The cost of banking services and the technical efficiency of a South African bank

*C van Heerden and G van der Westhuizen**

Abstract

Does the performance of one of the four largest banks in South Africa justify the customers' complaints about the higher bank fees? Data Envelopment Analysis (DEA) was used to estimate the technical efficiency and returns to scale of one of the largest banks in South Africa. The intermediation approach was applied to classify the inputs and outputs and the analyses were conducted with both input- and output- orientation under variable returns to scale. Returns to scale efficiency and technical efficiency for 37 districts over a period of 22 months were estimated. The analyses indicated that 19 districts out of the 37 districts were never fully technically efficient during the 22 months (input- and output-orientated). It appears that customers' complaints about high service fees are justified.

Keywords: Banking services, Bank efficiency, Technical efficiency, Data Envelopment Analysis

Disciplines: Economics, econometrics, mathematics, statistics, management sciences.

1. Problem statement and motivation for this paper

The South African financial sector is currently being dominated by four large banks, namely First National Bank, Nedbank, Standard Bank and ABSA. According to Okeahalam (2006:105) these four banks are controlling over 85% of total deposits and assets in South Africa.

Over the past 10 years bank customers have experienced relative large increases in bank service fees, and customers would like to know whether these increases are justified due to higher production costs or is it a matter of banks being inefficient. According to Capgemini *et*

* Prof. Gert J van der WSesthuizen is professor in Economics in the School of Economic Sciences at the Vaal Triangle Campus of North-West University. Mr Chris van Heerden is a recent graduate of North-West University.

al. (2007) South Africans pay more to own a cheque account than anyone else in the rest of the world. South African cheque account holders are paying, on average, R1 863 a year in fees while the global average is R732.07, the report said. Cheque account holders pay more for payment services than consumers in almost all other countries paid for all banking services in total.

The lack of providing financial services to everyone (Hawkins, 2004:199); the inability to introduce new financial products (Akinboade & Makina, 2006:116); the operating costs outgrowing the bank's income during 2000 (Hawkins, 2004:196); increasing staff costs (Banking Supervision, 2002:50); savings accounts being costly (Hawkins, 2004:200) and customers not paying fair prices for financial services (Hawkins, 2004:197) led to the debate about how efficient the South African banks really are?

Exhibit 1 indicates the efficiency of South African banks over the period 1997 to 2002. Since the year 2000, there has been a continuous increase in the cost-to-income ratio of the banks, meaning that there has been a decline in the efficiency of banks since 2000. The introduction of new enhanced service facilities, like ATM's and internet banking, and the shedding of jobs in the financial sector since 1997, may be an indication of why the operating cost-to-income ratio have been rising above the international benchmark of 60% (Hawkins, 2004:192, 196, 200).

Exhibit 1: Efficiency of South African banks[1]
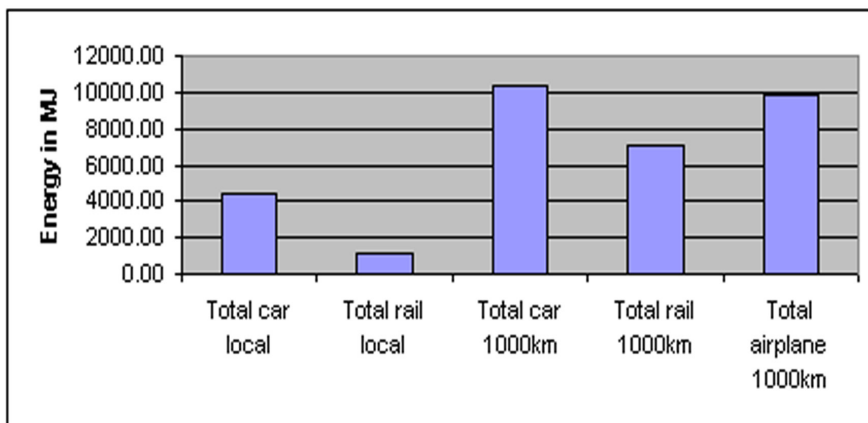
1 . Source: Bank Supervision (various years).

One can estimate the functionality or performance of a bank by estimating the efficiency of the bank and one possibility is to measure it in terms of the intermediation cost between the investor and the saver (Hawkins, 2004:183). Technical efficiency can be measured by expressing the operating expenses as a percentage of the total income (cost-to-income ratio) (Hawkins, 2004:196). Data also showed that operating expenses have outgrown the industry income. Thus the current cost-to-income ratio may have a negative impact on the consumer, with the fact that bank interest margins have been increasing to 3.8% on average, with non-interest income increasing to about 50% (Hawkins, 2004:197).

This paper attempts to determine whether customer complaints about bank costs are justified by estimating the performance of one of the four largest banks in South Africa. Efficiency may lead to a reduction in costs, while inefficiency may lead to an increase in costs.

