Machine Learning for monitoring the twin goals – Pitfalls and Solutions

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

June 19, 2019

@Tokyo office World Bank

Monitoring of poverty and shared prosperity (assuming a linear model)

- Data generation process (a linear model)

$$e_i = X_i^* \beta + u_i$$

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

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

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

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 (Xi)

$$e_i = X_i \ \beta + u_i$$

- Using the imputed expenditures, we estimate poverty rates and shared prosperity index
- Accuracy of the imputed expenditures is critical for estimation of poverty and shared prosperity indices

Machine Learning (ML) for poverty projections

- ML is becoming very popular in the field of development as well (especially among development practitioners)
- ML is a useful approach for predicting any indicators that are very costly to estimate directly
- ML is now used for poverty projections; but there are many possible pitfalls

Machine learning techniques (LASSO; RIDGE; Elastic Net)

- Since Elastic Net (EN) is a nested model of LASSO and RIDGE, we show only EN's optimization
- EN follows the two stage optimization process

1. $\beta(\lambda_1, \lambda_2) = argmin_\beta \sum_{i=1}^{N_1} (e_i^1 - X_i^{1\prime}\beta)^2 + \lambda_1 \sum_{k=1}^{K} (\beta_k)^2 + \lambda_2 \sum_{k=1}^{K} |\beta_k|$
2. $min_{\lambda_1, \lambda_2} \sum_{i=1}^{N_2} (e_i^2 - X_i^{2\prime}\beta(\lambda_1, \lambda_2))^2$

$where \ \{(e_i^1, X_i^1)\}_{i=1}^{N_1} are\ training\ data\ and\ \{(e_i^2, X_i^2)\}_{i=1}^{N_2}\ are\ testing\ data$

Notes

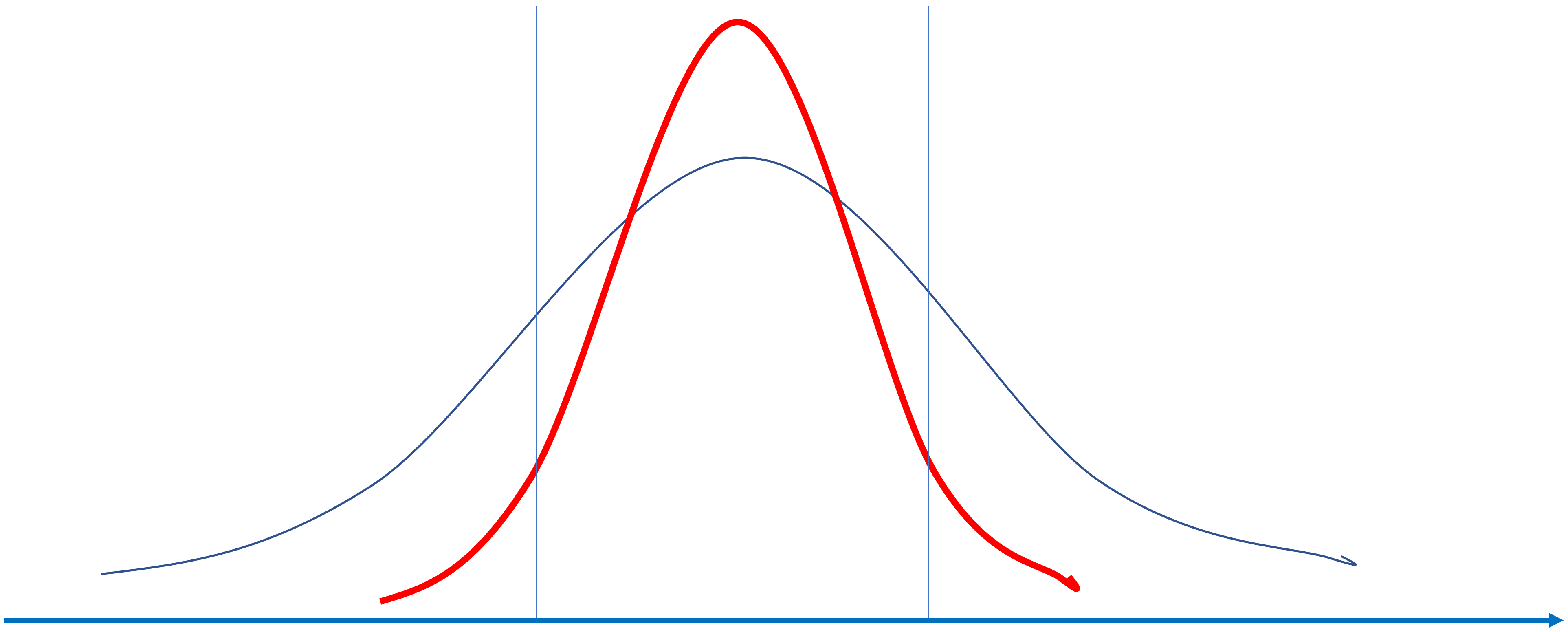
- LASSO ($\lambda_1 = 0$ & $\lambda_2 > 0$); RIDGE ($\lambda_1 > 0$ & $\lambda_2 = 0$); OLS ($\lambda_1 = 0$ & $\lambda_2 = 0$)
- If training data = testing data, OLS is optimal
- If OLS is not selected, absolute values of coefficients are smaller than OLS estimates – called "Shrinkage"
- ML produces predicted values: $\widehat{e_i} = X_i{}'\hat{\beta}_{EN}$

Issues of ML

- ML – use of predicted values only – potentially large bias in poverty rates

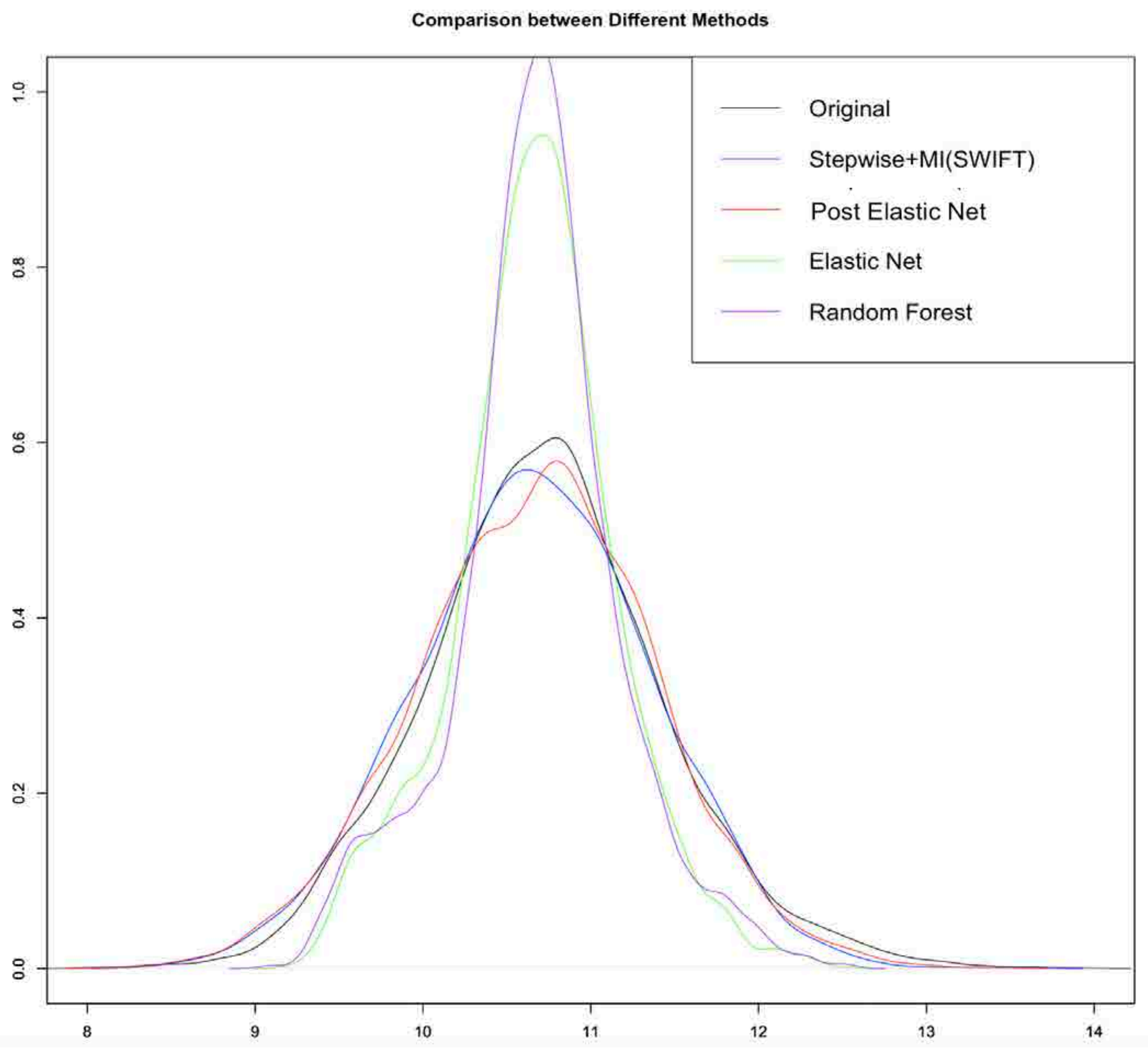
$$\widehat{e_i} = X_i'\hat{\beta}_{EN}$$

