A framework for evaluating IT security investments in a banking environment

EH Smith
BIQueue
Johannesburg
South Africa
Eugenesmith1@gmail.com

HA Kruger
School of Computer Statistical and Mathematical Sciences
North-West University
Potchefstroom, South Africa
Hennie.Kruger@nwu.ac.za

Abstract— **The amount of effort that can be expended on information security depends on funds available and management decisions. Organisations therefore have to prepare an annual budget for the maintenance and improvement of their information security systems. Two of the key issues that confront IT management, when dealing with IT security investments, are how to spend the IT security budget most effectively, and how to make the case for an increase in funds to maintain and further enhance information security. The aim of this paper is to present a quantitative framework as an alternative way of analysing IT security investments in a banking environment in order to address the two issues mentioned above. A two step framework is proposed. The first step utilizes a cluster analysis (CA) technique and the second step employs a linear programming technique called data envelopment analysis (DEA). The purpose of the clustering step is to ensure that evaluations are carried out in groups of homogenous bank branches while the purpose of the DEA model is to determine which of the branches make efficient use of the IT security resources available to them. Following a brief discussion of the proposed framework and techniques used, an illustrative example, based on a well known South African financial institution, is presented.**

Keywords - IT security investment; cluster analysis; data envelopment analysis

I. INTRODUCTION

The management of cyber security is a demanding task and investing in IT security is not just a matter of comparing technology needs with available technological function [1]. To manage the economic and technical impediments of an information security plan, Bodin *et al* [2] argued that there are two key issues that confront management; how to spend a limited information security budget effectively and how to make the case to top management for an increase in funds to further enhance an organization's information security. Another aspect that adds to the problem is the so called 'security paradox' as reported in The Security Paradox report issued by McAfee [3]. The paradox refers to the acknowledgement by organizations that cyber attacks, especially against midsize organizations, have increased, yet at the same time most organizations have frozen or cut their IT security budgets.

There are many sources available in the literature on IT security investments and text books usually provide a lot of theory on the making of an economic case for IT security and how to determine the economic value through measures such as the net present value, internal rate of return, return on investment etc. The authoritative work by Pfleeger and Pfleeger [1] serves as a good example of a source that deals with these types of issues. However, despite these theoretical explanations there is still a shortage of reliable quantitative models that can provide enough information to analyze IT security investments. Cavusoglu *et al* [4] stated for example, that the lack of a comprehensive model that incorporates the specific features of IT security technologies has prevented firms from applying rigorous quantitative techniques to make security investment decisions. In an effort to address this lack of comprehensive quantitative models in the area of IT security investments researchers like Cavusoglu *et al* [4] and others have started to investigate the use of different techniques that rely on measurable or quantifiable metrics that can be used to assist in making sound decisions when dealing with IT security investments. Examples of other researchers in the literature who worked in this area can be found in [2], [5], [6] and [7].

In this paper a quantitative framework or model is developed and suggested as an alternative way of analysing IT security investments. A two step framework (figure 1) is proposed and the bank branch network of a well known South African financial institution was used to develop, test and validate the proposed methodology. The first step of the proposed methodology utilizes a cluster analysis (CA) technique and the second step employs a linear programming technique called data envelopment analysis (DEA). The purpose of the clustering step is to ensure that evaluations are carried out in groups of homogenous branches while the purpose of the DEA model is to determine which of the branches make efficient use of the IT security resources available to them. A bank branch network was selected for the study because banks and financial institutions are normally organisations that are making intensive use of technology

