Chapter 15: Microfinance in Vietnam

The Efficiency of Microfinance in Vietnam: Evidence from NGO Programs

Hong Son Nghiem, Tim Coelli, and Prasada Rao

Abstract: A large amount of donor money and government money is spent on microfinance programs in developing countries around the world. However, there is very little quantitative research available on the relative efficiency of these programs. This research investigates the efficiency of the microfinance industry in Vietnam through a survey of 46 schemes in the north and the central regions. Data Envelopment Analysis (DEA) methods are used to assess the technical efficiency and scale efficiency of the microfinance schemes. Given the lack of previous studies in this industry, we review the various approaches to variable selection used in the financial institutions literature and amend the so-called “production” approach to accommodate the poverty reduction focus of microfinance. The empirical results reveal that the average technical efficiency scores of schemes surveyed is 80%. A second stage regression analysis is used to assess the impact of a variety of environmental variables upon the efficiency of the schemes. The age and the location of the scheme are found to have a significant influence upon efficiency.

Keywords: Microfinance, Efficiency, NGOs, Poverty Reduction

Introduction

Improving efficiency is one of the major challenges of the microfinance industry but little attention has been paid to efficiency assessment (Morduch, 1999). The
efficiency of microfinance schemes is an important topic because the degree to which a dollar spent on microfinance programs converts into poverty alleviation will be in part a function of the degree to which these programs operate in an efficient manner. A commonly used approach to the measurement of microfinance efficiency involves the use of financial ratios, which were originally designed for conventional financial institutions. Accounting ratios provide only a partial measure of efficiency, and are unable to take into account issues such as economies of scope and scale. Therefore, this study takes a different approach to efficiency measurement, utilising production frontier, which can provide more useful managerial information than available from traditional methods.

After the financial reforms introduced in the early 1990s the Vietnamese microfinance industry developed rapidly, although 70 per cent of the poor in the country still do not have access to reliable financial services (GSO, 2005). Improving the efficiency of microfinance institutions (MFIs) is one possible way in which to increase the outreach of this sector, as improvements in efficiency will mean that more clients can be served from the available resources. However, there has been no previous research into the efficiency of the Vietnamese microfinance industry. This study seeks to fill the gap in the research by analysing the efficiency of 44 microfinance schemes in the north and central regions of Vietnam. Determinants of efficiency in these programs are also examined using second stage regressions, where efficiency scores are regressed against environmental variables such as infrastructure, locations, and the maturity levels of NGO microfinance programs (NMPs).

The remainder of this chapter is organised into five sections. Section 2 provides a discussion of efficiency measurement methods. Section 3 reviews several previous efficiency studies in microfinance. The data and results of the efficiency analysis are then presented in Section 4, with some concluding remarks made in Section 5.

Methodologies

This section describes the two main approaches to the measurement of efficiency, namely parametric and non-parametric approaches.

The non-parametric approach

The non-parametric approach uses mathematical programming techniques to estimate “best practice” frontiers as the “real” frontier is typically unknown. The Data Envelopment Analysis (DEA) technique, popularised by Charnes et al. (1978), has been the most commonly used non-parametric method in empirical studies. DEA involves the calculation of efficiency by constructing a piecewise frontier surface (which represents efficient operations) using linear programming applied to the input-output vectors of a sample of firms. Efficiency is then measured as the distance that each inefficient firm lies below this frontier. DEA efficiency scores are normally measured using an input-oriented approach (i.e., reducing inputs while maintaining a particular level of outputs), or an output-
oriented approach (i.e., expanding outputs while using the same level of inputs) or non-orientation (i.e., both output expansion and input reduction).

This section briefly describes the DEA technique, as presented in Coelli et al. (2005). We use the notation $X$ to represent the $K \times N$ matrix of inputs, consisting of $K=\{1,2,\ldots,k\}$ inputs from $N=\{1,2,\ldots,n\}$ firms; $Y$ denotes the $M \times N$ matrix of outputs, consisting of $M=\{1,2,\ldots,m\}$ outputs from $N=\{1,2,\ldots,n\}$ firms. The input-oriented constant returns to scale (CRS) DEA frontier is defined by solutions to $N$ linear programs:

$$
\begin{align*}
\min_{\theta, \lambda} & \quad \theta \\
\text{subject to:} & \quad -y_i + Y \lambda \geq 0, \\
& \quad x_i \theta - X \lambda \geq 0, \\
& \quad \lambda \geq 0,
\end{align*}
$$

where $\lambda$ is a $N \times 1$ vector of weights, $0 \leq \theta \leq 1$ represents the technical efficiency (TE) score, and $1-\theta$ is the proportional reduction in inputs that could be achieved by the $i$-th firm to produce the given level of output.

Similarly, the output-oriented CRS DEA frontier is defined as follows:

$$
\begin{align*}
\max_{\phi, \lambda} & \quad \phi \\
\text{subject to:} & \quad -y_i \phi + Y \lambda \geq 0, \\
& \quad x_i - X \lambda \geq 0, \\
& \quad \lambda \geq 0,
\end{align*}
$$

where $0 \leq \phi \leq \infty$, $1/\phi$ represents the TE score, and $\phi-1$ is the proportional expansion in outputs that could be produced using the given level of inputs. Other parameters are defined as previously.

The CRS frontier assumes that all firms operate at their most productive scale (Coelli et al., 2005). This assumption is relaxed in the variable returns to scale (VRS) model by adding the convexity constraint $N1' \lambda = 1$ (where $N1$ is the $N \times 1$ vector of ones) to the above linear programming problems. Therefore, by comparing TE scores under the CRS and VRS frontiers, one can decompose the TE under the CRS frontier into TE$_{VRS}$ efficiency (i.e., TE$_{CRS} \cdot$ Scale efficiency). In this setting, TE$_{CRS}$, TE$_{VRS}$ and scale efficiency vary between zero and one.

In order to determine whether the firm operates at increasing or decreasing returns to scale, the non-increasing return to scale (NIRS) frontier is constructed by adding the constraint $N1' \lambda \leq 1$ to the CRS linear programming problem. This constraint ensures a firm is not benchmarked against smaller firms, but can be benchmarked against substantially larger ones. If the TE score of a firm under a VRS frontier does not equal the NIRS TE score, it operates under increasing returns to scale (IRS), while the firm operates at decreasing returns to scale (DRS) if its TE scores under the two frontiers are equal.
One advantage of the DEA technique is that it can provide information about peers of each inefficient firm (peers are those firms with similar input and output mixes to the inefficient firm, but located on the frontier). This peer information is very useful for managerial purposes as managers can improve the efficiency of their institutions by learning from their efficient peers. The main drawback of DEA (and other non-parametric techniques) is that it assumes that there is no random noise in the data.

