Measuring the value of housing services in household surveys: an application of machine learning approach

Weldensie T. Embaye and Yacob A. Zereyesus*

*Department of Agricultural Economics, Kansas State University, Manhattan, KS, USA

Contact Author: wembaye@ksu.edu

Selected Paper prepared for presentation at the Southern Agricultural Economics Association (SAEA) Annual Meeting, Mobile, Alabama, February 4-7, 2017.

Abstract

Housing value is a major component of the aggregate expenditure used in the analyses of welfare status of households in the development literature. Accurate measurement of housing services is therefore important to obtain the value of housing in household surveys. Data shows that a significant proportion of households in a typical household survey adopted by the Word Bank are self-owned. The standard approach of imputing the housing value for such surveys is based on what it would cost to rent the house. This hedonic pricing is normally predicted using an Ordinary Least Squares (OLS) method. Literature shows that machine learning methods have better predictive power over OLS applied in other valuation exercises. We tested whether or not several machine learning methods (e.g. Ridge, Lasso, Tree, Random forest and Boost) provide superior prediction values of housing over OLS using cross country household survey data. In general, our results confirm that machine learning methods have better predictive capacity than OLS over the given data sets. While predicting housing values, recent developments in the machine learning approaches adopted in the study could be used to obtain improved predictive outcomes.

***Keywords**: Housing values, machine learning, OLS, cross country evidence*

Introduction

The sustainable reduction of poverty and hunger are among the Millennium Development Goals (MDG) as adopted by the World Bank and world community. According World Bank (2012), world poverty had declined from 1.94 billion in 1981 to 1.29 in 2008. To better understand the dynamics of poverty and to design appropriate and effective poverty reduction policies, it is imperative to have well documented poverty data at household level. Many NGOs in collaboration with the national statistics offices have been participating in household level data collection. The World Banks's Living Standards Measurement Study-Integrated Surveys on Agriculture (LSMS-ISA) and others such as the Bill and Melina Gates Foundation cooperate with the national statistics offices on collecting panel based household surveys on many countries counties in Sub-Saharan Africa.

Household aggregate expenditure is frequently used for welfare analyses. Housing service value is a major component of the household's aggregate expenditure. For example, Zereyesus et al., (2017) indicate that in Northern Ghana following food expenditure, housing rent accounted was ranked the second highest category forming 16% of household's annual expenditure. Data shows that a significant proportion of households in a typical household survey adopted by the Word Bank are self-owned. Accurate measurement of housing services is therefore important to obtain true assessment of economic well-being. The standard approach of imputing the housing service values for such surveys is based on what it would cost to rent the house (Garner, Short and Kogan, 2006; Straszheim, 1974). This hedonic pricing is normally predicted using an Ordinary Least Squares (OLS) method. Such regression model has been widely applied in and out of sample predictions. OLS with minimum Mean Square Error (MSE) may offers the best fitted values for in-sample predictions. In OLS type of regressions, increasing the number of variables may increases the overall fit of the regression model (i.e. adjusted R^2), which in turn may increase the in-sample predictive power.. However, such complex model may not perform well in terms of predicting the out-of-sample dependent values. .

Traditional models such as OLS are poorly suited for complex models (Song and Bickel, 2011), especially due to dimensionality. High-dimensional models comprise serial correlation, structural dependence among explanatory variables (spatial correlation), small sample size problem (small

sample size relative to large number of variables) (Belloni et al., 2012; Song and Bickel, 2011) or combination of all. In the world of high-dimensionality, OLS fails to select better predictors and leads to poor out-of-sample prediction. In our analysis, the accuracy of the housing service value for the majority of self-owned type of households is based on performance of the out-of-sample predictive power of the model. Machine learning approaches provide better out-of-sample predictive values considering higher dimensionality of data structures. As Vinod (1978), notes that experiments consistently supported a better prediction error by machine learning methods over OLS. Machine leaning approaches are widely used in computer science field and have recently been adopted in the statistics and economics fields.

The current study aims to evaluate multiple machine learning approaches (i.e. Ridge, LASSO, Tree, Random forest, Bagging and Boosting) to predict housing service values using two period household level survey data from three sub-Saharan countries (Uganda, Tanzania, and Malawi). The performance of these models are evaluated in reference to a standard OLS approach. Results show that in general the machine learning approaches were better in their predictive performance compared to an OLS approach. The different approaches vary from each other.

Methodology

According to Varian (2014), data analysis in statistics and econometrics can be classified into 1) prediction, 2) summarization, 3) estimation, and 4) hypothesis testing. The primary concern of machine learning is prediction. In machine learning, the focus is to find some function that provides a good prediction of Y as a function of X. The vector X is referred to as ‘predictors’ or ‘features’.

Machine learning methods are believed to perform better on modeling data with dimensionality problems such as serial correlation (Ng, 2013; Bai, and Ng, 2009; Andrews, 1991), spatial correlation (Caner and Fan, 2010; Lounici et al., 2009; Knight, 2008; Meinshausen and Bühlmann, 2006; Tibshirani, 1996), small sample size problem (Belloni et al., 2012; Huang, Horowitz and Wei, 2010; Meinshausen and Bühlmann, 2006) or combinational of all (Buehlmann, 2006). In general, these machine learning methods deal with high-dimensional data through shrinking some of the variables to zero and retaining the most important covariates.

The machine learning models generally partition the data into 'training' and 'testing' parts. The training data is used to build the model; while the testing data part is used to test the prediction power of the model. For convenience, the data in this study is partitions into two equal parts. A K-fold cross-validation is considered for model selection. In other applications, Akaike information criteria (AIC), Bayesian information criteria (BIC), and others can also be used for model selection. The advantage of cross validation over BIC, AIC and others is that it depends on fewer assumptions (Varian, 2014). Detailed procedure on how K-fold cross-validation works is contained in James et al. (2013).

Following is a brief review of the standard OLS model and the machine learning approaches. The explanatory variables included in the OLS model and the features or factors in the machine learning models are selected based on their perceived relationship with the house rental price as supported by economic theory and housing economic literature (e.g. Garner, Short and Kogan, 2006; Graves et al., 1988; Straszheim, 1974).

