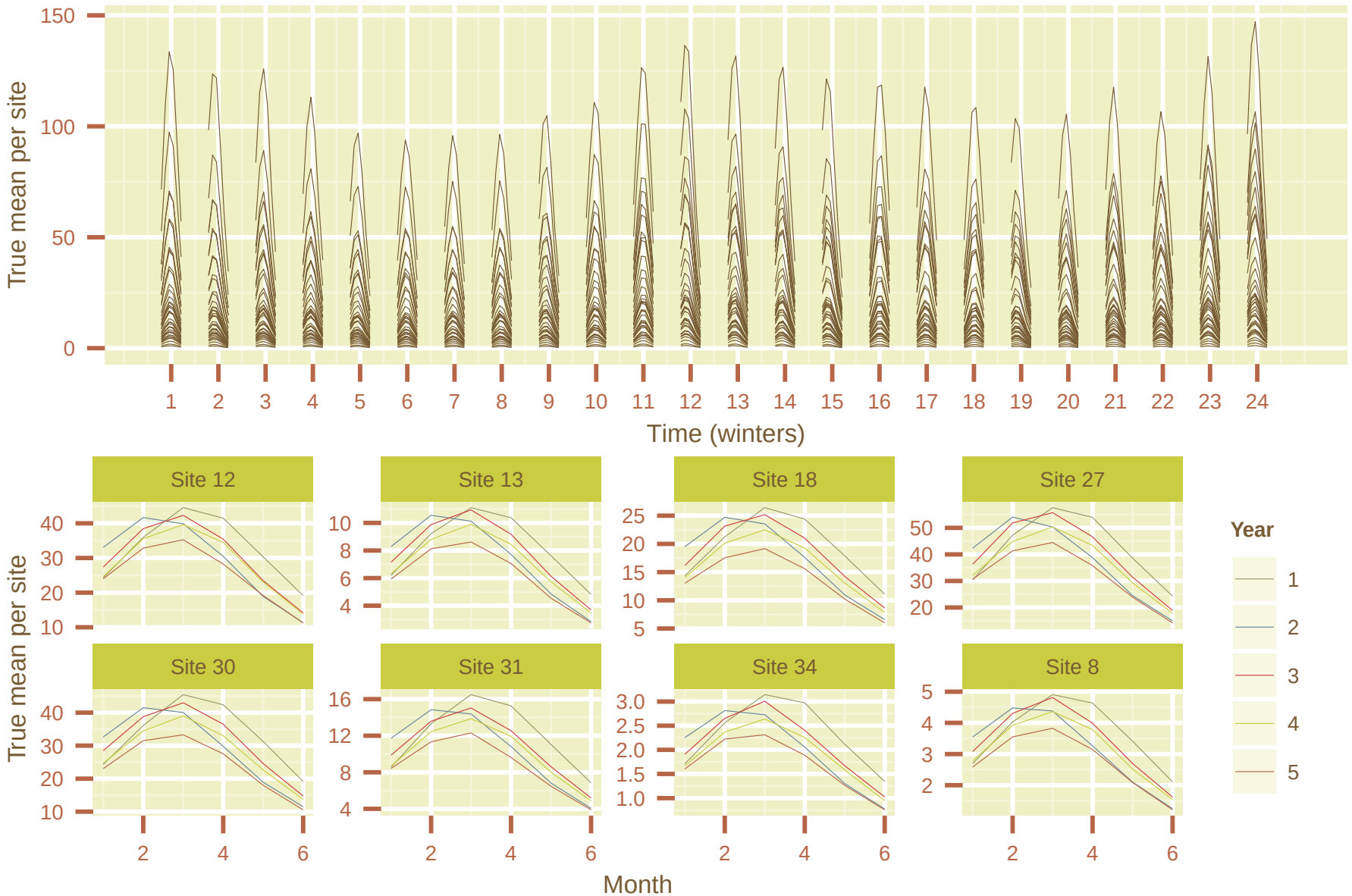
Altered index: replace imputation with rounded predictions

