

Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assessment

Thomas Pave Sohnesen¹ and Niels Stender²

¹Independent Consultant, Copenhagen, Denmark

²Innohead, Copenhagen, Denmark

Random forest (RF) is in many fields of research a common method for data-driven predictions. Within economics and prediction of poverty, RF is rarely used. Comparing out-of-sample predictions in surveys for the same year in six countries shows that RF is often more accurate than current common practice (multiple imputations with variables selected by Stepwise and Lasso), suggesting that this method could contribute to better poverty predictions. However, none of the methods consistently provides accurate predictions of poverty over time, highlighting that technical model fitting by any method within a single year is not always, by itself, sufficient for accurate predictions of poverty over time.

许多研究领域把随机森林作为一种用于数据预测的常用方法。在经济学和贫困预测中，随机森林却很少用到。通过对比 6 国同一年间由样本得出的调查预测，发现随机森林往往比当前通用的实践方法更准确（通用方法使用逐步回归法和套索算法筛选出的变量错误较多）。这表明使用随机森林得出的贫困预测更加准确。然而，随着时间推移，没有一种方法能够一直连续不断地给出准确的贫困预测，这着重说明：任何单一方法的技术模型拟合仅凭一己之力都不足以得出连续一年的准确贫困预测。

KEY WORDS: prediction methods, poverty, tracking poverty, Machine Learning, random forest, linear regression models

关键词: 随机森林, 贫困, 模型拟合, 套索算法, 逐步回归

Introduction

Poverty is a key policy concern in most countries, and in most countries, particularly in those with high levels, poverty is measured with consumption expenditures. Unfortunately, consumption expenditures are relatively complex to measure and monetary poverty is, therefore, considered complex and expensive to assess. For these reasons, poverty status is often assessed with poverty proxies,

with these proxies being indicators that are highly correlated with consumption and poverty, but easier to observe and collect data on. Ethiopia is an example of a country in such a situation. For the trend in poverty, Ethiopia relies on household consumption surveys every five years; however, ideally Ethiopia would have poverty numbers at a higher frequency, which could be attained through poverty proxies. Further, it has been proposed to address poverty in Addis Ababa by providing the poor with a cash transfer (World Bank Group, 2015). Here, households' poverty status is also assessed through poverty proxies. Hence, poverty proxies are in many cases a key instrument for both defining levels of poverty and designing instruments to reduce poverty.

The lessons from evaluations of tracking poverty over time through poverty proxies are positive and indicate that the approach is capable of tracking poverty over time (Christiaensen, Lanjouw, Luoto, & Stifel, 2012; Doudich, Ezzrari, Van der Weide, & Verme, 2013; Mathiassen, 2013; Sohnesen, 2014; Sumarto et al., 2007), but the evaluations also show that the method can be sensitive to the exact model applied. In many cases, the model used to predict over time is constructed in cross-section surveys from a single year. The assumption is that within-country variation in consumption and poverty proxies is informative on changes in consumption over time. This can be a strong assumption, which we return to below. In this context, model fit (measured by *r*-square) has been found to improve the accuracy of predictions (Christiaensen et al., 2012), though model fit has also been found insufficient by itself to guarantee high accuracy. In evaluations where year 1 is used to construct a model, and year 2 used to evaluate accuracy, around 50 percent of models produce predictions within the 95 percent confidence interval of the measured level in year 2 (Christiaensen et al., 2012; Dang, Lanjouw, & Serajuddin, 2014; Mathiassen, 2013; Sohnesen, 2014). Despite these evaluation efforts, including a focus on the selection of variables capable of tracking poverty over time, there is no consensus on an ideal variable selection for the prediction of poverty. Dang et al. (2014) provide a recent proposal with some formalization of a selection procedure. However, in their empirical evaluation using data from Jordan, they still find large variation in accuracy for different estimation models. In the case of Jordanian data, more elaborate models, including a larger range of variables, always perform better. They find that for estimations over only two years, in perfect settings, relatively elaborate models are needed for accurate predictions.

Mathiassen (2013) and Doudich et al. (2013) utilize several rounds of data and are, therefore, able to validate their model performance by predicting backwards and forwards. Doudich et al. (2013) seem to successfully track quarterly poverty in Morocco using labor market surveys. Similar efforts in Sri Lanka, however, were unsuccessful (Newhouse, Shivakumaran, Takamatsu, & Yoshida, 2014), arguably due to differences in questionnaire and sampling design. Hence, with lack of several rounds of data and clear guidance on selection of proxy variables, model fit is usually the preferred selection method, which is the aspect to which this article contributes.

