Assessing the effect of land-use diversity on farms' productive efficiency

Natalia Kuosmanen
MTT Agrifood Research Finland
February 2010

1. INTRODUCTION

Agricultural diversity is a managerial practice that can influence farms' performance and their productive efficiencies. Knowing and understanding the effect of diversity on farm's productive efficiency can help farmers and policymakers to improve managerial strategies in agriculture, to make decisions, adapt to various external changes that might influence farms' productive performance.

Operationalized measures of factors that influence productive efficiency (such as, agricultural diversity) in the productivity literature are commonly referred to as contextual variables (see e.g., Banker and Natarajan 2008). The effect of contextual variables on efficiency can be estimated by several approaches. However, the two-stage data envelopment analysis (DEA) firstly applied by Ray (1988, 1991) is one of the common approaches used to assess the impact of contextual variables on productive efficiency. In the two-stage DEA, the efficiency scores are estimated by non-parametric DEA method in the first stage, and in the second stage, the estimated DEA efficiency scores are regressed on the contextual variables.

Currently, two-stage DEA method is subject to debate and critique for the bias, serial correlation of the efficiency estimates and lack of coherent data generating process (Hoff 2007, McDonald 2009, Simar and Wilson 2007 among others). The main argument is that the conventional methods of statistical inference, such as t-test for the significance and the confidence intervals, are invalid in the second stage of this method. However, in the recent study by Banker and Natarajan (2008), the authors provided a formal statistical basis for the two-stage DEA method for assessing the impact of contextual variables on productive efficiency. In particular, they showed that the two-stage DEA estimator, comprising a DEA model followed by ordinary least squares or maximum likelihood estimation, is statistically consistent estimator under certain assumptions and regularity conditions.

In their recent study, Johnson and Kuosmanen (2009) proposed an unbiased one-stage estimator, where the effects of contextual variables are estimated within the Convex Nonparametric Least Squares (CNLS) regression (Hidreth 1954, Hanson and Pledger 1976). They showed that the CNLS-estimator for the contextual variables is statistically consistent under more general assumptions than the two-stage DEA estimator. Moreover, the CNLS-estimator for the contextual variables has most important properties that ensure validity of the conventional methods of statistical inference for this estimator. Monte Carlo simulations, performed in a wide variety of scenarios, provide evidence that the CNLS-estimator yields consistently more precise estimates than two-stage DEA and provide strong support for using this estimator for estimation of the effects of the contextual variables.

The purpose of this paper is to investigate the effect of agricultural diversity on farms' productive efficiency. To measure agricultural diversity, in this study we consider diversity in land-use and calculate the Shannon-Weaver index [henceforth Shannon index] (Shannon and Weaver 1994) at the farm-level for eight different land-use types. Calculated Shannon index for each farm in the further analysis is referred to as a contextual variable. Finally, to estimate the effect of the contextual variable on farms' productive efficiency, we utilize the one-stage CNLS estimator (Johnson and Kuosmanen 2009, Kuosmanen and Johnson 2010).

The rest of the paper is organized as follows. Section 2 introduces the theoretical model to be estimated and the one-stage CNLS estimator together with some references. In section 3, we briefly recapture the calculation procedure for the common measure of diversity - Shannon index. Section 4 presents an empirical application to Finnish farms and results. Section 5 concludes.

2. MODEL

Our theoretical model is specified by single-output as a general function of multiple inputs and an error term. For formal representation, consider a group of firms $i = 1,\ldots,n$. The resource use by firm *i* is characterized by a vector $\mathbf{x}_i = (x_{i1} \ldots x_{iR})'$, and the economic output is denoted by y_i. Thus, following the traditional approach for efficiency measurement (Farrell 1957), the economic output y may represent a farm's output measured in revenue terms, the inputs $\mathbf{x}$ may be labor, land and farm capital.

Hence, the cross-sectional production model is specified as follows (following Banker and Natarajan 2008, Johnson and Kuosmanen 2009):

(1) $y_i = f(\mathbf{x}_i) \cdot e^{\varepsilon_i}$, $i = 1,\ldots,n$.[1]

In model (1), the data on n firms are generated from the production function $f(\mathbf{x}_i)$, which is increasing and concave function characterizing the best-practice production technology, and an error term ε_i captures all deviations from the best-practice production function.

