Machine Learning for monitoring the twin goals – Pitfalls and Solutions

Kazusa Yoshimura (World Bank)
Nobuo Yoshida (World Bank)

Dec 5/6, 2019
@UNECE

WORLD BANK GROUP

Sequence of the presentation

- Poverty projection
- Introduction to ML
- Data structure for modeling and evaluation
- Issues of ML – Large bias
- Solution – ML in MI
- Refinement – ML in MI with further variable selections
- Selection of approaches via cross-validation

WORLD BANK GROUP

Poverty Projection

WORLD BANK GROUP

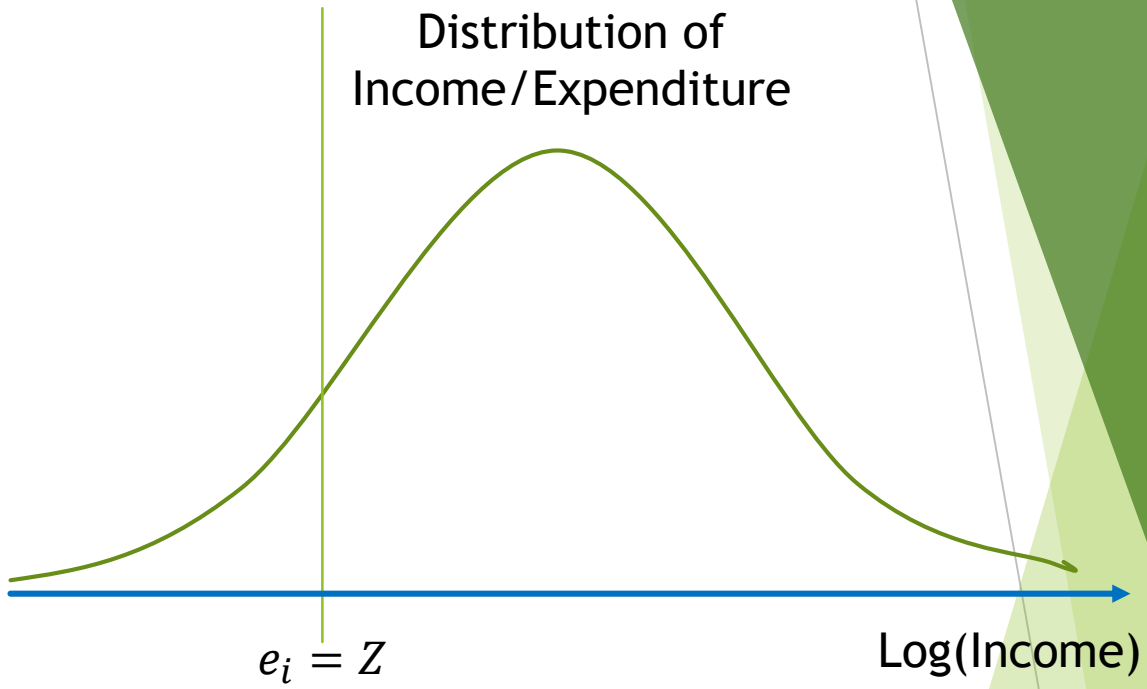
Monitoring of poverty and shared prosperity

- Data generation process:

$$e_i = X_i * \beta + u_i \text{ or } f(X_i; \theta) + u_i$$

$$D_i = 1 \; if \; e_i < \mathrm{Z} \quad (1)$$

$$D_i = 0 \; if \; e_i \geq \mathrm{Z} \quad (2)$$



- Poverty headcount rate = Average(D_i)
- Shared Prosperity index = Average(e_i) if e_i<40th percentile

WORLD BANK GROUP

Poverty projection

- Collecting household expenditure data (e_i) is very time-consuming and costly
- Instead, we try to project poverty rates by imputing household expenditures from non-expenditure data (X_i)

$$e_i = X_i\beta + u_i$$

- Using the imputed expenditures, we estimate poverty rates and shared prosperity index
- Poverty projection is known to be vulnerable to (i) overfitting and (ii) instability over time

Introduction to ML

Machine Learning (ML) for projections of poverty and shared prosperity indicators

- ML is a useful approach for predicting any indicators for two reasons:
  1. Variables needed for estimating household expenditures (X_i) are limited (10 to 15 questions) and simple (most are yes/no)
  2. It is extremely quick to transform the variables (X_i) to household expenditures (e_i) and poverty/shared prosperity indicators (in 1 – 2 minutes for 10,000+ observations)
  3. It is robust to overfitting (too good performance in-sample but bad performance out of sample)

WORLD BANK GROUP

Regularized regression Approaches (LASSO; RIDGE; Elastic Net)

- Regularized regression approaches
  - Impose the regularization on the magnitude of coefficients when minimizing the loss function
- In the normal regression, $\boldsymbol{y} = \boldsymbol{X\beta} + \epsilon$
  - The loss function is $L(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})$
  - In the regularized regression, we have a penalty term $p_\lambda(\beta)$

$$\text{Min}_\beta(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}) + p_\lambda(\boldsymbol{\beta})$$

  - Some popular forms include ridge, $p_\lambda(\beta) = \lambda\sum_{j=1}^{p}\beta_j^2$, lasso, $p_\lambda(\beta) = \lambda||\beta||$, elastic net, $p_\lambda(\beta) = \lambda_1\big|\big|\beta\big|\big| + \lambda_2||\beta||^2$
  - All these regularization functions add penalties against OLS estimators, such as λ and as the penalty becomes bigger, the regularized regression coefficients shrink toward 0

WORLD BANK GROUP

Random Forest (RF)

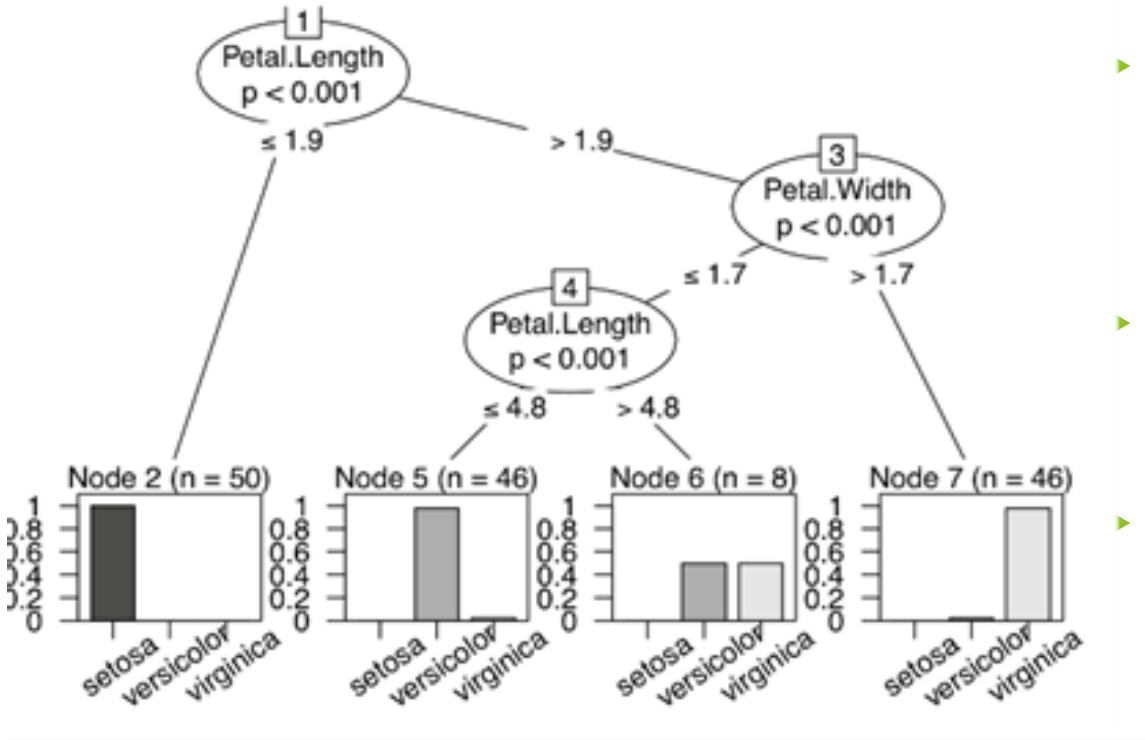


Figure 1: Example of CART

- A group of approaches under Elastic Net needs to know a functional form but in reality, we do not know it in advance
- A group of approaches under RF (Decision Tree and Bagging) can address it to a certain extent
- RF is an extended version of CART (classification and regression tree)
- RF gives us a predictor of household expenditures in a highly non-linear way:

$$\widehat{e_i} = f_{RF}\,(X_i; \hat{\theta})$$

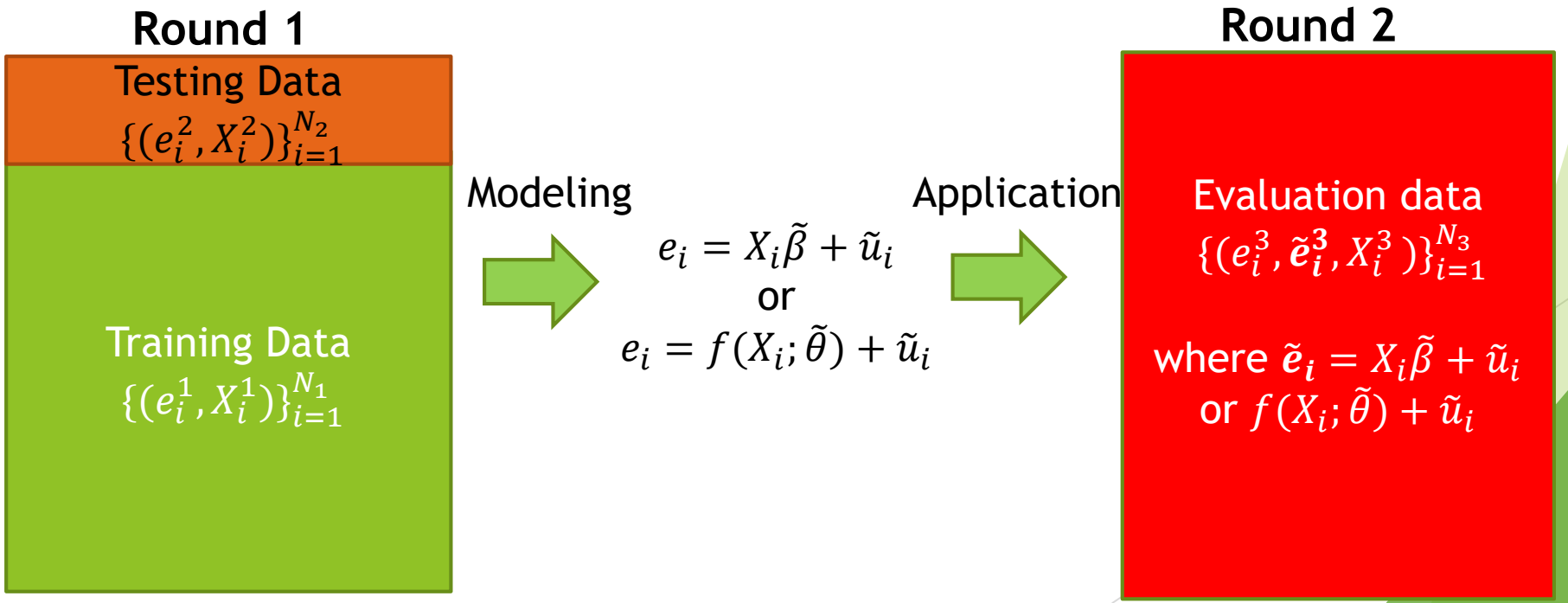
WORLD BANK GROUP

Data Structure for modeling and evaluation

WORLD BANK GROUP

Data structure for modeling and evaluation of different methodologies

- We use the data of rural Uganda, which has two rounds of data
- The data in round 1 are split between training and testing data. Models are constructed from the training data and evaluated by the testing data (Cross-Validation for minimizing overfitting)
- The performance of the models is also evaluated in round 2 data



$P - \hat{P}$;