In this article, we contribute to the existing literature on poverty proxies by considering an alternative algorithm for model selection and prediction of poverty status. The current common practice for estimating the trend in poverty relies on different variations of linear regressions. The imputation method developed for small area estimation by Elbers, Lanjouw, and Lanjouw (2003; known as the ELL method) is commonly used by the World Bank. The Multiple Imputation (MI hereafter) command in Stata is similar to ELL¹ and is also commonly used for survey-to-survey predictions of poverty. We evaluate the MI method as the current common practice, while also noting that the ELL and MI methods are very similar.

As an alternative to linear regression-based methods, we evaluate the random forest (RF hereafter) method (Breiman, 2001). The RF method is part of the Machine Learning literature and has been applied for predictions in a wide range of research fields. See, for instance, the review in Verikas, Gelzinis, and Bacauskiene (2011) that among other areas mentions: prediction of long disordered regions in protein sequences, classification of agricultural practices based on satellite imagery, spatially distributed measurements of environmental conditions, recognition of handwritten digits, and segmentation of video objects. It is notable that the method is commonly used in several areas of research, including the medical literature, but is largely absent in the economics literature. Varian (2014), an economist specializing in predictions, describes the advantages of Machine Learning as being better at predictions, while not able to do estimations and hypothesis testing. Machine Learning's better ability to predict seems rooted in better handling of nonlinearities, whereas, on the other hand, it works less well with linear variables.

Within the poverty prediction literature, the application of RF is very recent and still scant. In Indonesia, Otok and Seftiana (2014) find that an RF method is very accurate in identifying poor households eligible for social assistance packages, while an application in Mauritius uses RF to find poverty predictors and finds that RF predicts poverty accurately (Thoplan, 2014). McBride and Nichols (2015) analyze RF's predictive performance compared to existing regression-based models for developing proxy-means-test targeting models. Comparing out-of-sample accuracy in three countries (Bolivia, Timor-Leste, and Malawi), they find that quantile RF is not substantially better at predicting the overall poverty status of households. Quantile RF is, however, better at correctly estimating a poor household as poor, but it also has higher leakage (wrongly classifying a non-poor household as poor). The assessment is made for USAID, which in their valuation of methods punish errors in identifying the poor higher than other errors, and McBride and Nichols (2015) therefore conclude that RF can significantly improve out-of-sample performance by 2–18 percent.

To compare the accuracy of MI and RF, we apply both methods to two rounds of consumption expenditure surveys that measure poverty in six different countries. We implement two evaluations. First, we take one year of data and split it into two random samples. We generate a prediction model in the first half of the sample and evaluate the model's accuracy in the second half of the sample.

Second, we implement the same setup, but using two years of data, where year 1 is used for modeling and year 2 is used for evaluation of prediction accuracy. The first evaluation compares technical model fit and prediction only, whereas the second evaluates the prediction in the context of a changing economic environment.

We find that RF often has higher accuracy in predicting poverty, particularly at the rural/urban levels (as opposed to the national level). However, RF is not always more accurate, and for predictions at the national level, the differences in accuracy between methods are small. Hence, the very automated RF tool, which requires very little knowledge on behalf of the user, performs as well as or better than current common practice. However, none of the combinations of selection and estimation methods consistently predicts poverty accurately over time, highlighting that technical model fitting within one year can be insufficient for accurate predictions over time.

The Implementation of Methods section outlines the methodology applied, the Data section describes the data, and the Results section evaluates prediction outcomes, while the Conclusion section concludes.

Implementation of Methods

A technical comparison of RF to linear predictions is not straightforward, as the two approaches come from different strands of literature, and generally have little in common, except that they can both be used to achieve the goal of interest: predicting poverty. This section highlights some key differences between the linear regression-based imputations and RF, but, for an introduction and a more in-depth description on the methods themselves, see James et al. (2013) and Hastie, Tibshirani, and Friedman (2009) for RF and similar methods, and Stata’s description of its MI package. Table 1 lists the programs and settings utilized for modeling the predictions.

For the RF method, please consider the following example and outline of the procedure. RF consists of a number of decision trees; each tree consists of a number of decision nodes. In each node of a decision tree, data are split according

Table 1. Programs and Settings Used for Modeling Predictions

Estimation Method	Program	Settings	Variable Selection
MI	Stata—MI package	The predictive mean matching option and 100 repetitions to impute log consumption	Stepwise command with a .01 significance level for addition to the model and .1 for staying LASSO using Stata’s LARS package
RF	Python, Anaconda package	The standard settings, with 500 trees and a minimum of four observations in each leaf	Entropy loss function Gini loss function

to how well a variable can predict poverty. As an illustrative example of a decision tree, consider the following: Whether a household is larger than three predicts poor and non-poor households with the lowest error, and household size 3 and above therefore becomes the first splitting node. Hence, the sample is split in two groups, those with a household size above 3 and those size 3 and below. In the first group (those with household size 3 and below), household ownership of a TV is then the variable with the lowest prediction error, and it becomes the next splitting node. Households with a household size above 3 having a household head working as a farmer is the variable with the lowest prediction error, and this variable becomes the splitting variable. The sample is now split in four groups, and the process continues until a minimum of observations are left in each group (called a leaf in the literature). Utilizing the entire sample and all possible splitting variables would likely lead to overfitting, and each tree, therefore, only relies on a subsample of observations and each node only relies on a subsample of variables. To gain robustness, RF in turn relies on a large number of trees. The assessment of prediction error—or loss function—can vary, and the two most common ones (Gini impurity and entropy) are both evaluated below.