Data generating model

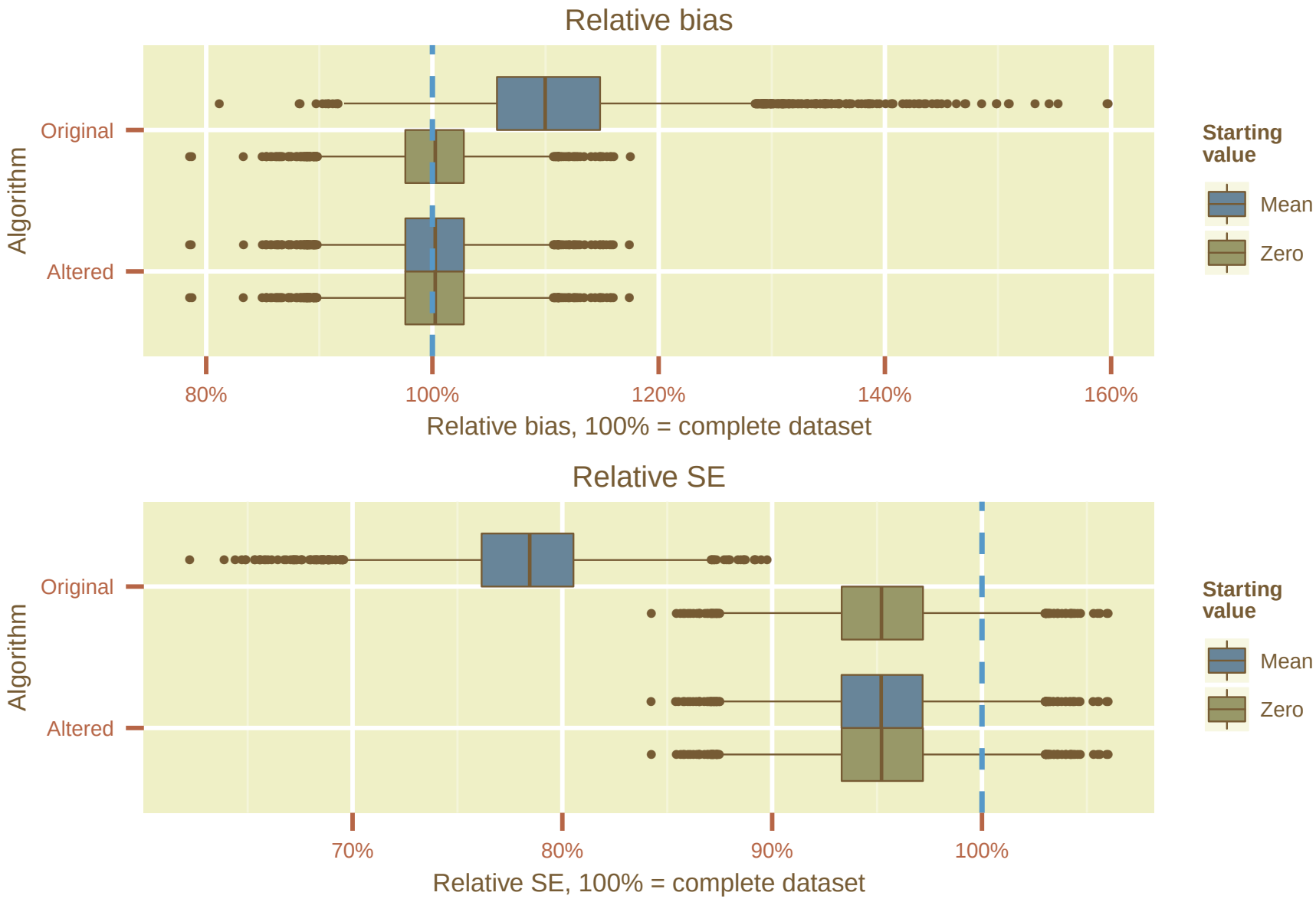
Counts follow negative binomial distribution