The parametric approach

a) Single output production functions

The parametric approach uses econometric techniques to construct the production frontier. One generally needs to specify distributional forms for the inefficiency error term and a functional form for the parametric production frontier. The inefficiency error component is often assumed to have a half-normal distribution, but this distribution can be criticised on the basis that it implicitly assumes that the modal outcome is the full efficiency. Other more general distributional forms, such as a gamma distribution (Yuengert, 1993) or a truncated normal distribution (Berger and DeYoung, 1997) have been tried, but revealed little difference compared to the results obtained from the half-normal distribution.

Parametric frontier productions were first used in efficiency measurement by Aigner and Chu (1968) with a deterministic specification as presented in equation (3), where y_i denotes output, x_{ki} denotes the k-th input, and u_i represents the inefficiency component of the i-th firm (i=1,2,...,N).

\[ \ln y_i = \alpha + \sum_{k=1}^{K} \alpha_k \ln x_{ki} - u_i \] (3)

The technical efficiency of an i-th firm is defined as the ratio of the observed output to its potential output.

\[ TE_i = \frac{y_i}{\exp(x_i \beta)} \] (4)

where the potential output level, \( \exp(x_i \beta) \), is estimated by solving the problem of minimising \( \sum_{i=1}^{N} u_i \), subject to the constraint \( u_i \geq 0 \) (i=1,2,...,N). This model was also estimated by other techniques such as maximum likelihood (Afriat, 1972) and corrected least squares (COLS) (Richmond, 1974), which is basically the ordinary least squares (OLS) estimate adjusted up so that estimated frontiers envelop all observations. However, the model of Aigner and Chu (1968) did not take into account noise and measurement errors. This motivated the development of the stochastic frontier analysis (SFA) technique, which has the ability to include random errors. A SFA may be defined as:

\[ \ln y_i = \alpha + \sum_{k=1}^{K} \alpha_k \ln x_{ki} + v_i - u_i \] (5)
where $y_i$ and $x_{ki}$ are as defined above; $u_i$ represents the strictly non-negative vector, representing the inefficiency component with a pre-assumed distributional form, such as half normal, truncated normal or exponential; and $v_i$ is a random error, which is generally assumed to be normally distributed.

The technical efficiency of the $i$-th firm can be estimated using the following conditional expectation:

$$TE_i = E[\exp(-u_i) | (v_i - u_i)]$$  \hspace{1cm} (6)

The two common functional forms used in SFA studies are the Cobb-Douglas and translog, which is quadratic in logs. The Cobb-Douglas functional form is generally less demanding to both estimate and interpret. In addition, it requires fewer parameters, which is useful for applications with a modest number of observations. The main disadvantage of the Cobb-Douglas functional form is that it assumed implicitly that all firms have the same production elasticities and that the elasticity of substitution between inputs equals one. The translog functional form is less restrictive on production and substitution elasticities but it is more difficult to interpret and involves the estimation of many parameters, which make it inconvenient for applications with few observations (Coelli and Perelman, 1999).

Attempts to weaken the distributional assumptions in the SFA approach were proposed in the distribution free analysis (DFA) method (Berger, 1993) and thick frontier analysis (TFA) method (Berg and Kim, 1994). The DFA method makes no assumption on the distributional form of either the error term or the inefficiency component. However, it requires access to panel data and assumes that the error component has a zero mean and that the inefficiency component is stable over time. The efficiency level of each firm is estimated as the difference between its average residual and that of the firm on the frontier. One disadvantage of the DFA method is that when efficiency is not stable (e.g., shifting over time due to technical changes or other factors), the difference between average residuals of any firm on the frontier does not represent the efficiency of that firm at any one point in time.

The TFA method also assumes no distributional form of the random error and the inefficiency component. Instead, it assumes that randomness is represented by deviation “within” the lowest and highest quartile performance values, whilst the inefficiency component is represented by the difference “between” these quartiles. With this method, best-practice firms do not necessarily lie on the frontier but are close to the frontier. Therefore, the TFA method reduces the effects of outliers but it provides only estimates of the overall efficiency instead of point estimates of efficiency for individual firms.

Although the DFA and TFA approaches do not require assumptions about distributions for the inefficient and noise components, they require new assumptions such as that the efficiency level of a firm remained constant overtime (DFA), or that the efficiency level varies within the first and the third quartiles of the residuals (TFA), both of which may be arguable. In addition, the DFA and TFA approaches require the availability of panel data, which is very rare in the world microfinance industry; no such data exist in Vietnam, to the best of our knowledge; therefore these techniques are beyond the scope of this study.
b) Multiple output industries and distance functions

The estimation of production frontiers using parametric approaches becomes more complex in multiple output industries, especially when the aggregation of outputs is not possible. For example, the main outputs of the microfinance industry are number of clients and loans volumes, which are both important and are difficult to aggregate. One can measure the efficiency of multiple output firms using cost or profit functions, where efficiency is measured as the ratio of observed costs and minimised costs (if the cost function is used), or the ratio of observed profits over maximised profits (if the profit function is used). However, the estimation of cost and profit functions requires assumptions of cost minimisation or profit maximisation, which may not be relevant behaviour for public and not-for-profit services such as microfinance. Fortunately, this multi-output issue can be solved using the distance function concept, pioneered by Shephard (1953). Distance functions can be specified with an input-oriented or output-oriented approach. Input distance functions, given the output vector, focus on minimal proportional contraction of the input vector, while output distance functions, given an input vector, seek a maximal proportional expansion of the output vector.

To define the distance functions one must first define the production technology. Based on the notations described in Coelli et al. (2005), we define $x \in \mathbb{R}^k$ as a vector of inputs, $y \in \mathbb{R}^m$ as a vector of output, and the production technology $T$ as:

$$T = \{ (x, y) : x \text{ can produce } y \}$$

(7)

Some assumptions are made with the production technology $T$ as follows:

- $T$ is closed and convex;
- Inputs and outputs are strongly disposable: if $(x, y) \in T$ and $(x^*, -y^*) \geq (x, -y)$, then $(x^*, y^*) \in T$;
- Inaction is possible: $(x, 0) \in T$;
- There is no free lunch: $(0, y : y > 0) \not\in T$; and
- Unlimited quantity of output cannot be produced: $P(x) = \{ y : (x, y) \in T \}$ is bounded from above.

Using these notations, input and output distance functions, respectively, are:²

---

² An alternative assumption is “weak disposability” of inputs and outputs, in which “$\geq$” is used instead of “$>$”. Since $x$, $x'$, $y$ and $y'$ are vectors, the notation “$\geq$” between two vectors means that all elements on one vector are greater than or equal to the corresponding elements of the other vector with at least one strict inequality.
Some properties of distance functions follow directly from assumptions of
the production technology $T$:  
1) $D_i(x, y) \geq 1 \iff (x, y) \in T$; and it is linearly homogeneous in $x$; 
2) $D_o(x, y) \leq 1 \iff (x, y) \in T$; and it is linearly homogeneous in $y$; 
3) Distance functions are non-decreasing in $x$ and $y$; and 
4) Distance functions equal one if the firm is located on the efficient frontier.

