Efficiency Aggregation in Stochastic Frontier Analysis with Hierarchical Data

Yashree Mehta, Bernhard Bruemmer

Department of Agricultural Economics and Rural Development, University of Goettingen

Abstract

Data regarding agricultural production often have a natural hierarchical structure. Ownership of multiple plots by a farmer is one such case. When there is more than one level of observation at which technical efficiency can be estimated, the process of its aggregation from a micro unit of analysis to a higher, aggregate level, poses a topic for a methodological debate. Having used Stochastic Frontier Analysis on data concerning maize production, with a hierarchical structure, we compare scaling up of technical efficiency scores from a plot-level stochastic frontier model, and the Linear Mixed Effects model. With Monte Carlo simulation, we conclude that if monotonicity in the ranking of farm households is to be preserved, the Linear Mixed Effects Model performs slightly better than aggregation indices applied after plot-level estimation. In maintaining the Cumulative Density of the true aggregated efficiency scores, unlike aggregation indices, the Linear Mixed Effects Model performs accurately.

Keywords: Technical efficiency, Hierarchical modelling, Stochastic Frontier Analysis

1. Introduction

A hierarchical data structure is commonplace in agricultural surveys. Ownership of more than one plot by a farm household is one case, where, the collected data on production has a hierarchical structure. Notwithstanding that the data is cross-sectional, without a time dimension, a farm household can form a cluster, with data on the plots owned by it assuming the role of repeated measurements. Output and input usage are directly measured at the plot-level. Data on socio-economic variables which determine the efficiency of the household as a producer, are collected at the farm-level. These variables do not vary across plots which belong to the same household but vary between households. Plot-level technical efficiency can be directly estimated by classifying the production inputs as the explanatory variables of the frontier, and the household-level inputs as Z-variables, by using Stochastic Frontier Analysis (SFA).

Theoretically, the production frontier is attached to the notion of a producer as a Decision Making Unit (DMU). Quoting from Fried et al. (1993),

"..in practice, one has only data—a set of observations for each decision-making unit (DMU) corresponding to achieved output levels for given input levels". (p. 121)

The inefficiency term in SFA captures the effect of managerial ability of the concerned DMU (here, the producer). Thus, estimation of technical efficiency should ideally produce the efficiency score of the producer. Given the naturally occurring hierarchical structure of survey data regarding production, there arises an estimation and inference anomaly with respect to

technical efficiency. The estimated efficiency at the plot-level by SFA does not fit in the otherwise unanimous understanding of "efficiency". Plot-level efficiency is, thus, difficult to interprete and cannot directly be used to conclude about the performance of a DMU. This calls for a methodologically sound aggregation strategy for navigating from plot-level efficiency estimates to the higher level in the data hierarchy, who, in this case are the producers/farm households. One may also intend to make policy recommendations which affect DMUs so as to facilitate an improvement in their efficiency and this requires information about their efficiency performance.

Cook et al. (1998) recognize the need for an aggregation method in efficiency estimation when DMUs occur naturally in groups. They distinguish between pure hierarchies and levels. The former refers to hierarchies formed based on a particular attribute of the DMUs whereas the latter refers to groupings at one level which may be formed based on multiple attributes. They propose a method for synchronizing DMU ratings based on relative efficiency according to those received by their respective groups in the application of Data Envelopment Analysis (DEA). Blackorby and Russell (1999) extend this approach by deriving conditions under which efficiency index aggregation can be carried out consistently across different levels of DMUs, using DEA.

Brorsen and Kim (2013) examine the consequences of data aggregation on the estimation of a stochastic frontier, when dealing with hierarchical data, using a cost function approach. In the light of the closed skew normal being the true distribution of the aggregated data, they find that misspecification

caused by using the standard stochastic frontier model leads to an incorrect conclusion of diseconomies of scale and and higher inefficiency of smaller units. The hierarchical structure studied by them concerns aggregation across DMUs, whereas our study, which is regarding multiple plots per farm, is concerned with a hierarchy across parts of a DMU.

