1. Introduction to the Research

1.1 Introduction

The main purpose of this study is to investigate some of the productivity and efficiency issues of crop farming in Mongolia in the period 1976-1989. In this chapter a general outline of the study is presented. A brief historical overview and an outline of the current economic reform of crop farming in Mongolia are given in Section 1.2. A justification of the study is presented in Section 1.3. The objectives and hypotheses of the study are stated in Section 1.4. The analytical methods and the data employed are described in Section 1.5. Finally, the organisation of the thesis is described in Section 1.6.

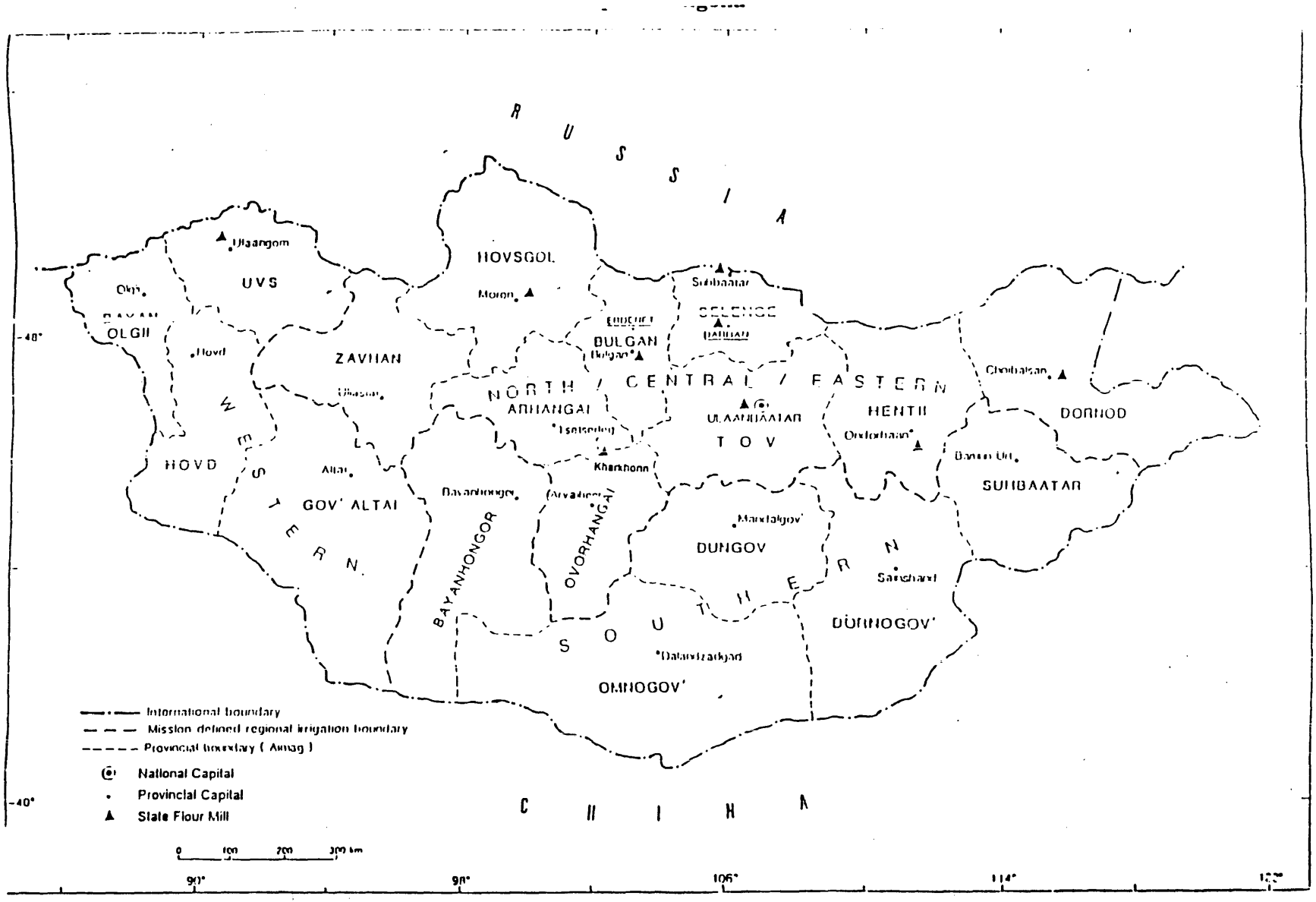
1.2 Historical Overview and Current Economic Reform

Mongolia is a land-locked country in Central Asia. It is bordered by the People's Republic of China to the south and east and Russia to the north and west (Figure 1.1).

The territory is approximately half the size of India, having about 1.6 mln sq. km. The country is situated between the longitudes of 87.4° and 119.6° to the west, and the latitudes of 52.1° and 41.4° to the north. The main land mass is elevated at an average of 1580 metres above sea level.

Agricultural land, of which the majority is pasture, represents 80 per cent of the national territory and is estimated to be 120 mln ha (Asian Development Bank, 1992a, p. 16). The population was estimated to be 2.2 mln in 1993 (State Statistical Office, 1994). Mongolia pursued a centrally-planned economic system for 70 years until 1990. It was the second oldest communist regime after that of the former Soviet Union. Under the centrally-planned economic system all land was owned by the state.

Figure 1.1 Administrative Map of Mongolia



Agriculture has been the core of the economy for millennia. In 1993, agriculture employed 39 per cent of the work force and generated 35 per cent of gross domestic product (State Statistical Office, 1994). Animal husbandry, which is the dominant sector in agriculture, is based largely on semi-nomadic open grazing methods. As of 1990, the total number of livestock was estimated to be 56 mln head in dry sheep-equivalent units[1] (Asian Development Bank, 1992b). The principal animals are sheep (38 per cent[2]), goats (20 per cent), cattle (11 per cent), horses (9 per cent) and camels (2 per cent) (Asian Development Bank, 1992a, p. 16). The livestock sector produces over three-quarters of the total agricultural output (in value terms) and is the single most important agricultural source of export earnings of the country (Asian Development Bank, 1994, p. 35).

Crop farming is not traditional, although grain, vegetables and potatoes have long been the essential elements in the diet of Mongolians (Chalmers, 1993). Since World War II, rapid population growth, increasing concern over food security and overall shortages of cereals in the Eastern Bloc gave thrust to the development of an entirely new crop sector in Mongolia (Chalmers, 1993). In the period between 1958 and 1960 the biggest initiative to harness virgin land took place. During this time around 250 000 ha of land were cultivated primarily for grain production. This crop land steadily increased to a peak of some 1.3 mln ha in 1990 (Asian Development Bank, 1994, p. 7). Although this amount (1.3 mln ha) represents merely one per cent of total agricultural land, according to the Land Management Institute of Mongolia (Asian Development Bank, 1992a) this is the maximum amount of land suitable for crop farming in the country.

---

[1] A standard conversion basis for animals based on output and feed consumption rates: 1 head of cattle = 7 dry sheep-equivalents, 1 head of camel = 7 dry sheep-equivalents, 1 head of horse = 6 dry sheep-equivalents, 1 head of goat = 0.8 dry sheep-equivalents.

[2] This is the percentage of a sheep in total number of livestock converted in sheep units. The same interpretation applies to other animals.

About two-thirds of the crop land is annually under crops and the balance lies under fallow (United Nations Development Program, 1992). The principal crops are grain (wheat 85 per cent; the balance is barley and oats), potatoes and other vegetables (referred to hereafter simply as vegetables).

Mongolian crop farming is characterised by low productivity and riskiness due to high altitude, harsh and uncertain climatic conditions and the absence of domestic production of such important agricultural inputs as machinery, fertiliser and chemicals (Chalmers, 1993). The main cropping areas are in the central, northern and eastern parts of the country (see Figure 4.1, in Chapter 4). On this area, the average growing season lasts 100 to 120 days and annual precipitation ranges from 250 to 350 mm (World Bank, 1995).

State farms and agricultural co-operatives were developed under communism. The structure and functioning of state farms followed that of Soviet Sovkhozy (Chalmers, 1993). They accounted for about 81 per cent of total national crop land (see Table 4.2 in Chapter 4) and were the main producers of crops in the country. In contrast, the agricultural co-operatives were mostly engaged in livestock production.

Despite its many technical and economic limitations, the Ministry of Agriculture allocated substantial resources to crop production during the three decades ending in 1990 (Ulziibat, 1992, p. 42). As a result, in the second half of the 1980s, self-sufficiency was achieved in grain production and a substantial part of the total vegetable and potato requirements were supplied from domestic production (United Nations Development Program, 1992, p. 1; Ulziibat, 1992, p. 42).

The dramatic political, social and economic reforms of former centrally-planned economies (including Mongolia) that began in the early 1990s have since had a substantial effect upon Mongolian agriculture. As a result of the privatisation process, the former state farms, which were primarily involved in crop production, have been broken up into smaller share-holding companies with their workers becoming the shareholders. The new competitive market environment, to which the individual companies have been exposed, has put enormous pressure onto

farms. The enhancement of efficiency and productivity of crop farming has become a vitally important issue, both to the farmers themselves and to a Ministry of Agriculture concerned about declining production levels.

1.3 Justification of the Study

This study is concerned with efficiency and productivity issues of Mongolian crop farming in the immediate pre-reform period rather than the current post-reform period for three main reasons. Firstly, it would be too difficult and costly to collect an adequate set of post-reform farm data relative to the financial constraints faced by the author. The limited post-reform data that was collected as part of the current research was found to be of very poor quality due to the disruptions both to record keeping and to farm input markets during the reform process. Secondly, it was expected that the analysis of 1991-1996 data is more likely to reflect the costs of the reform process rather than the benefits of a market economy. Thirdly, it was observed that following privatization, most grain farms continue to function in a similar way to the old state farms in terms of structure and technology, but with reduced size of units. Hence, it was believed that the analysis of pre-reform data is likely to provide information which will be relevant to the Ministry of Agriculture in the formulation of development strategies for the grain sector.

A number of development initiatives were introduced in Mongolian crop farming in the two decades leading up to 1990. (Ulziibat, 1992). In the crop sector, these initiatives basically fell into three categories:

(i) increased use of conventional inputs such as land, labour, machinery and fertilisers;

(ii) the development and importation of new technology; and

(iii) a series of policy reforms aimed at improving farm efficiency.

As shown in Table 1.1, in the period between 1970 and 1989, total sown area increased for grain by 61 per cent, for potato by 300 per cent and for vegetables by 164 per cent.

Table 1.1 Sown area and harvest for grain, potato and vegetable, selected years 1970-1989

Year	Sown area (000 ha)			Harvest (000 t)		
	Grain	Potato	Vegetable	Grain	Potato	Vegetable
1970	419.6	2.9	1.4	184.8	20.8	11.7
1980	575.6	7.5	2.4	259.1	37.9	26.3
1985	634.6	9.0	3.0	890.2	106.3	36.7
1986	631.5	9.3	3.5	869.6	123.9	43.8
1987	523.0	11.1	3.6	689.7	138.0	45.7
1988	641.0	11.8	3.6	814.4	97.9	51.2
1989	673.9	11.5	3.7	839.0	148.0	54.9
1989/1970 (per cent)	160.6	396.6	264.3	454.0	711.5	469.2

Source: Ministry of Food and Agriculture (1994a), Ulaanbaatar.

