Tanzania Journal of Forestry and Nature Conservation, Vol 90, No. 2 (2021) 115-133

DEVELOPMENT OF WOOD FUEL CONSUMPTION PREDICTIVE MODEL IN TANZANIA

[1]L.P. Lusambo and [2]G.E. Mbeyale

[1]Department of Forest and Environmental Economics
[2]Department of Forest of Forest Resources Assessment and Management,
College of Forestry, Wildlife and Tourism
Sokoine University of Agriculture Morogoro, Tanzania

Corresponding author: *mbeyale@sua.ac.tz*

ABSTRACT

This study aimed to develop a wood fuel *predictive model* that could be used to give information which can be used to manage woodfuel supply with a view foster forest resources stewardship. The paper has briefly defined *predictive modelling* concepts, highlighted the significance of predictive modelling and described the salient steps involved in constructing predictive models. The paper has explicitly described how the predictive model was developed and validated. In light of the validation results, the paper also highlights the adjustment that has been made to the model to make it more plausible. It is concluded that in the current Tanzanian situation where there is no any model that can be used to predict and/or estimate wood fuel consumption, the developed wood fuel consumption predictive model can be useful in sustainable forest management strategies. Prior to its use, however, the constructed model needs to be further validated and adjusted accordingly using newly collected longitudinal data from the study area. Sufficient data should be collected from the *strata* (locations) commensurate with those used in the present study.

Keywords: Wood fuel, consumption, predictive model, Tanzania, miombo woodlands

INTRODUCTION

Contextual definition of wood fuel as applies to this study

According to FAO (2004), woodfuel is defined as all types of fuels originating directly or indirectly from woody biomass. The main types of woodfuel in less-developed regions of the world are fuelwood and charcoal. Fuelwood is woodfuel in which the original composition of the wood is preserved; it includes wood in its natural state and residues from wood-processing industries. Charcoal is the solid residue derived from the carbonization, distillation, pyrolysis and torrefaction of wood. Sepp and Mann (2009) define woodfuel as firewood and charcoal. Most-commonly used forms of woodfuel include firewood. Firewood represents the largest share in wood energy fuels production and consumption (UNEP 2019). According to Njenga *et al*. (2018), firewood and charcoal constitute sustainable woodfuel. Darko-Obiri *et al.* (2015) asserted that fuelwood is used synonymously as firewood and defined woodfuel as firewood and charcoal. In the context of this study, charcoal was defined as a carbonaceous material obtained by heating wood in the earth-mound kiln, in the absence of air. Firewood was defined as Wood intended to be burned, typically for heat. Woodfuel was defined as **firewood and/or charcoal.** It implies that if a household consumed only firewood, computation of woodfuel consumption involved conversion of firewood (kg) to round wood equivalent (m^3) using appropriate conversion factor.

Similarly, for a household that consumed only charcoal, computation involved conversion of charcoal consumed (kg) into round wood equivalent (m^3) using appropriate conversion factor. For those households which consumed both firewood and charcoal, the woodfuel consumption was a sum of round wood equivalent from both firewood and charcoal. It is apparent therefore that household in the study area consumed woodfuel either in the form of firewood or charcoal

Overview of predictive modelling

Predictive modelling is the process by which a model is *created* or *chosen* to try to best predict the probability of an outcome. According to Mosley (2005), predictive modelling is a form of "*data mining*". Data mining is *analysis of observational datasets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner*. Predictive modelling takes these relationships and uses them to make inference about the future (ibid). Essentially, prediction is all about using *historic experience to attempt to predict the future outcome*s (ibid). Effective predictive modelling can and does enhance planning, decision making and natural resource management (Mosley2005, Harrell 2008, Dorazio and Johnson2003). A good manager is not so much one who can *minimise the effects of the past mistakes*, but rather the one who *can successfully manage the future* (Gilchrist 1978).

There are several types of models that can be fit to the data including *linear models, logistic regression, Markov models, neutral networks, Bayesian networks, regression splines, decision tree analysis*, and *classification and regression trees* (Zukerman and Albrecht 2001, Mosley 2005). Crawley (2009, p387) defined regression analysis as a statistical method used when both response variable and explanatory variables are continuous variables, and grouped them into seven categories: *linear regression* (the simplest and most frequently used), *polynomial regression* (often used to test for non-linearity in a relationship), *piecewise regression* (two or more adjacent straight lines), *robust regression* (models that are less sensitive to outliers), *multiple regression* (where there are numerous explanatory variables), *non-linear regression* (to fit a specified non-linear model to data), and *non-parametric regression* (used when there is no obvious functional form).

Generally, however, literature (e.g., Crawley 2009, Fisher n.d, Greene 2008) suggest that canonical understanding in the field of *statistical modelling* is that models can be broadly grouped into two strands: *general linear models* (GLMs) and *generalised linear models* (GLZ). The GLMs can further be sub-divided into two categories, namely *general linear univariate models* (GLUMs) which include *simple* and *multiple* regression techniques, analysis of variance (ANOVA), analysis of covariance (ANCOVA), T-test, and F-test; and *general linear multivariate models* (GLMMs), which occurs when one attempts to explain variation in more than one response variable simultaneously. Included in GLMMs are *multivariate analyses of variance* (MANOVA), *multivariate analysis of covariance* (MANCOVA), *discriminant function analysis* (DFA), *canonical correlation analysis* (CCA), and *principal component analysis* (PCA). The *least squares criterion* is used to obtain the estimates of parameters in general linear models (GLMs), and specific assumptions should be met: *independency* of observations; *normality* of the response variable (s); and constancy (*homogeneity*) of the variance. The general linear model can be algebraically presented as:

$$y = b_0 + bx + \varepsilon \qquad (1)$$

Where:

y = a set of outcome variables,

b_0 = a set of intercepts,

b = a set of coefficients, and x is a set of covariants.

Generalised linear models (GLZ) on the other hand are the extension of linear modelling process that allows models to be fitted to data that follow probability distributions other than *normal distribution*. GLZs also relax the requirement of equality or constancy of variance that is required in hypothesis testing in *general linear models* (GLMs). Parameter estimates in generalised linear models are obtained using the principle of *maximum likelihood*; therefore, hypothesis testing is based on comparisons of likelihoods or deviances of nested models. A generalised linear model has three components: (a) *a random component* – this is the dependent variable y_i which, conditional on independent variables, follows one of the distributions in exponential family including *normal, Poisson, binomial, gama or inverse-Gaussian*; (b) *linear predictor*, $\eta_i = X_i\beta$ on which the dependent variables depend; and (c) *link function*, $L(.)$ that transforms the expectation of the dependent variable $\mu_i \equiv E(Y_i)$ to the linear predictor η_i. Common link functions include *identity link*: $L(\mu_i) = \mu_i$; the *log link*: $L(\mu_i) = \log \mu_i$; the logit link: $L(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$; and probit link: $L(\pi_i) = \Phi(\pi_i)$. Included in this group of generalised linear models are: *logistic regression, general linear model* and *Poisson regression.* The general algebraic expression of generalised linear model is:

$$E(Y) = g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j \quad (2)$$

Where:

$g(\mu)$ is a non-linear link function that links the random component $E(Y)$ to the linear predictor $(\beta_0 + \beta_1 X_1 + ... + \beta_j X_j)$