978-1-4244-5494-5/10/$26.00 ©2010 IEEE

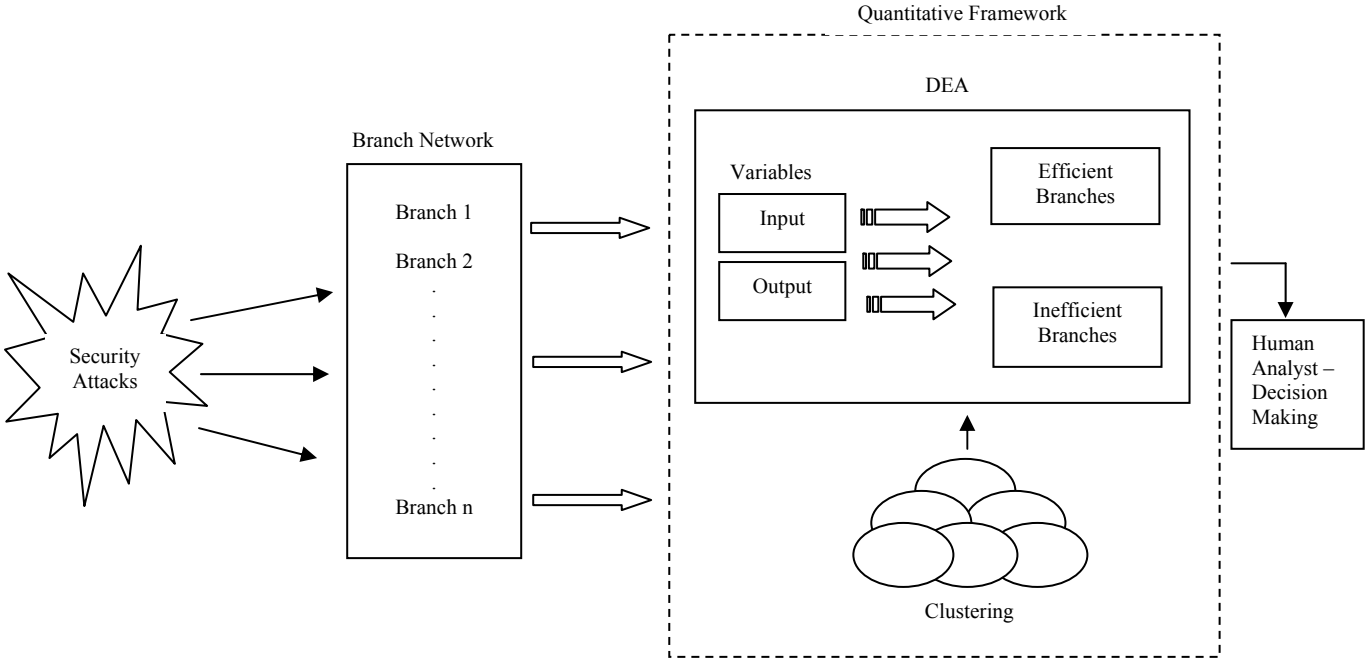

Figure 1. Proposed framework

and as such are subjected to high risks in terms of IT security and cyber attacks. In addition, the use of the DEA technique in a banking environment is not something new and there are numerous examples where DEA was used to determine the general efficiency of bank branches. Examples can be found in [8], [9] and [10].

The remainder of the paper is structured as follows. Sections II and III present a brief overview of clustering and data envelopment analysis respectively. The application of the methodology and results will be given in section IV while section V concludes the paper with some general comments.

II. CLUSTER ANALYSIS (CA)

Cluster analysis is a popular undirected data mining technique that is used to identify homogeneous objects [11]. It is defined by Jain *et al* [12] as the separation of a heterogeneous data set into groups of data, where each group is called a cluster, so that objects within a cluster are similar but different from objects in other clusters. Samoilenko and Osei-Bryson [13] stated that amongst the many reasons for performing clustering, the following two are particularly significant when dealing with predictive modeling in the presence of sample heterogeneity.

- To find a set of natural groups (i.e. segmentation)
- To improve the performance of predictive modeling and data mining techniques.

The above two reasons are also applicable in this study. The objective of the first step in the proposed framework is to identify natural groups of bank branches based on similarity principles. Bank branches of a financial institution differ extensively in services they offer and customers they serve. Branches located in rural areas would offer vastly different services to a particularly different customer base than a branch located in a highly populated, commercialized area. Therefore, in order for DEA (the 2nd step in the proposed framework) to produce meaningful results, a high level of homogeneity within branches compared is essential. It is of course possible to identify similar branches of a financial institution simply by asking an expert within the financial institution that has a thorough understanding of the current branch network. Creating clusters containing ten or even twenty similar branches may be possible; however, logical groupings and clusters identified by expert opinion would largely be subjective to his/her personal interpretation of the individual branches under investigation. Jain *et al* [12] argue that humans can, without a doubt, perform competitively against automatic clustering algorithms in a two dimensional space however, they assert that with the increase in the dimensionality of the problem, so does the difficulty levels of intuitively interpreting the data increase. With well over four hundred branches and a large number of attributes describing each individual branch in this study, it becomes an impossible task for even the most seasoned expert to resolve without the use of a clustering algorithm.