Bias due to the use of ML predictor

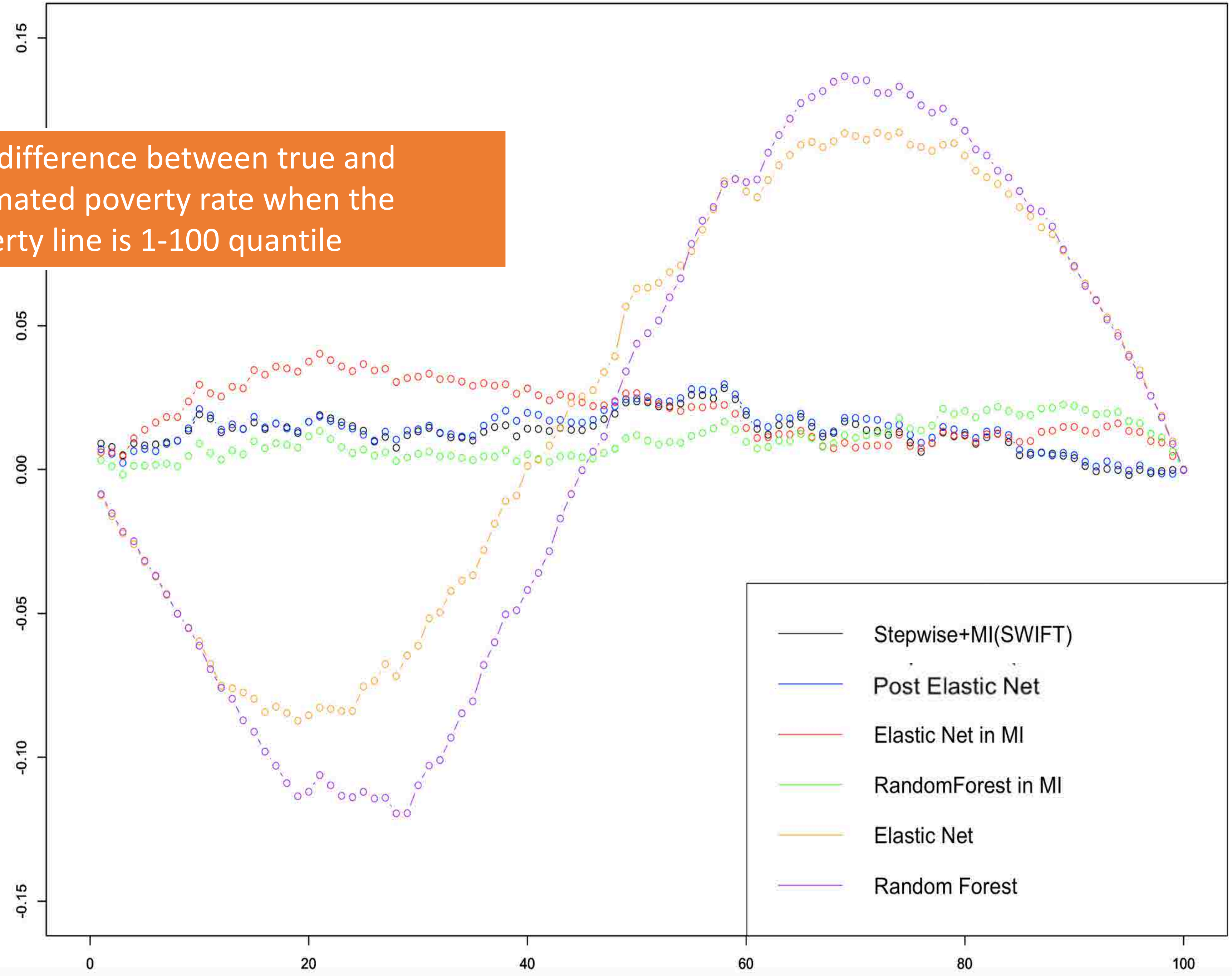


If ML predictor is used for estimating poverty rates, the estimates are usually biased heavily

Uganda Rural
2009 data – Training
2012 data – Testing



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



Possible solutions – Combining Multiple Imputations (MI) with ML

- MI – Adding error terms on predicted values

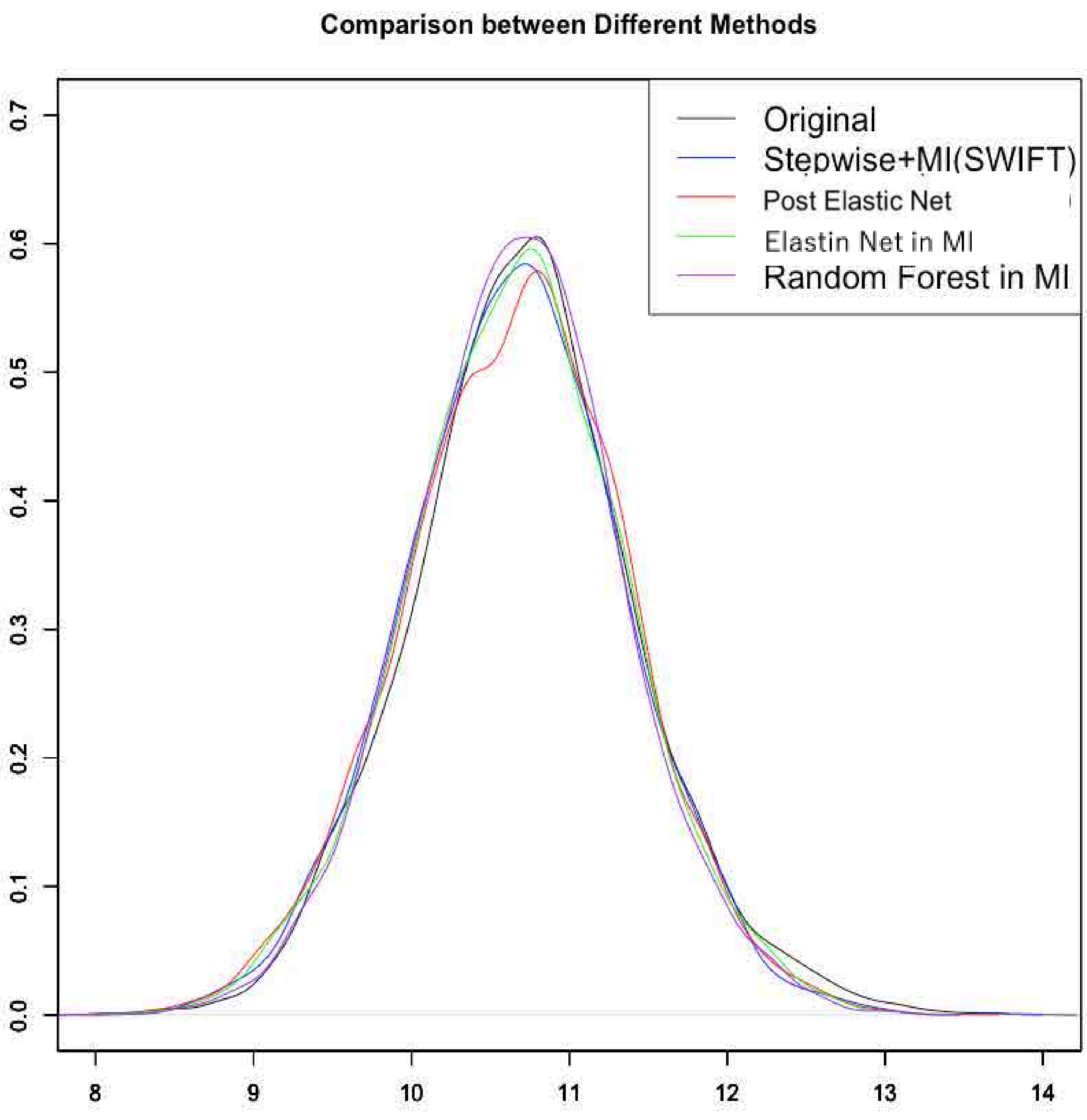
$$\tilde{e_i} = X_i\tilde{\beta} + \tilde{u_i}$$

- In the World Bank, 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)

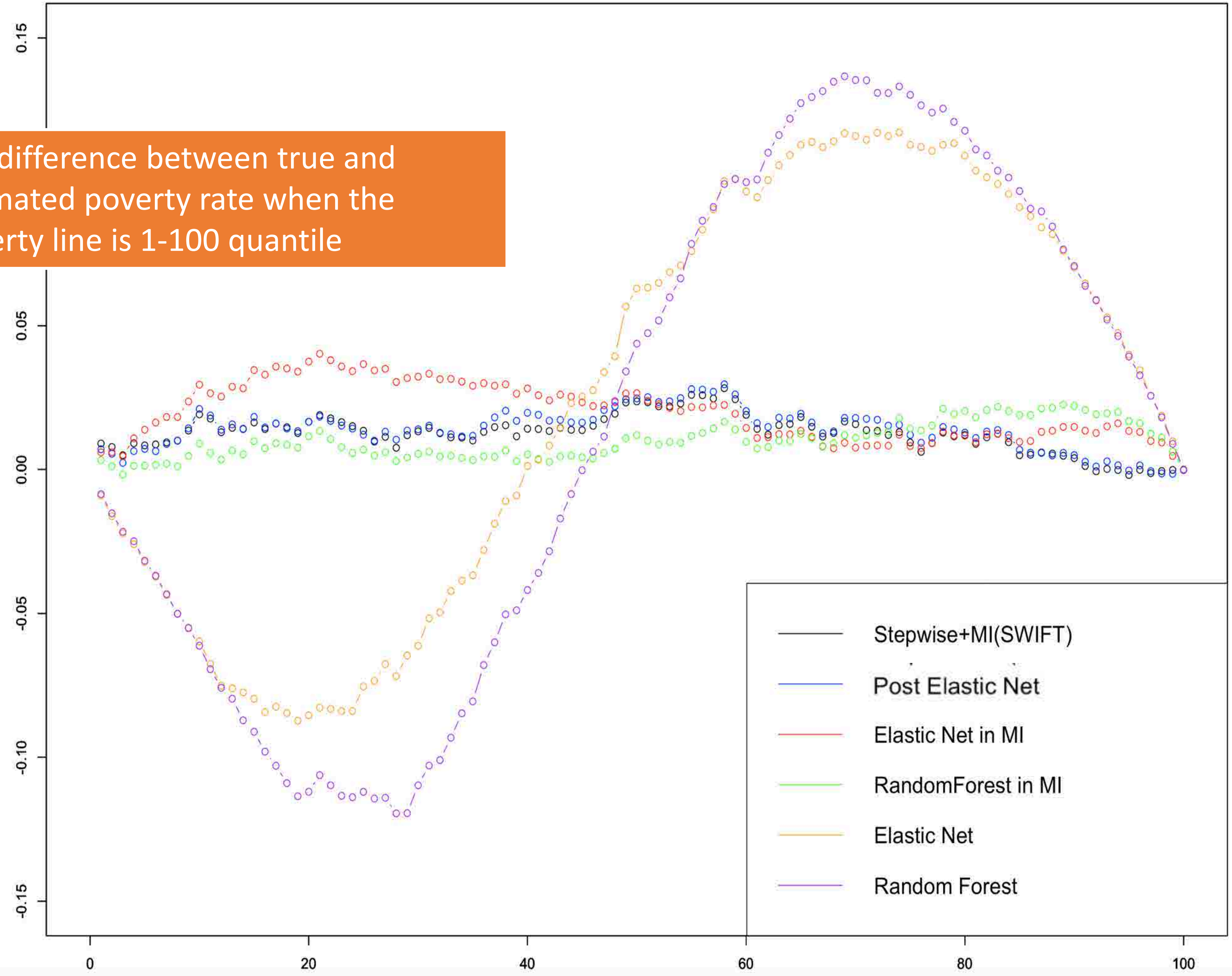
Some proposed approaches in the World Bank