P: True Poverty Rate

$\hat{P}$:Estimated Poverty Rate

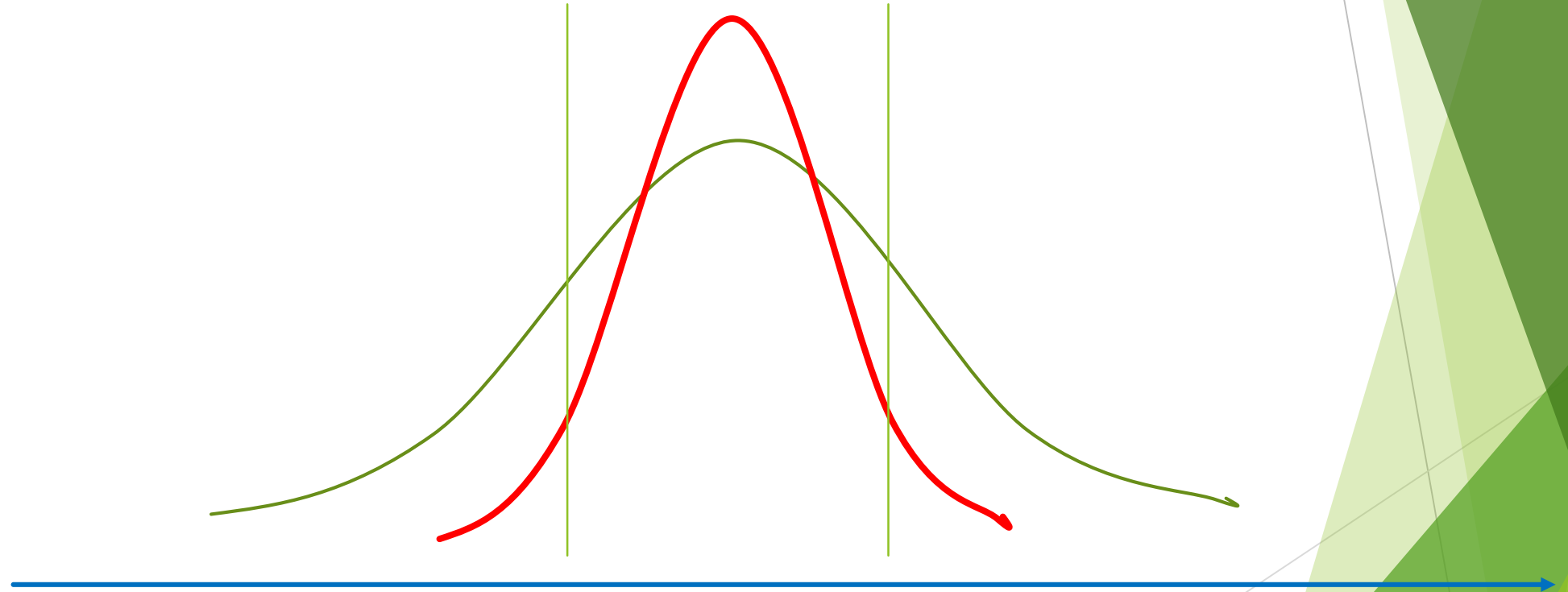
Issues of ML

WORLD BANK GROUP

Issues of ML

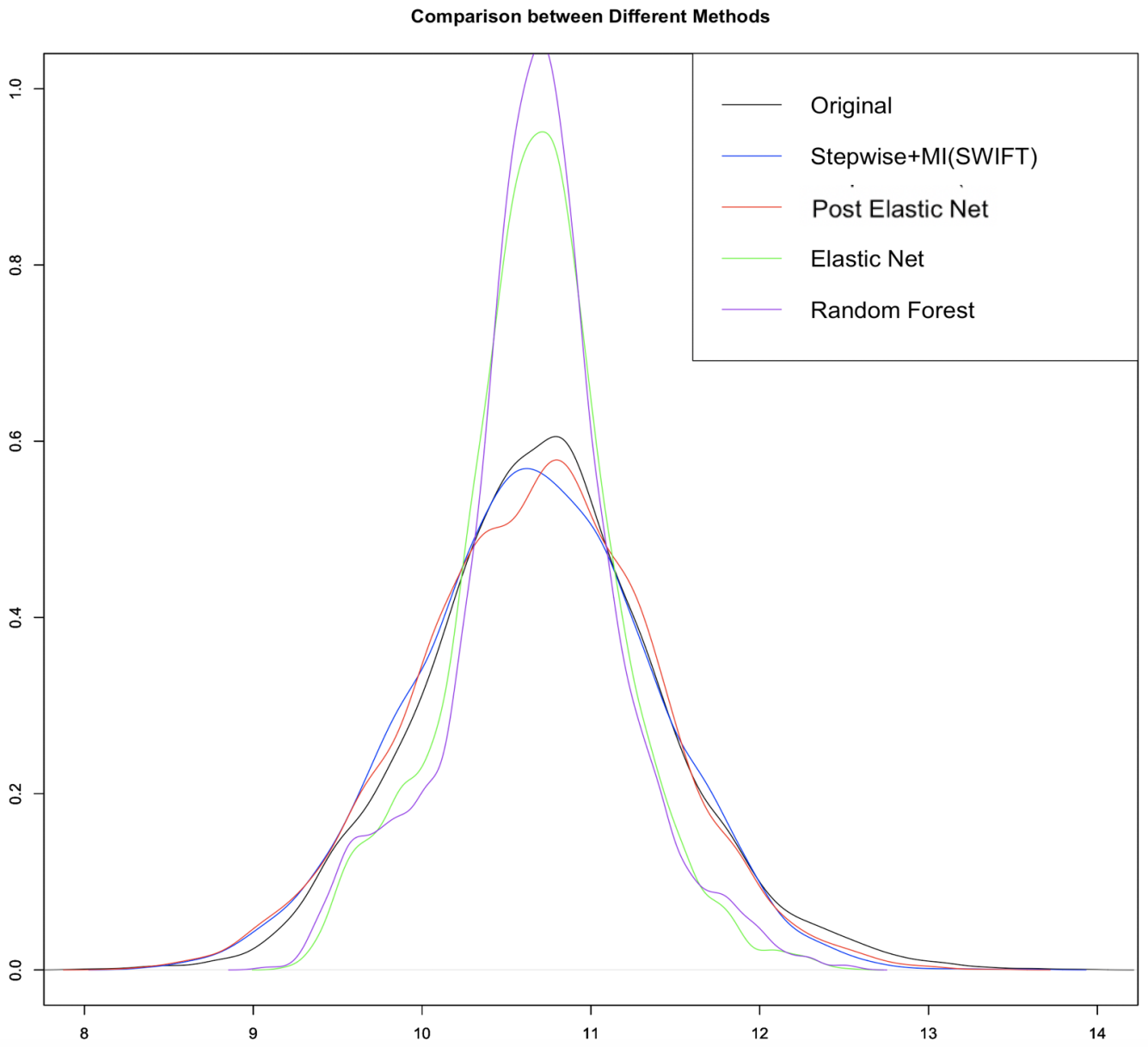
- ML produces predicted values only, not distribution.
- If we use them, we produce potentially large bias in poverty rates

$$\widehat{e_i} = X_i'\hat{\beta}_{EN} \text{ or } f_{RF}\,(X_i;\hat{\theta})$$



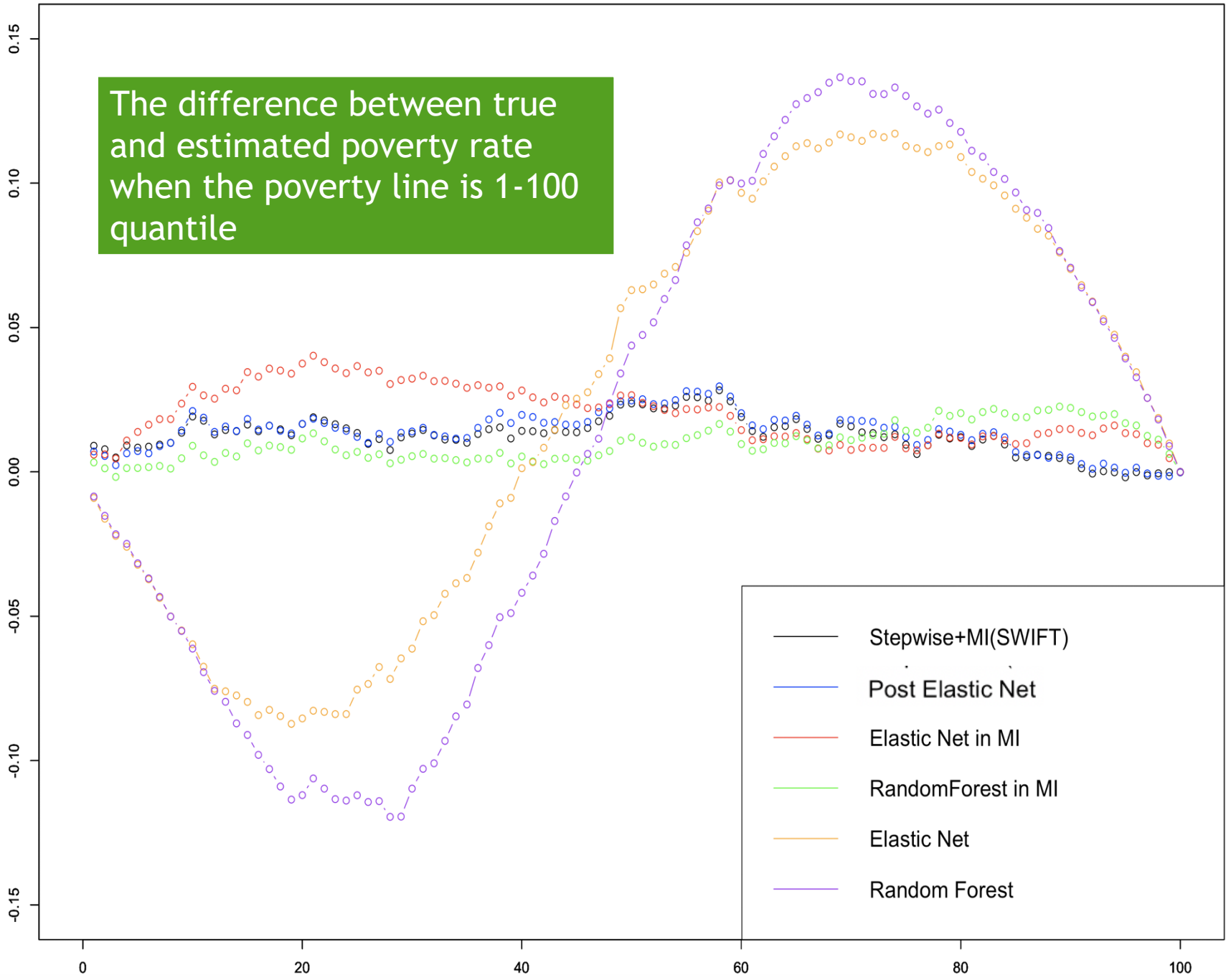
If ML predictor is used for estimating poverty rates, the estimates are usually biased heavily

WORLD BANK GROUP



WORLD BANK GROUP

The difference between true and estimated poverty rate when the poverty line is 1-100 quantile



WORLD BANK GROUP

Solution – ML in MI

WORLD BANK GROUP

Multiple Imputation is a key technique

- Considering a univariate variable x=(x_1,x_2,…,x_n) with p variables Z=(z_1,z_2,…,z_p) that follows a normal linear regression model;

$$x_i|z_i \sim N(z_i'\beta, \sigma^2) \qquad \text{(A)}$$

Let Z_o denote the observed components of Z and Z_m denote the missing components.

