Multiple Imputation of Hierarchical Nonlinear Time Series Data: an Application to School Enrollment Rates

Daphne H. Liu and Adrian E. Raftery

August 8, 2024

Motivation

- In data from surveys and administrative records, can often encounter highly related measures of similar underlying concepts with differing amounts and patterns of missing data
- Secondary school enrollment rates from the UNESCO Institute for Statistics
  - Country-level estimates harmonized to be comparable across countries and times
  - Compiled from surveys and administrative records
  - Two different measures of school enrollment that have a nonlinear relationship
- How can we leverage information from one measure to help impute the other?

School Enrollment Rates

- Net Enrollment Rate (NER) =

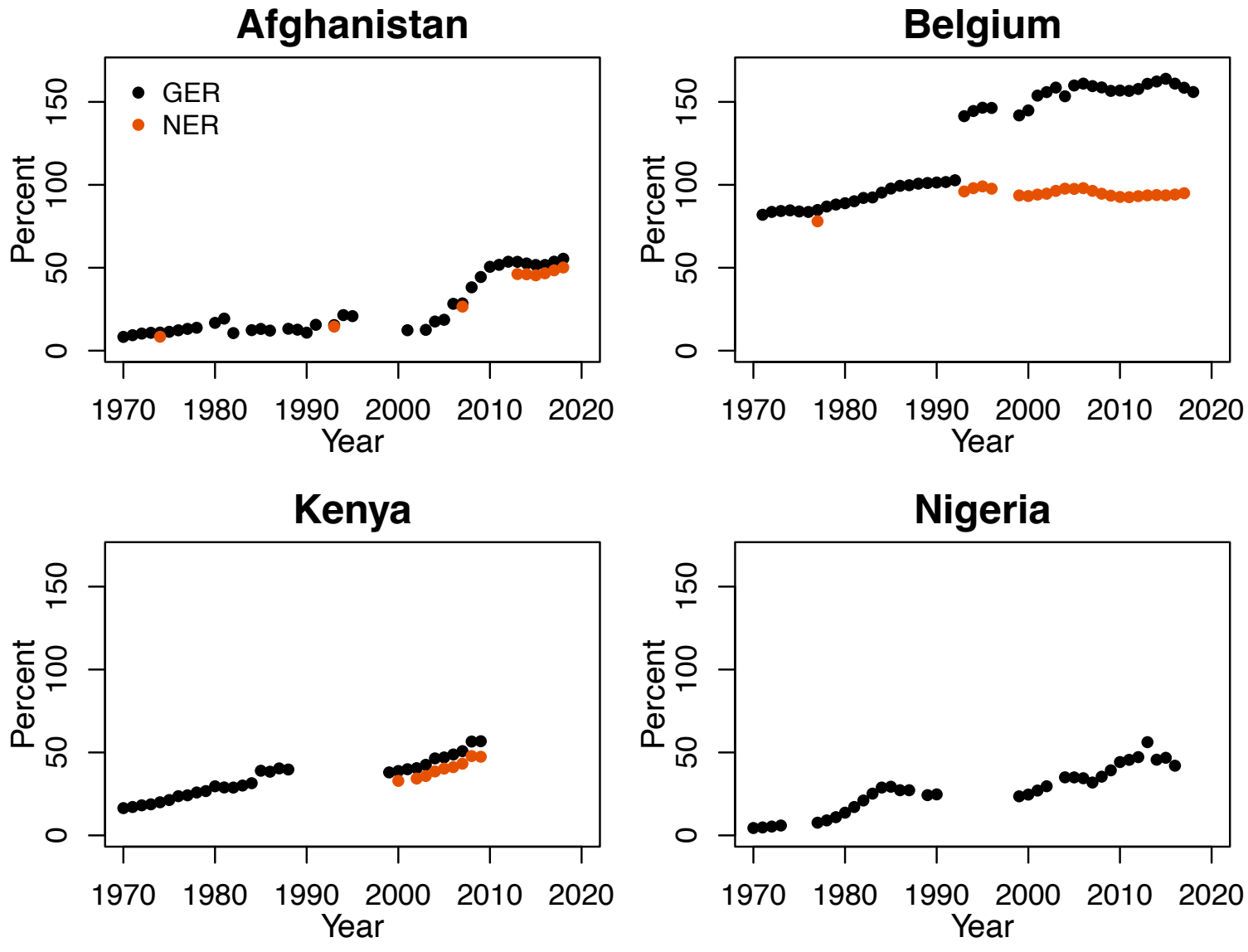
$$\frac{\text{number of children of official school age who are enrolled}}{\text{population size of official school age}}$$