- Fixed size
- Variable mean (defined on log-scale)
  - Intercept
  - Linear trend and random walk along winter
  - Random intercept and random walk along winter per site
  - Sine wave within a winter with variable phase among winters
  - Gaussian noise at observation level
- Different dataset per simulation
- All datasets based on the same hyperparameters

Example of simulated dataset



Evaluation of Underhill index



Requirements

- Choose model that doesn't require starting values
  - We choose a negative binomial GLMM
    - Winter and month as fixed effect (factors)
    - Site as random intercept
    - Fitted in R (R Core Team 2014) with INLA (Rue et al. 2009)

Requirements

- Choose model that doesn't require starting values
  - We choose a negative binomial GLMM
    - Winter and month as fixed effect (factors)
    - Site as random intercept
    - Fitted in R (R Core Team 2014) with INLA (Rue et al. 2009)
- Take the uncertainty of predictions into account
  - Sample from negative binomial distribution
  - Size
    - Sample from distribution of hyperparameter
  - Mean
    - Sample from gaussian distribution
    - Mean and SE of prediction on the link scale

Test setup

- Same test datasets as for testing Underhill-index
- Fit INLA model to observed dataset
- Generate M sets of imputed values
- For each set m
  - Calculate total per winter and month over all sites
  - Model totals
  - Save the regression parameters $\beta_{i\,m}$ and their SE $\sigma_{i\,m}$