Duality theorems presented in Diewert (1992) and Chambers et al. (1998) show that, under certain conditions, output distance functions are equivalent to the cost or profit functions, which are commonly used in efficiency studies in the banking sector. The advantage of distance functions is that they do not need price information, and hence, avoid possible issues of price endogeneity or a lack of price information.

The concept of distance functions was further developed into various branches such as the gauge function proposed by McFadden (1978) and directional distance functions, proposed by Chambers et al. (1996; 1998). The McFadden gauge function measures the largest radial expansion of a netput vector, which can be specified as $H(-x, y) = \inf_{\theta} \{\theta > 0 : (-x, y, \theta) \in T\}$. Directional distance functions measure the amount that one can translate the input and output vectors from the current position to the technology frontier in pre-assigned directions $(g^x \in \mathbb{R}^M, g^y \in \mathbb{R}^K)$, and can be defined as $\tilde{D}(x, y, g^x, g^y) = \sup_{\theta} \{\theta : (x - \theta g^x, y + \theta g^y) \in T\}$. However, this study will focus on discussing Shephard distance functions since the netput expansion concept presented in gauge functions, which is similar to the profit efficiency concept, may not apply in the microfinance sector. The pre-assigned direction of inputs and outputs in this industry is also difficult due to its multiple objective nature (i.e., the choice of the direction may vary according to different objectives). In addition, by using non-radial contraction of inputs or expansion of outputs, directional distance functions do not have the ability to decompose the overall efficiency into useful managerial concepts of technical and allocative efficiency.

---

$^{2}$ “Sup” stands for “Supremum” to allow the case when a maximum does not exist (i.e., $\delta=\infty$) and “Inf” stands for “Infimum” to allow the case when a minimum does not exist (i.e., $\theta=\infty$).
As mentioned previously, there are two functional forms commonly used to represent productions: Cobb-Douglas and translog. A translog input distance function for the case of K inputs and M outputs is specified as:

\[
\ln D_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m \ln y_{mi} + 0.5 \sum_{m=1}^{M} \sum_{n=1}^{M} \alpha_{mn} \ln y_{mi} \ln y_{ni} + \sum_{k=1}^{K} \beta_k \ln x_{ki} + 0.5 \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_{ki} \ln x_{li} + \sum_{k=1}^{K} \sum_{m=1}^{M} \delta_{km} \ln x_{ki} \ln y_{mi} \\
i = 1, 2, ..., N;
\]

where \( i \) denotes the \( i \)-th firm in the sample and \( D_i \) represents an input distance.

The required restrictions for homogeneity of degree +1 in inputs are:

\[
\sum_{k=1}^{K} \beta_k = 1
\]

\[
\sum_{k=1}^{K} \beta_{kl} = 0, \quad k = 1, 2, ..., K
\]

\[
\sum_{k=1}^{K} \delta_{km} = 0, \quad m = 1, 2, ..., M
\]

and the required restrictions for symmetry are

\[
\alpha_{mn} = \alpha_{nm}, \quad m, n = 1, 2, ..., M
\]

\[
\beta_{kl} = \beta_{lk}, \quad k, l = 1, 2, ..., K
\]

Distance functions can be used in efficiency measurement under parametric (e.g., SFA), non-parametric (e.g., DEA) and combined approaches (Coelli and Perelman, 1999). Some of these techniques are presented below (we focus on presenting the input-oriented approach because the output oriented approach is conducted analogously).

b1) Stochastic Frontier Analysis (SFA)

The application of the SFA technique to distance functions is conducted by introducing a random error component to the distance functions as in equation (10). In addition, the homogeneity of degree one in an input distance function specified in equation (11) is used to transform it to a relevant format that can be estimated using the SFA method. Recall that a function of homogeneity of degree \( \omega \) is presented as:

\[
D(\omega x, y) = \omega D(x, y) \quad \text{for all } \omega > 0
\]

Thus, if we choose one input arbitrarily (e.g., the \( k \)-th) and replace \( \omega = 1/x_K \), this will result in:

\[
D(x/x_K, y) = D_i(x, y)/x_K
\]

Representing the right hand side of the translog input distance function as represented in equation (10) by a brief format of \( TL(x, y) \), we have:
\[
\ln\left(\frac{D_i}{x_{ki}}\right) = TL\left(\frac{x_i}{x_{ki}}, y_i\right)
\]

or

\[
-ln(x_{ki}) = TL\left(\frac{x_i}{x_{ki}}, y_i\right) - \ln(D_i)
\]

With the introduction of the random error \(\varepsilon_i\), the transformed translog input-oriented distance function in the final form of equation (15) becomes:

\[
-ln(x_{ki}) = TL\left(\frac{x_i}{x_{ki}}, y_i\right) - \ln(D_i) + \varepsilon_i
\]

It is shown that equation (16) can be estimated by the SFA method with the composite error term including a non-negative component \(-\ln(D_i)\), representing efficiency, and a random error \(\varepsilon_i\).

b2) Corrected Ordinary Least Squared (COLS)

The COLS method was proposed by Greene (1980), while its application to distance functions was introduced later by Lovell et al. (1994) and Grosskopf et al. (1997). This method applied OLS to the transformed distance function using the homogeneity of degree one characteristics as presented in equation (15). In addition, the largest positive residual is added to the intercept, making the frontier envelops all the data points. The inefficiency of each firm is calculated as the distance between observed data and its radial expansion to the frontier.

b3) Parametric linear programming (PLP)

The PLP technique was proposed by Aigner and Chu (1968) with single-output deterministic Cobb-Douglas production functions. The technique was expanded to translog output distance functions by Fare et al. (1993). Coelli and Perelman (1999) argued that the translog was less restrictive on elasticity of substitution and scale properties, and hence preferable to the Cobb-Douglas functional form. Applying the PLP technique to an input-oriented translog distance function specified in equation (10) requires the solving of the following linear programming problem:

\[
\min \sum_{i=1}^{N} \ln D_i
\]

Subject to:

\[
\ln(D_i) \geq 0, \quad i = 1, 2, \ldots, N
\]

plus the homogeneity and the symmetry constraints in equations (11) and (12).

In summary, there are several methods for efficiency analysis using the production frontier approach. A number of the above-mentioned methods have been applied widely in the banking industry while only a few studies were conducted in the microfinance sectors. More details of efficiency studies in financial institutions are presented in the next section.
A review of efficiency studies of financial institutions

Microfinance institutions offer services (e.g., credit, savings and insurance) similar to those of other types of financial institutions such as commercial banks and credit unions. However, there is a sharp contrast in the number of efficiency studies in microfinance versus other types of financial institutions: for example, while the amount of research on financial institutions is very large, there are only a handful of studies on microfinance efficiency. Therefore, it is useful to review efficiency measurement studies on commercial banks and credit unions in addition to studies on microfinance institutions, with the view to applying relevant lessons to the microfinance sector.

