Development Engineering 2 (2017) 1–11

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

On LPG usage in rural Vietnamese households

Niklas Vahlne

Chalmers University of Technology Department of Energy and Environment Division of Energy Technology, SE-412 96 Göteborg, Sweden

A B S T R A C T

Cooking with solid biomass fuels, a common practice in the developing world, is associated with numerous problems. Hence policy makers wish to facilitate households to switch to more modern fuels. To better understand the potential for policy interventions, an enhanced understanding of household fuel choices and the process of fuel switching is paramount.

The primary aim of this paper is to perform an exploratory data analysis in order to obtain a set of factors associated with rural households that are using liquefied petroleum gas (LPG). This is achieved through *Random Forest analysis*, a statistical technique commonly employed for solving classification problems. In this context, fuel choice defines groups to which households belong, while the random forest modelling is used to determine the importance of the variables on the correct classification. The results from this study can be used for constructing further statistical models, as basis for experimental work as well as for questioning previous models.

This study ranks the overall predictive importance of a great number of variables previously used in literature on fuel switching. High importance is given to variables coupled to household wealth and history of income together with various commune level variables such as distance to nearest town. The results indicate that whether households have or will undergo fuel switching can be predicted based on area characteristics associated with rurality together with a history of household income and wealth, and furthermore there appears to be an interaction effect between these characteristics.

Current income is unable to fully account for a household's wealth and history of income, which in turn appear to be associated with both the current and future use of LPG. The likeliness that a household is or will start using LPG increases with increased wealth if the household resides in a less rural environment. Furthermore, some previously used variables for modelling fuel switching may instead be explained by their association with either wealth, stable income or with level of rurality.

1. Introduction

In many developing countries the primary fuel used for cooking is biomass (Foell et al., 2011). However, due to inefficient combustion, numerous problems are associated with this usage, among them severe health issues (Bruce et al., 2000; Desai et al., 2004; Smith and Peel, 2010). It is estimated that every year two million deaths are caused by the indoor air pollution originating from the combustion of solid fuels (WorldBank, 2011). Many of the compounds formed under inefficient combustion, such as methane and black carbon, are also contributing to global warming (Bond and Sun, 2005; Bond et al., 2011; Ramanathan and Carmichael, 2008).

Households have been noted to switch to or include cleaner and more convenient fuels in their fuel mix as their income and socio-economic status rise. This process is often called fuel switching (Heltberg, 2005) or fuel stacking (Masera et al., 2000), to reflect the fact that several fuels are often used concurrently in the same household. A number of studies have examined the process of fuel switching (Masera et al., 2000; Barnes et al., 2004; Campbell et al., 2003; Heltberg, 2004; Hosier and Dowd, 1987; Leach, 1992). For a comprehensive review, see van der Kroon et al. (2013)

Fuel switching is generally occurring faster in urban areas compared to rural areas (Heltberg, 2004, 2005; Hosier and Dowd, 1987; Gundimeda and Köhlin, 2008). Possible explanations for the lower rate of fuel switching in the rural areas include a lack of infrastructure for modern fuels (Leach, 1992), lower or non-monetary sporadic income, a traditional lifestyle and a lower opportunity cost of time in addition to the higher availability of collectible fuels. Additionally, the availability of biomass strongly influences the path of urban fuel switching (Barnes et al., 2004).

Several studies have found a correlation between the use of liquefied petroleum gas (LPG) and electrification (Barnes et al., 2004;

E-mail address: Niklas.vahlne@chalmers.se.

http://dx.doi.org/10.1016/j.deveng.2016.12.003
Received 25 June 2015; Received in revised form 5 October 2016; Accepted 21 December 2016
Available online 23 December 2016
2352-7285/

Campbell et al., 2003; Heltberg, 2004; Davis, 1998). It is, however, unclear if there is any causation or whether this correlation originates from the mutual dependence of the two on other factors. In general, more densely populated areas get electrified first. Thus, electrification might instead just be correlated with household density which in turn may be a proxy variable for further factors such as access to collectable biomass and stores selling LPG. There are several explanations offered for this correlation. Barnes et al., (Barnes et al., 2004) gives two possible explanations: "access [to electricity] proxies for market development. In that case, fewer barriers would constrain other modern fuels in cities where electricity is available" and "availability of lighting and other appliances spurs people to a greater acceptance of modernity and modern fuels". Heltberg (2005) adds two more: "areas that are in some sense more 'modern' (for example large as opposed to small towns and places with better infrastructure) get connected first to the grid and "assume that energy needs are organized in a hierarchy where electricity is the most desired innovation and modern fuels follow further down the list of priorities".

A possible explanation for the found correlation between electrification and LPG usage is that denser populated areas receive electricity and businesses selling LPG before sparsely populated areas. Furthermore, if there is an initial fee for a household to connect to the grid, electrification may also be a proxy for high income. A World Bank study focusing on the benefits of rural electrification found that electrification improved farm-based households' incomes (Khandker et al., 2009). The suggested mechanism for raising the income is the utilization of electric pumps for irrigation which in turn leads to higher yields from agriculture. Thus, electrification may, to some extent, induce fuel switching by raising household incomes.

Two recent reviews of rural fuel choices showed that the set of variables used to model fuel usage varied extensively between different studies (van der Kroon et al., 2013; Lewis and Pattanayak, 2012).

The primary aim of this paper is to perform an exploratory data analysis in order to obtain a set of variables associated with households that are or will start using LPG, utilizing variables from previous studies as a starting point.

The traditional econometric approach of linking fuel usage to factors that could influence it employs economic theory to derive a model for statistical regression. However, econometric approach may fail to capture all factors that influence rural household's choices in regards to cooking fuel. Although the approach in this paper may not be able to fully describe the decision process behind rural household's fuel choices either, it may complement previous methods. Many household surveys collect an extensive number of variables which are used in predicting household fuel choices. Although the inclusion of these variables in models of household's fuel use would not necessarily increase the accuracy of interpretation, heavy dependence on factors previously not included may lead to alternative interpretations and hypotheses. Furthermore, what influences a household's willingness to adopt LPG may be due to a number of factors and involve complex interplay between them, possibly leading to non-linear relations. For these reasons, previous approaches may benefit from being complemented by alternative methods from the field of machine learning. Instead of assuming certain relationship forms, properties are detected in the data and are then generalized and used for prediction. A relatively new algorithm from the field of machine learning, *Random Forests* (Breiman, 2001), has been successfully used in a variety of fields including ecology (Cutler et al., 2007), gene selection (Díaz-Uriarte and De Andres, 2006) and criminology (Berk et al., 2009).

The Random Forest algorithm, as proposed by Breiman (2001), displays many properties that make it suitable for exploratory data analysis. In addition to being one of the most efficient classifiers across many different data types and being able to handle complex relationships, including unknown interactions, it also provides measurements of variable importance, i.e. the impact that these variables have on the classification. This will be further explained in Section 3.1. Random Forest is a non-parametric method, in the sense that no assumptions on the form of the relationship between the response, in the present case fuel usage, and the explanatory variables, i.e. income and education, are needed. The results can then be used as guidance for constructing parametric models which may be compared to Random Forest as a benchmark (Strobl et al., 2009a) or suggest further research.