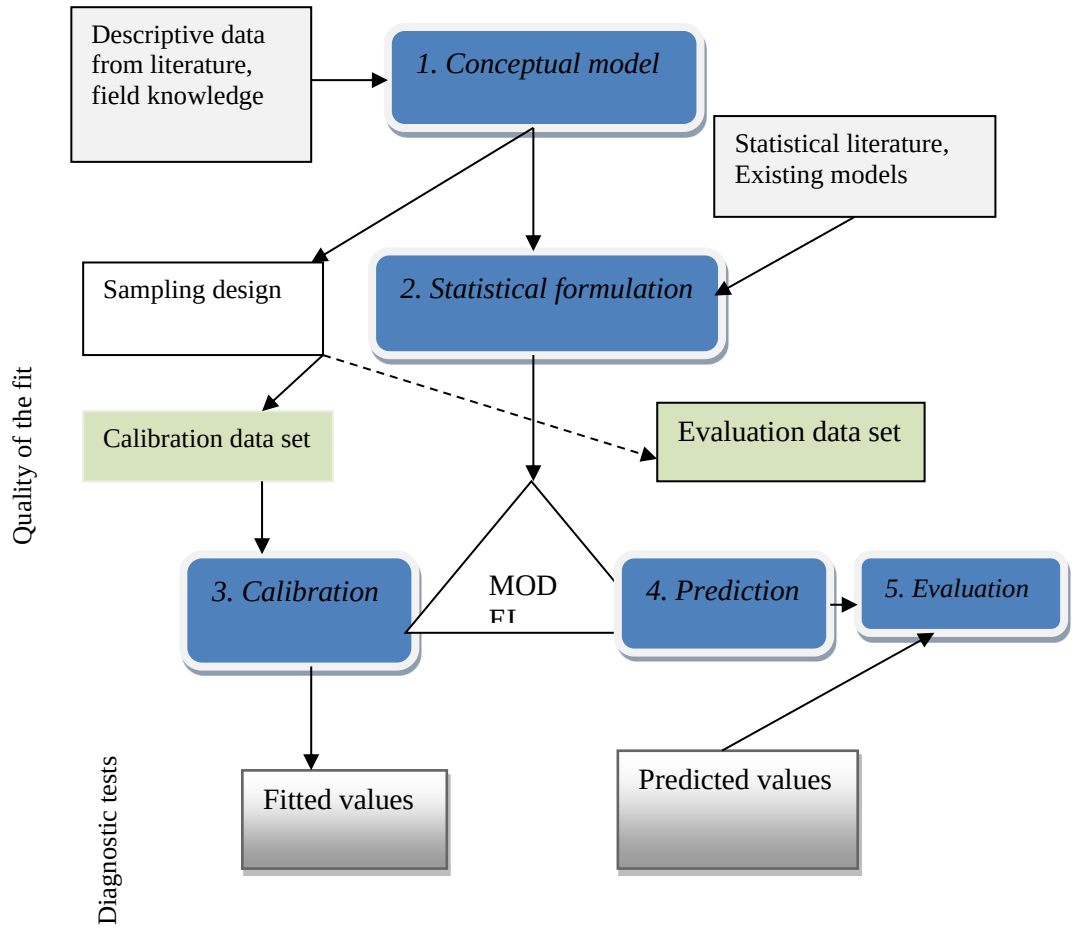

Figure 1: Successive steps of the model building process. Source: Adapted from Guisan and Zimmermann (2000).

Guisan and Zimmermann (2000) posit that there are five successive steps of the model building process (Figure 1): *conceptualisation model, statistical formulation, model calibration, prediction*, and *evaluation.*

Assessing goodness-of-fit of models

Assessment criterion for goodness-of-fit for regression model depends on the purpose of the model (Chatterjee and Price 1977): "purely *descriptive***:** for this particular use, there are conflicting requirements which are to explain as much the variation as possible which means inclusion of large numbers of variables, and to adhere to a principle of parsimony, which suggests that we try, for ease of understanding, to describe the process with as few variables as possible. So, the aim is to choose the smallest number of variables that explain the most substantial part of variation in the dependent variable; *estimation and prediction***:** a regression equation is sometimes constructed for prediction. From the regression equation, the aim is to predict the value of future observations, or to estimate the mean response corresponding to a given observation. When regression is used for this purpose, the variables are selected with an eye towards *minimising* the *mean square error (MSE)* of prediction; *control***:** a regression model may be used as a tool for control. The purpose for constructing the equation may be to determine the magnitude by which the value of an independent variable must be altered to obtain a specified value of dependent variable (target response). For this purpose, it is desired that the coefficients of variables in the model be measured accurately, that is, the standard errors of regression coefficients are small".

According to the authors (ibid), occasionally these functions overlap and an equation is constructed for some or all of these purposes. The main point to be noted is that the purpose for which the regression model is constructed determines the criterion that is to be optimised in its formulation. It follows that a subset of variables that may be best for one purpose may not be best for another. The concept for the "best" subset to be included in an equation always requires an additional qualification.

Validation of predictive models

Validation of the constructed predictive and estimation model can be carried out on the basis of the same data as the model was set up with to determine the model performance *(internal validation)* or can be carried out using new data obtained from storage to assess the quality of the model predictions, the process called *external validation* (Giffel and Zwietering 1999). The performance of a predictive model can be measured by statistical indices. *Bias* and *accuracy* factors are the common statistical indices used to assess the performance of predictive model (Giffel and Zwietering 1999, Koutsoumanis 2001, Skandamis and Nychas 2000, Ross 1996). According to the authors, *bias* is a multiplicative factor that compares the *model predictions* and is used to determine whether the model over- or under-predicts the response. Perfect agreement between predictions and observation produces a bias factor equal to 1. A bias factor (B) > 1 is called *fail-dangerous* while B < 1 is called *fail-safe* (Ross 1996). Dalgaard (2003) posits that a suitable predictive model should have a bias factor between 0.75 and 1.25. *The accuracy factor* is defined as the sum of absolute differences between the predictions and observations, and it *measures the overall model error.* According to Ross (1996) the *accuracy* factor provides an indication of the spread of results about the predicted values. Mellefont *et al.* (2003) argue that an accuracy factor of 1 represents perfect agreement between the observed and predicted values. The larger than one the value is, the less accurate, the average estimate is between observed and predicted values. According to Baranyi *et al.* (1999) and Liu and Puri (2008) the bias and accuracy factors are determined as follows:

Bias factor $(B_f) =$

$$exp\left(\frac{\sum_{k=1}^{m}(ln\,Y_{predicted} - ln\,Y_{observed})}{m}\right) \quad (3)$$

Accuracy factor

$$(A_f) = exp\left(\sqrt{\frac{\sum_{k=1}^{m}(ln\,Y_{predicted} - Y_{observed})^2}{m}}\right) \quad (4)$$

Where *m* is the validation sample size.