Commercial banks

A comprehensive review of efficiency measurement studies in financial institutions provided by Berger and Humphrey (1997) found 130 banking studies conducted mainly in the 1990s. The authors also revealed relatively equal proportions of non-parametric and parametric methods used, dominated by DEA and SFA techniques, respectively. In addition, the authors found that non-parametric methods have lower average efficiency scores than those of parametric estimates. Random noise in data may contribute to this difference. The relationship between results obtained from parametric and non-parametric techniques revealed generally weak correlation, although the correlation is much stronger when comparing results obtained from the different techniques within each category.

One of the earliest studies of efficiency measurement in the banking sector is that of Sherman and Gold (1985), who applied the DEA method to measure the efficiency of 14 bank branches. The authors emphasised the advantage of DEA compared with the traditional accounting standard in measuring efficiency of multiple-input, multiple-output industries. They described two approaches in classifying outputs in the banking sectors: production and intermediation. The production approach considers financial institutions as the producers of services to their clients. The intermediation approach argues that financial institutions play the role of intermediating funds between savers and borrowers. The authors argued that the production approach provides more useful information to the operations of banks in their study.

Within the above two approaches in banking efficiency studies, one of the most controversial issues is the classification of deposits, which arguably has the characteristics of both inputs and outputs. For example, deposits can be classified as an output if the production approach is used (i.e., deposits are the result of saving services) but it is considered an input if the intermediation approach is selected (i.e., deposits are part of input funds). To overcome this issue, Hunter and

---

3 Among 69 non-parametric studies surveyed, 62 applied DEA, five applied free disposable hull (FDH), one used the Index Numbers (IN) method, and one study used the Mixed Optimal Strategy (MOS) method. The 60 parametric studies included 24 SFA, 20 DFA and 16 TFA.
Timme (1995) and Berger et al. (1997) analysed two separate models and compared their respective efficiency measures. They found a strong correlation between the efficiency scores evaluated by the two approaches. However, the efficiency scores were higher when deposits were classified as an output, which is not surprising because, ceteris paribus, the production approach will result in a larger amount of outputs and less input quantity (i.e., deposits are shifted from the input side to the output side when moving from the intermediation to the production approach). In addition, the banking sector may have the dual roles of both producers of services and intermediators between savers and borrowers. Therefore, some studies such as Berger and Humphrey (1991) and Bauer et al. (1993) included both characteristics of deposits (i.e., interest paid on deposits are included in inputs while deposit quantity is included in outputs). One arguable point in this choice is that it may take similar amounts of labour to process deposits of different sizes: and therefore the number of deposits accounts may be a better proxy for deposit outputs.

Since the mid 1980s, the number of efficiency studies in the banking sector increased rapidly, making it one of the most popular areas in efficiency measurement research (Tavares, 2002). One point to keep in mind when reviewing efficiency studies is that regional or cross-country comparisons may produce misleading information unless a common frontier is constructed. For example, results from studies reviewed by Berger and Humphrey (1997) showed that the average efficiency scores of commercial banks in the United States (US) and other countries were 79 and 75 per cent, respectively. However, this does not necessarily mean that US banks were more efficient since the average figure for other countries included country-specific studies (i.e., using one frontier for each country). Using a common frontier, Fecher and Pestieau (1993) and Pastor et al. (1997) found that US banks were least efficient among the respective set of 11 OECD and 8 developed countries. This is a useful lesson for our study, that a common frontier should be used to compare the efficiency levels of microfinance programs in different regions.

Another useful lesson learnt from banking studies was that it is necessary to check for the robustness of results among different methods and input/output combinations. In addition, it is necessary to use flexible functional forms in the parametric approach and integration of random noise in the non-parametric approach to mitigate their respective limitations. Banking studies that used flexible functional forms include Berger and DeYoung (1997), and Berger and Mester (1997); those employing bootstrap techniques include Ferrier and Hirschberg (1997) and Simar and Wilson (1998).

Credit unions

Credit unions can be considered a transition between microfinance and commercial financial institutions. Credit unions provide similar financial services as commercial banks but as their roles focus on intermediating between savers and borrowers among members, profit maximisation assumption may not be relevant (Smith et al., 1981). Compared to credit unions, microfinance institutions
often have a smaller scale and focus more on rural areas. However, lessons learnt from credit union studies could be useful for microfinance applications.

One of the earliest production frontier studies on credit unions was that of Fried et al. (1993; 1996), who applied the FDH method to examine the efficiency of credit unions in the US. The authors argued that credit unions aim to maximise services given their resource constraints, and hence, the production approach to output variable selection was arguably more relevant. They found that credit unions were relatively efficient in providing financial services to their clients, with average scores ranging from 80 to 83 per cent. The relatively high efficiency scores estimated by the FDH were not surprising as this method enveloped the data closer, compared to DEA frontiers. However, some authors are sceptical about the choice of the FDH method in their analysis because there is no strong economic ground to relax the convexity assumption (Thrall, 1999).

In contrast to Fried et al. (1995) selected the intermediation approach for credit unions, arguing that the market value of outputs may not be higher than that of inputs, and thus the notion of “production” may not be relevant for this industry. They applied a translog cost function to measure the economies of scale of credit unions in Australia. The study found significant evidence of diseconomies of scale for small credit unions (ranked by asset sizes) but no such evidence was found in medium and large credit unions. One possible reason for the diseconomies of scale issue could be the subsidy from host companies in the form of provision of office space and arrangements for payroll deductions. However, the use of cost functions may be questionable because cost-minimising assumptions may conflict with the multiple-objective nature of credit unions.

A recent study conducted by Paxton (2006) considered the dual roles of credit unions as producers of services and intermediators between savers and producers. She applied the DEA technique to 350 non-bank financial institutions in rural Mexico, most of which were credit unions. The author found that the efficiency of credit unions was sensitive to choice of approach. On average, the technical efficiency of credit unions was 58 and 14 per cent under the intermediation and production approaches, respectively. This finding was in contrast to our expectation that the production approach would generate higher efficiency scores, ceteris paribus. The possible reason was that her choice of inputs and outputs were not comparable (the intermediation approach involved 3 inputs and 3 outputs while the production approach included only 2 inputs and 2 outputs). Mathematically, the intermediation approach produced higher efficiency scores due to its relatively larger input-output dimension in linear programming. It is also possible that the low efficiency scores in this study were due to the use of a single frontier for different types of financial institutions. Because of issues such as information asymmetry, efficient financial institutions may not wish to provide their services to some segments of the population, leaving room for those having an informational advantage to sell their “inefficient” services; therefore using a separate frontier for each type of financial institution may result in higher efficiency scores.
Microfinance

The measurement of efficiency using production frontiers is still quite rare in microfinance. To date, there has been only a handful of efficiency studies in the microfinance sector, including a study of the Grameen Bank by Hassan and Tufte (2001) and two international studies by Gutierrez-Nieto et al. (2006; 2007).