2. Background

Greater competition in the domestic and international markets and the fundamental structural changes forced organizations to be more flexible and productive, focusing more on serving the customer (Hartle, 1997:45). Kimball (1997:24) stated that the greater specialization and focus in commercial banks were permitted by the 30 years of evolution in bank structures and management. During the 1980s organizations became more performance-orientated. An era of 'management by objectives' began (Hartle, 1997:46).

Mester (2003:3) stated that the organization's performance should be measured against an objective, because without an objective the company will have no criterion on choosing investment strategies and projects. 'The main purpose of performance measurement is to align the goals of individual employees and the bank as a whole' (McDonell & Rubin, 1991:56). Mester (2003:3) stated that efficiency is a measure of the deviation between actual performance and desired performance. For a bank to achieve outstanding performance means that it must achieve in maximizing the shareholders' wealth, thus the market value of a firm's common stock (Mester, 2003:3). How to measure performance still depends on each individual organization, because of different situations and circumstances (Berry *et al.,* 2005:94).

Most bank managers quote Return On Equity (ROE) or Return On Assets (ROA) or other growth rates when asked about the performance of the bank (Avkiran, 1997:224). ROA measures net income per dollar of average assets owned during the period (Koch and Macdonald, 2003:112).. However, ROA provides no direct information concerning how or which of the bank's activities contribute to the creation of shareholder value. It ignores other performance benchmarks that customer-focused managers must consider to identify the best strategies for the future (Koch & MacDonald, 2003:170).

ROE measures the percentage return on each dollar of stockholder's equity (Koch and Macdonald, 2003:112). The drawback of accounting ROE and ROA measures are that they do not include any risk adjustments (Bessis, 2005:10). They can vary substantially and will be highly dependent on the industry; this is why it is best to compare them against the ROE and the ROA values of a similar company. This drawback is the origin of the concept of risk-adjusted performance measures. This is a reason for moving to 'economic values', 'mark-to-market' or 'mark-to-model' values, because these are both risk- and revenue-adjusted (Bessis, 2005:10).

Kaplan and Norton (2001a:3), De Young (1997:21), Eccles (1991:132), Davenport and Sherman (1987:35), Chelst *et al.* (1988:7), Smith and Schweikart (1992:54), Avkiran (1997:224), Clark (1997:25) and Kaplan and Norton (1996a:25) stated that financial measures have flaws and should not be used as the only measurement of bank performance. Financial measures should be used as an additional component to overall performance measurement. Berry *et al.* (2005:93-94) stated that no single measure provides consistent evidence of the correlation between stockholders' satisfaction and firm performance. Verma (1992:279) stated also that traditional measures of profit were no longer appropriate for the motivation of measuring bank performance.

Kaplan and Norton (2001:3) stated that performance indicators should be a combination of financial and non-financial measures. However the Balanced Scorecard (BSC), as the common non-financial measure, has some of the following limitations. The BSC will be useless if the included non-financial measures are not linked to the firm's strategic objectives (Kaplan & Norton, 1996:55-56, 77); and the BSC is weak if too many performance indicators are included (Gering & Rosmarin, 2002:18-19).

Gardner and Mills (1994:667) stated that most performance evaluators use accounting and other data to calculate financial conditions. During the research study for this paper it was found that the required data is not always available or the required financial statements are impossible to obtain. To overcome these limitations relative bank efficiency measurement was used. The next section discusses the concept of relative bank efficiency.

3. Relative bank efficiency

When economists point to efficiency they focus on scale, scope and X-efficiency, which are different aspects of performance (Mester, 2003:2). Scale efficiency determines if an organization is providing the most cost efficient level of outputs (Mester, 1994:3). Scope efficiency refers to whether a firm is producing the most cost efficient combination of products (Mester, 1994:4). X-efficiency determines if an organization is using its inputs, like capital and labour, in a cost effective way to produce its outputs as cheap as possible (Mester,

1994:5). Firms that exhibit X-inefficiency are either using the wrong combination of inputs to produce outputs (allocative inefficiency), or wasting their inputs (technical inefficiency), or both (Mester, 2003:3). X-efficiency can therefore be divided into two groups, technical efficiency and allocative efficiency. Thanassoulis (1999:4) stated that efficiency of any firm consisted of two components, production efficiency and intermediation efficiency. Production and intermediation efficiency indicate technical and allocative efficiency and is also known as the sum of total or economic efficiency (Coelli *et al.,* 1998:5).

To explain technical and allocative efficiency in more detail consider Figure 1. In Figure 1 the two inputs X1 and X2 are utilized to produce a single output y, under the assumption of constant returns to scale (Coelli *et al.,* 1998:134). Curve SS' is the unit isoquant, which indicate all possible relative input combinations that yield maximum outputs. Any point on this isoquant is technically efficient, where it is not possible to reduce the two inputs proportionately and still produce the same level of output. Firms Q and Q' are both technically efficient, while firm P is technically inefficient (Farrel, 1957:254).