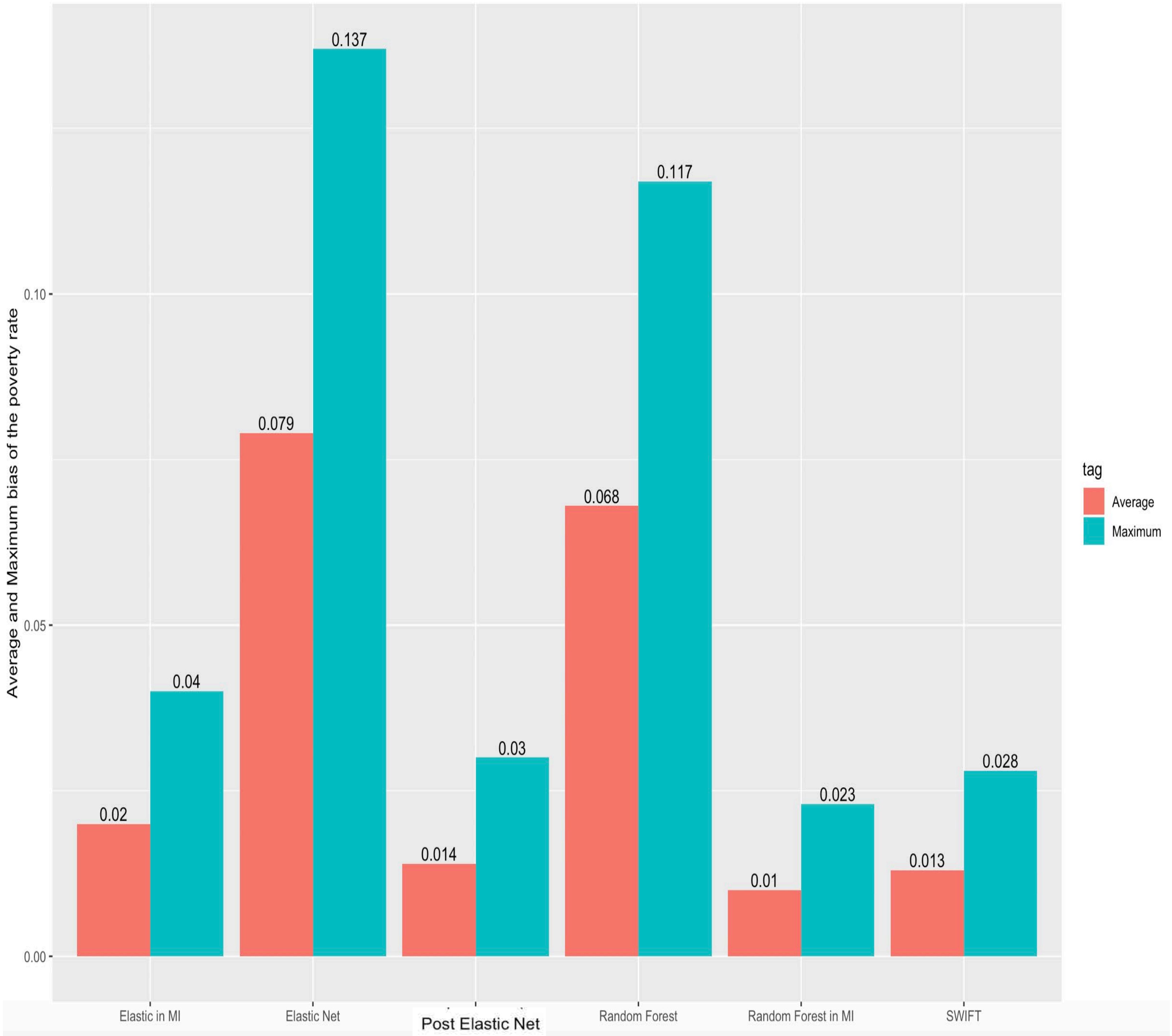
- All approaches combines ML with MI
  - Post EN – First select variables by EN and run MI for simulating household expenditures with OLS coefficients
  - SWIFT (stepwise + MI) – Select variables with statistical significance and run MI for simulating household expenditures with OLS coefficients
  - EN in MI – Select variables by EN and run MI for simulating household expenditures with EN coefficients
  - RF in MI – Select dummies and coefficients by RF and draw errors randomly

Rural Uganda (2009 – training data; 2012 – testing data)


Comparison between Different Methods
Original
Stepwise+MI(SWIFT)
Post Elastic Net
Elastin Net in MI
Random Forest in MI
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
8
9
10
11
12
13
14