To obtain RF predictions, the trees are built based on the steps outlined below and implemented in the Anaconda package in Python:

1. Split the data set in half, into a learning and an evaluation sample.
2. Draw a random sample of two-thirds of the observations with replacement for each tree from the learning data set.
3. Grow a tree based on the selected random sample:
 - 3.1. For every node in the decision tree, randomly sample the square root of the total number of variables (called features within the literature) as potential splitting variables.
 - 3.2. Select the variable used for splitting in each node that has the lowest Gini impurity or entropy value.
 - 3.3. Grow each tree until a split leads to fewer than four observations in a leaf, or until no split leads to further decrease in Gini impurity or entropy value.
4. Repeat Steps 2 and 3 a total of 500 times.

Based on these 500 trees, a main output is the share of times (called the score function) that a household is found to be poor or not. Hence, each tree (which is a prediction model by itself) predicts the status for each household, and the score function is the average prediction for 500 different trees (or prediction models). The large number of trees is, in part, why some argue that RF is a more robust predictor, as it does not rely on a single prediction model. The head count is obtained by taking the weighted mean of the score function for all households.

The predictions from RF are compared to predictions from two variable selection methods, both estimated by MI. There are many different variable selection methods to choose from; see, for instance, Castle, Qin, and Reed (2009) for comparison of 21 different ones. Stepwise and LASSO are used in this evaluation. The reliance on standardized and automated variable selection

methods provides the cleanest comparison, with minimal involvement from researchers. Further, some also use RF as a variable selection method. Based on the 500 trees, you can gauge which variables are more important than others by counting the number of times each variable is selected for a tree. This is known as the importance score and can be used as a selection method by itself. The variable selection from RF is compared to Stepwise and LASSO by comparing predictions from MI and RF, based on the 25 variables with the highest importance score. This latter evaluation, with only 25 variables, also shows whether RF is a useful tool to track poverty in a realistic setting, where cost of data collection is a critical aspect.

In the most common applications, even the explanatory variables vary between these two approaches. MI usually predicts the consumption distribution from which the poverty head count is calculated, whereas RF usually predicts a poor/non-poor dummy. As such, the MI method, unlike the application of RF, can produce additional welfare indicators that rely on the distribution of income or consumption, as depth of poverty and a number of different inequality measures. Both methods are, however, capable of predicting both explanatory variables.^{2,3}

Though both approaches utilize bootstrapping elements, they also differ in this aspect. MI relies on one model with a set of selected variables, whereas RF estimations are based on a large set of different models. MI estimates distributions of coefficients and errors and makes draws from these distributions in each replication, but always uses the same underlying model with the same set of explanatory variables.

The prediction model is at the household level, even though the outcome indicator of interest—poverty head count—is a national population statistic. Reflecting this, evaluations were undertaken with household, population, and no weights. In Stata, neither the Stepwise nor the LARS package allows using weights at the variable selection stage. The RF algorithm, on the other hand, allows weights that affect the splitting of variables and therefore alter the loss function. The weights in RF are not used when samples of observations are drawn for each tree, whereas MI, on the other hand, does allow the use of weights at this stage.

Data

The evaluation utilizes data from six countries: Albania, Ethiopia, Malawi, Rwanda, Tanzania, and Uganda. Poverty is defined as monetary poverty, and the evaluation follows the definition as applied by the government and statistical agencies in each of the countries. In all countries, poverty is defined based on consumption expenditure aggregates.

These countries are considered to have comparable consumption data for at least two years, but also represent good variation in number of years between surveys and level and trend in poverty.⁴ This allows testing of different estimation models in different circumstances.

The level of poverty varies from 12.5 percent in Albania to 50.2 percent in Malawi in the most recent year (Table 2). Over four and five years, respectively, Rwanda and Uganda both experienced substantial and significant poverty reduction in both rural and urban areas. In both cases, poverty was reduced more in rural areas than in urban areas, while urban areas continued to have lower levels of poverty. In Albania, poverty fell at the national level over three years, but was driven by significant poverty reduction in rural areas, while urban poverty did not change significantly. In Malawi, urban poverty fell substantially over five years, though it was insufficient to significantly change poverty at the national level. Finally, in Tanzania, poverty in rural areas increased by 5 percentage points over just two years, while poverty did not change significantly in urban areas.⁵ These diverse trends in poverty test whether the approaches are capable of replicating poverty at different levels and with different trends.