Further define the error term ε_i consisting of three distinct components:

(2) $\varepsilon_i = s_i \beta - u_i + v_i$,

[1] e^x is the exponential function, where the base e is the irrational number, expressed mathematically,

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n \approx 2.71828 .$$

In (2), s_i represents a contextual variable; $u_i \geq 0$ is a random inefficiency term of firm *i* that represents technical inefficiency; and v_i is a random noise term that represents the effects of omitted factors, measurement errors, and other stochastic noise in data. Variables v_i and u_i are assumed to be independently distributed random variables that are uncorrelated with the input variables $\mathbf{x}_i$, the contextual variable s_i, and with each other. Note, except for the contextual variable component, the specification of the error term in (2) is analogous to error term formulations in parametric stochastic frontier models (Aigner *et al*. 1977, Meeusen and Vanden Broeck 1977).

In specification (2), pursuing the example of farms, the contextual variable s_i that affects the firms' productive efficiencies may be represented by the Shannon index. The Shannon index is a common measure for an agricultural diversity at the farm-level and may influence farms' productive performance. We discuss the Shannon index in more details in the next section.

Implying previous insights, rewrite model (1) as follows:

(3) $$y_i = f(\mathbf{x}_i) \cdot e^{s_i\beta - u_i + v_i}, \text{ for } i = 1, \ldots, n.$$

In equation (3), the component $(s_i\beta - u_i)$ characterizes the total technical inefficiency of firm *i*; the term $s_i\beta$ represents technical inefficiency that is explained by the contextual variable *s*; and the term u_i represents the proportion of inefficiency unexplained by factor s_i. The coefficient β in equation (3) is an unknown parameter to be estimated. It represents the average effect of the contextual variable s_i on firm *i*'s performance.

For further discussion of the effect of contextual variables on firms' productive efficiency and, in particular, interpretation of the coefficient β for the contextual variable *s*, recall what do we mean by a firm's productive efficiency. Productive (or technical) efficiency is just one component of overall economic efficiency. However, to be economically efficient, a firm must first be technically efficient: it must produce the maximum output from a given set of input resources. The level of technical efficiency of firm i is characterized by the relationship between observed production and some ideal (or potential) production. The measurement of a firm's technical efficiency is based upon deviations of observed output y_i from the best (efficient) production frontier $f(\mathbf{x}_i)$.

Specifically, a measure of productive (or technical) efficiency of the i's firm is defined as the ratio of the observed output y_i for the i's firm, relative to the potential output, defined by the production (or frontier) function, given the input vector, $\mathbf{x}_i$. Applying formula (3), the productive efficiency of firm i can be represented as follows:

(4) $$\frac{y_i}{f(\mathbf{x}_i)} = e^{s_i\beta - u_i + v_i}, \text{ for } i = 1,\ldots,n.$$

Expression (4) provides an intuition for the interpretation of the coefficient β and hence for the effect of the contextual variable s_i on firm i's productive efficiency. Particularly, positive coefficient $\beta > 0$ indicates positive effect: larger the value of the contextual variable s, larger the productive efficiency. Vice versa, negative coefficient $\beta < 0$ of the contextual variable s indicates inverse effect: larger the value of the contextual variable smaller the value of firm specific productive efficiency.

Next, we consider the estimation of model (3). By taking logarithms of both sides of equation (3), the model can be rewritten as:

(5) $$\ln y_i = \ln f(\mathbf{x}_i) + s_i\beta - u_i + v_i,\ i = 1,\ldots,n.$$

Denoting $\gamma_i = \ln y_i$, $\varphi_i = \ln f(\mathbf{x}_i)$, and the composite residuals $\tilde{\varepsilon}_i = v_i - u_i$, model (5) can be represented as:

$$(6) \qquad \gamma_i = \varphi(\mathbf{x}_i) + s_i\beta + \tilde{\varepsilon}_i, \; i = 1,\ldots,n.$$

To estimate model (6), we employ the Convex Nonparametric Least Squares (CNLS) that is the oldest approach in the productivity literature, dating back to the work of Hildreth (1954). CNLS does not assume any particular functional form for f a priori. Instead, it postulates that f belongs to the set of continuous, monotonic increasing and globally concave functions, denoted by F_2. The CNLS problem is to find $f \in F_2$ that minimizes the sum of squares of the residuals, formally, for model (6):

$$(7) \qquad \min_{f,\beta} \sum_{i=1}^{n} \left(\gamma_i - \varphi(\mathbf{x}_i) - s_i\beta\right)^2 \text{ s.t. } f \in F_2.$$