Figure 1:   Allocative efficiency and technical efficiency[2]
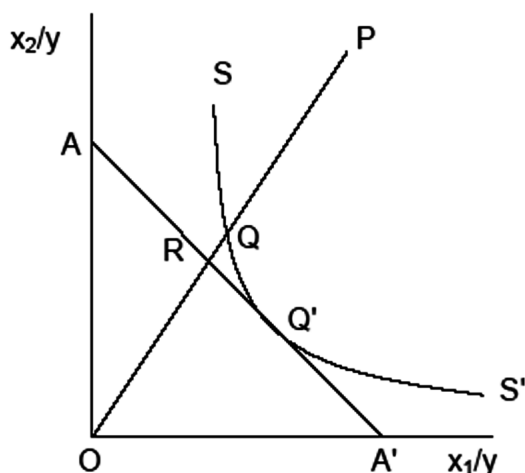
A firm that is operating at point P has technical efficiency, described by the ratio, which is the proportional reduction in all inputs without any reduction in output. This ratio has a value that lies between zero

2 . Source: Farrell (1957:135) & Coelli *et al.* (1998:135).

(total inefficient) and one (total efficient) (Coelli *et al.,* 1998:134-135). **L**ine AA' is an isocost line that indicate all possible input combinations that have the same cost. The allocatively efficient point, that minimizes cost, is the tangent-point between the isocost line AA' and the isoquant SS'. The optimal production point is at point Q'. The cost of inputs used at point Q' and point R is the same and is less than the cost of inputs used at points Q and P. Point Q is technically efficient, because it lies on the efficient isoquant (Coelli *et al.,* 1998:135; Farrell, 1957:135-136). Point Q is not on the isocost line and is therefore all**ocatively inefficient.**

Allocative efficiency of the firm operating at point P is defined to be the ratio, where the distance RQ is the reduction in production costs which would occur if production occurred at Q'. This ratio also has a value between zero and one. Economic efficiency is the sum of technical efficiency and allocative efficiency. Firm Q' is the only firm that is fully cost efficient (Coelli *et al.,* 1998:135-136; Farrel, 1957: 136). Efficiency measurement is one aspect of firm performance, which is measured with respect to an objective. The firm's objective can be measured with respect to the maximization of output, the maximization of profits or the minimization of costs (Mester, 2003:2).

Mester (1994:8) stated that it was recognized that there might have been unpredicted factors that affected a bank's cost over the period when data were collected. Some of these unpredicted factors included the possibility that the data are measured with error. More focus was therefore placed upon the degree of X-inefficiency in banking. The technique required was to estimate the best-practice cost function that predicted the cost function of banks that were X-efficient (Mester, 1994:9). The required technique also had to measure the degree of inefficiency relative to the best-practice, chosen as the appropriate benchmark. One of the common methodologies was therefore developed, namely the Data Envelopment Analysis (DEA) (Mester, 1994:9). Avkiran (1997:232) stated the main steps that should be followed in developing a performance model are as follows:

- Review the corporate objectives and strategies;

- Define branch performance in terms of corporate objectives. Identify those performance variables that are critical to the bank's success. Reflect the mix of personal banking business at branches;

- Identify the potential variables associated with branch performance. Identify those potential variables controllable by management;

- Determine the sources of data for performance and potential variables;

- Develop multivariate measures for data collection;

- Analyze data through multiple regression and other techniques. Determining the sets of potential variables explaining each of the

performance variables; and

- Examine the results of the analysis and raising performance through reconfiguring branches.

As stated earlier, the problem with measuring bank efficiency is data availability. The databases of banks accommodate only accounting procedures and not combined analyses of operational, marketing and financial data. DEA was found to be an appropriate method for overcoming these problems (Vassiloglou & Giokas, 1990:591-592). The reason is the modelling flexibility of the DEA and its ability to address qualitative and quantitative data, as well as non-discretionary and discretionary inputs (Golany & Storbeck, 1999:15). The DEA will now be discussed in the following section.

4. Data Envelopment Analysis (DEA)

'DEA involves solving linear programming problems that generate a non-parametric, piecewise linear convex frontier that envelops the input and output data relative to which cost is minimized' (Färe *et al.*, 1985:193). Coelli *et al.* (1998:140) stated that the DEA involves the use of linear programming methods to construct a non-parametric piece wise frontier across the data. This frontier is therefore used to measure the relative efficiency or productivity in terms of the inputs and outputs selected by the organization (Avkiran, 1999:206). This technique is called the Data Envelopment Analysis because the data on the best-practice banks cover or envelop the data completely from the rest of the banks in the sample (Avkiran, 1999:207). According to Sherman and Ladino (1995:63) DEA determines the following:

- the best-practice and most productive group of service units;

- the less-productive service units compared to the best-practice units;

- the amount of excess resources used by each of the less-productive units;

- the amount of excess capacity or ability to increase service outputs in less-productive units without utilizing added resources; and

- the set of best-practice service units most similar to the less-productive units.