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

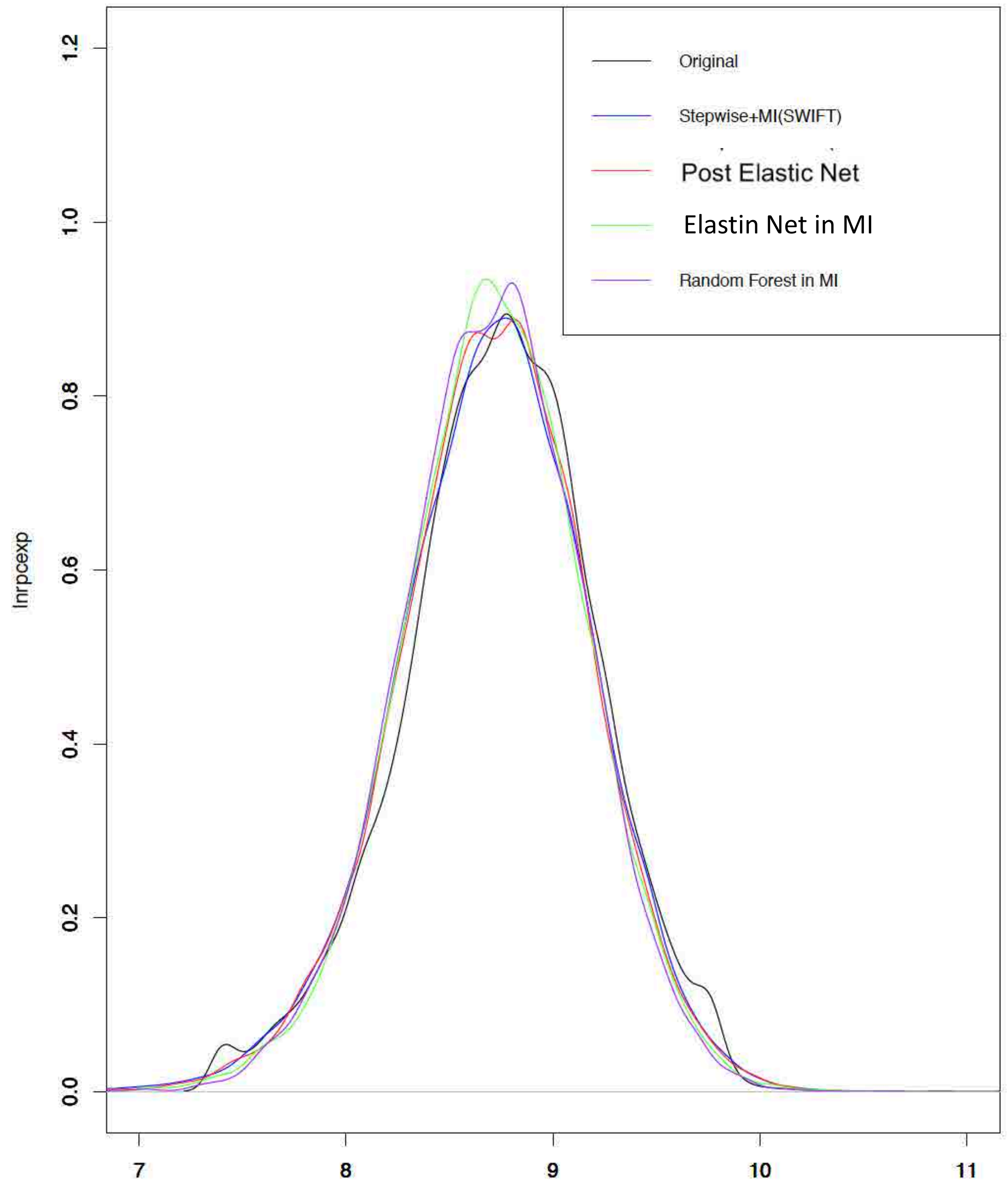


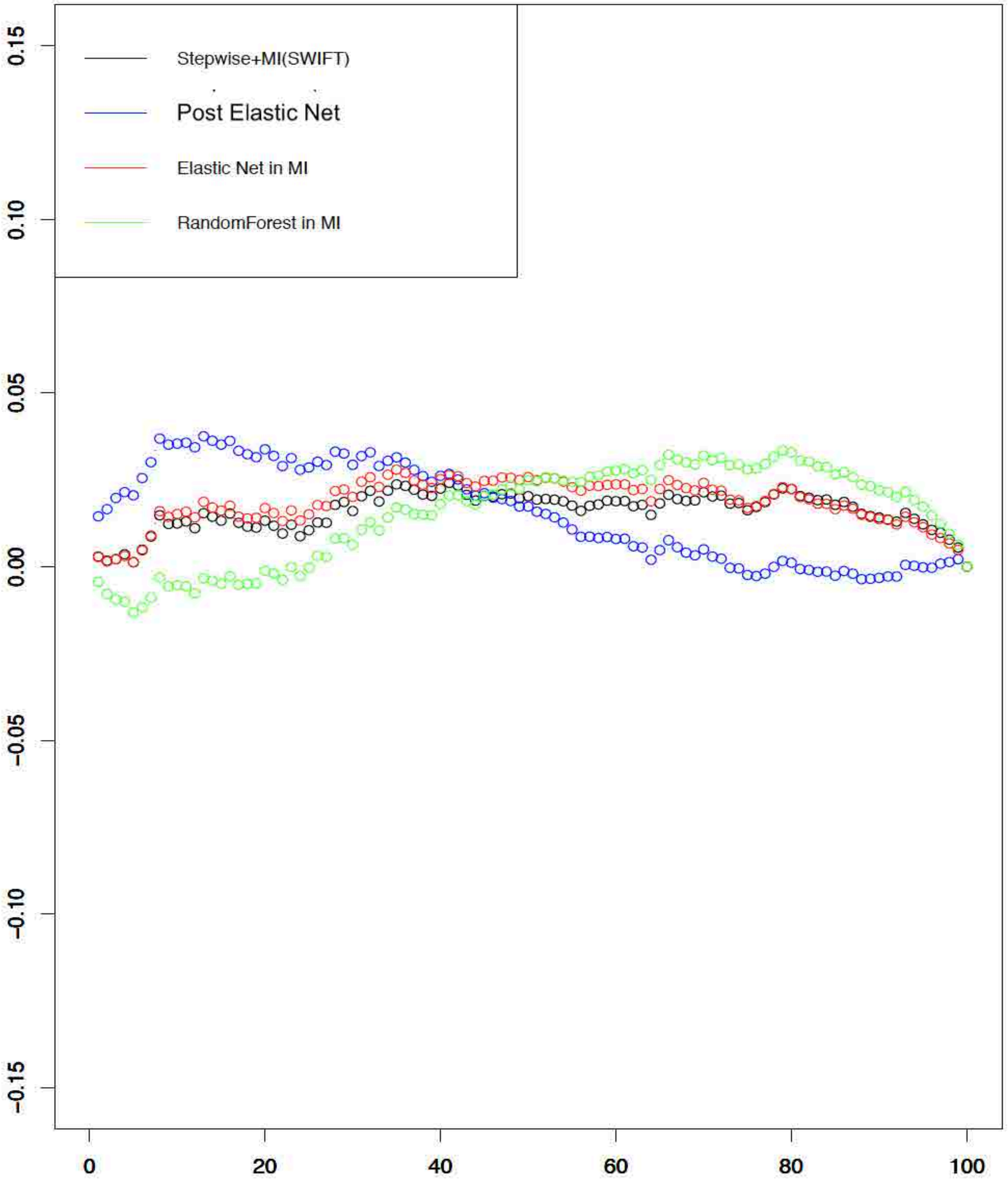


Average and Maximum bias of the poverty rate
0.00
0.05
0.10
0.02
0.04
0.079
0.137
0.014
0.03
0.068
0.117
0.01
0.023
0.013
0.028
Elastic in MI
Elastic Net
Post Elastic Net
Random Forest
Random Forest in MI
SWIFT
tag
Average
Maximum

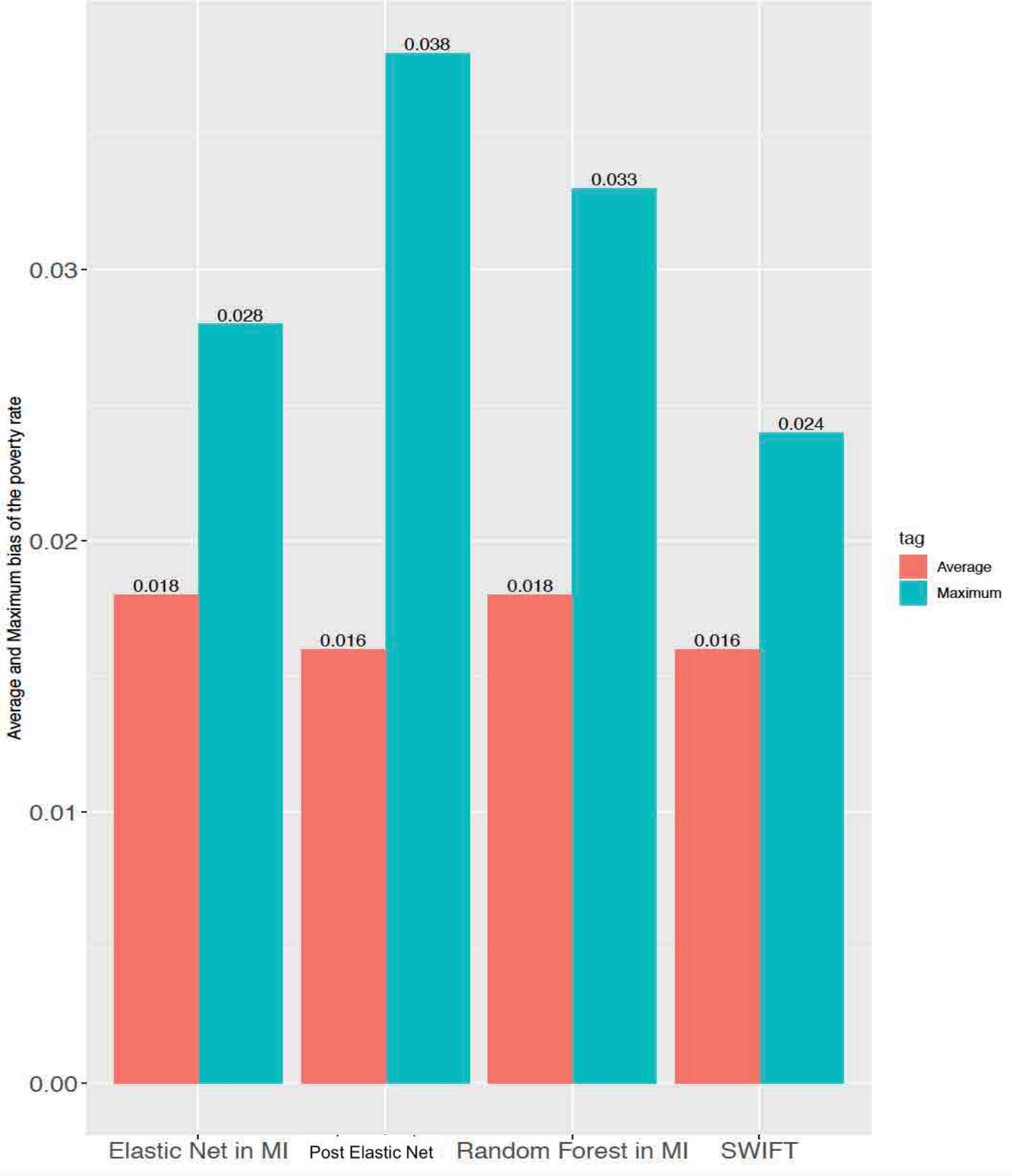
Rural Romania (2009 – Training data; 2012 – Testing data)

Comparison between Different Methods





Stepwise+MI(SWIFT)
Post Elastic Net
Elastic Net in MI
RandomForest in MI
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
20
40
60
80
100

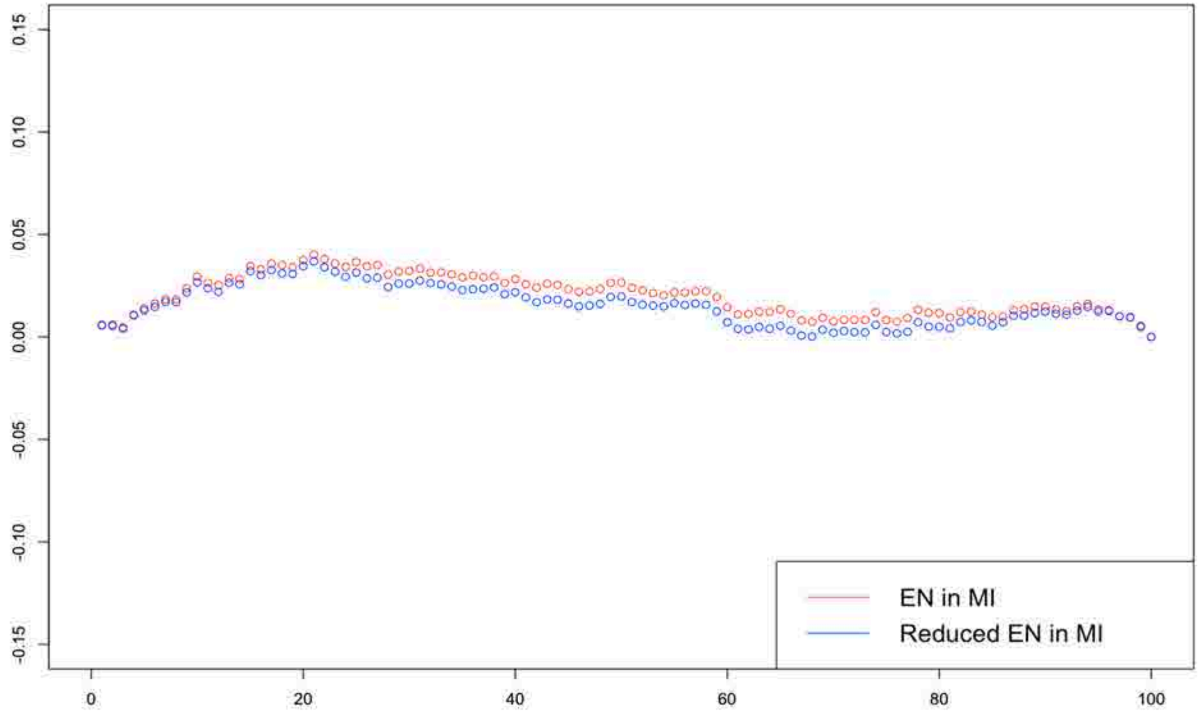

0.038
0.033
0.028
0.024
0.018
0.016
0.018
0.016
0.03
0.02
0.01
0.00
Average and Maximum bias of the poverty rate
tag
Average
Maximum
Elastic Net in MI
Post Elastic Net
Random Forest in MI
SWIFT