The CNLS problem (7) identifies the best-fitting function f from the family F_2, which includes an infinite number of functions. This makes problem (7) generally difficult to solve. However, the family F_2 can be equivalently represented by a family of piece-wise linear functions characterized by the Afriat's theorem (Afriat 1967, 1972) as has been shown in Kuosmanen (2008). The infinite dimensional problem (7) can be transformed into an equivalent finite-dimensional quadratic-programming (QP) problem that can be solved by standard mathematical algorithms. In the presence of contextual variables, the one-stage CNLS estimator estimates the shape of the production function f and the effects of contextual variable s:

$$(8) \qquad \min_{\boldsymbol{\alpha},\beta,\boldsymbol{\delta},\hat{\mathbf{y}}} \sum_{i=1}^{n} \left(\ln y_i - \ln \hat{f}(\mathbf{x}_i) - s_i\beta\right)^2 \left| \begin{array}{l} \hat{f}_i = \alpha_i + \mathbf{x}_i\boldsymbol{\delta}_i \;\; \forall i = 1,\ldots,n; \\ \alpha_i + \mathbf{x}_i\boldsymbol{\delta}_i \le \alpha_h + \mathbf{x}_i\boldsymbol{\delta}_h \;\; \forall h,i = 1,\ldots,n; \\ \boldsymbol{\delta}_i \ge 0 \;\; \forall i = 1,\ldots,n. \end{array} \right.$$

In QP problem (8), the first constraint estimates α_i and $\boldsymbol{\delta}_i$ parameters for each observation: n different regression lines that can be interpreted as tangent lines are fit to the unknown function f. The second constraint imposes concavity through a system of inequality constraints on tangent lines, known as the Afriat inequalities (see Kuosmanen 2008 for details). The third constraint imposes monotonicity.

The CNLS estimator (8) for the contextual variables is unbiased, asymptotically efficient, asymptotically normally distributed, and converge at the standard parametric rate of order $n^{-1/2}$. Hence, the conventional methods of statistical inference, such as t-test for the significance and the confidence intervals, can be applied to the CNLS estimator [2].

3. SHANNON INDEX

In this study we consider land-use diversity and measure it by the Shannon index that is the common approach to measure diversity. The Shannon index quantifies the diversity of the countryside based on two components: the number of different land-use types and the proportional area distribution among land-use types. It is calculated by adding for each land-use type present the proportion of area covered, multiplied by that proportion expressed in natural logarithm, according to the equation:

(9) $$s_i = -\sum_{k=1}^{K} \frac{w_{ik}}{W_i} \ln \frac{w_{ik}}{W_i}, \text{ for } i = 1,\ldots,n \text{ farms.}$$

In equation (9), $k = 1,\ldots,K$ refers to the number of land-use types; w_{ik} stands for the area covered by land-use type *k* of farm *i*; W_i represents the total area of farm *i*; and $\frac{w_{ik}}{W_i}$ is proportion of area covered by land-use type *k*. Shannon index s_i equal to zero implies that total area of farm *i* contains only one land-use type *k*; whereas, Shannon index increases as the number of different land-use types increases.

[2] For statistical properties of the CNLS estimator in the presence of contextual variables, see Johnson and Kuosmanen (2009), Kuosmanen (2008), and Kuosmanen and Johnson (2010).

Note, the Shannon index calculated according to formula (9) addresses only the number and relative abundance of land-use types, it does not include how these types are distributed (Rosenzweig 1995).

4. APPLICATION

We next apply the one-stage CNLS estimator described in section 2 to the empirical data of 199 farms located in southern Finland. The purpose of this exercise is to investigate the effect of the contextual variable (i.e., the Shannon index representing land-use diversity) on the productive efficiency of farms. For comparison we also present the results from the stochastic frontier model of Cobb-Douglas form focusing on the coefficient for the Shannon index variable.

We first calculate the Shannon index according to equation (9) for each farm in the sample. It measures the land-use diversity for the following eight land-use types:

1) Cereals (SE035)
2) Other field crops (SE041)
3) Vegetables and berries (SE046*)
4) Flowers and ornamental plants (SE046**)
5) Seeds
6) Perennial crops (SE054-55, SE065)
7) Fodder crops and fallow (SE071-73)
8) Other area

Next, the calculated Shannon indices are analyzed in terms of linear dependence (correlation) with other variables, such as total UAA, labor, farm capital and output.

Finally, given the calculated values of the Shannon's indices, we apply the one-stage CNLS estimator and, for comparison, the Cobb-Douglas model.

4.1. Data

In this study we utilize empirical data of 199 farms located in southern Finland with the total utilized agricultural area (UAA) more than 50 hectares (ha). The data were collected from "EconomyDoctor" (Taloustohtori) farm-level database for the accounting year 2004. The farms' production activities are described by the economic output that is the total revenue from crops and crop products, livestock and livestock products and other output (SE131), expressed in Euros (€); and input resources, which include UAA in ha (SE025), labor input that is time worked in hours (hr) by total labor input on holding (SE011), and farm capital in € (SE510). An overview of key characteristics of the data is presented in Table 1 in the form of mean, standard deviation, minimum and maximum values.