2. Data

The paper is based on data originally collected for the evaluation of the Rural Electrification Program in Vietnam. The evaluation was undertaken jointly by the World Bank and the Institute of Sociology (IOS) of the Vietnam Institute of Social Sciences. The evaluation was initiated in 2001 and household level data was collected from 42 rural communes in seven provinces at 3 time points, in 2002, 2005 and 2008. Provinces included in the study were drawn from six out of eight regions of Vietnam – from the southern tip to the northernmost mountains. Both the IOS (2009) and the World Bank (Khandker et al., 2009) have released reports based on this data. Neither of these reports focused on household fuel usage for cooking.

Of the 42 participating communes, 22 were in the process of being electrified as part of the electrification program, 13 were not part of the electrification program and seven were already electrified (IOS, 2009). The chosen arrangement enabled comparison of the development in the different kinds of communes in the original studies based on this data (Khandker et al., 2009; IOS, 2009). From Vietnam's eight regions, six were considered for sampling; the remaining two were not part of the studied electrification program. In a stratified sampling approach communes were chosen in a following way: one commune that has already been electrified, three communes that were part of the electrification program and two communes that were assumed to receive no electrification by the end of the study (Khandker et al., 2009). In each commune 30 households were sampled, stratified according to income category: ten of the poorest, ten middle income households and ten of the richest households. The sampling can thus be considered representative for most of Vietnam, except the excluded regions; the Red River delta and the North East region.

The total sample size is 1207 households in 2008 and 1259 in 2005. The data collection experienced a relatively low level of attrition with only 53 households missing in the 2008 data of those that were present in 2005. The questionnaire covered the following areas: household composition, socio-economic questions, agriculture, private infrastructure, health, credits and savings and energy (fuel consumption). The Institute of Sociology in Vietnam administered the questionnaire.

3. Methods

The combustion of biomass is still the most common way for obtaining the necessary energy to prepare meals in rural Vietnam. Although many households with access to electricity use electrical rice cookers, cooking is seldom performed solely using electricity. Instead households often continue to use biomass alongside their electrical rice cookers. Additionally, some households employ either coal or LPG. In this paper, fuel switching is defined to have occurred if a household uses LPG at least occasionally. Households are consequently divided into two groups, LPG users and non-LPG users. It is, however, important to note that LPG users include all households that use LPG, including those using it occasionally, partially or for all their cooking.

In this study a modification of the original Random Forest algorithm, namely Random forests constructed from conditional trees (Strobl et al., 2009b), is used to classify the households into either LPG or non-LPG users. This section continues with a short description of the Random Forest algorithm before moving on to how the Random Forest algorithm is applied in this study. This section ends with a discussion of the variables included in the analysis.

3.1. Random forests and bagging

Random Forest is an ensemble method based on Classification and Regression Trees (CART). In CART, a search algorithm finds an optimal cutting point in one of the variables, which is used to divide the data set into two subsets. The two subsets are then further subdivided, until a stopping criterion is reached (Breiman, 2001; Strobl et al., 2009a). The final result is a tree structure with "nodes" and "leaves," in which the data is divided into purer groups within each node, until a final grouping, the "leaves," is reached. In each such group the number of observations that belong to a certain class is counted, and the class belonging of this group is then decided to be the class that most observations belong to. To make a prediction from such a tree, one need only to check which "leaf" the new observation belongs to, and, consequently, which class it should have. However, a single tree is very sensitive to small changes in the data. That is, repeating the experiment by collecting new data or bootstrapping, i.e. resampling from the original data, often leads to completely different trees.

Random Forest can be viewed as a method of stabilizing the CART algorithm. It grows a large number of trees, i.e. a forest, where each tree is based on a bootstrap sample of the original data, and then classifies a new observation based on how the majority of the trees classify the observation.

In the original Random Forest, CART were used as trees (Breiman, 2001; Breiman et al., 1984). In this paper a modification of the original CART trees, based on conditional inference trees (Hothorn et al., 2006) is used. The conditional inference trees are more suitable when the predictor variables includes both continuous and factor variables (Strobl et al., 2009b). In CART the search for an optimal split is performed in all considered variables concurrently, hence variables with many possible splits, i.e. continuous and factor variables with many levels, are more likely to be chosen by chance. In conditional trees this is avoided by separating the procedure of selecting a variable and the choice of split in this variable. For an extensive introduction to Random forests and related algorithms, see for example Strobl et al. (Strobl et al., 2009a).

Each tree in the forest is built using only a subsample and to decrease the correlation between the trees, only a subset of the included variables is considered for the current split in each node (Breiman, 2001). The number of variables to consider in each split is a tuning parameter and has to be sought for each data set. If this number is equal to the total number of variables, the procedure is known as bagging (Breiman, 1996).

Since each tree is built from using about two thirds of the data points, when the forest is constructed, for each data point there exists trees that did not use this point for construction. Hence, this enables an internal cross validation procedure in which each data point is predicted using only the third (approximately) of the forest that did not have access to this data point during construction. The resulting prediction error is referred to as the out of bag error (OOB-error) (Breiman, 2001).

One of the most important outcomes in a study such as this is the ranking of variables according to their importance for prediction. In this paper, permutation based importance is used (Strobl et al., 2009a). Permutation based importance means that, after the forest is constructed, each independent variable is permutated so that the link between it and the dependent variable is lost. The decrease in predicting capability of the forest after the variable has been permutated is then the importance value. However, when the predicted classes are imbalanced, instead of focusing the decrease in prediction after permutation, the variable importance is calculated as the decrease in the area under the curve (AUC), where the curve is the Receiver operation characteristics (ROC) after permutation (Janitza et al., 2013).

Because, in each node, only a subset of the variables are considered it is possible that variables that are correlated with other variables that are better predictors also receive a high importance. A conditional importance measure, that performs the permutation only within intervals corresponding to splits in correlated variables and therefore reduces the importance given to variables that are not needed when better predictors are available have been proposed (Strobl et al., 2008). However, simulation have shown that importance based on bagging also approach a conditional interpretation (Grömping, 2009). This gives the possibility for certain variables to decrease the importance of other correlated variables should it describe the household propensity to use LPG better than the correlated variables.

Although the assumed functional form need not be specified, this is not a straightforward result from the algorithm either. Rather the outcomes, besides a black box classifier, include a ranking of the importance of the variables considered and the so-called partial dependence for each variable. The partial dependence describes the influence of a variable on a certain outcome.

The partial dependence is defined as (Berk et al., 2009):

$$\widetilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}),$$

Where x is the variable for which partial dependence is of interest and x_{iC} include the other variables.

For classification purposes, the function f is:

$$f(x) = \log p_k(x) - \frac{1}{K}\sum_{j=1}^{K} \log p_j(x)$$

where K represents the number of classes, k the class for which the partial dependence of variable x is sought and p_j the proportion of votes for class j. A new data set is created for each value of variable of interest, in which this variable is constant in each data set whereas all other values take on their original values. Each data set is then used to predict a single value. These values from each data set are then used to construct a graph predicting the function of the variables assigned different values between data sets. No adherence to possible correlations is included and generated data points used for creating the partial dependence may include data combinations never observed in the original data. The partial dependence cannot be interpreted as a causal relationship and is not based on a statistical model (Berk et al., 2009). However, in an exploratory study, the ability to detect possible nonlinearities and interactions can be very useful for further analysis. To detect possible interactions, the partial dependence can be plotted conditional on different values of other variables; differences in the shape of the relationship can then indicate interaction effects.