- Gross Enrollment Ratio (GER) =

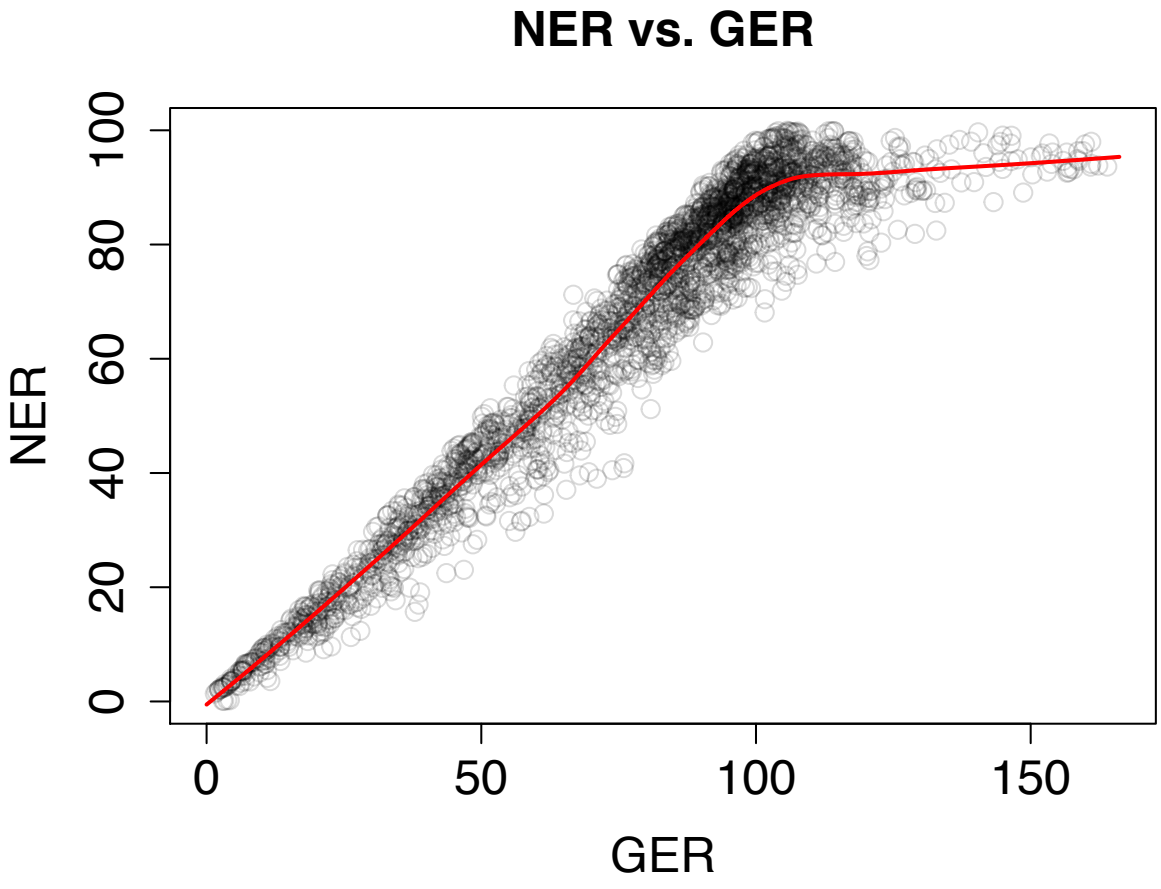
$$\frac{\text{number of children who are enrolled, regardless of age}}{\text{population size of official school age}}$$

- NER is our variable of interest for subsequent analyses, but has a larger amount of missing data than GER
- NER requires knowledge of the ages of children enrolled in school $\rightarrow$ more difficult to measure than GER