Ordinary Least Square model

Assume that we have the following OLS model:

$$Y = \beta X + \varepsilon,\ \varepsilon \sim (0, \sigma^2) \quad \ldots\ldots (1)$$

Where Y is the outcome variable, X represents a matrix of explanatory variables, ε represents identically and independently distributed (IID) error term, and β refers to the matrix of parameters. . We run the linear regression model to find the optimal estimates that gives least squared errors.

$Minimum\ SE(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$, where E is expected value.

Bias-variance tradeoff

Based on the OLS estimations of the variables coefficients, suppose that our fitted model is given by $\hat{Y} = \hat{\beta} X$. Relative to the actual values of Y, the Prediction error (PE) at particular point of X_0 is given by:

$$PE(X_0) = E_{Y/X=X_0}\left\{(Y - \hat{Y})^2 \middle| X = X_0\right\}$$

After some manipulation of the above prediction error equation, it can be decomposed into the following parts as:

$$PE(X_0) = {\sigma_\epsilon}^2 + Bias^2\big(\hat{Y}(X_0)\big) + Var\big(\hat{Y}(X_0)\big).$$

This decomposition is called bias-variance tradeoff. As more terms are added to the model, the estimates are influenced by high variance. Hence, reducing some variables from the model may add little bias in the model but it may decrease the variance significantly, which in turn reduce the PE. This bias-variance decomposition is at the heart of machine learning models such as Ridge and LASSO (sometime referred as regression regularization methods) which attempt to introduce bias into the regression solution, but reduce the variance significantly as compared to the OLS solutions. The idea is that while OLS regressions provide unbiased coefficient estimates, the lower variance from such machine learning methods produces better Mean Squared Error results.

Machine Learning Methods

Ridge regression

Ridge regression is similar to OLS in that it minimizes the MSE by applying a penalty parameter. The size of this parameter depends on the complexity of the model. Ridge model is specified as:

Minimize $\sum_i (Y_i - \beta X_i)^2 + \lambda \sum_j {\beta_j}^2$

where λ represents penalizing parameter. At $\lambda = 0$, ridge equals to OLS.

Least Absolute Shrinkage and Selection Operator Regression:

The Least Absolute Shrinkage and Selection Operator (LASSO) and ridge are of the same family, where instead of penalizing square of coefficient in ridge, we penalize the summation of the absolute value of the coefficients in Lasso. The Lasso model is:

Minimize $\sum_i (Y_i - \beta X_i)^2 + \lambda \sum_j \left|\beta_j\right|$

Unlike ridge, lasso reduces some coefficients to zero. Neither of them is superior to another. Ridge is chosen over Lasso when so many explanatory variables have small impact on the outcome

variable. Similarly, lasso is chosen over ridge when some explanatory variables have large effect on the outcome variable. Cross validation (Belloni, Chernozhukov and Hansen, 2012) is commonly used to select λ.

Regression Tree:

Unlike the previous machine learning methods, regression tree is a non-parametric approach that does not require specification of any particular functional form. It splits the data in to subtrees based on the variable that best explains the data. It keeps dividing the data in to subtrees until (1) there is single observation in each subtree; (2) all data in the subtree are identical, and (3) number of subtree can be determined by the practitioner.

The model for the Regression tree is given by:

$$Y = \sum_{n} \beta_n \, 1(X \epsilon \, R_n)$$

The indicator $1(X \epsilon \, R_n)$ is a dummy variable with a value of X =1 if the variable is with subtree n, and 0 otherwise. The β_n is the mean value of the data in subtree n.

Bagging

Bagging (Breiman, 1996) is another type of regression tree applied to reduce potentially high variance using a bootstrapping method. As Breiman (1996) states, the bagging predictor is a method for generating multiple versions of a predictor to obtain an aggregated predictor. One of the drawbacks for bagging is that it may not perform well when the predictive power of some of the bagging regressions is much better than the others.

Random Forests

Random forest is similar procedure to bagging, however, unlike bagging; there is the possibility that less strong covariates could be chosen at random. Random forest performs better prediction when we have highly correlated explanatory variables. James et al. (2013) discuss that averaging various uncorrelated quantities offer much less variance compared to highly correlated quantities.

Inclusion of inappropriate variables and not sufficiently small selected variables in random forest results with poorer prediction.

Boosting

Boosting is another type of regression tree that follows slightly different mechanism from random forest. Unlike random forest that build trees independent of each other, boosting builds the trees sequentially. Each tree is built using the information from the previous created tree. Later, a tree is fitted to the residuals, instead of the outcome variable of the former tree and then the errors are updated by adding the new decision tree to the fitted function. This strategy avoids fitting large single tree in to the data that potentially reduces the possibility of overfitting the data.

Data

Data used in the study comes from the nationally representative Living Standards Measurement Study (LSMS) of World Bank household level surveys conducted in 2009/10, 2011/2012 and 2013/2014 in Uganda, Tanzania and Malawi, respectively. Information from surveyed households pertaining to rental values and other variables of interest describing the location, quality of housing, and other features were extracted from the survey data for use in the current study. The number of households that were rented out during the survey in each survey year by country form the basis of the sample sizes for the analyses. In Uganda, the sample sizes are 315 and 389 in the years 2009 and 2011, respectively. In Tanzania, the sample sizes are 699 and 956 in the years 2011 and 2013, respectively. In Malawi, the sample sizes are 1492 and 676 in the years 2011 and 2013, respectively (Table 1). The rental monetary values from the domestic currency in each of countries is converted to a US dollar to facilitate easier comparison and analyses of results.

Table 1 presents descriptive statistics of key variables used in the analyses. In Uganda, households who live in urban area form 54 percent in years 2009 and 59 percent in 2011. Likewise, in Tanzania, over 70 percent of the households live in urban areas in both years, and in Malawi, households who live in urban area constitute 77 and 66 percent in years 2011 and 2013, respectively. The annual rent is almost constant over the years across all three countries. The rental rate per year in US dollars is around 300 in both years for Uganda and Malawi, where as in Tanzania, the rental rates are 168 in year 2009, and 260 in year 2013. The majority of households in all countries have roofs, floors, and external walls made up of mud. The average number of

rooms per a house ranged from 1.67 in Tanzania in year 2009 to 2.53 in Malawi in 2011. Percentage of households with access to electricity ranged from 37 percent to 54 percent. The type of water sources includes: private tap water, public tap water, bore-hole water, protected well water and unprotected water. There is no consistent pattern in the type of water source used in the three countries. Public tap water in Uganda (47%) and Malawi (27%) and protected well water in Tanzania (25%) are relatively dominant source of water used. The type of toilets available are covered private and shared toilets, VIP and uncovered latrine toilets, and flush toilets. The main type of toilet in both years is cover shared toilet (61%) in Uganda but it varies by year in Tanzania and Malawi. In the year 2013, VIP latrine type of toilet turned out to be the main type of toilet in Tanzania and Malawi.