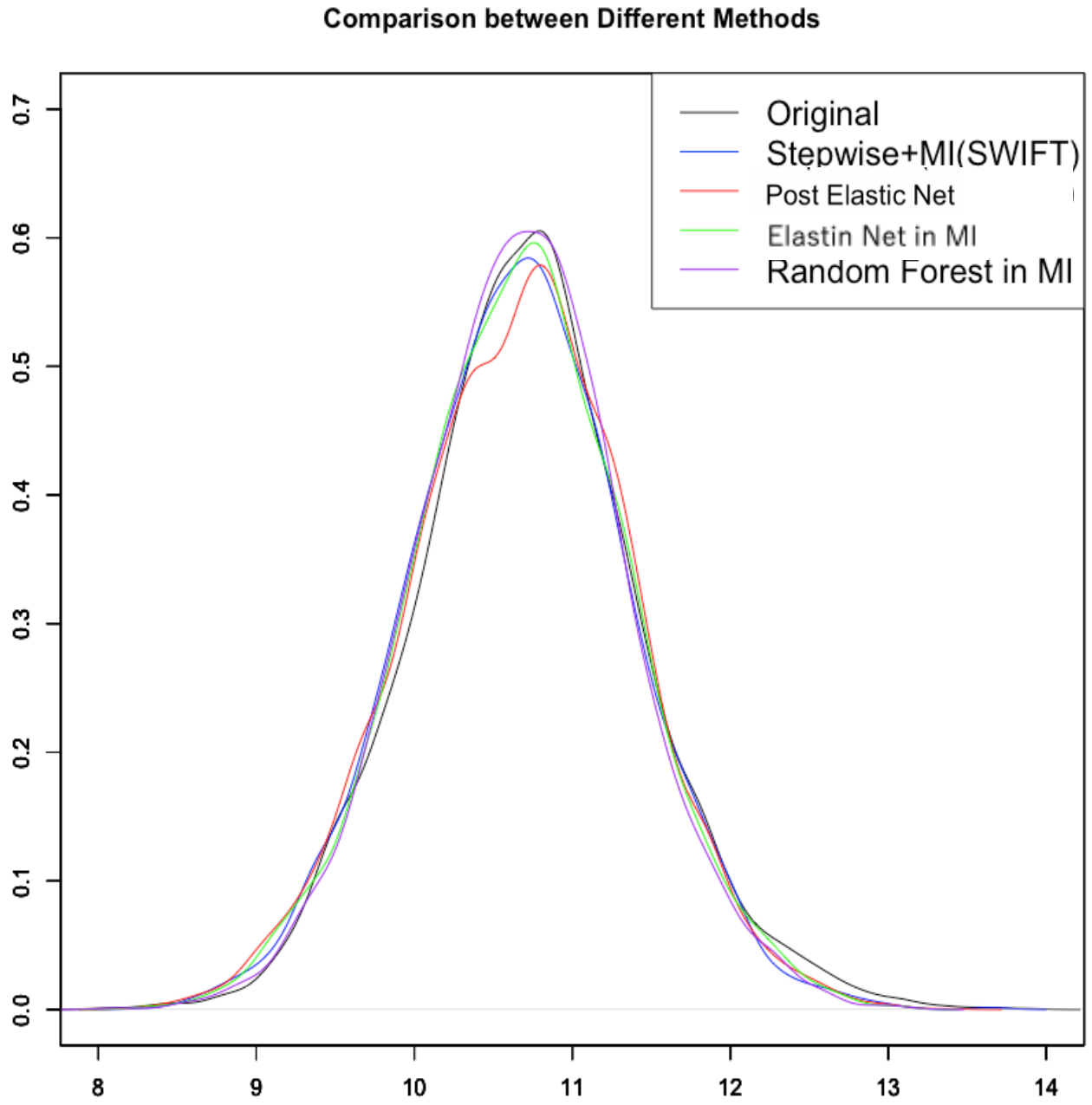
- 1) Fit a regression model (A) to the observed data (x_0, Z_0) to obtain estimates β and σ^2 of the model parameters.
- 2) Simulate new parameters β_* and σ_*^2 from their joint posterior distribution;
  - $\sigma_*^2 \sim \hat{\sigma}^2(n_0 - q)/X_{n0-q}^2$
  - $\beta_*|\sigma_*^2 \sim N\{\hat{\beta}, \sigma_*^2(Z_o'Z_o)^{-1}\}$
- 3) Obtain one set of imputed values, X_m^1 by simulating from $N(Z_m\beta_*, \sigma_*^2 I_{n1\times n1})$

WORLD BANK GROUP

- We have multiple options to combine MI with ML
- Differences are:
  - (i) how to select variables (X_i)
  - (ii) how to estimate coefficients (β)
  - (iii) how to draw error terms (u_i)

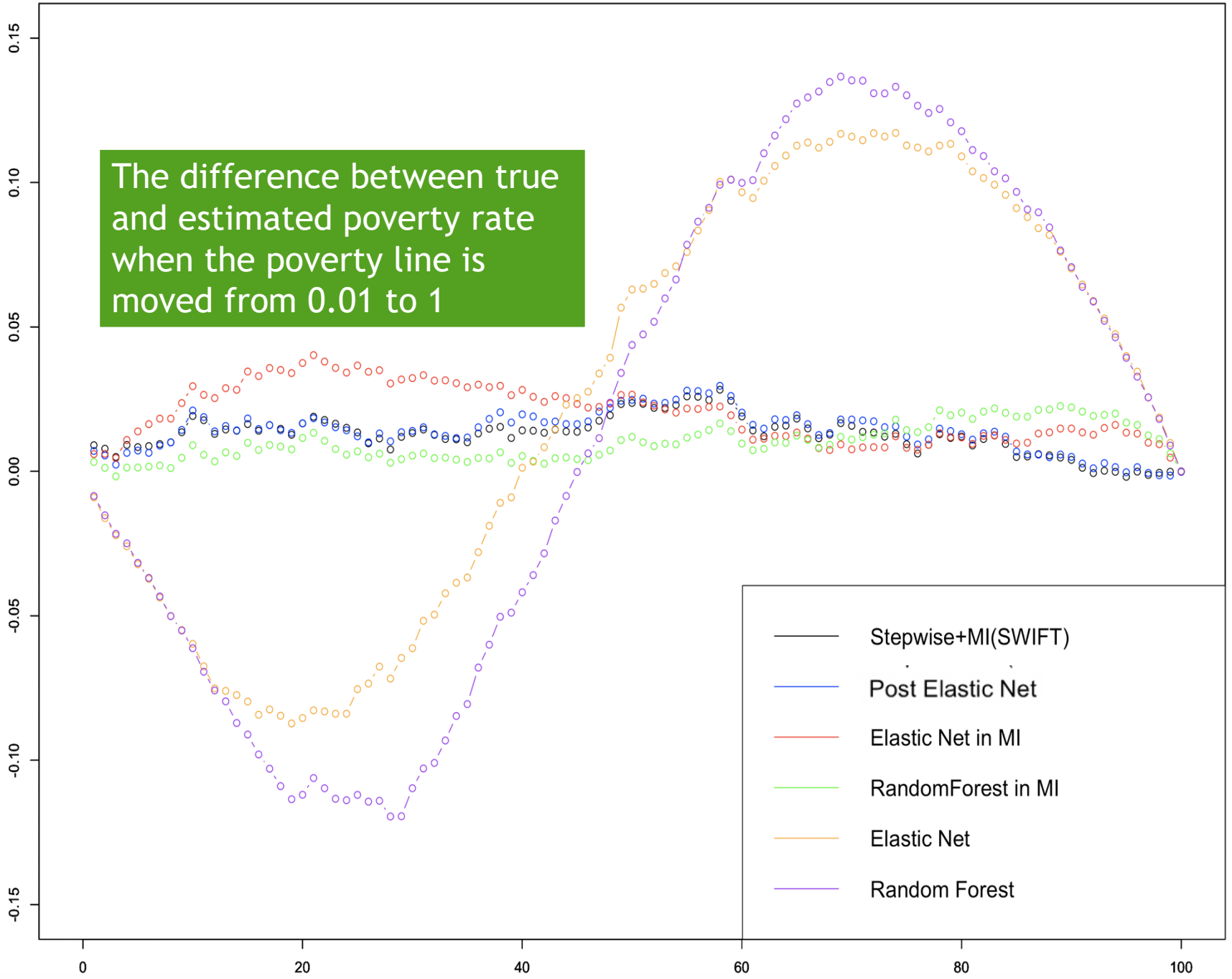
- All approaches combines ML with MI
  - Post EN – First select variables by EN and run MI for simulating household expenditures with OLS coefficients
  - SWIFT (stepwise + MI) – Select variables with statistical significance and run MI for simulating household expenditures with OLS coefficients
  - EN in MI – Get the distribution of $\beta \& \sigma$ using bootstrapped data by EN and randomly draw from the predictive distribution $N(X\beta, \sigma^2)$
  - RF in MI – Calculate $f_{RF}(X_i)$ and σ by Random Forest and randomly draw from the predictive distribution $N(f_{RF}(X_i), \sigma^2)$

Rural Uganda (2009 – Data for Modeling; 2012 – Data for Model Evaluation)



WORLD BANK GROUP

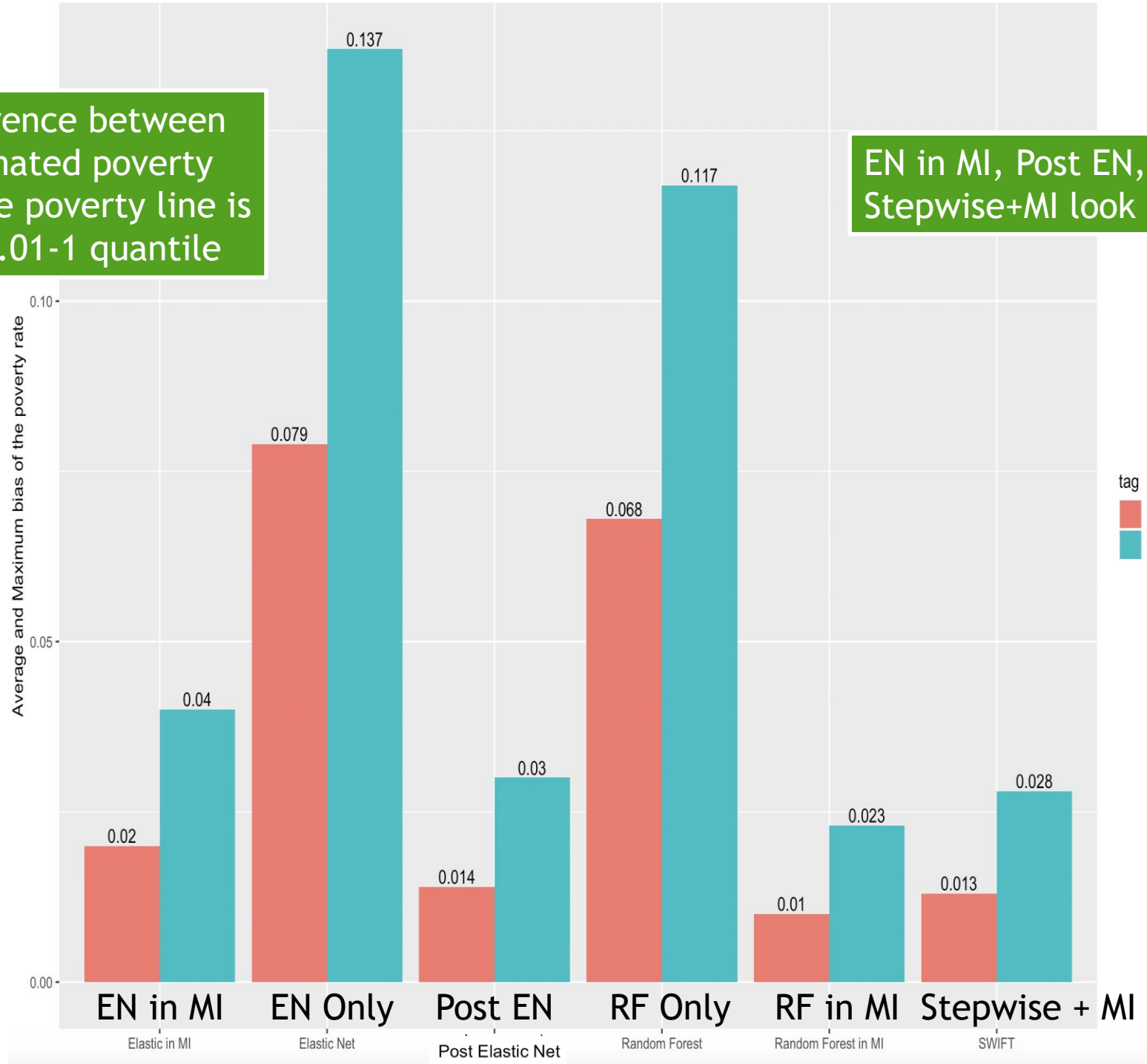
The difference between true and estimated poverty rate when the poverty line is moved from 0.01 to 1