According to the authors the *percent discrepancy* factor between the predictive model and the observation, and *percent bias* respectively, are computed as follows:

Percentage discrepancy factor

$$(\%D_f) = (A_f - 1) \times 100\% \quad (5)$$

Percentage bias factor

$$(\%B_f) = sgn(ln\,B_f) \times (exp|ln\,B_f| - 1) \times 100\% \quad (6)$$

Where the $sgn(ln\,B_f)$ is the function interpreted as:

$$sgn(ln\,B_f) = \begin{Bmatrix} +\ if\ B_f > 0 \\ 0\ if\ B_f = 0 \\ -1\ if\ B_f < 0 \end{Bmatrix}$$

The role of the *sign* (lnB_f) is to indicate whether the overall bias is negative or positive. If $\%B_f > 0$ then the model *over-predicts*; if the $\%B_f < 0$ then the model *under-predicts*. Other authors (e.g., Mellefont *et al.* 2003, Dalgaard 2003, Ross 1996, Giffel and Zwietering 1999, Koutsoumanis 2001) use different equation forms for calculating *bias*- and *accuracy*-factors which according to Baranyi *et al.* (1999) results in the same answer as using the *above-presented* equations (equation 4 and 5):

Bias factor

$$(B_f) = 10^{\left(\frac{\sum log(Y_{predicted}/Y_{observed})}{m}\right)} \quad (7)$$

Accuracy factor

$$(A_f) = 10^{\left(\frac{\sum|log\ (Y_{predicted}/Y_{observed})|}{m}\right)} \quad (8)$$

Nonetheless, Zeuthen (2003) argues that there is *currently no set of criteria* which can enable a model to be described as valid. According to Steyerberg *et al.* (2001) the performance of a predictive model tends to be *overestimated* when simply determined on sample of subjects that was used to construct the model (*internal validation*). Okafor (2007) argues that any model that has explained more than 75% of variation in the curve (i.e., $R^2 > 75\%$) can be used for prediction purposes. Hämäläinen (2006) argues that model *complexity* has an effect on both *accuracy* and *robustness* of the model: too complex models do not *generalise* to other data sets, whereas too simple model cannot model essential features in the data – such a model is said to have *low representational power*. According to Hämäläinen (2006), a model is said to be *robust* if it is *insensitive* to *small changes in the data*. The author points out that the aim of *model validation* is to give insurance that the model is *a good one* or at least that a *poor model* is not accepted. The author differentiates techniques used for *descriptive model validation* and *predictive model* validation. In descriptive model validation, techniques used are *statistical significance tests* with the aim to verify that the discovered patterns are meaningful and not only due to chance. According to the author, typical levels of significance *p* are: 0.05 (*nearly significant*), 0.01 (*significant*), and 0.001 (*very significant*). Hämäläinen (2006) argues further that in predictive models, validation (*prediction accuracy tests*) aims to ensure that the model has not *over-fitted* the data and generalises well. The most popular techniques for measuring prediction error are the *sum of squared error* (SSE):

$$SSE = \sum_{i=1}^{n}[(y_i - f(x_i)]^2 \quad (9)$$

Where y_i is a real value, $f(x_i)$ is the predicted value of data point x_i ($i = 1, \ldots, n$). And *mean squared error* (MSE):

$$MSE = \frac{SSE}{n} \quad (10)$$

Where *n* is the sample size and SSE is *sum of squared error*. Nonetheless, Hämäläinen (2006) argues that the final test of the model is *how well it works in practice*. The

relationship between *statistical significance*, *statistical power*, and *model predictive performance* is unclear. There is no guarantee that a model with *statistically significant* terms will give *good predictive performance* (Wintle *et al*. 2005). According to Carrasco *et al* (2006) performance of the predictive model is assessed by its *coefficient of determination* (R^2), *root of mean square error* (RMSE) and *standard error of prediction percentage* (SEP). According to the author, SEP is computed using *equation 12*. Evans (2008) underlines that *MSE* is the useful statistic for assessing the *predictive accuracy* of a model: a good model will predict with an average error close to zero, and with only small over/under prediction around this average (i.e., MSE).

$$SEP = \left(\frac{100}{Y_{observed}}\right)\sqrt{\frac{\Sigma(Y_{observed}-Y_{predicted})^2}{n}} \quad (11)$$

While making explicit distinctions between the statistics concerning the *calibrated* and *predictive* models, Konovalov *et al.* (2008) argue that *the gold standard* of model validation is the *blind fold prediction* when the model's predictive power is assessed from how well the model predicts the activity (response) value which was not considered in any way during model development (calibration). Table 1 distinguishes various *statistics* which can be used to assess the predictive accuracy of the model, where SSE, SSEP, SST, and SSTP are defined as follows:

$$SSE = \sum_{i=1}^{n_c} {e_i}^2 \quad (12)$$

$$SSEP = \sum_{i=n_c+1}^{n} {e_i}^2 \quad (13)$$

$$SST = \sum_{i=1}^{n_c}\left(y_i - \left\{\sum_{i=1}^{n_c}\frac{y_i}{n}\right\}\right)^2 \quad (14)$$

$$SSTP = \sum_{i=1}^{n_c}\left(y_i - \left\{\sum_{i=n_c+1}^{n}\frac{y_i}{n}\right\}\right)^2 \quad (15)$$

Konovalov *et al.* (2008) argue further that *by definition*, the fitting (*calibration*) ability statistics presented in Table 1 have *nothing* to do with the measurements of the *model's predictive power*.

Table 1: Statistics for calibrated and predictive models

Statistic	Explanation	Computational formula
MSE	The mean square of calibration error	$MSEP = \sum_{1 \le i \le nc}^{nc} {e_i}^2/nc$
MAE	The mean absolute error of calibration	$MSEP = \sum_{1 \le i \le nc}^{nc} \lvert e_i \rvert/nc$
MedAE	The median absolute error of calibration	$\underbrace{MED}_{1 \le i \le n_c} \lvert e_i \rvert$
R^2_c	The coefficient of determination for calibration	${R_c}^2 = 1 - \left(\frac{SSE}{SST}\right)$
MSEP	The mean square of prediction error	$MSEP = \sum_{nc \le i < n}^{n} {e_i}^2/n_v$
MAEP	The mean absolute error of prediction	$MAEP = \sum_{n_v \le i < n}^{n_v} \lvert e_i \rvert/n_v$
MedAEP	The median absolute error of prediction	$\underbrace{MED}_{n_c \le i \le n} \lvert e_i \rvert$
R^2_v	The coefficient of determination for prediction	${R_v}^2 = 1 - \left(\frac{SSEP}{SSTP}\right)$

Source: Adapted from Konovalov *et al* (2008).

When the fitted model has "*good fitting ability*" while possessing no *predictive power* (or generalisation) at all, the condition is called model *over-fitting*. As a rule of thumb, according to Konovalov *et al.* (2008), the *more flexible* a model is, the *less correlated* the *fitting ability* and *predictive power* statistics become. Konovalov *et al. (*2008) argue furthermore that in the case of simple linear regression (SLR) and multiple linear regressions (MLR), it is assumed that *over-fitting is much more avoided.* As such, it is generally assumed that *there is a correlation* between the corresponding *fitting* and *predictive* statistics: *MSE* and *MSEP*; *MAE* and *MAEP*; *MedAE* and *MedAEP*; and R^2_c and R^2_v. While this assumption is known to be false for more flexible (e.g., nonlinear) models, the assumption is rarely questioned for SLR and MLR models (Konovalov *et al.* 2008).

Konovalov *et al.* (2008) asserted that cross-validation has two variations: "most existing cross-validation (CV) technique could be reduced to some form of the *leave-group-out cross-validation* (LGO-CV) where a sample of *n* observations is portioned (i.e., splited) into *calibration* (i.e., training) and *validation* (i.e test) subsets. As implied by their names, the calibration subset (with n_c data points) is used to train a model, while the validation subset (with $n_v = n - n_c$ data points) is used to test how well the model predicts the new data, that is, the data points not used in the calibration procedure. In an attempt to improve upon the hold-out cross-validation, the *leave-one-out* (LOO) cross-validation was developed. The LOO cross-validation (LOO-CV) consists of running the LGO-CV *n* times using each of the observations as a validation subset of size n_v =1." Nevertheless, the authors (ibid) cautioned that *leave-one-out cross validation* **must not** be used for assessing the predictive power of models or for model selection. A challenge with a *hold-out cross validation* is to determine a validation sample size. Various recommendations have been put forward: World Bank and Erickson (1995) recommend a validation sample size (n_v) = 7; Benigni and Bossa (2008) recommend that a validation sample (n_v) should be 10% of the total sample size (n); Konovalov *et al.* (2008) recommend that (n_v) = n_c = 50% of sample total size. Shao (1996) recommends that $n_c = n^{¾}$ (implying that, sample for validation should be $n - n^{¾}$).