During the period 1980-1988, the number of workers in the crop sector increased by 16.5 per cent, in spite of an 8.3 per cent decline in the total agricultural workforce (see Table 1.2).

Table 1.2 Total agricultural workforce, selected years 1980-1988

	1980 (person)	1985 person)	1987 (person)	1988 (person)	1988/1980 (per cent)
Crop production	26 600	30 300	30 800	31 000	116.5
Total	197 000	181 500	178 500	180 600	91.7

Source: State Statistical Office (1990).

As Table 1.3 shows, in 1988 compared to 1980 the stock of agricultural machinery on state farms had increased considerably, with the number of tractors increasing by 56 per cent. Furthermore, as Figure 1.2 shows, in the period between 1976 and 1990 the total import of chemical fertiliser more than doubled (and then collapsed to essentially nothing in 1991 and 1992).

As a result of these substantial development initiatives in the crop sector, by the end of the 1980s the country claimed it was self-sufficient in staple crops such as grain, potato and vegetables (United Nations Development Programme, 1992). During the period 1970-1989 grain output increased by 354 per cent and in 1989 it reached 839 000 t. During the same period, potato output increased by 611 per cent and vegetable output increased by 369 per cent, reaching 148 000 t, and 54 900 t, respectively in 1989 (Table 1.1). As a result, domestic consumption of crop products reached historic record levels in the late 1980s. For instance, in 1988, national annual per capita consumption of cereal was 406.5 kg, and that of potato and vegetables was 51.5 kg and 28.1 kg, respectively (State Statistical Office, 1988).

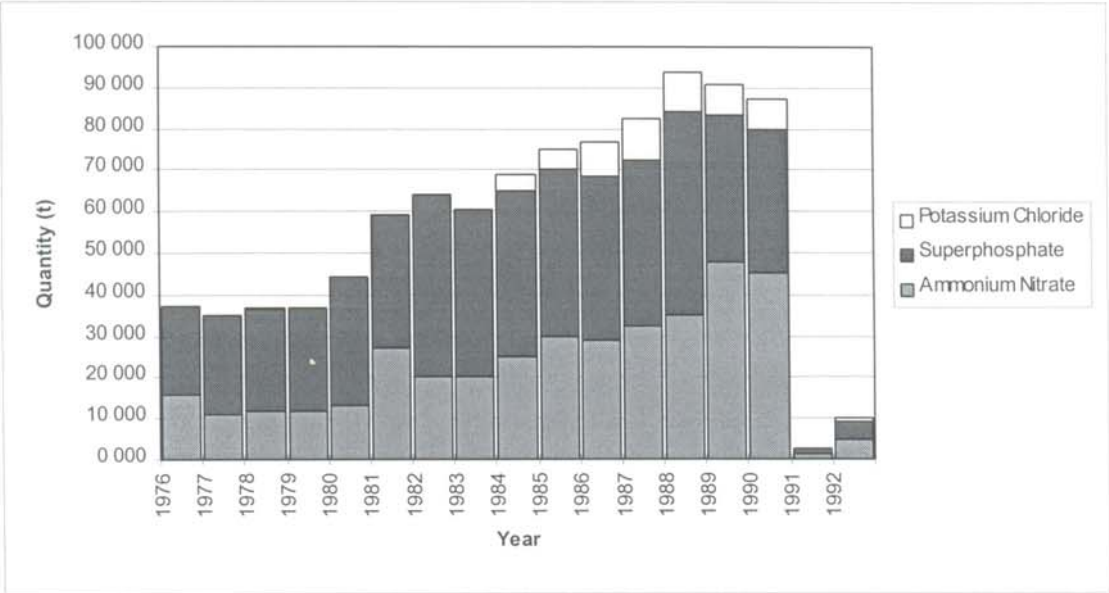
Table 1.3 Agricultural machinery stocks of state farms, selected years, 1960-1988

Year	Tractors (physical units)	Tractors (15 horse-power units)	Ploughs (physical units)	Cultivators (physical units)	Seeders (physical units)	Harvesters (physical units)
1960	1100	2 5)0	1100	200	1200	800
1980	4100	11 6)0	1500	2100	3400	1800
1985	4500	13 9)0	1000	4200	4700	1800
1987	6200	N/A[a]	1100	4000	5200	2100
1988	6400	N/A	1200	4400	5400	2200
1988/1980 (per cent)	156.1	N/A	80.0	209.5	158.8	122.2

Source: State Statistical Office, Ulaanbaatar (1990).

[a] Not available

Figure 1.2 Fertiliser imports, 1976-1992



Source: Ministry of Food and Agriculture (1994b), Ulaanbaatar.

The emphasis of the development initiatives differed in individual policy sub-periods.[3] While during the first sub-period (1976-1980) output growth was ensured mostly by increased use of conventional inputs, during the second (1981-1985) and third (1986-1989) sub-periods, the Ministry of Agriculture began shifting its policy from a so-called "extensive" to an "intensive" growth strategy (Ulziibat, 1992). The emphasis of the new approach was the increased role of new technology (Unen, 1981), investment in human resources and the introduction of incentive systems (Unen, 1986) with the aim of improving farm productivity levels. In particular, during the last four years (1986-1989) of the centrally-planned economic regime, several new forms of farm incentive systems aimed at improving farm performance were experimented with within the state farm structure (Ministry of Food and Agriculture, 1990b).[4] This was a reflection of the

[3] The current study covers the fifth (1976-1980), sixth (1981-1985) and seventh (1986-1989) five-year plan periods of the national economy of Mongolia. These are referred to here as the first, second and third sub-periods respectively. The seventh plan period was meant to be 1986-1990 but was shortened to 1989 by the reform.

[4] More detailed discussion of Government policy on the introduction of new technology and gradual economic reforms in the crop farming sector is presented in Chapter 4.

new wave of Gorbachev's economic reform, "Perestroika", which was carried out throughout the Eastern Bloc.

Despite all the intensification measures and various reform attempts and, ironically, some improvements in farm productivity and performance indicators recorded in official Government documents (Ministry of Food and Agriculture, 1990b; 1991 and Table 1.4) the unprecedented radical economic reforms took place after 1990 and were based on economic arguments.

Hence, it remained unclear what the major driving forces behind the growth strategy in the crop sector were in the last two decades of central planning and what the impact was, on farm productivity, of Ministry of Agriculture efforts to invest in new technology, improved human resources and new incentive systems.

While some recent studies reported that existing technology in crop farming was obsolete and inadequate (Ulzibat, 1992; Ulrich, 1994), others found that the main problem facing the crop sector was poor management and organisational deficiency within the farm (Dixon, 1989; United Nations Development Programme, 1992). For instance, Ulziibat (1992) questioned the standard and adequacy of Soviet technology used in the crop sector and Ulrich (1994) reported that the existing technology in the crop sector was well below the western standards.

In contrast, several studies by experts of crop farming in the pre-reform period concluded that the primary problem with farm performance was not technological but was associated with poor management, organisational and incentive issues (implying low efficiency levels).

For instance, United Nations Development Programme (1992, p. 2) reports:

The problems of agriculture are organisational rather than technical ... However, the underlying technology employed was essentially sound. The problems of the system were more a lack of economic criteria to allocate resources, and more importantly, the lack of a human dimension to encourage effort and initiative ... .

Table 1.4 Selected performance indicators of state farm sector, 1981-1987

Performance indicators	Unit	198	1982	1983	1984	1985	1986	1987
Total crop land	000 ha	749.)	750 3	793.5	784.2	796.2	793.3	655.7
Capital	mln tgs[a]	800.)	867 8	929.9	1044.5	1100.3	1097.5	1117.6
Workers	persons	913)	9403	10 421	10 137	10 838	11 132	99 65
Total output (in fixed prices)	mln tgs	184.4	253.0	364.9	270.4	362.6	459.0	402.2
Labor productivity	000 tgs per person	20.	26.9	35.0	26.6	33.4	41.2	40.4
Return on capital	per cent	17.0	20.0	23.0	20.0	23.0	30.0	28.0
Capital per person	000 tgs	87.0	92.3	89.2	103.0	101.5	98.6	112.1
Capital investment (whole-farm)	mln tgs	193 4	247.1	260.6	314.9	255.2	251.0	266.3
Grain production	000 t	266 1	371.8	584.7	371.4	580.6	551.4	437.8
Potato production	000 t	28.0	55.5	66.6	82.9	69.4	82.1	90.0
Vegetable production	000 t	21.0	25.0	23.8	21.1	25.0	27.4	23.6

Source: Ministry of Food and Agriculture (19:9), Ulaanbaatar.

[a] "tgs" stands for Tugrigs, the Mongolian currency. The official exchange rate of Tugrigs against US dollars was fixed until 1989 to the rate 1 USD = 3.1 Tgs (Chalmers, 1993, p. 7).

Also Dixon (1989, p. 15) wrote:

> *Inadequate management, particularly at the farm level, is the most critical constraint limiting growth in agricultural production... This needs to be tackled by strengthening management skills, making available better agricultural information and decentralising decision-making.*

No detailed analysis of total factor productivity and efficiency in Mongolian crop farming has yet been conducted. Hence, the influence of the development and importation of new technology and the successive economic reforms in Mongolian crop farming that took place in the period 1976-1989 need to be quantified. Therefore, the present work is founded on the belief that the identification and quantification of the driving forces behind Mongolia's output growth strategy and exploration of the main factors causing variability in farm performance in the pre-reform period will help policy makers to focus better on formulating and implementing the current on-going policy reform measures.

In the wider context, the actual reasons behind the economic failure of centrally-planned economies are far from clear and are still under debate (Ofer, 1987; Bergson, 1987, 1992; Easterly and Fischer, 1995).

Only a few studies related to productivity and efficiency issues have been carried out for the former centrally-planned agricultures, except for China. These studies often had controversial findings and frequently found against the prevailing course of development (Carter and Zhang, 1994; Johnson *et al.*, 1994; Brada and King, 1993; Koopman, 1989).[5] Moreover, the majority of these analyses looked at agricultural efficiency and productivity issues using aggregated national-level data (Gemma, 1991). Hence, the current analysis of farm-level data may provide valuable insights that may have been masked by aggregation effects in previous studies. Furthermore, given the fact that Mongolian crop farms were almost exact prototypes of Soviet Sovkhozy (i.e., state farms) in terms of structure and

[5] Detailed review of these studies is given in Chapter 3.

functioning, and also that there has been a striking similarity in policy change patterns in Mongolian and Soviet agriculture, this analysis of Mongolian farm-level data may also provide some additional knowledge of the characteristics of pre-reform Soviet-style agricultural enterprises.

The current study covers the 14-year period 1976-1989 leading up to the end of Mongolia's centrally-planned economic system primarily for the reason that consistent comparable data sets were available only for that period.

1.4 Objectives and Hypotheses

Objectives

The main objectives of the current study of Mongolian crop farming are to:

(i) measure the extent of, and changes in technical efficiency, technical progress and total factor productivity (TFP) of grain and potato farms[6] for the period 1976-1989;

(ii) measure and analyse factors affecting efficiency levels;

(iii) investigate farm scale economies; and

(iv) draw policy implications related to the current development of Mongolian grain and potato farming.