WORLD BANK GROUP

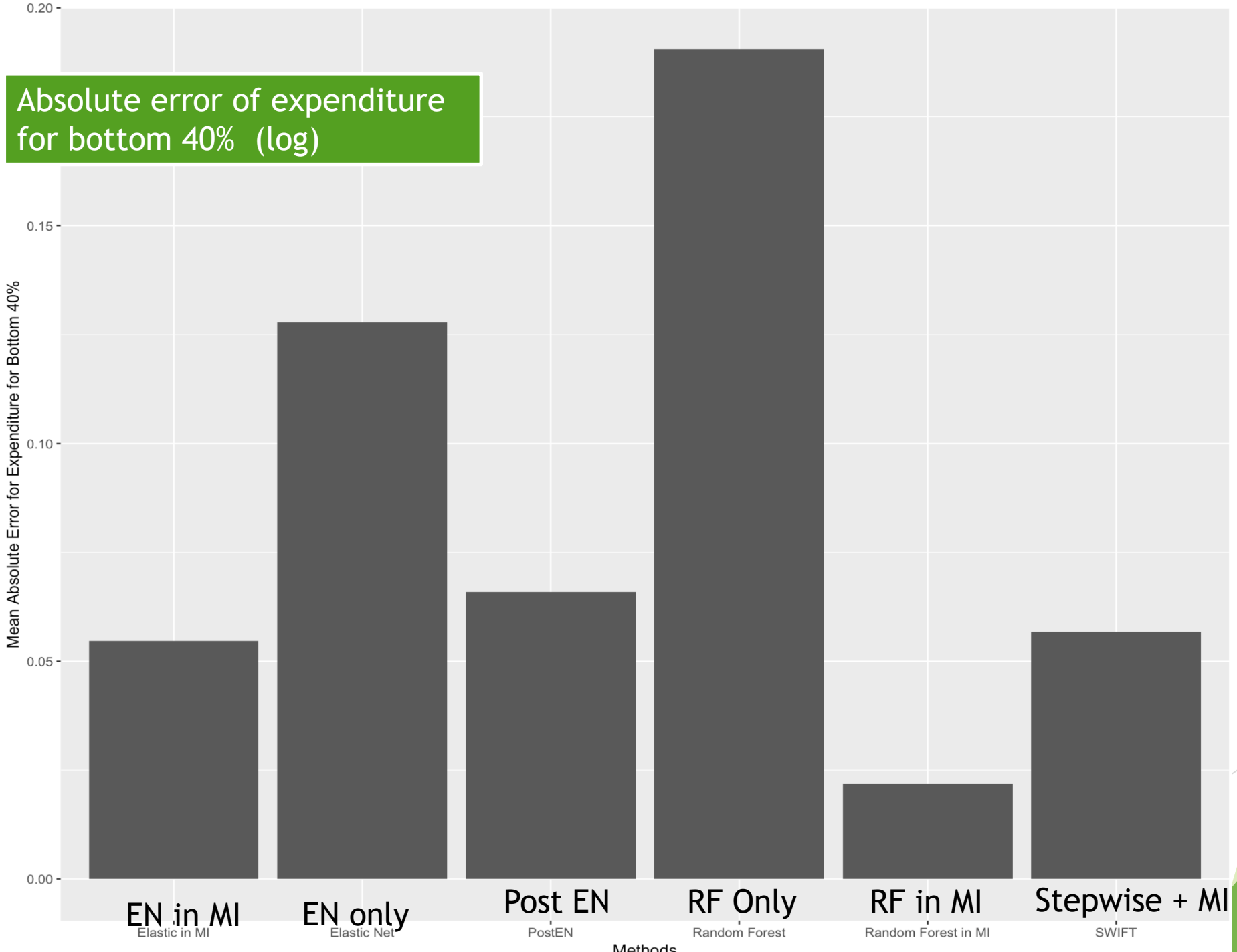
Average difference between true and estimated poverty rate, when the poverty line is moved from 0.01-1 quantile

EN in MI, Post EN, RF in MI and Stepwise+MI look promising!



WORLD BANK GROUP

Absolute error of expenditure for bottom 40% (log)



WORLD BANK GROUP

Refinement for ML in MI

WORLD BANK GROUP

Further modifications – additional variable selections

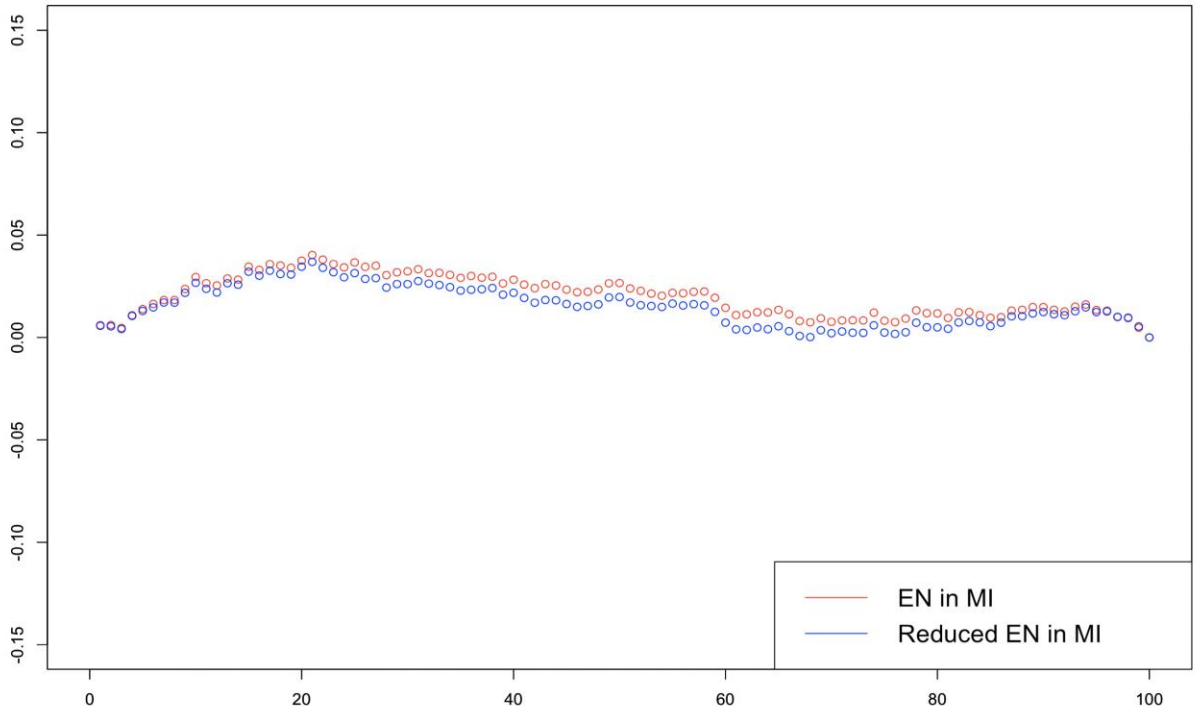
- Both EN in MI and RF in MI work well but they might need lots of variables
  - EN in MI with statistical significance tests (Reduced EN in MI) – After applying EN and obtaining the distribution of the β_i, drop variables whose coefficients are not statistically significant at 5%
  - RF in MI with variable importance tests (Reduced RF in MI) – After applying RF and obtaining the variable importance, drop variables whose variable importance is low
- In this way, the final variable set is manageable but the performance of projections is still good

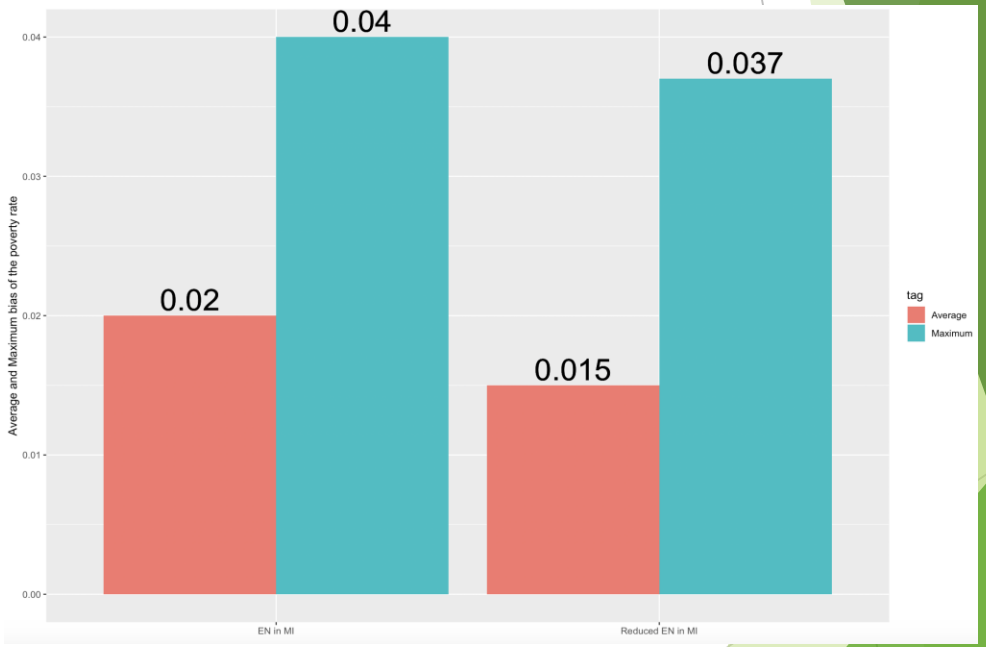
Number of variables		
Approach	**Rural Uganda**	**Rural Romania**
Total # vars	55	70
SWIFT	31	39
Post EN	41	51
EN in MI	55	70
RF in MI	55	70
Reduced EN in MI	20	24
Reduced RF in MI	10 or 20	10 or 20

WORLD BANK GROUP

Elastic Net in MI with variable selection (Rural Uganda 2009 to 2012)

Average difference between true and estimated poverty rate, when the poverty line is moved from 0.01-1 quantile