Table 1

Descriptive statistics: year 2004, sample size n = 199.

Variable	Mean	St. Dev.	Min	Max
Total output, €	101,263	93,562	9,161	669,620
Labor, hr	3,620	2,219	361	13,458
Farm capital, €	343,404	248,300	42,929	1,770,813
UAA, ha	89.1	39.0	50.4	324.3

4.2. Results

4.2.1. Shannon index

To measure land-use diversity, we calculated the Shannon indices for each farm according to equation (9). Interestingly, only two farms from the sample of 199 farms had maximum of five different land-use types, the largest proportion of the farms (101 farms) had three different land-use types, and only one farm had one land-use type (see Table 2).

Table 2

Number of farms containing from 5 to 1 land-use types.

N. of land-use types	5	4	3	2	1	Total:
N. of farms	2	27	101	68	1	**199**

Fig. 1 gives a visual intuition of the relationship between the total UAA (horizontal axis) and the Shannon index (vertical axis). Each diamond in the scatter represents one observation. Using a simple linear regression, we fit a straight line through the set of observed values for the Shannon's indices and corresponding values of the UAA by minimizing the sum of squared residuals of the linear regression model.

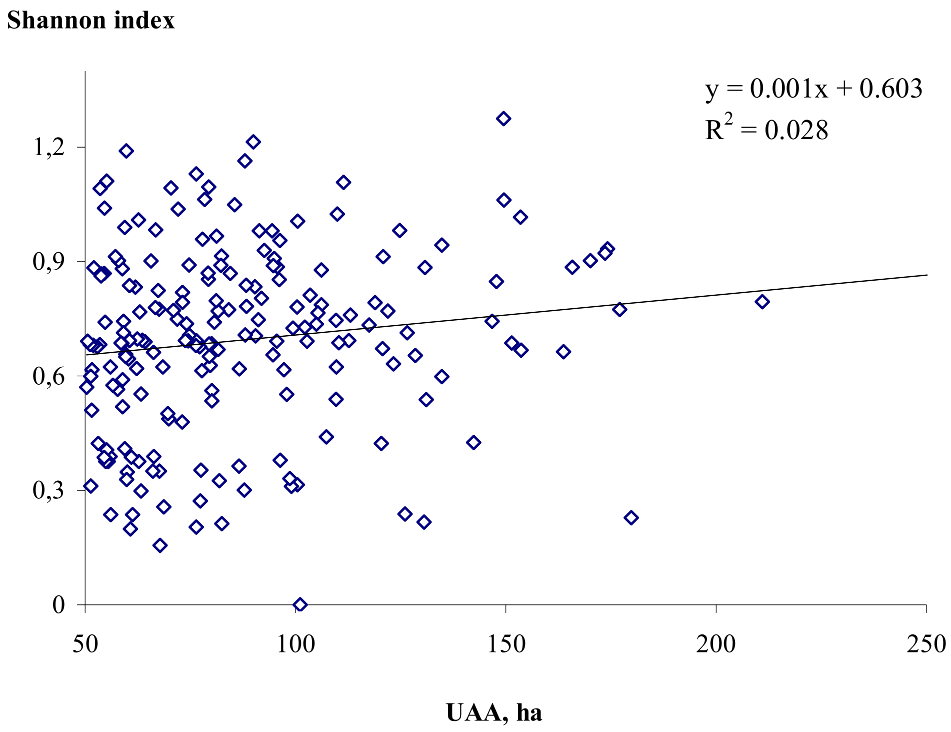

Fig. 1. The relationship between UAA and Shannon index on Finnish farms (n = 199).

As visible from Fig. 1, the relationship between the Shannon index and UAA is positive but rather week (slope of the line is positive and equals 0.001). The coefficient of determination R^2 is the standard goodness of fit statistics in the regression analysis. R^2 statistics measure how good the estimated regression equation is, the higher R^2 more confident one might be in the equation. The coefficient of determination has the range of values between 0 and 1; R^2 = 1 indicates that the regression line perfectly fits the data and R^2 = 0 indicates no linear relationship between dependent and independent variables. In our exercise, the coefficient of determination of R^2 = 0.03, which is rather small (nearly

3%). We cannot conclude that larger UAA will be associated with the larger Shannon index and hence larger diversity.

We may also assess correlation or linear dependence between the variables. Table 3 reports the Pearson's product moment correlation coefficients between the Shannon index, output and inputs (labor, UAA, and farm capital). The value of correlation coefficient varies between -1 and 1 and measures the degree of association (or the strength of linear relationship) between two variables. A positive value for the correlation implies a positive association, which means that larger values of one variable tend to be associated with larger values of other variable. By contrast, a negative value for the correlation implies an inverse association. In Table 3, the Shannon index is only positively correlated with the total UAA, but the correlation is rather small, r = 0.169. Some minor tendency exist, that larger land area is associated with the larger value of the Shannon index, and hence, larger diversity, though this association is rather weak.