Hassan and Tufte (2001) applied a Cobb-Douglas stochastic cost function to estimate the cost efficiency of the Grameen Bank and then used Tobit regressions in the second stage to identify determinants of the efficiency. They used the production approach, which defines microfinance as a process that uses standard inputs such as labour and capital to produce outputs such as number of loans and number of clients. In particular, the three output variables used in their study were the numbers of loans, savings volumes, and the number of members, and the two input variables were labour and capital. The study found that most of the 186 branches of the Grameen Banks in the 1989–1991 period were highly efficient, with average technical efficiency scores varying from 94 to 97 per cent. Determinants of cost efficiency were investigated using second-stage regressions with regressors including infrastructure (electricity, road, and density of banks and schools) and characteristics of MFIs (age of institutions, asset values, and sex of target clientele). The study found that female-only branches were more efficient than male-only and mixed branches. As one may expect, branches located in areas with good infrastructure are more efficient than branches located in remote areas. Education for members significantly affects the efficiency of a branch because it allows members to have more control over their own transactions. The age and size of a branch have insignificant influences on its cost efficiency. The cost minimisation assumption used in Hassan and Tuft (2001) may not reflect the multiple objectives of microfinance, especially the social objective of serving the poor. As forcefully argued by Pestieau and Tulkens (1994), assessing and explaining the performance of multiple-objective institutions is more complex, and production frontier techniques are generally a more adequate choice for use in measuring the performance of such organisations.

Gutierrez-Nieto et al. (2007) used DEA in combination with the principal component analysis (PCA) technique to measure the efficiency of 124 MFIs from around the world. The production approach was selected as the authors argued that deposit service, the main component of the intermediation approach, plays minor roles in MFIs studied. Their DEA model included two inputs, namely labour and operating expenses; and three outputs, namely interest rate income, loan volume and the number of loans. They also conducted DEA models with all combinations of inputs and outputs to check the robustness of the main results. The determinants of the efficiency of microfinance institutions were analysed using PCA with all DEA models. The results revealed that location and type of MFIs are the main components affecting the efficiency levels of MFIs. An arguable point in the study by Gutierrez-Nieto et al. (2007) is their use of all possible combinations of the selected two inputs and three outputs, in which many combinations were formed by just one input and one output; thus the use of
the DEA method will provide limited additional benefit relative to accounting ratios, other than adjusting for scale.

Gutierrez-Nieto et al. (2006) expanded their data to 430 MFIs and used separate models for financial and social objectives. The two models shared three inputs (total assets, operating costs, and number of employees) but used a different set of outputs. The two outputs for the social objective model included the number of women borrowers and the number of poor clients, while the two outputs for the financial objective model included the total loan volume and the financial revenue. Although the authors revealed that most MFIs were efficient in both aspects, the use of separate models for each objective weakens the main advantage of DEA in benchmarking multiple-objective production units.

Gutierrez-Nieto et al. (2006) also tried two specifications using one output (either number of women clients, or number of poor clients) and the same set of three inputs. They found a high correlation between efficiency scores of the two models, which suggests that MFIs that efficiently serve women also have positive contribution to poverty reduction. The authors found no clear relationship between the age of MFIs and their social efficiency level, which is against the expectation of the conceptual model proposed in Chapter 3 of this study that mature MFIs could be more efficient in serving the poor.

In summary, efficiency studies using the production frontier approach were widely applied in the banking sector whilst such studies in credit unions and microfinance limited. An important factor in efficiency measurement of financial institutions was the choice of production or intermediation approaches, which was often made on a case-by-case basis. The characteristic of microfinance is to deliver financial services on a small scale to the poor rather than maximising financial transactions between all savers and borrowers; hence the production approach is arguably more relevant than the intermediation approach. In addition, methods involving behavioural assumptions such as profit maximisation may not be relevant to the not-for-profit nature of credit unions or microfinance programs.

Empirical Analysis

The descriptive statistics of the main variables from the institutional survey, presented in Table #1, show that the operational scale varies greatly among NMPs surveyed due to a large variation in key variables, such as number of members, and loans and saving volumes. The loan interest rates and saving interest rates do not vary greatly across schemes. On average, the loan interest rate is 1.28 per cent per month, which is higher than the interest rate charged by the VBP at 0.6 per cent per month. However, compared with the interest rate charged by local moneylenders (which varies from two to five per cent per month), the interest rates charged by NMPs were reasonable.

Table #1: Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unit</th>
<th>Median</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups</td>
<td>Numbers</td>
<td>89</td>
<td>381.84</td>
<td>710.52</td>
<td>5</td>
<td>3900</td>
</tr>
<tr>
<td>Members</td>
<td>Persons</td>
<td>776</td>
<td>2413.77</td>
<td>3707.72</td>
<td>68</td>
<td>19508</td>
</tr>
<tr>
<td>Borrowers</td>
<td>Persons</td>
<td>681</td>
<td>2259.11</td>
<td>3663.22</td>
<td>48</td>
<td>19608</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
<td>----</td>
<td>-------</td>
</tr>
<tr>
<td>Savers</td>
<td>Persons</td>
<td>660</td>
<td>2337.77</td>
<td>3732.83</td>
<td>18</td>
<td>19508</td>
</tr>
<tr>
<td>Loan interest rate</td>
<td>% per month</td>
<td>1.2</td>
<td>1.28</td>
<td>0.22</td>
<td>0.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Deposit interest rate</td>
<td>% per month</td>
<td>0.6</td>
<td>0.61</td>
<td>0.13</td>
<td>0.4</td>
<td>0.85</td>
</tr>
<tr>
<td>Outstanding loans</td>
<td>VND’000</td>
<td>673000</td>
<td>2.06E+06</td>
<td>3.43E+06</td>
<td>31000</td>
<td>2.00E+07</td>
</tr>
<tr>
<td>Saving volume</td>
<td>VND’000</td>
<td>48697.5</td>
<td>445782.11</td>
<td>1.05E+06</td>
<td>2225</td>
<td>6.35E+06</td>
</tr>
<tr>
<td>Income</td>
<td>VND’000</td>
<td>108720</td>
<td>335296.29</td>
<td>596547.02</td>
<td>5580</td>
<td>3.60E+06</td>
</tr>
<tr>
<td>Interest cost</td>
<td>VND’000</td>
<td>2947.39</td>
<td>36455.66</td>
<td>92120.54</td>
<td>172.6</td>
<td>571206</td>
</tr>
<tr>
<td>Other financial cost</td>
<td>VND’000</td>
<td>9540</td>
<td>10368.68</td>
<td>6503.92</td>
<td>382</td>
<td>35000</td>
</tr>
<tr>
<td>Wages</td>
<td>VND’000</td>
<td>127200</td>
<td>151658.2</td>
<td>106028.78</td>
<td>40800</td>
<td>564000</td>
</tr>
<tr>
<td>Other operating cost</td>
<td>VND’000</td>
<td>14550</td>
<td>10528.74</td>
<td>6408.92</td>
<td>130</td>
<td>22846</td>
</tr>
<tr>
<td>Head quarter staff</td>
<td>Persons</td>
<td>3</td>
<td>2.82</td>
<td>1.69</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Local staff</td>
<td>Persons</td>
<td>8</td>
<td>20.91</td>
<td>37.75</td>
<td>2</td>
<td>220</td>
</tr>
<tr>
<td>Year in operations</td>
<td>Years</td>
<td>7</td>
<td>6.61</td>
<td>2.92</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Poorest areas</td>
<td>1=yes, 0=no</td>
<td>0.5</td>
<td>0.5</td>
<td>0.51</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Car accessibility</td>
<td>1=yes, 0=no</td>
<td>1</td>
<td>0.89</td>
<td>0.32</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Electricity</td>
<td>1=yes, 0=no</td>
<td>1</td>
<td>0.82</td>
<td>0.39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Distance to township</td>
<td>Km</td>
<td>5</td>
<td>6.33</td>
<td>4.28</td>
<td>0</td>
<td>16.75</td>
</tr>
<tr>
<td>Northern regions</td>
<td>1=yes, 0=no</td>
<td>1</td>
<td>0.61</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ethnic minority</td>
<td>1=yes, 0=no</td>
<td>0.00</td>
<td>0.39</td>
<td>0.40</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Sustainability objectives</td>
<td>1=yes, 0=no</td>
<td>0.00</td>
<td>0.32</td>
<td>0.47</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Max loan size</td>
<td>VND’000</td>
<td>2000</td>
<td>2819</td>
<td>3042</td>
<td>500</td>
<td>15000</td>
</tr>
<tr>
<td>Repeating clients</td>
<td>Per cent</td>
<td>95.00</td>
<td>75.10</td>
<td>37.19</td>
<td>0.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Note: Unless otherwise specified, all results in this chapter are own calculations from the survey data.