EN in MI

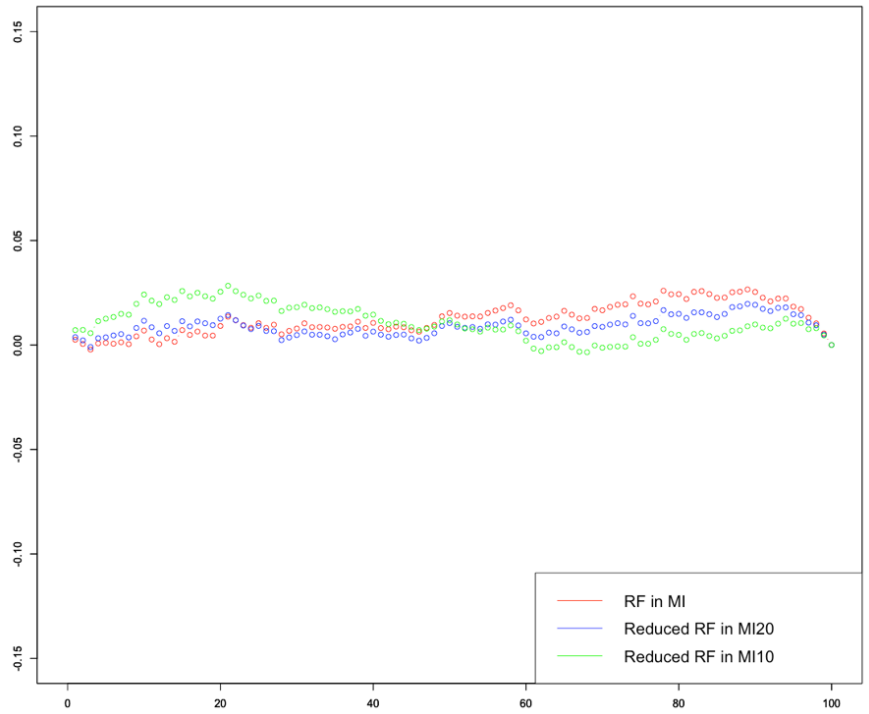
Reduced EN in MI

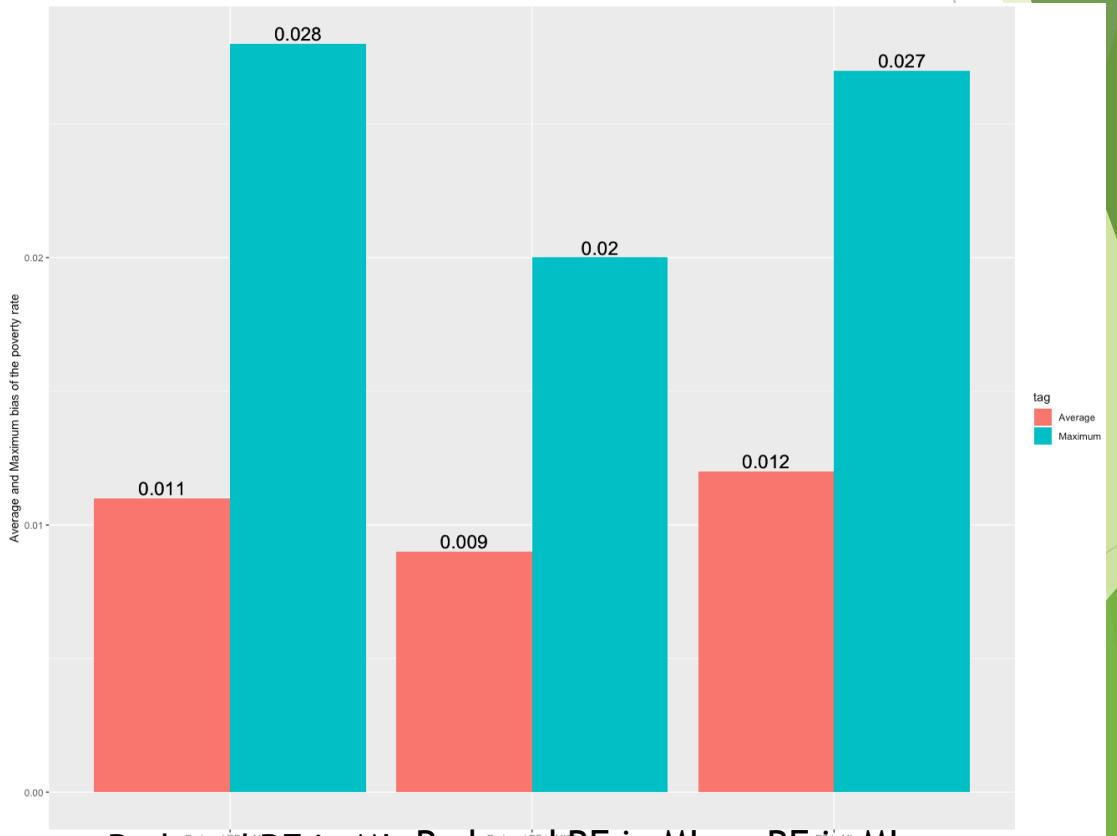
"Reduced EN in MI": Elastic Net with variable selection

WORLD BANK GROUP

Random Forest in MI with variable selection (Rural Uganda 2009 to 2012)

Average difference between true and estimated poverty rate, when the poverty line is moved from 0.01-1 quantile





Reduced RF in MI with 10

Reduced RF in MI with 20

RF in MI

WORLD BANK GROUP

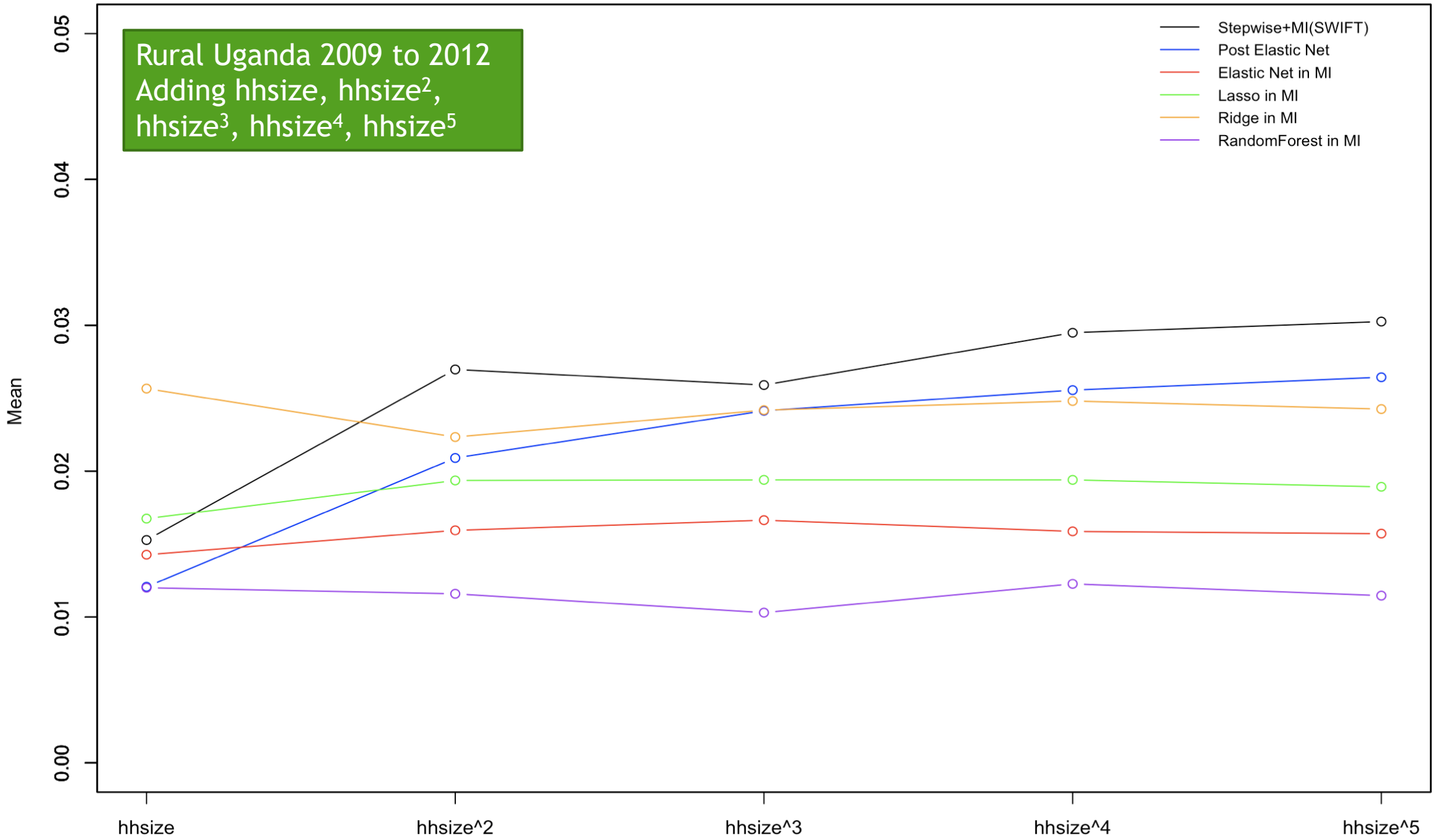
Effect of the strong collinearity on the performance

WORLD BANK GROUP

Multicollinearity

- When the explanatory variables are corelated with each other, it makes the coefficients unstable for the linear regression
- We compared the performance of different methodology accumulating the term of household size
  - Only hhsize + other variables
  - hhsize + $hhsize^2$ + other variables
  - hhsize + $hhsize^2$ + $hhsize^3$ + other variables
  - hhsize + $hhsize^2$ + $hhsize^3$ + $hhsize^4$ + other variables
  - hhsize + $hhsize^2$ + $hhsize^3$ + $hhsize^4$ + $hhsize^5$ + other variables
- Stepwise and Post EN were vulnerable to the multicollinearity, while other machine learning methodologies were quite robust

WORLD BANK GROUP

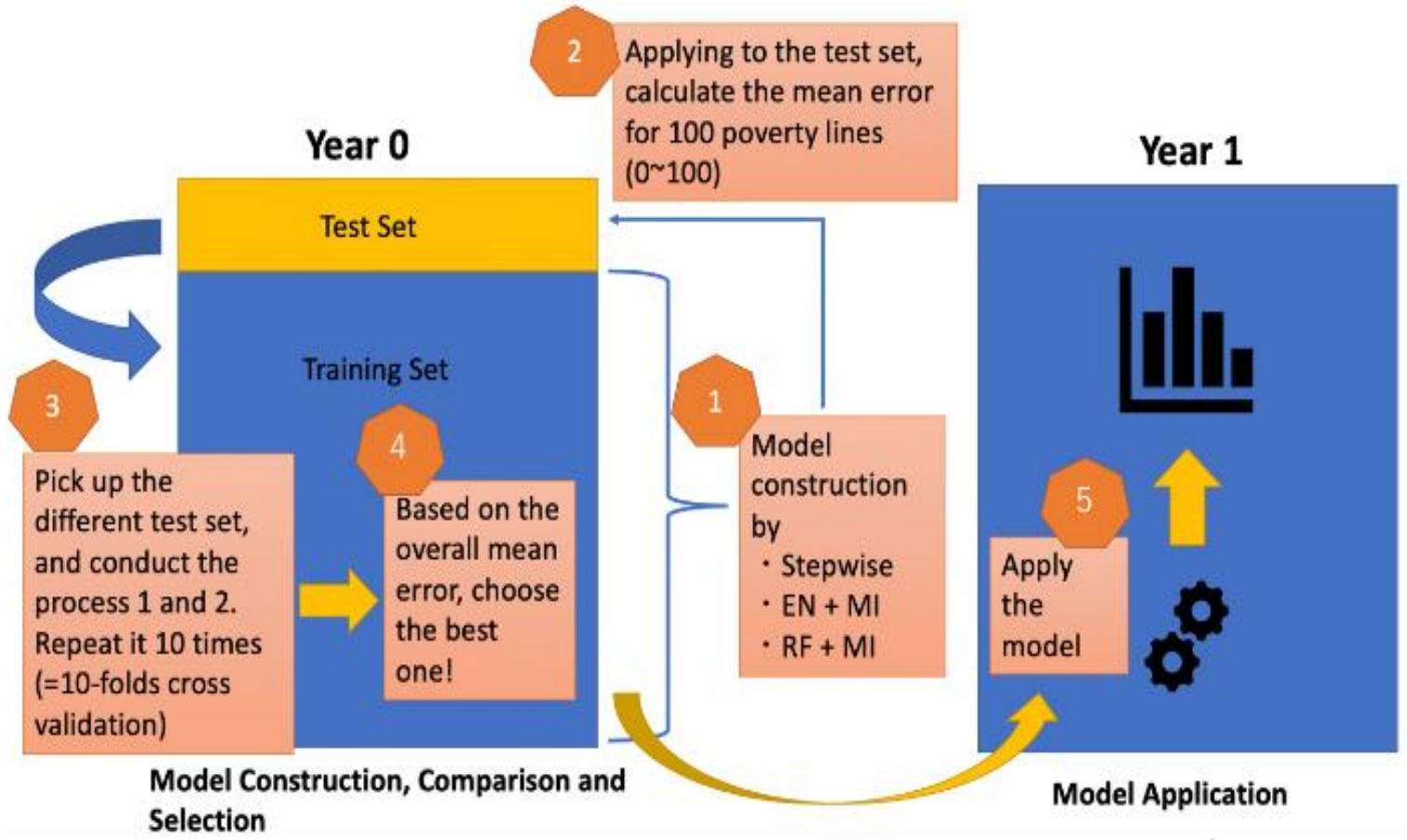

Rural Uganda 2009 to 2012
Adding hhsize, hhsize2, hhsize3, hhsize4, hhsize5
Stepwise+MI(SWIFT)
Post Elastic Net
Elastic Net in MI
Lasso in MI
Ridge in MI
RandomForest in MI
Mean
0.00
0.01
0.02
0.03
0.04
0.05
hhsize
hhsize^2
hhsize^3
hhsize^4
hhsize^5

Selection of approaches via cross-validation

WORLD BANK GROUP

How to determine the best methodology only from modeling data?

In reality, there is no evaluation data and we need to decide the best methodology only from the modeling data. We usually do that using cross-validation (CV) technique – splitting data between training and testing data.



WORLD BANK GROUP

Limitations and future research

- Limitations
  - All distributions are normal distributions
  - Only a few ML approaches are tested
- Future research
  - More examples with non-normal distribution
  - More ML approaches – Neural Network (or deep learning) etc.
  - Research on the use of multi-country database like GMD – research on “Transfer Learning”
- Collaborations with academia
  - Osaka University, Tokyo University, and Boston University

WORLD BANK GROUP

Key takeaways

- Direct application of ML for measuring poverty might not work well
- Careful combination between ML and MI works well but too many variables might be selected
- Reducing the number of variables by significance tests for EN and variable importance for RF might even improve the performance of ML
- Machine learning techniques combined with MI are robust to the multicollinearity

Thank you!