Further modifications – additional variable selections

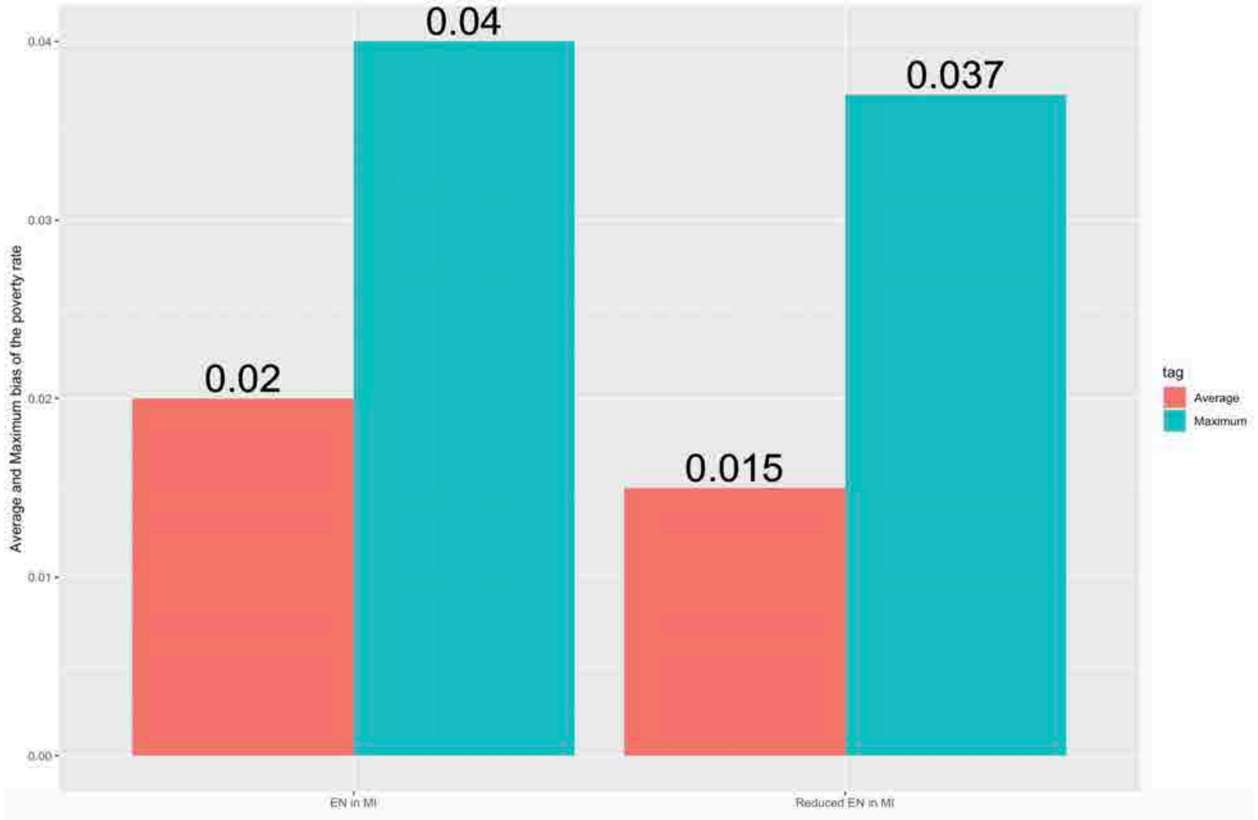
- Both EN in MI and RF in MI work well but they might need lots of variables
  - EN in MI with statistical significance tests (Reduced EN in MI) – After selecting variables by EN, drop variables whose coefficients are not statistically significant at 5%
  - RF in MI with variable importance tests (Reduced RF in MI) – After selecting dummies (or variables) by RF, 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

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



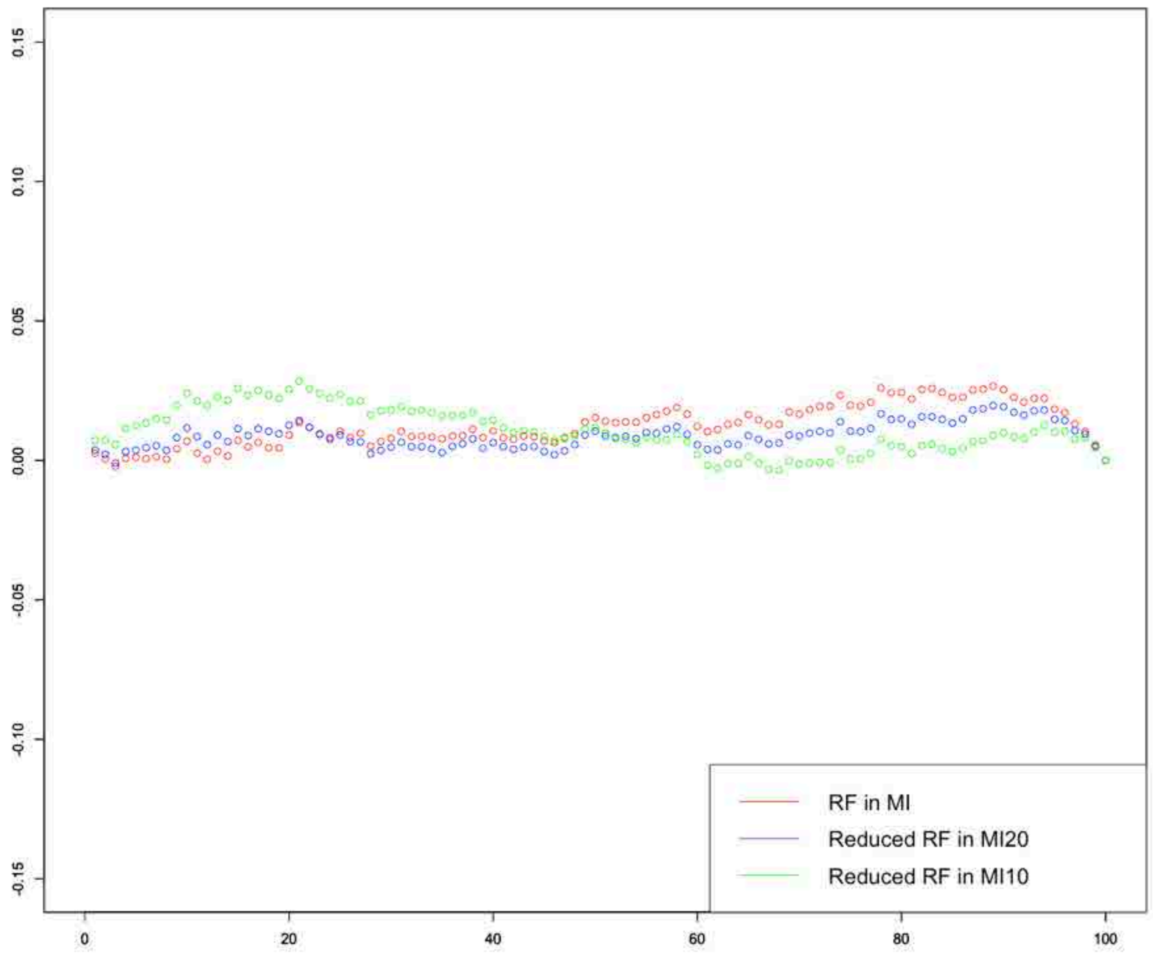
"Reduced EN in MI": Elastic Net with variable selection