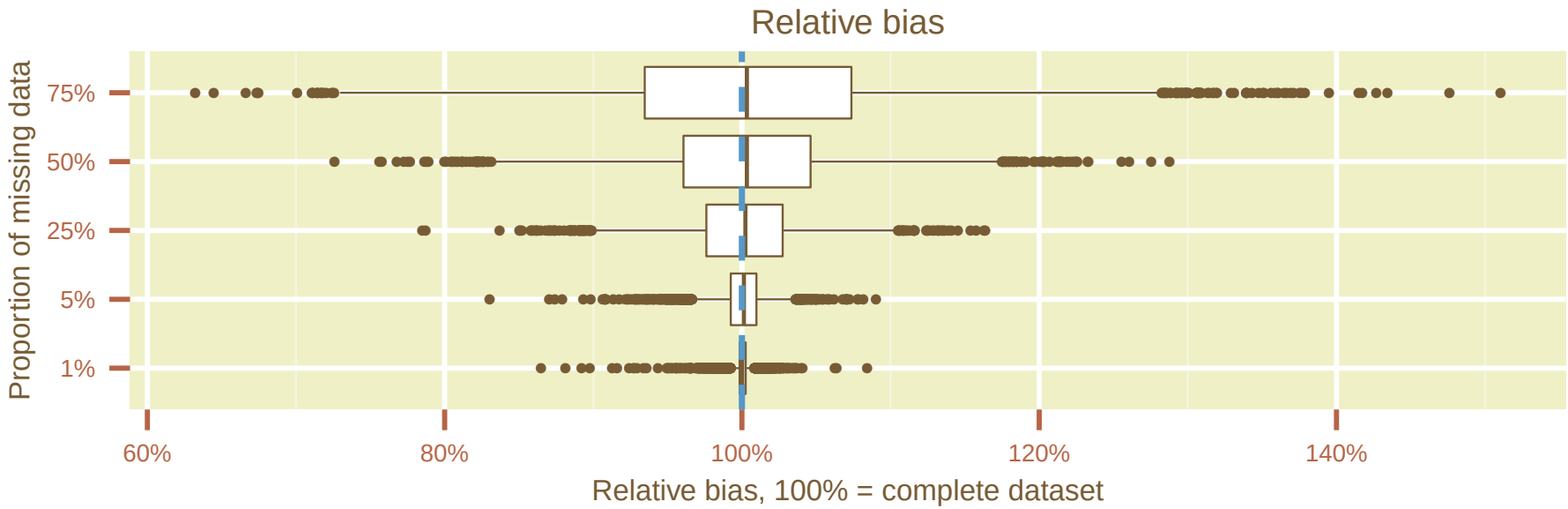
Test setup

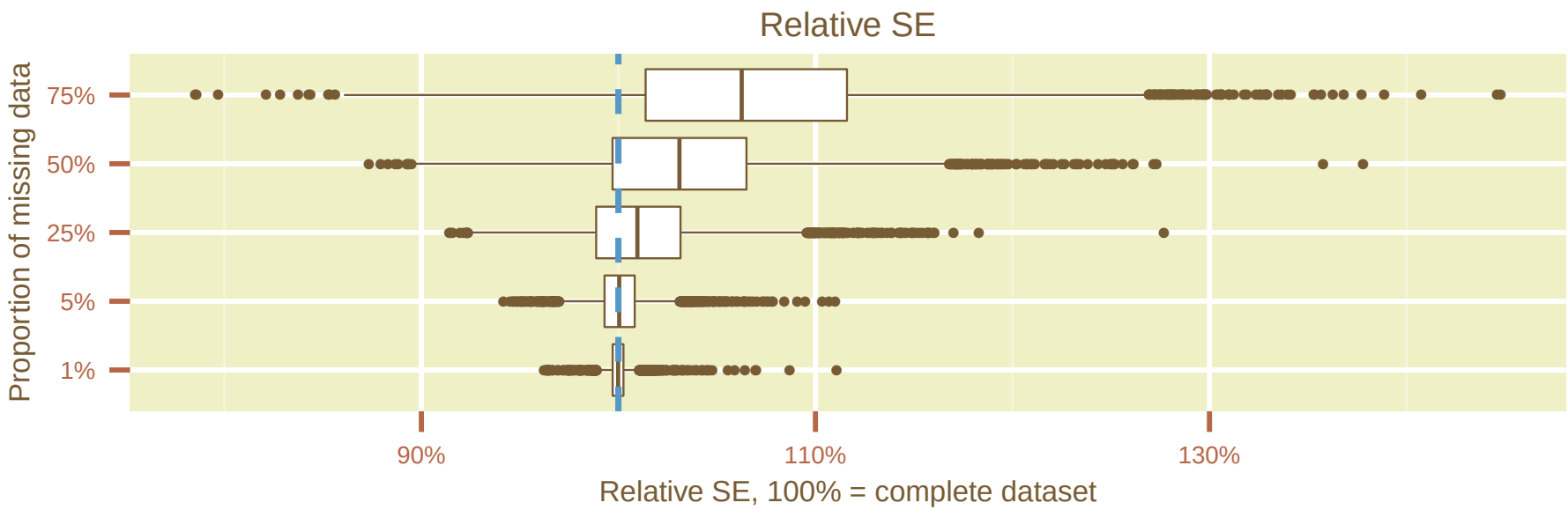
- Same test datasets as for testing Underhill-index
- Fit INLA model to observed dataset
- Generate M sets of imputed values
- For each set m
  - Calculate total per winter and month over all sites
  - Model totals
  - Save the regression parameters $\beta_{i\,m}$ and their SE $\sigma_{i\,m}$
- Aggregate over all M sets (Rubin 1987)