DEA is also used because it lends itself more easily to multiple output analysis. DEA is useful in cases where the behavioural objective of the organization is clouded by, for example, government regulations or operating constraints (Van der Westhuizen, 2006:4). Sherman and Ladino (1995:61-62) stated that the DEA is a powerful tool for the service businesses, because it helps the organization to identify the best-practices in complex service operations. This includes service operations that are too complex for traditional analytical techniques

and observations (Sherman & Ladino, 1995:62).

The following graph is being used to illustrate DEA. Figure 2 shows a DEA model with a solid line connecting the efficient decision making units (DMUs) L, M and N that represent achieved efficiency. For example, DMU, K is classified as inefficient and needs to move to K' on the frontier before it can be classified as efficient. Unit K will be compared to units M and N on the efficient frontier with the calculation of its efficiency score (Avkiran, 1999:207). In comparing one bank's efficiency to that of another bank, one must compare each bank's cost of producing the same outputs (Mester, 1994:7).

Golany and Storbeck (1999:16) stated that the following three steps must be followed in order to construct a DEA model. Firstly, the selection of a decision-making unit (DMU) which means choosing the best-practice bank that will be used as a benchmark. Secondly, choosing the appropriate input and output combination and thirdly, to choosing a mathematical formulation for the DEA model.

Figure 2: A DEA model showing an efficiency frontier[3]
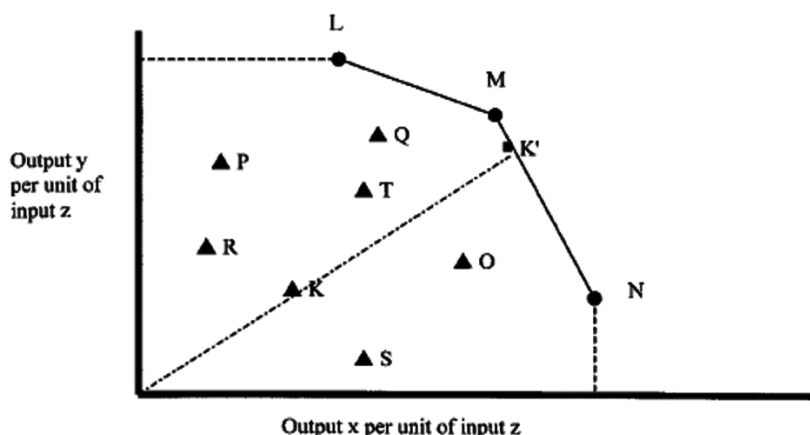
5. The model

DEA was used to estimate the efficiency of one of the largest banks in South Africa. The lack of data about each individual branch led to the use of districts. Data collected included 37 districts for 22 months covering the period July 2005 to April 2007. The DEAP (version 2.1) program developed by Coelli (1998) was used. The following DEA model developed by Charnes *et al.* (1978:430) was used:

---

3. Source: Avkiran (1999:207).

Objective function

$$\max E_o = \frac{\sum_{i=1}^{k} u_i \psi_{io}}{\sum_{j=1}^{m} v_j x_{jo}} \quad , \tag{1}$$

where

o = the region being assessed from the set of r = 1, 2, ..., n districts;

k = the number of outputs at the districts;

m = the number of inputs at the districts;

ψ_{ir} = observed output i at district r;

x_{jr} = observed input j at district r.

Constraints

$$\frac{\sum_{j=1}^{m} u_i \psi_{ir}}{\sum_{j=1}^{m} v_j x_{jr}} \quad > 1 \; r = 1, ..., \boldsymbol{n} \tag{2}$$

$$u_i , v_j > 0, i = 1, ..., k, j = 1, ..., m \tag{3}$$

The above analysis is performed repetitively, with each district in the objective function, producing efficiency ratings for each of the n districts. The solution sought is the set of (u_i, v_i) values that maximise the efficiency ratio E_o of the district being rated, without resulting in an output/input ratio 1 when applied to each of the other districts in the data set. Equation 1 can be interpreted as

Maximise $\frac{output\ index}{input\ index}$, equation 2 as a boundary constraint and equation 3 as a non-negativity constraint.