Results and Discussions

Determinants of housing values

Table 2-4 present the results from the OLS and various machine-learning approaches used in the study to estimate the determinants of housing service values in Uganda, Tanzania and Malawi. The significance of the explanatory variables in the OLS models are consistent in the two periods for the three countries. In Uganda, factors that are significantly associated with housing rental values are location in urban, number of rooms, availability of electricity, and private tap water. In Malawi, Urban location, number of rooms, electricity, and flash toilet are significantly associated with the housing rental values. In Tanzania, Urban location, number of rooms, external wall, availability of electricity, and protected well water are significantly associated with the house rental values. None of the other explanatory variables is significantly associated with the dependent variables in any of the three countries.

Of the significantly associated variables, for example, Urban location is positively associate with rental values implying that rent is higher for houses located in urban than rural. On average, house rent in urban location is higher by $10-14 in Uganda, $5 in the year 2013 in Tanzania and $9 in the year 2011 in Malawi. Many households want to live in urban area because urban are associated with more employment opportunities, and better public services that lead to higher living standards (Sahn and Stifel, 2003). House rent increases with the increase of number of rooms. Increasing number of rooms by 1, increases on average monthly house rent by about $5 in Uganda, $7-13 in

Tanzania, and $11 in Malawi. The effect of electricity on house rent is also positive. Rental values of houses with access to electricity are higher by $19 per month in Uganda, $6-13 in Tanzania and $7 in the year 2013 in Malawi. These relationships suggest that electricity is essential for household's livelihood in that it can be used for household purposes (lighting, cooking, etc.), mechanizations (farming and non-farming) and communication and other purposes (Bernard, 2010; Khandker et al., 2009). Private tap water is also positively associated with the value of house rent in Uganda. On average, a house rents in dwellings with a private water tap are higher by $28 in Uganda. North and Griffin (1993) found that households value in-house piped water sources higher than any other house characteristics. Rental houses increases with access to clean water because more households want to stay closer to clean water (Nauges and Whittington, 2010; Quigley, 1982). In Tanzania, however, protected well type of water source is negatively associated with house rent.

Results from the ridge machine learning regression models show that most of the estimated coefficients of the explanatory variables are smaller in magnitude compared to the OLS coefficients (see Table 2-4). This is consistent with the literature (James et al., 2013) in that ridge regression shrinks the estimated coefficients and reduces the variance resulting in increased predictive power of the covariates. Similarly, the LASSO regression approach shrinks the variable estimates and in some cases, the coefficients are reduced to zero. When the estimated variables coefficients are zero, then the variables may not play a role in predicting the value of the dependent variables. In Uganda. the variable estimates that were reduced to zero under LASSO regression are dwelling, floor, bore hole water, unprotected water sources, covered private and shared type of toilet, pit and uncovered latrine in 2009; and dwelling, roof, floor, public tap, bore hole and unprotected water sources, pit and uncovered latrine in 2011. In Tanzania variable estimates for roof, floor, public tap, protected and unprotected well water sources, VIP latrine type of toilet in 2011 and floor, public tap and unprotected well water sources, and VIP latrine type of toilet in 2013 are also reduced to zero under the LASSO regression. For Malawi, the variables with zero estimated coefficients under LASSO regression are dwelling, floor, public tap, bore hole, protected and unprotected well water sources and VIP latrine type of toilet in 2011 and dwelling, public tap, protected and unprotected well water sources, VIP and pit latrine type of toilet in 2013 in.

Unlike OLS, Ridge, and LASSO, the other machine learning approaches (i.e. the Regression trees, bagging, random forest and boosting) do not provide the estimated coefficients between the covariates and predicted variables.

In regression tree, the prediction of the house rental values is done by using the mean/mode of the group at each node of the tree. For example, in Uganda, variables selected in the regression tree by rank are flush toilet, availability of electricity, protected water (well) and number of rooms in the year 2009; and electricity, number of rooms, protected well and private tap water sources in the year 2011 (Figure 1_u_a). The summary statistics indicate that only 4 percent of the households have access to flush toilet. The largest group, those without access to flush toilet are further split using electricity, protected water and number of rooms variables. Regression tree is highly criticized that changing the data slightly brings substantial change in tree construction (James et al., 2013). Likewise, variables used in the tree regression in Tanzania are number of rooms, electricity, pit latrine, private water tap, and external wall in the year 2009 and number of rooms, electricity, external wall, flush toilet and urban in the year 2013. In Malawi, variables such as flush toilet, electricity and number of rooms in 2011, and flush toilet, electricity and number of rooms in 2013 are used in the tree regression.

As stated above, the bagging method is an extension of the regression tree. We construct numerous trees simply through resampling multiple times from the same data (i.e. bootstrapping) and then we average the entire prediction. In each bagging, 500 trees are created using twelve variables. The results from the bagging regression are sorted according to their significance in the prediction process. For example, in Uganda, the three top most variables used during the bagging regression by rank are flush toilet, protected water sources, and electricity; whereas the dwelling variables appears in the bottom most level (Figure 2_u_a). The arrangement of variables vary by year of analysis and country of analysis. Details of these results are presented in the appendix.

Random forest is similar to bagging, but instead of averaging all the predictors, we take some of the predictors via random selection to perform the prediction. For example, out of the 12 bagging variables used, random forest picked 4 variables randomly. The top 4 variables selected in random forest in Uganda are private water tape sources, flush toilet, public water tap, and electricity in the

year 2009 (Figure 3_u_a). The rest of the results for the other models by year and by country is provided in the appendix.

The results from the boosting regression, which is another extension to the regression tree are presented in table 5. The estimated from the boosting regression are arranged in order of their importance to the prediction process. For instance, In Uganda in the year 2009, the top three variables used in boosting by order of importance are electricity, number of rooms, and public water tape sources; the four of the variables that are not used for boosting are unprotected water sources, flush toilet, and roof. A number of variable are not used in the boosting regression of different countries and in different years as shown in the table.

