Elsevier Editorial System(tm) for Expert Systems With Applications

Manuscript Draft

Manuscript Number:

Title: Increasing the Discriminatory Power of DEA in the Presence of the Sample Heterogeneity with Cluster Analysis and Decision Trees

Article Type: Full Length Article

Section/Category:

Keywords: Data Envelopment Analysis; Heterogeneity; Homogeneity; Clustering; Decision Tree Induction; Data Mining

Corresponding Author: Dr. Sergey Valery Samoilenko, PhD

Corresponding Author's Institution: Virginia Commonwealth University

First Author: Sergey Valery Samoilenko, PhD

Order of Authors: Sergey Valery Samoilenko, PhD; Kweku-Muata Osei-Bryson, PhD

Manuscript Region of Origin:

Abstract:

Increasing the Discriminatory Power of DEA in the Presence of the Sample Heterogeneity with Cluster Analysis and Decision Trees

Sergey Samoilenko and Kweku-Muata Osei-Bryson
Department of Information Systems & The Information Systems Research Institute
VIRGINIA COMMONWEALTH UNIVERSITY
Richmond, VA 23284, U.S.A.

ABSTRACT:
Data Envelopment Analysis (DEA) is a widely used non-parametric data analytic tool discriminatory power of which is dependent on the homogeneity of the domain of the sample. In many real-life cases, however, the sample of the Decision Making Units (DMU) could consist of two or more naturally occurring subsets, thus exhibiting clear signs of heterogeneity. In such situations, the discriminatory power of DEA is limited, for the nature of the relative efficiency of a DMU is likely to be influenced by its membership in a particular subset of the sample. In this study, we propose a three-step methodology allowing for increasing the discriminatory power of DEA in the presence of the heterogeneity of the sample. In the first phase, we use Cluster Analysis (CA) in order to test for the presence of the naturally occurring subsets in the sample. In the second phase DEA is used to calculate the relative efficiencies of the DMUs, as well as averaged relative efficiencies of each subset identified in the previous phase. Finally, we utilize Decision Tree (DT) induction in order to inquire into the subset-specific nature of the relative efficiencies of the DMUs in the sample. Illustrative example is provided.

KEYWORDS:
Data Envelopment Analysis, Heterogeneity, Homogeneity, Clustering, Decision Tree Induction; Data Mining

1. INTRODUCTION

Data Envelopment Analysis (DEA) is a non-parametric data analytic technique that is extensively used by various research communities (e.g. Hong et al., 1999; Sohn and Moon, 2004; Seol et al., 2006) since its introduction by Charnes et al. (1978). The domain of inquiry of DEA is a set of the entities, commonly called Decision Making Units (DMUs), which receive multiple inputs and produce multiple outputs. Given a sample of the DMUs, the purpose of DEA is establishing of the relative efficiencies of each DMU within a sample. By collapsing the multiple inputs and outputs into the meta input and meta output, DEA employs linear programming techniques to establish an input-output based ratio that is then used to derive and assign a relative efficiency score to each DMU in the sample.

As any non-parametric method, DEA eschews the hard-to-satisfy distributional assumption of data normality, thus allowing to investigator to inquire into the nature of the relative efficiency of the sample of convenience. This comes with the price, however, for DEA relies on a different set of assumptions that must be satisfied for the results to be valid.

One of the fundamental assumptions of DEA is that of a functional similarity of the DMUs in a sample. This simply means that in order to compare, meaningfully, the relative efficiencies of DMUs in the sample (or dataset), these DMUs must be similar in terms of utilization of the inputs and production of the outputs. Thus, DEA requires us to make sure that we compare apples and apples, and not apples and oranges.

The difficulty arises when DMUs are represented not by apples and oranges, but by somewhat less homogenous entities such as people, companies, schools, countries, etc. In general, departments are more complex than people, companies are more complex then departments, and industries and countries are more complex than companies. As a complexity of a DMU increases, it gives rise to the increase in diversity in the set of DMUs, and this leads, problematically for DEA, to heterogeneity, rather than homogeneity, of the sample.

One of our motivations for addressing is this issue is that it arises in various applications of the DEA methodology to real world problems. In our case we were

exploring the relative efficiencies of a set of countries that were transitioning from centralized, planned economies to the market economies. To the extent that they are often considered by the international community, including significant actors such as the World Bank and IMF, to be a single group under the label *Transition Economies* (TE), there appears to be good reason to meaningfully compare them using DEA. However, in the situations where a domain of DMUs has been superimposed by the existing classifications, i.e., TEs, firms, departments, industries, one may end up with the sample consisting of DMUs that are functionally similar, yet heterogeneous. The DMUs are functionally similar because they receive, according to the specified DEA model, the same set of the inputs and produce the same set of the outputs, and such DMUs are possibly heterogeneous because no DMU could be specified completely by its inputs and the outputs. Again, complexity of the DMUs gives rise to the heterogeneity of the relevant set.

It would appear that while the functional similarity of DMUs in a sample is assumed, the homogeneity of the DMUs is taken for granted. This, of course, severely limits the discriminatory power of DEA results. On the other hand that the explicit consideration of the possible heterogeneity in a sample would increase the discriminatory power of the results of the DEA.

Currently, it would appear that the test of the assumption of the homogeneity of the DMUs in the sample represents the less rigorous, if not non-existent, part of DEA, for it is implicitly assumed that an investigator made sure that this important assumption holds. We suggest that the rigor of DEA, as well as its discriminatory power, could be increased by making the process of the assumption checking more explicit and objective.

If one intends to inquire into the differences between efficient and inefficient DMUs in the sample, then, fundamentally, there are only two possible routes. First, if the assumption of the homogeneity holds, then the differences in the scores of the relative efficiency could be investigated without any adjustments. Or, second, if the assumption of the homogeneity of the DMUs in the sample does not hold, then the differences in the scores of the relative efficiency should be investigated with the appropriate adjustments. The question is then becomes: what are the appropriate adjustments?

It is not our intent to propose that DEA should only be performed in the situations where homogeneity of the DMUs has been decisively established. After all, a homogeneity, or, heterogeneity, of the DMUs in a sample is a matter of a degree. Sometimes the differences are minor and the researchers should declare that essentially the assumption of the homogeneity of the DMUs holds. Other times, however, the researchers might want to take into consideration the differences between the DMUs, while still comparing their relative efficiencies within the same sample.