The intermediation approach was chosen to identify the inputs and outputs used. To define inputs and outputs in the banking industry may be the greatest problem of efficiency measurement (Mlima & Hjalmarsson, 2002:12). With the multi-product nature of a banking firm there is still no agreement as to the definition and measurement of banks' inputs and outputs (Girardone *et al.*, 2004:217). Berger and Humphrey (1997:34) pointed out that the intermediation approach is the most appropriate approach for evaluating the entire bank. The reason is because this approach includes interest expense, which accounts for up to two-thirds of total costs. Molyneux *et al.* (1996:152) and Mester (1996:1033-1034) stated that the intermediation approach is used because it views financial institutions as mediators between the supply of and the demand for funds. Elyasiani and Mehdian (1990:543) also stated that the intermediation approach is preferred

because the quality of data benefits the intermediation approach.

An ongoing debate is about deposits, whether deposits should by classified as an input or an output (Resti, 1997:224). Elyasiani and Mehdian (1990:543) stated that deposits are better categorized as inputs than as outputs, because banks buy deposits rather than sell deposits (Elyasiani & Mehdian, 1990:543).

With the available data the following inputs and outputs were used to analyze the 37 districts and the 10 provinces:

Inputs	Outputs
- Deposits - Interest expenses - Non-interest expenses	- Total loans - Interest income - Non-interest income

Interest expenses include cost of funds (COF) charges, which are the interest the branch must pay for loans. Non-interest expenses include operating expenses, labour and fixed asset costs. Operating expenses include maintenance costs and service level agreement (SLA) expenditures. SLA expenditures are for example costs for moving money from one branch to another. Total loans include mortgage, agricultural, business, commercial and consumer loans. Non-interest income includes sales commission and other non-interest income.

Yue (1992:36) and Kao and Liu (2004:2355) included deposits and non-interest expenses as inputs. Rangan *et al.* (1988:172) included labour as an input. Yue (1992:36) stated that non-interest expenses include taxes, salary expenses associated with premises and fixed assets and other expenses. Yue (1992:36) also stated that these measures represent the bank's capital, labour and operational cost. Berger and Humphrey (1997:34) and Elyasiani and Mehdian (1990:543) stated that to leave interest expenses out of the model can distort the empirical results, because it accounts for up to two-thirds of total costs. Both Kao and Liu (2004:2355), Wheelock and Wilson (1995:693), Favero and Papi (1995:390), Kaparakis *et al.* (1994:884), Rangan *et al.* (1988:172) and Yue (1992:36) used total loans as an output under the intermediation approach. Yue (1992:36) and Kao and Liu (2004:2355) included interest income as an output. Kaparakis *et al.* (1994:884), Favero and Papi (1995:390), Yue (1992:36) and Kao and Liu (2004:2355) included non-interest income as an output.

The final step is to formulate the mathematical formula of the DEA model. Input-orientated, output-orientated and variable returns to scale were used to formulate the DEA model. Input-orientated characterizes the production technology of the organization, for producing a given output mix with the minimum inputs (Coelli *et al.*, 1998:62). Output-orientated characterizes an organization in pursuit

of producing the maximum output bundle with the given inputs mix (Coelli *et al.*, 1998:62).

Constant returns to scale will not function under conditions such as imperfect competition and constraints on finance. Firms may therefore be unable to operate at an optimal scale (Coelli *et al.*, 1998:150). Banker *et al.* (1984:1086-1088) introduced the variable returns to scale (VRS) as the answer to this problem. VRS is the disproportional increase or decrease in outputs when the inputs are increased (Avkiran, 1999:211). The best approach may be to run the DEA model under constant returns to scale (CRS) and variable returns to scale (VRS) separately and then to compare the efficiency scores (Coelli *et al.*, 1998:150). The CRS efficiency score represents technical efficiency that measures the inefficiencies due to the size of operations and the organization's input/output combination. The VRS efficiency score represents pure technical efficiency that measure efficiency without scale efficiency (Avkiran, 1999:211).

6. Results

Table 1 exhibits the technical efficiency estimates and the returns to scale for the 37 districts under both the input- and output-orientated approach. A DEA analysis generates efficiency estimates that range from one (fully efficient) to zero (fully inefficient) (Coelli *et al.*, 1998:134-135). The technical efficiency estimates contain the minimum, maximum and mean values for each district.

Under the input-orientated approach district 26 (the best performing district under the input orientated approach) had the highest average technical efficiency estimate of 98.9%, the highest minimum technical efficiency estimate of 94.9% and was the most times (13 times) fully technically efficient during the 22 months. On average, district 26 can increase its technical efficiency by proportionally reducing its input quantities by 1.1% without a reduction in output. District 2 (the worst performing district under the input orientated approach) had the lowest average technical efficiency estimate of 72.9%, the lowest minimum technical efficiency estimate of 64.8%. This means that on average district 2 could increase its technical efficiency by reducing its input quantities by 27.1% without a reduction in output. This can result in the reduction of the number of personnel, smaller number of ATM's or the reduction in other inputs.

