WPS7612

Policy Research Working Paper 7612

# Is Random Forest a Superior Methodology for Predicting Poverty?

## An Empirical Assessment

*Thomas Pave Sohnesen*
*Niels Stender*

WORLD BANK GROUP
Poverty and Equity Global Practice Group
March 2016

Policy Research Working Paper 7612


## Abstract

Random forest is in many fields of research a common method for data driven predictions. Within economics and prediction of poverty, random forest is rarely used. Comparing out-of-sample predictions in surveys for same year in six countries shows that random forest 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.


---

This paper is a product of the Poverty and Equity Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at tpavesohnesen@worldbank.org.

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

Thomas Pave Sohnesen[1]

Niels Stender

JEL: I320, D6, C8, C140

Keywords: Prediction Methods, Poverty, Tracking Poverty, Machine Learning, Random Forest, Linear regression models.

[1] Thomas Pave Sohnesen (corresponding author; tpavesohnesen@worldbank.org): consultant with Development Research Group, World Bank. Niels Stender, Director at Innohead. This work was undertaken as part of small area estimation of poverty implemented by the World Bank for the Poverty and Social Impact Analysis grant, "spatial poverty, inequality and agricultural growth in Ethiopia". The authors would like to thank; Phyllis Ronek for editorial assistance; Ruth Hill and Ana Carolina Areis for valuable comments.

# *1. 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 a cash transfer (World Bank, 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 of poverty over time through poverty proxies are positive and indicate that the approach is capable of tracking poverty over time (Sumarto et al. (2007), Christiaensen et al. (2012), Mathiassen (2013) , Douidich et al. (2013), and Sohnesen (2014)), 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 one is used to construct a model, and year two used to evaluate accuracy, around 50 percent of models produce predictions within the 95 percent confidence interval of the measured level in year two (Christiaensen et al. (2012), Mathiassen (2013) , Sohnesen (2014), and Dang et al. (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 Douidich et al. (2013) utilize several rounds of data, and are, therefore, able to validate their model performance by predicting backwards and forwards. Douidich et al. (2013) seem to successfully track quarterly poverty in Morocco using labor market surveys. Similar efforts in Sri Lanka, however, were unsuccessful (Newhouse, 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 paper contributes.

In this paper, 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 et al. (2003) (known as the ELL method) is commonly used by the World Bank. The Multiple Imputation, (MI hereafter) command in Stata is similar to ELL[2] 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 et al. (2011) that among other areas mention: 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. Hal 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 non-linearities, while, 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 use RF to find poverty predictors and find 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

[2] 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 MI command in Stata.

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, while 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 therefore conclude that RF can significantly improve our-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 one is used for modeling, while year two is used for evaluation of prediction accuracy. The first evaluation compares technical model fit and prediction only, while 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 is small. Hence, the very automated RF tool, that requires very little knowledge on behalf of the user, performs as well 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.

Section 2 outlines the methodology applied, Section 3 describes the data, and Section 4 evaluates prediction outcomes, while Section 5 concludes.

# *2. 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 more in depth description on the methods themselves, see James et al. (2013) and Hastie et al. (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.

**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 0.01 significance level for addition to the model and 0.1 for staying. |
| | | | LASSO using Stata's LARS package |
| RF | Python, Anaconda package | The standard settings, with 500 trees and a minimum of 4 observations in each leaf. | Entropy loss function |
| | | | Gini loss function |

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 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 predict poor and non-poor households with the lowest error and household size three and above therefore becomes the first splitting node. Hence, the sample is split in two groups, those with a household size above three and those size three and below. In the first group (those with households size three and below), household ownership of a TV is then the variable with the lowest prediction error and it becomes the next splitting node. For households with a household size above three 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 over fitting and each tree therefore only rely on a subsample of observations and each node only rely on a subsample of variables. To gain robustness, RF in turn relies on a large number 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 evaluation sample.

2. Draw a random sample of ⅔ of 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 4 observations in a leaf, or until no split leads to further decrease in Gini impurity or entropy value.

4. Repeat steps 2 and 3, 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 headcount 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 et al. (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 in-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 if 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 headcount is calculated, while 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.[3,4]

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

The prediction model is at household level, even though the outcome indicator of interest – poverty headcount - 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 allow 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, while MI on the other hand does allow the use of weights at this stage.

# *3. 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.[5] This allows testing of different estimation models in different circumstances.

[3] 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.

[4] The ELL method can also produce standard errors of these welfare measures, something not incorporated into the standard application of RF or MI.