A large number of clustering algorithms are available and reported on in the literature [14]. New algorithms are regularly introduced and older algorithms updated, with every algorithm performing better in its own right when presented with appropriate data. Clustering algorithms are generally grouped into hierarchical or partitioning algorithms and in this study *k*-means clustering, which is a partitioning clustering

978-1-4244-5494-5/10/$26.00 ©2010 IEEE

technique, was used and will briefly be described in the next paragraphs.

K-means clustering is one of the most popular clustering algorithms used to group similar objects together [15], [16]. The *k*-means algorithm groups a multi-dimensional data set into a predetermined number of groups called clusters. Objects within a cluster represent similar characteristics and dissimilar characteristics to objects in different clusters.

Once the user specified the appropriate *k*-value, the algorithm starts by selecting *k* arbitrary data points as the initial cluster centres, also called centroids. All data points are then assigned to the nearest arbitrary selected cluster centre and can only belong to one cluster. Subsequently, a new mean is calculated for every cluster and this becomes the new cluster centre or centroid. All the data points are then reassigned to the nearest centroid. The mean for every cluster is then re-calculated. At any point the *k*-means are actually the arithmetic mean of the groups of clusters it represents. This is an iterative process and continues until data points stop moving between clusters. Figure 2 is a graphical illustration of the *k*-means clustering process. With reference to the successive iterations, observe the cluster centre moving as objects get assigned to new clusters.

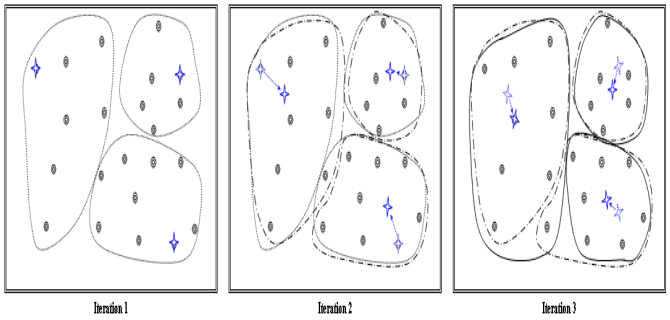

Figure 2. Clustering of a set of objects with the *k*-means method as illustrated by Han and Kamber [17]

Analysing every possible subset of clusters is computationally infeasible and various greedy heuristics are used for iteration optimisation. The squared error clustering algorithm is generally applied, and this algorithm minimises the squared error. Dunham [18] describes the squared error of a cluster as the sum of the squared Euclidean distances between each object in the cluster and the cluster centroid C_k. Given a cluster k_i, let the set of objects mapped to that cluster be $\{t_{i1}, t_{i2}, ..., t_{im}\}$. The squared error is defined as:

$$seK_i = \sum_{j=i}^{m} \left\| t_{ij} - C_k \right\|^2$$

Given a set of clusters $K = \{K_1, K_2, ..., K_k\}$, the squared error for *K* is defined as:

$$seK = \sum_{j=i}^{k} seK_i$$

It should be noted that different *k*-values will produce different results and the selected *k*-value should correspond to the natural structure of the data. Finding an appropriate *k*-value might be difficult. A *k*-value that is too high might lead to over fitting whereas a *k*-value that is too low might lead to patterns being undetected. The *k*-means algorithm may also be sensitive to outliers in the data.

III. DATA ENVELOPMENT ANALYSIS (DEA)

Data envelopment analysis (DEA) is a non-parametric linear programming methodology that is used to determine how efficiently an operating unit converts inputs to outputs when compared to other units. The technique was introduced by Charnes *et al* in 1978 [19] and has since been used extensively to measure and evaluate the relative efficiency of decision making units such as, for example, branches of the same bank, universities, hospitals and electric utilities.