If you have any question, please feel free to contact us:

Kazusa Yoshimura
(kyoshimura1@worldbank.org)

Nobuo Yoshida
(nyoshida@worldbank.org)

WORLD BANK GROUP

Annex

WORLD BANK GROUP

Post EN and Stepwise + MI

$$e_i = X_i\beta + u_i$$

1. Variable selections (X_i^*)
   - Post EN – Elastic Net for selecting variables
   - Stepwise + MI – Stepwise for selecting variables
2. Estimation of distributions of β and u using OLS
   - It is easy to derive the distributions of β and u as long as we use OLS
   - That is not necessarily the case if we use EN or RF
3. Simulate household expenditures by drawing β and u from the distributions

$$\tilde{e}_i = X_i^*\tilde{\beta} + \tilde{u}_i$$

EN in MI

Modeling Data
$\{e_i^1, X_i^1\}$

Bootstrapped data 1
$\{e_{i1}^1, X_{i1}^1\}$

Bootstrapped data 2
$\{e_{i2}^1, X_{i2}^1\}$

.....

Bootstrapped data B
$\{e_{Bi}^1, X_{iB}^1\}$

$X_i * \hat{\beta}_{EN1} + u_i$
With $\hat{\sigma}_1^{\,2}$

$X_i * \hat{\beta}_{EN2} + u_i$
With $\hat{\sigma}_2^{\,2}$

$X_i * \hat{\beta}_{ENB} + u_i$
With $\hat{\sigma}_B^{\,2}$

Random draw from $N(X\hat{\beta}_j, \hat{\sigma}_j^{\,2})$

Evaluation Data
$\{e_i^3, X_i^3\}$

$\tilde{e}_{i1}^3 = \tilde{X}_i^3 * \hat{\beta}_{EN1} + \tilde{u}_{i1}^3$

$\tilde{e}_{i2}^3 = \tilde{X}_i^3 * \hat{\beta}_{EN2} + \tilde{u}_{i2}^3$

$\tilde{e}_{iB}^3 = \tilde{X}_i^3 * \hat{\beta}_{ENB} + \tilde{u}_{iB}^3$

Comparison between $\{\tilde{e}_i^3\}_{i=1}^N$ and $\{\{\tilde{e}_{ik}^3\}_{i=1}^N\}_{k=1}^B$ in terms of prediction of poverty and shared prosperity indices

WORLD BANK GROUP

RF in MI

Modeling Data $\{e_i^1, X_i^1\}$

Bootstrapped data 1 $\{e_{i1}^1, X_{i1}^1\}$

Bootstrapped data 2 $\{e_{i2}^1, X_{i2}^1\}$

.....

Bootstrapped data B $\{e_{iB}^1, X_{iB}^1\}$

$f_{RF}(X_i, \hat{\theta}^1)$ with $\hat{\sigma}_1^2$

$f_{RF}(X_i, \hat{\theta}^2)$ with $\hat{\sigma}_2^2$

$f_{RF}(X_i, \hat{\theta}^B)$ with $\hat{\sigma}_B^2$

Random draw of $\tilde{u}_{ij}^3$ from $N(0, \hat{\sigma}_j^2)$

Evaluation Data $\{e_i^3, X_i^3\}$

$\tilde{e}_{i1}^3 = f_{RF}(\tilde{X}_i^3, \hat{\theta}^1) + \tilde{u}_{i1}^3$

$\tilde{e}_{i2}^3 = f_{RF}(\tilde{X}_i^3, \hat{\theta}^2) + \tilde{u}_{i2}^3$

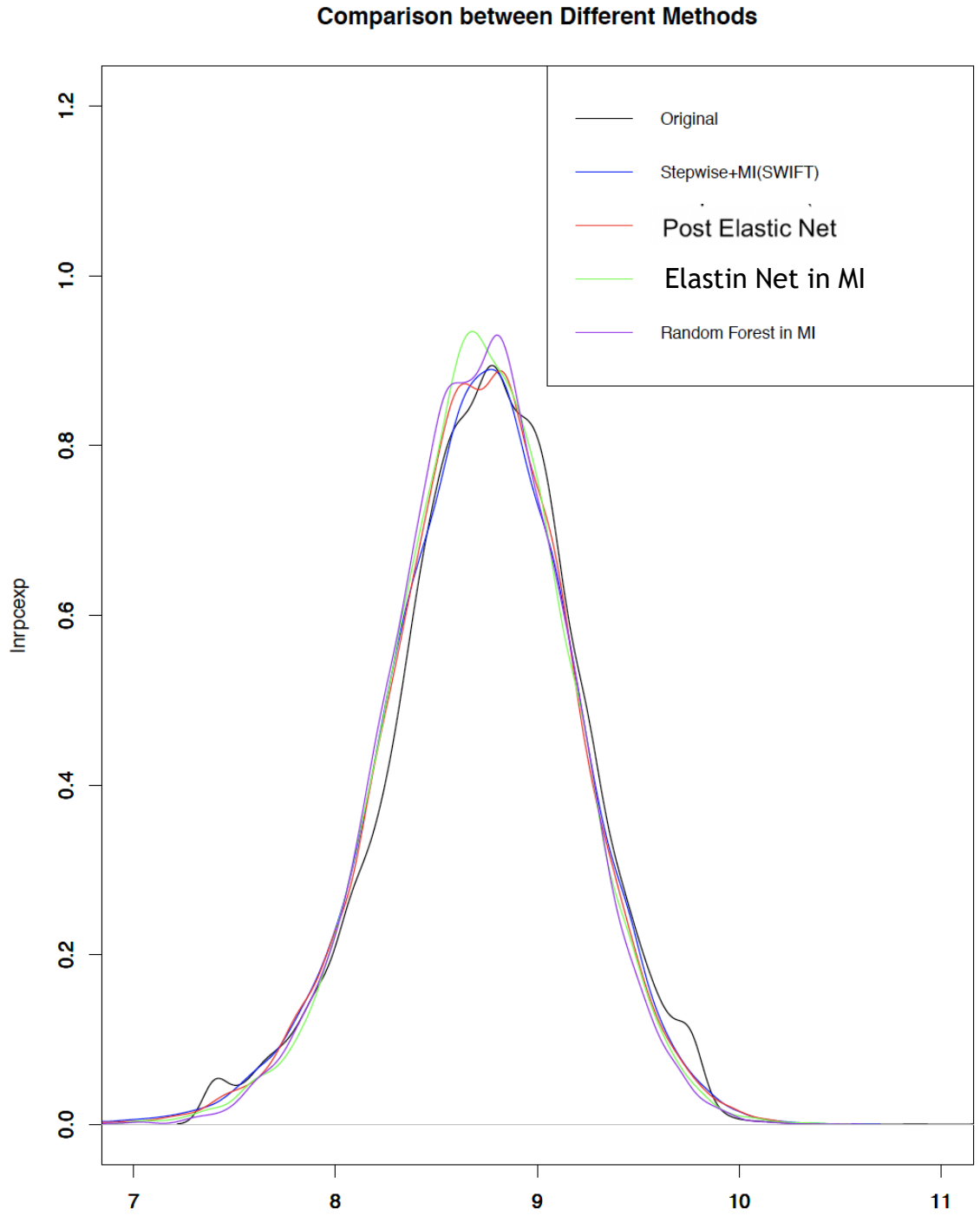
$\tilde{e}_{iB}^3 = f_{RF}(\tilde{X}_i^3, \hat{\theta}^B) + \tilde{u}_{iB}^3$

Comparison between $\{e_i^3\}$ and $\{\{\tilde{e}_{ik}^3\}\}_{k=1}^B$ in terms of prediction of poverty and shared prosperity indices

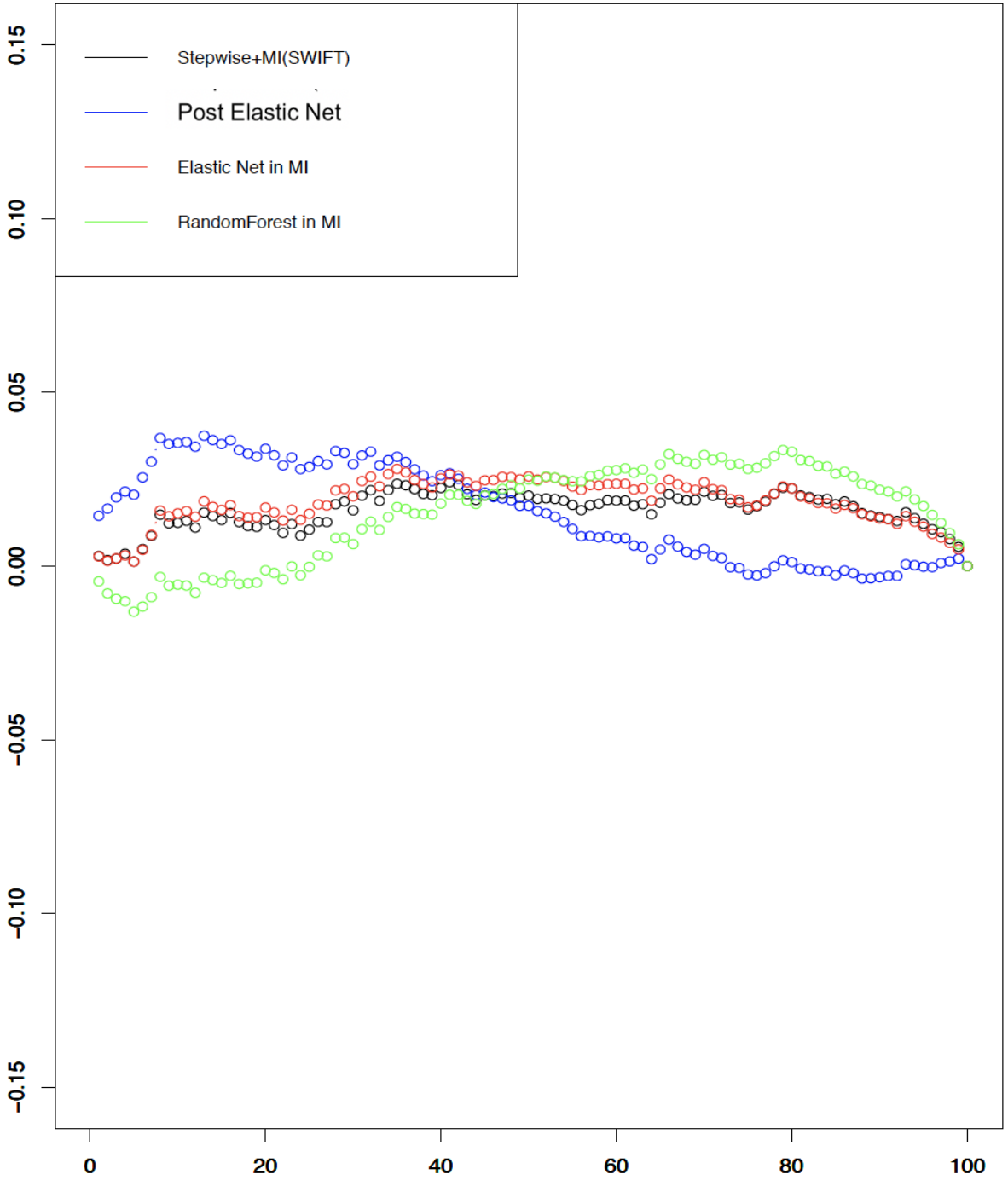
WORLD BANK GROUP

Rural Romania (2011 – Data for Modeling; 2012 – Data for model evaluation)