School Enrollment Rates



School Enrollment Rates



- Commonly used multiple imputation methods cannot accommodate this nonlinear relationship

Motivation

- Goal: develop a multiple imputation method for continuous hierarchical time series data that can account for nonlinear relationships between variables
  - Bivariate setting
  - One variable of interest for subsequent analyses and one auxiliary variable
- Proposed method is called MINTS for Multiple Imputation of hierarchical Nonlinear Time Series data

Notation and Assumptions

- Complete data $(\mathbf{X}, \mathbf{Y})$
  - $Y =$ variable of interest for analyses $=$ NER
  - $X =$ auxiliary variable $=$ GER
- $Y_{c,t}$ is the value of Y in country c and time t
  - $c \in 1, ..., C = 202$
  - $t \in 1, ..., T = 51$ for 1970-2020
- Assume missing data mechanism is ignorable and X is easier to measure than Y

Multiple Imputation Framework

- Sequential decomposition (Lipsitz & Ibrahim 1996)
- Joint distribution of $(\mathbf{X}, \mathbf{Y}) = (\text{GER}, \text{NER})$ is decomposed as

$$p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_Y) p(\mathbf{X}|\boldsymbol{\theta}_X)$$

- The distribution is further decomposed by time as

$$p(\mathbf{X}|\boldsymbol{\theta}_X) = \left(\prod_{t=1}^{T} p(\mathbf{X}_t|\mathbf{X}_{t-1}, \boldsymbol{\theta}_X) \right) p(\mathbf{X}_0|\boldsymbol{\theta}_X)$$

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_Y) = \left(\prod_{t=1}^{T} p(\mathbf{Y}_t|\mathbf{Y}_{t-1}, \mathbf{X}_t, \boldsymbol{\theta}_Y) \right) p(\mathbf{Y}_0|\mathbf{X}_0, \boldsymbol{\theta}_Y)$$

Model

- $\mathbf{X}_t|\mathbf{X}_{t-1}$ is modeled as a random walk with a country-specific drift term γ_c
- For country c and $t \in 1, ..., T$,

$$X_{c,t}|X_{c,t-1}, \boldsymbol{\theta}_X \sim TN_{[0,\infty)}(X_{c,t-1} + \gamma_c,\ \sigma_X^2)$$

$$\gamma_c \sim N(\mu_{drift}, \sigma_{drift}^2)$$

- Prior distributions chosen to be conjugate and diffuse

Model

- $\mathbf{Y}_t|\mathbf{Y}_{t-1}, \mathbf{X}_t$ is modeled with:
  - country-specific intercept α_c
  - nonlinear function f of $\mathbf{X}$ with coefficient β
  - AR(1) term with autoregressive parameter ρ
- Heteroscedasticity modeled as a function h of $\mathbf{X}$
- For country c and $t \in 1, ..., T$,

$$Y_{c,t}|\, Y_{c,t-1}, X_{c,t}, \boldsymbol{\theta}_Y \sim TN_{[0,\min(X_{c,t},100)]}\left(\mu_Y,\ \sigma_Y^2\, h(X_{c,t})\right)$$