Currently, to our knowledge, there exists no methodology allowing a researcher to investigate the differences between the relatively efficient and inefficient DMUs while taking into consideration heterogeneity of the DMUs in the sample. In this research we aim to address this problem. Namely, we propose and illustrate a three-step methodology allowing a researcher, first, to inquire into the differences between the DMUs in the sample, second, conduct DEA, and third, inquire into the differences between the relatively efficient and inefficient DMUs while taking into consideration the differences uncovered in the first step.

The first step of our methodology utilizes Cluster Analysis (CA), the second step employs Data Envelopment Analysis (DEA), and the third step relies on the Decision Tree (DT) analysis. We provide brief overview of these data analytic techniques next, followed by the description of the proposed methodology.

2. OVERVIEW OF THE RESEARCH TECHNIQUES

2.1 Data Envelopment Analysis (DEA)

Data Envelopment Analysis (DEA) is a nonparametric method of measuring the efficiency of a decision-making unit (DMU). Any collection of entities that operates with the same set of the inputs and produces the same outputs, be it a firm or a country, could be designated as DMUs.

This method, which is nonparametric in the sense that DEA is entirely based on the observed input-output data, was originated as a collection of techniques for measuring the relative efficiency of a set of DMUs with unknown or unavailable price values for data inputs and outputs (Sengupta 1996). While it is beyond the scope of this paper to provide an overview of the theory behind the computations in DEA, we would like to

direct the interested reader to the comprehensive presentation of the theoretical underpinnings of the DEA by Dula (2002b).

One of the benefits of DEA lies in its flexibility, for a researcher could take advantage of several models and orientations that this method has to offer. Thus, for example, the choice of a given DEA model would depend on the underlying economic assumptions about the returns to scale of the process that transforms the inputs into the outputs (Dula 2002a). Consequently, the different sets of assumptions would yield the different models. As a result, instead of forcing a single perspective, DEA offers multiple vantage points in the form of the several models and orientations available to the researcher.

The three commonly mentioned orientations of DEA model are the following: input-oriented, output oriented, and base-oriented (Charnes, Cooper, Lewin and Seiford 1994). An input-oriented model is concerned with the minimization of the use of the inputs for achieving a given level of the output. Output-oriented DEA model, on the other hand, is concerned with the maximization of the level of the outputs per given level of the inputs.

The base-oriented model, unlike the first two, has dual orientation and is concerned with the optimal combination of the inputs and outputs. Therefore, this type of DEA model deals with the efficiency of the input utilization and efficiency of the output production, having control over both inputs and outputs within the model.

In this research, we employ the input- and output-oriented DEA models, descriptions of which are provided in the Table 1.1. We reason that a policy maker might use the input-oriented model to get the insights regarding the efficiency of the utilization of the resources (e.g., investment in ICT), and the output-oriented model to evaluate the effectiveness of the transformation of the given level of the inputs into the outputs (e.g., revenues from ICT) in a given year. However, taking into the consideration the context of this study, it is hard to imagine that a policy maker would have a control over both inputs and outputs (e.g., both investments in and revenues from ICT). Thus, we decided against using a base-oriented DEA model.

The original DEA model was introduced in 1978 by Charnes, Cooper and Rhodes and it is commonly called CCR (an abbreviation consisting of first letters of the authors' names). This model allows representing multiple inputs and outputs of each DMU as a single abstract "meta input" and "meta output." Consequently, the efficiency of each DMU can be represented as a ratio of the abstract input to the abstract output, and the resulting efficiency value can then be used for comparison with other DMUs in the set.

Mathematically, this ratio can be expressed as the following objective function:

max $h_0(u,v_i) = \Sigma_r u_r y_{ro} / \Sigma_i v_i x_{io}$

where,

the u_r is the variable representing the output value

the v_i is the variable representing the observed input value

the y_{ro} is the observed amount y of output r produced by DMU_0 from the input x_{io}

the x_{io} is the observed amount of input i consumed in order to produce y amount of output r by the DMU_0 (the DMU to be evaluated).

Adding the normalizing constraint, according to which ratio of virtual input to virtual output for each DMU must be less than one, the following linear programming (LP) problem can be formulated:

max $h_0(u,v) = \Sigma_r u_r y_{ro} / \Sigma_i v_i x_{io}$

subject to

$\Sigma_r u_r y_{rj} / \Sigma_i v_i x_{ij} \leq 1$ for $j = 1,\ldots, n$

$u_r, v_i \geq 0$ for all i and r,

where in the case of fully rigorous development $u_r, v_i \geq 0$ would be replaced with the constraint including a *non-Archimedean element* ε such as

$$\frac{u_r}{\sum_{i=1}^{m} v_i x_{i0}}, \frac{v_r}{\sum_{i=1}^{m} v_i x_{i0}} \geq \varepsilon > 0,$$ and ε is smaller than any positive real number.

(adapted from Cooper et al. (2004)).

However, the given above formulation yields infinite number of solutions, thus, a transformation of the formulation has been developed by Charnes and Cooper(1962),

which changes (u, v) variables to (μ, v) and yields the equivalent linear programming problem to the one above.

For the convenience of the reader, we provide a summary consisting of a Charnes-Cooper transformation model and the corresponding to it LP Dual (sometimes referred to as the "Farrell model") in the "relaxed" form in the Table 1.1 below.

Table 1.1: DEA Model Types(adapted from Cooper et al. (2004).

	Charnes-Cooper transformation	LP Dual ("Farrell model")	LP Dual Solution (Score)
Input-Oriented DEA model	$\max z = \sum_{r=1}^{s} \mu_r y_{r0}$ Subject to $\sum_{r=1}^{s} \mu_r y_{r0} \quad \sum_{i=1}^{m} v_i x_{ij} \le o$ $\sum_{i=1}^{m} v_i x_{ij} = 1$ $\mu_r, v_i \ge 0$	$\Theta^* = \min \Theta$ Subject to $\sum_{J=1}^{n} x_{ij} \lambda_j \le \Theta x_{i0} \qquad i = 1,2,....,m;$ $\sum_{J=1}^{n} y_{rj} \lambda j \ge y_{r0} \qquad r = 1,2,....,s;$ $\lambda_j \ge 0 \qquad j = 1,2,....,n;$	*Solution:* $\Theta^* \le 1$ *Score:* If $\Theta^* < 1$, DMU is inefficient If $\Theta^* = 1$, DMU is efficient.
Output-Oriented DEA model	$\min q = \sum_{i=1}^{m} v_i x_{i0}$ Subject to $\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \ge o$ $\sum_{r=1}^{s} \mu_r y_{r0} = 1$ $\mu_r, v_i \ge \varepsilon$	$\Theta^* = \max \Theta$ Subject to $\sum_{j=1}^{J} z_j x_{jn} \ge \Theta u_{jm} \qquad m = 1,2,....,M;$ $\sum_{j=1}^{J} z_j x_{jn} \le x_{jn} \qquad n = 1,2,....,N;$ $z_j \ge 0 \qquad j = 1,2,....,J;$	*Solution:* $\Theta^* \ge 1$ *Score:* If $\Theta^* > 1$, DMU is inefficient If $\Theta^* = 1$, DMU is efficient.