4. Bagging and random forest application in this study

Because the advantage of bagging over random forest when interpreting the importance values, the starting point was bagging, i.e. all variables were considered in each node. However, to check that any predictive capacity was not lost due to this approach, which could have indicated that bagging did not capture as many relations as random forest, the predictive power was also compared to the random forest approach, i.e. considering only a subset of variables for each split. The number of variables to consider in each split was explored through testing all possible values and examining the prediction error (PE), true positive rate (TPR) and false positive rate (FPR).

All the included variables were ranked according to their importance, described in Section 3.1. To evaluate the stability of the results, multiple forests were constructed using subsamples of the data. A hundred data sets were assembled through subsampling of the original data, and random forest models created for each of them, and subsequently importance values for each forest were recorded. The importance values for the hundred forests that use all variables are presented as boxplots, where each box reflects the dispersion of importance values of a certain variable over the forest.

Furthermore, two different subsampling strategies were employed. One strategy was simply to draw two thirds of the complete data set,

while the other strategy was to first draw two thirds of the communes and then two thirds of the households in each of the selected communes in each draw. This was done in order to explore possible effects of variations in data collection both assuming this to occur on household level and commune level. All draws were made through subsampling, i.e. without replacement. OOB errors are used for importance calculation which is the standard random forest method. However, for estimating the prediction error, the non-selected data points for each subsampling round were used.

Besides ranking the variables according to their importance values a variable selection method was also used, the Diaz-Uriarte method, a backward selection algorithm developed for Random Forests (Díaz-Uriarte and De Andres, 2006; Diaz-Uriarte, 2007). The Diaz-Uriarte variable selection algorithm uses the importance ranking and then removes the least important variables until the OOB error increases. The same two bootstrap strategies as used for the importance ranking were also used for the variable selection procedure.

Partial dependence plots for certain relationships that were ranked as highly important were also produced. Possible interactions between variables were detected, by plotting the partial dependence conditional on values in other variables.

Besides the prediction of current LPG use in 2008, using the variables listed in Table 1, data from 2005 was used to predict which households that would start to use LPG by 2008. In this case households that were already using LPG in 2005 were removed from the sample.

The R software with the package “party” was used for the modelling; please see (Strobl et al., 2009b) for further details.

4.1. Variables

The data were used for two separate predictions with slightly different data input. First, current (2008) use of LPG was predicted using various variables from 2008, together with the household's income levels in 2005 and 2002. Secondly, with the use of household data from 2005, as well as income level in 2002, the model was set up to predict whether a household would have started to use LPG by 2008. The second analysis is limited to the households that did not use LPG in 2005.

For the first model, the prediction of LPG use in 2008 based on the household data collected in 2008, the data contains 1207 households of which 215 are LPG users. For the second model, the prediction of which households would start to use LPG by 2008 there are 1064 households in the data that did not use LPG in 2005 and were present in the 2008 data set, of which 101 started to use LPG between 2005 and 2008.

A recent review of seven papers revealed a large difference in the type of variables that have been used in models that describe fuel switching (van der Kroon et al., 2013). In this paper, many of these variables, as well as several related variables, were tested for their predictive abilities and were ranked accordingly. The independent variables are listed in Table 1 together with a classification inspired by van der Kroon et al. (2013). The last column indicates whether the variable is a factor variable, and if so the number of levels, or if it is a continuous variable.

In the previous studies (e.g. see van der Kroon et al. (2013), Lewis and Pattanayak (2012)), the most commonly used variable for modelling fuel switching has been income. Three income related variables were included in this study, including the total income as well as income from wage work, as opposed to self-empland agriculture and finally income of the spouse of the head of the household. Wage income was included because it can signify a steadier stream of income (compared to income from agriculture) and more time spent away from home. Furthermore, it has been reported that the burden of fuel wood collection and/or cooking often falls on women in the family whereas a male head of the household usually is in charge of household finance. This has been proposed to be one of the barriers to fuel switching and the adoption of improved cooking stoves because the head of household may not value time savings and a cleaner cooking environment for other family members (Miller and Mobarak, 2011). Hence the wage of the spouse was included as a separate variable as well as the gender of the head of the household. The same reasoning was applied to the various education measurements. Note that several of the variables are dependent on each other, which, in an ordinary regression analysis, may cause the problem of multicolinearity. For example, the number of appliances is a likely indicator of previous household wealth. Furthermore, total land holdings of a household can be seen both as an indicator of wealth, but this private land area may also provide possibilities for the collection of biomass fuels, beyond that which is possible on public land.

Table 1
Variables included in the analysis.

Category[a]	Variable	Levels (C is continuous)
Household business	Household business	2
	Hours worked in household business	C
	Restaurant	2
	Shop	2
Education	Highest education in household	C
	Education of head of household	C
	Education of spouse	C
Household composition	Number of family members	C
	Share of women	C
	Share of children	C
	Share of adults with income	C
	Gender of head of household	2
External decision factors and environment	Collection rate	C
	Relative fuel price	7
	Distance to paved road	C
	Commune wealth ranking	3
	Distance to nearest town	42
	Village average land	42
Age	Age of head of household	C
	Age of spouse	C
Incomes	Total income	C
	Income from wage	C
	Income of spouse	C
Appliances	Number of appliances	C
	Rice cooker	2
	Electric fans	2
	Water heater	2
	Refrigerator	2
	TV	2
Occupation	Farm	2
	Industry	2
	Office	2
Additional wealth measurements	Type of house	5
	Total land of household	C
	Total income in 2005	C
	Total income in 2002	C
Electrification	Electrified	2
	Years electrified	C

[a] The categorization of variables is partly inspired by the categories used in van der Kroon et al. (2013) for easier comparison.

The previously discovered difference between urban and rural fuel switching (Heltberg, 2004) together with an acknowledgement areas classified as rural can still be more or less rural calls for some sort of description of the degree of the rurality that may not be captured by household variables to allow for the possibility of this affecting fuel choices. Two variables intuitively related to the degree of rurality were included. The first is the mean area of land holdings of the households that belong to the same commune, henceforth label village average land. Another variable that intuitively is related to rurality is the distance to the nearest town. This variable was not recorded as part of the commune questionnaire, but commune names were used to localize them and find the distance to the nearest town by means of Google maps. Village average land and the distance to nearest town may all be seen as related to fuel wood collection, although collection rate is also included as a separate variable. However, these variables may describe a variety of changes related to fuel choices as the environment becomes more rural, such as availability of LPG, market exposure and various social factors. However, the method employed in this paper is not be able to differentiate between the different explanations why different degrees of rurality, as described by these two variables may be important for fuel choice. Neither can the description of rurality by these two variables be considered to be a complete description of rurality, neither is there any guarantee that these two variables capture all the variation in degree of rurality that may be relevant in connection with household's fuel choice.

Electrification was included as a predictor variable. However, very few households were still not electrified in 2008, hence this may have caused this variable's importance to not be properly reflected in the random forest modelling. To compensate for this fact, the number of years that a household has been electrified was also included as an explanatory variable. However, if the effect of electrification is immediate, also this variable may fail to capture the full extent that electrification has on LPG usage.