Harrell (2008) asserts that the model validation can be (a) *external* (best using newly corrected data from another location at another time), or (b) *internal*: which is subdivided into *apparent validation* – evaluating fit on the same data that was used to create the model; *data splitting*; *cross-validation*; and *bootstrap*. Harrell posits further that the two main types of aspects to validate are *calibration* or *reliability* which refers to the ability of the model to make unbiased estimates of response, and *discrimination* which refers to the ability of the model to separate responses. In ordinary least squares (OLS) models, model *discrimination* is measured by R^2 value, while in binary logistic regression model it is measured by the area under ROC curve (ibid). Hurme *et al.* (2005) point out that *model evaluation* with newly collected data is recommended as the most preferred method of model validation.

Rationale for constructing household wood fuel predictive model

Literature review, field experience and empirical evidence from the present study all unequivocally indicate that wood fuel has perilous environmental consequences, and is the most dependable household fuel in Tanzania, and will remain so for the foreseeable future.

The quantification of households' fuel requirements is thus a cardinal aspect in *sustainable forest management*. particularly for natural forests. Quantification of households' fuel requirements could equally be useful when planning for woodfuel-related afforestation/*reforestation* programmes. When undertaking such programmes, it is reasonably imperative to have in mind the *amount* of wood fuel that

will be needed by a particular community, when the planted trees reach their rotation age (for fuel purposes). This calls for the availability of *wood fuel consumption predictive models*.

Sometimes the forest management goal may be towards evaluating effectiveness of a particular programme(s) intended to reduce the amount of household wood fuel consumption. In this case, *wood fuel consumption predictive models* will be useful in computing the *expected* wood fuel consumption (given the set of household socio-economic and demographic characteristics). This consumption is then compared with the *actual* household wood fuel consumption in the presence of wood fuel saving programmes. The difference between two consumptions will illuminate the significance of respective programme(s). In these two cases (i.e., *wood fuel plantation* programmes and *wood fuel saving* programmes) the use of a wood fuel predictive model is, *reasonably speaking*, inevitable. At times, the aim might be painting a picture of wood fuel consumption trends. While collecting information to address this particular need may be expensive and time consuming (Edward *et al.* 2003), the use of a wood fuel consumption predictive model is often relatively cheaper and less time consuming: *one needs only to gather information on variables contained in the predictive model*. In summary, the predictive model developed in this paper will be useful for an array of functions: *planning afforestation/reforestation* programmes, *evaluation* of wood fuel saving programmes, and determination of wood fuel *consumption trends*.

METHODOLOGY

The study was conducted *Morogoro* and *Songea* districts, Tanzania. The design of the study was a *descriptive* and *analytic* cross-sectional survey. The sample design for the present study strove to have a study sample which is *sufficient* and *representative* of the target population. The *target populations* for this study were communities in *Morogoro* and *Songea* districts. *The sampling frame* was in three types depending on the sampling phase. During sampling of *villages* in rural areas and *wards* in peri-urban and urban areas, the sampling frame was the *list of villages* bordering the selected forests and *list of wards* in the municipalities respectively. During sampling of *hamlets* in rural areas and *streets* in peri-urban and urban areas, the sampling frame was the *list of all hamlets* in the selected villages and *list of all streets* in the selected wards respectively. When sampling households for the study, the sampling frames that were used are the *updated lists of households registers* in the sampled hamlets and streets. All chairpersons and executive officers in the selected study sites were asked to update lists of households in their respective areas by excluding households which no longer existed and/or adding those ones which were missing in their lists. *Stratified random sampling* design was used in the present study. Stratification was carried out at two levels: (a) stratification of study sites in the study districts into *rural*, *peri-urban* and *urban* areas, and (b) stratification of respondents into wealth categories: *low*, *medium* and *high*. Stratification of respondents into respective wealth categories was done in a participatory manner during focus group discussions (FGD), which were conducted in respective strata (rural, peri-urban and urban). A following question was posed by a researcher: "I want all *households* in this area *stratified* into three main wealth categories *(life standard)*: *low*, *medium* and *high* wealth categories. What are the household attributes you would use to allocate the households in their respective categories?". The *criteria* for households' stratification into three wealth categories were then *harmonised* and *standardised* for each stratum (Table 2).

Table 2: Standardised criteria for household's categories in different study strata*

Category	Stratum		
	Rural	*Peri-urban*	*Urban*
Low	▪ Poor housing ▪ Food insecurity ▪ Less than 2 meals a day ▪ Works as causal labourer ▪ Physically disabled ▪ No bicycle/No radio	▪ Works as causal labourer ▪ Poor housing ▪ Physically disabled ▪ Not sure of his meals	▪ Unemployed ▪ Unreliable income sources ▪ Living in poor dwelling
Medium	▪ Physically able and smart ▪ Modestly decent dwelling ▪ Modest land holdings ▪ Few animals (esp. goats/chickens) ▪ Sure of 3 meals a day	▪ Petty business ▪ Own fairly decent houses ▪ Sure of 3 meals a day	▪ Petty business ▪ Live in modern house ▪ Sure of 3 meals a day
High	▪ Government employee ▪ Has a shop ▪ Have animals (cattle) ▪ Grinding machines ▪ Big farms ▪ Modern house	▪ Government employee ▪ Has own-transport ▪ Have modern house	▪ Government employee ▪ Whole sale shop ▪ Retail shop ▪ Own Guest house/hotel ▪ Has transport business

*The household should have *one* or *several* of the criteria to qualify into a given category

Rural areas in the context of the present study refer to communities *bordering* the *forests*. *Urban areas* refer to the community residing *fairly* in the *centre* of *municipality*. *Peri-urban* areas refer to the areas geographically located within the municipality, but lying on its periphery.

A total of 568 respondent households were involved in this study **(Table 3**). The sample size for the study was computed using equestions 8 and 9 as recommended by Bartlett *et al.* (2001):

$$n = \left(\frac{n_0}{1 + \frac{n_0}{N}} \right) \qquad (16)$$

The computation of sample size for *categorical* data, according to Bartlett *et al.* (2001), follows the same way as in *continuous* data, except in the computation of n_0, which is:

$$n_0 = \left(\frac{t^2 \times pq}{d^2} \right) \qquad (17)$$

Where: p is the proportion of respondent that will give you information of interest (the proportion *confirming*), q viz (1-p) is the proportion not giving you information of interest (proportion *defective)*, and p*q is the estimate of variance (*which is maximum when p = 0.50 and q=0.50)*.

The maximum population variance of 0.25 will give the maximum sample size. Consequently, the formula used to determine sample size (n) from a population (N) is:

$$n = \frac{384}{1 + \frac{384}{N}} \qquad (18)$$

Data was collected using direct measurements of households' firewood and charcoal consumption and researcher's direct observation. Data analysis was carried out using *SPSS* and *Excel* statistical computer programmes. Conversion of firewood (kg) and charcoal (kg) in woodfuel (m^3) was effected using appropriate conversion factors (Table 4). *Log-linear regression model* was used to construct a predictive model of household woodfuel consumption.