Hypotheses

The principal hypotheses that are considered in this study are as follows:

---

[6] It should be noted that even though the grain and potato farms produced some vegetables other than potato, the vegetable component was left out of the current study for the following reasons: (i) input-output data for vegetable production were incomplete (ii) it is also relatively insignificant compared to grain and potato production in terms of both sown area and total production (see Table 1.1).

(i) the farms were technically efficient

(ii) average technical efficiency did not change over time;

(iii) there was no technical change over time;

(iv) total factor productivity did not change over time;

(v) five farm-specific factors (technical education, experience of grain-farm mechanizators, Soviet investment, farm incentive systems and favourable agro-ecological conditions) had no impact on efficiency;

(vi) farms exhibit no economies or diseconomies of scale, i.e., the elasticity of production with respect to scale is unity.

1.5 Analytical Methods and Data

For the purpose of measuring the extent of, and changes in technical efficiency and technical progress, and for estimating scale economies in grain and potato production, two separate measurement techniques were used: the stochastic frontier production function (SFPF) (Battese and Coelli, 1992) and data envelopment analysis (DEA) (Färe, Grosskopf, Norris and Zhang, 1994). Then the efficiency scores and technical changes measured were used to calculate TFP change. Two separate approaches were employed to calculate TFP change, one based on a parametric frontier (Nishimizu and Page, 1982; Perelman, 1995) and the other based on a non-parametric frontier (Färe, Grosskopf, Norris and Zhang 1994).

For the purpose of identifying and quantifying the main factors causing efficiency variation in grain farming, the SFPF with a technical inefficiency-effects model for panel data (Battese and Coelli, 1995) was used.

Farm-level input and output data on 48 farms over the 14-year period 1976-1989 were obtained from the farm annual financial and economic reports kept at the Ministry of Food and Agriculture. Additional data on farm-specific characteristics

for the final three years of the study period, 1987-1989, were obtained from the farm human resources reports (Ministry of Food and Agriculture, 1990a).

1.6 Organisation of the Thesis

This thesis is organised as follows. A literature review of efficiency and productivity measurement techniques is presented in Chapter 2. Following this a literature review on empirical studies of productivity and efficiency of agriculture in the centrally-planned economies is provided in Chapter 3. The natural environment, farm structure, functioning and the development alternatives practised in Mongolian crop farming in the pre-reform period are investigated in Chapter 4. In this chapter the data employed in the efficiency and productivity analyses are also discussed.

In Chapter 5, efficiency and productivity analyses of Mongolian grain farms for the period 1976-1989 are conducted using the stochastic frontier production function models. Here attempts were also made to explain the efficiency variation among grain farms for the period 1987-1989 in terms of farm-specific explanatory variables using the technical inefficiency-effects model of Battese and Coelli (1995). Also, the various statistical tests used for the selection of the models and functional forms are discussed here. In Chapter 6 the same analysis as in Chapter 5 is repeated for the case of potato farms.

In Chapter 7, an alternative, non-parametric efficiency measurement method – data envelopment analysis (DEA) – is applied to the same set of data on grain and potato farms as considered in Chapters 5 and 6 to analyse the efficiency and productivity changes in grain and potato farms. The primary purpose of this chapter is to check the robustness of the SFPF results in Chapters 5 and 6. Finally, Chapter 8 synthesises the results of the study and draws some policy implications for the current development of crop farming in Mongolia. This chapter also makes some suggestions as to further research.

2. Literature Review: Methodology

2.1 Introduction

The primary purpose of this chapter is to review the literature on efficiency and productivity measurement techniques with an emphasis on frontier methodology. The chapter is divided into two major parts. In the first part, the concepts and alternative measurement techniques of efficiency of a production unit are discussed and, in the second part, the concepts and measurement techniques of productivity of a production unit are reviewed. While the stochastic frontier approach is the focus of the first part of the review, the frontier-based total factor productivity (TFP) measurement option is the focus of the second part of the chapter.

2.2 Efficiency Measurement

2.2.1 Concepts of efficiency

Efficiency is a vital economic concept and is important in assessing a producer's performance. Efficiency measurement is important for various reasons, but the two principal ones are, as Lovell (1993, p. 5) argues:

> *First, they (efficiency scores) are success indicators, performance measures by which production units are evaluated. Second, only by measuring efficiency ... and separating their effects from the effects of the production environment, can we explore the hypothesis concerning the sources of efficiency ... differentials. In some cases measurement enables us to quantify differentials that are predicted qualitatively by theory.*

The literature based on efficiency arguments for better performance of production units includes the effect of market structure, economic regulation and ownership on production performance (Lovell, 1993, p. 6). Furthermore, by estimating an unobservable frontier function as part of efficiency analysis, additional

information about the technology of best practising firms is obtained (Coelli, 1995b).

The technical efficiency of any productive unit is defined as a comparison between observed and optimal values of its output and input (Lovell, 1993) and it essentially implies certain benchmarks against which the individual units are assessed. In the context of production, it may be defined as the ratio of observed to maximum attainable output from the given input (output-orientated case). This case assumes output-maximising behaviour for the production units. If additional behavioural goals are assumed, such as profit-maximisation or cost-minimisation, then economic efficiency comes into consideration.

Efficiency of a production unit was defined first by Koopmans (1951, p. 60) in a general theoretical framework as:

A producer is technically efficient if an increase in any output requires a reduction in at least one other output or an increase in at least one input, and if a reduction in at least one input requires an increase in at least one other input or a reduction in at least one output.

This definition of efficiency is fairly strict and alternative definitions matched by actual efficiency measurement techniques only partly fulfil the Koopmans' requirement.

One alternative definition for the efficiency of a production unit (along with an actual measurement technique) was proposed by Debreu (1951) and Farrell (1957). This is discussed below.

Economic theory defines a production function as the maximal output attainable from a given set of inputs, implying an upper-bound function for all production possibilities (for example, see Beattie and Taylor, 1985). If one specifies an upper-bound function, then the economic performance of individual producers against those standards can be measured and assessed.

Farrell's (1957) seminal paper, which was largely inspired by an earlier work of Debreu (1951), first offered a unique procedure for both estimating the frontier as well as measuring the productive efficiency (*PE*) of individual production units against this frontier. According to Farrell, the productive efficiency (*PE*) of production units consists of two elements, technical efficiency (*TE*) and allocative efficiency (*AE*). Technical efficiency (*TE*) relates to the distance of individual production units from the frontier while allocative efficiency is associated with the optimal combination of inputs given the prevailing market prices. Farrell's exposition was based on the following upper-bound function with two inputs, X_1 and X_2 and a single output, Y

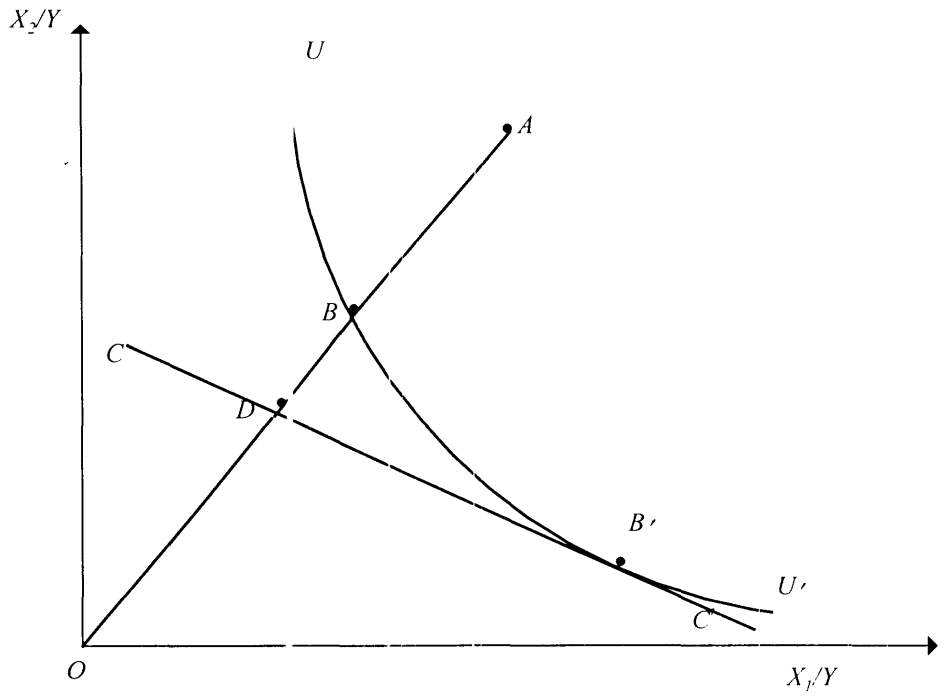
$$Y = f(X_1, X_2). \qquad (2.1)$$

If constant returns to scale is assumed, the above formula can be written as:

$$1 = f(X_1/Y, X_2/Y). \qquad (2.2)$$

In Figure 2.1 an efficient unit isoquant (EUI) is drawn. It is depicted by UU'. UU' is the set of all technically efficient combinations of X_1 and X_2 which may be used produce one unit of output Y.

Figure 2.1 Farrell productive efficiency



Points *A* and *B* are two different production units using the same proportionate mix of inputs X_1 and X_2 (because they lie on the same ray originated from *O*) to produce one unit of *Y*. However, as A lies well above the efficient isoquant *UU'*, it uses more inputs than *B* and consequently is labelled as a technically inefficient firm. Note that *B* is a technically efficient firm because it operates on the frontier. So, the ratio *OB/OA* is defined as the (input-orientated) technical efficiency (*TE*) of firm *A*. Alternatively, since this efficiency score varies between 0 and 1, (1 - *OB/OA*) expresses the technical inefficiency level of firm *A* and shows by how much inputs can be proportionately reduced for the given amount of output. This can be re-stated, for the constant returns-to-scale case, as firm A can increase its output by *OA/OB* for the given amount of inputs (Farrell, 1957).

After introducing a behavioural assumption of cost minimisation of firms and market prices for inputs, efficiency measurement can be extended to include allocative efficiency (*AE*) (Farrell, 1957).

Alternatively, allocative inefficiency may be expressed as (*1 - OD/OB*) indicating the potential cost reduction from using optimal input proportions.

Following Farrell (1957), the overall productive efficiency (*PE*) is expressed as a product of *TE* and *AE*:

PE = (OB/OA) x (OD/OB).

In order to measure the efficiency of production units, two stages need to be completed: construction of an efficient frontier and the measurement of distances of the individual production units to that frontier.

While introducing the conceptual framework for efficiency measurement, Farrell (1957) also proposed two different estimation methods: (i) a non-parametric piecewise-linear convex isoquant constructed so that all the actual observations are bounded from the left and below and (ii) a parametric production function of a form such as Cobb-Douglas is estimated, again so that the unit isoquant from this function lies to the left and below all the observations. Since Farrell's work, studies of efficiency measurement based on frontier estimation have followed these separate tracks, i.e., parametric and non-parametric approaches. Although both are based on Farrell's conceptual framework, their frontier construction and measurement techniques for assessing the efficiency of individual production units are different.