[5] 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 (2015) for details and some sensitiveness test of the impact from these variations. The sensitivity tests suggest the variations do not have a large

The level of poverty varies from 12.5 % in Albania to 50.2 % 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 five percentage points over just two years, while poverty did not change significantly in urban areas.[6] These diverse trends in poverty test if the approaches are capable of replicating poverty at different levels and with different trends.

**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 |
| Albania 2008 | 12.5% | 10.2% | 14.7% | by large poverty reduction in rural areas |
| Ethiopia 2010 | 29.6% | 25.7% | 30.4% | |
| Malawi 2004/05 | 52.0% | 25.2% | 56.0% | No significant change at national level, but |
| Malawi 2009/10* | 50.2% | 17.0 % | 56.4% | significant poverty reduction in urban areas |
| Rwanda 2006 | 56.7% | 28.5% | 61.9% | Significant poverty reduction at national level and in |
| Rwanda 2011 | 44.9% | 22.1% | 48.7% | urban and rural areas |
| Uganda 2005/06 | 29.6% | 12.8% | 32.8% | Significant poverty reduction at national level and in |
| Uganda 2009/10 | 23.0% | 8.0% | 25.7% | urban and rural areas |
| Tanzania 2008 | 15.0% | 6.1% | 17.5% | Significant increase in poverty at national level*, |
| Tanzania 2010 | 17.9% | 5.2% | 22.3% | driven by changes in the rural areas |

Notes: Numbers might not match official numbers as sample at times varies from those used in official reports, due to missing observations n some explanatory variables. *In 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.

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, while the second year is a panel

impact on the consumption aggregate or its distribution. To be conservative, the Ethiopean data is only used in the analysis relying on a single year.

[6] In the published data, the increase in poverty is only significant at the 10 % level, while here it is found to be significant at 5 % level. The exclusion of households with missing explanatory variables could drive this difference.

following a part of the first year of data. In both panels, split and moving households were tracked.[7] Even though split and moving households were tracked, panel surveys can "age" and/or become a non-representative 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 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, while 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 do 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, non-food consumption, housing quality, ownership of durable goods, employment, and location. All categorical variables have been turned into dummies and extremely skewed variables were excluded.

# *4. 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 same year based on a randomly split sample, where one-half is used to build a model, the other half used to evaluate accuracy of predictions. The second section evaluates prediction accuracy over time based on two years of data, where year one is used for modeling and year two used to evaluate

[7] See complete documentation of surveys on www.worldbank.org/lsms.

prediction accuracy. The later 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.[8] Mean square error (MSE) between the predicted poverty rate and the 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 (Column 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, while Stepwise is not (Column 1 and 2, Table 3). High accuracy at 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 national level, but poorly predicted at urban and rural levels. RF does not seem sensitive to variation in loss function, as both Gini or entropy loss functions leads to very similar results in all countries (Table 3).

[8] 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.

**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 | MI | MI | RF | RF | 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 | 0.01 | 8.75 | 22.76 | 2.42 | 2.69 | 1.97 | 2.51 |
| Malawi | 1.52 | 0.84 | 23.57 | 21.18 | 0.32 | 0.39 | 1.72 | 1.91 |
| Uganda | 2.41 | 4.97 | 2.32 | 3.94 | 1.83 | 2.16 | 5.72 | 4.04 |
| Albania | 0.77 | 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 | 0.79 | 0.72 | 1.03 | 0.66 |
| Average | 2.81 | 2.09 | 8.41 | 9.45 | 1.71 | 1.94 | 2.58 | 2.58 |

Source: Authors' calculations. Notes :Table shows mean square error 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.

As mentioned, RF can also been 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 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 0.01 significance level for addition to the model and 0.1 for staying), while 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, column 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 collects data on a limited number of aspects.[9] 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 of poverty with predictions.

**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 | 0.77 | 0.31 | 0.01 |
| Malawi | 0.32 | 0.72 | 0.24 | 0.84 |
| Uganda | 1.83 | 2.34 | 1.16 | 4.97 |
| Albania | 3.61 | 2.99 | 2.87 | 2.46 |
| Tanzania | 1.26 | 1.55 | 0.53 | 2.50 |
| Rwanda | 0.79 | 0.09 | 1.13 | 1.76 |
| Average | 1.71 | 1.41 | 1.04 | 2.09 |

Source: Authors' calculations. Notes :Table shows mean square error for national poverty predictions out of sample, for same year, based on a split sample. MI results shown are for a models selected with Lasso in Stata with population weights. RF results are based on a Gini loss function and uses population weights.