Table 3

Correlation matrix of the Shannon index, output and inputs. Sample size n = 199

	Shannon Index	Output	UAA	Labor	Farm capital
Shannon Index	1				
Output	**-0.191**	1			
UAA	**0.169**	0.182	1		
Labor	**-0.084**	0.610	0.077	1	
Farm capital	**-0.154**	0.861	0.364	0.630	1

Surprisingly, the correlation between Shannon's indices and other variables, such as total output, labor and farm capital is negative. Negative correlation is evidence of a general tendency that large value of the Shannon index, and hence larger diversity, is associated with small values of output, labor and farm capital. In spite of negative sign of the correlation coefficients between the Shannon index and output, labor and farm capital, the coefficients are close to zero, which means that the association is rather weak (or no association at all).

Table 3 also reports correlations between inputs (such as, UAA, labor and farm capital) and output that is total revenue from production. All correlation coefficients are positive. Noticeably, the correlation between farm capital and total revenue is quite high, $r = 0.86$.

4.2.2. CNLS estimator

This section presents the results obtained by the one-stage CNLS estimator for the contextual variables discussed in Section 2. We first estimated model (3) by solving the QP problem (8). Here, we try to assess the statistical significance of the effect of the contextual variable (i.e. the Shannon index that represents the land-use diversity) on farms' productive efficiency. Table 4 reports the estimated coefficient β on the Shannon index along with other statistics.

Table 4

CNLS results: $R^2 = 0.026$.

	Coefficients	*Std Err.*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.5E-07	0.098	-4.6E-06	1.000	-0.194	0.194
Shannon index	**-0.307**	0.134	-2.299	**0.023**	-0.571	-0.044

The resulted coefficient of the Shannon index is negative $\beta = -0.307$ and significant: p-value is less than 0.05. However, the coefficient of determination, R^2, or the proportion of variation in the total revenue explained by changes in the Shannon index, is only 0.024, which means that only about 3% of the variability in the total revenue can be explained by the Shannon index. Hence, using just the Shannon index to explain the differences in farms' productive efficiency is not enough. Therefore, next we consider the agricultural specialization of farms and introduce dummy variables in our analysis.

Specifically, farms in our application belong to different types of farming, and there is a set of factors that are likely to be different between these farms. Since we cannot measure all of those differences within the confines of the study we are doing, we can use a dummy variable to capture these effects, to improve the fit of the model and try to obtain "pure" effect of the Shannon index on the farms' performance.

Agricultural specialization is a categorical variable; it consists of a series of categories (or codes that represent the type of farming) that are both exhaustive and mutually exclusive such that each observation is assigned to one and no more than one category. The most common method of creating variables to represent categories is to create dummy variables: a series of binary variables that identify whether or not each observation is a member of a specific category. A binary variable is a variable that is coded either as a one or as a zero. If an observation is classified as a member of a particular category, then it is coded as a one on the binary variable representing that category. Otherwise, this observation is coded as a zero on this binary variable.

In our application, the categorical variable "agricultural specialization" includes six categories (or types of farming)[3]:

1) Specialist cereals, oilseeds and protein crops (COP)
2) General field cropping (Field crops)
3) Specialist dairying (Milk)
4) Specialist granivores (Granivores)
5) Field crops and grazing livestock combined (Mixed crops*)
6) Various crops and livestock combined (Mixed crops**)

Table 5 reports the number of farms included in each type of farming category. The largest category of 73 farms represents specialist cereals, oilseeds and protein crops type of farming, and the smallest category of 15 farms represents specialist granivores type of farming.

Table 5

Types of farming and number of farms in each category.

Type of farming	COP	Field crops	Milk	Granivores	Mixed crops*	Mixed crops**	Total
N. of farms	73	29	39	15	18	25	199

[3] The codes for the types of farming according to TF8/TF14 classification are: specialist cereals, oilseeds and protein crops (13), general field cropping (14), specialist dairying (41), specialist granivores (50), field crops and grazing livestock combined (81), and various crops and livestock combined (82).

Agricultural specialization assumes one of six categories for each farm in the sample and is represented by five dummy variables s1-s5 (COP is a reference category). For each type of farming, we assign the value 1 for the category that matches the observation, and assign 0 for all other categories. This means that observations that match the reference category will have values of 0 for every dummy variable. Finally, we include the created dummy variables in model (3) and estimate it in the similar fashion by solving the QP problem (8).