2.2.2 Parametric frontiers

2.2.2.1. Introduction

The development of parametric frontiers has gone through two distinct phases, firstly deterministic frontiers and then stochastic frontiers. These differ in the assumptions imposed on the error term, which is supposed to contain inefficiency elements of production. Under the deterministic frontier framework, any deviation of firms from the frontier is identified as inefficiency, while the stochastic frontier framework assumes that any deviation of firms from the frontier can be attributed to two parts, *viz.*, stochastic elements outside the control of firms and inefficiency elements within the control of firm management.

2.2.2.2. Frontier production functions

Deterministic frontier production functions

The deterministic frontier model for cross-sectional data is defined as:

$$Y_i = f(X_i ; \beta) \exp(-U_i), \qquad i = 1,2,...,N \tag{2.3}$$

where Y_i is the actual production level of the i-th firm, $f(.)$ is a suitable functional form, X_i is the vector of inputs of the i-th firm, β is the vector of unknown parameters to be estimated, U_i, is a non-negative random error associated with the inefficiency of individual firms and N is the number of firms.

The inefficiency term, U_i, is bound between the values of 0 and 1 in its log form, thus restricting individual units' outputs to be less than or equal to the deterministic frontier $f(X_i ; \beta)$, i e., the following inequality holds:

$$Y_i \leq f(X_i ; \beta), \quad i = 1,2,...,N. \tag{2.4}$$

Inspired by Farrell's pioneering work, the first deterministic frontier function was estimated by Aigner and Chu (1968) assuming a Cobb-Douglas functional form. They employed both linear and quadratic programming procedures. In their linear programming method the values of the β's were obtained, such that $\sum_{i=1}^{N} U_i$ is minimised subject to the constraints $U_i \geq 0$, $i = 1,2,3,..., N$. This approach relaxed the constant returns to scale assumption of the original Farrell approach and also had the advantage of robustness to specification errors of inefficiency (Greene, 1993). However, because the estimators do not have any statistical properties, one is never sure how reliable they are, and the method can be quite sensitive to outliers (Førsund *et al.*, 1980).

Aigner and Chu (1968) also suggested that by using chance-constrained programming techniques one could handle outlier problems, allowing some of the observations to be above the frontier. Timmer (1971) followed this suggestion and calculated a "probabilistic frontier" where he estimated the model repeatedly,

gradually discarding the most efficient observations until the parameter estimates were stabilised. The weakness of this work is that one has to arbitrarily choose the number of observations to be omitted from the sample. Another drawback of the programming approach is that the parameter estimates obtained are lacking statistical properties, so one can not be confident about the reliability of either the estimated parameters of the frontier or the efficiency measures (Schmidt, 1976).

Afriat (1972) attempted to solve this issue by imposing a specific distributional assumption (gamma distribution) on the one-sided error term U_i and estimating the parameters using maximum likelihood estimation (MLE) techniques. Here, a certain functional form for the non-positive efficiency component, U_i, is assumed first and all the parameters of the frontier function and the distribution of efficiency components are simultaneously estimated. MLE estimates envelop all the observations and the residual is used to derive efficiency estimates.

Two alternative estimation techniques have been proposed since then: corrected ordinary least squares (COLS) and modified ordinary least squares (MOLS).

COLS was first introduced by Winsten in 1957 (Lovell, 1993). The procedure works as follows: first, the model is estimated using ordinary least squares (OLS) and then the intercept is corrected by upward shifting until all residuals are non-positive and at least one equals zero. The corrected residuals are then used to calculate the technical efficiencies of individual firms. This method does not assume any functional forms for the efficiency distribution.

MOLS, introduced by Richmond (1974), makes distributional assumptions such as half-normal and exponential about the efficiency terms, U_i, and estimates the model first using OLS. Then the intercept is adjusted by the mean of U_i which is derived from the moments of the OLS residuals.

Despite all these advances in the literature, the above-mentioned two problems (namely, that the resulting efficiency scores lack statistical properties and are sensitive to measurement errors) were not solved satisfactorily until the stochastic frontier production function was developed.

Stochastic frontier production functions

The basic assumption underlying the deterministic frontier functions discussed in the previous section implies that all firms in the sample share the same production frontier and that all the deviation of a firm's output from the frontier is due to inefficiency. This assumption has been questioned from the empirical point of view as real-world production is often influenced by stochastic elements such as bad weather, machinery breakdown and input supply as well as by measurement error. It was Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) who independently proposed the stochastic frontier production function as:

$$Y_i = f(X_i ; \beta) \exp(V_i - U_i), \qquad i = 1.....N, \tag{2.5}$$

where

V_i are assumed to be independent and identically distributed random errors which have normal distribution with mean zero and variance σ_V^2;

U_i are non-negative random variables associated with the technical inefficiency of production; and all other notation is as defined earlier.

Based on this stochastic frontier, output-orientated radial efficiency (Debreu-Farrell type) can be calculated as:

$$TE_i = [Y_i / f(X_i ; \beta) \exp(V_i)] = \exp(U_i) \tag{2.6}$$

However, in these early stochastic frontier papers the authors were only able to calculate the average efficiency level of the whole sample, not the specific efficiency scores of individual firms. This was a major weakness of the models (Førsund *et al.*, 1980) until Jondrow *et al.* (1982) suggested a solution involving the decomposition of the error term. Jondrow *et al.* (1982) derived the conditional distribution $(U_i \mid V_i - U_i)$ assuming a half-normal distribution for the inefficiency term U_i. Then the mean of this conditional distribution is inserted in (2.6) for U_i to derive an estimator of TE_i. However, as noted by Greene (1993, p. 81), these estimators of Jondrow *et al.* (1982) are, although unbiased, still inconsistent, because regardless of the number of observations in the sample, the variance of the estimate remains non-zero.

Distributions for the one-sided error term representing inefficiency

The initial stochastic frontier production function model considered by Aigner, Lovell and Schmidt (1977) assumed a half-normal distribution [Figure 2.2(a)] for the one-sided error term. In this half-normal distribution $N(0, \sigma^2)$, the one-sided error term U_i is randomly selected from a half-normal distribution with mean $\mu = 0$ and variance σ^2. However this assumption can be unnecessarily restrictive in that it assumes the majority of inefficient firms should be working within close proximity of efficiency scores equal to one, thus possibly overstating efficiency scores of individual firms.[1] This was first pointed out by Stevenson (1980) who noted that this distributional form implies the probability of occurrence of inefficient behaviour decreases as the inefficiency level increases. He further argued that the factors believed to be associated with efficiency levels of individual production units such as education, training and experience of managers do not likely follow such a pattern. Instead, he proposed a truncated-normal distribution [Figure 2.2(b) and (c)] for the one-sided error, $N(\mu, \sigma^2)$, where the mean of the distribution, μ, is non-zero.

Alternatively, Greene (1990a) suggested a two-parameter gamma distribution for the one-sided error. However, the overly complicated procedure for deriving it prohibits its empirical use as the author himself recognised (Greene, 1993, p. 81). Despite the generalisations for one-sided error term distributions such as truncated-normal and gamma distributions, the original half-normal distribution was still found to be the most useful formulation in practical applications (Greene, 1993, p. 115).

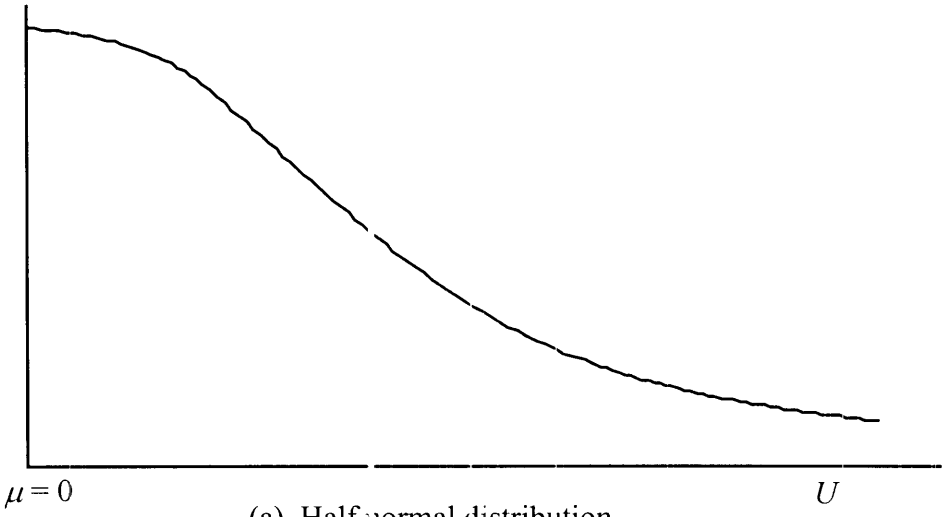
Stochastic frontiers in a panel-data context

In a panel-data context, when each individual firm is observed over a number of time periods, several improvements can be made against the cross-sectional

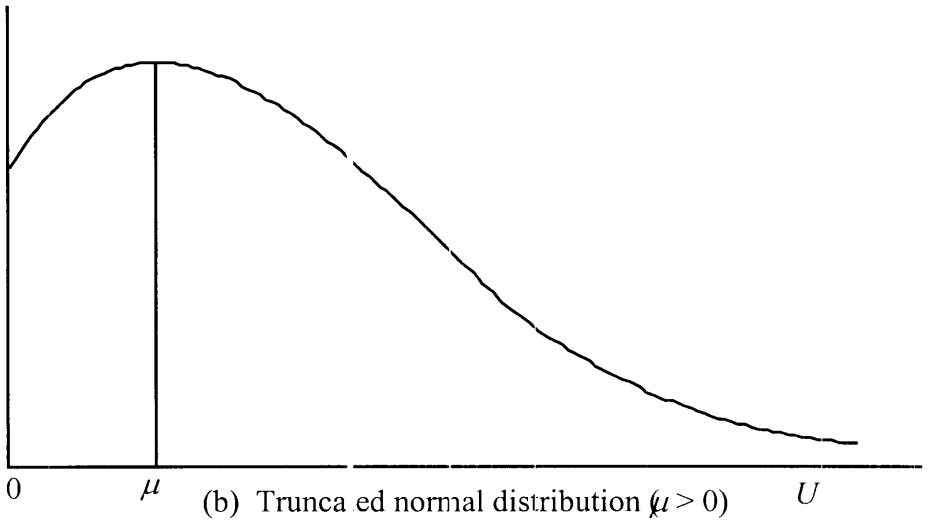
---

[1] Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) also considered the exponential distribution. The exponential distribution also suffers from this problem.