The need for consistent efficiency estimation at different levels of observation in a hierarchical data structure gives rise to the need for a method of aggregation. Using hierarchical maize production data of smallholders in Kenya, we address the unresolved question of being able to infer about aggregate efficiency from lower-level estimates from a data hierarchy, using SFA, by specifying two models: plot-level stochastic frontier model and the Linear Mixed Effects (LME) model. We verify if a fundamentally correct distribution of efficiency can be arrived at, at the household level from plot-level estimates of technical efficiency. We compare the performance of the plot stochastic frontier model and the LME model in deriving efficiency estimates at an aggregate level, with respect to estimating the true scores, using Monte Carlo simulation. We also examine the role of the plot-level statistical error term in maintaining robustness of the aggregation process.

The rest of the paper is organized as follows: In Section 2, the methodological background is provided for setting the context of hierarchical modelling. In Section 3, the methodology used for aggregation in the two models is explained, respectively, after a description of their specification. Section 4 provides a description of the variables used in the study. In Section 5, we present the characteristics of the data used. Section 6 describes the pro-

cedure followed for Monte Carlo simulation, along with the result obtained from it. In Section 7, we present the empirical application of the two models on the existing data, along with a discussion of its result. Section 8 presents the conclusion.

2. Methodological background

Data arising from repeated measurements of different plots belonging to the same farmer implies a clustered structure. It is a data structure in which a unit of observation is nested within another higher-level unit of observation. Such data can be analyzed using a cross-sectional multilevel modeling approach. In multilevel modeling, estimation and inference at one level of an observed unit often depends on the estimation and inference of parameters (random coefficients) at a higher level. It is this property of conditional modeling due to which it is also known as "hierarchical" (Gelman and Hill, 2007). It retains the identity of being cross-sectional but implies a hierarchical structure.

Clark (2016) provides an explanation of the different modeling approaches in clustered data analysis. There are several terms used for models applied to clustered data: Variance components, Random Intercept, Random effects, Hierarchical model, Multilevel model, Mixed models, and so on. These refer to the same modeling scheme, viewed from a different purpose of analysis and treatment of the random components. To start with, the classical linear model is expressed in terms of the data generating process, before considering the clustered structure of the data. Having introduced the cluster structure,

a model is specified in which the coefficients vary by cluster. These coefficients can be varying intercepts as well as varying slopes. Since our study uses varying intercepts and not varying slopes, we focus on the typology of clustered data analysis with respect to a random intercepts model.

We use the "Multilevel model" classification (Clark, 2016), which is given as follows:

$$y_{ij} = \alpha_i + \beta X_{ij} + \epsilon_{ij} \tag{1}$$

$$\alpha_i = \beta_o + \gamma_{oi} \tag{2}$$

$$\gamma_{oi} \sim N(0, \tau^2)$$

In equation (2), y_{ij} is the dependent variable, which in our case is the output. x_{ijk} is the k^{th} covariate and the βs are the estimated coefficients corresponding to the covariates. γ_{oi} measures the extent to which the cluster i differs in its "base" level of production from the population fixed intercept β_o. Thus, the sum of β_o and γ_{oi} is the cluster-specific random intercept, denoted by α_i. The random deviation γ_{oi} is assumed to follow a normal distribution with zero mean and variance τ^2. ϵ_{ij} is the random error term.

Substituting equation (2) in equation (1), we get equation (3) as follows:

$$y_{ij} = (\beta_o + \gamma_{oi}) + \beta X_{ij} + \epsilon_{ij} \tag{3}$$

Alternatively, the random deviation γ_{oi} can be summed with the random error term, ϵ_{ij} as given in equation (4).

$$y_{ij} = \beta_o + \beta X_{ij} + (\gamma_{oi} + \epsilon_{ij}) \tag{4}$$

Equations (3) and (4) correspond to the different treatment of the random component γ_{oi} in terms of the purpose of modeling. Equation (3) considers the random deviation as a variable of interest which forms the cluster-specific intercept. Equation (4) regards the random deviation as a nuisance parameter. In our study, we will be using cluster-specific intercepts for estimating technical efficiency. Hence, random effect is a substantive parameter, and we use the Multilevel model specification in equations (1) and (2)[1].

Since we are interested in cluster-specific random intercepts, we acknowledge their process of prediction as the Best Linear Unbiased Predictor (BLUP). Since the random deviation is a random variable, we "predict" them instead of "estimating" them. Its prediction is about realizing its conditional mean, based on the data at hand. For an explanation of the method of prediction of the random effects, one can refer to Fitzmaurice et al. (2011).