The relative fuel price was defined as the cost per kg of fuel wood divided by the cost per kg of LPG. Again, these data were not collected in the commune questionnaire but this value was constructed based on households purchasing fuel wood and LPG. Because of this construction, this value was only obtainable on the province level, i.e. not all communes included sampled households purchasing both LPG and fuel wood. The variable “collection rate” is the mean kg wood collected per person-hour for each commune.

Education is measured in years of schooling. The commune wealth ranking is an official government ranking with three levels and not calculated based on the questionnaire. Type of house includes the following levels ranked from one to five: multi-storied house, permanent single storied house (brick), semi-permanent house (wood), wood frame with leaf roofing, temporary house made of straw with leaf roofing.

The method employed in this paper does not rely on a pre-specified model, instead a large number of variables have been included, of which most has been included (or closely related variables) in previous econometric studies (van der Kroon et al., 2013; Lewis and Pattanayak, 2012). Variables will be further discussed in Section 5 in relation to their respective importance and whether they were selected in the variable selection process.

5. Results and analysis

The tuning parameter, how many variables to be tried in each split was explored for optimal value. Starting with all variables considered in each split, no increase in prediction capabilities could be detected when decreasing the number of variables considered in each split. Hence all the results presented in this section are for when the number of variables considered in each split was set to the total number of variables available for the forest (i.e. bagging).

This section begins with the presentation of the results for the case where LPG usage in 2008 is predicted using data for 2008. The results from the 100 random forest models is presented by means of plotting variable importance (for the full models) and tabulating variable selection (for reduced models). Then the section proceeds to a presentation of the evaluation of the performance of the models by considering the classification rate of LPG usage, both for the full and reduced models. The section then moves on to a presentation of the partial dependence plots of the most important variables.

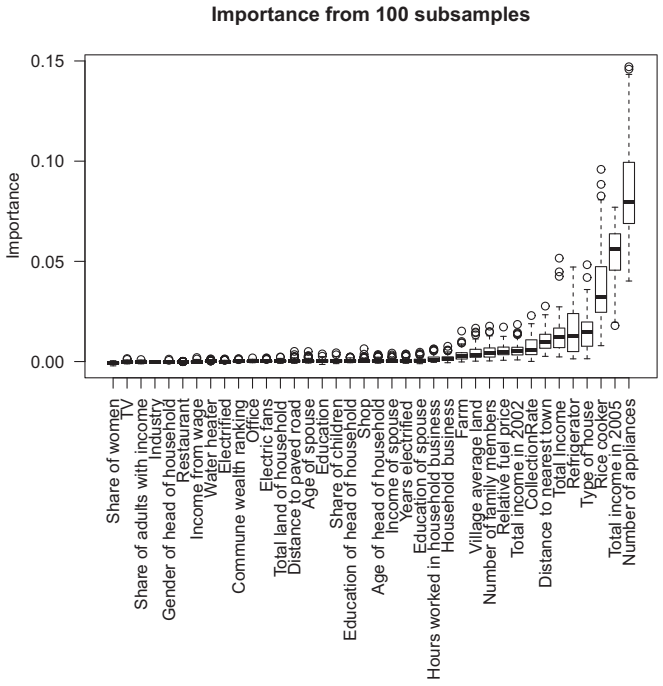

Fig. 1. Bagging AUC-based importance values from 100 subsamples. Subsampling scheme is households from total population. Each subset is two thirds of the data points (805 of 1207). The variables are ordered according to their mean importance values across the forests.

This section ends with a presentation of the variable importance of the case where the 2005 household data are used for predicting whether households will start to use LPG before 2008.

The importance ranking shows that several variables that can be associated with wealth and income are judged to be influential for the correctness of household classification, see Fig. 1. It's interesting to note that total income, income in 2005 and income in 2002 are all deemed to be important for current LPG use. It's worth noting here that the correlations between the income levels for the different years are quite low (< 0.5), a sign of relatively large fluctuations between income levels over different years. Various appliances, such as whether the household owns a rice cooker or a refrigerator are also usable predictors for LPG usage. Further important variables are the distance to nearest town, village average land, collection rate and relative fuel price, all of which can be associated with aspects of the degree of rurality, see section 3.2. It is also of interest that several variables previously used to model fuel switching, such as education, the different occupations and whether the household is electrified or not, are judged to be of low importance, i.e. they are suppressed by the other variables. Similar results as Fig. 1, but with the second bootstrap strategy, randomly sampling both villages and households within villages, can be found in the appendix, please see Fig. A1. It may also be of interest to compare the results in Fig. 1 with the univariate linear correlations between the predictor variables and the, see Fig. A2. Although many wealth related variables, such as income, previous incomes, type of house and certain appliances are both highly correlated with LPG usage and receive high importance, other highly associated variables such as household business, shop, years electrified,

Table 2
Results from variable selection using from 100 bootstrap samples. Sub. HH and Sub. C stands for subsampling on household and commune level respectively as explained in Section 6.

Category	Variable	Sub. C	Sub. HH
Household business	Household business	19	41
	Hours worked in household business	16	10
	Restaurant	0	0
	Shop	14	14
Education	Highest education in household	2	0
	Education of head of household	3	0
	Education of spouse	3	0
Household composition	Number of family members	48	62
	Share of women	1	0
	Share of children	4	0
	Share of adults with income	1	0
	Gender of head of household	0	0
External decision factors and environment	Collection rate	60	74
	Relative fuel price	39	57
	Distance to paved road	16	3
	Commune wealth ranking	0	0
	Distance to nearest town	96	100
	Village average land	57	65
Age	Age of head of household	0	0
	Age of spouse	3	0
Incomes	Total income	98	97
	Income from wage	1	0
	Income of spouse	5	2
Appliances	Number of appliances	100	100
	Rice cooker	100	100
	Electric fans	26	18
	Water heater	0	4
	Refrigerator	95	100
	TV	17	22
Occupation	Farm	17	35
	Industry	0	0
	Office	0	0
Additional wealth measurements	Type of house	93	100
	Total land of household	10	8
	Total income in 2005	98	100
	Total income in 2002	89	95
Electrification	Electrified	0	0
	Years electrified	9	3

the various education measurements are highly correlated but are ranked lower according to importance in the bagging procedure. Some variables receive relatively higher importance compared to their ranking according to the correlations; these are to a large extent the area descriptions, village average land, distance to town and relative fuel price.

The results from the variable selection are presented in Table 2. These results compare well with the importance ranking of variables in the random forest and bagging procedures. Similarly, to the previous results, we can see that the current income, income 2005 and 2002 and other wealth measurements such as number of appliances, refrigerator, rice cooker and type of house, are almost always present in the reduced models. Other variables that often appear are the distance to nearest town, village average land, collection rate and relative fuel price. The distance to nearest town is chosen in almost all reduced models, regardless of subsampling strategy.