The DEA technique converts multiple inputs and outputs into a single comprehensive measure of efficiency for each of the decision making units. This measure can lie anywhere between zero (meaning the unit is totally inefficient) and one (meaning the unit is technically efficient). Another advantage of DEA is that beyond simply discovering inefficient branches, additional insight can also be provided relating to the magnitude of inefficiency associated with the inputs and outputs. This ability of DEA makes it possible to pinpoint particular areas for improvement of efficiency. The original model and how the analysis is performed is briefly summarised by Vassilogou and Giokas [20] as follows.

$$\text{Maximise} \quad E_o = \left(\sum_{i=1}^{k} u_i \psi_{io} \right) \Big/ \left(\sum_{j=1}^{m} v_j x_{jo} \right)$$

$$\text{subject to} \quad \left(\sum_{i=1}^{k} u_i \psi_{ir} \right) \Big/ \left(\sum_{j=1}^{m} v_j x_{jr} \right) \le 1 \qquad r = 1, ..., n$$

$$u_i, v_j > \varepsilon \qquad i = 1, ..., k \qquad j = 1, ..., m$$

where

o = the index of the unit being assessed from the set of r = 1, …, n units

k = the number of outputs at the units

m = the number of inputs at the units

ψ_{ir} = observed output i at unit r

x_{jr} = observed input j at unit r

ε = small positive number

The above analysis is performed for the different units producing an efficiency rating for each of the n units. The required solution is the set of (u_i, v_j) values that maximise the

efficiency ratio E_o of the unit being rated without resulting in an input/output ratio exceeding 1 (100% efficiency). Consequently, if a relative efficiency rating of 100% is not attained under this set of weights, it cannot be attained under any other set (for the same sample of units). This fractional programming problem is replaced with a linear programming equivalent through a series of transformations, which are set out in detail in [19].

One of the fundamental assumptions of the DEA technique is that of functional similarity (homogeneity) of decision making units under review [21]. According to Samoilenko and Osei-Bryson [13] it simply means that in order to compare, meaningfully, the relative efficiencies of units in a sample or data set, the units must be similar in terms of utilisation of the inputs and production of the outputs. In this paper, the possible non-homogeneity of bank branches is addressed through the first step (clustering analysis) in the proposed framework.

The mathematical details of the DEA technique do not form part of this paper. A good exposition of technical details and the various types of DEA models can be found in [22] and [23].

IV. Application and Illustrative Example

This section presents an application and illustrative example of the proposed framework. The two steps in the framework were applied to the branch network of a well known South African financial institution and results indicated that the use of the methodology is feasible.

A. Application

The first step of the proposed methodology is to perform a cluster analysis of bank branches into homogeneous sets. The motivation for this step lies in the homogeneity assumption about units under assessment in a DEA model together with the fact that bank branches may differ in services offered to different communities. The financial institution operates close to four hundred branches across South Africa. While some of the branches operate in urbanised areas such as Sandton, which is at the heart of the South African economy, other branches operate in rural towns, for instance Ventersdorp, within a much smaller society. It is therefore clear that when comparing bank branches they should be similar by nature. Data for the clustering was provided by the financial institution and comprises information from 398 retail branches across South Africa for the period 1 January 2007 to 31 December 2007.

There were literally hundreds of variables available that described the branches and the services they deliver. To reduce the dimension of the problem it was decided to select only a few variables for clustering purposes. An intense process was then followed to ensure that the most appropriate variables are selected for the clustering exercise. This process included, amongst other things the discarding of descriptive variables e.g. branch name; the use of expert opinion; the use of correlation analysis to exclude redundant variables; removal of variables with missing data values etc. The variable selection process resulted in the following eleven variables used for the final clustering.