$$\mu_Y = \alpha_c + \beta\, f(X_{c,t}) + \rho\, Y_{c,t-1}$$

$$\alpha_c \sim N(\mu_0, \sigma_0^2)$$

- Prior distributions chosen to be conjugate and diffuse

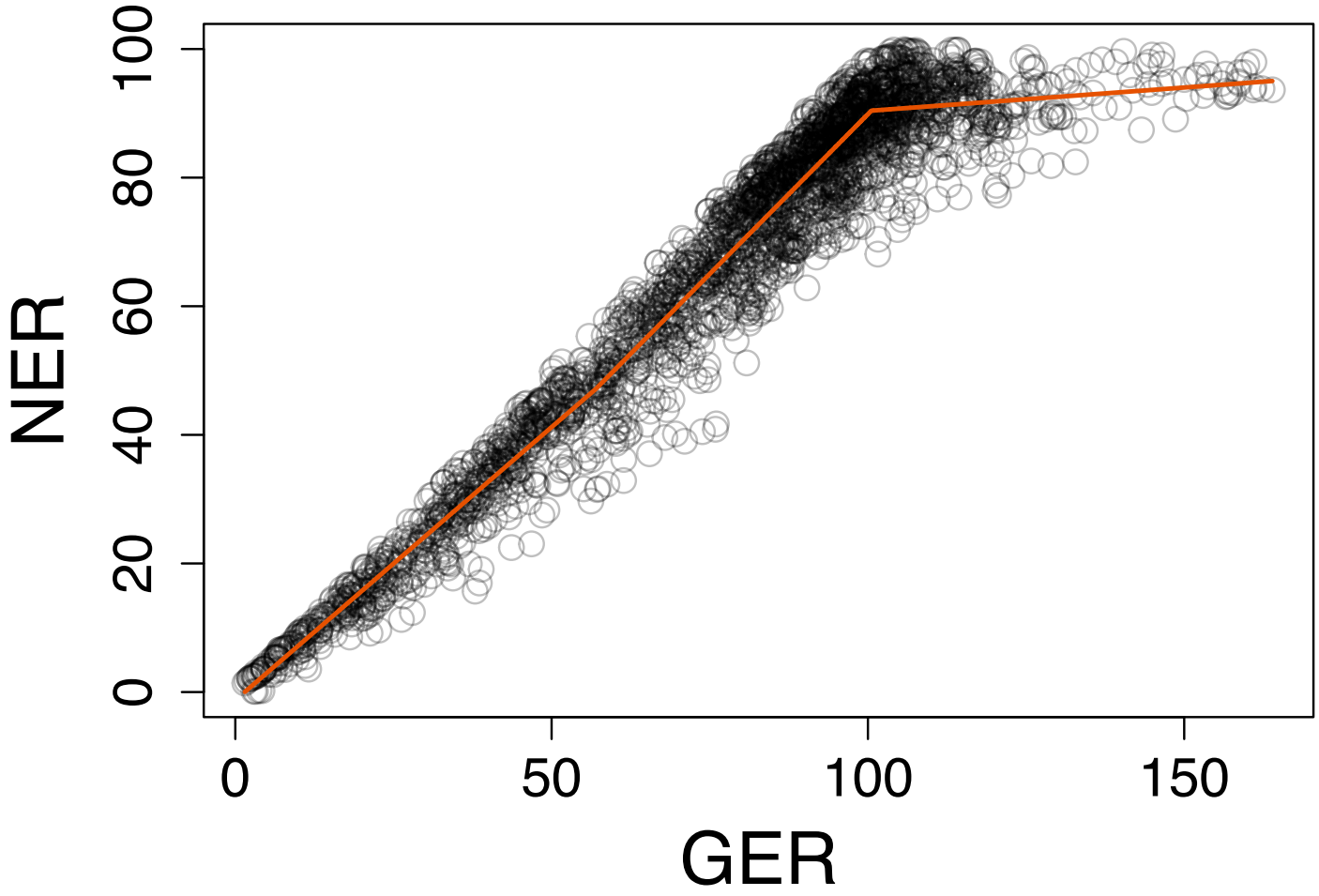
Nonlinear Functions f and h

- Estimated using adaptive splines (A-splines)
  - Goepp et al. (2018)
  - B-splines with automated selection of number and placement of knots