The multilevel modeling approach can also be viewed as a combination

[1]The nomenclature and interpretation of "Fixed effects" and "Random effects" is not uniform and is often a source of confusion among researchers. An account of their various interpretations is given in the footnote of Gelman and Hill (2007), p. 245. They identify five different definitions of these terms, out of which, we adopt the first one. Accordingly, "fixed effects" are fixed across all individuals whereas random effects are individual-specific. In the context of our model, the "fixed effects" are parameters (β)s, including the coefficients of the X-covariates as well as the overall population intercept (which can be interpreted as the expected value of the random intercepts in multi-level modeling), estimated from regression, and "random effects" refer to the random deviation estimated to capture subject heterogeneity. These fixed and random effects, together, form the Linear Mixed Effects model.

of regressions, conditional and marginal model, correlated error model, multivariate normal model, penalized regression and Bayesian mixed model. For a complete overview of the different modeling approaches, one can refer to Clark (2016).

3. Methodology

A variant of clustered data is longitudinal data, wherein, the ordering of the repeated measures is to be preserved for analysis (Fitzmaurice et al., 2011). Several methods of longitudinal data analysis come under the purview of those which are used in the more general case of clustered data, as given in Fitzmaurice et al. (2011).

Various models have been proposed for efficiency estimation with longitudinal data in SFA. Schmidt and Sickles (1984) provide a framework for estimating the production frontier, wherein, inefficiency is assumed to be time-invariant. It can be estimated by way of fixed effects or random effects. Models addressing time-varying efficiency estimation with longitudinal data were proposed thereafter - each building upon the previous in order to separate inefficiency, as distinct from heterogeneity. Inefficiency estimation has further been bifurcated into persistent and transient – giving rise to another class of models which estimate it.

We estimate technical efficiency scores at the farm level independently from two models: (i) Plot-level Stochastic Frontier, and (ii) the Random Intercept Model (henceforth, the Linear Mixed Effects (LME) Model). Having arrived at two efficiency scores for the same farm household from the two

models, we measure Spearman's rank correlation coefficient and Kolmogorov-Smirnov D statistic for comparing their performance. Estimation of the correlation coefficient is for checking if the ranking of farm households, based on their efficiency, is in accordance with the true ranking. The Kolmogorov-Smirnov D statistic, which measures the maximum difference between two distributions, has been estimated to check if the full distribution of the true efficiency can be arrived at.

3.1. Plot-level Stochastic Frontier Model

We use the stochastic frontier model proposed by Aigner et al. (1977) and Meeusen and van den Broeck (1977) for plot-level stochastic frontier estimation, as given by equation (5).

$$Y_j = \beta_o + \beta_k X_{jk} + \upsilon_j - u_j \tag{5}$$

$$j = 1, 2, ..., n$$

$$\upsilon_j \sim N(0, \sigma_\upsilon^2) \tag{6}$$

$$u_j \sim N^+(0, \sigma_u^2) \tag{7}$$

j is an index for a plot. Y_j is the output of plot j. X_{jk} is the k^{th} input applied on plot j. β_k is the estimated coefficient corresponding to input k. u_j is the one-sided inefficiency at the level of plot j and υ_j is the symmetric statistical noise term, which is meant to capture measurement error at the plot level. υ_j is assumed to follow the normal distribution with zero mean

and variance, σ_v^2. u_j is assumed to follow a half-normal distribution with zero mean and variance, σ_u^2.

From the plot-level Stochastic Frontier model, we estimate the technical efficiency of each plot based on Jondrow et al. (1982). For deriving efficiency estimates at the farm level from the plot efficiency scores, we use four composite indices: Arithmetic Mean (AM), Output-Weighted Arithmetic Mean (WAM), Geometric Mean (GM) and Output-Weighted Geometric Mean (WGM). The weight refers to the share of the plot's output in the total output of the respective farm household and is applied to the plot's efficiency score.

We estimate a Cobb-Douglas production frontier. Our main interest is in studying aggregation of technical efficiency, and less in the fit of the functional form. Previously, Ruggiero (1999), Ondrich and Ruggiero (2001), and Banker et al. (1993) have used the Cobb-Douglas specification for comparison of different methods of efficiency estimation. The dependent variable is output of maize of each plot.