Out-of-sample prediction

The main strength of machine learning approaches is their predictive capability. The Mean Squared Error is one of the methods used to compare between the machine learning approaches and the standard OLS regression. The results of such comparison of these models is presented in table 6. Our results indicate that the prediction power of machine learning is superior to OLS. But, it matters which type of machine learning we are comparing with OLS. For example, in Uganda in the year 2009, the prediction methods by rank are boosting, ridge, lasso, forest, OLS, bagging and tree; and bagging, forest, lasso, OLS, ridge, tree and boosting in 2011. In Uganda, OLS performed better in prediction than bagging and tree in 2009 and ridge, tree and boosting in 2011. However, OLS is less preferred than boosting, ridge, lasso and forest in the year 2009 and bagging, forest and lasso in the year 2011.

In Tanzania, the prediction methods in order of preference judged by minimum MSE are forest, tree, boosting, OLS, lasso, ridge, and bagging in the year 2009 and forest, ridge, lasso, OLS, boosting, bagging and tree in the year 2013. This implies that forest, tree and boosting in the year 2009, and forest, ridge and lasso in the year 2013 performed better for out-of-sample prediction compared to OLS. On the other hand, OLS is superior to lasso, ridge and bagging in the year 2009, and boosting, bagging and tree in the year 2013. Importantly, the rank of the prediction methods from top to bottom in Malawi are forest, boosting, bagging, tree, OLS, ridge and lasso in the year 2011 and boosting, tree, bagging, forest, OLS, lasso, and ridge in the year 2013. Likewise, in

Malawi, in both years, OLS performed better in prediction than ridge and lasso but poorer prediction than boosting, bagging, tree and forest.

Conclusion

Generally, the results indicated that the machine learning methods have better prediction power than OLS. The use of cross-country data shows that the results vary by country as well as year of analysis. This is in agreement with James et al. (2013) conclusion that there is no method that dominates all other methods over any data set. Besides, the advantage of machine leaning models over OLS type regressions is significant when the number of predictor variables (features) is significant larger.

References:

Andrews, D. W. (1991). Asymptotic optimality of generalized C L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, *47*(2), 359-377.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *The American Economic Review*, *105*(5), 481-485.

Bai, J., & Ng, S. (2009). Boosting diffusion indices. *Journal of Applied Econometrics*, *24*(4), 607-629.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369-2429.

Bernard, T. (2010). Impact analysis of rural electrification projects in sub-Saharan Africa. *The World Bank Research Observer*, lkq008.

Breiman, L. 1996 "Bagging Predictors" *Machine Learning* 24 (2) : 123-140

Buehlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 559-583.

Caner, M., & Fan, Q. (2010). *The adaptive lasso method for instrumental variable selection*. Working Paper, North Carolina State University.

Garner, T. I., Short, K. S., & Kogan, U. (2006). What Do We Know About the Value of Owner Occupied Housing Services? Rental Equivalence and Other Approaches.

Graves, P., Murdoch, J. C., Thayer, M. A., & Waldman, D. (1988). The robustness of hedonic price estimation: urban air quality. *Land Economics*, *64*(3), 220-233.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Jimenez, E. (1982). The value of squatter dwellings in developing countries. *Economic Development and Cultural Change*, *30*(4), 739-752.

Khandker, S. R., Barnes, D. F., Samad, H. A., & Minh, N. H. (2009). Welfare impacts of rural electrification: evidence from Vietnam. *World Bank Policy Research Working Paper*, (5057).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485-585). Springer New York.

Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, *38*(4), 2282.

Knight, K. (2008). Shrinkage estimation for nearly singular designs. *Econometric Theory*, *24*(02), 323-337.

Lounici, K., Pontil, M., Tsybakov, A. B., & Van De Geer, S. (2009). Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 1436-1462.

Nauges, C., & Whittington, D. (2010). Estimation of water demand in developing countries: An overview. *The World Bank Research Observer*, *25*(2), 263-294.

Ng, S. (2013). Variable selection in predictive regressions. *Handbook of Economic Forecasting*, *2*(Part B), 752-789.

North, J. H., & Griffin, C. C. (1993). Water source as a housing characteristic: Hedonic property valuation and willingness to pay for water. *Water Resources Research*, *29*(7), 1923-1929.

Quigley, J. M. (1982). Nonlinear budget constraints and consumer demand: An application to public programs for residential housing. *Journal of Urban Economics*, *12*(2), 177-201.

Sahn, D. E., & Stifel, D. C. (2003). Urban–rural inequality in living standards in Africa. *Journal of African Economies*, *12*(4), 564-597.

Song, S., & Bickel, P. J. (2011). Large vector auto regressions. *arXiv preprint arXiv:1106.3915*.

Straszheim, M. (1974). Hedonic estimation of housing market prices: A further comment. *The Review of Economics and Statistics*, 404-406.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Zereyesus, Y. A., Embaye, W. T., Tsiboe, F., & Amanor-Boadu, V. (2017). Implications of Non-Farm Work to Vulnerability to Food Poverty-Recent Evidence from Northern Ghana. *World Development*, *91*, 113-124.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3-27.

Vinod, H. D. (1978). A survey of ridge regression and related techniques for improvements over ordinary least squares. *The Review of Economics and Statistics*, 121-131.

Tables and figures:

Table 1: Summary statistics of rental house characteristics

Variable	Definitions	Mean values					
		Ugandan		Tanzania		Malawi	
		2009	2011	2009	2013	2011	2013
Urban	1 if a house is located in Urban, 0 otherwise	0.54	0.59	0.94	0.75	0.77	0.66
		(0.50)	(0.49)	(0.45)	(0.43)	(0.42)	(0.47)
Annual rent	Amount of rent paid per year in dollar	290	290	168	260	300	300
		(37.19)	(37.19)	(35.24)	(37.06)	(50.15)	(44.51)
Dwelling	1 if the room is located within a house, 0 otherwise	0.69	0.70	-	0.45	0.87	0.92
		(0.46)	(0.46)	(0.49)	(0.49)	(0.34)	(0.28)
Roof	1 if roof type if mud, 0 otherwise	0.92	0.92	0.96	0.93	0.82	0.86
		(0.28)	(0.26)	(0.23)	(0.25)	(0.39)	(0.35)
floor	1 if floor is mud, 0 otherwise	0.70	0.70	0.97	0.77	0.74	0.79
		(0.46)	(0.46)	(0.41)	(0.41)	(0.44)	(0.41)
External wall	1 if external wall is in mud, 0 otherwise	0.72	0.74	0.67	0.72	0.05	0.05
		(0.45)	(0.44)	(0.47)	(0.45)	(0.22)	(0.21)
Number of rooms	Number of rooms rented	1.80	1.77	1.67	1.68	2.53	2.11
		(1.11)	(1.16)	(0.97)	(0.91)	(1.09)	(0.97)
Electricity	1 if the house has electricity, 0 otherwise	0.37	0.41	0.54	0.54	0.37	0.43
		(0.48)	(0.49)	(0.50)	(0.50)	(0.48)	(0.50)
Water sources							
Private tap water	1 if private tap water source, 0 otherwise	0.08	0.07	0.35	0.12	0.13	0.14
		(0.27)	(0.25)	(0.26)	(0.32)	(0.34)	(0.35)

Variable	Definitions	Mean values					
		Ugandan		Tanzania		Malawi	
		2009	2011	2009	2013	2011	2013
Public tap water	1 if public tap water source, 0 otherwise	0.49	0.47	0.39	0.12	0.27	0.27
		(0.50)	(0.49)	(0.38)	(0.33)	(0.44)	(0.44)
Bore hole water	1 if bore hole water source, 0 otherwise	0.19	0.19	-	0.07	0.33	0.31
		(0.39)	(0.39)	-	(0.25)	(0.47)	(0.46)
Protected well water	1 if protected well water source, 0 otherwise	0.14	0.13	0.11	0.25	0.04	0.04
		(0.34)	(0.34)	(0.43)	(0.43)	(0.22)	(0.28)
Unprotected water	1 if unprotected well water, 0 otherwise	0.10	0.10	0.15	0.13	0.21	0.20
		(0.34)	(0.25)	(0.24)	(0.34)	(0.41)	(0.40)
Covered private toilet	1 if covered private toilet, 0 otherwise	0.16	0.17	-	0.04	0.14	0.16
		(0.37)	(0.37)		(0.19)	(0.35)	(0.37)
Cover shared toilet	1 if covered shared toilet, 0 otherwise	0.61	0.62	-	0.25	0.09	0.05
		(0.49)	(0.48)		(0.43)	(0.29)	(0.22)
VIP latrine toilet	1 if VIP latrine toilet, 0 otherwise	0.07	0.08	0.17	0.05	0.67	0.64
		(0.25)	(0.26)	(0.41)	(0.22)	(0.47)	(0.48)
Uncovered latrine	1 if uncovered latrine, 0 otherwise		0.10	-	0.20	-	
			(0.29)	-	(0.40)	(0.27)	
Pit latrine	1 if pit latrine, 0 otherwise	-	-	0.70			
				(0.45)			
Flush toilet	1 if flush toilet, 0 otherwise	0.04		0.13			
		(0.19)		(0.33)			
Observation		315	389	525	956	1492	676

Table 2: Determinants of housing rental values based on OLS, Lasso and Ridge regressions models in Uganda

Variables	2009			2011		
	OLS	**Ridge**	**Lasso**	**OLS**	**Ridge**	**Lasso**
Constant	-10.76 (42.03)	11.95	3.06	-2.76 (9.12)	6.62	-1.34
Urban	13.92* (4.02)	8.03	8.21	9.80* (2.62)	5.44	5.53
Dwelling	2.69 (4.50)	-0.71	-	-1.58 (3.06)	-1.46	-
Roof	-13.58 (8.56)	-7.97	-1.32	-3.81 (5.68)	-3.22	-
floor	-0.26 (5.47)	2.04	-	1.96 (3.74)	-0.66	-
External wall	9.93 (5.36)	6.62	3.64	7.90 (3.65)	4.72	3.47
Number of rooms	5.75* (1.79)	3.89	3.42	4.56*** (1.04)	3.00	3.74
Electricity	18.80** (4.05)	12.56	15.79	19.23* (2.91)	10.93	16.76
Private tap water	20.23 (30.43)	9.69	10.01	28.42** (6.79)	18.70	23.96
Public tap water	4.59 (29.74)	--5.25	-1.13	-0.73 (3.96)	-0.75	-
Bore hole water	11.16 (29.96)	-4.27	-	4.07 (4.21)	-2.57	-
Protected well water	36.10 (29.98)	15.04	19.72	15.29 (7.35)	6.67	9.32
Unprotected water	16.86 (30.29)	-1.59	-	3.09 (7.96)	-2.22	-
Covered private toilet	-0.88 (27.94)	-2.87	-	-12.45 (12.01)	-4.14	-0.57
Cover shared toilet	2.60 (25.99)	-0.55	-	-4.20 (7.99)	0.24	1.55
VIP latrine toilet	-11.72 (32.08)	-7.38	-3.29	-14.08 (12.16)	-3.20	-0.11
Flush toilet	45.61 (30.43)	31.66	42.86	14.23 (25.68)	17.60	21.54
Pit latrine	-	-	-	-	-	-
Uncovered latrine	3.09 (26.04)	-0.93	-	-9.25 (8.73)	-2.10	-
R^2	0.32			0.29		