Ethiopia, Malawi, Albania, and Rwanda are cross-section household surveys with large samples. The first year of data for Uganda and Tanzania is a large, cross-section sample, whereas the second year is a panel following a part of the first year of data. In both panels, split and moving households were tracked.⁶ Even though split and moving households were tracked, panel surveys can “age” and/or become a nonrepresentative sample. This happens if tracking of split and moving households is incomplete (Himelein, 2014). Such changes to the sample would further challenge the assumption of steady relationships over time

Table 2. Poverty Trends

	Poverty Rates (%)			Trend in Poverty
	National	Urban	Rural	
Albania 2005	18.5	11.2	24.2	Significant poverty reduction at national level driven by large poverty reduction in rural areas
Albania 2008	12.5	10.2	14.7	
Ethiopia 2010	29.6	25.7	30.4	
Malawi 2004/2005	52.0	25.2	56.0	No significant change at national level, but significant poverty reduction in urban areas
Malawi 2009/2010 ^a	50.2	17.0	56.4	
Rwanda 2006	56.7	28.5	61.9	Significant poverty reduction at national level and in urban and rural areas
Rwanda 2011	44.9	22.1	48.7	
Uganda 2005/2006	29.6	12.8	32.8	Significant poverty reduction at national level and in urban and rural areas
Uganda 2009/2010	23.0	8.0	25.7	
Tanzania 2008	15.0	6.1	17.5	Significant increase in poverty at national level, driven by changes in the rural areas
Tanzania 2010	17.9	5.2	22.3	

Source: Authors' calculations.

Note: Numbers might not match official numbers, as sample at times varies from those used in official reports due to missing observations in some explanatory variables.^aIn Malawi, the sampling for urban areas was changed from only sampling four major urban areas to sampling urban areas in all districts. The change can jeopardize comparability and poverty rates for the four major urban areas, as well as for all urban areas presented. Appendix Table A1 shows full references to data sources.

between proxy variables and outcome variables, as the model was built on a representative sample of the country, but applied to only a specific subsample of the population. In the case of Uganda, 16 percent of households and 20.5 percent of individuals were not tracked in 2010, whereas in Tanzania, only 3 percent of households and 10 percent of individuals were not tracked successfully over the two years between surveys. In a different panel data set from Uganda, Kasirye and Ssewanyana (2010) find that the observed 25 percent attrition does lead to significant bias in consumption regression coefficients.

In all six countries, data are collected from the following eight sections of the questionnaire: demographics, education, food consumption, nonfood consumption, housing quality, ownership of durable goods, employment, and location. All categorical variables have been turned into dummies, and extremely skewed variables were excluded.

Results

To compare the two approaches, the results are based on a comparison of the prediction accuracy of the following six models:

1. RF using Gini impurity loss function.
2. RF using Entropy loss function.
3. MI with Stepwise variable selection.
4. MI with Lasso variable selection.
5. MI with 25 variables based on importance score from RF
6. RF with 25 variables based on importance score from RF.

To evaluate more closely the methodological aspects only, the first section shows results based on predictions within the same year based on a randomly split sample, where one half is used to build a model and the other half is used to evaluate accuracy of predictions. The second section evaluates prediction accuracy over time based on two years of data, where year 1 is used for modeling and year 2 is used to evaluate prediction accuracy. The latter illustrates a realistic setting in which the economic environment changes over time.

The Gini impurity loss function is not to be confused with the Gini coefficient measure of inequality, which might be more familiar to many. The Gini impurity loss function is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. For further details, see the documentation in the python package.

Predictions Within Same Year

Comparing prediction accuracy from the linear regression-based models and RF, using one year of data randomly split into two halves, shows that both approaches do well at the national level, though RF models are generally more accurate.⁷ Mean square error (MSE) between the predicted poverty rate and the

Table 3. MSE of Poverty Predictions for Different Variable Selection Methods and RF Loss Functions

Location	National		Urban/Rural		National		Urban/Rural	
Estimation Method	MI		MI		RF		RF	
Variable Selection/Loss Function	Stepwise	LASSO	Stepwise	LASSO	Gini	Entropy	Gini	Entropy
Country/column	1	2	3	4	5	6	7	8
Ethiopia	6.99	.01	8.75	22.76	2.42	2.69	1.97	2.51
Malawi	1.52	.84	23.57	21.18	.32	.39	1.72	1.91
Uganda	2.41	4.97	2.32	3.94	1.83	2.16	5.72	4.04
Albania	.77	2.46	1.37	2.73	3.61	3.60	3.68	3.75
Tanzania	3.17	2.50	10.86	3.85	1.26	2.05	1.35	2.61
Rwanda	1.98	1.76	3.60	2.23	.79	.72	1.03	.66
Average	2.81	2.09	8.41	9.45	1.71	1.94	2.58	2.58

Source: Authors' calculations.