A relative efficiency of a DMU can be characterized as being *strong* or *weak*. The provided above models operates under assumption of *strong disposal* by ignoring the presence of non-zero slacks, allowing, consequently, for the solutions with *weakly efficient* DMUs. For example, a DMU is considered to be *strongly* (fully) efficient if $\Theta^* = 1$ and all slacks are equal to zeroes. On another hand, a DMU could be *weakly* efficient if it obtained the same score of $\Theta^* = 1$, but some slacks are not equal to zero. As it was shown above, the type of relative efficiency of a given DMU is determined by the constraint of the LP problem: if the constraint utilizes a *non-Archimedean element* ε, the solution would allow for *fully efficient* DMUs, and if not, then all relatively efficient DMUs must be qualified as *being weakly* efficient. Consequently, a provided in the Table 1.1 relaxed form of the Farrel model would allow for the presence of the weakly efficient DMUs.

An obtained solution to the appropriate LP Dual results in the assignment of a score to each of the DMU in the set. The comparison of the scores that are assigned to DMUs allows for the efficiency ranking of each DMU in the given set, where the highest-ranking DMU is considered to be 100% efficient (Sengupta 1996).

An assigned score of "1" indicates relative efficiency of a DMU, while any other received score indicates relative inefficiency of a DMU. A score of less other than one indicates relative inefficiency means that some other unit(s) from the sample could produce the given level of outputs using less inputs (in output-oriented model), or, could utilize the given level of the inputs more efficiently by produce higher level of the outputs(in the case of input-oriented model).

Because multiple DMUs could receive the same score, there could be multiple 100% efficient DMUs in the given set. As a result, DEA 'envelops' the data set with the boundary points represented by the 100% efficient DMUs, which are assigned scores of "1", while the rest of the DMUs, being relatively inefficient, are located off the boundary.

While the enveloping surface of the boundary consists of the relatively efficient DMUs, the shape of the resultant surface largely depends on the assumption of the DEA model regarding the return to scale. Two of the classical DEA models are the CCR (after ***C***harnes, ***C***ooper and ***R***hodes) model, which assumes constant return to scale (CRS), and

the more flexible BCC (after ***B***anker, ***C***harnes, and ***C***ooper) model, which assumes variable returns to scale (VRS).

Unlike the CCR ratio model, which evaluates the overall efficiency, the BCC model distinguishes between technical and scale inefficiencies. In this context 'scale efficiency' represents a measure of the deviation from the constant return to scale, and 'technical efficiency' represents a greatest ratio of the maximum possible output to the actual produced output (in the case of output-oriented model), or the minimal feasible input to the actual used input (for input-oriented model). Consequently, BCC model allows for estimating of the technical efficiency under the conditions of the constant or variable (i.e., increasing or decreasing) return to scale.

Hence, a given DMU is considered to be efficient by CCR model only if is both scale and technically efficient, while for the same DMU to be considered efficient by BCC model it must only be technically efficient (Bowlin 1998). Thus, if a DMU is considered efficient by CCR model, it will also be considered as such by BCC model, while reverse not necessarily being true. Similarly to CCR model, for a given DMU to be qualified as efficient under the input oriented model of BCC, it must simultaneously qualify as an efficient under the output oriented model and vice versa.

The different assumptions regarding the return to scale are achieved by restricting the intensity variable (λ_j and z_j in the Table 1.1) for the given set of DMUs. In our study we determine the relative efficiency scores under the conditions of CRS ($z_j > 0$), VRS ($z_j = 0$), and Non-Increasing Return to scale (NIRS) ($1 > z_j > 0$). We also determine a scale efficiency (SE), which is the measure of the deviation from the constant return to scale.

2.2 Cluster Analysis (CA)

Clustering is a popular data mining technique (e.g. Rai et al., 2005; Okazaki, 2005; Wallace et al., 2004; Cristofor and Simovici, 2002; Dhillon, 2001; Ben-Dor and Yakhini, 1999; Huang, 1997; Fisher, 1997; Benfield and Raftery, 1992) that involves the partitioning of a set of objects into a useful set of mutually exclusive clusters such that the similarity between the observations within each cluster (i.e. subset) is high, while the

similarity between the observations from the different clusters is low. There are different reasons for doing clustering, two of which are:

1. Finding a set of natural groups (i.e. segmentation), and the corresponding description of each group. This is relevant if there is the belief that there are natural groupings in the data.
2. Improving the performance of other predictive modeling and DM techniques when there are many competing patterns in the data.

Our interest in this paper is in the first of the reasons, that is identifying natural groups of DMUs based on their structural similarity with regards to the levels of the inputs and outputs that DMUs receive and produce. In particular with regards to a given DEA model, we are interested in knowing if the set of DMUs is homogenous, and if it is not homogenous then what are the appropriate meaningful groups.

There are numerous algorithms available for doing clustering. They may be categorized in various ways such as: hierarchical (e.g. Murtagh, 1983; Ward, 1963) or partitional (e.g. Mc Queen, 1967), deterministic or probabilistic (e.g. Bock, 1996), hard or fuzzy (e.g. Bezdek, 1981; Dave, 1992). We will now provide overview of three general approaches to clustering: hierarchical clustering, partitional clustering (e.g. k-means, k-median), and two-step clustering.

Hierarchical clustering could form clusters by one of the two methods, agglomerative or divisive. Agglomerative method assumes that each data point is its own cluster, and with each step of the clustering process, these clusters are combined to form larger clusters, which are eventually combined to form a single cluster. Divisive method of the hierarchical clustering, on the other hand, starts with the single cluster encompassing all data points within the sample and proceeds to divide it into the smaller dissimilar clusters. Unlike hierarchical clustering, K-means clustering requires the number of resulting cluster, k, to be specified prior to analysis. Thus, k-means clustering will produce *k* different clusters of greatest possible distinction.