TABLE I. Variables used for clustering

1. Branch number	7. Number of credit card accounts
2. Number of current accounts	8. Number of high value customers
3. Number of home loan accounts	9. Number of asset based finance accounts
4. Number of small business customers	10. Number of investment accounts
5. Number of affluent customers	11. Number of money market accounts
6. Number of savings accounts	

The SAS 5.2 Enterprise Miner software package was used to perform the clustering. The procedure used in this software package finds clusters of objects, in this case homogeneous bank branches, by using the *k*-means clustering algorithm. It was decided to run the cluster analysis with different *k*-values in order to determine the most appropriate number of clusters. Results obtained from these empirical experiments revealed that a value of k=12 produced the best clustering results. These results were confirmed as being the "best" by statistical measures such as, for example, mean square errors, standard deviations and cluster radius. The results were also presented to senior management who confirmed that the clusters are realistic. From the above results, a cluster containing 109 branches was selected for further analysis. This cluster of branches, together with appropriate input and output variables, were taken as input for the DEA model (2^{nd} step of the proposed framework) and represents the required homogeneous set of units to be evaluated.

DEA is based on observed input and output data that is used to construct a measure of efficiency. The choice of input and output variables for this study relied on the availability of data. Unfortunately the data related specifically to IT security were classified as sensitive and confidential and the financial institution under review did not want to make the data available for publication. Data was therefore generated through a simulation process in order to be able to illustrate the suggested model. The next paragraph explains the choice of input and output variables used.

Input variable. Only one input variable, the *cost* of IT security at each branch, was selected for the DEA model. An average cost for IT security was available for each branch and included costs such as costs for physical security and costs recovered from branches for logical IT security. Unfortunately

978-1-4244-5494-5/10/$26.00 ©2010 IEEE

these cost figures were highly confidential and for illustrative purposes a cost figure was generated (based on the real data) from a normal distribution with mean 20 and a standard deviation of 10 (times R10000).

Output variables. Two output variables were selected. First, the *number of IT security incidents* at each branch was chosen. This data was also available but, as with the cost data, the data was too sensitive to publish. The number of incidents therefore had to be simulated as well. A Poisson distribution with a mean value of 20 was used to simulate the data. The second output variable used in the model was a *customer satisfaction rating*. Customer satisfaction ratings per branch were available and was calculated based on security aspects such as trust and confidence, services, safety and security, how safety staff conduct themselves etc. To protect the confidentiality of the data, a scaled version of the data was used in the model.

B. Results

The DEA model was applied to each of the 109 branches identified during the clustering analysis and a concise extraction of the results are presented below in table II. Each row of table II represents the solution to a linear program, which maximises the efficiency rating of the corresponding bank branch, under the constraints dictated by the output/input relationships operating in the complete data set.

TABLE II. BANK BRANCH EFFICIENCY RATINGS

Branch number	Efficiency rating	Reference set
1	0.4966	9, 51
2	1	
3	0.7144	9, 29
.	.	.
51	1	
52	0.5828	9, 29
.	.	.
91	0.7530	2, 51
92	0.9394	14, 29
.		
.		

In the third column, next to the relatively inefficient bank branches, appears the corresponding efficiency reference set. This is the subset of relatively efficient bank branches to which the bank branch in question has been most directly compared in deriving its efficiency rating. In this example only 6% of the branches were evaluated as fully efficient. This relative low number of efficient branches should not be a cause of concern as the results of any number of efficient branches can be used as a benchmark to assist management with future IT security investment decisions. Also, when the number of input and/or output variables is increased, the linear programming model would be able to discriminate easier amongst efficient and inefficient branches.

Results in table II already provide a lot of information for decision making. Not only does it give a list of efficient and inefficient branches, but the average use of input and output levels of efficient and inefficient branches can now be calculated and compared to assist in decisions pertaining to IT security investments (input) and the return (number of security breaches and customer satisfaction) on the investments. Table III presents a summary of the basic information that can be obtained from a simple analysis based on information in table II (note that the customer satisfaction rating is a scaled value).

TABLE III. ANALYSIS OF EFFICIENT AND INEFFICIENT BRANCHES

Variable	Average Value for efficient branches	Average Value for inefficient branches	Dif-ference	% difference
Cost	R204870	R256030	R51160	24.97
Number of breaches	18.73	24.57	5.84	31.18
Customer satisfaction	0.0127	0.0098	0.0029	29.60

This simple analysis shows that, on average, the inefficient branches used 24.97% resources (cost) more than the efficient branches. On average, the customer satisfaction rating at efficient branches was also 29.6% higher than those at the inefficient branches and the average number of security breaches at the efficient branches was 18.73 while for the inefficient branches the average number of breaches was 24.57; a difference of 31.2%.