Table 6 reports the estimated coefficients on the Shannon index and five dummy variables, together with the standard errors, t-ratios, p-values and the 95% confidence intervals. The statistical significance of the coefficients at the 5% significance level can be directly inferred from the confidence intervals. Note that the confidence intervals for all dummy variables (except for s4) in Table 6 do not include zero, these coefficients are statistically significant. The t-ratios for the coefficients of dummies *s*1-s3 and s5 are more than the critical value of 1.96[4] at the 5% significance level, and P-values are less than 0.05, which means that there is less than a 5% chance that the true values of these coefficients are zero. Hence, there is statistical evidence that dummy variables s1-s3 and s5 affect the total revenue (the value y). Moreover, the coefficients for these dummies have positive sign and hence have a positive effect on the revenue; larger the value of the coefficient, larger the revenue of the farms belonging to this type of farming.

Table 6

CNLS results with dummy variables: $R^2 = 0.304$.

	Coefficients	*Std. Err.*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.330	0.101	-3.276	0.001	-0.528	-0.131
Shannon index	**-0.174**	0.129	-1.345	**0.180**	-0.429	0.081
s1	0.423	0.090	4.701	4.94E-06	0.246	0.601
s2	0.389	0.078	5.018	1.19E-06	0.236	0.542
s3	0.693	0.114	6.089	6.08E-09	0.469	0.918
s4	-0.082	0.103	-0.798	0.426	-0.286	0.121
s5	0.432	0.092	4.689	5.19E-06	0.250	0.614

4 The critical value is obtained from the Student's t distribution table with *n-p* degrees of freedom, where *n* is the number of observations and *p* is the number of coefficients estimated, including the constant.

For the Shannon index and dummy s4, the lower bound of the confidence interval is negative, these coefficients are statistically insignificant and the evidence is inconclusive with respect to the Shannon index and s4. As well, t-ratios for these variables are less than the critical value of 1.96 and P-values are more than 0.05, there is more than 10% chance that the true values of these coefficients are zero. In conclusion, we do not have enough statistical evidence to reject the null hypothesis that the coefficient of the Shannon index equal to zero, but we also cannot say that there is no effect of the Shannon index on farms' productive efficiency; the effect is negative, but insignificant.

Comparing the results obtained with the CNLS estimator with and without dummies reveals some interesting insights. Note the coefficient of determination R^2 has increased from 3% in the model without dummies to nearly 30% in the model with dummies. By including dummy variables representing different types of farming, some effect that arises from the differences between the farm types is captured and explained by the dummies: the coefficients for the dummy variables have positive sign and statistically significant (except for dummy s4).

4.2.3. Cobb-Douglas regression

For comparison, we also applied a Cobb-Douglas stochastic frontier model (Aigner *et* al. 1977, Meeusen and Vandenbroeck 1977, Cobb and Douglas 1928) for estimating the coefficients of the parameters. Cobb-Douglas is a common functional form, and the model for three inputs and single output is defined by:

$$(10) \qquad y_i = e^{\alpha} x_{1i}^{\beta_1} x_{2i}^{\beta_2} x_{3i}^{\beta_3} e^{s_i \delta - u_i + v_i},$$

where y_i is the output of *i*'s firm, $x_{1i}, x_{2i},$ and x_{3i} are the inputs of *i*'s farm representing the variables: labor, land and farm capital; s_i is the Shannon index of *i*'s farm; $\varepsilon_i = v_i - u_i$ is the error term; $\beta_1, \beta_2, \beta_3,$ and δ are unknown parameters to be estimated.

Taking the natural logarithm of both sides of equation (10), we obtain:

(11) $\ln y_i = \alpha + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \beta_3 \ln x_{3i} + s_i\delta - u_i + v_i$, for $i = 1,\ldots,n$.

The logarithmic transformation (11) of the production function provides a log-linear form which is convenient and commonly used in econometric analyses using linear regression techniques (for example, employing a more general form of the function can allow for estimation of the coefficient values).

Firstly, we performed the regression of $\ln y$ on $\ln x_1$, $\ln x_2$, $\ln x_3$ and the Shannon index; and secondly, we did the same but included five dummy variables in model (11). The statistical results listed in Tables 7 and 8 are the ordinary least squares (OLS) estimates for the parameters of model (11). These results assume that there are no inefficiency effects, i.e., the term u_i is not included in the equation (11).

Table 7

Cobb-Douglas regression results: $R^2 = 0.735$.