Note: Table shows MSE for poverty predictions out of sample, for same year, based on a split sample. MSE for urban/rural is a simple average of MSE for urban and rural.

measured poverty rate for RF and MI estimations is on average very similar. However, looking at MSE for urban and rural areas combined, RF has notably higher accuracy in four out of six countries, and is better at the mean (columns 7 and 8 vs. 3 and 4, Table 3). The results are consistent with RF being perceived as a more robust method. The pattern of RF and MI being similar at the national level, with higher accuracy at the urban/rural levels for RF, is observed for both variable selection methods (Stepwise vs. Lasso), and for different loss functions for RF (Gini vs. Entropy) (Table 3).

The results are not systematically better for MI using Lasso or Stepwise, though accuracy can be model dependent. See, for instance, predictions for Ethiopia, where Lasso is very accurate, whereas Stepwise is not (columns 1 and 2, Table 3). High accuracy at the national level is also not a guarantee for high accuracy at lower levels. In Malawi, for instance, poverty is accurately predicted with both Lasso and Stepwise at the national level, but poorly predicted at urban and rural levels. RF does not seem sensitive to variation in loss function, as both Gini and entropy loss functions lead to very similar results in all countries (Table 3).

As mentioned, RF can also be seen as a variable selection method. Variables that enter more trees can be seen as stronger predictors of poverty than those that enter fewer trees. The share of trees each variable enters into is called the importance score. You can gauge the relative strength of RF as a variable selection method, as opposed to an estimation method, by selecting the variables with the highest importance score and predicting poverty with MI. Table 4 shows the results when RF and MI are restricted to the 25 variables with the highest importance score. A model with 25 variables would, by most, be considered a small model, particularly in contrast to the large models selected by both Lasso and Stepwise. Stepwise and Lasso find valuable information in a large set of variables, and have selected 81 and 132 variables on average across the six

Table 4. MSE of Poverty Predictions, Large and Reduced Models

Country	RF Full Model	RF Top 25	MI Top 25	MI Lasso
Ethiopia	2.42	.77	.31	.01
Malawi	.32	.72	.24	.84
Uganda	1.83	2.34	1.16	4.97
Albania	3.61	2.99	2.87	2.46
Tanzania	1.26	1.55	.53	2.50
Rwanda	.79	.09	1.13	1.76
Average	1.71	1.41	1.04	2.09

Source: Authors' calculations.

Note: Table shows MSE for national poverty predictions out of sample, for same year, based on a split sample. MI results shown are for a model selected with Lasso in Stata with population weights. RF results are based on a Gini loss function and use population weights.

countries (Appendix Table A4). Many would find such large models over fitted. Further, for tracking of poverty over time, the net gain through lower data-collecting costs might be limited with so many variables. In Albania and Tanzania, for instance, Lasso selects 140 and 152 variables in surveys with around 1,400 and 1,700 observations, respectively. Stepwise selects 70 and 92 variables for the same data. For Stepwise, smaller models can be obtained by lowering the significance levels (results are based on a .01 significance level for addition to the model and .1 for staying), but the Lasso command in the Stata LARS package does not have such a tuning option.

Predicting poverty with MI based on the 25 variables with the highest importance score from RF leads to improved accuracy in four out of six countries and a lower average error, indicating that RF variable selection might have some advantages (Table 4, columns 4 and 5) compared to Stepwise and Lasso, though all models perform well.

Restricting RF predictions to a limited set of variables also illustrates a realistic setting in which poverty is tracked with smaller surveys that only collect data on a limited number of aspects.⁸ There is no loss of accuracy in predicting poverty using the restricted model with only 25 variables compared to the full model without restrictions (Table 4). There are minor variations in each country between the full and restricted models, with the restricted models slightly more accurate on average. Hence, the RF approach also predicts poverty accurately using only a small model, and can therefore be a real alternative for tracking poverty with predictions.

Poverty is a population statistic, and all final estimates are calculated using appropriate population weights from the surveys. The underlying model is generally a household model, as all members of households are given the same poverty status. As mentioned in the Introduction section, applying household weights in RF alters the loss function, thereby influencing the estimation models, whereas neither Stepwise nor Lasso in Stata have this option. Appendix Table A3 shows that prediction accuracy does not systematically depend on the type of weight or whether weights are applied at all, when using MI, though there is some notable variation in accuracy for the same country and model. For RF, this

sample of countries indicates that population weights marginally improve performance over household weight or no weights, though the variations in accuracy are too small to draw firm conclusions.