Apart from identifying inefficient bank branches, DEA can also provide additional insight about the degree of inefficiency. This ability of DEA makes it possible to pinpoint particular areas for improvement of inefficiency and is derived from each inefficient branch's reference set (refer to table II). The reference subset consists of a branch or branches, producing a better level of output (number of security breaches and customer satisfaction) with fewer inputs (cost). Consider, for example, branch 3. The efficiency rating is 0.7144 (table II) and the reference set consists of branches 9 and 29. The solution to the linear program (not presented in this paper) for branch 3 gave dual prices of 0.242 and 0.567 for the reference branches 9 and 29 respectively. Suppose we create a composite branch by combining 0.242 of branch 9 with 0.567 of branch 29, i.e. $0.242[x_1, x_2, \ldots, x_n]^T + 0.567[x_1, x_2, \ldots, x_n]^T$ where x_i = input *i*. Doing this, for the input

variable cost, the following result (table IV) is obtained for branch 3.

TABLE IV. ANALYSIS OF BRANCH 3

Input	**Actual value (R10000)**	**Value if efficient (Target value) (R10000)**	**Difference (R10000)**	**Percentage difference**
Cost	26.322	18.806	7.516	39.96

The result shows that the actual input (26.322) for branch 3 was greater than the derived composite (18.806) efficiency set. Specifically, the output achieved for branch 3 could have been achieved using R75160 less costs. Similar analyses for each inefficient branch can now be done to determine a cost value for IT security in each of the bank branches.

To summarise, the proposed framework provides a technique for assessing the relative efficiency of a homogenous group of bank branches where there are multiple outputs and inputs related to IT security investments. It also provided an indication as to how inefficient branches should attempt to vary its IT security inputs and outputs so as to achieve a performance comparable to the best observed. It is this additional insight that will enable managers to make the case for an increase in funds for IT security investment or to ensure that budgets are spent effectively.

In general, the use of a DEA model has the following advantages [24] which are of specific significance in the work presented in this paper.

- Multiple outputs and inputs (related to IT security investment) may be included simultaneously in a single measure and because efficiency is often not a simple or singular measure, DEA can help to avoid the ambiguity that can arise with simple measures such as ratio analysis, cost variances etc.
- A priori weights are not required for the input and output variables.
- If controllable inputs (e.g. IT security cost) are included in the DEA model, a management strategy can be developed to improve efficiency.
- DEA focuses on achievable best performance. Each unit (bank branch) is compared to other units and not to a hypothetical ideal or average performance. The model provides a meaningful and defensible standard based on best observed practice.
- The reduction of multiple variables to a single performance measure reduces cognitive complexity.

V. CONCLUSION

Baker *et al* [5] stated that although there is no shortage of security standards and research business leaders often admit they have no reliable methodology for measuring the effectiveness of their security initiatives or collecting the data needed to make strategic decisions and determine the financial value of their efforts.

In this paper a quantitative framework for evaluating IT security investments in a banking environment was presented. The proposed framework consists of two steps. In the first step clustering analysis is performed to ensure that evaluations are carried out in groups of homogenous bank branches while the second step entails the application of a linear programming model called data envelopment analysis to assess the efficiency of IT security investments. To demonstrate the framework it was applied on a limited scale to a South African financial institution. Due to the sensitive nature of data, certain data had to be simulated to illustrate the framework and model. Results have indicated that the proposed framework provides easy to understand information on the efficiency of bank branches where there are multiple input and output variables related to IT security investments. Additional information such as how to vary inputs (e.g. cost) will enable management to make important security investment decisions such as to justify an increase in funds.

Based on the results and advantages offered by the proposed methodology, it is believed that the approach will alleviate uncertainty that is often linked to IT security investments and that an improvement in analysis and decision making can be expected.

REFERENCES

[1] C. P. Pfleeger and S. L. Pfleeger, Security in Computing. 4th edition, Prentice Hall, Upper Saddle River, NJ, 2007

[2] L. D. Bodin, L. A. Gordon, and M. P. Loeb, "Evaluating information security investments using the analytic hierarchy process," Communications of the ACM, vol. 48(2), pp.79-83, February 2005.