Cost and income information shows that wage costs are almost three quarters of total costs. Interest costs contribute only a minor proportion of the total costs.
because most loan funds are supplied by donors with no interest. Information on staff numbers is provided; however, this information is questionable in many cases because some microfinance workers also have another full-time job and hence are unlikely to work full time on microfinance activities. For example, most local staff of NMPs are also staff of the Women's Union; therefore, they only receive bonuses from NMPs as additional income from their main jobs. The variation in the number of groups reflects the fact that NMPs operating in remote regions tend to have smaller group sizes than those operating in densely populated areas.

Input-output choice

As mentioned previously there are two common approaches used in efficiency measurement of financial institutions, namely, the production approach and the intermediation approach. We follow the production approach, essentially because it is a more appropriate choice to make in the case of microfinance as the focus is more on providing a service to a large number of poor people, rather than trying to negotiate large loans with rich clients. In addition, the production approach considers deposits as an output since deposits are a product of the savings mobilisation activities (in the intermediation approach it is considered as an input since deposits are part of the loanable fund). Also, the intermediation approach considered volume of transactions (e.g., volumes of savings and loans) as outputs, which may create a downward bias for the efficiency of some NMPs that focus on serving smaller transactions. Since it takes a similar amount of time and resources to process small transactions as to process larger transactions, those serving small transactions look less efficient if the volume of transaction is selected as outputs.

Despite considerable time and effort, the dataset from the institutional survey still suffers from incomplete data on several important variables, such as subsidies received and depreciation on fixed assets. Some financial inputs, such as bad debts, contained many zero values, as NMPs have no bad debt and these zero-value input variables can cause computation problems in the DEA method. In addition, there is a very high correlation between the number of clients and the number of female clients, and between the number of savers and the number of frequent savers, so that it may not provide additional benefits to estimate different models with alternative choices between these variables. The final choice of variables for this study, therefore, includes two inputs, namely labour costs and non-labour costs, and three outputs, including number of savers, number of borrowers and number of groups (see Table #2). Labour cost is selected as an input instead of number of employees because it accounts for possible differences in labour quality across schemes and mitigates the difficulties of taking into account part-time labour.

Among the three selected outputs, the number of groups is included in an attempt to accommodate the social objective of microfinance: to reach the poor who are often located in isolated villages in difficult terrain. In these remote areas

---

4 The use of this labour cost variable includes an implicit assumption that wage rates do not differ significantly across regions, which would be a reasonable assumption.
it is more difficult for microfinance members to travel to attend regular group meetings. Hence, the number of groups tends to be greater in schemes that operate in remote areas (compared with schemes having the same numbers of clients operating in more convenient locations).

The environmental factors that may explain efficiency differences include age of the program and characteristics of the village, including poverty status, availability of grid electricity, accessibility by car, distance to town, and the dummy variable for the North region. It is expected that the age of NMPs would have a positive effect on efficiency levels as the high set-up cost has diminished and the savings revenue would rise as they reach more clients. More importantly, it is expected that together with the passage of time, the proportion of members deposits to total loanable funds would increase, which would create a greater feeling of ownership in NMPs, and hence, could contribute to improvements in efficiency, as suggested by the quotation of Milton Friedman stated in the beginning of this chapter.

Table 2: Variables selected and their definitions

<table>
<thead>
<tr>
<th>Names</th>
<th>Definitions</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labour costs (X1)</td>
<td>Wages for head-quarters staff, and allowances for local staff.</td>
<td>VND’000</td>
</tr>
<tr>
<td>Non-labour costs (X2)</td>
<td>All other operational and financial costs, measured in thousands of VND.</td>
<td>VND’000</td>
</tr>
<tr>
<td>Borrowers (Y1)</td>
<td>Total number of clients, who borrowed from NMPs at least once in the survey period.</td>
<td>Persons</td>
</tr>
<tr>
<td>Savers (Y2)</td>
<td>Total number of members deposited voluntary and/or compulsory savings at least once in the survey period.</td>
<td>Persons</td>
</tr>
<tr>
<td>Groups (Y3)</td>
<td>Total groups available at the NMPs.</td>
<td>Groups</td>
</tr>
<tr>
<td>Age of NMPs</td>
<td>Time since NMPs have operated officially to the time the survey was conducted.</td>
<td>Years</td>
</tr>
<tr>
<td>Poor areas</td>
<td>If the areas are in the Government’s list of poorest areas.</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Electricity</td>
<td>If the areas are connected to the national electricity network.</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Car accessibility</td>
<td>If the project areas of the NMPs can be accessed by cars.</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Distance to town</td>
<td>The distance from the project areas to the district capital.</td>
<td>Km</td>
</tr>
<tr>
<td>Ethnic minority clients</td>
<td>If the project serves clients of ethnic minority groups.</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Financial sustainability</td>
<td>A dummy variable representing if obtaining financial sustainability is one of the objectives of the programs.</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Repeating borrowers</td>
<td>The percentage of members who borrowed more than once in the last 12 months.</td>
<td>Per cent</td>
</tr>
<tr>
<td>Max loan size</td>
<td>The maximum amount that a member can borrow.</td>
<td>VND’000</td>
</tr>
</tbody>
</table>

Regarding the environmental variables, it is expected that NMPs operating in more favourable conditions, such as access to electricity, good roads and
Results and discussions