Table 3: Sampled households in the study sites

District	Stratum	Sampled village/ward	Sampled hamlet/street	Households in sampled hamlet/street	Sampled households	Sub-total for stratum
Morogoro	Rural	Fulwe	Dindili	39	35	167
			Ulundo	68	58	
		Maseyu	Kitulangalo	45	41	
			Ng'ambala	36	33	
	Peri-urban	Kingoluwira	Mahakamani	86	70	115
			Tambuka reli	51	45	
	Urban	Kihonda	Kilombero	104	82	82
Songea	Rural	Mtyangimbole	Kanisani	45	40	91
		Ndilimalitembo	Ndilimalitembo	59	51	
	Peri-urban	Mshangano	Mshangano	74	62	62
	Urban	Songea mjini	CCM	59	51	51
GRAND TOTAL					**568**	**568**

Table 4: Conversion factors for firewood and charcoal into wood volume

Firewood	kg	$1m^3$ of wood = 725kg of firewood	Kaale (2005), Amous (1999)
Charcoal	kg	$1m^3$ of wood = 165kg of charcoal	Amous (1999)

The functional form used in developing the wood fuel predictive model is *log-linear regression* whose mathematical formula is:

$$ln\,Y = C + \sum \beta_i \, ln\, X_i + \gamma_j X_j + \varepsilon \qquad (19)$$

Where: Y is the annual amount of wood fuel consumed by a household; C is the constant term; β_i and γ_i are the coefficients; X_i and X_j are socio-economic variables considered to influence quantity of household fuel consumption; and ε is a random error term.

The candidate variables which were selected for model building, and the correlation matrix for the candidate variables are presented in Table 5 and Table 6 respectively. It is worth mentioning that household *income category* was used instead of *household income* because it was noted during data collection that it is easier to get a response using income category than by asking a respondent to provide information on his/her actual income. Putting it in a slightly different way: *asking in which category a respondent's income falls is less sensitive than asking respondent's actual income level.*

Backward elimination procedure was used for variables reduction. The correlation matrix for candidate variables for regression model (Table 6) is important because the backward elimination procedure for variables reduction/selection (Table 7) requires the absence of *multicollinearity* among the candidate variables (Chatterjee and Price 1977). Garson (2007) asserts that as a rule of thumb, inter-correlation among independent variables > 0.80 signals a multicollinearity problem. The results (Table 6) suggest that there is not sufficient evidence for a multicollinearity problem since none of the correlation is above 0.80

Table 5: Description of variables used in the Log-Linear Regression Model

Variable	Description
Y	ln [Household wood fuel consumption (m^3/household/year)]
X_1	ln [Total household size]
X_2	ln [Location of the household (1= rural; 2 = peri-urban; 3= urban)]
X_3	ln [Age (mid-point of age class) of the household head]
X_4	ln [Dwelling size (number of rooms in the main house)]
X_5	ln [Household monthly income category. 1: ≤Tshs 30,000; 2: Tshs. 31,000 – 60,000; 3: ≥Tshs. 61,0000)]
X_6	ln [Price of charcoal (Tshs/kg)]
X_7	ln [Price of kerosene (Tshs/litre)]
X_8	Dwelling category (1 = modern; 0 = traditional)
X_9	Education level of household head (1= educated; 0 = illiterate)
X_{10}	Gender of the household head (1 = female; 0 = male)

Table 6: Correlation matrix for candidate variables for regression model

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1		–0.014	0.018	0.394	0.046	0.046	–0.003	0.208	0.112	–0.110
X_2			0.106	–0.068	–0.148	0.526	–0.184	0.289	0.012	0.063
X_3				0.130	–0.122	0.103	0.046	–0.094	–0.375	0.045
X_4					0.159	–0.063	0.006	0.078	0.140	0.071
X_5						–0.242	0.015	0.107	0.186	–0.150
X_6							–0.030	0.160	–0.051	0.001
X_7								–0.072	–0.042	–0.152
X_8									0.246	–0.014
X_9										0.012
X_{10}										

Table 7: Variables selected by backward elimination method

Independent Variables	P	RMS	R^2	R^2_{adj}	D-W	N
$X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}$	11	0.029	0.766	0.747	1.980	138
$X_1\ X_2\ X_4\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}$	10	0.028	0.766	0.749	1.980	138
$X_1\ X_2\ X_5\ X_6\ X_7\ X_8\ X_9\ X_{10}$	9	0.028	0.766	0.751	1.980	138
$X_1\ X_2\ X_5\ X_6\ X_7\ X_9\ X_{10}$	8	0.028	0.765	0.752	1.956	138
$X_1\ X_2\ X_5\ X_6\ X_7\ X_9$	7	0.028	0.763	0.753	1.912	138
$X_2\ X_5\ X_6\ X_7\ X_9$	6	0.028	0.760	0.751	1.863	138
$X_2\ X_5\ X_7\ X_9$	5	0.039	0.753	0.750	1.939	421
$X_2\ X_5\ X_7$	4	0.038	0.753	0.751	1.942	422
$X_2\ X_7$	3	0.038	0.752	0.751	1.938	422
X_7	2	0.040	0.745	0.744	1.881	422

Key:

RMS = Residual Mean Square
P = Term equation (number of parameters in the equation)
R^2 = Coefficient of determination
R^2_{adj} = Adjusted coefficient of determination
D-W* = Durbin-Watson Statistic
N = Valid cases (sample size) used in the model

*The Durbin-Watson Statistic (coefficient) tests whether the observations are independent, an assumption which is made by many statistical procedures including multiple regression. According to Garson (2007), the D-W statistic should be between 1.5 and 2.5 for independent observations. Therefore, the *recommended* model is the one with minimum RMS taking into account a principle of parsimony.

RESULTS

Woodfuel consumption in the study area

Wood fuel consumption in the study area is presented in Table 8.

Parameter estimates

Two predictive models were constructed: the first one using the GLM and the second one using the *regression analysis model*. The variables which were used in the GLM model and model parameter estimates are, respectively, presented in Table 9 and Table 10.

Table 8: Wood fuel consumption in the study area (mean ± s.e)

Stratum	Morogoro District			Songea District			Pooled sample		
	% in use	(m^3/hh/yr)	(m^3/capita/yr)	% in use	(m^3/hh/yr)	(m^3/capita/yr)	% in use	(m^3/hh/yr)	(m^3/capita/yr)
Rural	93.2	4.1 ± 0.2	0.82 ± 0.04	98.9	6.3 ± 0.4	1.26 ± 0.08	88.8	5.0 ± 0.2	1.00 ± 0.04
P/urban	87.8	5.0 ± 0.4	1.25 ± 0.10	95.2	6.9 ± 0.5	1.38 ±0.10	90.4	5.7 ± 0.3	1.14 ± 0.06
Urban	96.3	5.9 ± 0.4	0.98 ± 0.07	100	9.9 ± 0.7	1.65 ± 0.12	97.7	7.5 ± 0.4	1.25 ± 0.07
Overall	95.2	4.8 ± 0.2	0.96 ± 0.04	98	7.4 ± 0.3	1.48 ± 0.06	91.4	5.8 ± 0.2	1.16 ± 0.04

Table 9: Description of variables used in the *Generalied Linear Model*

Variable	Description
Y	ln [Household wood fuel consumption (m^3/household/year)]
X_2	Location of the household (1= rural; 2 = peri-urban; 3= urban)]
X_5	Household monthly income category. 1: ≤Tshs 30,000; 2: Tshs. 31,000 – 60,000; 3: ≥Tshs. 61,0000)
X_9	Education level of household head (1= educated; 0 = illiterate)
X_6	ln [Price of charcoal (Tshs/kg)]
X_7	ln [Price of kerosene (Tshs/litre)]

Table 10: *Generalised linear regression model* parameter estimates for wood fuel consumption prediction model