18 Districts out of the 37 districts were at least one time fully technically efficient during the 22 months, while 19 districts out of the 37 districts were not once fully technically efficient during the 22 months. District 2 had the lowest maximum technical efficiency estimate of 78.7%. More districts were operating at increasing returns to scale than decreasing returns to scale, meaning that the scale they were operating at was too small.

Table 1: Technical efficiency estimates and the returns to scale for the 37 districts (input- and output-orientated)

Districts	Technical Efficiency Input-orientated			Returns to scale Input-orientated	Technical Efficiency Output-orientated			Returns to scale Output-orientated
	Mean	Min	Max		Mean	Min	Max	
1	0.776	0.699	0.948	Decreasing	0.837	0.781	0.972	Decreasing
2	0.729	0.648	0.787	Decreasing	0.812	0.678	0.883	Decreasing
3	0.909	0.791	1.000	Decreasing	0.902	0.824	1.000	Decreasing
4	0.968	0.916	1.000	Decreasing	0.963	0.895	1.000	Decreasing
5	0.902	0.856	0.966	Increasing	0.875	0.806	0.955	Increasing
6	0.770	0.725	0.828	Increasing	0.731	0.681	0.803	Increasing
7	0.850	0.769	0.970	Decreasing	0.901	0.836	0.978	Decreasing
8	0.813	0.732	0.970	Decreasing	0.840	0.762	0.978	Decreasing
9	0.899	0.801	0.975	Increasing	0.862	0.716	0.961	Increasing
10	0.930	0.851	1.000	Increasing	0.908	0.791	1.000	Increasing
11	0.795	0.734	0.852	Increasing	0.750	0.651	0.840	Increasing
12	0.852	0.800	0.942	Decreasing	0.854	0.775	0.949	Decreasing
13	0.838	0.774	0.944	Decreasing	0.810	0.713	0.945	Decreasing
14	0.901	0.817	0.975	Increasing	0.798	0.681	0.932	Increasing
15	0.942	0.890	1.000	Increasing	0.908	0.825	1.000	Increasing
16	0.852	0.686	1.000	Decreasing	0.857	0.707	1.000	Decreasing
17	0.960	0.874	1.000	Increasing	0.955	0.878	1.000	Increasing
18	0.933	0.887	1.000	Increasing	0.910	0.860	1.000	Increasing
19	0.890	0.810	1.000	Increasing	0.873	0.767	1.000	Increasing
20	0.841	0.770	0.924	Decreasing	0.865	0.749	0.963	Decreasing
21	0.890	0.791	1.000	Decreasing	0.905	0.783	1.000	Decreasing
22	0.807	0.762	0.865	Decreasing	0.784	0.720	0.872	Decreasing
23	0.874	0.815	0.930	Increasing	0.856	0.774	0.934	Increasing
24	0.808	0.735	0.957	Decreasing	0.831	0.720	0.966	Decreasing
25	0.965	0.899	1.000	Increasing	0.935	0.833	1.000	Increasing
26	0.989	0.949	1.000	Increasing	0.982	0.921	1.000	Increasing
27	0.874	0.780	1.000	Decreasing	0.871	0.762	1.000	Decreasing
28	0.921	0.811	1.000	Decreasing	0.936	0.838	1.000	Decreasing
29	0.960	0.901	1.000	Increasing	0.948	0.878	1.000	Increasing
30	0.917	0.840	1.000	Decreasing	0.917	0.816	1.000	Decreasing
31	0.965	0.894	1.000	Increasing	0.960	0.879	1.000	Increasing
32	0.796	0.727	0.895	Decreasing	0.818	0.737	0.917	Decreasing
33	0.773	0.717	0.844	Increasing	0.693	0.617	0.801	Increasing
34	0.763	0.666	0.876	Increasing	0.624	0.509	0.708	Increasing
35	0.874	0.830	0.924	Increasing	0.846	0.771	0.925	Increasing
36	0.945	0.845	1.000	Increasing	0.938	0.811	1.000	Increasing
37	0.966	0.921	1.000	Increasing	0.943	0.887	1.000	Increasing

Under the output-orientated approach district 26 (the best performing district under the output orientated approach) had the highest average technical efficiency estimate of 98.2%, the highest minimum technical efficiency estimate of 92.1% and was the most times (13 times) fully technically efficient during the 22 months. District 26 can, on average, increase its technical efficiency by increasing its output quantity by 1.8% without any increase in its input quantities but only by utilizing its inputs more efficiently. This means making better use of inputs like personnel, ATM's, computer facilities and other types of inputs.

District 34 (the worst performing district under the output orientated approach) had the lowest average technical efficiency estimate of 62.4% and the lowest minimum technical efficiency estimate of 50.9%. This district also had the lowest maximum technical efficiency estimate of 70.8% and was not once fully technically efficient during the 22 months.