Predictions Over Time

Results thus far have focused on prediction within the same time period, allowing a cleaner comparison of different approaches; this part focuses on predictions over time. Predictions over time show that none of the methods are consistently accurate enough (Table 5). Here, accurate enough is seen as national estimates that fall within the 95 percent confidence interval of the measured level in year 2. The boldfaced results are national poverty estimates that are outside the measured 95 percent confidence interval of the survey poverty head count. For RF, three out of five estimates are within this threshold, while five out of ten are within this threshold for MI. Hence, only about half of the models predict poverty accurately.

The relatively favorable results found for RF in urban/rural areas in three countries are not found for predictions over time, and no selection or any prediction method consistently performs better than the others. It is noteworthy that the variation in prediction error between Lasso and Stepwise in some cases is quite large. This illustrates how results can be sensitive to the exact model applied.

The two countries with panel data sets (Uganda and Tanzania) have especially poor predictions. Predictions in these two countries are, in fact, the least accurate of all countries, which could indicate that potential aging of the sample leads to additional inaccuracy. Disregarding the panel data sets, the accuracy is fairly good, as only Rwanda falls outside the 95 percent confidence interval, and only marginally so. A country like Albania, which went through an economic crisis including large-scale migration (Carletto & Kilic, 2009), is, in fact,

Table 5. MSE of Poverty Predictions Over Time

Country	National			Urban/Rural		
	MI Lasso	MI Stepwise	National RF	MI Lasso	MI Stepwise	RF
Malawi	.06	.28	1.41	8.31	8.60	24.86
Uganda	22.69	.08	28.13	18.25	6.76	22.19
Albania	.44	.32	2.66	8.33	6.87	10.65
Tanzania	37.63	19.73	17.89	30.90	16.77	18.55
Rwanda	3.32	3.46	.52	10.63	9.38	4.10
Average	12.83	4.77	10.12	15.28	9.67	16.07

Source: Authors' calculations.

Note: Table shows MSE for poverty predictions out of sample, for same year, based on a split sample. MSE for urban/rural is a simple average of MSE for urban and rural. Boldfaced estimates are outside the 95 percent confidence interval of the measured level of poverty in year 2. Boldfacing is only applied to national estimations, and not urban/rural estimations.

still predicted accurately. For RF only Tanzania and Uganda are predicted poorly; all other countries are within the 95 percent confidence interval.

The mixed performance of the prediction models that are based on fitting of data across households within a single period mirrors the results in other evaluations (Christiaensen et al., 2012; Mathiassen, 2013; Sohnesen, 2014). The assumption is that the structural relationship between consumption and proxies within a year can be used to predict the movements in consumption over time. There are no guarantees that such an assumption is true, particularly as the economic environment and prices change over time. The evaluations have also shown that deselecting variables that are not “stable” over time are critical, even at the cost of within-model performance, as a single variable can throw off model predictions, while at the same time many different model specifications can produce very similar accurate results (Christiaensen et al., 2012; Sohnesen, 2014). Sohnesen (2014), for instance, highlights two examples of durable assets that could not be expected, *a priori*, to be stable over time due to changes in prices and technology. The first example is cell phone ownership in Malawi, which increased from 4 to 41 percent of households over five years. The second example is cassette players. In both cases, having a cell phone or a cassette player was associated with higher consumption in 2004, particularly for cell phones, which was only a privilege of the rich. Five years later, 4 out of 10 households have a cell phone, and it is doubtful that these households had the same high consumption level as the 4 percent among the richest had five years earlier. Cassette players were also associated with higher consumption, but due to technological development, this need no longer be the case. In general, the correlation between consumption and a certain variable can change over time, which would not be reflected in the predictions. Christiaensen et al. (2012) show how predictions fail in Vietnam when rice (a very common food item) is included in the model. Doudich et al. (2013), in their successful predictions of poverty at a quarterly basis for several years, have the advantage of several rounds of data, allowing them to test the models by predicting both backwards and forwards in time.

Conclusions

RF has been used successfully as a data-driven prediction method in different fields of research. This article shows that RF is also a good predictor of poverty; in some cases, it is a better predictor than current commonly applied methods. RF is not the most accurate method in all cases; it is, however, more robust and does not make as large prediction errors at rural/urban levels as commonly applied linear regression models. This is fully in line with the RF literature that emphasizes that the reliance on many models makes RF a more robust predictor. The reliance on multiple models is an integrated part of the RF approach, and though such iterative approaches and resilience testing by applying multiple models could be applied to linear regression methods, it is rarely done in any systematic way. The lack of any industry standard and lack of readily available

programs for such iterative approaches within a linear regression framework likely hamper widespread and systematic use. RF is simple and automated to use and could be used instead of, or as a complement to, other methods currently in use.