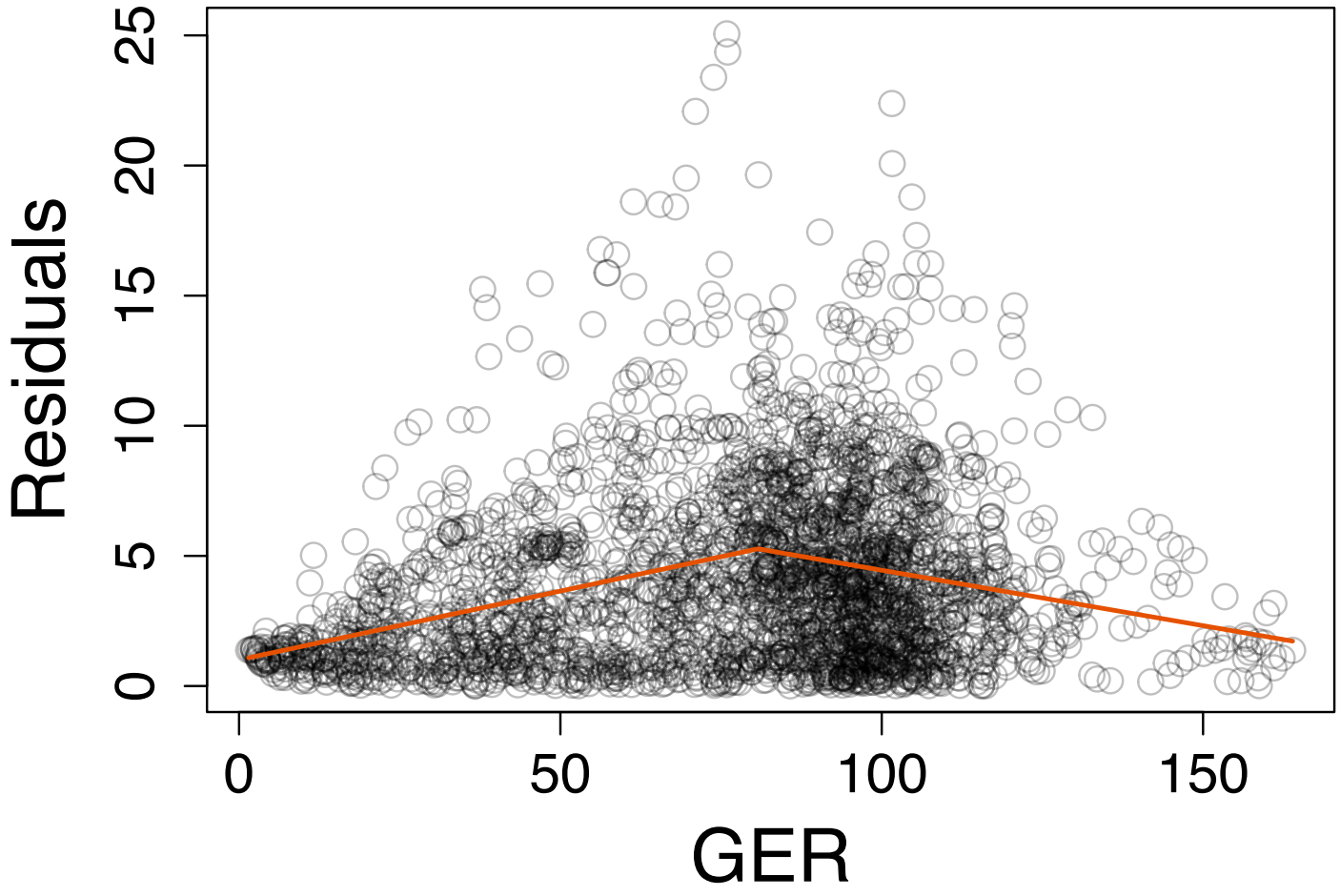
- f estimated using complete cases $(\mathbf{X}, \mathbf{Y})$ using a B-spline of degree 1

- h estimated using the residuals from the estimation of f using a B-spline of degree 1