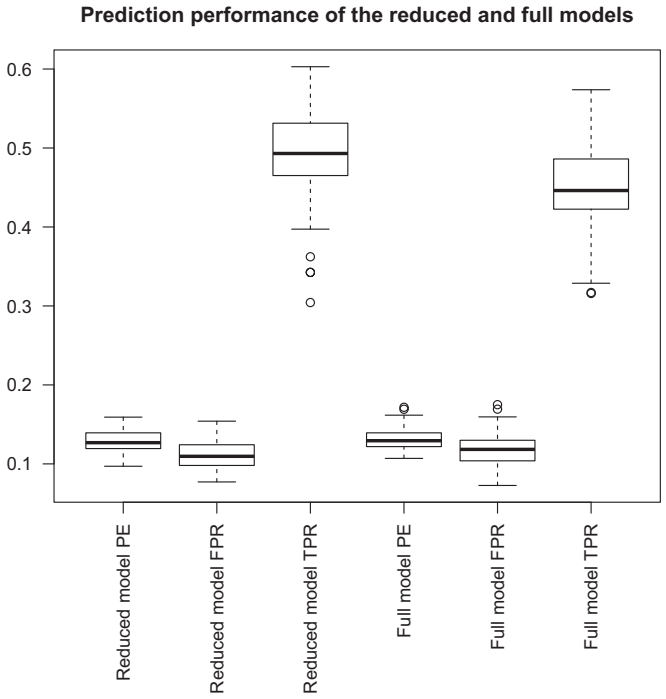

Fig. 2. Prediction performance of the reduced models and full model. Values are based on 100 hundred subsamples of the total data with one forest built on each subsample that attempts to predict the out of subsample data.

The prediction performance of the algorithm before (full model) and after variable selection (reduced models) is presented in Fig. 2. The values here are not based on the OOB-error but the actual out of subsample for each forest. Note that the prediction capabilities were increased after variables selection. The reduced models are forests constructed after variables selection. The true positive rate is only around 50% for the reduced model, and slightly lower for the full model, Fig. 2.

This section now continues with a presentation of partial dependencies of some selected variables with high importance. The partial dependencies are from the reduced model, i.e. after variable selection (included variables are the 12 rightmost variables in Fig. 1), and using the complete data set. The variables that are presented here are income, the most commonly used variable for household fuel use and the rurality variables, village average land and distance to town. However, the other variables used in the reduced model, all has partial dependencies similar to the either income or the rurality variables. The partial dependencies are displayed in Fig. 3 and 4.

The partial dependence of income at different levels of income in 2005 and the number of appliances are shown in Fig. 3. The differences between the partial dependencies, conditional to changes in income in 2005 and number of appliances, are most pronounced at low levels of current income. Note that the income distribution is skewed towards the left, i.e. the range where past income makes the greatest difference is also the range in which the majority of household incomes are observed. The partial dependencies of all income and wealth related variables are similar as the partial dependency for income, i.e. a sharp increase at a certain level, followed by a more moderate increase. Furthermore, the effect of conditioning on different levels of other wealth related variables give a similar effect in the shift of the partial dependence curve as observed in Fig. 3, in the sense that the difference is most pronounced at lower levels of the analyzed variable.

The propensity to use LPG for cooking is declining with an increased village average land and distance to the nearest town, Fig. 4. A striking difference between the shapes of the partial

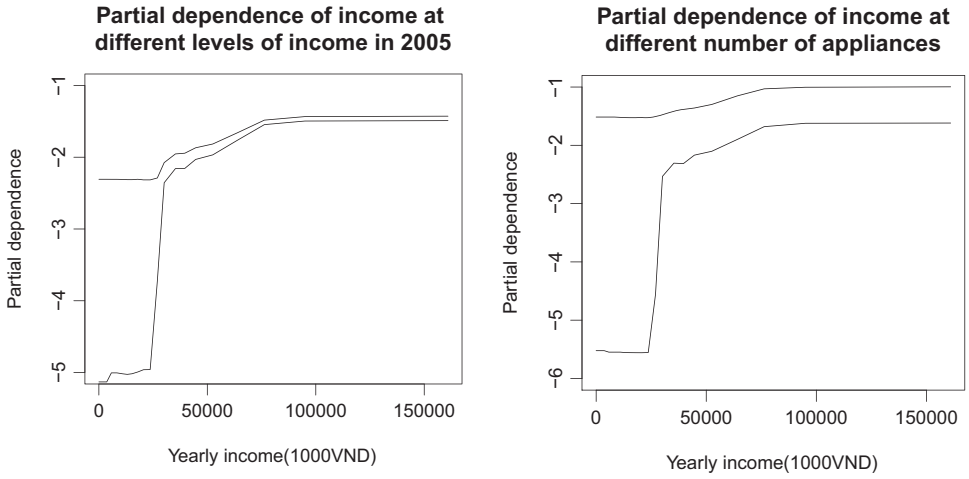

Fig. 3. Partial dependence on income at fixed levels of income in 2005 and at a fixed number of appliances. In left panel, bottom line the income in 2005 has been set to the first quartile for all households, while for top line this has been set to the third quartile. In the right panel is the households number of appliances has been set to first quartile (bottom line) and third quartile respectively (top line).

dependency is visible, for both distance to town and village average land, depending on if income has been set to either the first or the third quartile. In the former, a sharp decline is visible at certain values, while for the high income case, a decline is much more gradual and slow. The partial dependence curve for collection rate has a similar shape as the one constructed for village average land and distance to town.

Although the importance levels and partial dependencies presented in this section are for using LPG at least occasionally, the importance measures and partial dependencies were found to be similar when only households engaging in full fuel switching were considered, i.e. LPG used for cooking all meals.

The results from the random forest when set up to predict which households that would have started to use LPG between 2005 and 2008 are presented in Fig. 5. Current income (2005) together with wealth related variables rice cooker, number of appliances and previous income (2002) were important for determining whether household would have started using LPG by 2008. The area description variables village average land, distance to town, are still ranked high. In these aspects the results compare well with those of the current LPG usage, Fig. 1, however there are some discrepancies as well. A variable that was only deemed moderately important for current LPG usage is now the fourth most important, Farm, which indicates whether a household is mainly occupied within farm or agricultural based activities. Furthermore, the income of the spouse also received relatively higher importance. A farm based household is less likely to start to use LPG while a high income of the spouse increases this likelihood. When considering all current LPG users, the variables chosen in Table 2, are chosen first and appear to encompass Farm and Income of Spouse, i.e. these variables do not add to the prediction, however for the subset of households that started to use LPG recently, these are chosen more often.