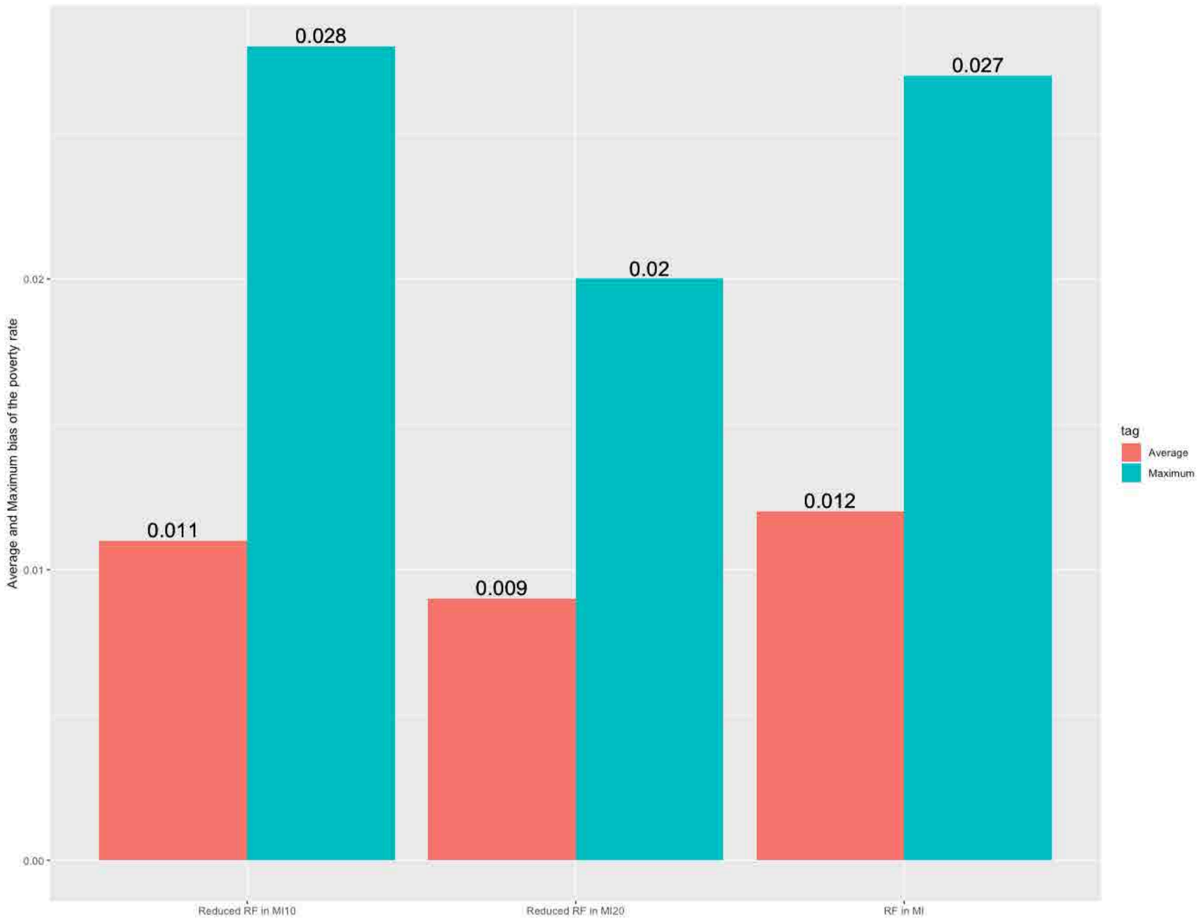


EN in MI

Reduced EN in MI

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



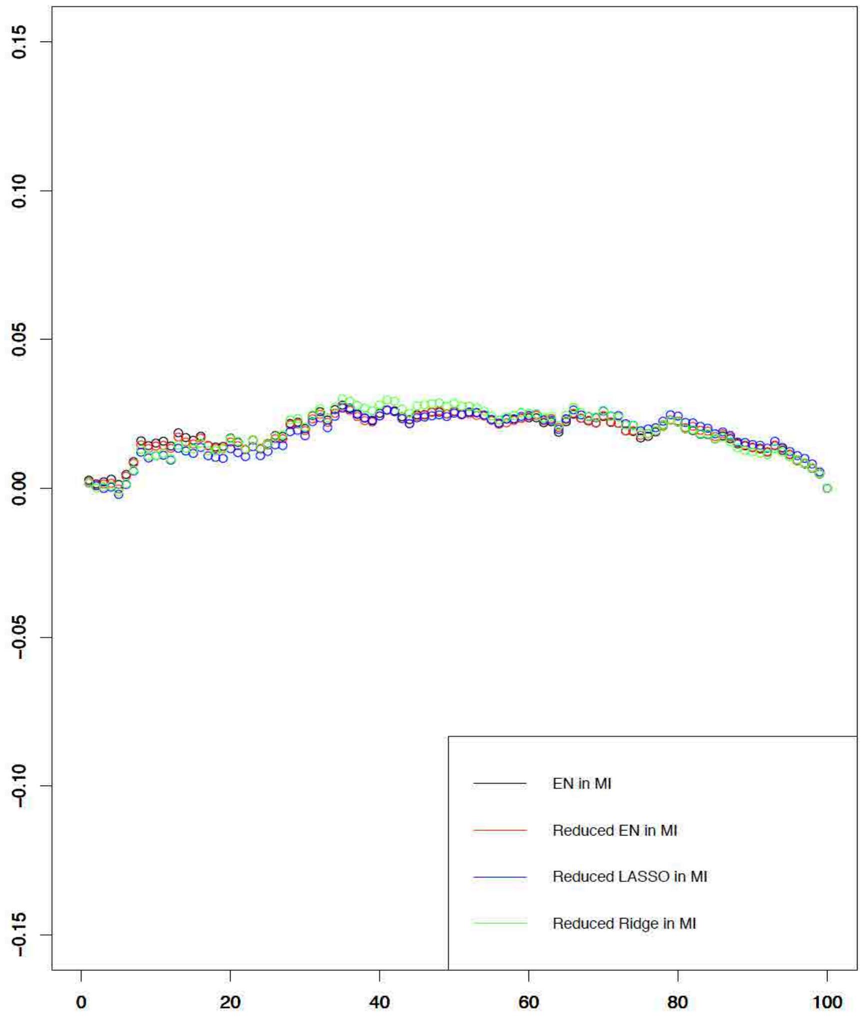


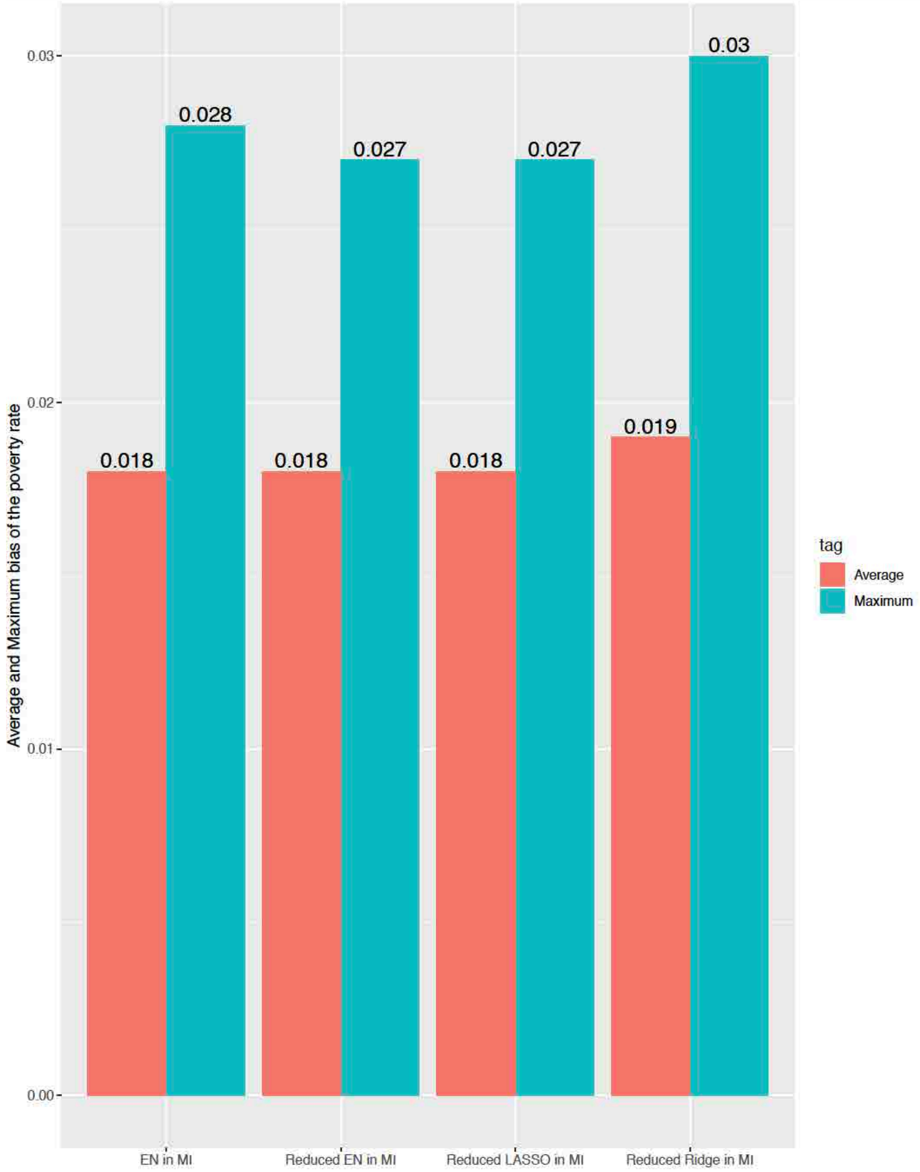
Reduced RF in MI with 10

Reduced RF in MI with 20