WORLD BANK GROUP

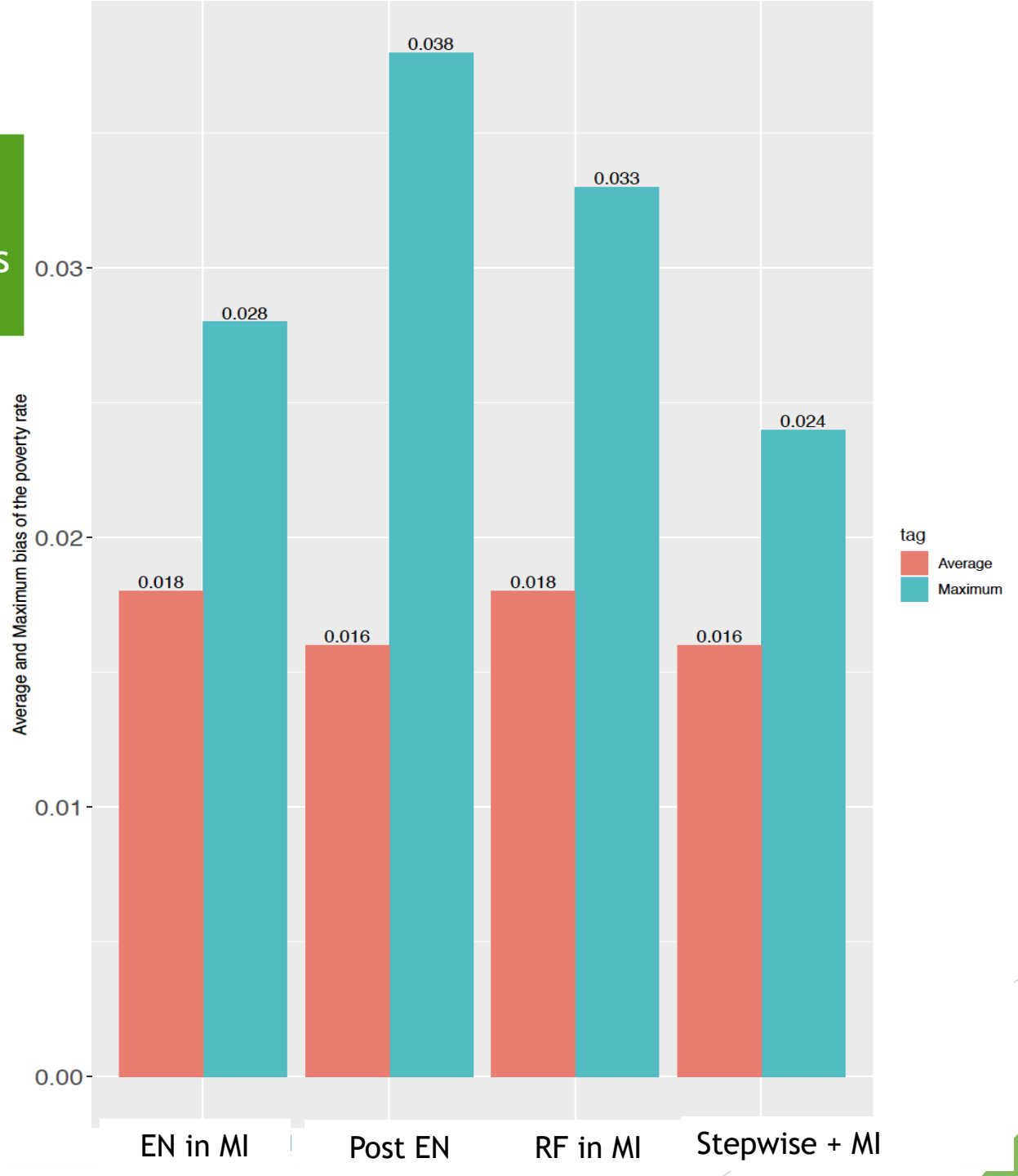


WORLD BANK GROUP



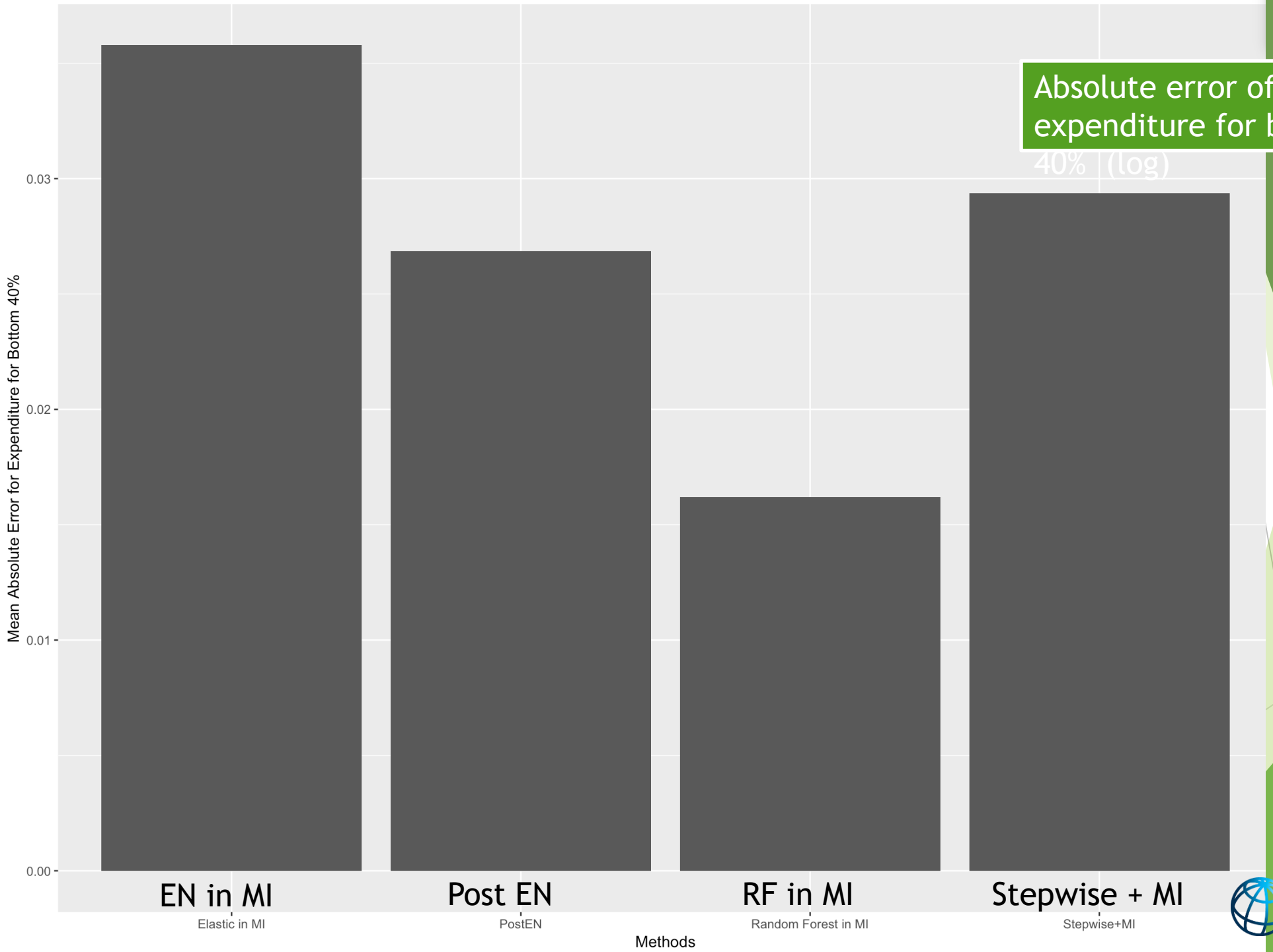
WORLD BANK GROUP

Average difference between true and estimated poverty rate, when the poverty line is moved from 0.01-1 quantile



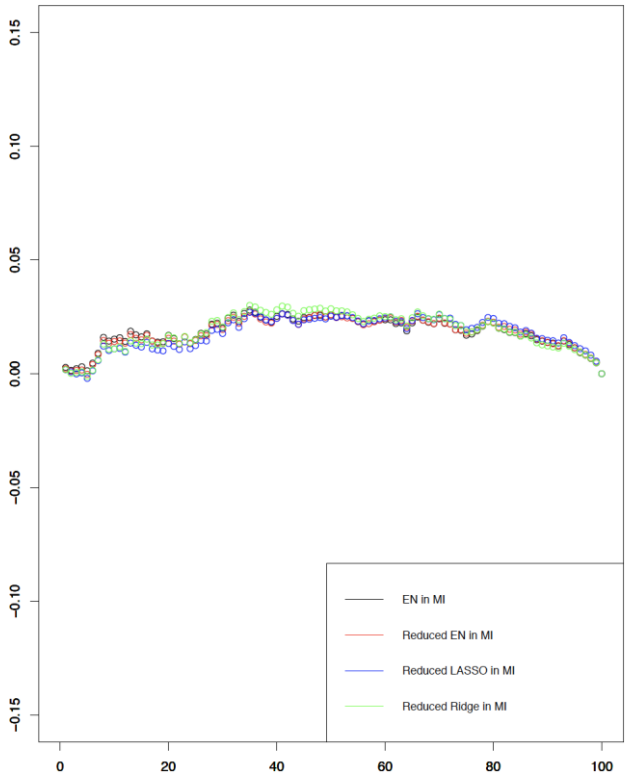
WORLD BANK GROUP

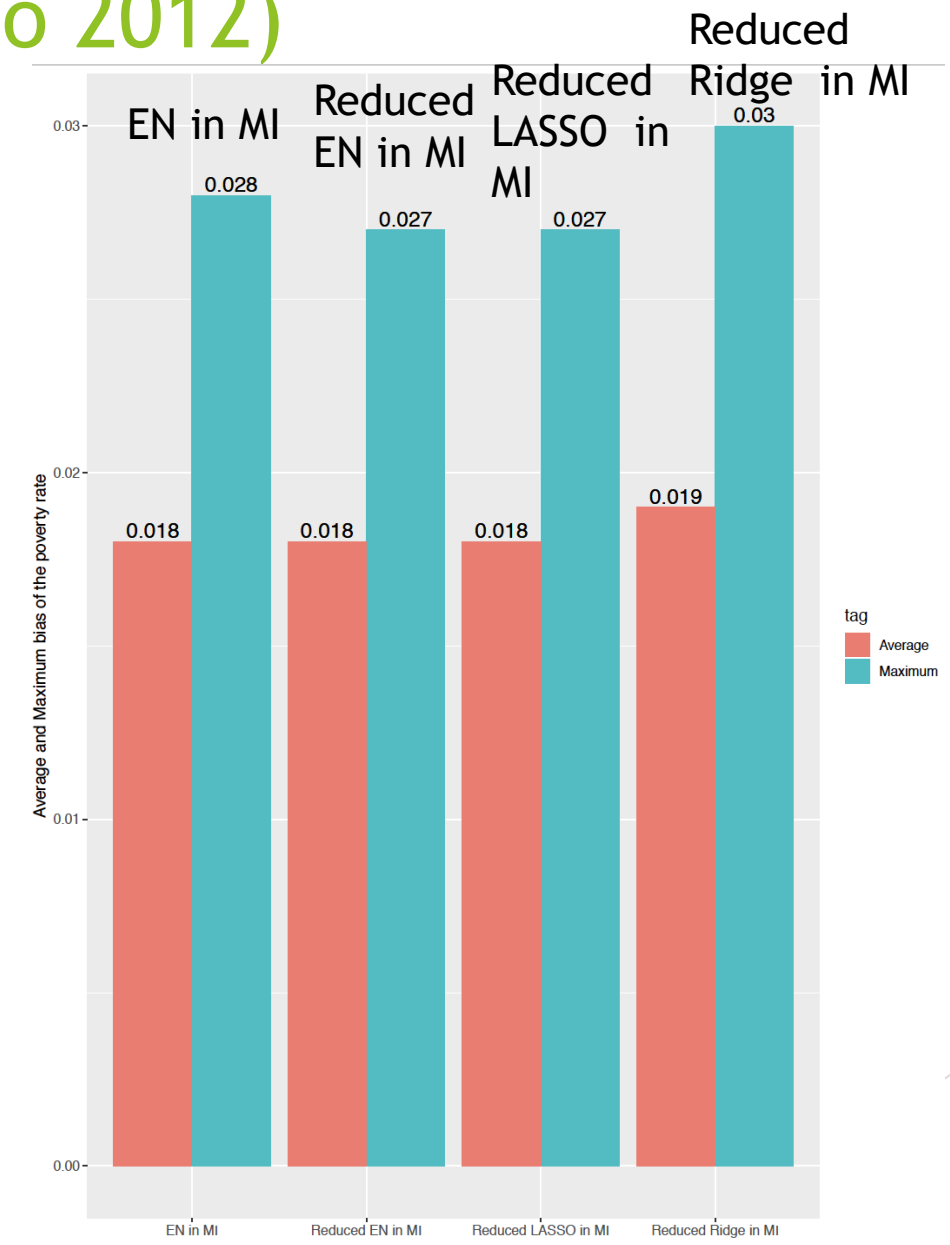
Absolute error of expenditure for bottom 40% (log)