Parameter	B	Std. error	t	Sig.	95% Confidence Interval	
					Lower Limit	Upper Limit
Intercept	8.931	0.502	17.800	0.0001***	7.940	9.923
[Income=1.00]	0.024	0.035	0.702	0.4840Ns	–0.044	0.093
[Income =2.00]	0.015	0.032	0.476	0.6350Ns	–0.048	0.078
[Income =3.00]	0	.	.	.	.	.
[Location=1]	–0.086	0.040	–2.134	0.0350*	–0.165	–0.006
[Location=2]	0.016	0.033	0.502	0.6160	–0.048	0.081
[Location=3]	0	.	.	.	.	.
[Education=.00]	–0.133	0.050	–2.658	0.0090**	–0.232	–0.034
[Education=1.00]	0	.	.	.	.	.
Ln Kerosene price	–1.398	0.066	–21.122	0.0001***	–1.529	–1.267
Ln Charcoal price	–0.014	0.042	–0.338	0.7360Ns	–0.096	0.068

R^2 = 0.780

R^2_{adj} = 0.769

NS: Not statistically significant at α = 0.05;

*: Statistically significant at α = 0.05;

**: Statistically significant at α = 0.01;

***: Statistically significant at α = 0.001

Consequently, the proposed structural form of the predictive model (using GLM) is:

$$\ln Y = C + loc_i + inc_j + edu_k + \beta_1 \ln\ P_{charcoal} + \beta_2 \ln\ P_{ker\,osene} \quad (20)$$

Where:

$P_{charcoal}$ = Price of charcoal (Tshs/kg)

$P_{ker\,osenel}$ = Price of kerosene (Tshs/litre)

i: 1 = *rural area*; 2 = *peri-urban;* 3 = *is urban area*

j: 1 = *low income*; 2 = *medium income;* 3 = *high income* (as defined in Table 5)

k: 0 = *illiterate*; 1 = *literate* (with formal education)

Using the information available in Table 6., the resulting predictive model (GLM) is:

$$\ln Y \rightleftarrows = 8.931 + \begin{pmatrix} -0.086:\ if\ i=1 \\ 0.016:\ if\ i=2 \\ 0.00:\ if\ i=3 \end{pmatrix} + \begin{pmatrix} 0.024:\ if\ j=1 \\ 0.015:\ if\ j=2 \\ 0.00:\ if\ j=3 \end{pmatrix} + \begin{pmatrix} -0.133:\ if\ k=0 \\ 0.00:\ if\ k=1 \end{pmatrix} - 0.014\ \ln P \rightleftarrows\rightleftarrows\rightleftarrows_c - 1.398\ \ln P_k \quad (21)$$

Using a standard multiple regression analysis, the following functional form was proposed:

$$\ln Y = C + \beta_2 \ln X_2 + \beta_5 \ln X_5 + \beta_6 \ln X_6 + \beta_7 \ln X_7 + \beta_9 X_9 \quad (22)$$

Where:

Y = *Wood fuel (m^3/household/yr)*

X_2 = *Location of household (as defined in the present study)*

X_5 = *Income category of household*

X_6 = *Price of Charcoal (Tshs/kg)*

X_7 = *Price of kerosene (Tshs/litre)*

X_9 = *Education level* (*as defined in the present study)*

The model parameter estimates are presented in Table 11.

Table 11: *Classical multiple regression model* parameter estimates for wood fuel prediction model

						Correlation		Collinearity statistic	
	on	β	s.e	t-value	p-value	Zero-order	Partial	VIF	Tolerance
Regression of Annual wood fuel consumption (ln) (m^3/household/year)	Location (ln) (X_2)	0.082	0.040	2.076)	0.0400*	0.256	0.178	1.443	0.693
	Income category (ln) (X_5)	–0.027	0.034	–0.781	0.4360Ns	–0.039	–0.068	1.099	0.910
	Price of charcoal (ln) (X_6)	–0.018	0.044	–0.401	0.6890Ns	0.064	–0.017	1.448	0.691
	Price of Kerosene (ln) (X_7)	–1.367	0.071	–19.192	0.0001***	–0.860	–0.858	1.043	0.959
	Education level (dummy) (X_9)	0.120	0.052	2.324	0.0220*	0.132	0.198	1.040	0.962
	Constant	4.414	0.550	15.575	0.0001***				

Key:

NS: Not statistically significant at α = 0.05

*: Statistically significant at α = 0.05

* **: Statistically significant at α = 0.001

VIF: Variance-inflation factor. It tests multicollinearity problem (Garson, 2007; Greene, 2003). According to Garson (2007), as a rule of thumb, multicollinearity is a problem when: VIF > 4; tolerance < 0.20

∴ The model is:

$$\ln Y = 4.414 + 0.082 \ln X_2 - 0.027 \ln X_5 - 0.018 \ln X_6 - 1.367 \ln X_7 + 0.120 \ln X_9 \quad (23)$$

Model validation

Both GLM and regression models were subjected to model validation. As mentioned earlier, a *cross-validation* approach was adopted in the present study. Of 56 samples selected for the validation process, only 14 cases were valid and thus used for validation (Appendix 1). The first and foremost validation technique, which was simultaneously used to select between the two constructed models, was determination of residual values: *residual* =

(measured/actual wood fuel) – (model-predicted wood fuel). The residual values for the same data set for GLM and *Classical multiple regression model* are, respectively, presented in Table 12 and Table 13.

It was found, looking at *residual values* (as evidenced by Table 12 and Table 13) that GLM is not a suitable predictive model for wood fuel consumption, at least for this particular study. Consequently, further validation was carried out for *classical multiple regression analysis*-derived prediction model, as shown in Table 14.

Table 12: *Generalised linear regression model* estimated wood fuel versus field-measured (actual) wood fuel

S/N	loc_i	inc_j	edu_k	–0.014 ln P_C	–1.398 ln P_K	Constant	Ln Y	Estimated WF	Actual WF	Residual
1	–0.086	0	0	–0.0742	–10.2194	8.931	–1.4486	0.235	6.64	6.405
2	–0.086	0	0	–0.0742	–10.2194	8.931	–1.4486	0.235	5.55	5.315
3	0.016	0	0	–0.07728	–9.786	8.931	–0.91628	0.400	7.83	7.430
4	0.016	0	0	–0.07728	–10.2194	8.931	–1.34968	0.259	6.64	6.381
5	0.016	0	0	–0.07728	–10.2194	8.931	–1.34968	0.259	4.42	4.161
6	0	0	0	–0.0798	–9.66018	8.931	–0.80898	0.445	8.85	8.405
7	0	0.015	0	–0.0798	–10.2194	8.931	–1.3532	0.258	4.42	4.162
8	0	0.024	0	–0.08386	–10.2194	8.931	–1.34826	0.259	6.64	6.380
9	0	0.015	0	–0.0798	–10.2194	8.931	–1.3532	0.258	6.64	6.382
10	0	0	0	–0.0742	–9.66018	8.931	–0.80338	0.448	6.64	6.192
11	0	0.015	0	–0.0798	–10.2194	8.931	–1.3532	0.258	6.64	6.382
12	0	0.024	0	–0.0798	–9.66018	8.931	–0.78498	0.456	7.76	7.304
13	0	0	0	–0.07938	–10.1774	8.931	–1.32578	0.266	9.5	9.234
14	0	0.015	0	–0.07672	–9.91182	8.931	–1.04254	0.353	8.85	8.497

Table 13: *Classical regression model*- estimated wood fuel versus field-measured (actual) wood fuel