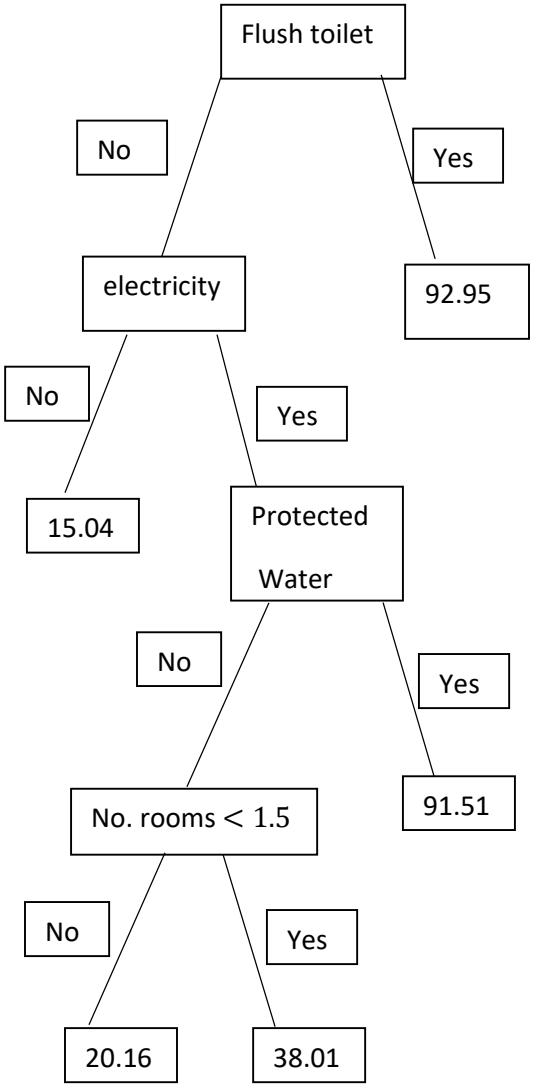
Table 3: Determinants of housing rental values based on OLS, Lasso and Ridge regressions models in Tanzania

	2009			2013		
Variables	**OLS**	**Ridge**	**Lasso**	**OLS**	**Ridge**	**Lasso**
Constant	-11.63	-5.05	-5.20	-15.27**	-14.42	-14.31
	(16.34)			(6.12)		
Urban	5.55**	4.67	2.78	4.55*	4.36	3.51
	(5.27)			(4.42)		
Dwelling	-	-	-	-	-	-
Roof	3.94	-0.28	-	-3.62	-3.16	-0.87
	(20.91)			(7.45)		
floor	-5.92	-2.15	-	2.44	1.52	-
	(22.45)			(5.09)		
External wall	7.92**	6.67	6.61	10.81***	10.11	9.58
	(3.15)			(3.68)		
Number of rooms	6.93***	5.84	6.24	13.27***	12.44	12.61
	(1.39)			(1.76)		
Electricity	5.90*	5.59	5.59	12.95***	12.12	12.67
	(3.05)			(3.83)		
Private tap water	3.30	3.57	4.72	-0.44	0.45	0.83
	(4.43)			(3.20)		
Public tap water	-2.79	-2.43	-	-2.75	-2.03	-
	(3.97)			(6.59)		
Bore hole water	-	-	-	-	-	-
Protected well water	-2.75	-2.16	-	-7.65*	-6.78	-3.83
	(4.56)			(8.31)		
Unprotected water	-	-0.23	-	-3.32	-2.72	-
				(4.03)		
Covered private toilet	-	-	-	-	-	-
Cover shared toilet	-	-	-	-	-	-
VIP latrine toilet	3.64	0.97	-	0.48	1.19	-
	(15.79)			(7.54)		
Flush toilet	15.40	11.56	10.84	0.96	2.05	0.36
	(15.94)			(6.64)		
Pit latrine	-3.15	-5.81	-5.93	-3.71	-2.95	-3.41
	(15.40)			(6.39)		
Uncovered latrine	-	-	-	-		-
R^2	0.32			0.44		

Table 4: Determinants of housing rental values based on OLS, Lasso and Ridge regressions models in Malawi

Variables	2011			2013		
	OLS	**Ridge**	**Lasso**	**OLS**	**Ridge**	**Lasso**
Constant	-30.49 (36.76)	-5.44	-16.95	-6.47 (9.53)	4.81	-10.44
Urban	8.96** (4.43)	8.30	7.06	1.64 (4.93)	2.08	0.78
Dwelling	-3.66 (10.94)	-1.76	-	-5.13 (11.67)	-2.12	-
Roof	7.67 (8.54)	5.91	2.32	5.83 (9.13)	3.70	1.40
floor	-3.03 (5.58)	-2.21	-	-1.70 (5.30)	-2.05	-0.88
External wall	9.67 (9.05)	9.01	5.71	8.23 (7.76)	6.92	4.30
Number of rooms	10.54*** (1.98)	9.96	9.97	10.90*** (1.87)	9.47	9.99
Electricity	6.80** (5.04)	6.91	6.17	7.17* (4.68)	7.62	7.43
Private tap water	4.03 (21.98)	2.98	2.46	-	5.11	5.03
Public tap water	0.99 (20.32)	-0.83	-	-5.68 (18.97)	-1.53	-
Bore hole water	2.29 (12.27)	-0.14	-	-9.80 (19.21)	-4.62	-3.05
Protected well water	0.51 (13.44)	-1.59	-	-5.30 (26.37)	-1.18	-
Unprotected water	0.45 (21.23)	-1.47	-	-4.05 (18.93)	-0.41	-
Covered private toilet	-	-	-	-		-
Cover shared toilet	-	-	-	-		-
VIP latrine toilet	10.49 (15.09)	-11.37	-	0.66 (16.61)	-11.67	-
Flush toilet	60.81 (35.60)	36.31	50.04	40.87*** (20.61)	23.18	39.58
Pit latrine	6.81 (35.32)	-15.30	-2.85	1.03 (12.39)	-13.68	-
Uncovered latrine	-		-	-		-
R^2	0.37			0.39		