f for Enrollment Data



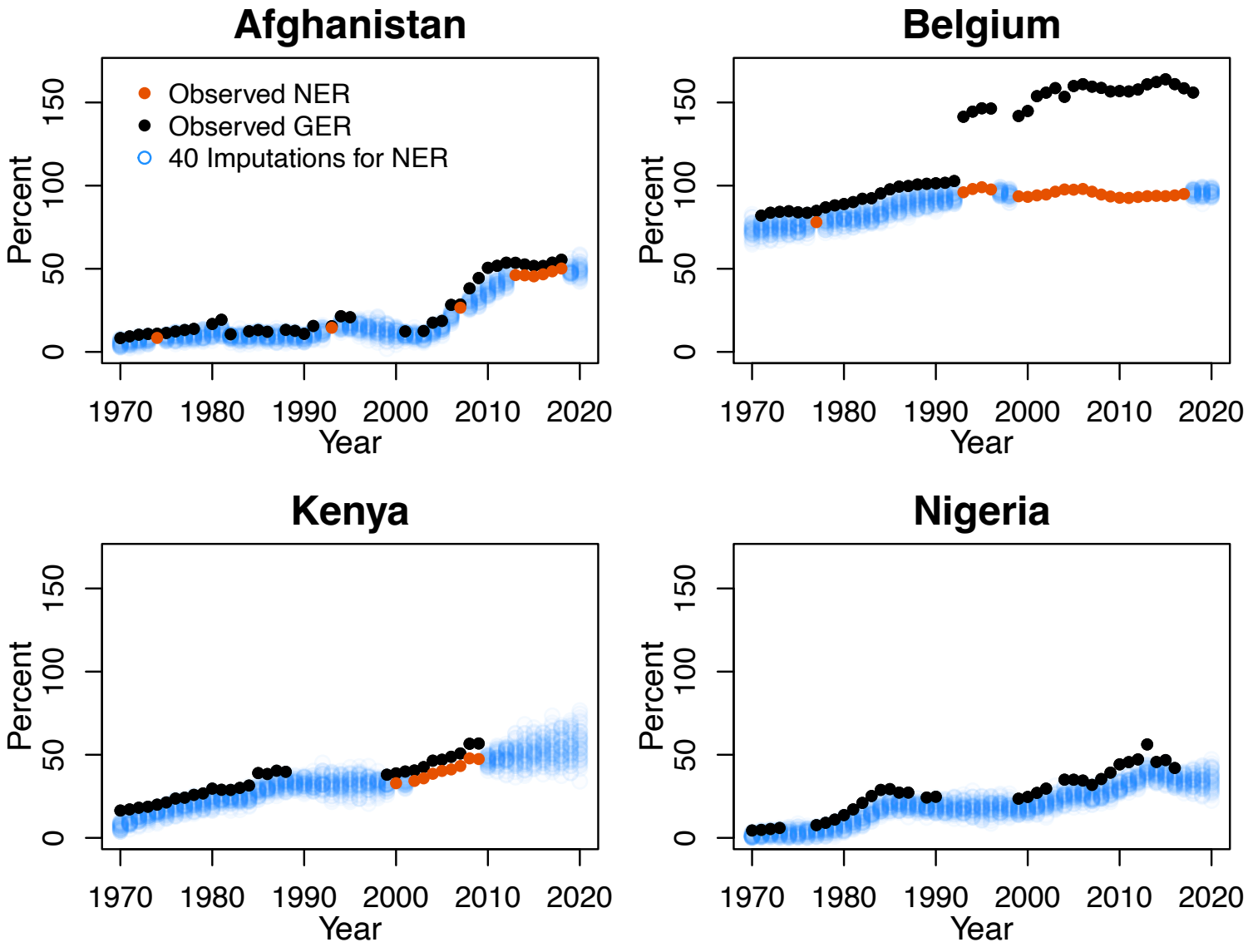
h for Enrollment Data



Model Estimation and Imputation Procedure

- Multiple imputations are created in two phases
- Estimation phase: parameters of the MINTS imputation model are estimated using a Markov chain Monte Carlo (MCMC) algorithm
  - Iterate between estimation of imputation model parameters and missing values in a similar fashion as the data augmentation algorithm of Tanner & Wong (1987)
- Imputation phase: additional iterations of the same MCMC algorithm are run and used to create multiple imputations

Imputation Results for Example Countries



Validation Exercises

- Conducted several validation exercises using simulated data and by simulating additional missing values in the school enrollment data
- Compared MINTS with existing methods for multiple imputation of hierarchical time series data
  - Models based on the Multiple Imputation by Chained Equations (MICE) methodology of van Buuren & Groothuis-Oudshorn (2011)
  - Models based on the Amelia II methodology of Honaker & King (2010)

Validation Exercises

- For multiply imputed estimation of parameters in analysis models, MINTS generally results in
  - Smaller MAE
  - Close to nominal coverage for 95% intervals
  - Smaller fraction of missing information
- For prediction of individual missing values, MINTS generally has
  - Smaller MAE
  - Best balance between coverage and width of 95% intervals

Summary

- Developed the MINTS method for multiple imputation of hierarchical nonlinear time series data
- Applied the MINTS method to a secondary school enrollment rate data set
- Through validation exercises, found MINTS can lead to substantial gains in performance compared to existing methods when variables in the imputation model have a nonlinear relationship