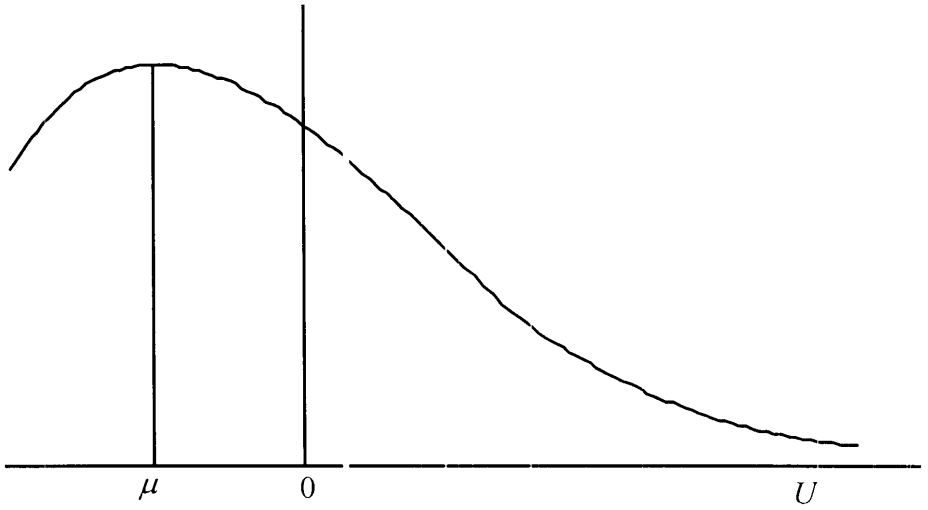
Figure 2.2 Distributional assumptions of the one-sided error term, U



(a) Half normal distribution



(b) Truncated normal distribution ($\mu > 0$)



(c) Truncated normal distribution ($\mu < 0$)

approach. As Coelli (1995b) noted, these include the following features:

1) the necessity to make specific distributional assumptions regarding the U_i is removed;

2) the assumption that the inefficiency term U_i is not correlated with the regressors can be relaxed;

3) consistent estimators of firm efficiencies may be possible;

4) the degrees of freedom for estimation of parameters are increased;

5) it is possible to investigate simultaneously both technical change and efficiency change over time.

Pitt and Lee (1981) redefined the original stochastic frontier model of Aigner, Lovell and Schmidt (1977) in the panel data context as:

$$Y_{it} = f(X_{it}\,;\,\beta)\exp(V_{it} - U_{it}),\quad i = 1, \ldots, N,\quad t = 1, \ldots, T. \tag{2.7}$$

where

Y_{it} denotes the production level for the i-th firm in the t-th year;

X_{it} is a vector of inputs associated with the production of the i-th firm in the t-th year;

$f(.)$ is a suitable function describing the production technology;

V_{it}s are assumed to be independent and identically distributed random errors which follow a normal distribution with mean zero and variance σ_V^2;

U_{it}s are non-negative random variables associated with the technical inefficiency of production; and

β is a vector of unknown parameters to be estimated.

Pitt and Lee (1981) assumed a half-normal distribution on the inefficiency component, U_{it}, and estimated different versions of the above model, each with one of the following assumptions:

1) U_{it}s are not correlated over time and between firms;

2) U_{it}s are constant (invariant) over time.

The time-invariant version of Pitt and Lee's (1981) model takes the form:

$$Y_{it} = f(X_{it}\,;\,\beta)\exp(V_{it} - U_i), \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{2.8}$$

Battese and Coelli (1988) considered a general truncated-normal distribution for time-invariant inefficiency effects, U_i, in model (2.8) and generalised the cross-sectional inefficiency predictor of Jondrow *et al.* (1982) into the panel-data case. This model was extended to accommodate unbalanced data by Battese, Coelli and Colby (1989).

All these stochastic frontier panel-data models had assumed time invariance of inefficiency until the work of Kumbhakar (1990). He imposed on the inefficiency effects U_{it} in (2.7) an additional time-varying structure:

$$U_{it} = [1+\exp(bt+ct^2)]^{-1} U_i \tag{2.9}$$

where the individual firm effects, U_{it} are functions of an exponential time variable and firm-specific random variable U_i. The sizes and signs of the parameters b and c of the deterministic function of time in (2.9) allows inefficiencies of firms to monotonically increase or decrease or to increase and then decrease. However, the disadvantage of this model is its dependence on restrictive distributional assumptions about the technical inefficiency component (Heshmati and Kumbhakar, 1994) and no empirical application has yet been made of this model (Coelli, 1995a).

Cornwell *et al.* (1990) model ed the inefficiency effects through the intercept term in the following way:

(2.10) $$\beta_{it} = \beta_0 - U_{it} = \alpha_i + \beta_i t + \gamma_i t^2$$

where β_0 is an intercept term of the frontier function.

This formulation allows one to calculate firm-specific levels of efficiency which also vary over time. This formulation has been criticised on the ground of interpretational difficulty (Lovell, 1996). The major problem with this approach is that by modelling time-varying inefficiency effects this way, one excludes the possibility of identifying technical change, often modelled as a time trend in the frontier function.

Battese and Coelli (1992) modelled the time-varying inefficiency effect as the product of a deterministic time-varying element and a firm-specific random element:

(2.11) $$U_{it} = \eta_{it} U_i = \{exp[-\eta(t-T)]\} U_i, \quad t \in \tau(i),$$

where η is an unknown parameter to be estimated and U_i are independent and identically distributed non-negative random variables obtained by truncation (at zero) of the normal distribution with unknown mean μ and variance σ^2; and $\tau(i)$ is the sub-set of T_i time periods of the total set of T periods for which the observations for the i-the firm are available.

In this formulation, the value of the parameter η indicates the trend in inefficiency. Its interpretation is as follows: when $\eta > 0$, U_{it} (inefficiency) decreases as t increases; when $\eta < 0$, U_{it} increases; and when $\eta = 0$, U_{it} is constant through time. By modelling the inefficiency trend through the error term, one is able to identify technical change by including a time trend in the frontier function. It also should be noted that in this model only one more parameter is required to be estimated in addition to those for the time-invariant model (Battese, 1992). This model has been applied in a number of agricultural cases including Indian paddy farms (Battese and Coelli, 1992), Ukrainian agriculture (Johnson *et*

al., 1994), and U.S. dairy farms (Ahmad, 1994). The MLE of various versions of stochastic frontier production functions has been automated in the computer program, FRONTIER, Version 4.1 – see Coelli (1994).

Heshmati and Kumbhakar (1994) and Heshmati, Kumbhakar and Hjalmarsson (1995) proposed a firm-effects model where the error term e_{it} is modelled as:

$$e_{it} = V_{it} - (U_{it} + \mu_i) \tag{2.12}$$

where μ_i is a firm-specific component, which captures the firm-specific fixed effects; U_{it} is an asymmetric non-negative random effect representing inefficiency; and V_{it} is a symmetric error term representing stochastic elements in production. In this model the time trend among the regressors is interpreted as technical change. A truncated-normal distribution was assumed for U_{it}.

A two-step estimation procedure was proposed for the Heshmati, Kumbhakar and Hjalmarsson (1995) model.

First step:

1) estimate the parameters of the frontier using the within transformation[2] and with OLS on the transformed data.

2) calculate the residual, m^0_{it} in the following way:

$$m^0_{it} = Y_i - f(X_i ; \beta) \tag{2.13}$$

which in turn has the form,

$$m^0_{it} = \beta_0 + \mu_i + V_{it} + U_{it} \tag{2.14}$$

Second step:

---

[2] See Heshmati (1994, p.34) for technical details on conducting "within transformation" operation.

1) regress m^0_{it} on N dummy variables to obtain $\beta_0 + \mu_i$.

2) calculate the composed error part of the error term in the following way using all the parameters (β) of the function and the N farm-specific dummies:

(2.15) $$V_{it} + U_{it} = m^0_{it} - (\beta_0 + \mu_i).$$

3) disentangle the inefficiency elements from the stochastic element using the approach of Jondrow *et al.* (1982).

This model thus allows one to identify separately the effects of efficiency change, technical change and scale change. It has been used to analyse the Swedish dairy and pork industries (Heshmati, Kumbhakar and Hjalmarsson, 1995).

After deriving individual efficiency scores of the u_{it}, one can also calculate the percentage change of inefficiency as:

(2.16) $$\Delta TEF = 100 \times [\exp(u_{it}) - \exp(u_{i,t-1})]/\exp(u_{i,t-1}).$$

However, a weakness of this firm-specific fixed-effects model is that it is often impossible to determine whether μ_i is a firm-specific time-invariant effect or the input effects which vary across the firms (Lovell 1996).

2.2.2.3. Explaining inefficiency effects

From a policy perspective, the identification and quantification of the factors affecting efficiency scores of individual firms is quite important. However, this issue has not been sufficiently addressed in the past despite its importance (Farrell, 1988).

The so-called "two-stage" approach has most widely been used in identifying inefficiency effects. Under this approach, first, efficiency scores of individual production units are calculated using frontier analysis. Second, these efficiency scores are regressed against the variables which supposedly affect farm efficiency such as age and education of farmers, extension services, land tenure and credit

availability. The early literature which employed this approach includes Timmer (1971), Pitt and Lee (1981) and Kalirajan (1981). More recent applications of the approach are Ali and Flinn (1989), Kalirajan and Shand (1989) and Caves and Barton (1990). However, there are several drawbacks associated with this approach. First, there is ambiguity as to which inputs should be treated as first-stage regressors and which inputs should be treated as second-stage regressors. Second, the efficiency scores used as dependent variables in the second stage are bounded by one and zero. Therefore, a transformation of the dependent variable is necessary so as to avoid bias (Lovell, 1993). Third, and a major drawback associated with this approach, is the statistical inconsistence regarding the inefficiency effects. In the first stage, the inefficiency effects are assumed to be independently and identically distributed. But regressing them against some other explanatory variables in the second stage implies that those inefficiency effects are not identically distributed (Coelli, 1995b). Solutions to this problem came with Kumbhakar *et al.* (1991) and Reifschneider and Stevenson (1991) in the cross-sectional context. These authors modelled both frontier function and inefficiency effects as a single-stage problem and estimated them simultaneously. Battese and Coelli (1995b) applied a similar model to panel data. This model also enables one to identify efficiency change and technical change separately (In the context of the current study, more detailed discussion of this model is presented in Chapter 5.).

Duality

If one is willing to make an additional assumption of profit-maximisation, cost-minimisation or revenue-maximisation, then the dual approach is another possible extension to cross-sectional stochastic frontier analysis. The development of this approach is particularly motivated by the following reasons (Lovell, 1993):

1) it avoids the multi-collinearity problem which is common in the single-equation case when flexible functional forms (e.g., translog) are used;

2) it easily handles the multiple-output problem;

3) it can calculate overall economic efficiency and its components, i.e., technical and allocative efficiency.

3) it can calculate overall economic efficiency and its components, i.e., technical and allocative efficiency.

Depending on whether the firm is assumed to have cost-minimising, profit-maximising or revenue-maximising behaviour, stochastic cost, profit or revenue-functions can be constructed Because of its popularity in empirical applications, the stochastic frontier cost function case is considered here.

A stochastic frontier cost function can be written as (Greene, 1993):

$$C_i = C(Y_i, W_i, \alpha) + V_i + U_i, \qquad i = 1 \ldots N,$$

where C_i is the observed cost for the i-th firm, $C(.)$ is a suitable functional form describing production technology, W_i is the vector of input prices for the i-th firm, α is the vector of parameters to be estimated; V_i is assumed to be independent and identically distributed random error and U_i is a non-negative random variable associated with the cost inefficiency of production.

Two estimation methods, i.e. the extended COLS or MOLS method and the MLE method, have been suggested in the literature (Lovell, 1993). The dual systems approach has the advantage of explicitly accounting for allocative inefficiency, which is reflected in the error terms of the factor demand equations and it also treats output as exogenous and inputs as endogenous (Coelli, 1995b). However, as noted by Coelli (1995b), this approach has two serious drawbacks. First, estimation of the cost frontier requires input price data with sufficient variation across firms, but as the firms in the same industry tend to face similar prices, it is often difficult to obtain satisfactory input price data. Second, in the case of more flexible functional forms such as translog, which are not self-dual, the modelling of the relationships between the error terms of the cost function and the input-share equations is quite difficult. Detailed discussion on this issue is presented in Bauer (1990a).