A two-way (or two-step) approach could involve using a partitional approach such as k-means to generate the maximum possible number of clusters (i.e. k_{Max}) that would be of interest to the decision-maker followed by the application of an agglomerative clustering

method to combine pairs of clusters until the specified minimum number of clusters (i.e. k_{Min}) is obtained. In this paper for our illustrative we use a two-step approach that involves the use of k-means but we are not claiming that this is the only or always best approach, particularly since for a given dataset it is never clear which approach is the most appropriate.

Clustering approaches that generate partitions (i.e. sets of clusters) of different sizes require some method for determining the most appropriate partition. This may require the use of domain expert knowledge, and/or a simulation based criterion (e.g. Cubic Clustering Criterion) and/or a user specified threshold (e.g. a cluster is an outlier if the percentage of the objects that it includes is less than $\tau_{Outlier}$ of the objects in the entire dataset). Given our interest in determining whether a set of DMUs is homogenous or heterogeneous, we will use a user-specified threshold on outlier size to assess whether a given partition contains outlier clusters, and also use expert knowledge to further assess whether the partition is meaningful.

We shall assume that at the end of the clustering process each object is assigned to a specific cluster, and so each object has an additional cluster identifier attribute (say *ClusterNum*). For our purposes we will assume that this is a unique discrete number (e.g. "*Majority*" for the first cluster, "*Leaders*" for the second cluster). We also assume that the domain expert may associate a a parallel description (say ClusDescr) for each cluster (e.g. 1 for the first cluster, 2 for the second cluster)

2.3 Decision Tree Induction

A decision tree (DT) is a tree structure representation of the given decision problem (e.g. Razi and Athappilly, 2005; Sohn and Moon, 2004;Wu et al, 2006) such that each non-leaf node is associated with one of the decision variables, each branch from a non-leaf node is associated with a subset of the values of the corresponding decision variable, and each leaf node is associated with a value of the target (or dependent) variable. There are two main types of DTs: 1) classification trees and 2) regression trees. For a classification tree, the target variable takes its values from a discrete domain (e.g. Efficient or Inefficient), and for each leaf node the DT associates a probability (and in some cases a value) for each class (i.e. value of the target variable). The class that is

assigned to a given leaf node of the classification tree results from a form of majority voting in which the winning class is the one that provides the largest class probability even if that probability is less than fifty percent (50%). For a regression tree, the target variable takes its values from an interval domain (e.g. Relative Efficiency Score is between 0 and 1). In this paper we will focus on the classification tree, which is the most commonly used type of DT.

Once the decision tree is constructed, it is presented in an easily understandable visual form, which then could be converted into the equivalent, yet more readable set of the *decision rules*. Decision Tree modeling could be used in cases when the dependent variable of the data set is categorical (e.g., "1" if the DMU is "efficient"; "0" if the DMU is "inefficient") or continuous (e.g. any value between "0" and "1"). Classification Tree' models are constructed in the cases when the dependent variable of the data set is categorical, while regression trees are used in the cases of the continuous dependent variable. If used for the purposes of classification, DT allows predicting the membership of a particular case to a group, while in the case of regression DT predicts a value.

In the case of this research, we could potentially use both versions of the DT, for the dependent variable of the data set is the efficiency of a given DMU. In the case of the using direct relative efficiency scores provided by DEA, we would use Regression Tree to create a predictive model. However, we could binarize the assigned relative efficiency scores to create a classification model. We take the latter approach in this paper.

3.0 THE PROPOSED METHODOLOGY

In this section we describe, in step-by-step fashion, the sequence of the procedures constituting the proposed methodology. Before describing our methodology we provide an overview of the dataset that is used in our running illustrative example.

3.1 Overview of Dataset of Illustrative Example:

To illustrate our methodology in action, we have chosen the following problem.

Given a 10-year data set on 18 TEs, spanning a period form 1993 to 2002, we want to find out what are:

1. The differences in the relative efficiencies of these economies regarding their investments in telecoms , and
2. Some of the factors that contribute to the differences in the relative efficiencies.

The data for this study were obtained from two sources. The first source was represented by the database of World Development Indicators, which is the World Bank's comprehensive database on development data. The second source of the data was represented by the *Yearbook of Statistics,* which is published yearly by International Telecommunication Union (ITU). In our choice of variables, we were greatly restricted by the availability of the data. For example, while the development data of the World Bank's database covers more than 600 indicators for 208 economies, data on many of the indicators relevant to our research were not available, or were available only for a few countries, or contained too few data points to be useful in statistical analysis. In terms of the length of the time series, we were restricted to the period from 1992 to 2002, data for which were provided by *Yearbook of Statistics* of ITU.

In our choice of TEs we were guided by the intent to isolate a group of countries that started the process of transition in approximately the same time. As a result, we have decided to concentrate on the 25 countries of the former Soviet block. Based on the availability of the data, the following 18 transitional economies out of 25 have been selected for this research: Albania, Armenia, Azerbaijan, Belarus, Bulgaria, Czech Republic, Estonia, Hungary, Kazakhstan, Kyrgyz Republic, Latvia, Lithuania, Moldova, Poland, Romania, Slovak Republic, Slovenia, and Ukraine. Despite the original intent, the data offered for 7 out of 25 TEs, namely, Tajikistan, Turkmenistan, Uzbekistan, Georgia, Macedonia, Russian Federation and Croatia, turned out to be insufficient to allow the inclusion of these economies in this study. For the DEA part of the methodology we have identified a model consisting of the input variables and output variables that are listed in Table 2.1:

Table 2.1: List of Variables for DEA Models

Role	Subset of Variables
Input	GDP per capita (in current US $), Full-time telecommunication staff(% of total labor force), Annual telecom investment per telecom worker, Annual telecom investment(% of GDP in current US $), Annual telecom investment per capita, Annual telecom investment per worker
Output	Total telecom services revenue per telecom worker, Total telecom services revenue(% of GDP in current US $), Total telecom services revenue per worker, Total telecom services revenue per capita

3.2 Description of the Methodology:

Our methodology has the following three major steps that will be described in detail:

1. Determine the Structural Homogeneity Status of the Dataset
2. Determine the Relative Efficiency Status of each DMU
3. Describe the Relative Efficiency Categories

These steps require the initialization of relevant parameters including:

- k_{Max}: the maximum possible number of clusters that would be of interest to the decision-maker. This parameter is required in Step 1.
- $\tau_{Outlier}$: the threshold that is used to determine if a cluster is an outlier. This parameter is required in Step 1.
- *DMU_Goal*: This could be "Input Orientation" or "Output Orientation". This parameter is required in Step 2.
- *DMU_Criterion*: This could be CRS, VRS, or NIRS. This parameter is required in Step 2.