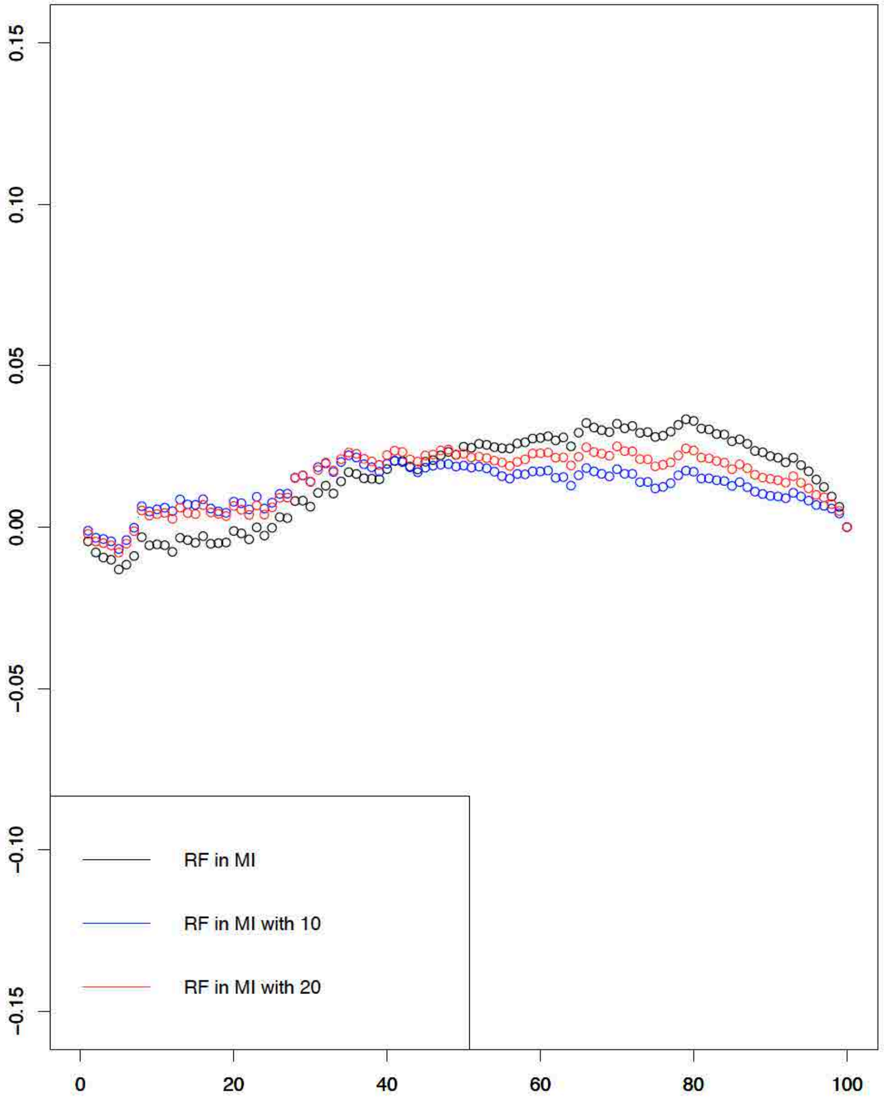
RF in MI

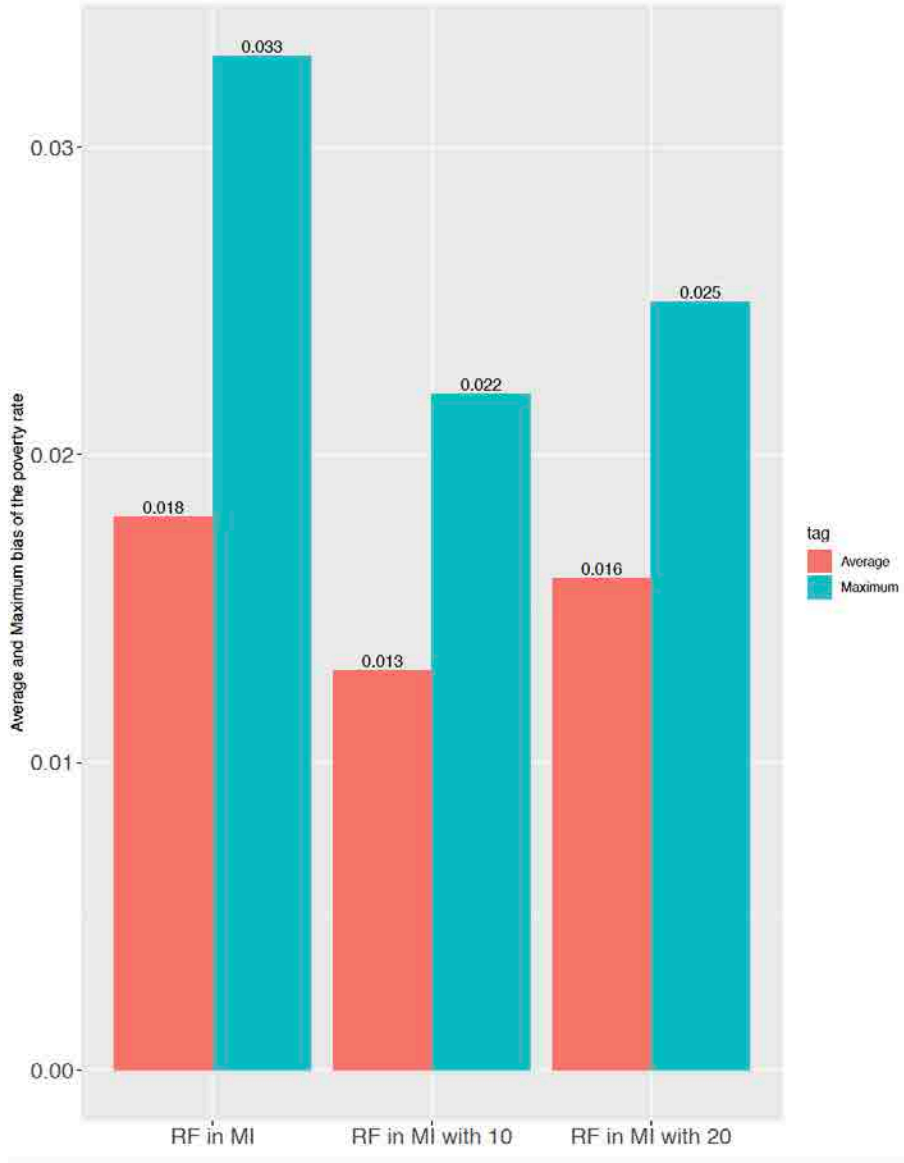
Elastic Net in MI with variable selection (Rural Romania 2011 to 2012)



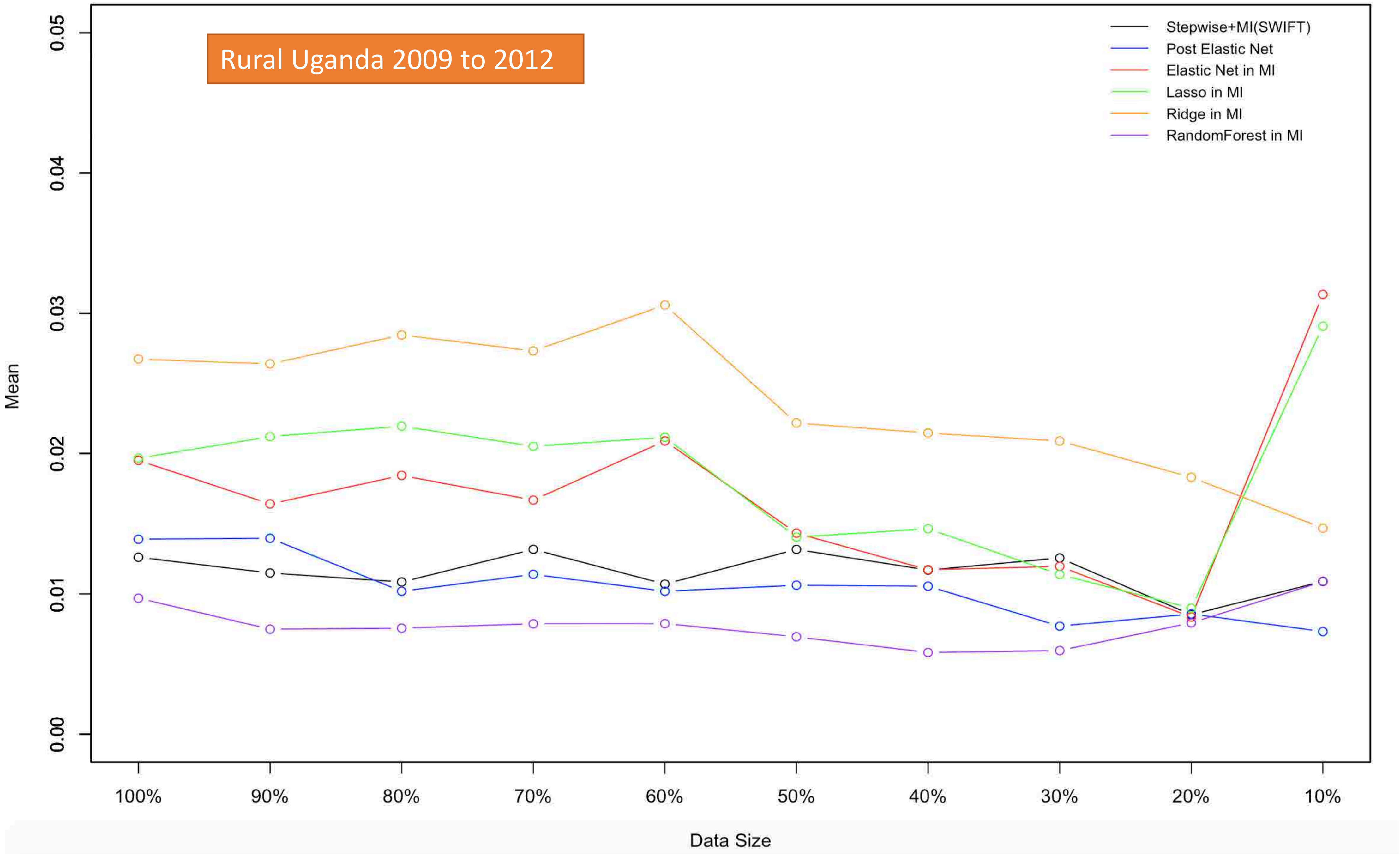


Elastic Net in MI with variable selection (Rural Romania 2009 to 2012)