S/N	0.02 ln X_2	–0.027 ln X_5	–0.018 ln X_6	–0.367 ln X_7	0.12 X_9	Constant	Ln Y	Estimated WF	Actual WF	Residual
1	0	–0.02966	–0.0954	–2.68277	0.12	4.414	1.7	5.62	6.64	1.02
2	0	–0.02966	–0.0954	–2.68277	0.12	4.414	1.7	5.62	5.55	–0.07
3	0.013862944	–0.02966	–0.09936	–2.569	0.12	4.414	1.8	6.36	7.83	1.47
4	0.013862944	–0.02966	–0.09936	–2.68277	0.12	4.414	1.7	5.67	6.64	0.97
5	0.013862944	–0.02966	–0.09936	–2.68277	0.12	4.414	1.7	5.67	4.42	–1.25
6	0.02197224	–0.02966	–0.1026	–2.53597	0.12	4.414	1.9	6.60	8.85	2.25
7	0.02197224	–0.01871	–0.1026	–2.68277	0.12	4.414	1.8	5.77	4.42	–1.35
8	0.02197224	0	–0.10782	–2.68277	0.12	4.414	1.8	5.84	6.64	0.80
9	0.02197224	–0.01871	–0.1026	–2.68277	0.12	4.414	1.8	5.77	6.64	0.87
10	0.02197224	–0.02966	–0.0954	–2.53597	0.12	4.414	1.9	6.65	6.64	–0.01
11	0.02197224	–0.01871	–0.1026	–2.68277	0.12	4.414	1.8	5.77	6.64	0.87
12	0.02197224	0	–0.1026	–2.53597	0.12	4.414	1.9	6.80	7.76	0.96
13	0.02197224	–0.02966	–0.10206	–2.67176	0.12	4.414	1.8	5.77	9.5	3.73
14	0.02197224	–0.01871	–0.09864	–2.60203	0.12	4.414	1.8	6.28	8.85	2.57

Table 14: Determination of bias and accuracy factors of prediction model

$WF_{pred.}$	$WF_{obs.}$	lnWF(p)	lnWF(o)	[lnWF(p)– lnWF(o)]	[lnWF(p)– lnWF(o)]^2
0.62	6.64	1.726332	1.893112	–0.1667803	0.027815668
5.62	5.55	1.726332	1.713798	0.012533736	0.000157095
6.36	7.83	1.850028	2.057963	–0.207934133	0.043236604
5.67	6.64	1.735189	1.893112	–0.157922846	0.024939625
5.67	4.42	1.735189	1.48614	0.249049422	0.062025614
6.60	8.85	1.88707	2.180417	–0.29334781	0.086052938
5.77	4.42	1.752672	1.48614	0.266532384	0.071039512
5.84	6.64	1.764731	1.893112	–0.128381167	0.016481724
5.77	6.64	1.752672	1.893112	–0.140439883	0.019723361
6.65	6.64	1.894617	1.893112	0.001504891	2.2647E–06
5.77	6.64	1.752672	1.893112	–0.140439883	0.019723361
6.80	7.76	1.916923	2.048982	–0.132059722	0.01743977
5.77	9.50	1.752672	2.251292	–0.498619718	0.248621623
6.28	8.85	1.83737	2.180417	–0.343047479	0.117681573
			Σ	–1.67935	0.754941
			$\frac{\Sigma}{m=14}$	–0.11995	0.053924
B_f				0.88696	
A_f				1.26	
$\%D_f$				(1.26-1) x 100% =26%	
$\%B_f$				(+1) x [(0.11995 -1)] x 100% = 88% < 0	

Key:
$WF_{pred.}$ = Predicted wood fuel consumption (m^3/household/year)
$WF_{obs.}$ = Observed (measured) wood fuel consumption (m^3/household/year)
lnWF (p) = Natural logarithm of predicted wood fuel consumption
lnWF (o) = Natural logarithm of observed/ measured wood fuel consumption

Using equation 6.4 the *bias factor* = exp (–0.11995) = 0.88696 was computed, and its corresponding *percent bias* (% B) was computed using equation 6.7 and found to be $(+1)\times(0.11995–1)\times100\%$ = –88% <0. The *accuracy factor* for the developed model was computed using equation 6.5 and found to be = exp $\{\sqrt{0.053924}\}$ = 1.26. Using an equation 6.6, *percent discrepancy* (%D) = $(1.26\text{-}1)\times100\%$ = 26%. The model validation findings point out that: the developed prediction model is not perfectly accurate (*because accuracy factor ≠ 1*), and under-predicts the household wood fuel consumption (*because the percentage bias factor < 0*). Nonetheless, the constructed model seems to be plausible because its bias factor is such that:

$0.75 < B_f\,(= 0.88696) < 1.25$, therefore within the range of plausible predictive models.

DISCUSSION

Both the *discriminative properties* (model fit) of the predictive model for households' wood fuel consumption presented in this study (R^2 = 0.76) as well as its *calibration* (predictive power) appears to be fairly good. The constructed model (as might possibly be expected) is not perfectly accurate (accuracy factor is approximately 1.26). The findings also revealed that the constructed predictive model is biased: *the bias factor and corresponding percent bias are, respectively,* 0.88696 and –88%. This implies that the

predicted wood fuel consumptions are *undervalued.* The actual wood fuel consumption is supposed to be *accuracy factor* (1.26) multiplied by *the predicted value*. The validation results suggest that in order to obtain a more plausible predictive model, a *correction factor* is imperative. Accordingly, the *corrected* household wood fuel predictive model is:

$$W = 1.26 \times e^{(4.414 + 0.082\,ln\,X_2 - 0.027\,ln\,X_5 - 0.018\,ln\,X_6 - 1.367\,ln\,X_7 + 0.120X_9)} \quad (25)$$

Where:

W	= wood fuel (m^3/household/year)
1.26	= accuracy factor of the constructed predictive model
X_2, X_5, X_6, X_7, X_9	= predictor variables as previously defined

Nevertheless, the above correction in the constructed predictive model notwithstanding, I recommend, as many authors have pointed out (e.g. Hurme *et al.*, 2005; Harrell, 2008), that the corrected model be *externally validated* using the newly corrected data from the study area and adjusted accordingly before it can ultimately be put in use. Furthermore, in order to have a more robust predictive model, data to be used for external validation should be *longitudinally* collected so as capture the *temporal* variations in households' wood fuel consumption. External validation of this model before ultimately using it, is particularly important because the validation sample size used was very small (n=14).

CONCLUSION

The model validation findings point out that: the developed prediction model is not perfectly accurate (*because accuracy factor* $\neq 1$), and under-predicts the household wood fuel consumption (*because the percentage bias factor* < 0). Nonetheless, the constructed model seems to be plausible because its bias factor is such that: $0.75 < B_f\,(= 0.88696) < 1.25$, therefore within the range of plausible predictive models. It is reasonable therefore to argue that in the current Tanzanian situation where there is no any model that can be used to predict and/or estimate wood fuel consumption, the present wood fuel consumption predictive model (equation 25) can be useful in sustainable forest management strategies. However, it is prudent that prior to its use, the constructed model needs to be further validated and adjusted accordingly using newly collected longitudinal data from the study area. Sufficient data should be collected from the *strata* (locations) commensurate with those used in the present study.

ACKNOWEDGEMENTS

I owe so much-and therefore am exceedingly grateful - to my supervisor Prof. Colin Price for his outstanding guidance, advice, assistance, encouragement and constructive criticism throughout the study. I am indebted to my employer, the Sokoine University of Agriculture for granting me the study leave and Association of Commonwealth University Sponsorship for financing my studies and the British Council for effectively administering the scholarship. My heart-felt gratitude also go to the WWF for their partial financial support during fieldwork, without which data collection would be absolutely difficult undertaking due to my then financial constraints for fieldwork.

Declaration of Conflicting Interest: "The authors declares that there is no conflict of interest regarding the publication of this article."