Existing evaluations show that you can fit a model on cross-section data and use it to predict poverty over time, but also that technical model fit, in itself, is an insufficient decision criterion. Large variations in predictions for different models indicate that some models are incapable of taking price and economic changes into account. This evaluation supports these observations, as only about half of the models accurately predict poverty over time. However, the prediction errors are mostly found in Tanzania and Uganda, both—and the only—countries that rely on panel data. This could indicate that the attrition between survey rounds adds a systematic bias to the predictions.

Notes

1. One key difference is that ELL allows a decomposition of the error term into a location and household effect. This is not possible with the current version of the MI command in Stata.
2. We also had RF estimate a continuous consumption distribution, but achieved similar results, and choose to stick to the industry standard of using a dummy.
3. The ELL method can also produce standard errors of these welfare measures, something not incorporated into the standard application of RF or MI.
4. There was, however, a notable variation in the questionnaire design used in Ethiopia, with a change in the recall period and fewer household visits between the HECS 2005 and 2010. See World Bank Group (2015) for details and some sensitiveness test of the impact from these variations. The sensitivity tests suggest the variations do not have a large impact on the consumption aggregate or its distribution. To be conservative, the Ethiopian data are only used in the analysis relying on a single year.
5. In the published data, the increase in poverty is only significant at the 10 percent level, whereas here it is found to be significant at the 5 percent level. The exclusion of households with missing explanatory variables could drive this difference.
6. See complete documentation of surveys on www.worldbank.org/lms.
7. Appendix Table A2 shows the measured and estimated levels of poverty at national, urban, and rural levels, which MSE is based on for a subset of the predictions.
8. Kilic and Sohnesen (2015) find that changing the questionnaire from a long to a short version can, by itself, lead to different predictions of poverty.

References

- Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5–32.
- Carletto, C., & T. Kilic. (2009). *Moving up the ladder? The impact of migration experience on occupational mobility in Albania*. World Bank Policy Research Working Paper No. 4908.
- Castle, J. L., X. Qin, & W. R. Reed. (2009). *How to pick the best regression equation: A review and comparison of model selection algorithms*. Working Papers in Economics 09/13, University of Canterbury, Department of Economics and Finance.
- Christiaensen, L., P. Lanjouw, J. Luoto, & D. Stifel. (2012). "Small area estimation-based prediction methods to track poverty: Validation and applications." *Journal of Economic Inequality*, 10(2), 267–297.

- Dang, H.-A. H., P. F. Lanjouw, & U. Serajuddin. (2014). *Updating poverty estimates at frequent intervals in the absence of consumption data methods and illustration with reference to a middle-income country*. World Bank Policy Research Working Paper 7043.
- Doudich, D., A. Ezzrari, R. Van der Weide, & P. Verme. (2013). *Estimating quarterly poverty rates using labor force surveys: A primer*. World Bank Policy Research Working Paper 6466.jj.
- Elbers, C., J. O. Lanjouw, & P. Lanjouw. (2003). "Micro-level estimation of poverty and inequality." *Econometrica*, 71(1), 355–364.
- Hastie, T., R. Tibshirani, & J. Friedman. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer-Verlag.
- Himelein, K. (2014). "Weight calculations for panel surveys with subsampling and split-off tracking." *Statistics and Public Policy*, 1(1), 40–45. doi:10.1080/2330443X.2013.856170
- James, G., D. Witten, T. Hastie, & R. Tibshirani. (n.d.). *An introduction to statistical learning: With applications in R*. Springer Texts in Statistics 103. doi:10.1007/978-1-4614-7138-710
- Kasirye, I., & S. N. Ssewanyana. (2010). *Impacts and determinants of panel survey attrition: The case of Northern Uganda Survey 2004–2008*. Research Series 127536, Economic Policy Research Centre.
- Kilic, T., & T. P. Sohnesen. (2015). *Same question but different answer: Experimental evidence on questionnaire design's impact on poverty measured by proxies*. World Bank Policy Research Working Paper No. WPS 7182. Washington, DC: World Bank Group.
- Mathiassen, A. (2013). "Testing prediction performance of poverty models: Empirical evidence from Uganda." *Review of Income and Wealth*, 59, 91–112. doi:10.1111/roiw.12007
- McBride, L., & A. Nichols. (2015). "Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools." Memo.
- Newhouse, D. L., S. Shivakumaran, S. Takamatsu, & N. Yoshida. (2014). *How survey-to-survey imputation can fail*. World Bank Policy Research Working Paper No. 6961.
- Otok, B. W., & D. Seftiana. (2014). "The classification of poor households in Jombang with random forest classification and regression trees (RF-CART) approach as the solution in achieving the 2015 Indonesian MDGs' targets." *International Journal of Science and Research (IJSR)*, 3(8), 1497–1503.
- Sohnesen, T. (2014). "Tracking poverty via consumption proxies." Memo.
- Sumarto, S., D. Suryadarma, & A. Suryahadi. (2007). "Predicting consumption poverty using non-consumption indicators: Experiments using Indonesian data." *Social Indicators Research*, 81, 543–578.
- Thoplan, R. (2014). "Random forests for poverty classification." *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, North America, 17.
- Varian, H. R. (2014). "Big data: New tricks for econometrics." *Journal of Economic Perspectives*, 28(2), 3–28.
- Verikas, A., A. Gelzinis, & M. Bacauskiene. (2011). "Mining data with random forests: A survey and results of new tests." *Pattern Recognition*, 44, 330–349.
- World Bank Group. (2015). *Ethiopia poverty assessment 2014*. Washington, DC: World Bank. Retrieved from <https://openknowledge.worldbank.org/handle/10986/21323> License: CC BY 3.0 IGO