6. Discussion

The variables that were ranked the highest according to the importance measure provided by Random Forest and also chosen in the variable selection for current LPG use can be categorized as measurements of wealth together with various commune level descriptions. The wealth variables are income during 2002 and 2005, the type of house, various appliances and the total number of appliances. The commune level variables- the distance to town, village average land and the collection rate can be associated with the opportunity cost of collected fuels and the access to modern fuels, but can also signify further differences between more or less rural areas in terms of fuel use. These results compare well with the results of which households that would start using LPG between 2005 and 2008; however, some additional variables received relatively higher importance, i.e. Farm, which indicates whether a household is mainly occupied within farm or

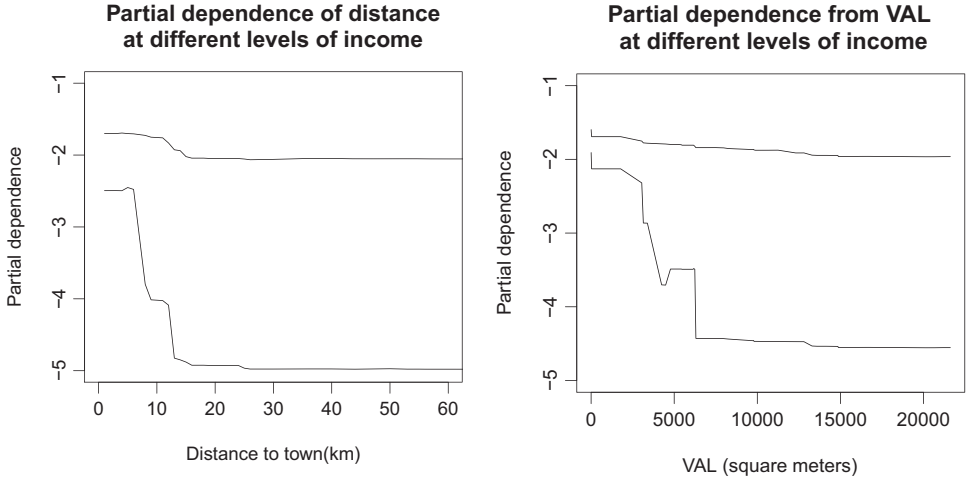

Fig. 4. Partial dependence on Distance to town and village average land(VAL) at fixed levels of income. Left panel is distance to town while the right panel show the partial dependence for village average land. Bottom lines are when income has been set to the first quartile and top line are the third quartile, for both panels.

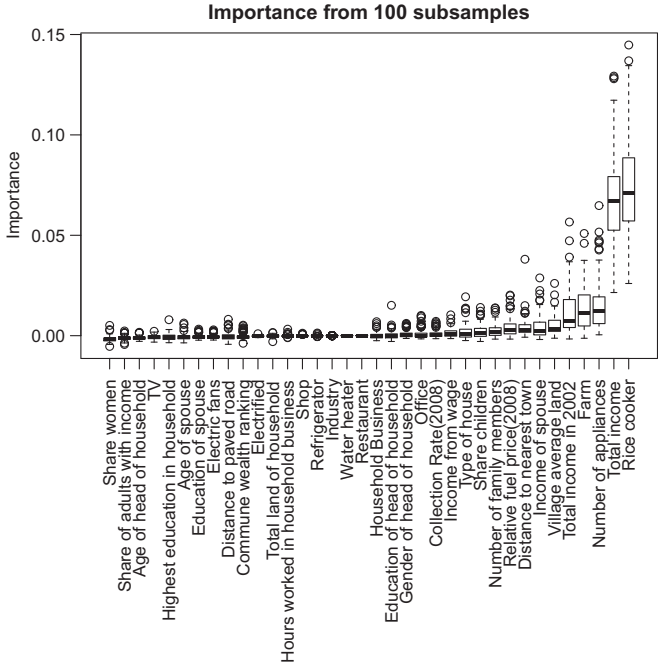

Fig. 5. Bagging AUC-based importance values from 100 subsamples. Subsampling scheme is households from total population. Each subset is two thirds of the data points (709 of 1064). The variables are ordered according to their mean importance values across the forest.

agricultural based activities and the income of spouse.

One of the indicators of wealth that received a high importance is the number of household appliances. Possible explanations for the importance of this parameter are that it signifies a high and stable income over a long period of time; this can be compared to the high importance placed on measurements of incomes during 2002 and 2005. Another variable that received high rankings in both importance measurements is whether households owned a refrigerator, ranked as the third most important parameter according to conditional importance. Similar to the number of appliances, this could be interpreted as a wealth factor and might signify a wealthy household. It is also interesting from the perspective of possible connections between electrification and LPG uptake. The open fire is used not only as a means of cooking but also as a food preserver in certain areas of Vietnam and China. Hanging the food in the ceiling above the fire, thereby exposing the food to smoke, enhances preservation. A refrigerator together with an LPG stove may provide satisfying results both in terms of cooking and food preservation; after purchase of a refrigerator, one of the benefits of the open fire would be reduced. However, the results presented in this paper cannot distinguish between these explanations. Because of an unmeasured potential confounder, it would also be difficult to use regression techniques to measure any possible causation; i.e. it is possible that regardless of any co-benefits, households that prioritize kitchen improvements invest in both refrigerators and LPG stoves if they have the ability to do so. This also holds for the characteristic of owning a rice cooker, despite showing up as important also for which households that would later obtain LPG, i.e. whether this signifies wealth, an unknown household characteristic, or whether rice cookers have another connection with LPG. Although, in the light of which type of other variables that where shown to be important, the interpretation of these appliances as signifying wealth seems plausible and possibly a preference order in which household obtain certain goods, as suggested by Heltberg (2005). This is further underscored through the partial dependence on income for different values of past incomes and the number of appliances as shown in Fig. 3. If there are signs of past high household income at low income levels, the probability of a household using LPG was substantially increased. These households may have had the opportunity to purchase an LPG stove in the past and despite lower current incomes, they continued to use LPG. However, the past income and appliances are indicative also for making future fuel switching (see Fig. 5), indicating the possibility that households need a stable high income in order to make an investment in an LPG stove.

Indicators of wealth, apart from current income, are seldom to any large extent included in studies of fuel switching (van der Kroon et al., 2013). When they are, they are sometimes interpreted as drivers of fuel switching instead of being interpreted solely as indicators of wealth. Although the results in this paper do not refute such assumptions, the high importance of past income levels in combination with these wealth factors may support the interpretation that it is a stable economic situation for a longer period of time or a previous high income that enabled households to make the investment. This finding can also be compared to previous noted strategy among poor households to behave in a risk averse manner (O'Keefe and Munslow, Understanding).

Village average land, distance to town and collection rate variables were also found to be usable predictors of LPG usage. A household with constraints on liquid funds may have to consider the cost and benefits of an LPG stove before purchase. For a household that have access to collectable wood and with a low opportunity cost for time, the difference in cost between cooking with wood and LPG is larger than for a household that find it harder to collect wood or have to purchase wood from a market. It is possible that village average land, distance to town and collection rate all capture this effect when predicting LPG usage. Furthermore, this effect seemed to be less accentuated for higher incomes, see Fig. 4, i.e. the partial dependence of village average land and distance to town were not as drastic for high income levels, indicating that at a certain level of wealth and income the effect of cost of fuel wood collection becomes less important. Note however, that these location variables may capture not only effects regarding fuel wood collection but also level of access to LPG, social practices as well as further unknowns.

Because more densely populated areas closer to towns generally obtain general development before rural and sparsely populated areas, doubts can be cast on the findings that development in other areas (Heltberg, 2004), would necessarily cause fuel switching. Electrification has been found by several studies to correlate with LPG usage (e.g. (Barnes et al., 2004; Heltberg, 2004) and no firm conclusion concerning this effect has been reached (Köhlin et al., 2011). Though the number of years electrified is correlated with LPG usage, Fig. A2, its importance came close to zero, both for current LPG usage, Fig. 1, and fuel switching, Fig. 5, indicating that there exist other variables in the set that better describe the relationship, in the random forest model. As noted above, a possible explanation for the correlation between LPG usage and electrification found in previous studies is that areas closer to a town with a higher population density get connected before more sparsely populated and distant areas. Furthermore, more well-off households in richer areas are also likely to be electrified before others. However, Khandker et al. (2009) found that electrification and the number of years electrified increased household income, mainly attributable to an increased usage of pumps for irrigation in the agricultural sector, which in turn could affect fuel choice. Thus, it is possible that electrification influence fuel choice by increasing household income. Furthermore, as mentioned in the previous paragraph, there is a possible connection between certain electric appliances and LPG.