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, applying household weights in RF alters the loss function, thereby influencing the estimation models, while 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 if 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

[9] Kilic and Sohnesen (2014) find that changing the questionnaire from a long to a short version can, by itself, lead to different predictions of poverty.

the methods is 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 two. The results shaded in grey are national poverty estimates that are outside the measured 95 percent confidence interval of the survey poverty headcount. 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 3 are not found for predictions over time, and no selection nor any prediction method consistently perform better than the others. It's 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.

**Table 5: MSE of poverty predictions over time**

| Country | National MI Lasso | National MI Stepwise | National RF | Urban/rural MI Lasso | Urban/rural MI Stepwise | Urban/rural RF |
|---|---|---|---|---|---|---|
| Malawi | 0.06 | 0.28 | 1.41 | 8.31 | 8.60 | 24.86 |
| Uganda | 22.69 | 0.08 | 28.13 | 18.25 | 6.76 | 22.19 |
| Albania | 0.44 | 0.32 | 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 | 0.52 | 10.63 | 9.38 | 4.10 |
| Average | 12.83 | 4.77 | 10.12 | 15.28 | 9.67 | 16.07 |

Source: Authors' calculations. Notes: Table shows mean square error for poverty predictions out of sample, for same year, based on a split sample. Notes: MSE for urban/rural is a simple average of MSE for urban and rural. Shaded estimates are outside the 95 percent confidence interval of the measured level of poverty in year 2. Shading is only applied to national estimations, and not urban/rural estimations.

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 ageing 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 and Kilic, 2009), is, in fact, 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), and Sohnesen (2014)). The assumption is that 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 is 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, four out of ten households have a cell phone, and it is doubtful if 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. Douidich 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.

# *5. Conclusion*

Random Forest has been used successfully as a data driven prediction method in different fields of research. This paper shows that Random Forest is also a good predictor of poverty; in some cases, a better predictor than current commonly applied methods. Random Forest 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 criteria. 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.

## *References*


Breiman, Leo (2001). "Random Forests". Machine Learning **45** (1): 5–32.

Carletto, Calogero, Kilic, Talip (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., Qin, X., and Reed, W. R. (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, Lanjouw, P, Luoto, J., Stifel, D, 2012. "Small area estimation-based prediction methods to track poverty: validation and applications," Journal of Economic Inequality, Springer, vol. 10(2), pages 267-297, June

Dang, Hai-Anh H. Lanjouw, Peter F. and Serajuddin, Umar. 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

Douidich Mohamed; Abdeljaouad Ezzrari; Roy Van der Weide; Paolo Verme. 2013. Estimating Quarterly Poverty Rates Using Labor Force Surveys: A Primer. World Bank Policy Research Working Paper 6466.

Elbers, Chris, Lanjouw, Jean O., and Lanjouw, Peter, 2003. Micro-Level Estimation of Poverty and Inequality, Econometrica, 71-1: 355-364.

Hastie, Trevor, Tibshirani Robert, and Friedman, Jerome.The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag,2 edition, 2009.

Himelein. Kristen (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, et al., An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics 103, DOI 10.1007/978-1-4614-7138-7 10.

Kasirye, Ibrahim & Ssewanyana, Sarah N., 2010. "Impacts and determinants of panel survey attrition: The case of Northern Uganda survey 2004-2008," Research Series 127536, Economic Policy Research Centre (EPRC).

Kilic, T, Sohnesen, T. P., 2015. Same question but different answer: experimental evidence on questionnaire design's impact on poverty measured by proxies. World Bank Policy Research Working Paper.

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 and Nichols, A. 2015. Improved poverty targeting through machine learning: An application to the USAID Poverty Assessment Tools. Memo

Newhouse, David Locke and Shivakumaran, S. and Takamatsu, S. and Yoshida, N., How Survey-to-Survey Imputation Can Fail (July 1, 2014). World Bank Policy Research Working Paper No. 6961.

Otok B. W., Seftiana, D, 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) Volume 3 Issue 8, August 2014

Sohnesen, 2014. Tracking Poverty via Consumption Proxies. Memo.

Sumarto, S., et al. (2007). "Predicting Consumption Poverty using Non-Consumption Indicators: Experiments using Indonesian Data." Social Indicators Research 81(3): 543-578.

Thoplan, R.2014. Random Forests for Poverty Classification. International Journal of Sciences: Basic and Applied Research (IJSBAR), North America, 17, aug. 2014.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives, 28(2): 3-28.