Recent studies which use the dual approach include a stochastic cost frontier analysis by Bauer (1990b) for the case of U.S. banking; and stochastic profit

frontier analyses by Kumbhakar, Ghosh and McGuckin (1991) for the case of U.S. dairy farms, and by Ali and Flinn (1989) for the case of Pakistan's rice producers.

The literature in this area is still in its infancy (Lovell, 1993, p. 23). In the context of the current study, this alternative has not been considered a plausible option, primarily because the rational economic assumptions of profit-maximisation or cost-minimisation underlying the dual models are not relevant to the behaviour of centrally-planned state farms.

Shortcomings of stochastic frontiers

The stochastic frontier production function framework is, despite having advantages such as the accommodation of production noises and scope for the conduct of statistical tests on the results, subject to a number of shortcomings. First, unless additional assumptions such as profit maximisation or cost minimisation are made, the stochastic frontier production function cannot easily handle multiple-output cases. Exceptions include a few recent papers on accommodating multiple outputs in composed error models using distance functions (Grosskopf and Hayes, 1993; Coelli and Perelman, 1996). Modelling of a multiple-output case could be especially crucial in cases where making the assumption of jointness in production is not valid. For instance, if a production unit produces several outputs and the decisions related to the input allocation to each individual enterprise are inter-related, then the production analysis based on individual outputs will be biased.

Second, a parametric description of a production technology (i.e., specifying a functional form) unnecessarily restricts the production technology. Although more flexible functional forms (e.g., translog) reduce the restrictions on the production technology assumed in the Cobb-Douglas function (such as constant returns to scale and unitary elasticity of substitution) they come at the cost of increased estimational difficulty and reduced degrees of freedom.

Third, in the process of disentangling the inefficiency effects from stochastic elements in the error term, one must make an arbitrary distributional assumption

on the efficiency effects. However, more general distributional forms have been suggested for the error term to minimise possible biases that might have resulted from the assumptions of restricted distributional forms (Greene, 1993).

2.2.3 Non-parametric frontiers

2.2.3.1. Introduction

The non-parametric approach to efficiency is largely represented by the data envelopment analysis (DEA) method. As the title describes, this linear programming-based method envelops the observed data, constructing a frontier against which the distances of individual observations are measured. Compared to the parametric approach discussed earlier, this method has several attractive features. First, it does not parametrise the production technology while constructing the frontier, thus avoiding biases due to restricted forms. Second, it easily accommodates the multiple-output, multiple-input case without requiring additional behavioral assumptions like profit maximisation or cost minimisation. Furthermore, the fact that it employs linear programming techniques, which are widely used in economic analyses, may make it easier for people to conceptualise and utilise it. Finally, as the method itself is based on frontier notions, in a panel-data context, it enables researchers to calculate total factor productivity and its components without resorting to prices or any other aggregating elements entailing additional assumptions on the behaviour of decision-making units.

Since Farrell's (1957) work on frontier estimation using a piecewise-linear convex-hull approach, a number of papers using the linear programming framework have been published, including Bressler (1966), Boles (1966), Seitz (1966), Sitorus (1966), Aigner and Chu (1968) and Afriat (1972). However, the method became widely known following papers by Charnes, Cooper and Rhodes (1978, 1981). In their original work, Charnes, Cooper and Rhodes (1978) reformulated Farrell's efficiency measurement in a multiple-output case using linear programming techniques and introduced the term *data envelopment analysis*. Since then the DEA method has been used widely applied in many fields including management, economics and operational research. The development of

DEA has occurred in terms of both theory and practice. Comprehensive reviews of the methodology are found in Seiford and Thrall (1990), Ali and Seiford (1993) and Seiford (1996).

DEA models can be classified into groups based on (i) model orientations and (ii) assumptions related to returns to scale.

In terms of model orientations, the DEA models can be further classified into input-orientated, output-orientated or non-orientated models. The choice of one model over another depends on the particular problem in focus. If a firm is given a certain production target and is set to minimise the inputs, then an input-orientated model is appropriate. On the other hand, if a firm is asked to maximise output for a given set of inputs then an output-orientated model is the preferred option. However, if a firm can not be characterised by either of these options, then a non-orientated model can be chosen.

In terms of assumptions related to returns to scale, there are constant returns-to-scale DEA models (Charnes, Cooper and Rhodes, 1978) and variable returns-to-scale DEA models (Banker, Charnes and Cooper, 1984).

In the following section, the input-orientated constant and variable returns-to-scale DEA models are discussed.[3]

2.2.3.2. Input-orientated DEA models

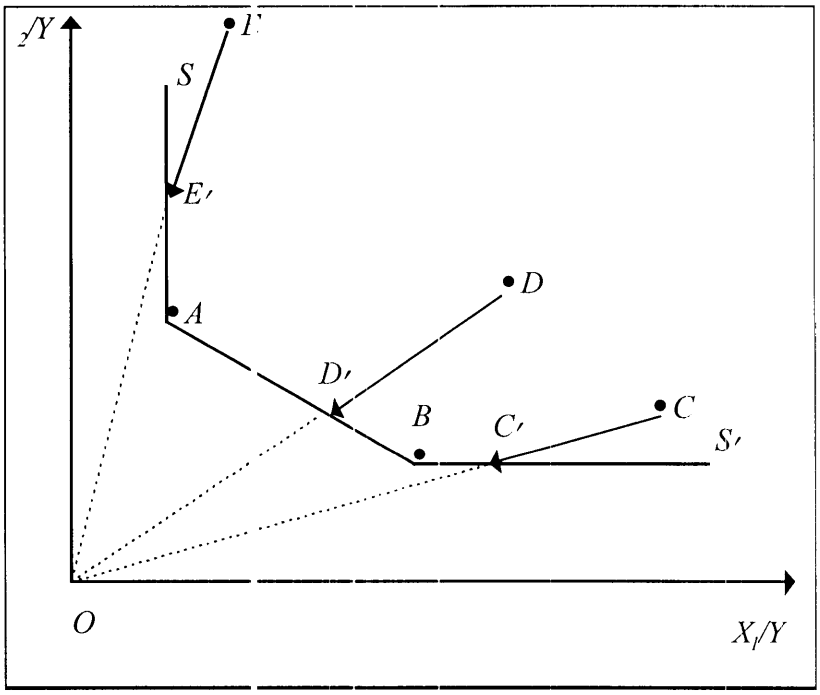
The input-orientated model for a two input (X_1, X_2) and one output (Y) case can be explained via Figure 2.3. The unit isoquant $SABS'$ is the lowermost frontier of input combinations for producing one unit of output.

The points A, B, C, D and E are five data points each representing a firm. While the points A and B are the most efficient firms through being located on the

[3] Note that output-orientated DEA models are used in the application in Chapter 7. Input-orientated DEA models are discussed below because of their prevalence in the early DEA literature.

frontier, points *C*, *D* and *E* are inefficient firms because of their deviation from the efficient frontier. According to Farrell's principles, all three inefficient firms can reduce their inputs equi-proportionately so that these inefficient firms are brought to new points *C'*, *D'*, and *E'*. The distance between their original points and projected points are interpreted as inefficiencies associated with those firms. While only *D'* is considered as a truly efficient point in a Koopmans' (1951) sense because there is no further way of reducing inputs for given output, *C'* and *E'* points are not truly efficient although they are on the isoquant boundary.

Figure 2.3 Input-orientated efficiency measurement in DEA context



For instance, there is still room to reduce X_1 input by the amount $C'B$ for firm C; and reduce X_2 input by the amount $E'A$ for firm E. So, only those points projected from the origin to the area within the points AB are efficient in both Farrell's and Koopmans' definition, while those projected outside that area are efficient only in Farrell's sense. There is ample literature on correcting these slacks, but all alternatives come at certain costs and none of them seems to have had overwhelming success (Lovell, 1993, p. 14). The usual way to deal with this

problem of slack is either to ignore it or to report it along with efficiency scores (Lovell, 1993).

As the number of inputs and outputs increases, it becomes impossible to illustrate the problem in simple two-dimensional geometric illustration. Only mathematical formulations can handle the cases with more than two inputs and outputs.

Input-orientated constant returns-to-scale DEA model

The original formulation of DEA (Charnes *et al.*, 1978) was input-orientated and assumed constant returns to scale. The mathematical formulation of this DEA model is discussed here.

Let us assume there are N firms to be evaluated, each having K inputs and M outputs. The i-th firm utilises input vector, x_i to produce output vector, y_i. The $K \times N$ input matrix, X, and the $M \times N$ output matrix, Y, represent the data of all N firms.

The linear programming formulation of the envelopment problem can be written as:

(2.17) $\min_{\theta,\lambda} \theta$

subject to:

(2.18) $-y_i + Y\lambda \geq 0$

(2.19) $\theta x_i - X\lambda \geq 0,$

(2.20) θ *free*, $\lambda \geq 0.$

The constraint (2.18) implies that the reference (frontier) firm should produce at least as much as firm i and the constraint (2.19) implies that the input use of firm i, after being adjusted by efficiency coefficients, should be at least as much as the input use of the reference firm. The optimal solution θ^*_i represents an efficiency score of firm i.

In this envelopment problem of input-orientated DEA, the performance of a firm is assessed by how much it can reduce its inputs equi-proportionally subject to the constraints imposed on them.

Relaxing constant returns to scale

Constant returns to scale (CRS) is too stringent an assumption in many cases. So, the next significant extension was a variable returns-to-scale (VRS) DEA model (Banker *et al.*, 1984). Also, non-increasing returns to scale (NIRS) DEA model can be calculated to permit the identification of scale economies of particular firms at certain input or output levels (Färe, Grosskopf and Lovell, 1985). The descriptions of the scale effects, in input-saving orientation, on the efficiency scores of individual firms are shown in Figure 2.4 below. The points *P, Q, R, S* are identified as individual firms. The ray, *OQ'*, passing through the origin, describes the constant returns-to-scale technology. Given this constant returns-to-scale technology, only firm *Q* is efficient as it lies on the frontier. The efficiency level of firm *Q* is measured as $X_Q'/X_Q = 1$. All other firms, *P, R, S* are inefficient as they deviate from the frontier and the efficiency levels are less than one. For instance, the efficiency level for firm *P* is calculated as $X_P'/X_P < 1$.