This section presents our results on the efficiency analysis of NMPs surveyed using the DEA technique and second stage analysis on efficiency determinants. In addition, there is a sensitivity analysis using alternative estimation techniques, choices of data and treatment for sample variability. In general, NMPs surveyed were relatively efficient, recording an average technical efficiency score of 75 per cent, and the results are relatively robust among alternative techniques and input/output combinations.

a) Efficiency estimates

Results from the DEA models using two inputs (labour and non-labour costs) and three outputs (number of borrowers, number of savers and number of groups) are presented in Table (detailed results for all NMPs are presented in Table A2.1 in Appendix 2). As can be seen, the overall efficiency of the NMPs surveyed was very modest at 20.8 per cent, suggesting that NMPs can improve their performance by 80 per cent. Table also shows that scale inefficiency plays a significant role in the low performance of most NMPs. On average, NMPs surveyed achieved only 29 per cent efficiency with regard to their operational scale. The large gap in the operational scale of NMPs (the largest NMP in the sample serves 20,000 clients while the smallest has only 70 clients), may contribute to an “unrealistic” figure that the least inefficient NMP can improve their scale efficiency by 98.7 per cent (i.e., the minimum scale efficiency is 1.3 per cent). The detailed results show that with the exceptions of NMP No. 2 and No. 21, other programs operate at increasing returns to scale, and hence, can improve performance by expanding operations. This finding is consistent with studies reviewed by Berger and Humphrey (1997) where scale differences often account for very low average efficiency performance of banks and credit unions.

Although there is substantial room for improvement in scale efficiency, it may be difficult for inefficient NMPs to adjust their scale, at least in the short run. Therefore, the technical efficiency measure provides more relevant information for NMPs. Table #.3 shows that the average technical efficiency of NMPs

---

5 Results of the main DEA model were obtained using the DEA computer Program (DEAP) version 2.1, written by Coelli (1996a).

6 Because of the confidentiality agreement, we do not disclose names of NMPs associated with their efficiency performance in this book.
surveyed is 75 per cent; hence on average inefficient NMP can save up to 25 per cent of inputs while being able to produce the existing level of outputs. However, the third quartile of the TE score is very high at 94.6 per cent, suggesting that a number of NMPs operate closely to the efficient frontier. In addition, the first quintile of TE score is 56.8 per cent, which is reasonably close to the central tendency, given that the DEA method does not take into account measurement errors. Therefore, the level of technical competence may not vary considerably among NMPs surveyed. This finding is not surprising because most NMPs surveyed share similar characteristics such as operating in poor, rural areas and often using staff from the local Women’s Union.

Table 3. Efficiency of microfinance programs (main results)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Overall efficiency</th>
<th>Technical efficiency</th>
<th>Scale efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.208</td>
<td>0.748</td>
<td>0.291</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>0.249</td>
<td>0.203</td>
<td>0.301</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.013</td>
<td>0.413</td>
<td>0.016</td>
</tr>
<tr>
<td>1st quartile</td>
<td>0.057</td>
<td>0.568</td>
<td>0.061</td>
</tr>
<tr>
<td>Median</td>
<td>0.112</td>
<td>0.755</td>
<td>0.147</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>0.241</td>
<td>0.946</td>
<td>0.458</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

One advantage of the DEA technique is that it can provide information about peers (i.e., those NMPs on the efficient frontier) that an inefficient NMP could learn from to improve its efficiency. The results of this study revealed that five NMPs (No. 2, 5, 9, 21 and 35) are often referred to as peers for inefficient NMPs. Among the five influential players, only NMP No. 2 and No. 21 also achieve scale efficiency. However, it is interesting to note that NMP No. 2 is referred to as a peer only seven times (see Table #4), while NMPs No. 5 and No. 9, which are relatively smaller scale, were the peer of 30 and 29 inefficient NMPs, respectively. Therefore, it is possible that the input-output combination of NMP No. 2 was not relevant for many inefficient NMPs: NMP No. 21 was the most influential, appearing 38 times as a peer for inefficient NMPs despite it being only 50 per cent the size of NMP No. 2. Therefore, the size of NMP No. 21 and its input-output combination may be most relevant for inefficient NMPs. A closer examination revealed that this influential NMP has been in operation for more than 10 years, and more importantly, the principal aim of its donors was to build a successful demonstration model; hence, its efficiency level at the survey period may be due to the large amount of support it received in the set-up period.

Table 4. The most influential NMPs

<table>
<thead>
<tr>
<th>NMPs No.</th>
<th>Peer counts</th>
<th>Overall efficiency</th>
<th>Technical efficiency</th>
<th>Scale efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>7</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>0.041</td>
<td>1.000</td>
<td>0.041</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>0.016</td>
<td>1.000</td>
<td>0.016</td>
</tr>
<tr>
<td>21</td>
<td>38</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>35</td>
<td>9</td>
<td>0.060</td>
<td>1.000</td>
<td>0.060</td>
</tr>
</tbody>
</table>
b) Second-stage regressions

Factors determining the efficiency of NMPs can be analysed using several methods such as sub-samples, one-stage estimation (i.e. include environmental variable in efficiency estimates), and two-stage estimation (i.e. regress efficiency scores against environmental variables) (Coelli et al., 2005). This study focuses on the two-stage approach but the one-stage estimation is implemented for sensitivity analysis.

Before conducting the second stage analysis, it is necessary to discuss some issues related to this approach. Tobit regressions have been applied widely in second stage analysis because a number of DEA efficiency scores equal one. However, a Monte Carlo experiment by Simar and Wilson (2007, p. 48) showed that estimates from truncated regressions were fairly close to the truth whilst results obtained from Tobit regressions were, according to the authors, “catastrophic”. Therefore, truncated regressions are selected for the second stage analysis in this study.

Another concern with the second stage approach is the interdependency of efficiency estimates (i.e., changing the position of one firm in the frontier can cause changes in the efficiency estimates for other firms), and hence, the standard regression approach is irrelevant. This issue can be mitigated using the bootstrap method to construct estimates from a pseudo population. There are a few studies in the literature, particularly Xue and Harker (1999), Hirschberg and Lloyd (2002), Casu and Molyneux (2003) and Simar and Wilson (2007), that have applied this approach. The bootstrap procedure in the first three studies applied a data generating process (DGP) proposed by Xue and Harker (1999), which considered that random noise has a normal distribution.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Truncated regressions</th>
<th>Tobit regressions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficients</td>
<td>z-ratios</td>
</tr>
<tr>
<td>Age</td>
<td>-0.054</td>
<td>-1.330</td>
</tr>
<tr>
<td>Age squared</td>
<td>0.002</td>
<td>0.710</td>
</tr>
<tr>
<td>Poorest areas (Y/N)</td>
<td>***-0.257</td>
<td>-3.970</td>
</tr>
<tr>
<td>Electricity (Y/N)</td>
<td>*0.155</td>
<td>1.650</td>
</tr>
<tr>
<td>Car accessibility (Y/N)</td>
<td>*-0.201</td>
<td>-1.850</td>
</tr>
<tr>
<td>Town distance (km)</td>
<td>***0.039</td>
<td>3.800</td>
</tr>
<tr>
<td>Ethnic minority (Y/M)</td>
<td>**0.150</td>
<td>2.040</td>
</tr>
<tr>
<td>Sustainability (Y/N)</td>
<td>*-0.113</td>
<td>-1.810</td>
</tr>
<tr>
<td>Repeating borrowers</td>
<td>0.001</td>
<td>1.370</td>
</tr>
<tr>
<td>Max loan size</td>
<td>-9.2E-06</td>
<td>-0.930</td>
</tr>
<tr>
<td>Constant</td>
<td>***0.860</td>
<td>5.460</td>
</tr>
</tbody>
</table>