3.2. The Linear Mixed Effects Model

The LME uses the hierarchical structure of the data concerning maize production on multiple plots owned by farm households. Hierarchical data tend to exhibit (positive) correlation within repeated measurements of a cluster. If the presence of this correlation is not accounted for, it leads to erroneous statistical inference as the resultant standard errors are too high. The statistics for hypothesis testing such as the p-value will be flawed (Fitzmaurice

et al., 2011). The LME model remedies the problem with the help of a random effects induced covariance structure. It also facilitates the inclusion of covariates which vary at the household (cluster) level and not the plot level - the Z-variables specified in the stochastic frontier model can be classified as such group-level predictors.

The LME model has been estimated with the following specification:

$$y_{ij} = \beta_o + \beta_k X_{ijk} + \gamma_{oi} + \epsilon_{ij} \tag{8}$$

$$i = 1, 2, ..., m$$

$$j = 1, 2, ..., n_i$$

$$y_{ij} = \alpha_i + \beta_k X_{ijk} + \epsilon_{ij} \tag{9}$$

i refers to a farm household out of a total of m observed farm households. j refers to a plot which belongs to the farm household i. The total number of plots owned by a farm household, n_i, is not the same for all households. Hence, the subscript i has been assigned to n, for denoting the total number of plots owned by a specific household i. Analogous to equation (5), y_{ij} is

the maize output of plot j which belongs to farm household i, x_{ijk} is the k^{th} input applied on plot j of farm i and β is a vector of estimated coefficients corresponding to the inputs. γ_{oi} measures the extent to which household i differs in its "base" level of production from the population fixed intercept β_o. Thus, the sum of β_o and γ_{oi} is the household random intercept, denoted by α_i. ϵ_{ij} is the plot level random error.

Having estimated the random intercept for each farm household, we estimate the farm efficiency score, as proposed by Schmidt and Sickles (1984). Aggregation of efficiency indices is carried out as given in Equations (10) and (11).

$$u_i = \max(\alpha_i) - \alpha_i \tag{10}$$

$$TE_i = exp(-u_i) \tag{11}$$

The random intercept estimated at the household level is transformed to arrive at household-level technical efficiency, denoted by TE_i.

4. Variable Description

The selection of inputs as frontier covariates follows Liu and Myers (2009), who, in a bid to introduce a model choice procedure across different specifications of the stochastic frontier model, also estimate the model for maize production from a survey of smallholders in Kenya. Similar to Liu and Myers (2009), we also distinguish between inputs which would determine the physical output of maize and those which are expected to affect production by

operating as farm management characteristics i.e. Z variables. There is, however, some dissimilarity in the measurement of some variables as compared to our study.

The inputs which are included in the estimation of the production frontier are plot size, seed usage, labour (pre-harvest as well as post-harvest, family as well as hired) and the quantity of fertilizer, pesticide, and manure. . An interaction term of seed usage and fertilizer application has been included for estimating the differential impact of fertilizer, given a unitary increase in seed usage. Additionally, we incorporate a dummy variable for the soil quality of each plot, viz., poor, medium and good. Medium soil type is the reference category and the effect of poor and good soil is captured through dummy variables. Similarly, the season of cultivation is controlled for by introducing a dummy variable for long rains (March-April, 2012). The season of short rains (October-November, 2011) is the reference category.

The six Agro-Ecological Zones (AEZs) to which the plots belong have been split into five dummy variables, with Coastal Lowland being the reference category. These AEZ dummies would account for the difference in environmental conditions, the omission of which, would result in an omitted variable bias as they determine input level decisions (Liu and Myers, 2009). Additional dummy variables for certain inputs have been included in order to accommodate for zero input values for some plots under a Cobb-Douglas specification (Battese, 1997).

The set of Z-variables comprises of the farm-level inputs which are expected to affect efficiency. The maximum level of education among the mem-

bers of a household would affect the efficiency of the household in production. Similarly, the distance to the nearest agricultural extension service center from the household residence is expected to inversely affect production efficiency. The type of land ownership affects the incentive structure for investment through the notion of tenure security (Liu and Myers, 2009). Therefore, the proportion of land owned out of the total land cultivated by the household has been included. The measurement of this variable differs from Liu and Myers (2009) as they create a dummy variable, depending upon whether the concerned field was owned or rented. We also include a dummy variable which is indicative of whether the farm household tried to avail credit and was unsuccessful in doing so, as this is expected to reduce efficiency by distorting the timing of input usage.