3.2.1 Step 1: Determine the Structural Homogeneity Status of the Dataset

3.2.1.1 Description of Step 1:

SubStep 1a:

a) Apply two-step approach to generate segmentattions of sizes k_{Max} through k_{Min}.

b) Set $k = k_{\text{Max.}}$

SubStep 1b:

Examination the segmentation with k clusters.

IF $k > 1$ and there is at least one cluster is that consists of less than $\tau_{Outlier}$ percent of the DMUs

THEN

Set $k = k - 1$;

Repeat Substep 1b

ELSE

Current Segmentation with k clusters provides the 'natural' groupings of the DMUs;

Terminate Step 1.

3.2.1.2 Justification of Step 1:

The intended purpose of this Cluster Analysis step 1 of our methodology is to investigate a 'structural similarity' of the dataset. Structural similarity of the DMUs reflects not the types, not the transformation of the inputs into outputs, but the levels of the inputs and outputs that DMUs receive and produce. Consequently, in the first step of our methodology we aim to determine whether all DMUs in the sample are similar in terms of the levels of the received inputs and the levels of the produced outputs. It could be suggested that the purpose of CA is to test the assumption of the homogeneity of the domain given the chosen DEA model. Consequently, homogeneity of the domain of the DMUs is always going to be relative to the given DEA model.

Thus, the required prerequisite to the first step in our methodology is that an investigator has identified a DEA model, i.e., a set of the inputs and a set of the outputs, which are going to be used in the step 2. Once the model is determined, the actual data set that is going to be used to perform DEA is subjected to CA. During the first stage of CA we suggest to start, assuming that an investigator uses a software package allowing to conduct CA, with generating automatically a baseline clustering solution. Once it is done and the certain large number of the clusters has been generated, an investigator should evaluate the membership of each cluster. We suggest, as a rule of thumb, not to retain any cluster with the number of DMUs less than $\tau_{Outlier}$ percent of the total data set.

By gradually decreasing the number of clusters over the iterations of CA, an investigator would get one of the two types of the CA solutions. First type is reflected by the presence of a single large cluster, containing, as a rule of thumb, (100 - $\tau_{Outlier}$) percent or more of the DMUs, and one or two small clusters, with combined membership of (again as a rule of thumb) $\tau_{Outlier}$ percent or less of the sample. If this is the case, we suggest that an investigator should treat the sample of DMUs as homogenous.

In the case of the second type of CA solution, however, an investigator ends up with two or more clusters with the membership of more than 10% of the sample in each cluster. A possible case in this scenario is when the final solution is presented in the form of two or three large clusters and a single small cluster (less than 10 % of the sample). In this situation we suggest a visual inspection of the solution (the diagrams are provided by most of the software packages), or examination of the distances between the clusters; based on the results of the evaluation a small cluster should be combined with the closest large cluster.

If the CA yields a second type of a solution, we suggest that the sample should be treated as heterogeneous and this heterogeneity should be reflected in labeling the resulting clusters as “Cluster1”, “Cluster 2 “ etc. Consequently, an investigator should document the membership of the each cluster, by noting which DMUs belong to which cluster. After it is done, we proceed to the next step, DEA.

3.2.1.2 Illustration of Step 1:

For our illustration of this step we decided to use following eight variables to inquire into the homogeneity of our dataset:

1. Total telecom services revenue(% of GDP in current US $),
2. Total telecom services revenue per capita (Current US $),
3. Total telecom services revenue per worker (Current US $),
4. Total telecom services revenue per telecom worker (Current US $),
5. Annual telecom investment per capita (Current US $),
6. Annual telecom investment (% of GDP in current US $),
7. Annual telecom investment per worker (Current US $),
8. Annual telecom investment per telecom worker (Current US $).

The parameter $\tau_{Outlier}$ was set to 10%, k_{Max} was set to 5 and k_{Min} to 2. We used SAS Enterprise Miner(EM) to perform cluster analysis of the data set. The variables that we used are not measured on the same scale, so, prior to cluster analysis we transformed the data by standardizing the variables. We started our analysis by choosing "Automatic" setting, which did not require any input regarding the desired number of clusters from the researcher. Summary information regarding each obtained solution is compiled in the Table 2.2a.

Table 2.2a: Summary Output of Clustering

Number of Clusters	Number of DMUs in each Cluster	Top-level Split Variables
5 clusters	10, 11, 20, 3, 136	Total telecom service revenue per worker Annual telecom investment per worker Annual telecom investment (% of GDP)
4 clusters	10, 32, 3, 135	Total telecom service revenue per worker Annual telecom investment per worker
3 clusters	30, 3, 147	Total telecom services revenue per worker
2 clusters	72, 108	Annual telecom investment per telecom worker

Clusters with less than $\tau_{Outlier}$ percent of the DMUs are underlined

By using cluster analysis, we were able to come up with a solution that partitions our data set into two clusters. The membership of each cluster is provided in the Table 2.2b. Based on the compiled information we can see, that while some of the TEs are 'permanent residents' of one cluster, other TEs are 'migrants', i.e., they change the cluster membership depending on a year.

Table 2.2b: Cluster Membership when Number of Cluster = 2

Contents of the 1st cluster	**Contents of the 2nd cluster**
Albania (1993-2002)	Czech rep (1993-2002)
Armenia(1993-2002)	Estonia (1994-2002)
Azerbaijan(1993-2002)	Hungary (1993-2002)
Belarus(1993-2002)	Bulgaria (2002)
Bulgaria(1993-2001)	Latvia (1994, 1995, 1997-2002)
Slovak Rep(1993,1994, 1999)	Lithuania (1999-2002)
Kazakhstan(1993-2002)	Slovenia (1993-2002)
Kyrgyz Rep (1993-2002)	Poland (1993-2002)
Latvia (1993, 1996)	Slovak Rep(1995-1998, 2000-2002)
Lithuania (1993-1998)	
Moldova (1993-2002)	
Romania(1993-2002)	
Ukraine (1993-2001)	

Finally, we should ask ourselves the following question: What is the significance of the separation of 18 transitional economies into the two clusters? One of the possible answers is provided in the research by Piatkowski (2003b), who concluded that in the period "between 1995 and 2000 ICT capital has most potently contributed to output growth in the Czech Republic, Hungary, Poland, and Slovenia." Thus, it could be suggested that we were able to separate 18 transitional economies into the two groups, one group of transitional economies that benefits the most from the investments in telecom, and another group where the benefits are less pronounced. Consequently, we labeled the members of the Cluster 1 as "*Majority*" and the members of the Cluster 2 as "*Leaders*".