Village average land, distance to town and collection rate were not as important for the prediction of which households that would start to use LPG between 2005 and 2008. Instead the algorithm gave more importance to Farm. Note however that the influence of the location description variables is most pronounced at low values, i.e. when removing households that already had LPG in 2005 many households from these areas are excluded from the analysis. The income of the spouse also received higher importance when the algorithm tried to

predict future LPG usage. There are several possible explanations why the income of the spouse may be important beyond adding to total income; the spouse may have more influence over decisions, have a higher opportunity cost for time, i.e. more expensive to collect wood and to spend long hours cooking, and more income sources may mean a higher stability in household income. A possible explanation for why this variable was not chosen in the 2008-year data set can partly be because, as for total income, the past years' values are more correlated with 2008 use of LPG than the income in 2008 (and spouse's income in 2008) but also because it is partly covered by the appliances.

Because poor households are not able to make lump sum payments and cannot afford high upfront costs (Leach, 1992), poor households in developing countries often end up paying a higher price for energy services. This is often the case for fuel wood which in many cases is more expensive than LPG, on a per meal basis, but where large lump sums, when changing canisters, and upfront costs for new stoves are avoided (Heltberg, 2005). The importance of the history of household income and wealth for both current LPG use and future fuel switching points towards the possibility of subsidizing LPG stoves, or fund payment plans, rather than the LPG itself. Although additional factors are pointed out as drivers, such as regulated price of kerosene in Ethiopia and rising incomes in Grenada, there exists successful implementations of such polices in Ethiopia (kerosene stoves) and Grenada (LPG stoves) (Leach, 1992). Innovative schemes involving cooperative based solutions for households to obtain LPG have lately been proposed and initial pilot programs have met with positive results (Nayak et al., 2014).

The data set used in this study was collected in order to compare Vietnamese communes that were part of the rural electrification project to non-project communes. It could therefore be disputed that the communes were not necessarily sampled in a random fashion, questioning the representativeness of this study. However, there are no apparent reasons to assume that the fuel switching process in these communes is different from other Vietnamese communes. Further possible flaws in this study include the imperfect measurement of the variables such as relative fuel price, collection rate, as well as village average land and distance to nearest town. The relative fuel price is measured on a province level whereas village average land, distance to town and collection rate are measured on a commune level. There is thus a possibility that village average land and the distance to town capture the effect on the commune level which is not properly reflected by the relative price variable.

It is important to note that the results in this paper indicate a minimal set of variables needed to achieve high prediction as supported by the data, and are not a proof of causality. Neither do the results prove the absence of causal effect on fuel choice from variables that are not chosen. Such omission can be explained by limited data, but also the fact that the algorithm had access to intermediates or descendants of intermediates in a causal chain. As an example many of the variables suppressed in the importance ranking, but correlated with fuel choice, are likely causes of a higher and more stable income, such as education, household business, and possibly electrification. It should also be noted that the importance ranking may be influenced by the absence of certain variables, i.e. a variable that is ranked low here may be ranked higher if it is important by virtue of interacting with variables not included in this analysis. However, the variables included in this paper cover most categories considered in previous studies (van der Kroon et al., 2013). The lack of importance for certain variables and the high importance of others does neither prove nor disprove any actual cause and effect, however pose the question as to why the chosen variables are able to discriminate between the types of households in a more efficient way.

A downside in using a non-parametric model include the interpretation of partial dependencies which cannot be interpreted in any causative or absolute way, but only as descriptive (Berk et al., 2009). However, an exploratory analysis based on parametric techniques would not provide any useful causal relationships either if not firmly motivated by theory due to potentially omitted variables or included intermediates. Compared to such approaches, the Random Forest algorithm provides more stable results, which depend less on the researchers' assumptions and which can be used to guide further modelling and research.

7. Conclusion

Current income is unable to fully account for a household's wealth and history of income, which in turn appear to be associated with both the current and future use of LPG. This could also indicate that households need a stable income over a longer period of time in order to consider fuel switching, either because of the risk of adverse behavior or because of a common preference order in which appliances are purchased.

The likeliness that a household is or will start using LPG increases with increased wealth if the household resides in a less rural environment. A measurement of rurality could therefore be included in further projections of fuel switching to capture variations in this process for different areas. This could also signify the possibility that policies may reach different level of effects in different areas. However, more research is needed to fully understand the connection between different measurements of rurality and fuel use for the purpose of predicting future fuel choices.

An alternative explanation for some of the variables found to correlate with fuel switching in previous studies, but suppressed by other variables in the random forest/bagging model, is the association of these variables with wealth and stable income such as education, occupation and household business, or with their association with rurality such as electrification.

Acknowledgements

The following are gratefully acknowledged: Rebecka Jörnsten, Vera Liskovskaja, Olle Nerman. Erik Ahlgren for commenting on earlier versions of this manuscript, The Institute of Sociology, VASS, for kindly sharing their data and thehe anonymous reviewers for improving the manuscript through comments and suggesting better ways of explaining certain issues.

Appendix A

See Fig. A1 and A2 here.

In this section the boxplots of the importance values for the variables based on the alternative bootstrap strategy are presented.

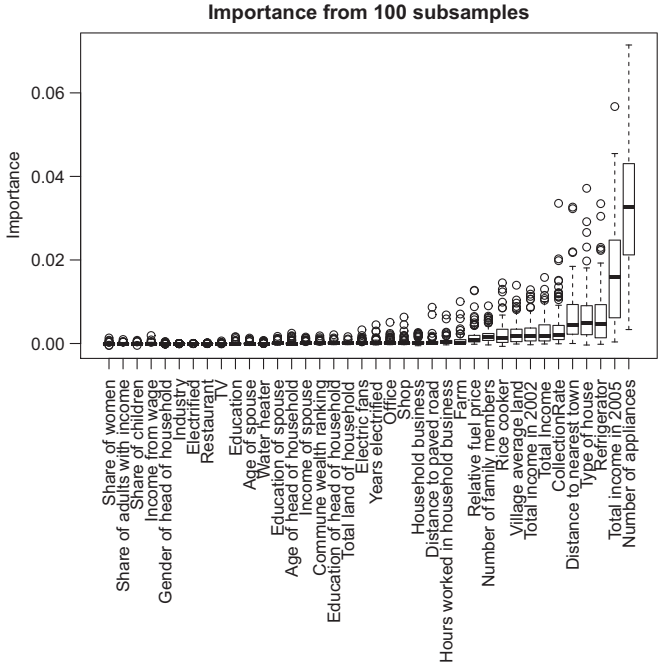

Fig. A1. Bagging based importance values from 100 subsamples for households currently using LPG. Subsampling on commune and household level; two thirds of the communes and two thirds of the households were sampled in each sampled commune for each round.