Figure 1: Tree Regression in Uganda in year 2009



Bagging regression results

Figure 2_u_a

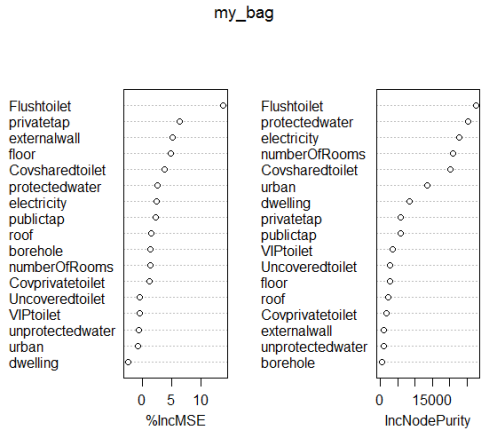


figure 2_u_b

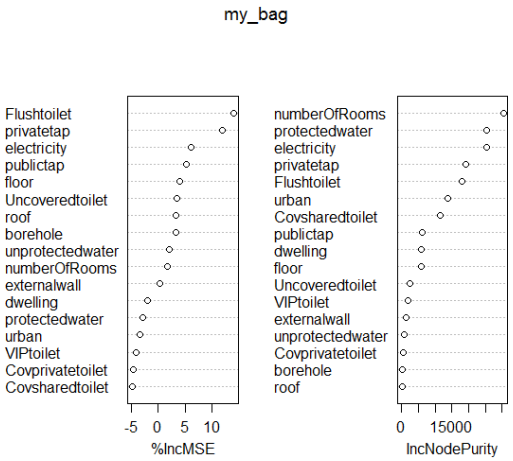


Figure 2_t_a

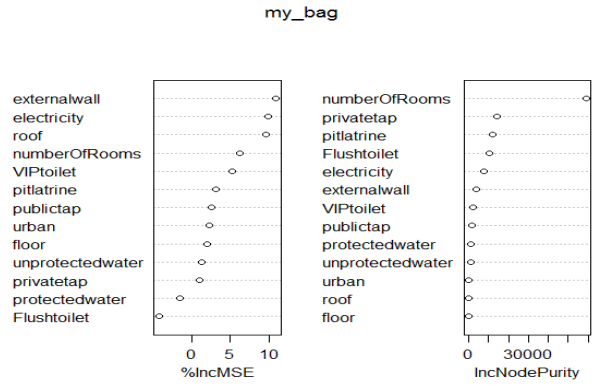


Figure 2_t_b

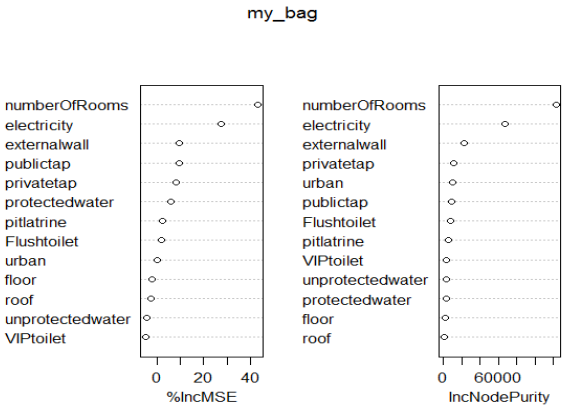


Figure 2_m_a

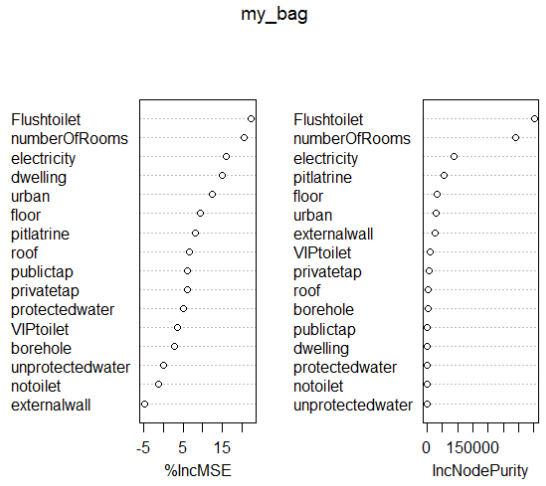
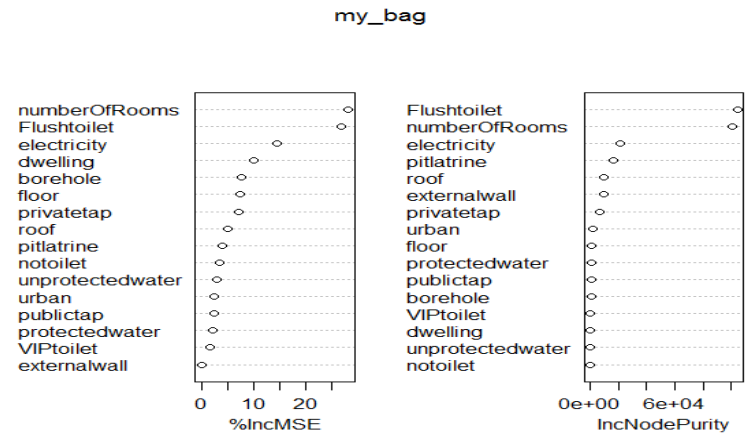