$$\bar{\beta}_i = \frac{1}{M} \sum_{m=1}^{M} \beta_{i\,m}$$

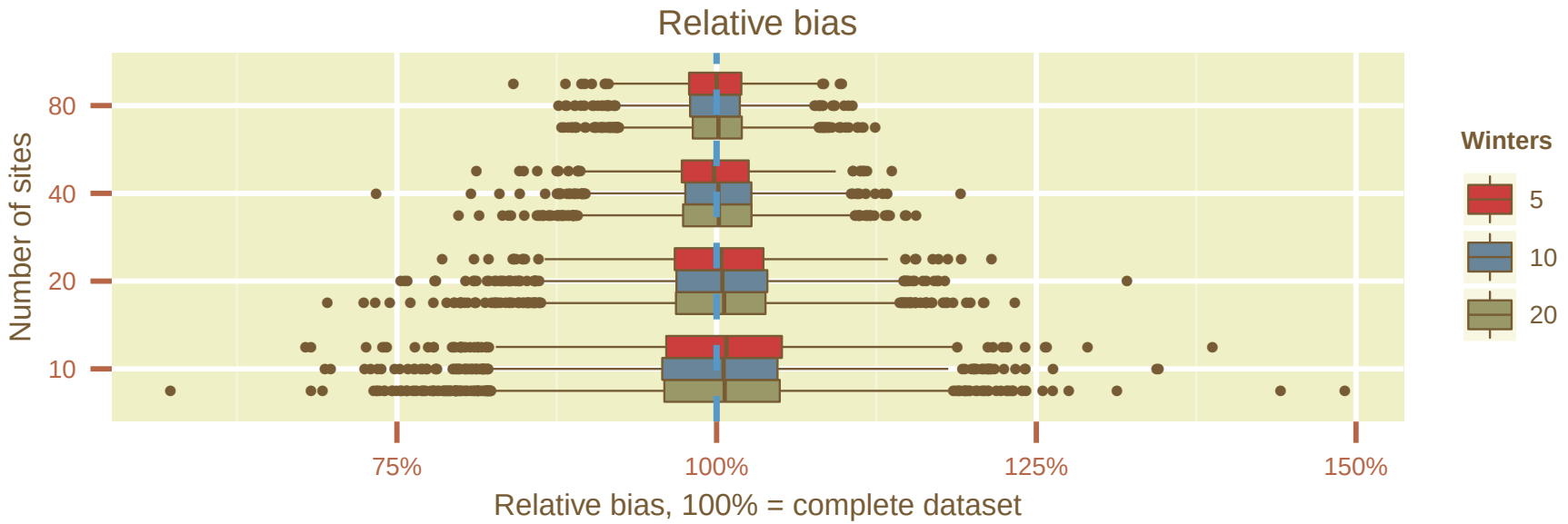
$$\bar{\sigma}_i = \sqrt{\frac{1}{M} \sum_{m=1}^{M} \sigma_{i\,m}^2 + \frac{M+1}{M} \sum_{m=1}^{M} \frac{(\beta_{i\,m} - \bar{\beta}_i)^2}{M-1}}$$