[3] McAfee, "The security paradox," McAfee, Inc., Santa Clara, CA, 2009.

[4] H. Cavusoglu, B. Mishra, and S. Raghunathan, "A model for evaluating IT security investments," Communications of the ACM, vol. 47(7), pp.87-92, July 2004.

[5] W. H. Baker, L. P. Rees, and P. S. Tippett, "Necessary measures metric-driven information security risk assessment and decision making," Communications of the ACM, vol. 50(10), pp.101-106, October 2007.

[6] J. J. C. H. Ryan and D. J. Ryan, "Expected benefits of information security investments," Computers & Security, vol. 25, pp.579-588, 2006.

[7] R. Bojanc and B. Jerman-Blazic, "Towards a standard approach for quantifying an ICT security investment," Computer Standards & Interfaces, vol. 30, pp.216-222, 2008.

[8] T. T. Lin, C. Lee and T. Chiu, "Application of DEA in analyzing a bank's operating performance," Expert Systems with Applications, vol. 36, pp.8883-8891, 2009.

[9] C. J. O'Donnell and G. van der Westhuizen, "Regional comparisons of banking performance in South Africa," South African Journal of Economics, vol. 70(3), pp.485-518, March 2002.

[10] L. P. Fatti and K. Clarke, "Use of data envelopment analysis and regression for establishing manpower requirements in a bank," Orion, vol. 14(1), pp.57-66, 1998.

[11] I. Kononenko and M. Kukar, Machine learning and data mining: introduction to principles and algorithms. Chichester, Horwood, 2007.

978-1-4244-5494-5/10/$26.00 ©2010 IEEE

[12] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM computing surveys, vol. 31(3), pp.264-323, September 1999.

[13] S. Samoilenko and K. Osei-Bryson, "Increasing the discriminatory power of DEA in the presence of heterogeneity with cluster analysis and decision trees," Expert Systems with Applications, vol. 34, pp.1568-1581, 2008.

[14] R. Gelbard, O. Goldman, and I. Spiegler, "Investigating diversity of clustering methods: an empirical comparison," Data & Knowledge engineering, vol. 63(1), pp.155-166, October 2007.

[15] S. Liao and C. Wen, "Artificial neural networks classification and clustering methodologies and applications – literature analysis from 1995 to 2005," Expert Systems with Applications, vol. 32, pp.1-11, 2007.

[16] S. A. Mingoti and J. O. Lima, "Comparing SOM neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms," European Journal of Operational Research, vol. 174(3), pp.1742-1759, 2006.

[17] J. Han and M. Kamber, Data mining: concepts and techniques. 2nd edition, Kaufmann, San Francisco, 2006.

[18] M. H. Dunham, Data mining introductory and advanced topics. Prentice Hall, Upper Saddle River, NJ, 2003.

[19] A. Charnes, W. W. Cooper, and E. L. Rhodes, "Measuring the efficiency of decision making units," European Journal of Operational Research, vol. 2(6), pp.429-444, 1978.

[20] M. Vassiloglou and D. Giokas, "A study of the relative efficiency of bank branches: an application of data envelopment analysis," Journal of the Operational Research Society, vol. 41(7), pp.591-507, 1990.

[21] R. G. Dyson, R. Allen, A. S. Camanho, V. V. Podinovski, C. S. Sarrico and E. A. Shale, "Pitfalls and protocols in DEA," European Journal of Operational Research, vol. 132(2), pp.245-259, 2001.

[22] A. Charnes, W. W. Cooper, A. Y. Lewin and L. M. Seiford, Data envelopment analysis theory, methodology and applications. Kluwer Academic, London, 1994.

[23] L. M. Seiford and R. M. Thrall, "Recent developments in DEA: the mathematical programming approach to frontier analysis," Journal of Econometrics, vol. 46, pp.7-38, 1990.

[24] L. M. Metzger, "Opeartional auditing and DEA: measuring branch office efficiency," Internal Auditing, pp.3-12, Fall 1994.

978-1-4244-5494-5/10/$26.00 ©2010 IEEE