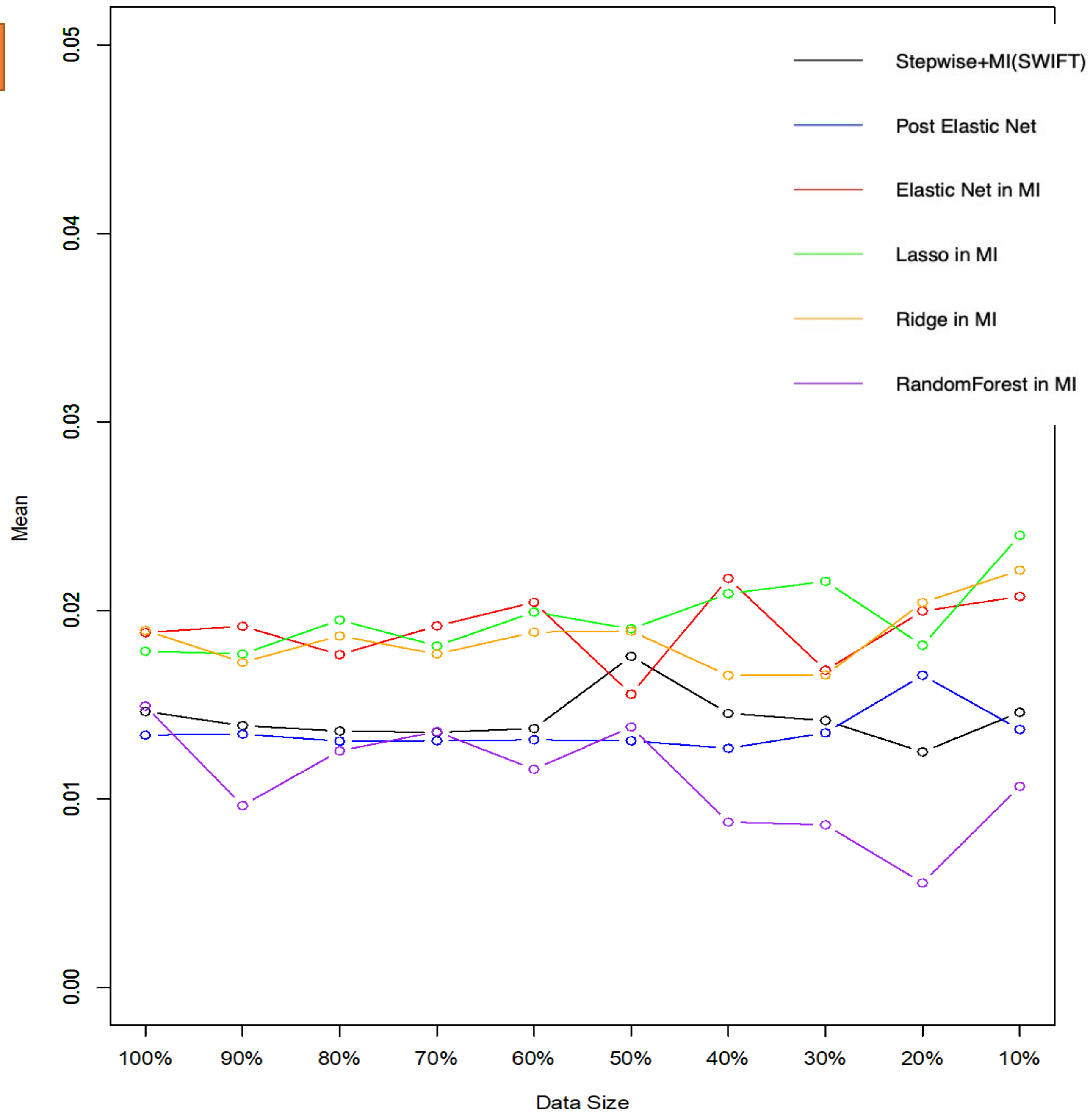


Effect of the size of training data on the performance

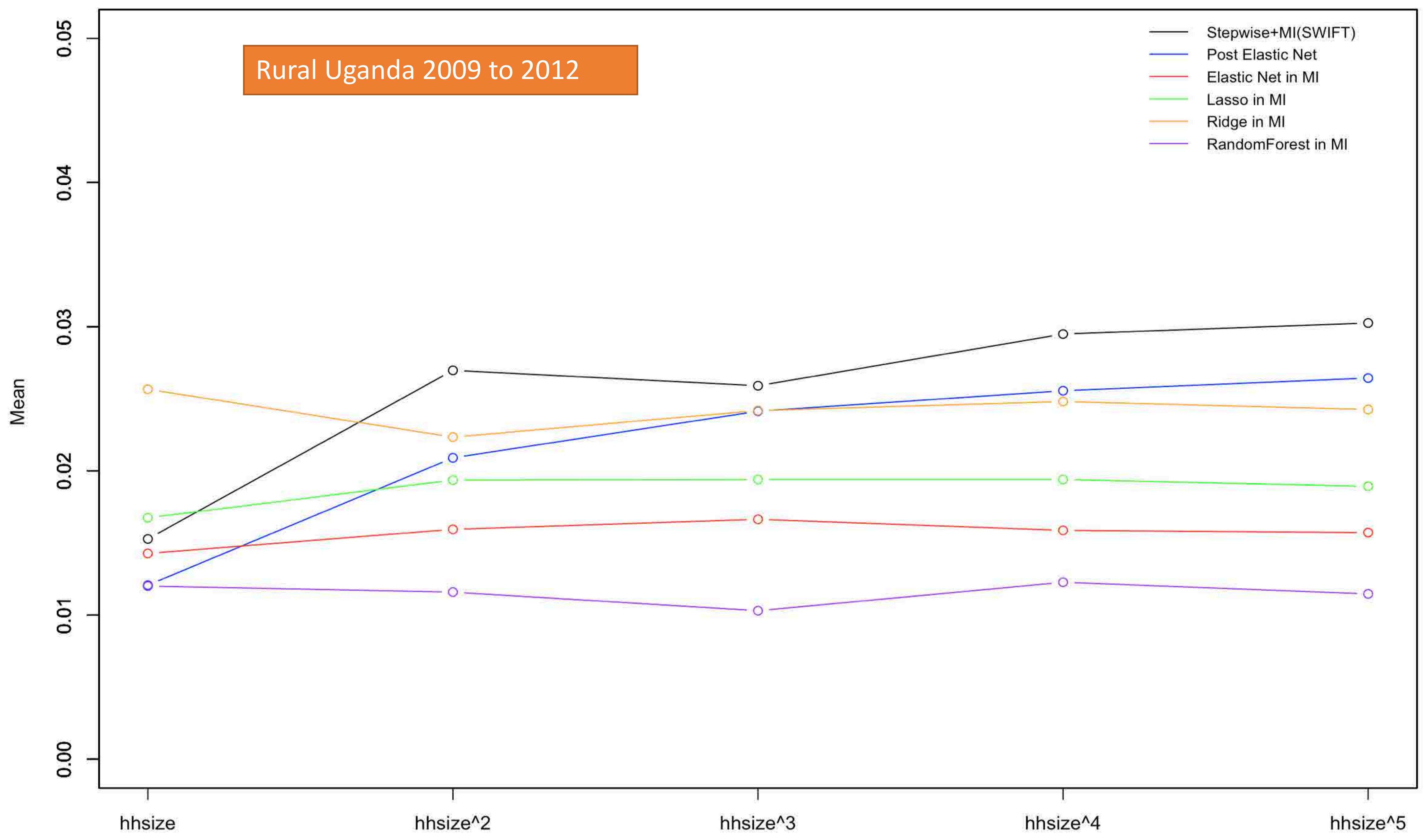


Rural Uganda 2009 to 2012
Stepwise+MI(SWIFT)
Post Elastic Net
Elastic Net in MI
Lasso in MI
Ridge in MI
RandomForest in MI
Mean
0.00
0.01
0.02
0.03
0.04
0.05
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
Data Size

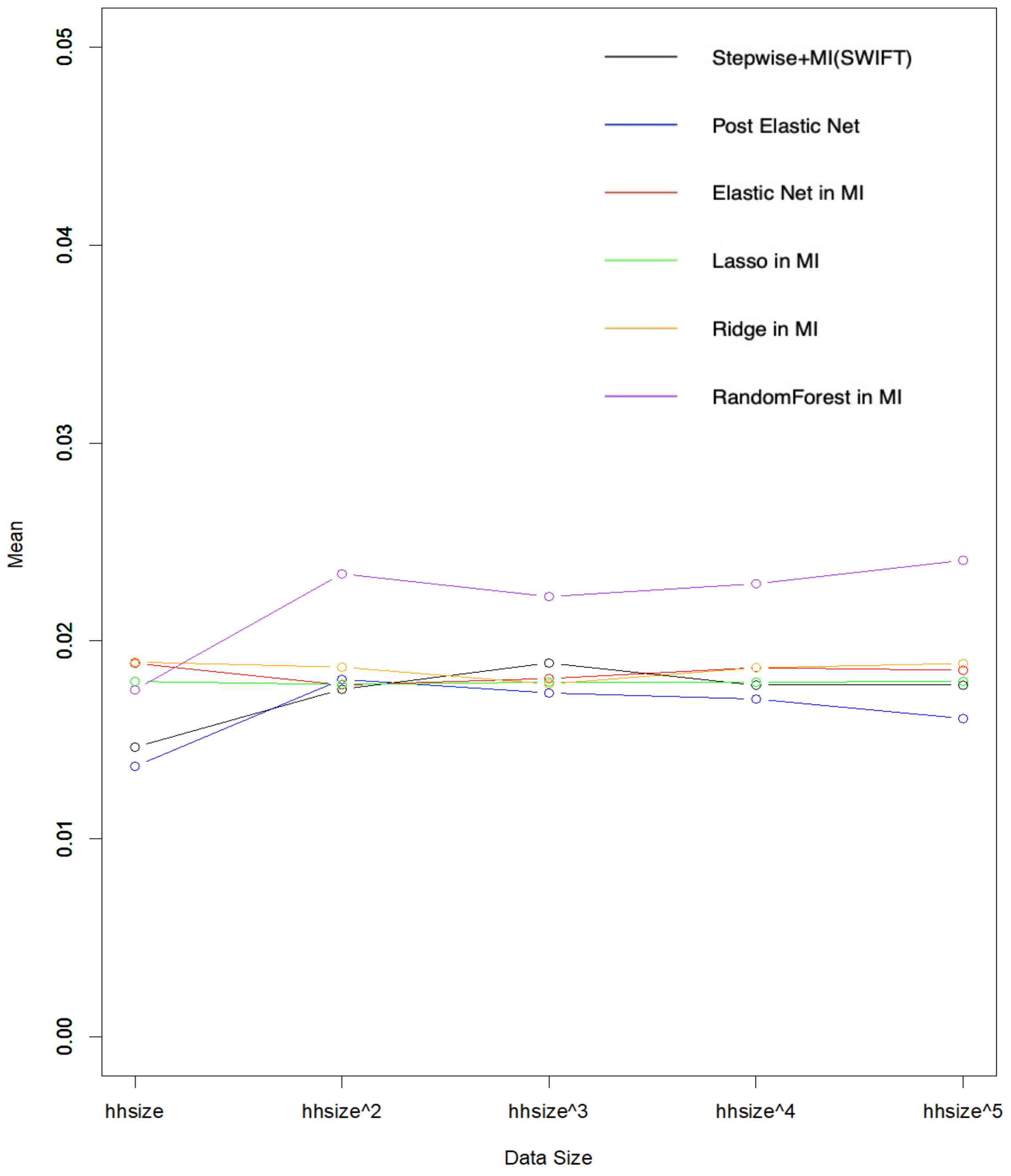
Rural Romania 2011 to 2012



Effect of the strong collinearity on the performance


Rural Uganda 2009 to 2012
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

Rural Romania 2009 to 2012



Summary of assessments

Approaches	Variable selection	# of variables	Coefficients	Error terms	Bias	Multi-collinearity	Small samples
EN	EN	large	EN	None	large		
RF	RF	large	RF/OLS	None	large		
Stepwise+MI	Stepwise	small	OLS	MI	small	weak	Fair
Post EN	EN	small	OLS	MI	small	weak	Fair
EN in MI	EN	large	EN	MI	small		
RF in MI	RF	large	RF/OLS	MI	small		
Reduced EN in MI	EN	small	EN	MI	small	Robust	weak
Reduced RF in MI	RF	small	RF/OLS	MI	small	Fair	Fair

Next steps

- Evaluate all approaches with their performance in estimating the shared prosperity index
- How to identify good approaches from training data only
- Theoretical evaluation of each approach (stepwise + MI; Post EN; Reduced EN in MI; and Reduced RF in MI)
- More details on Random Forest (RF)