Following Charnes *et al.* (1984), by imposing the additional convexity constraint, $N1'\lambda = 1$ to equations (2.17) - (2.20), a VRS DEA model can be obtained, the envelopment problem of which is:

(2.21) $\min_{\theta,\lambda} \theta$

subject to:

(2.22) $-y_i + Y\lambda \geq 0$

(2.23) $\theta x_i - X\lambda \geq 0,$

(2.24) $N1'\lambda = 1,$

(2.25) θ free, $\lambda \geq 0.$

Also, following Charnes *et al.* (1984), the envelopment problem of DEA in the case of NIRS can be written by replacing the constraint (2.24) with $N1'\lambda \leq 1$ to obtain:

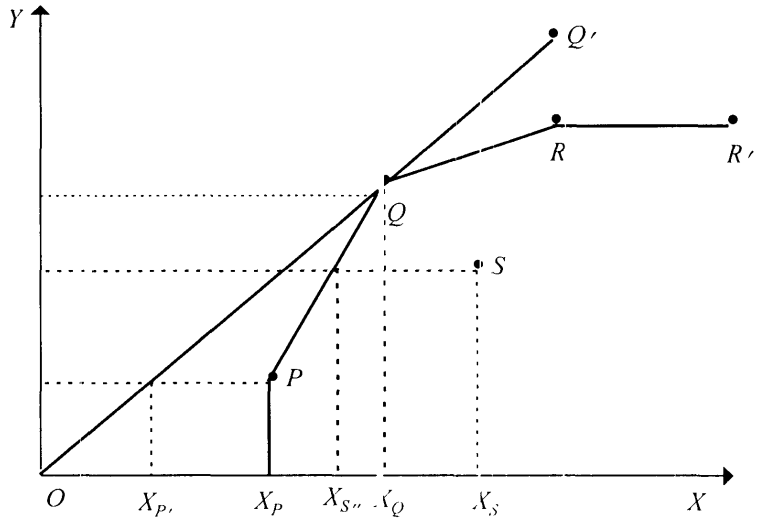
(2.26) $min_{\phi,\lambda}\ \theta$

subject to:

(2.27) $-y_i + Y\lambda \geq 0,$

(2..28) $\theta x_i - X\lambda \geq 0$

Figure 2.4 Scale effects of DEA



Note: The points OQQ' represent CRS DEA technology;
The points X_PPQRR' represent VRS DEA technology;
The points $OQRR'$ represent NIRS DEA technology.

(2.29) $N1'\lambda \leq 1,$

(2.30) $\theta free,\ \lambda \geq 0.$

where *N1* is a *Nx1* vector of ones.

As can be seen from Figure 2.4, the DEA model under VRS technology envelops the data in a much tighter way than DEA under CRS or NIRS technology, thus identifying fewer firms as being inefficient than the latter two.

Once the efficiency scores of individual firms are calculated under the two different technologies, i.e., CRS and VRS, the scale efficiencies of individual firms can be calculated. The scale efficiency (*SE*) of a firm *i*, is defined as the ratio of technical efficiency scores under CRS and VRS (Färe, Grosskopf and Lovell, 1994). That is:

$$(2.31) \qquad SE_i = TE_{i,crs}/TE_{i,vrs}$$

where $TE_{i,crs}$ is the technical efficiency score of firm *i* calculated under the CRS technology and $TE_{i,vrs}$ is the technical efficiency score of firm *i* calculated under the VRS technology.

Although the DEA model, as such, does not give any information about scale elasticities of production, by running a third DEA model under NIRS technology for the same data set and then by comparing the efficiency scores under all three technologies (CRS, VRS and NIRS), one can identify the scale regions at which individual firms are operating (Färe *et al.*, 1985). For instance, if the efficiency score of a VRS DEA model is equal to that of a NIRS DEA model, this implies that decreasing returns to scale is present. But if the efficiency score of a VRS DEA model is not equal to that of a NIRS DEA model, then increasing returns to scale is present. Furthermore, if the efficiency scores under CRS and VRS technology are equal, then constant returns to scale is applicable.

2.2.3.3. Extensions to basic DEA models

In the past two decades, a substantial number of extensions have been made to the DEA literature. As Seiford (1996) characterises it, the development of DEA has grown from the initial simple constant returns-to-scale model into a whole range of widely applied models covering economics, management science, operations research, public and non-profit organisations. Within the substantial number of

new developments in DEA modelling, the following papers can be singled out as major advances: Charnes *et al.* (1982, 1983) proposed a multiplicative DEA model where Cobb-Douglas type of production technology was described. The additive model of Charnes *et al.* (1985) suggested a Pareto-efficiency equivalent measurement of efficiency. Next, a cone-ratio model of Charnes *et al.* (1989) delivered an improvement in efficiency measurement for the case of a large number of firms. Banker and Morey (1986a,b) suggested ways of including non-discretionary and categorical variables into DEA. Incorporation of judgemental variables into DEA (Dyson and Thanassoulis, 1988) opened up another dimension into DEA modelling. The bridging between non-parametric efficiency measurement methods and the Malmquist index of productivity change (Färe, Grosskopf, Norris and Zhang, 1994) enables one to calculate productivity change without resorting to price information and furthermore, to decompose it (productivity change) into efficiency change and technical change. Relaxing of convexity constraints in the basic DEA model, as has been done by Petersen (1990), eliminates another unrealistic assumption from DEA modelling.

Despite all these attractive features and the advances made, DEA modelling can be criticised on several grounds. These include its failure to account for errors in the data (Lewin *et al.* 1982), and the sensitivity of efficiency results either to changes in variable selection (Epstein and Henderson, 1989) or to omitted variables (Ahn and Seiford, 1993). Due to their deterministic nature, most DEA models lack statistical testing although a few recent papers (Land *et al.*, 1993; Banker, 1996; Grosskopf, 1996; and Simar 1996) have attempted to address this issue. Above all, the issues of accommodating stochastic elements and introducing statistical properties have become the most critical and difficult task faced by DEA modellers (Seiford, 1996, p. 107).

2.3 Productivity Measurement

2.3.1 Total factor product vity (TFP)

The focus of the previous section was on discussion of the concepts and measurement of technical efficiency of a production unit, which is a major but only one component of a broader notion, productivity.

The productivity of a production unit is generally defined as "a ratio of output to input of a production unit" (Lovell, 1993, p. 3). If a production unit employs only one input to produce one output, then the calculation of the productivity of that unit is quite straightforward, output is divided by input. However, as in the real world, when the production process is characterised by multi-input and multi-output operation, productivity measurement becomes more complicated. In general there are two options for calculating the productivity of a production unit, viz., partial factor productivity (PFP) or total factor productivity (TFP). Partial factor productivity is calculated as a ratio of output (often in value terms) and one input. The typical examples are labour productivity and land productivity. This approach is particularly useful in cases where someone wants to assess the performance of a unit in relation to a single most important input. It is also very simple to calculate these partial indicators. However, this approach has several serious shortcomings in that a single partial-productivity measure will not give an overall picture of the performance of a production unit. Even if several partial-productivity measures are obtained, one for each input, the individual results may point in different directions, failing to give a consistent performance assessment. Moreover, PFP does not differentiate the effects of input substitutions or shifts in the production function (Germa, 1991, p. 2). Thus, partial productivity measures may lead to biased policy directives.

The second option is the calculation of total factor productivity. This option seeks to assess the performance of a production unit in terms of all useful inputs and outputs involved in the production process (Lovell, 1993). Grosskopf (1993, p. 162) defines TFP as "an index of output divided by an index of input usage".

The traditional TFP measurement methods, which do not use frontiers, fall into two sets of alternatives, growth accounting and econometric. The next section will discuss these two approaches briefly in a non-technical manner.

2.3.2 Traditional TFP measurement approaches

The growth accounting approach initially proposed by Solow (1957) seeks to disaggregate output growth into input growth and a residual which is identified as technical change. Then the technical change is considered as TFP change. On the production function surface, input growth is seen as movement along the production function, and technical change as shifts in the production function (Solow, 1957). The actual measurement of TFP was done using index number theory.

Index number theory suggests various alternative ways of aggregating outputs and inputs; among the most popular indices are the arithmetic, geometric and the Divisia index. These indices differ one from another by the weights chosen for the aggregation of outputs and inputs, which are crucial in that each of them implicitly assumes a different functional form of production. Arithmetic aggregation methods which include the Laspeyres and Paasche indices, implicitly assume a linear (Diewert, 1974) or Leontief (Diewert, 1976) functional form. A geometric index assumes a Cobb-Douglas production function (Diewert, 1976), and the Tornqvist-Theil index (a discrete approximation of the Divisia index) assumes a homogeneous translog production function (Christensen, 1975). More detailed discussions of index numbers can be found in Allen (1975), Diewert (1974, 1976, 1981) and Selvanathan and Rao (1994).

Despite its simplicity of calculation, less demand on data and a strong theoretical economic basis, the index number approach has several weaknesses. In the aggregation process index numbers use either (input and output) prices or value shares, thus an (implicit) assumption of competitive input and output markets is made. Furthermore, by assuming that firms are operating on the frontier, the index number approach ignores the efficiency component while calculating TFP. Next, the implicit assumption of various production functions imposes restrictions on

production technology. Finally, statistical tests cannot be conducted on the result, so one can not be sure how reliable is the outcome.

The econometric approach to TFP measurement is based on the estimation of production functions explicitly assuming a functional form. The technical change here is defined as any shift in the production function over time, and is often represented by including a time-trend variable into the production function. Taking the first derivative of the estimated production function with respect to the time variable will provide an estimate of the technical change. This approach has several advantages over index numbers. First, there is no need to aggregate inputs using price weights. Second, various statistical tests can be done both on the adequacy of functional forms and the reliability of the parameter estimates. Third, in the production function framework (though not in the case of profit or cost functions), the behavioural assumptions related to markets, such as competitive input and output markets, do not have to be made. Lastly, a number of important economic criteria can be tested such as non-neutral vs. neutral and disembodied vs. embodied technical change and scale economies of production. However, this (econometric) approach has a number of shortcomings too. For instance, in cases of more general functional forms and small sample sizes, problems associated with degrees of freedom often arise. Also, technical change is often calculated as identical for all firms in a particular period (if neutral technical change is imposed). Last, this approach, like the index number approach, ignores the efficiency component of TFP by assuming that all firms are technically efficient.

2.3.3 TFP measurement approaches based on frontier techniques

This group of approaches based on frontier techniques defines the TFP in a different, broader term. For instance, as Lovell (1993, p. 3) defines it:

> *Productivity varies due to the differences in production technology, differences in the efficiency of the production process, and the differences in the environment in which production occurs.*

Grosskopf (1993, p. 160) defines TFP growth in an inter-temporal context as:

> *"... the net change in output due to change in efficiency and technical change, where the former is understood to be the change in how far an observation is from the frontier of technology and the latter is understood to be shifts in the production frontier."*

Ignoring one or the other component while measuring productivity levels will not only bring biased results, but will also be not satisfactory in terms of the policy implications drawn from it. This is because the policy implications drawn from each component of the TFP change differ from one another significantly.

For instance, while the improvement in the efficiency component would suggest (Nishimizu and Page, 1982, p. 921) "learning by doing, diffusion of new technological knowledge, improved managerial skills as well as short-run adjustment to shocks external to enterprises", the technical progress would be due to "... change in the best practice frontier.", in other words, the investment into and importation of new technology, etc.

It was Nishimizu and Page (1982) who first decomposed the TFP effects into efficiency and technical change using the deterministic, parametric estimation techniques of Aigner and Chu (1968) and Timmer (1971). After estimating individual efficiencies and the rates of technical change, they added these two components to obtain TFP changes.

The TFP measurement literature which explicitly includes efficiency changes can be broadly classified into those approaches based on econometric methods and those on mathematical programming frontier techniques. This classification is consistent with the frontier-based efficiency literature discussed earlier.