5. Data

The survey was concentrated in the areas which mainly grow maize, spread across the six AEZs of Kenya. The classification of AEZs is based on the one given by Hassan et al. (1998). These AEZs were the strata from which rural sublocations were sampled using the probability proportionate to size method. Households were randomly sampled from these sublocations. The reference year for recall was 2012. The data used is a subsample[2] comprising of 2799 plots, owned by 1050 households. The count of plots from each AEZ is given in Table 1. Some plots are repeated in the data in order to

[2]Plots which reported crop failure, as indicated by zero harvest of maize and zero harvest labour were excluded.

account for cultivation in two seasons, long and short rains. The number of observations according to the season is given in Table 2. The total number of households exceeds 1050 because there are some plots cultivated in both seasons by them.

Table 1: Count of plots by AEZ

AEZ	No. of plots
Highland tropics	234
Moist transitional	578
Dry transitional	638
Dry mid-altitude	532
Moist mid-altitude	619
Lowland tropics	198
Total	2799

Table 2: Count of observations by season

Season	No. of plots	No. of households
Long rains (March-April)	1576	921
Short rains (Oct-Nov)	1223	734
Total	2799	1655

The count of households according to plot ownership is given in Table 3.

Table 3: Count of households by plot ownership

No. of plots	No. of households	
	Long rains	Short rains
1	481	404
2	301	213
3	87	84
4	37	26
5	10	6
6 or more	5	1
Total	921	734

The descriptive statistics are given in Table 4.

Table 4: Descriptive Statistics

Variable	Unit	Mean	SD	Min	Max
Dry Harvest	Kg	363.6	534.05	1	5490
Plot size	Acre	1.05	0.79	0.05	4
Seed	Kg	6.98	6.3	0.5	60
Fertilizer	Kg	26.15	55.94	0	600
Pesticide	Liters	0.1	0.5	0	6
Manure	Kg	268.6	539.34	0	7000
Labour	Person-days	22.53	20.48	1	210
Poor soil	Dummy	0.12	0.33	0	1
Good soil	Dummy	0.36	0.48	0	1
Max education	Years	10.89	2.87	0	18
Credit Shortage	Dummy	0.12	0.33	0	1
Distance to extension	Km	7.63	8.4	0	80
Female headed HH	Dummy	0.16	0.37	0	1
Own cultivation	Proportion	0.85	0.26	0	1

[a] SD stands for the standard deviation.

The seed types were mainly recycled hybrids, local varieties or Open Pollinated Varieties (OPVs). Fertilizer mainly consists of quantities of DAP[3] and variants of NPK[4]. Fertility of plot soil was self-reported by the farmers.

These data constitute a hierarchy, wherein a farm-household/producer

[3]DAP stands for Diammonium Phosphate

[4]NPK stands for Nitrogen, Phosphorous and Potassium

owns cultivable plots. We use the single level structure of ownership of multiple plots (repeated measurements) by households (the group at a higher level).

6. Monte Carlo simulation

The purpose of carrying out Monte Carlo simulation is to check the performance of the two models in arriving at the true aggregated efficiency scores at the household level as well as observe the effect of changes in plot error on their performance.

We use an artificially created hierarchical set of data from our original data. We establish a balanced cluster of farm households by assigning them a random identification variable which is common across 3 plots per household[5] Thus, we have a cluster of 933 households who own 2799 plots, each supposed to be owning 3 plots. We generate the random deviation (γ_{oi}) at the household level from a skew-normal distribution with zero mean, standard deviation 1 and omega parameter as -2.

$$\gamma_{oi} \sim SN(0, 1, -2)$$

$$\alpha_i = 4.35 + \gamma_{oi} \tag{12}$$

[5]We extended the analysis procedure to 9 plots per household and found identical patterns in the results.

$$u_i = \max(\alpha_i) - \alpha_i \tag{13}$$

Having assumed the fixed population intercept β_o as 4.35 (the average value returned in model estimation), we compute the unique random intercept specific to each household, as given in equation (12). We use equation (13) to arrive at household inefficiency estimates, and equation (14) to generate the true efficiency score for each household.

$$TE_{true} = exp(-u_i) \tag{14}$$

We generate random numbers for the plot error term υ_j with different combinations of the parameters pertaining to the assumed normal distribution with parameters μ and σ_υ^2, as the mean and variance, respectively. Thus, σ_υ indicates the standard deviation of the plot error.