WORLD BANK GROUP

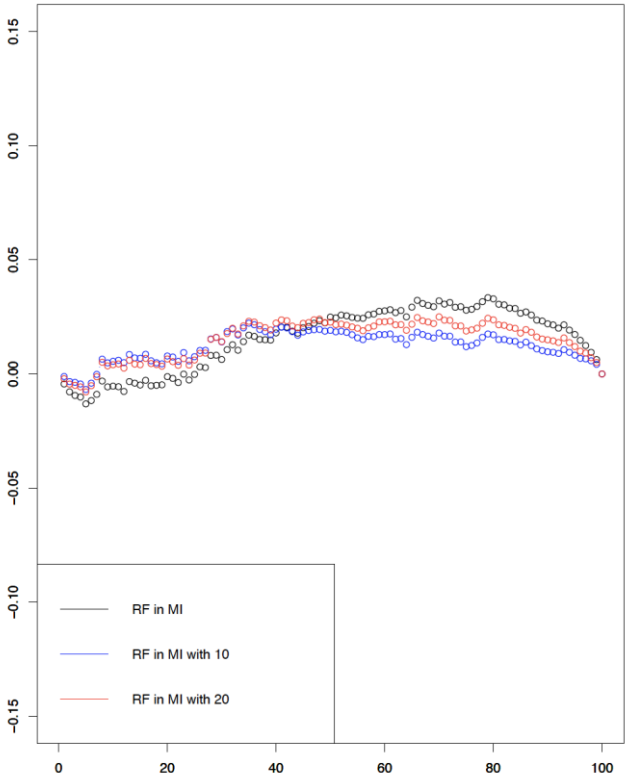
Elastic Net in MI with variable selection (Rural Romania 2011 to 2012)

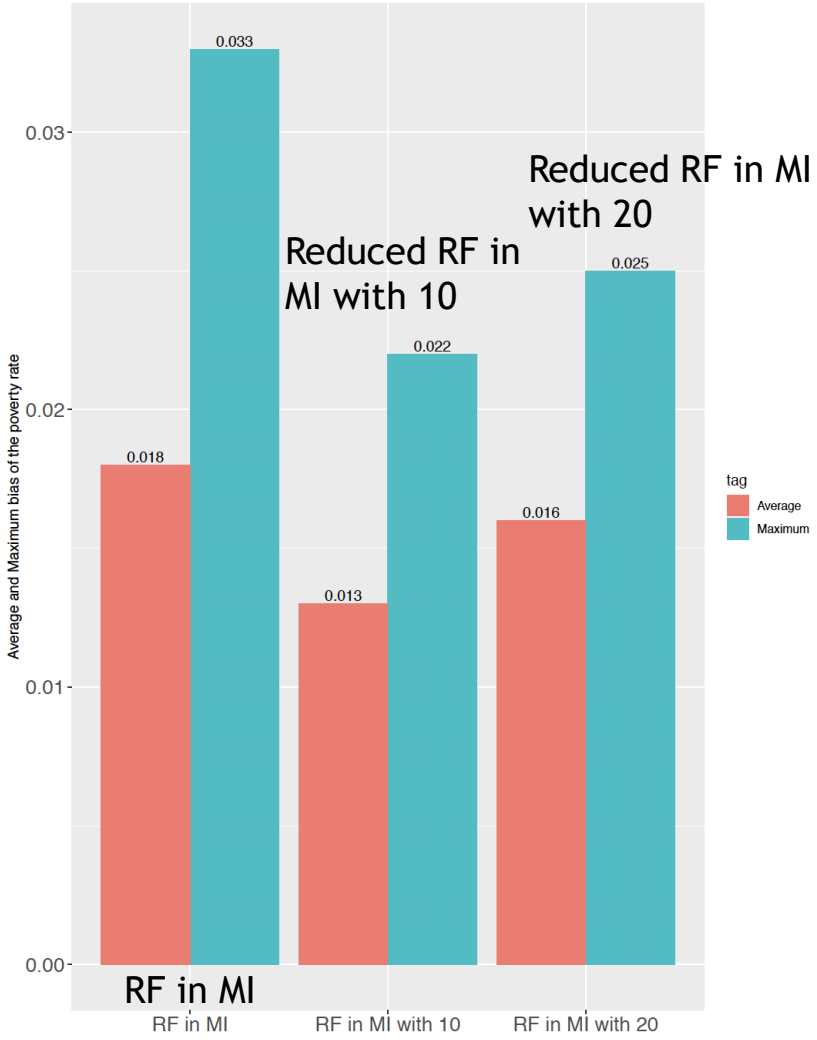




WORLD BANK GROUP

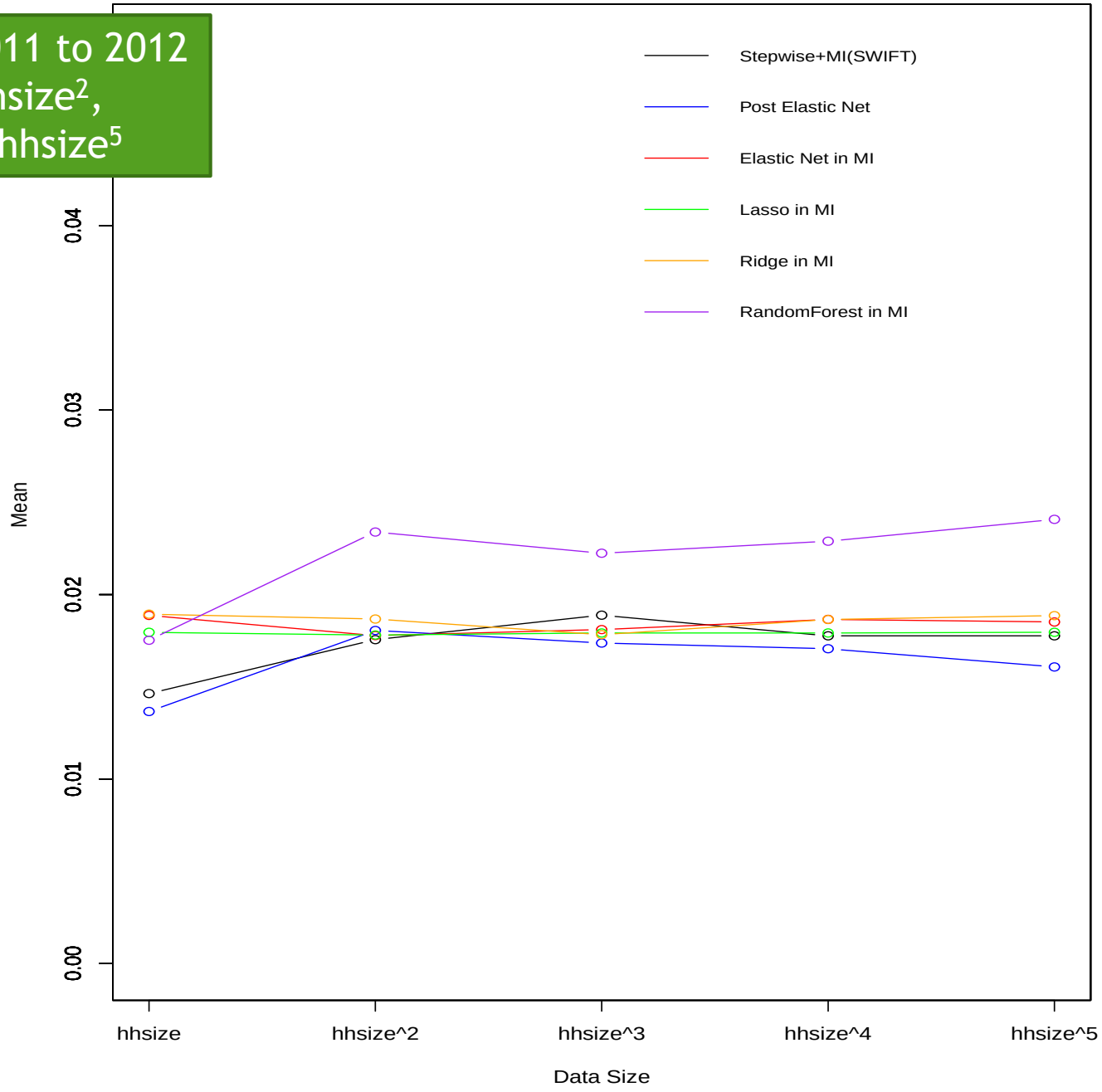
Random Forest in MI with variable selection (Rural Romania 2011 to 2012)





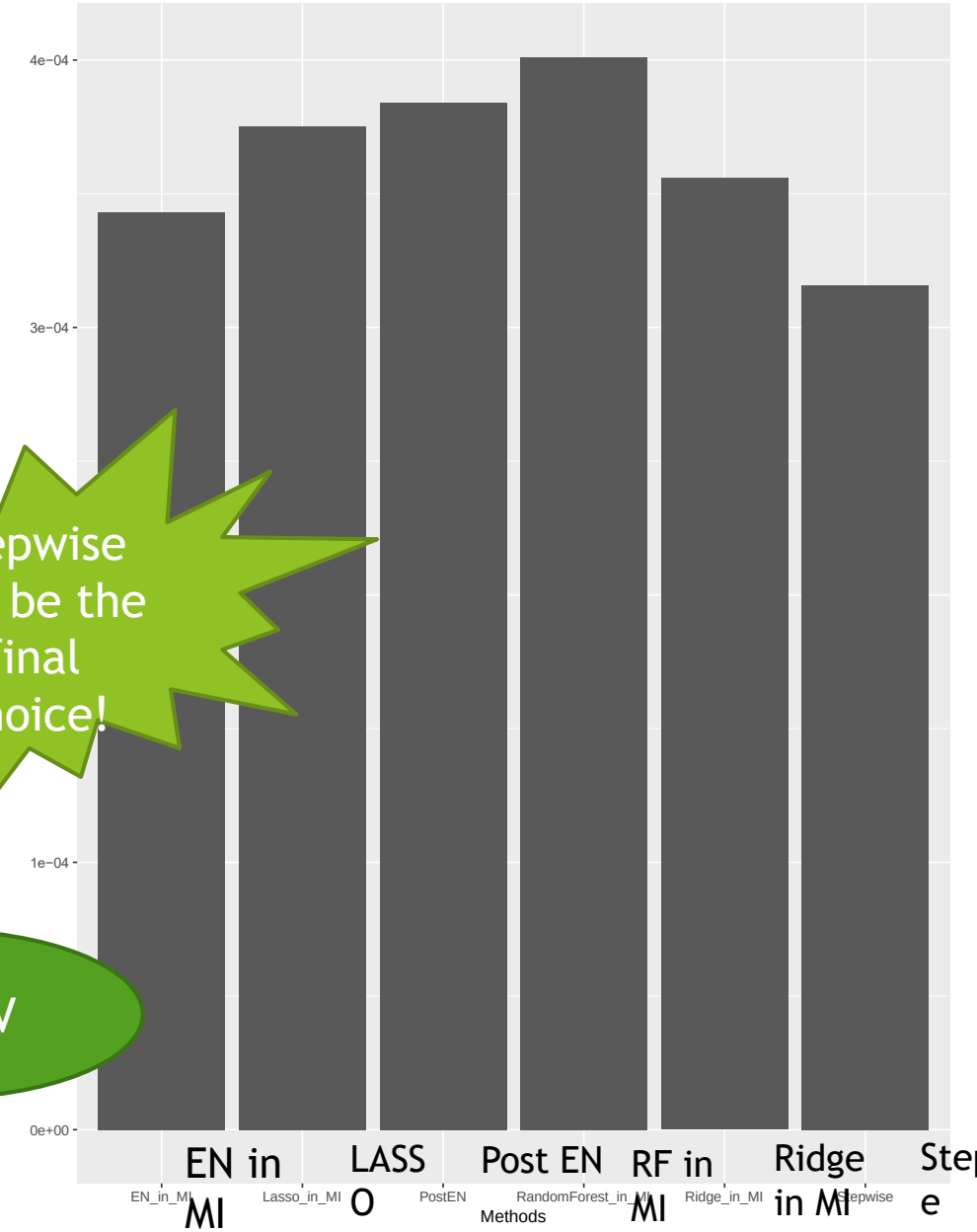
WORLD BANK GROUP

Rural Romania 2011 to 2012
Adding hhsize, hhsize^2, hhsize^3, hhsize^4, hhsize^5



WORLD BANK GROUP

Model selection with CV for Rural Romania and its performance



Stepwise will be the final choice!

2011 CV