3.2: Step 2 - Determine the Relative Efficiency Status of DMUs

3.2.2.1 Description of Step 2:

Based on the values of the parameters *DMU_Goal* (i.e. "Input Orientation" or "Output Orientation") and *DMU_Criterion* (i.e. CRS, VRS, or NIRS), the relevant DEA approach is performed in order to obtain the relative efficiency scores of the DMUs in the sample. We assume that at the end of the DEA process each object is assigned a specific relative efficiency status (relatively efficient, or relatively inefficient), and so each object has an additional relative efficiency status attribute (say *EfficiencyStatus* with the value "1" for Relatively Efficient, and the value "0" for Relatively Inefficient).

3.2.2.2 Illustration of Step 2:

For our illustration of this step, given the values of the parameters *DMU_Goal* and *DMU_Criterion* , the relevant DEA models need to be constructed.

Although not required, we generated the relative efficiencies of DMUs using both input-oriented and output-oriented DEA models, for three types of conditions, constant (CRS), variable (VRS), and non-increasing return to scale (NIRS).

Thus, we generated altogether six DEA models, with the following settings of the parameters *DMU_Goal* and *DMU_Criterion*:

1. *DMU_Goal* = "Input Orientation" and *DMU_Criterion* = CRS
2. *DMU_Goal* = "Input Orientation" and *DMU_Criterion* = VRS
3. *DMU_Goal* = "Input Orientation" and *DMU_Criterion* = NIRS
4. *DMU_Goal* = "Output Orientation" and *DMU_Criterion* = CRS
5. *DMU_Goal* = "Output Orientation" and *DMU_Criterion* = VRS
6. *DMU_Goal* = "Output Orientation" and *DMU_Criterion* = NIRS

This step does not differ in any way or form from the regular DEA, thus, no adjustments are necessary. We also computed the average relative efficiencies of the two clusters identified in the Step 1, and compare the averaged relative efficiencies of these two groups of TEs produced by the DEA. Table 2.3a demonstrates the differences in the

relative efficiency between the "Leaders" and the "Majority" in terms of the utilization of the inputs. The input-orientation does not concern itself with the maximization of the outputs, but rather with maximization of the utilization of the inputs. Thus, it is probably reflective of the perspective of the policy maker, especially in the case when the available resources are limited. Table 2.3b demonstrates the differences in the relative efficiency between the "Leaders" and the "Majority" in terms of the maximization of the outputs. Unlike in the case of the input-oriented model, the output-orientation does not concern itself with the efficient utilization of the inputs, but rather with the maximization of the outputs. Thus, it is probably reflective of the perspective of the investor, especially in the case when the resources are abundant and the primary goal is to obtain the maximum revenue.

Table 2.3a DEA: Comparison of the Clusters based on the Input-oriented Model

Criterion for Comparison	**"Leaders" Cluster**	**"Majority" Cluster**	**Difference**	**Difference %**
Average efficiency score, CRS	0.89	0.79	0.10	12.54%
Average efficiency score, VRS	0.95	0.88	0.07	7.48%
Average efficiency score, NIRS	0.89	0.80	0.09	11.63%
Average efficiency score, SE	0.94	0.89	0.04	4.96%

Table 2.3b: Comparison of the clusters based on the Output-Oriented DEA Model

Criterion for Comparison	**"Leaders" Cluster**	**"Majority" Cluster**	**Difference**	**Difference %**
Average efficiency score, CRS	**1.17**	**1.41**	**-0.24**	**-16.71%**

Average efficiency score, VRS	**1.15**	**1.29**	**-0.14**	**-11.00%**
Average efficiency score, NIRS	**1.16**	**1.36**	**-0.19**	**-14.29%**
Average efficiency score, SE	**1.02**	**1.10**	**-0.07**	**-6.81%**

Out of the six DEA models generated in the Step 2 we have arbitrarily chosen the model *DMU_Goal* = "Input Oriented" and *DMU_Criterion* = "CRS" to illustrate Step 3 of our methodology.

3.2.3 Step 3: Describe the Relative Efficiency Categories

3.2.3.1 Description of Step 3:

In this step we will use decision tree induction to generate rules that can describe the relative efficiency categories in terms of the input and output variables of the DEA models. This will require the inclusion of a target variable (say "*EfficiencyCategory*") that identifies the efficiency category. For the case of the homogeneous sample of DMUs, we would use a binary variable to indicate whether the DMU is relatively efficient. For a homogenous dataset, we have to use a categorical variable that allows for two efficiency categories per cluster. Thus, in the case if the CA in the Step 1 resulted in the solution with two clusters, the domain of values for our categorical target variable could be represented as follows:

"11" – Relatively Efficient DMU with membership in the Cluster 1,

"10" – Relatively Inefficient DMU with membership in the Cluster 1,

"21" – Relatively Efficient DMU with membership in the Cluster 2,

"20" – Relatively Inefficient DMU with membership in the Cluster 2.

If we do not take into consideration heterogeneity of the sample, than we end up with the following domain of values:

"1" – Relatively Efficient DMU in the sample,

"2" – Relatively Inefficient DMU in the sample.

In general we will assume that our target variable (say *EfficiencyCategory*) is a concatenation of the relevant cluster identifier attribute *ClusterNum* and relative efficiency status attribute *EfficiencyStatus*. Once the dataset has been amended to include the target variable *EfficiencyCategory*, DT induction is used to generate a DT that can be used to describe the efficiency categories.

3.2.2.2 Illustration of Step 3:

For our illustration of this step, since the result of Step 1 indicated that our dataset was heterogeneous with two groups, we first populated our target variable "Cluster Efficiency" with the following values:

- Value of "21" was assigned to the "*Efficient Leaders*", those TEs that belong to the "leaders" cluster and were assigned the score of "1" by DEA
- Value of "20" was assigned to the "*Inefficient Leaders*" , those TEs that belong to the "leaders" cluster and were assigned the score of less than "1" by DEA
- Value of "11" was assigned to the "efficient majority", those TEs that belong to the "majority" cluster and were assigned the score of "1" by DEA
- Value of "10" was assigned to the "*Inefficient Majority*", those TEs that belong to the "*Majority*" cluster and were assigned the score of less than "1" by DEA

The number (N) of DMUs in each of our four categories are: 38 *Efficient Leaders*, 31 *Inefficient Leaders*, 43 *Efficient Majority*, and 68 *Inefficient Majority*.