	Coefficients	*Std Err.*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*
Intercept	-2.136	0.665	-3.210	0.002	-3.449	-0.824
lnx1	-0.189	0.107	-1.769	0.078	-0.400	0.022
lnx2	0.380	0.064	5.923	1.42E-08	0.254	0.507
lnx3	0.895	0.076	11.800	1.42E-24	0.746	1.045
Shannon index	**-0.188**	0.138	**-1.356**	**0.177**	-0.460	0.085

Table 8

Cobb-Douglas regression results with dummy variables: $R^2 = 0.815$.

	Coefficients	*Std. Err.*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*
Intercept	0.465	0.682	0.682	0.496	-0.880	1.810
lnx1	0.225	0.106	2.110	0.036	0.015	0.434
lnx2	0.223	0.070	3.195	0.002	0.085	0.360
lnx3	0.610	0.073	8.351	1.42E-14	0.466	0.754
Shannon index	**-0.176**	0.129	**-1.371**	**0.172**	-0.430	0.077
s1	0.570	0.095	5.977	1.11E-08	0.382	0.758
s2	0.696	0.128	5.423	1.78E-07	0.443	0.949
s3	1.012	0.145	6.967	5.23E-11	0.725	1.298
s4	0.123	0.122	1.008	0.315	-0.118	0.363
s5	0.700	0.114	6.144	4.65E-09	0.475	0.925

Consider first the goodness of fit. The R^2 for the model without dummy variables is 0.74, and with dummy variable is somewhat larger 0.82. This means that 74% of the variation in $\ln y_i$ is explained by the combined variation in $\ln x_{1i}, \ln x_{2i}, \ln x_{3i},$ and s_i; and 82% of the variation in $\ln y_i$ is explained by the combined variation in $\ln x_{1i}, \ln x_{2i}, \ln x_{3i}, s_i$ and dummy variables s1-s5. In terms of goodness of fit both models performed quite well.

In both scenarios, the results show that the estimated coefficient β on the Shannon index is negative (the coefficients are very similar, -0.188 and -0.176, obtained by the model without and with dummies, respectively) and there is no evidence that it is statistically significant. Firstly, the *t-ratio* for the coefficient β is about -1.4 in both scenarios that is less than the critical value of 1.97 at the 5% significance level. Secondly, *p-values* are about the same in both models: 0.177 without dummies and 0.172 with dummies and are larger than 0.05. For that reason, we cannot reject the null hypothesis that the estimated coefficient $\beta = 0$, but we also cannot say that there is no effect of the Shannon index on farms' productive efficiency: the effect is negative but insignificant.

Interestingly, by comparing the results from Table 6 and 8, reveals that the CNLS estimator and the Cobb-Douglas regression provide very similar results for the estimated coefficients on the Shannon index and p-values, both estimates are negative and insignificant. Furthermore, the estimated coefficients on dummy variables s1-s3 and s5 obtained with both methods are positive and statistically significant, with the largest value assigned to the coefficient of dummy s3.

5. CONCLUSIONS

In this paper, our objective was to estimate the effect of the land-use diversity represented by the Shannon index (a contextual variable) on farms' productive efficiency. Productive or technical efficiency is a common measure of performance of productive units, such as farms. To assess the statistical significance and the effect of the Shannon index on farms' productive efficiency we applied the Convex Nonparametric Least Squares (CNLS) estimator for the contextual variables. This estimator has most important properties to ensure validity of the conventional methods of statistical inference for this estimator and for the coefficients of the contextual variables. For comparison, we also applied the stochastic frontier model of Cobb-Douglas form to estimate the coefficients of the parameters.

In particular, we first calculated the Shannon's indices for 199 farms located in southern Finland and performed a preliminary data analysis in terms of correlation of the Shannon index with other variables, such as land, labor, farm capital and total revenue from production. This analysis revealed that the Shannon index is positively correlated only with the land, but negatively correlated with labor, farm capital and revenue. Though, all correlation coefficients were close to zero, which implies that the association between the Shannon index and other variables is rather weak, or no association at all.

We next estimated the coefficients for the parameters of the model by the CNLS estimator and the Cobb-Douglas regression. By utilizing both models, we considered two scenarios: with and without dummy variables. By including dummy variables in the models we aimed to capture the effects of different types of farming. In all scenarios we did not find a significant effect of the Shannon index representing the land-use diversity on farms' productive efficiency. By including the dummy variables, the models performed better, but the estimated coefficient on the Shannon index stayed negative and statistically insignificant.

In conclusion, in this application to the empirical data, we did not obtain enough statistical evidence to reject the hypothesis that the "true" coefficient of the Shannon index is really zero, and therefore we did not have enough statistical evidence that the Shannon index has an effect on farms' productive efficiency.

References

Afriat, S.N. (1967): The construction of a utility function from expenditure data. *International Economic Review* 8, 67-77.

Afriat, S.N. (1972): Efficiency estimation of production functions, *International Economic Review* 13, 568-598.