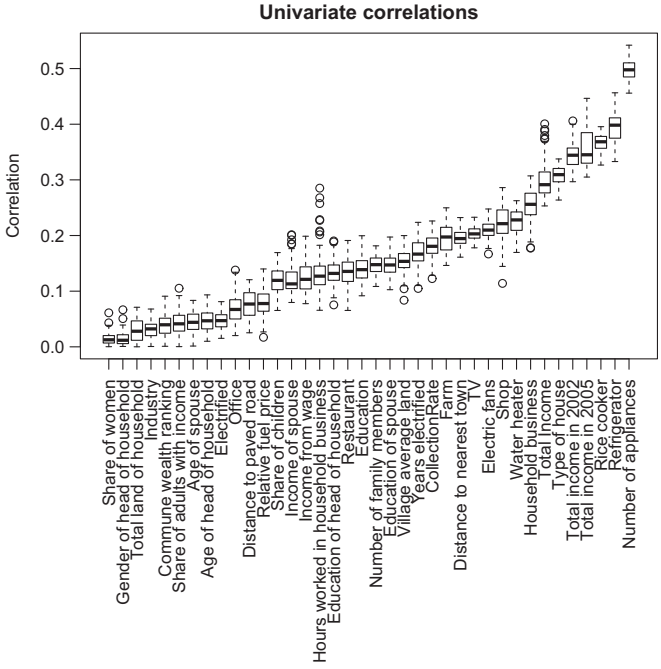

Fig. A2. Absolute values of univariate correlations from 100 subsamples. Subsampling on household level, cfr Fig. 1.

References

Barnes, D.F., Krutilla, K., Hyde, W., 2004. The Urban Household Energy Transition. Energy, poverty, and the environment in the developing worldWorld Bank Washington DC; ⟨https://www.esmap.org/sites/esmap.org/files/Rpt_UrbanEnergyTransition.pdf⟩.

Berk, R., Sherman, L., Barnes, G., Kurtz, E., Ahlman, L., 2009. Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. J. R. Stat. Soc.: Ser. A (Stat. Soc.). 172 (1), 191–211.

Bond, T.C., Sun, H., 2005. Can reducing black carbon emissions counteract global warming? Environ. Sci. Technol. 39, 5921–5926, (2005).

Bond, T.C., Zarzycki, C., Flanner, M.G., Koch, D.M., 2011. Quantifying immediate radiative forcing by black carbon and organic matter with the specific forcing pulse. Atmos. Chem. Phys. 11 (4), 1505–1525.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees (CART). Wadsworth International Group, Belmont, CA, USA.

Bruce, N., Perez-Padilla, R., Albalak, R., 2000. Indoor air pollution in developing countries: a major environmental and public health challenge. Bull. World Health Organ. 78, 1078–1092.

Campbell, B.M., Vermeulen, S.J., Mangono, J.J., Mabugu, R., 2003. The energy transition in action: urban domestic fuel choices in a changing Zimbabwe. Energy Policy 31 (6), 553–562.
Cutler, D.R., Edwards, Jr.T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., et al., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792.
Davis, M., 1998. Rural household energy consumption. The effects of access to electricity - evidence from South Africa. Energy Policy 26 (3), 207–217.
Desai, M.A., Mehta, S., Smith, K., 2004. Indoor smoke from solid fuels. Assessing the environmental burden of disease at national and local levels Environmental Burden of Disease Series(4).
Diaz-Uriarte, R., 2007. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinforma. 8 (1), 328.
Díaz-Uriarte, R., De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinforma. 7 (1), 3.
Foell, W., Pachauri, S., Spreng, D., Zerriffi, H., 2011. Household cooking fuels and technologies in developing economies. Energy Policy 39 (12), 7487–7496.
Grömping, U., 2009. Variable importance assessment in regression: linear regression versus random forest. Am. Stat. 63 (4), 308–319.
Gundimeda, H., Köhlin, G., 2008. Fuel demand elasticities for energy and environmental policies: indian sample survey evidence. Energy Econ. 30 (2), 517–546.
Heltberg, R., 2004. Fuel switching: evidence from eight developing countries. Energy Econ. 26 (5), 869–887.
Heltberg, R., 2005. Factors determining household fuel choice in Guatemala. Environ. Dev. Econ. 10 (03), 337–361.
Hosier, R.H., Dowd, J., 1987. Household fuel choice in Zimbabwe: an empirical test of the energy ladder hypothesis. Resour. Energy 9 (4), 347–361.
Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15 (3), 651–674.
IOS, 2009. Impacts of rural electrification in Vietnam. Institue of sociology in Hanoi.
Janitza, S., Strobl, C., Boulesteix, A.-L., 2013. An AUC-based permutation variable importance measure for random forests. BMC Bioinforma. 14 (1), 119.
Khandker, S., Barnes, D., Samad, H.A., Minh, N.H., 2009. Welfare impacts of rural electrification: evidence from Vietnam. world Bank policy. Res. Work. Pap. No 5057⟨http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1476699⟩.
Köhlin, G., Sills, E., Pattanayak, S., Wilfong, C., 2011. Energy, gender and development: what are the linkages? Where is the evidence? World Bank Policy Research Working Paper No5800; ⟨http://papers.ssrn.com/sol3/papers.cfm?Abstract_id=1931364⟩.
van der Kroon, B., Brouwer, R., van Beukering, P.J.H., 2013. The energy ladder: Theoretical myth or empirical truth? Results from a meta-analysis. Renew. Sustain. Energy Rev. 20 (0), 504–513.
Leach, G., 1992. The energy transition. Energy Policy 20 (2), 116–123.
Lewis, J.J., Pattanayak, S.K., 2012. Who adopts improved fuels and cookstoves? A systematic review. Environ. Health Perspect. 120 (5), 637.
Masera, O.R., Saatkamp, B.D., Kammen, D.M., 2000. From linear fuel switching to multiple cooking strategies: a critique and alternative to the energy ladder model. World Dev. 28 (12), 2083–2103.
Miller, G., Mobarak, A.M., 2011. Intra-Household Externalities and Low Demand for a New Technology: Experimental Evidence on Improved Cookstoves. Unpublished manuscript; ⟨http://casi.sas.upenn.edu/sites/casi.sas.upenn.edu/files/iit/Miller%20and%20Mobarak.pdf⟩
Nayak, B.P., Werthmann, C., Aggarwal, V., 2014. Trust and cooperation among urban poor for transition to cleaner and modern cooking fuel. Environ. Innov. Soc. Transit..
O'Keefe, P., Munslow, B., Understanding. fuelwood. Conference Understanding fuelwood, vol. 13. Wiley Online Library, p. 11–19.
Ramanathan, V., Carmichael, G., 2008. Global and regional climate changes due to black carbon. Nat. Geosci. 1 (4), 221–227.
Smith, K.R., Peel, J.L., 2010. Mind the Gap. Environ. Health Perspect. 118 (12), 1643–1645.
Strobl, C., Malley, J., Tutz, G., 2009a. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14 (4), 323.
Strobl, C., Hothorn, T., Zeileis, A., 2009b. Party on!. R. J. 1/2, 14–17.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinforma. 9 (1), 307.
WorldBank, 2011. Household Cookstoves, Environment, Health and Climate Change: a New Look at an Old Problem (63217). World Bank, Washington, DC, (2011).