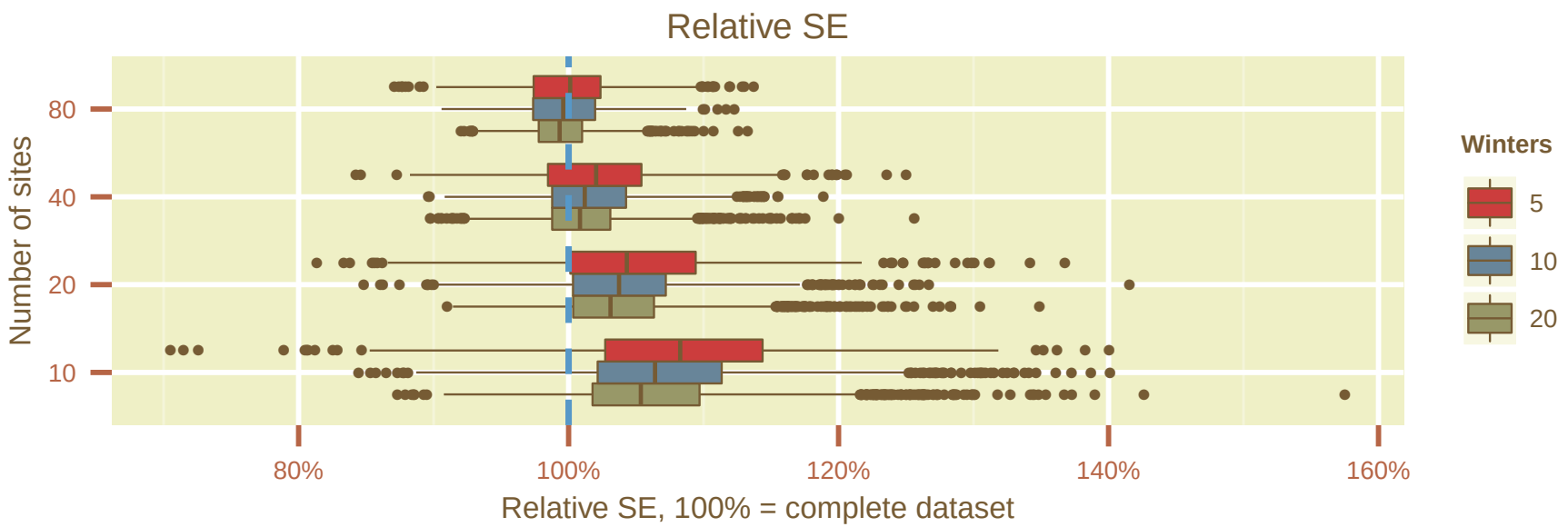
Evaluation of multiple imputation



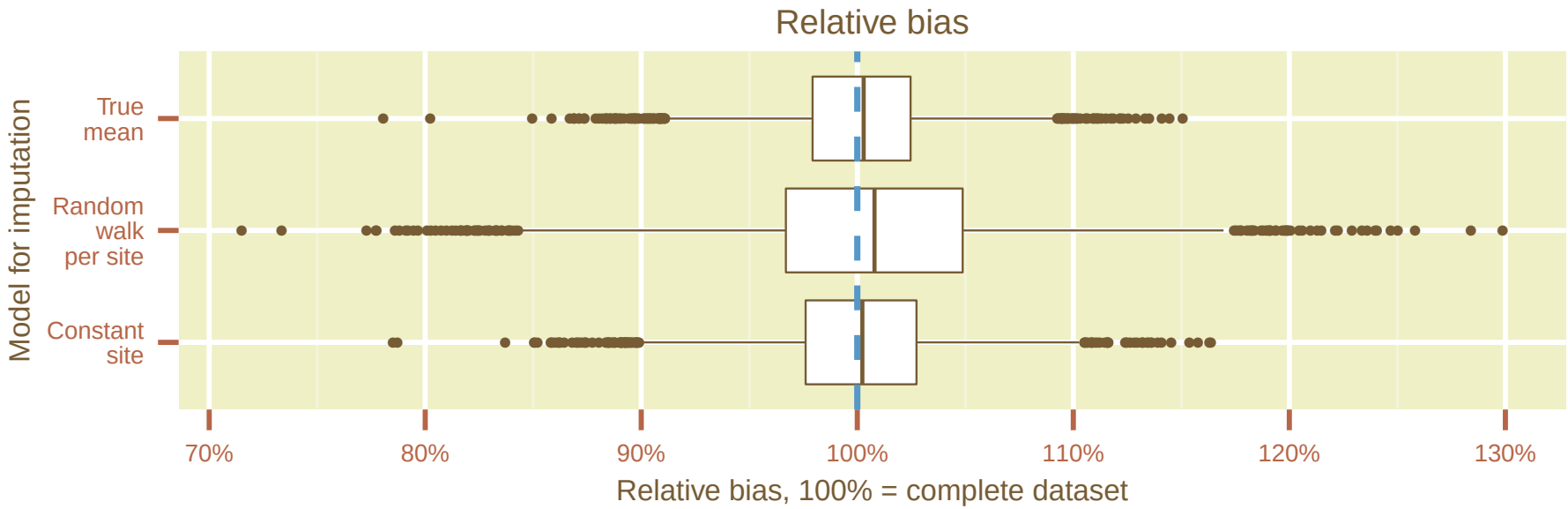


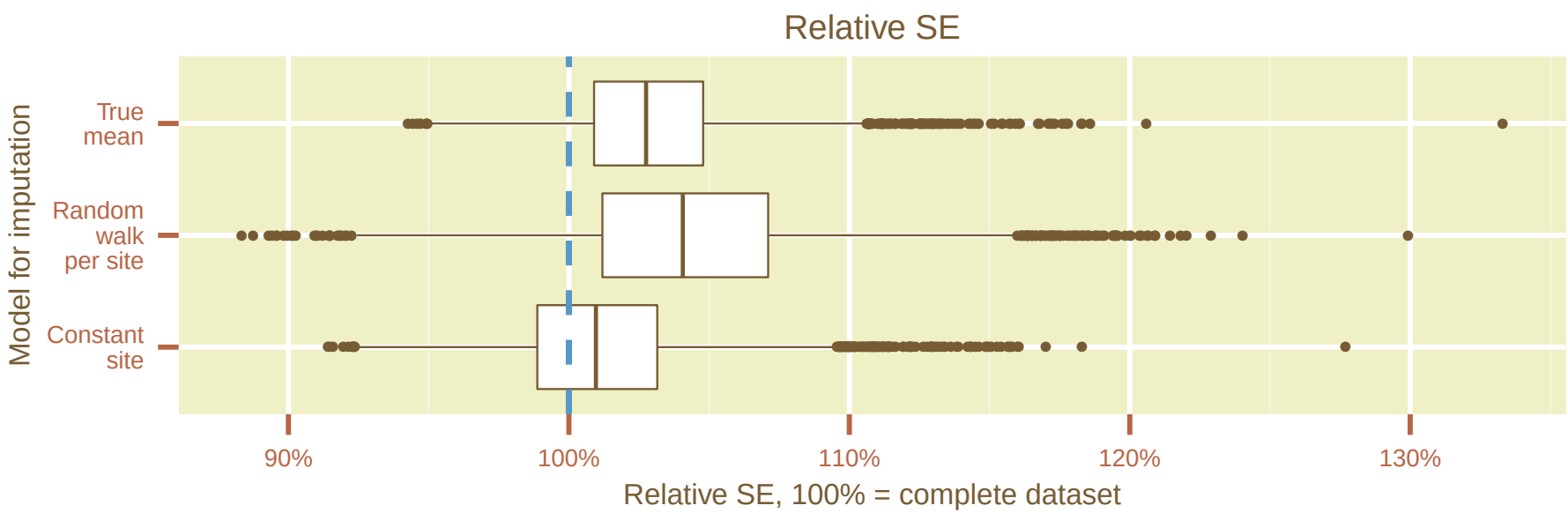
Effect of design





Effect of model for imputation





Real life examples

Underhill-index

- Must use zero as starting value
  - Otherwise biased upward

Underestimates standard errors

- **Incorrect Type I errors!**
- Too optimistic

Multiple imputation

- Unbiased estimates

Increased standard errors

- Imputation = more uncertainty
- Implies lower power
  - Sample size actual dataset $<$ sample size complete dataset

- Increase of standard error depends on
  1. **Proportion of missing data**
  2. *Size of dataset*
  3. Imputation model
- Marginal improvement with increased number of imputations

Questions?

References

R Core Team. 2014. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. `http://www.r-project.org/`.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons, Ltd.

Rue, Håvard, Sara Martino, Finn Lindgren, Daniel Simpson, and Andrea Riebler. 2009. *INLA: Functions Which Allow to Perform Full Bayesian Analysis of Latent Gaussian Models Using Integrated Nested Laplace Approximation*.

Underhill, L. G., and R. P. Prys-Jones. 1994. "Index Numbers for Waterbird Populations. I. Review and Methodology." *Journal of Applied Ecology* 31 (3): 463–80. doi:10.2307/2404443.