$$\upsilon_j \sim N(\mu, \sigma_\upsilon^2)$$

We use two of our X-covariates from the data, plot-size and labour, with their respective elasticities, 0.45 and 0.35, and calculate the true values of y_j through the data generating mechanism, given by equations (15) and (16).

$$\log y_j = \alpha_i + 0.45 \log plotsize_j + 0.35 \log labour_j + ploterror_j \tag{15}$$

$$y_j = \exp(\log y_j) \tag{16}$$

We apply the two models, plot-stochastic frontier and LME model, in their original specification as given in Section 3.1 and Section 3.2, on this

Table 5: Monte Carlo simulation statistics for $\mu = 0$

	σ_v	ρ	D	σ_v	ρ	D	σ_v	ρ	D
AM	0.2	0.98	0.74	0.4	0.94	0.78	0.6	0.88	0.82
WAM	0.2	0.97	0.74	0.4	0.93	0.79	0.6	0.85	0.84
GM	0.2	0.98	0.73	0.4	0.94	0.77	0.6	0.88	0.81
WGM	0.2	0.97	0.74	0.4	0.93	0.79	0.6	0.86	0.84
LME	0.2	0.98	0.06	0.4	0.94	0.08	0.6	0.88	0.10

newly generated dependent variable, y_j and compute aggregated farm efficiency by their respective aggregation strategies. We use 500 replications of the simulation procedure and compare the farm efficiency scores generated thus, from the two models, with the true values, using Spearman's rank correlation coefficient (ρ) and Kolmogorov-Smirnov test statistic (D). Tables 5 and 6 present the mean[6] of ρ and D, for different plot error parameter combinations, across the 500 replications. They are presented for the values, μ=0 and μ=2 [7].

[6]The standard deviation of ρ across 500 simulations was 0.00 and increased to a positive integer in the second decimal place as the plot error standard deviation increased.

[7]Further, Monte Carlo simulation was carried out for other values of the assumed mean of the plot error such as -2, 4, and -4. The pattern, as observed in Tables 5 and 6, did not change.

Table 6: Monte Carlo simulation statistics for $\mu = 2$

	σ_v	ρ	D	σ_v	ρ	D	σ_v	ρ	D
AM	0.2	0.98	0.74	0.4	0.94	0.78	0.6	0.88	0.82
WAM	0.2	0.97	0.74	0.4	0.93	0.79	0.6	0.85	0.84
GM	0.2	0.98	0.73	0.4	0.94	0.77	0.6	0.88	0.82
WGM	0.2	0.97	0.74	0.4	0.93	0.79	0.6	0.86	0.84
LME	0.2	0.98	0.06	0.4	0.94	0.09	0.6	0.88	0.09

An overall comparison of the ρ and the D statistic between the plot stochastic frontier model and LME reveals that both are able preserve the ranking of true efficiency scores, with minor differences between them. In each case, LME performs slightly better than the aggregation indices but the latter do produce high correlation as well. However, there is a stark contrast between the two models when one considers the D statistic. The aggregation indices lead to high values of the Kolmogorov-Smirnov D, most of them being close to one. The LME model produces low values of the D, most of them being close to zero. This indicates that LME model is well able to maintain the cumulative density of the true efficiency distribution.

An increase in the plot-level statistical error variability (σ_v) erodes the ranking of the efficiency scores, as ρ falls with an increase in the standard deviation. The plot stochastic frontier model as well as LME report a decrease in the correlation due to increase in σ_v.

As far as the choosing between the different aggregation indices is concerned, the Arithmetic Mean produces the highest correlation, as compared

to other indices. The Geometric Mean produces high correlation at lower levels of plot error variability but its performance drops to second to the Arithmetic Mean, when there is an increase in plot error standard deviation. However, although none of the indices are an appropriate choice according to D. Also, irrespective of whether it is WAM or WGM, the application of weights reduce the correlation, as compared to the unweighted means.

7. Empirical application

This section applies the two models on the existing maize data from smallholders in Kenya. Estimates are presented in tables 7 and 8.