REFERENCES

Amous, S. 1999. The role of wood in Africa. Wood energy today for tomorrow. Rivero, S.I and Flood, R. (eds.). Forestry Department, FAO, Rome Italy.

Baranyi, J., Pin, C. & Ross, T. 1999. Validating and comparing predictive models. *International Journal of Food Microbiology*, 48:159-166.

Bartlett, J.E., Kotrlik, J.W. & Higgins, C.C. 2001. Oroganizational Research: Determining Appropriate Sample Size in Survey Research. *Information Technology, Learning, and Performance Journal*, 19 (1): 43-50.

Chatterjee, S. & Price, B. 1977 *Regression by example*. Published by John Wiley & Sons, New York.

Climate Technology Centre and Network (CTC). 2018. Sustainable Woodfuel (Charcoal and Firewood) Systems in Tanzaniia. A grassroot Training Mannual.

Crawley, M.J. 2009. *The R Book*. John Wiley & Sons Ltd, England. pp 942.

Dalgaard, P. 2003. Predictive Microbiology. In: Huss, H.H., Ababouch, L. and Gram, L. (eds.,). Assessment and Management of Seafood Safety and Quality. FAO Fisheries Technical Paper 444. FAO, Rome.

Darko-Obiri. B, Owusu-Afriyie, K., Kwarteng E. & Nutakor E., 2015. Fuel Wood Value Chain Report. The USAID/Ghana Sustainable Fisheries Management Project (SFMP). Narragansett, RI: Coastal Resources Center, Graduate School of Oceanography, University of Rhode Island and SNV Netherlands Development Organization. GH2014_SCI011_SNV. 157 pp.

Dorazio, R.M. & Johnson, F.A. 2003. Bayesian inference and decision theory- A framework for decision making in natural resource management. *Ecological Applications*, 13(2): 556-563.

Edward, R.D., Smith K.R., Zhang, J. & Ma, Y. 2003. Model to predict emission of health-damaging pollutants and global warming contributions of residential fuel/stove combinations in china. *Chemosphere*, 50: 201-215.

Evans, M. 2008. Measuring the predictive accuracy of various models of formability of Corus Tubular Blanks. *J Mater Sci*, 43: 2562-2573.

FAO. 2004. *Unified bioenergy terminology –UBET*. Rome.

Fischer, C.S., n.d. Log-linear analysis and generalized linear model. Unpublished.

Garson, G.D. 2008. Scales and Standard Measures. (Online). Available at http://faculty.chass.ncsu.edu/garson/PA765/standard.htm. Visited on 13/11/2008.

George, E.I. 2000. The variable selection problem. University of Texas at Austin. Unpublished.

Giffel, M.C. & Zwietering, M.H. 1999.Validation of predictive models describing the growth of *Listeria monocytogenes*. *International Journal of Microbiology*, 46:135-149.

Gilchrist, W. 1978. A statistical Forecasting. Department of mathematics and statistics, Sheffield Polytechnic. A Wiley-Inter-science Publication.

Greene, W.H. 2008. *Econometric Analysis*. Sixth Edition. Edited by Alexander, D.; Leale, J., Zonneveld, C., & Feimer, M. Published by Pearson Prentice Hall., Upper Saddle River, New Jersey.

Guisan, A. & Zimmermann, N.E. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 125(2-3): 147-186.

Hämäläinen, W. 2006. Descriptive and Predictive Modelling techniques for educational technology. Licentiate thesis, Department of computer science, University of Joensuu, Finland. pp 183.

Harrell, F.E. 2008. Regression Modelling Strategies. Department of Biostatistics short-course, 14-18 January 2008. Online at http://biostat.mc.vanderbilt.edu/twiki/pub/main/rms/ . Accessed on 24/01/2009.

Hurme, E., Mönkkönen, M., Nikula, A., Nivala, V., Reunanen, P., Heikkinen, T. & Ukkola, M. 2005. Building and evaluating predictive occupancy models for Siberian flying Squirrel using forest planning data. *Forest ecology and Management*, 216: 241-256.

Kaale, B.K. 2005. Baseline Study on Biomass Energy Conservation in Tanzania. SADC Programme for Biomass Energy Conservation (ProBEC). Report. pp.55.

Konovalov, D.A., Llewellyn, L.E., Heyden, Y.V. & Coomans, D. 2008. Robust Cross-Validation of linear regression QSAR (Quantitative structure – activity relationship) Models. *Journal of Chemical and Information Modelling*, 48 (10): 2081-2094.

Koutsoumanis, K. 2001. Predictive Modelling of the Shelf life of fish under Nonisothermical Conditions. Applied and Environmental Microbiology. *American Society for Microbiology*: 1821-1829.

Liu, S. & Puri, V.M. 2008. pH spatial distribution model during ripening of Camembert cheese. *LWT-Food Science and Technology*, 41: 1528-1534.

Mellefont, L.A., McMeekin, T.A. & Ross, T. 2003. Performance evaluation of a model describing the effects of temperature, pH and Lactic acid on the growth of *Escherichia coli*. *International Journal of Food Microbiology*, 82:45-58.

Mosley, R.C. 2005. The use of predictive modelling in the insurance industry. Online at http://www.pinnacleaactuaries.com/pages/publication/ . Accessed on 24/01/2009.

Njenga, M., Gasaya, O., Sabrina, C., Pesha, I., Jilala, Z., Pangal, R. Frumence, R., Chikawe, M. & Kimaro, A. 2018. Sustainable woodfuel (charcoal and firewood) systems in Tanzania. A grassroots training manual.

Okafor, B. 2007. Modelling of metal wear in screw presses in palm oil mills. *Journal of Engineering and Applied Science*, 2 (3):481-484.

Oracle Cooperation. 2008. New Horizon in predictive modelling and risk analysis. Online at http://www.oracle.com/technology/products/ . Accessed on 24/01/2009.

Ross, T. 1996. Indices for performance evaluation of predictive models in Food Microbiology: *Journal of Applied Bacteriology*, 81:501-508.

Schuerman, J.R. 1983. *Multivariate Analysis in the Human services*. Edited by William, J. Reid. Published by Springer. pp 292.

Sepp, S., & Mann, S. 2009. Woodfuel Supply Intervention. Lessons Learned and Recommendations. Biomass Energy Strategy (BEST).

Shao, J. 1996. Bootstrap model selection. *J. Am. Stat. Assoc*, *91*: 655–665.

Skandamis, P.N. & Nychas, G.E. 2000. Development and Evaluation of a model predicting the survival of *Escherichia coli*: H7NCTC 12900 in Home-made egg plant salad at various temperatures, pHs, and organo essential oil conditions. Applied and Environmental Microbiology. *American Society for Microbiology.* 66(4): 1646-1653.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y. & Habbema, J.D.F. 2001. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis, *Journal of Clinical Epidemiology*, 54:774-781.

United Nations Environment Programme (UNEP). 2019. Review of Woodfuel Biomass Production and Utilization in Africa. A desk study.

Wintle, B.A., Elith, J. & Potts, J.M. 2005. Fauna Habitat modelling and mapping. A review and case study in the Lower Hunter Central Coast Region of NSW. Austral Ecology, 30:719-738.

Wold Bank & Eriksson, L. 1995. Statistical validation of QSAR results. In van de Waterbeemd, H., (ed.). *Chemometric Methods in Molecular Design*. VCH: Weinheim, Germany: pp 309-318.

Zeuthen, P. 2003. *Food Preservation Techniques*. P. Zeuthen & L. Bogh-Sorensen (eds.). Wood- head publishing Ltd. pp. 581.

Zukerman, I. & Albrecht, D.W. 2001. Predictive Statistical Models for user modelling. *User Modelling and User-adapted Interaction*, 11:5-18.