In the econometric approach, after estimating individual efficiencies and the rate of technical change, the TFP is defined as the sum of the percentage rates of the two. It is defined in a similar way as Nishimizu and Page (1982) did, except that in the latter case a deterministic frontier production function was estimated using linear programming techniques. The papers that accomplished the econometric decomposition of TFP changes are Fecher and Perelman (1989) and Perelman

(1995) in the stochastic frontier production function case; and Bauer (1990b) in the stochastic frontier cost function case.

On the other hand, the mathematical programming literature uses linear programming techniques to first calculate the efficiencies of individual firms in each year. Then, in a successive two year period context (i.e., t and $t+1$), the shift in frontier functions between the two years is calculated separately and the two elements (efficiency change and technical change) are aggregated to derive TFP change. This body of work includes TFP measurement based on estimation of a deterministic parametric production function (Nishimizu and Page, 1982) and Malmquist index of TFP change using DEA-based methods (Färe *et al.* 1989; Färe, Grosskopf, Norris and Zhang, 1994).

2.3.3.1. Malmquist index of TFP change

The Malmquist index has recently been used in the context of efficiency analysis literature (Färe *et al.* 1989, Färe, Grosskopf, Norris and Zhang, 1994). This approach has several advantages over conventional indices like the Fisher ideal index or the Tornqvist index:

1) it does not make any behavioural assumption about the production unit, such as profit maximisation or cost minimisation;

2) it does not need any additional information, such as prices or input shares; it is only a quantity index;

3) it decomposes the TFP change into efficiency change and technical change, thus expanding the traditional index numbers approach by another policy dimension.

The following section discusses the Malmquist index of productivity in detail.

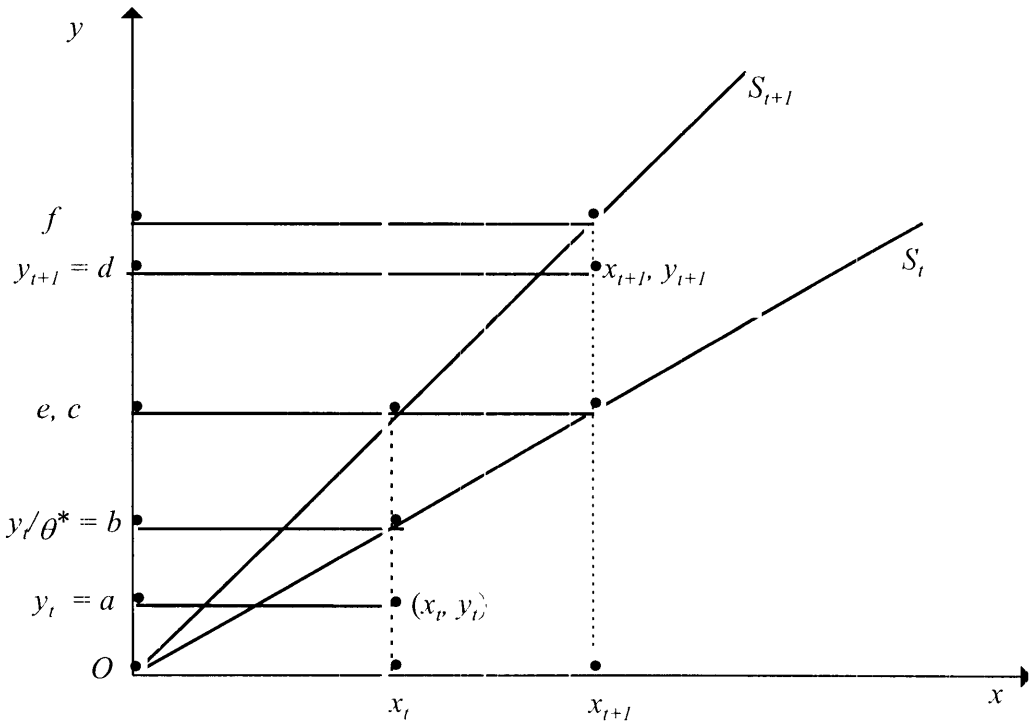
Caves *et al.* (1982b) proposed a productivity index using distance functions and termed it the "Malmquist productivity index" after Sten Malmquist who first used distance functions as a quantity index in consumer theory. However, it was Färe *et al.* (1989) who utilised the fact that the distance functions, the building

components of Malmquist productivity index, are the reciprocals of Farrell's radial efficiency measures, thus bridging the efficiency and index number literatures. Hence, Färe *et al.* (1989) have calculated the Farrell efficiency measures using non-parametric DEA methods. After Farrell efficiency measures were calculated (including inter-period measures), the reciprocals of these efficiency measures (as distance functions), were used to calculate the Malmquist productivity index. Furthermore, in Färe *et al.* (1989), the Malmquist productivity index was decomposed into efficiency change, technical change and scale effects. This Malmquist index approach was a significant contribution of the efficiency measurement literature to the productivity measurement literature, because in assessing a producer's performance, one does not have to resort to prices, costs or profit measures any more (Grosskopf, 1993).

There are thus two major advantages stemming from these new approaches over the traditional index numbers. First, as no price data or value share data are required (which is a must in conventional index number or growth accounting approaches) to calculate the Malmquist productivity index, it effectively avoids any biases resulting from price distortions. Second, as it introduces a new efficiency element into productivity measurement, it brings a new policy dimension into empirical productivity analysis, not to mention its flexibility in describing multiple-input, multiple-output technology.

As shown in Figure 2.5, the measurement of the Malmquist index of productivity change of a production unit over two time periods (t and $t+1$) can be illustrated in a simple two-dimensional (y-output, x-input) diagram as follows: The points (x_t, y_t) and (x_{t+1}, y_{t+1}) are the actual input-output combinations of a firm in two time periods t and $t+1$.

Figure 2.5 Malmquist ndex (output-based) of productivity change



Source: Färe, Grosskopf, Norris and Zhang (1994).

The corresponding output frontiers for the case of constant returns-to-scale (because the frontiers are ra/s which pass through the origin) technology in the two time periods, *t* and *t+1* are S_t and S_{t+1}. The shortfalls of output in each time period, *t* and *t+1*, to their respective frontiers (S_t and S_{t+1}) as inefficiencies are illustrated by distances *Oa/Ob*. The technical change (technical progress in this case) is expressed as a shift between S_t and S_{t+1}. However, depending on which point of input use one chooses, the two rates of technical changes can be calculated: *Oc/Ob* (for the input level of period *t*) and *Of/Oe* (for the input level of period *t+1*). To avoid ambiguity, the overall technical change of that particular firm from *t* to *t+1* is calculated as the geometric mean of the two technical changes, each evaluated at different input levels, i.e., the input levels of the period *t* and *t+1* (Färe, Grosskopf, Norris and Zhang, 1994).

In terms of output changes on the y axis, the Malmquist index of productivity change can be written the following way:

$$M_o(x_{t+1}, y_{t+1}, x_t, y_t) = \left(\frac{Od}{Of}\right)\left(\frac{Ob}{Oa}\right)\left[\left(\frac{Of}{Oe}\right)\left(\frac{Oc}{Ob}\right)\right]^{1/2}. \quad (2.32)$$

While the first expression, $\left(\frac{Od}{Of}\right)\left(\frac{Ob}{Oa}\right)$ reflects the changes in efficiency from t to $t+1$,

the second expression, $\left[\left(\frac{Of}{Oe}\right)\left(\frac{Oc}{Ob}\right)\right]^{1/2}$ reflects the technical change from t to $t+1$.

Detailed discussion of the Malmquist index of productivity change in terms of distance functions is given in Chapter 7 in the context of the current study.

2.4 Summary and Conclusions

The main purpose of this chapter was to review the literature on efficiency and productivity measurement and identify the suitable analytical options for the current study. The chapter was divided into two parts. The first part deals with alternative efficiency measurement techniques, whereas the second part deals with alternative productivity measurement techniques. All commonly used efficiency measurement techniques currently available are based on the basic concepts of Farrell's efficiency measurement in the context of a frontier technology. However, the individual techniques differ one from another in the way the frontier is constructed and the efficiencies of individual firms are measured.

The existing efficiency measurement techniques may be broadly divided into parametric and non-parametric methods. Under the parametric option, the frontier production technology is usually estimated using econometric techniques. The development of frontier function approach has gone through two major distinctive stages, i.e., deterministic and stochastic frontier functions. Under the deterministic

frontier approach, any deviation of a firm from frontier is attributed to inefficiency, whereas under stochastic frontier approach, any deviation of a firm form the frontier is attributed to a stochastic element (such as weather shock and data errors) and an inefficiency element associated with human factors (i.e. management). The advantage of the stochastic frontier approach is that it takes into account the reality of production which is subject to stochastic elements often outside someone's control. Since the initial stochastic frontier formulation (Aigner *et al.*, 1977; Meeusen and van den Broeck, 1977), a substantial development took place in this field, including relaxing the distributional assumptions regarding the one-sided error term, explaining the inefficiency variation in terms of socio-economic characteristics and extensions to accommodate panel data.

Amongst other advantages, panel data model allows one to investigate technical change and efficiency change over time, which in turn permits one to calculate TFP change, without needing to resort to price or value share information for aggregation purposes as traditionally was done. Despite these important developments, the stochastic frontier framework has several shortcomings which may affect the estimation of firm efficiency. In particular, the need to parametrise the production technology and to make an assumption regarding the distributional pattern for the inefficiency term are fairly restrictive.

The data envelopment analysis (DEA) method, which is a non-parametric efficiency measurement technique based on linear programming methods may be able to avoid the shortcomings associated with stochastic frontier function as mentioned above. DEA does not parametrise the production technology and does not make any assumption about the distribution of inefficiency term. One of the most important recent developments in DEA modelling is its extension to allow one to calculate TFP using the Malmquist index of productivity change. The Malmquist index has certain advantages over traditional TFP indices, such as Tornquist and Fischer indices. For instance, it does need any price or value share information to calculate TFP change, which in turn removes the competitive market assumption from the model. Next, the Malmquist index approach does not assume that firms are efficient. Thus, under the Malmquist index, not only can

TFP change can be calculated, but it also can be decomposed into efficiency change, technical change and scale effects (Färe, Grosskopf, Norris and Zhang, 1994). This approach is not, however, without shortcoming. For instance, it is more demanding in terms of data compared to traditional index numbers, because it requires firm-level data. Also, one can not conduct traditional hypothesis tests to assess the reliability of the results. Perhaps the most serious shortcoming of the method is its inability to accommodate noise. Despite some recent work (Land *et al.*, 1993; Banker, 1996; Grosskopf, 1996; Simar, 1996), this method is still considered primarily as deterministic. This was perhaps the main reason why very few applications of DEA-based methods were available in the field of agriculture, where the production process is often significantly influenced by stochastic elements. In contrast, there have been a substantial number of applications of the stochastic frontier function approach in agriculture (Battese, 1992; Bravo-Ureta and Pinheiro, 1993; Coelli, 1995b). In view of the fact that the Mongolian crop farming is characterised by substantial stochastic elements (both weather and input supply-related), the stochastic frontier function approach is selected as the preferred option. However, this study will also involve the application of DEA-based models to test the robustness of the stochastic frontier analysis results.