Table 8: Estimates of classical inputs in plot stochastic frontier and LME model

	Plot Stochastic frontier	Linear Mixed Model	
Log(Harvest)	Estimate	Estimate	t-value
Intercept	5.07***	4.35	15.14
	(0.17)	(0.29)	
Log(Size)	0.52***	0.4	9.23
	(0.03)	(0.04)	
Log(Seed)	0.13***	0.17	4.06
	(0.04)	(0.04)	
Log(Labour)	0.05	0.18	4.48
	(0.03)	(0.04)	
Fertilizer dummy	0.11	0.03	0.33
	(0.07)	(0.1)	
Log(Fertilizer)	0.07*	0.09	2.16
	(0.03)	(0.04)	
Log(Seed)*Log(Fertilizer)	0.04***	0.02	1.81
	(0.01)	(0.01)	
Pesticide dummy	-0.08	-0.09	-0.97
	(0.06)	(0.1)	
Log(Pesticide)	0.02	-0.02	-0.28
	(0.03)	(0.06)	
Manure dummy	0.49***	0.04	0.31
	(0.11)	(0.14)	
Log(Manure)	0.07***	0.01	0.37
	(0.02)	(0.02)	

[a] Figures have been rounded upto 2 decimal places.

[b] *, **, *** correspond to 0.1, 0.05 and 0.01 level of significance, respectively.

[c] Standard errors are reported in parenthesis.

Table 7: Estimates of dummy variables in plot stochastic frontier and LME model

	Plot Stochastic frontier	Linear Mixed Model	
Log(Harvest)	Estimate	Estimate	t-value
Poor soil	-0.27***	-0.16	-2.15
	(0.05)	(0.07)	
Good soil	0.15***	0.08	1.73
	(0.04)	(0.05)	
Long rains dummy	0.05	-0.03	-1.01
	(0.03)	(0.03)	
High Tropics	0.68***	0.59	3.59
	(0.09)	(0.17)	
Moist Transitional	0.35***	0.36	2.33
	(0.08)	(0.15)	
Dry Transitional	0.2**	0.1	0.62
	(0.08)	(0.16)	
Dry Mid-Altitude	0.2**	0.03	0.2
	(0.08)	(0.16)	
Moist Mid-Altitude	0.37***	0.41	2.71
	(0.08)	(0.15)	

[a] Figures have been rounded upto 2 decimal places.

[b] *, **, *** correspond to 0.1, 0.05 and 0.01 level of significance, respectively.

[c] Standard errors are reported in parenthesis.

In the estimation of plot stochastic frontier, monotonicity is globally satisfied as all output elasticities are positive. The largest output elasticity is

that of the plot size, followed by seed. Soil fertility plays a major role in determining production, as the coefficients of both, poor as well as good type are significantly different from medium category, which is the reference category. Also, they have opposite signs, as expected. The five AEZs included in the model perform significantly better than Coastal Lowland, which is the reference category. The LME model hints at a significant effect of labour on the frontier, with an output elasticity of 0.18 percent.

Table 9 provides the coefficients of Z-variables/group-level predictors, respectively, from plot stochastic frontier and the LME model. Education in the household and a female headed household are significant in explaining inefficiency, according to plot stochastic frontier. The former reduces inefficiency and the latter increases it, as expected. The negative effect of a female headed household is in lines with the result of Liu and Myers (2009). They explain this adverse effect on efficiency through the fact that it is difficult for women to possess land ownership rights, unlike men and this affects the incentive to work. The coefficients of group-level predictors in LME model indicate the effect of a unitary increase in the predictor on the household random deviation.

Table 10 presents model-specific results. In terms of AIC, BIC and Log-likelihood, the LME model fares better than the plot stochastic frontier. Further, the higher value of the standard deviation of the random effect, denoted by σ_γ, as compared to the residual standard deviation, confirms a high level of heterogeneity among the farm households in production. The estimates of Gamma and variance share of the inefficiency term in plot stochastic

Table 9: Estimates of coefficients of plot-invariant variables

	Plot Stochastic frontier	Linear Mixed Model	
Log(Harvest)	Estimate	Estimate	t-value
Max education	-0.21***	0.01	0.67
	(0.05)	(0.01)	
Female headed HH	0.81***	-0.15	-1.64
	(0.23)	(0.09)	
Cultivated land owned	-0.83	-0.36	-2.59
	(0.49)	(0.14)	
Distance to Extension	-0.01	0	-1.13
	(0.01)	(0)	
Credit Shortage	-0.53	0.13	1.26
	(0.3)	(0.1)	

[a] Figures have been rounded upto 2 decimal places.

[b] *, **, *** correspond to 0.1, 0.05 and 0.01 level of significance, respectively.