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

Banker, R.D. and R. Natarajan (2004): Statistical tests based on DEA efficiency scores. In *Handbook of Data Envelopment Analysis*. Kluwer, New York.

Banker, R.D. and R. Natarajan (2008): Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research* 56(1), 48-58.

Cobb, C.W. and P.H. Douglas (1928): A Theory of production, *American Economic Review* 18, 139-165.

Farrell, M.J. (1957): The measurement of productive efficiency, *Journal of the Royal Statistical Society Series A. General* 120, 253-282.

Hanson, D. L. and G. Pledger (1976): Consistency in concave regression. *Annals of Statistics* 4(6), 1038-1050.

Hildreth, C. (1954): Pont estimates of ordinates of concave functions. *Journal of the American Statistical Association* 49(267), 598-619.

Hoff, A. (2007): Second stage DEA: Comparison of approaches for modelling the DEA score. *European Journal of Operational Research* 181(1), 425-435.

Johnson A.L. and T. Kuosmanen (2009): How operational conditions and practices affect productive performance? Efficient semi-parametric one-stage estimators. Working paper.

Kuosmanen, T. (2008): Representation theorem for convex nonparametric least squares. *The Econometrics Journal* 11, 308-325.

Kuosmanen, T. and A.L. Johnson (2010): Data envelopment analysis as nonparametric least squares regression. *Operations Research* 58(1), 149-160.

McDonald, J. (2009): Using least squares and tobit in second stage DEA efficiency analyses. *European Journal of Operational Research* 197(2): 792-798.

Meeusen, W. and J. Vandenbroeck (1977): Efficiency estimation from Cobb-Douglas production functions with composed error, *International Economic Review* 18, 435-445.

Ray, S. C. (1988): Data envelopment analysis, nondiscretionary inputs and efficiency: an alternative interpretation. *Socio-Economic Planning Sciences* 22(4), 167-176.

Ray, S. C. (1991): Resource-use efficiency in public schools: A study of Connecticut data. *Management Science* 37(12), 1620-1628.

Rosenzweig, M. L. (1995): Species Diversity in Space and Time. Cambridge University Press, New York, NY.

Shannon, C.E. and W. Weaver (1949): *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois.

Simar, L. and P. W. Wilson (2007): Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics* 136(1), 31-64.

Appendix: Diversity effects on productive efficiency of farms: preliminary results

Data

In the preliminary analysis, we utilized the empirical data of 379 Finnish farms located all over the country with the total utilized agricultural area (UAA) more than 50 hectares (ha). The data set is obtained from FADN database for the accounting year 2004. The economic output is the total revenue from crops and crop products, livestock and livestock products and other output (SE131), expressed in Euros (€). Input resources include UAA in ha (SE025), labor in hours (hr) (SE011), and farm capital in € (SE510). An overview of key characteristics of the data is presented in Table I in the form of mean, standard deviation, minimum and maximum values.

Table I

Descriptive statistics: year 2004, sample size n = 379.

Variable	Mean	St. Dev.	Min	Max
Total output, €	107,341	91,993	4,847	685,301
Labor, hr	4,249	2,376	200	15,105
Farm capital, €	349,166	218,702	42,929	1,770,813
UAA, ha	83.1	34.1	50.1	324.3

CNLS results

The results obtained by the CNLS estimator for contextual variables are presented in Table II.

Table II

CNLS results: $R^2 = 0.0094$.

	Coefficients	*Std Err.*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-4.1E-07	0.062	-6.6E-06	0.999	-0.122	0.122
Shannon index	**0.188**	0.099	1.890	**0.059**	-0.008	0.383

The resulted coefficient of the Shannon index is positive $\beta = 0.188$ and almost significant: p-value is about 0.06.

Table III includes the results obtained with the CNLS estimator including dummy variables. However, in this application dummy variables are constructed for the farms with different land-use categories (8 land-use categories), but not for agricultural specialization.

Table III

CNLS results with dummy variables: $R^2 = 0.994$.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.480	0.087	17.058	1.14E-48	1.309	1.650
Shannon index	-4.797	0.078	-61.579	7.5E-198	-4.950	-4.643
s2	11.146	0.071	156.379	0	11.006	11.286
s3	11.054	0.064	171.583	0	10.927	11.181
s4	3.029	0.041	74.039	1.7E-225	2.949	3.110
s5	2.948	0.062	47.354	4.6E-160	2.826	3.070
s6	3.656	0.079	46.470	1.9E-157	3.501	3.810
s7	3.339	0.081	41.158	6E-141	3.180	3.499
s8	-10.099	0.141	-71.769	8.8E-221	-10.376	-9.822