Figure 2_m_b



Forest regression results

Figure 2_u_a

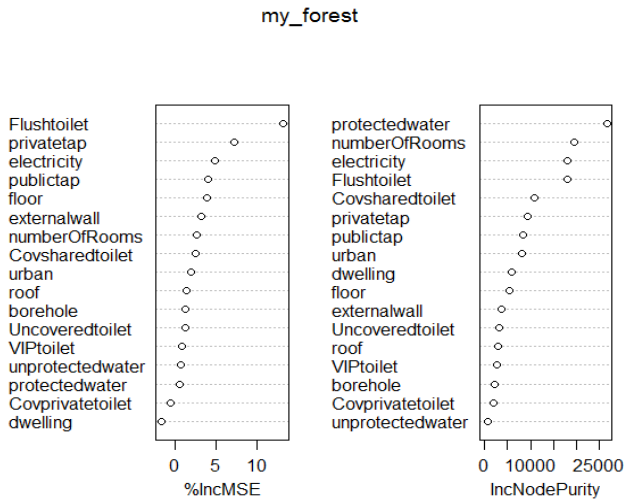


Figure 2_u_b

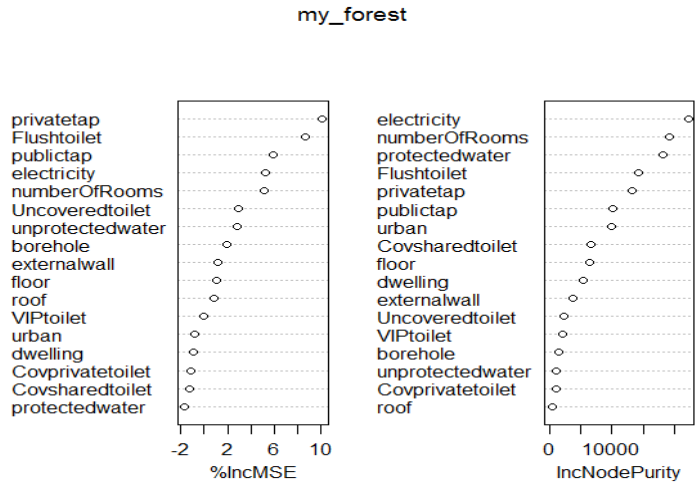


Figure 2_t_a

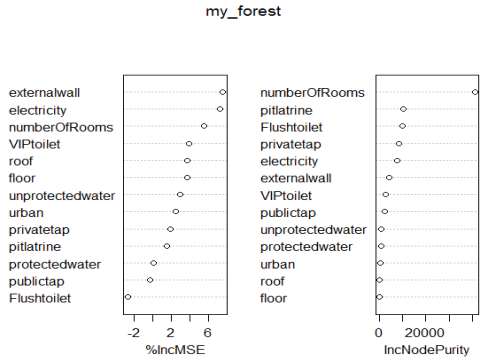


Figure 2_t_b

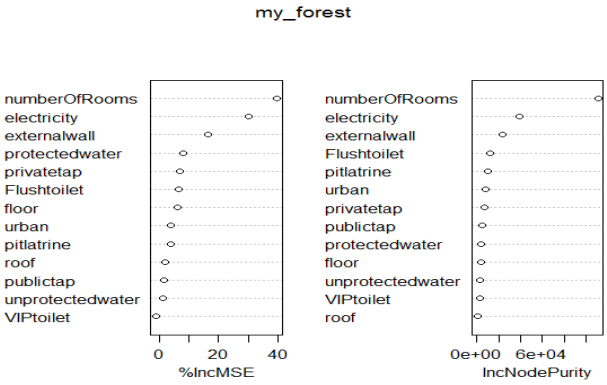
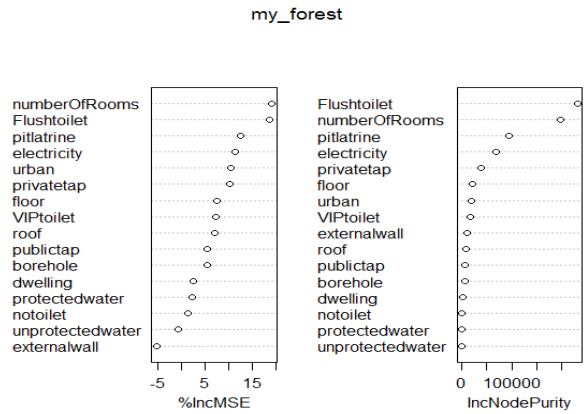


Figure 2_m_a

Figure 2_m_b



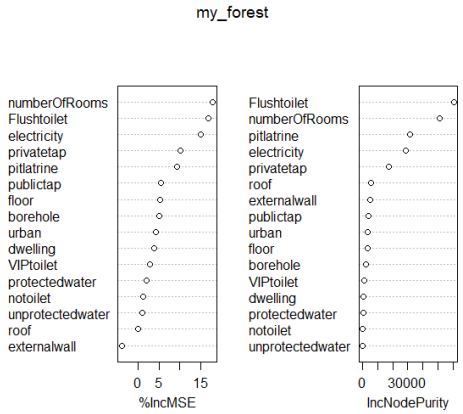


Table: Table 5: Modeling factors associated to housing rental values using boosting regression

Uganda				Tanzania				Malawi			
2009		**2011**		**2009**		**2013**		**2011**		**2013**	
Variable	Rel. info	Variable	Rel. info	Variable	Rel. info	Variable	Rel. info	Variable	Rel. info	Variable	Rel. info
Electricity	25.73	Number Of rooms	31.06	Number of rooms	47.71	Number of rooms	53.73	Flush toilet	52.64	Flush toilet	4.97
Number Of rooms	22.66	Electricity	27.83	Flush toilet	18.99	Electricity	24.95	Number of rooms	37.45	Number of rooms	3.94
Public tap water	18.07	Protected well	18.75	Pit latrine	10.99	External wall	10.63	Electricity	4.65	Electricity	6.76
Protected well	9.57	Urban	7.14	Private tap water	10.76	Flush toilet	2.76	Urban	2.03	Private tap water	2.60
Urban	6.83	Public tap water	4.10	Electricity	5.96	Pit latrine	2.71	Pit latrine	1.32	External wall	0.60
Private tap water	5.28	Dwelling	3.69	External wall	2.81	Urban	2.53	Private tap water	0.93	Pit latrine	0.35
Dwelling	3.99	Shared toilet	2.60	VIP toilet	2.29	Private tap water	1.68	roof	0.41	roof	0.33
Shared toilet	3.67	External wall	1.74	Public tap water	0.23	Public tap water	0.42	External wall	0.36	floor	0.15
External wall	1.49	floor	1.29	Urban	0.22	Protected well	0.25	floor	0.16	Bore hole	0.10
floor	1.04	Private tap water	0.71	Unprotected water	0.02	VIP toilet	0.24	VIP toilet	0.04	Urban	0.07
Bore hole	0.73	Private toilet	0.53	Protected well	0.02	roof	0.07	Bore hole	0.01	Dwelling	0.01
VIP toilet	0.55	Flush toilet	0.39	roof	0.00	floor	0.02	Dwelling	0.00	Public tap water	0.01
Private toilet	0.31	Unprotected water	0.06	floor	0.00	Unprotected water	0.02	Protected well	0.00	Protected well	0.00
Uncovered toilet	0.10	Bore hole	0.05	-	-	-	-	Public tap water	0.00	VIP toilet	0.00
Unprotected water	0.00	Uncovered toilet	0.05	-	-	-	-	Unprotected water	0.00	Unprotected water	0.00
Flush toilet	0.00	VIP toilet	0.00	-	-	-	-	No toilet	0.00	No toilet	0.00
roof	0.00	roof	0.00	-	-	-	-	-	-		

Table 6: Prediction mean square error of estimating housing values under various models

	Ugandan		Tanzania		Malawi	
	2009	**2011**	**2009**	**2013**	**2011**	**2013**
OLS	723.89	612.43	462.98	432.68	2086.01	1671.83
Ridge	623.38	659.13	467.95	404.94	2092.60	1699.03
Lasso	634.22	610.85	465.54	422.13	2092.83	1685.57
Tree	870.80	677.30	407.92	678.21	2033.67	1501.49
Bagging	854.01	576.08	488.18	459.15	1964.68	1503.17
Forest	658.49	579.64	388.60	370.62	1927.84	1525.52
Boosting	580.56	709.64	410.79	433.18	1955.77	1445.2