Appendix

Table A1. Data Sources

Country	Year	Survey
Albania	2005	Living Standard Measurement Survey
	2008–2009	Household Budget Survey
Ethiopia	2010–2011	Household Consumption Expenditure Survey
Rwanda	2005–2006	Enquete Intégrale sur les Conditions de Vie des ménages de Rwanda (EICV2)
	2010–2011	Enquete Intégrale sur les Conditions de Vie des ménages de Rwanda (EICV3)
Malawi	2004–2005	Malawi 2004–2005 Second Integrated Household Survey (IHS2)
	2010–2011	Malawi 2010–2011 Third Integrated Household Survey (IHS3)
Tanzania	2008–2009	Tanzania National Panel Survey (NPS)
	2010–2011	Tanzania National Panel Survey (NPS)
Uganda	2005–2006	Uganda National Household Survey (UNHS 2005/2006)
	2009–2010	Uganda National Household Survey (UNHS 2009/2010)

Table A2. Estimated Head Count Accuracy From Linear Regressions and RF, Same Year

Country	National			Urban			Rural		
	Survey	RF Est.	MI Est.	Survey	RF Est.	MI Est.	Survey	RF Est.	MI Est.
Ethiopia	.30	.30	.28	.31	.18	.23	.24	.32	.29
Malawi	.51	.50	.52	.25	.19	.24	.55	.55	.56
Uganda	.30	.27	.28	.11	.10	.14	.33	.31	.31
Albania	.19	.18	.18	.12	.10	.10	.25	.24	.23
Tanzania	.16	.14	.15	.07	.05	.06	.18	.17	.17
Rwanda	.56	.55	.55	.29	.27	.30	.61	.60	.60

Source: Authors' calculations.

Note: Table shows MSE for poverty predictions out of sample, for same year, based on a split sample. MSE for urban/rural is a simple average of MSE for urban and rural.

Table A3. National MSE for Different Weight Schemes

Country	MI Stepwise			MI Lasso			RF Gini		
	hh Weight	Pop Weight	No Weight	hh Weight	Pop Weight	No Weight	hh Weight	Pop Weight	No Weight
Ethiopia	1.67	6.99	.00	1.94	.01	1.25	.92	2.42	1.71
Malawi	2.38	1.52	3.47	2.85	.84	1.36	.02	.32	.00
Uganda	.50	2.41	.72	.05	4.97	.40	4.47	1.83	4.01
Albania	.45	.77	1.37	2.03	2.46	2.47	9.62	3.61	9.35
Tanzania	.28	3.17	.85	.05	2.50	.45	2.90	1.26	1.77
Rwanda	.35	1.98	.06	.26	1.76	.02	4.50	.79	5.22

Source: Authors' calculations.

Note: Table shows MSE for poverty predictions out of sample, for same year, based on a split sample. MSE for urban/rural is a simple average of MSE for urban and rural.

Table A4. Number of Variables in Models

Variable Selection Method	Same Time Period			Across Time	
	Stepwise	Lasso	HH Observations in Survey	Stepwise	Lasso
<i>Country</i>					
Ethiopia	70	115	12,312		
Malawi	57	68	5,082	66	81
Uganda	91	156	3,618	105	142
Albania	92	152	1,716	91	141
Tanzania	70	140	1,473	82	130
Rwanda	103	161	3,398	120	184
Average	81	132	4,600	93	136

Note: Table shows number of variables in each model selected by Stepwise and Lasso, respectively, using Stata. Stepwise has a setting of .01 for entering a model and .1 for staying. The number of household observations are those used for model selection and is half of the total sample in the survey.