We then generated our DT model that enabled us to identify conditions associated with our four categories. In Table 2.4 we display a subset of the rules that are associated with this DT. For each category, we selected a pair of rules, each of which had a strong probability (i.e. **Prob** > 0.90) for the occurrence of the associated category given the condition component of the rule. We provide the complete DT, as well as a complete set of the corresponding English rules, in the Appendix of this paper.

Table 2.4: Pairs of Rules that Describe the Efficiency Categories

Condition	Efficiency Category		N	Prob
	Group	Efficiency Status		
Productivity Ratio per Telecom Worker ≥ 4.1754445351 & Annual Telecom Investment per Worker ≥ $58	*Leader*	Efficient	14	1.00
Total Telecom Services Revenue per person ≥ $210 & Full-Time Telecommunication Staff % ≥ 0.0039016912 & Productivity Ratio per Telecom Worker < 4.1754445351 & Annual Telecom Investment per Worker ≥ $58	*Leader*	Inefficient	5	1.00
Full-Time Telecommunication Staff % < 0.0039016912 & Productivity Ratio per Telecom Worker < 4.1754445351 & Annual Telecom Investment per Worker ≥ $58	*Leader*	Inefficient	11	1.00
Full-Time Telecommunication Staff % < 0.0031414015 & Productivity Ratio per Telecom Worker ≥ 3.8043909395 & GDP per Capita ≥ $519 & Annual Telecom Investment per Worker < $33	*Majority*	Efficient	8	1.00
Total Telecom Services Revenue ≥ 0.0118204323 & GDP per Capita < $519 & Annual Telecom Investment per Worker < $33	*Majority*	Efficient	22	1.00
Productivity Ratio per Telecom Worker < 3.8043909395 & GDP per Capita ≥ $519 & Annual Telecom Investment per Worker < $33	*Majority*	Inefficient	39	1.00
Full-Time Telecommunication Staff % ≥ 0.0031414015 & Full-Time Telecommunication Staff %< 0.0054371357 & Productivity Ratio per Telecom Worker ≤3.8043909395 & GDP per Capita < $519 & Annual Telecom Investment per Worker < $33	*Majority*	Inefficient	12	0.92
Productivity Ratio per Telecom Worker < 2.002357802 & $33 ≤ Annual Telecom Investment per Worker < $58	*Majority*	Inefficient	6	1.00

4.0 CONCLUSION

DEA is a good data analytic tool discriminatory power of which, however, is somewhat dependent on the homogeneity of the sample of DMUs. Relative efficiencies of the NFL players, for example, could not be meaningfully compared unless the position played, which could be dependent on height, speed, and weight, is taken into consideration. Similarly, relative efficiencies of the designated hitters and catchers of

MLB in terms of the production of the home runs should not be compared directly either. Thus, necessity arises to find a way to conduct DEA while taking into consideration some of the important differences between the groups of the DMUs in the sample.

In the case of this study, we have identified two sub groups within our sample of 18 TEs, the "*Majority*" and the "*Leaders*." Having this insight in mind, we have three options of conducting DEA. First, we could disregard this information and proceed directly with DEA. Second, we may conduct separate DEA per each cluster. Third, we have an option of proceeding according to the proposed in this paper methodology.

We would like to argue that our methodology allows for achieving of a better discriminatory power of DEA than its two alternatives. In order to do so, we offer a comparison of the pairs of rules that were obtained by utilizing each of the three mentioned above options. In order to avoid repeating contents of the Table 2.4 in this section of the paper, we provide only the sets of the decision rules corresponding to two other options.

1st option : **DT based on DEA only** ("efficiency" encoded as "1"-efficient, "0" – inefficient), 57 efficient DMUs, 123 inefficient DMUs

Table 4.1 1st option: **Pairs of Rules that Describe the Efficiency Categories**

Condition	**Efficiency Category**		**N**	**Prob**
	Group	**Efficiency Status**		
Full-time telecommunication staff % < 0.0041924905 & Productivity ratio per telecom worker ≤ 5.4191734889	*Whole Set*	Inefficient	25	0.92
Total telecom services revenue ≤ $51,715,239 & Total telecom services revenue per telecom worker < $60,794 & Productivity ratio per telecom worker < 5.4191734889	*Whole Set*	Inefficient	91	0.956

The results of this DEA demonstrate that a number of transitional economies have obtained a rating of being relatively efficient. It does not mean, however, that all of the countries that were deemed relatively efficient are in fact efficient. A common characteristic of DEA models is that they tend to evaluate as efficient those DMUs that have the smallest input values, or, the DMUs with the largest outputs (Ali 1994). Thus, the approach consistent with the 1st option does not allow an investigator to determine

whether the relative efficiency of a DMU is caused indeed by its efficiency, or whether a DMU was awarded a relatively efficient status because it had the smallest level of the inputs in the sample.

Consequently, one of the shortcomings of the first DT model is that it does not allow an investigator to incorporate the results of the CA, i.e., an "*Efficient Leader*" is highly likely to be very different from an "*Efficient Majority*".

2nd option: **DT based on DEA, one DT model per cluster**

a) "*Leaders*" cluster ("efficiency" encoded as "1"-efficient, "0" – inefficient), 38 efficient DMUs, 31 inefficient DMUs

Table 4.2 2nd option: **Pairs of Rules that Describe the Efficiency Categories in "Leaders" cluster**

Condition	**Efficiency Category**		**N**	**Prob**
	Group	**Efficiency Status**		
Total telecom service revenue per capita ≤$101 & Productivity ratio per telecom worker < 4.1754445351 & Annual telecom investment %GDP < 0.0166 & Full-time telecommunication staff %≤0.0043712305	*Leaders*	Efficient	12	1.0
Productivity ratio per telecom worker ≥ 4.1754445351	*Leaders*	Efficient	17	1.0
Full-time telecommunication staff % < 0.0043712305 & Productivity ratio per telecom worker < 4.1754445351	*Leaders*	Inefficient	21	0.905

b) "*Majority*" cluster ("efficiency" encoded as "1"-efficient, "0" – inefficient), 43 efficient DMUs, 68 inefficient DMUs

Table 4.3 2nd option: **Pairs of Rules that Describe the Efficiency Categories in "Majority" cluster**

Condition	**Efficiency Category**		**N**	**Prob**
	Group	**Efficiency Status**		
Productivity ratio per telecom worker < 3.8106041166 & GDP per capita ≤ $519	*Majority*	Inefficient	47	1.0
Total telecom services revenue %≤0.0118204323 & GDP per capita < $519	*Majority*	Efficient	22	1.0

These models, consistent with the 2nd option, improve on the first model by being cluster-specific. Let us recall, nevertheless, that while some of the TEs are "permanent residents" of one cluster, the other TEs are "migrants," for they change their membership depending on the year. Consequently, none of the generated by this approach models would help us to inquire into the question why, for example, Lithuania was a member of the "*Majority*" cluster for the period from 1993 to 1998, but became a member of the "*Leaders*" in the period from 1999 to 2002.