18 districts out of the 37 districts were at least 1 time fully technically efficient during the 22 months. 19 districts out of the 37 districts were not fully technically efficient during the 22 months. More districts were operating at decreasing returns to scale than increasing returns to scale, meaning that the scale they were operating at was too large.

7. Conclusion

From the results presented in Table1, it is clear that the districts from this bank were not as efficient as it should be. The analyses indicated that 19 districts out of the 37 districts were not at least one time fully technically efficient during the 22 months (input- and output-orientated approach). This means that more than 50% of the number of districts from this bank was inefficient. Under the input orientated approach it means that these districts could proportionally reduce all its inputs without any reduction in output. Reduction in inputs means lower input costs that can lead to lower production costs.

Under the output orientated approach it means that these inefficient districts could expand its output without altering its input quantities, but by better utilizing its inputs. Better utilization of inputs can lead to lower per unit production costs.

On average, only a limited number of these districts were operating at the optimal scale. They were either operating at a scale that was too large (decreasing returns to scale) or they were operating at a scale that was too small (increasing returns to scale).

Only one of the four largest banks was included in this investigation, but with four banks controlling over 85% of total deposits and assets in South Africa, it is clear that banks continuously face fierce competition competing for customers and scarce resources.

Analyzing the efficiency of the districts of this bank, the conclusion can be made that a large portion of higher bank fees may be the result of inefficiency in banks, indicating that customer complaints about high bank fees are justified.

8. References

AKINBOADE, O.A. & MAKINA, D. 2006. Financial sector development in South Africa, 1970-2002. *Studies in Economics and Econometrics*, 30(1):101-127.

AVKIRAN, N.K. 1997. Models of retail performance for bank branches: predicting the level of key business drivers. *International Journal of Bank Marketing,* 15(6-7):224-237.

AVKIRAN, N.K. 1999. An application reference for data envelopment analysis in branch banking: helping the novice researcher. *International Journal of Marketing,* 17(5):206-220.

BANKER, R.D., CHARNES, A. & COOPER, W.W. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science,* 30(9):1078-1092.

BANK SUPERVISION. Various years. *Annual report.* Pretoria: South African Reserve Bank.

BERGER, A.N. & HUMPHREY, D.B. 1997. Efficiency of financial institutions: international survey and directions for future research*. European Journal of Operations Research*, 98(2):175–212.

BERRY, A.J., BROADBENT, J. & OTLEY, D. 2005. *Management control: theories, issues and performance*, 2nd ed. New York: Palgrave Macmilan.

BESSIS, J. 2005. *Risk management in banking*, 2nd ed. New York: Wiley.

CHARNES, A., COOPER, W.W. AND RHODES, E. 1978: Measuring efficiency of decision-making units", *European Journal of Operations Research*, 2, 429-444.

CHELST, K., SCHULTZ, J. & SANZHVI, N. 1988. Issues and decisions aids in designing branch networks. *Journal of Retail Banking,* 10(2):5-17.

CAPGEMINI, ING & THE EUROPEAN FINANCIAL MANAGEMENT AND MARKETING ASSOCIATION (EFMA). 2007. *World Retail Banking Report*. capgemini.com/m/en/n/pdf_World_Retail_Banking_Report.pdf. Accessed, 28 May 2008.

CLARK, P. 1997. The balanced scorecard. *Accountancy Ireland,* 29(6):25-26.

COELLI, T., RAO, D.S.P. & BATTESE, G. 1998. *An introduction to*

efficiency and productivity analysis. Boston: Kluwer Academic Publishers.

DAVENPORT, T.O. & SHERMAN, H.D. 1987. Measuring brance profitability. *The Banker's Magazine,* 170(5):37-38.

DE YOUNG, R. 1997. Measuring bank cost efficiency: don't count on accounting ratios. *Financial Practice and Education,* 7(1):20-31.

ECCLES, R. 1991. The performance measurement manifesto. *Harvard Business Review,* 69(1):131-137.

ELYASIANI, E. & MEHDIAN, S.M. 1990. A non-parametric approach to measurement of efficiency and technological change: the case of large US commercial banks. *Journal of Financial Services Research*, 4(2):157–168.

FäRE, R., GROSSKOPF, S. & LOVELL, C.A.K. 1985. *The measurement of efficiency of production.* Boston: Kluwer-Nijhoff.

FARREL, M. J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A*, 120(3):253-281.

FAVERO, C.A. & PAPI, L. 1995. Technical efficiency and scale efficiency in the Italian banking sector: a non-parametric approach. *Applied Economics,* 27(4):385-395.

GARDNER, M.J. & MILLS, D.L. 1994. *Managing financial institutions: an asset liability approach.* Orlando: The Dryden Press.