Note: ***, ** and * represent, respectively, 1, 5 and 10 per cent significant level.
We have tried a simulation experiment based on the DGP introduced by Simar and Wilson (2007), which is more relevant to the nature of efficiency scores (i.e., bounded between zero and ones). However, the results of both the single bootstrap procedure (i.e., resampling directly from original efficiency estimates) and the double bootstrap procedure (i.e., efficiency estimates were corrected using the bootstrap procedure in Simar and Wilson (1998) before bootstrapping the second stage estimates) do not generate significant estimates. In addition, we observed that the results from Tobit regressions provide consistent estimates as compared to the truncated regression. Therefore, we do not report the bootstrap results of second-stage regressions as they are the subject of further investigation.

As can be seen in column 1 of Table, both age and age-squared received unexpected signs, which suggested that NMPs can be less efficient as they become more mature; however, this reduction in efficiency over time may be plausible. From a practical viewpoint, NMPs often lack a funding capacity to meet the increasing credit demand of clients, and so they kept providing small loan sizes. As clients become wealthier, they may leave microfinance programs to seek larger loans from commercial financial institutions, thanks to their now improved financial knowledge and experience. In the meantime, the NMPs seek new clients, who may not have the equivalent confidence and capacity to make use of microfinance services. Even when new clients are not vulnerable, the theory on information asymmetry suggests that the peer-monitoring system becomes weaker when demand for funds become more satisfied (i.e., if members can easily switch to other institutions, borrowers have less incentive to monitor and repay for defaulting members). Economic theory also suggests that, as the level of competition increases (e.g., the arrival of commercial institutions in the microfinance market), NMPs will have to use more resources in order to keep existing clients, as well as serving newer clients. All these factors can lead to the possible decline of efficiency of NMPs.

The sign of the age-squared variable suggests that the marginal effect of technical efficiency increases over time (i.e., becomes less negative). In particular, when the age of a the NMP reaches $0.054/(2*0.002)=13.5$ years the marginal effect becomes zero. One possible factor making this happen is the increase of ownership over time (i.e., the share of members’ funds in the total loanable fund increase). Due to the low level of saving mobilisation, the restriction to mobilise saving within members, and the relatively young age of NMPs (less than 12 years), deposits from members were only 30 per cent of the total fund, on average. Therefore, most people still considered microfinance funds were a kind of external gift from NGOs. The weak level of ownership can lead to detrimental effects on efficiency, especially when NGOs hand over microfinance programs to local authorities.

The poor area dummy variable relates to a location in one of the poorest communities in Vietnam. It is expected that this would reduce efficiency levels because these areas are often located in remote areas, and this can lead to higher

---

7 The DGP of this bootstrap procedure will be discussed in Section 5.4.3.2, where various means of sensitivity analyses for efficiency estimates are presented.
transaction costs. Both truncated and Tobit regressions confirmed this expectation significantly.

A similar story is shown by the dummy variable for the availability of grid electricity, as both the truncated and Tobit regressions supported the expectation that NMPs in villages with grid electricity are more efficient. However, only results produced by truncated regressions were statistically significant. One possible factor supporting this result is that NMPs operating in areas connected to the grid electricity may be able to apply labour-saving equipments such as computers and fax machines and hence the productivity of services can be higher compared to those operating in other regions.

Three other environmental variables, including car accessibility, distance to township and the dummy variable for members of ethnic minority groups (i.e. those who are often poor and live in remote areas) also received significant but unexpected estimates. In particular, villages that are accessible by car have low efficiency in their NMPs. In contrast, NMPs are more efficient if they operate in areas further from townships and in areas with ethnic minority groups. These results are not as expected: it was thought that it would be more expensive to deliver financial services with difficult infrastructure and terrain (e.g., far away from township and no car access). Ethnic minority groups (with the exception of the Chinese) often lack essential skills to integrate into mainstream economic activities (Baulch et al., 2002). Therefore, NMPs operating in ethnic minority areas often need to spend more resources supporting activities such as training in numeracy, literacy and production techniques (Che, 2002), which should make them less efficient than others. One possible reason for this counter-intuitive result is that NMPs operating in remote areas, where ethnic minority groups often reside, face less competition from other financial service providers, and hence, they can attract larger number of clients with less effort than those operating in other areas, ceteris paribus.

The most important characteristics of NMPs, such as maximum loan size, the attitude towards financial sustainability and the proportion of repeating borrowers, have almost no influence on the efficiency of NMPs as the estimated parameters were very close to zero. This finding, similarly shown by all alternative estimates, may be due to the fact that NMPs in the survey share similar characteristics such as loan size and interest rates. It is also shown that NMPs with financial sustainability objectives are less efficient. One may argue that the sustainability objective drives NMPs to apply tougher screening and monitoring devices to select clients. In addition, NMPs that really care about sustainability would have to spend more resources on recruiting and training independent staff (i.e., not depending on local staff of MOs) and establishing a sustainable organisation (e.g., transfer NMPs to PCFs) after the termination of their microfinance projects. These effort are time and resource consuming, which could make them appear less efficient in the short-run (recall that NMPs in this study had been in operation for 12 years at most). It is also possible that mature NMPs become less efficient due to the reduction of subsidy from donors after the set-up period.
Conclusions

In this chapter, we have analysed the technical efficiency of microfinance schemes in the north and the central regions of Vietnam using the DEA technique. The study uses the traditional production approach in defining input and output variables for use in analysing the efficiency of NMP surveyed. However, we amended the approach by including an extra variable, the number of groups, to help capture the social aspect of microfinance activities.

DEA results produced an average technical efficiency score of 75 per cent, suggesting that there is scope for efficiency improvements in many of these schemes. In addition, most NMPs surveyed were very inefficient in operational scale but this issue is difficult to address given the limited funding from donors and poor performance on saving mobilisation from members.

A second stage analysis, using truncated and Tobit regressions, revealed that NMPs may be less efficient as they become more mature. It is possible that the departure of capable clients and the arrival of more vulnerable clients may have caused the decline of efficiency among mature NMPs. Other environmental variables (e.g., location and infrastructure) and program characteristics (e.g., maximum loan size and attitude for financial sustainability) confirmed the expectation that NMPs operating in favourable environments are more efficient.

References