Verikas, A. Gelzinis, A. and Bacauskiene, M. Mining data with random forests: A survey and results of new tests. Pattern Recognition, 44(2):330–349, 2011

World Bank Group. 2015. Ethiopia Poverty Assessment 2014. Washington, DC. © World Bank. 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 |
| | 2008-09 | Household Budget Survey |
| Ethiopia | 2010/11 | Household Consumption Expenditure Survey |
| Rwanda | 2005-06 | Enquete Intégrale sur les Conditions de Vie des ménages de Rwanda (EICV2) |
| | 2010-11 | Enquete Intégrale sur les Conditions de Vie des ménages de Rwanda (EICV3) |
| Malawi | 2004-05 | Malawi 2004-05 Second Integrated Household Survey (IHS2) |
| | 2010-11 | Malawi 2010-11 Third Integrated Household Survey (IHS3) |
| Tanzania | 2008-09 | Tanzania National Panel Survey (NPS) |
| | 2010-11 | Tanzania National Panel Survey (NPS) |
| Uganda | 2005-06 | Uganda National Household Survey (UNHS 2005/06) |
| | 2009-10 | Uganda National Household Survey (UNHS 2009/10) |

**Table A2: Estimated headcount accuracy from linear regressions and RF, same year**

| | National | | | Urban | | | Rural | | |
|---|---|---|---|---|---|---|---|---|---|
| Country | Survey | RF est. | MI est. | Survey | RF est. | MI est. | Survey | RF est. | MI est. |
| Ethiopia | 0.30 | 0.30 | 0.28 | 0.31 | 0.18 | 0.23 | 0.24 | 0.32 | 0.29 |
| Malawi | 0.51 | 0.50 | 0.52 | 0.25 | 0.19 | 0.24 | 0.55 | 0.55 | 0.56 |
| Uganda | 0.30 | 0.27 | 0.28 | 0.11 | 0.10 | 0.14 | 0.33 | 0.31 | 0.31 |
| Albania | 0.19 | 0.18 | 0.18 | 0.12 | 0.10 | 0.10 | 0.25 | 0.24 | 0.23 |
| Tanzania | 0.16 | 0.14 | 0.15 | 0.07 | 0.05 | 0.06 | 0.18 | 0.17 | 0.17 |
| Rwanda | 0.56 | 0.55 | 0.55 | 0.29 | 0.27 | 0.30 | 0.61 | 0.60 | 0.60 |

Source: Authors' calculations. Notes :Table shows mean square error 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**

| | MI Stepwise | | | MI Lasso | | | RF Gini | | |
|---|---|---|---|---|---|---|---|---|---|
| Country | hh weight | pop weight | no weight | hh weight | pop weight | no weight | hh weight | pop weight | no weight |
| Ethiopia | 1.67 | 6.99 | 0.00 | 1.94 | 0.01 | 1.25 | 0.92 | 2.42 | 1.71 |
| Malawi | 2.38 | 1.52 | 3.47 | 2.85 | 0.84 | 1.36 | 0.02 | 0.32 | 0.00 |
| Uganda | 0.50 | 2.41 | 0.72 | 0.05 | 4.97 | 0.40 | 4.47 | 1.83 | 4.01 |
| Albania | 0.45 | 0.77 | 1.37 | 2.03 | 2.46 | 2.47 | 9.62 | 3.61 | 9.35 |
| Tanzania | 0.28 | 3.17 | 0.85 | 0.05 | 2.50 | 0.45 | 2.90 | 1.26 | 1.77 |
| Rwanda | 0.35 | 1.98 | 0.06 | 0.26 | 1.76 | 0.02 | 4.50 | 0.79 | 5.22 |
| Average | 0.94 | 2.81 | 1.08 | 1.20 | 2.09 | 0.99 | 3.74 | 1.71 | 3.68 |

Source: Authors' calculations. Notes :Table shows mean square error 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**

| | Same time period | | | Across time | |
|---|---|---|---|---|---|
| Variable selection method | Stepwise | Lasso | HH observations in survey | Stepwise | Lasso |
| Country | | | | | |
| Ethiopia | 70 | 115 | 12312 | | |
| Malawi | 57 | 68 | 5082 | 66 | 81 |
| Uganda | 91 | 156 | 3618 | 105 | 142 |
| Albania | 92 | 152 | 1716 | 91 | 141 |
| Tanzania | 70 | 140 | 1473 | 82 | 130 |
| Rwanda | 103 | 161 | 3398 | 120 | 184 |
| Average | 81 | 132 | 4600 | 93 | 136 |

Table shows number of variables in each model selected by Stepwise and Lasso, respectively using stata. Stepwise has a setting of 0.01 for entering a model and 0.1 for staying. The number of household observations are those used for model selection and is half of total sample in survey.