GERING, M. & ROSMARIN, K. 2002. The do's and don'ts of balanced scorecard. *Accountancy SA,* p.18-19, Oct.

GIRARDONE, C., MOLYNEUX, P. & GARDENER, E.P.M. 2004. Analysing the determinants of bank efficiency: the case of Italian banks. *Applied Economics,* 36(3):215–227.

GOLANY, B. & STORBECK, J. 1999. A data envelopment analysis of the operational efficiency of bank branches. *Interfaces,* 29(3):14-26.

HARTLE, F. 1997. *Transforming the performance management process.* London: Clays Ltd, Kogan Page.

HAWKINS, P. 2004. South Africa's financial sector ten years on: performance since democracy. *Development South Africa*, 21(1):179-204.

KAO, C. & LIU, S.-T. 2004. Predicting bank performance with financial forecasts: a case of Taiwan commercial banks. *Journal of Banking and Finance,* 28(10):2353-2368.

KAPARAKIS, E.I., MILLER, S.M. & NOULAS, A.G. 1994. Short-run cost inefficiency of commercial banks: a flexible stochastic frontier approach. *Journal of Money, Credit and Banking,*

26(4):875–893.

KAPLAN, R.S. & NORTON, D.P. 1996. Linking the balanced scorecard to strategy. *California Management Review,* 39(1):53-80.

KAPLAN, R.S. & NORTON, D.P. 1996a. The balanced scorecard as a strategic management system. *Harvard Business Review,* 74: 75-79, Jan/Feb.

KAPLAN, R.S. & NORTON, D.P. 2001. *The strategy focused organization: how the balanced scorecard companies thrive in the new business environment.* Boston: Harvard Business School.

KIMBALL, R.C. 1997. Innovations in performance measurement in banking. *New England Economic Review*, 23-39, May/June.

KOCH, T. W. & MACDONALD, S. S. 2003. *Bank management.* 5th ed. London: Dryden Press.

MCDONELL, E.D. & RUBIN, M.J. 1991. A scorecard for branch performance. *Financial Managers' Statement,* 13(3):46-56.

MESTER, L.J. 1994. Efficiency of banks in the third Federal Reserve district. *Federal Reserve* Bank *of Philadelphia, Working Paper*, 94-1:3-18.

MESTER, L.J. 1996. A study of bank efficiency taking into account risk-preferences. *Journal of Banking and Finance,* 20(6):1025-1045.

MESTER, L.J. 2003. Applying efficiency measurement techniques to central banks. *Federal Reserve Bank of Philadelphia, Working Paper,* 03-13:1-40.

MLIMA, A.P. & HJALMARSSON, L. 2002. Measurement of inputs and outputs in the banking industry. *Tanzanet Journal,* 3(1):12-22.

MOLYNEUX, P., ALTUNBAS, Y. & GARDENER, E.P.M. 1996. *Efficiency in European banking.* England: Wiley.

OKEAHALAM, C.C. 2006. Production efficiency in the South African banking sector: a stochastic analysis. *International Review of Applied Economics,* 20(1):103-123.

RANGAN, N., GRABOWSKI, R., ALY, H.Y. & PASURKA, C. 1988. The technical efficiency of US banks. *Economics Letters,* 28(2):169-175.

RESTI, A. 1997. Evaluating the cost-efficiency of the Italian banking system: what can be learned from the joint application of parametric and non-parametric techniques. *Journal of Banking and Finance,* 21(2):221-250.

SHERMAN, D. & LADINO, G. 1995. Managing bank productivity using data envelopment analysis (DEA). *Interfaces*, 25(2):60–73.

SMITH, R.L. III & SCHWEIKART, J.A. 1992. A better way to assess branch profitability. *ABA Banking Journal,* 84(4):54-56.

THANASSOULIS, E. 1999. Data envelopment analysis and its use in banking. *Interfaces,* 29(3):1-13.

VAN DER WESTHUIZEN, G. 2006. Estimating technical and scale efficiency of banks using balance sheet data: a South African case. http://www.ekngvdw@puknet.puk.ac.za. Date of access: 20 May 2007.

VASSILOGLOU, M. & GIOKAS, D. 1990. A study of the relative efficiency of bank branches: an application of data envelopment analysis. *Journal of the Operational Research Society,* 41(7):591-597.

VERMA, K. 1992. Total factor productivity measurement, pertinent or passé? (*In* Bruns, W.J., Jr., *ed. Performance measurement, evaluation and investments.* Boston: Harvard Business School. p. 279-300.)

WHEELOCK, D.C. & WILSON, P.W. 1995. Explaining bank failures: deposit insurance, regulation, and efficiency. *Review of Economics and Statistics,* 77(4):689-700.

YUE, P. 1992. Data envelopment analysis and commercial bank performance: a primer with applications to Missouri banks. *Federal Reserve of St Louis,* p.31-45, Jan.