[c] Standard errors are reported in parenthesis.

frontier confirm the existence and high level of inefficiency in production by the households.

Table 10: Model-specific estimates

	Plot Stochastic frontier	Linear Mixed Model
AIC	8117	7210
BIC	8272	7364
Log likelihood	-4033	-3579
σ_γ		1.03
Residual σ		0.577
Mean efficiency	0.446	
Gamma	0.967	
Var(u) / Var(u)+Var(υ)	0.914	

[a] σ_γ denotes the standard deviation of the household random deviation γ_{oi}.

[b] Residual σ is the standard deviation of the residuals after LME estimation.

8. Conclusion

This is study is a first in examining the performance of the LME model and aggregation indices in estimating technical efficiency when there is a data hierarchy, using Stochastic Frontier Analysis. We perform Monte Carlo simulations with replications of the data generating process, using different parameter combinations of the plot error and observe the mean of correlation and Kolmogorov-Smirnov statistic of plot-level stochastic frontier and the Linear Mixed Effects model, in order to compare the accuracy in efficiency estimation at an aggregate (farm household) level. We observe that

both models maintain the ranking of households according to the true ranking. However, the LME also closely estimates the true efficiency distribution, unlike aggregation indices of the plot-level stochastic frontier. The variability of plot-level error plays a systematic role in affecting the performance of both models. As it increases, the comparability of both models with the true aggregate efficiency distribution is reduces progressively.

The empirical application of the two models on maize production data, collected from smallholders in Kenya, gives insight into the factors which play a role in determining production of maize and the efficiency of the concerned farm households. There is scope for increasing the production through improving the soil fertility. Higher education achieved in the household improves efficiency in production.

Further, one can explore the potential of the LME model by incorporating more levels in the data hierarchy and use Multilevel Modeling to check the robustness of the results of our study.

References

Aigner, D., Lovell, C. A. K., and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6:21–37.

Banker, R. D., Gadh, V. M., and Gorr, W. L. (1993). A Monte Carlo comparison of two production frontier estimation methods : Corrected ordinary

least squares and data envelopment analysis. *European Journal of Operational Research*, 67:332–343.

Battese, G. E. (1997). A note on the estimation of Cobb-Douglas production functions when some explanatory variables have zero values. *Journal of Agricultural Economics*, 48(2):250–252.

Blackorby, C. and Russell, R. R. (1999). Aggregation of Efficiency Indices. *Journal of Productivity Analysis*, 12:5–20.

Brorsen, B. W. and Kim, T. (2013). Data aggregation in stochastic frontier models: The closed skew normal distribution. *Journal of Productivity Analysis*, 39(1):27–34.

Clark, M. (2016). Thinking About Mixed Models. https://m-clark.github.io/docs/mixedModels/mixedModels.html, last accessed on 2020-04-20.

Cook, W. D., Chai, D., Doyle, J., and Green, R. (1998). Hierarchies and Groups in DEA. *Journal of Productivity Analysis*, 10(2):177–198.

Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, second edition.

Fried, H., Lovell, C. A. K., and Schmidt, S., editors (1993). *The Measurement of Productive Efficiency Techniques and Applications*. Oxford University Press.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Hassan, R. M., Njoroge, K., Njore, M., Otsyu!a, R., and Laboso, A. (1998). *Adoption Patterns and Performance of Improved Maize in Kenya in Maize Technology Development and Transfer: A GIS Application for Research Planning in Kenya*. CAB International.

Jondrow, J., Knox Lovell, C. A., Materov, I. S., and Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics*, 19(2-3):233–238.

Liu, Y. and Myers, R. (2009). Model selection in stochastic frontier analysis with an application to maize production in Kenya. *Journal of Productivity Analysis*, 31(1):33–46.

Meeusen, W. and van den Broeck, J. (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review*, 18(2):435–444.

Ondrich, J. and Ruggiero, J. (2001). Efficiency measurement in the stochastic frontier model. *European Journal of Operational Research*, 129:434–442.

Ruggiero, J. (1999). Efficiency estimation and error decomposition in the stochastic frontier model: a Monte Carlo analysis. *European Journal of Operational Research*, 115(3):555–563.

Schmidt, P. and Sickles, R. C. (1984). Production frontiers and panel data. *Journal of Business and Economic Statistics*, 2(4):367–374.