Considering the presented above alternatives, it would appear that the results of our methodology, presented in the Table 2.4, allow for achieving of a higher discriminatory power of DEA. While the results of the conventional DEA only yield the efficiency scores for each DMU in the sample, the results of our approach yield the efficiency scores and, in the case of the heterogeneous domain, the membership within the subset of the sample for each DMU. This (considering that the most decision makers are interested not so much in learning whether or not a given DMU is inefficient, but rather why it is inefficient) allows for a higher degree of granularity of the subsequent analysis, as it was demonstrated by comparing Tables 2.4, 4.1, 4.2, and 4.3.

Another benefit of our methodology is that it takes an approach of an 'external augmentation' of DEA, meaning, it does not require any changes to or alterations of DEA itself. As a result, our methodology is not model-specific and, consequently, could be applied to any DEA model. However, despite the contributions that this research makes, we must acknowledge that our study is not without its limitations.

First limitation of this research is associated with the use of CA. At this point, we cannot offer strict criteria determining whether the sample should be considered homogeneous or heterogeneous. Thus, despite the explicit nature of testing for heterogeneity by means of CA, the determining decision regarding the sample still lies with the decision maker.

Second limitation of our study is associated with the use of DT induction. At this point, we cannot suggest to a decision maker what splitting criteria and what settings yield a better tree. As a result, this issue as well resides in the domain of the responsibilities of the decision maker.

REFERENCES

1. Ali, A.I., (1994). "Computational aspects of DEA." In: Charnes, A., Cooper, W.W., Lewin, A., Seiford, L.M. (Eds.), Data Envelopment Analysis: Theory, Methodology and Applications. Kluwer Academic Publishers, Boston, pp. 63–88.
2. Banfield, J. and Raftery, A. (1992) "Identifying Ice Floes in Satellite Images", ***Naval Research Reviews 43***, 2-18.
3. Ben-Dor, A. and Yakhini, Z. (1999) "Clustering Gene Expression Patterns", ***Proceedings of the 3rd Annual International Conference on Computational Molecular Biology*** *(RECOMB 99)*, 11-14, Lyon, France.
4. Bezdek, J. (1981) ***Pattern Recognition with Fuzzy Objective Function Algorithms***. Plenum Press, New York, NY.
5. Bock, H. (1996) "Probability Models in Partitional Cluster Analysis", ***Computational Statistics and Data Analysis 23***, 5-28.
6. Bowlin, W. F. (1998). "Measuring Performance: An Introduction to Data Envelopment Analysis (DEA)", *Journal of Cost Analysis*, Fall 1998, pp. 3-27.
7. Charnes, A. and Cooper, WW. (1962). "Programming with linear fractional functionals", *Naval Research Logistics Quarterly*, Vol.9, pp.181–186.
8. Charnes, A., Cooper, W.W., Lewin, A.Y. and Seiford, L.M. (1994), "*Data Envelopment Analysis: Theory, Methodology and Applications",* Kluwer Academic Publishers, Norwell, MA.
9. Cristofor, D. and Simovici, D. (2002) "An Information-Theoretical Approach to Clustering Categorical Databases using Genetic Algorithms", ***Proceedings of the SIAM DM Workshop on Clustering High Dimensional Data***, 37-46. Arlington, VA.
10. Cooper, W.W., Seiford, L.M. and Zhu, Joe. (2004). "Data envelopment analysis: History, Models and Interpretations", in Handbook on Data Envelopment Analysis, eds W.W. Cooper, L.M. Seiford and J. Zhu, Chapter 1, pp.1-39, Kluwer Academic Publishers, Boston. 2004.
11. Dave, R. (1992) "Generalized Fuzzy C-Shells Clustering and Detection of Circular and Elliptic Boundaries", ***Pattern Recognition 25***, 713–722.

12. Dhillon, I. (2001) "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning", ***Proceedings of the 7th ACM SIGKDD***, 269-274, San Francisco, CA.
13. Dula, J.H., (2002a),"Data Envelopment Analysis (DEA)," in Handbook of Applied Optimization, P.M. Pardalos and M.G.C. Resende, (eds.), pp. 531--543, Oxford University Press, New York, 2002. ISBN 0-19-512594-0.
14. Dula, J.H., (2002b), "Computations in DEA", *Pesquisa Operacional*, Vol. 22, no.2, pp.165-182, Dec. 2002, ISSN 0101-7438.
15. Hong, H., Ha, S., Shin, C., Park, S., and Kim, S. (1999) "Evaluating the Efficiency of System Integration Projects using Data Envelopment Analysis (DEA) and Machine Learning", ***Expert Systems with Applications 16*****,** 283–296.
16. Huang, Z. (1997) "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining", ***Proceedings SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery***, Tech. Report 97-07, UBC, Dept. of CS.
17. McQueen, J. (1967) "Some Methods for Classification and Analysis of Multivariate Observations", In: Lecam, L.M. and Neyman, J. (Eds.): ***Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability***, 281-297.
18. Murtagh, F. (1983) "A Survey of Recent Advances in Hierarchical Clustering Algorithms which Use Cluster Centers", ***Computer Journal 26***, 354–359.
19. Okazaki, S. (2005) "What Do We Know About Mobile Internet Adopters? A Cluster Analysis", ***Information & Management***, in press.
20. Piatkowski, M. (2003b). "Does ICT Investment Matter for Output Growth and Labor Productivity in Transition Economies?" *TIGER Working Paper Series*, No. 47. December. Warsaw. Available at www.tiger.edu.pl
21. Rai, A., Tang, X., Brown, P. and Keil, M. (2005) "Assimilation Patterns in the Use of Electronic Procurement Innovations: A Cluster Analysis", ***Information & Management***, in press.

22. Razi, M. and Athappilly, K. (2005) "A Comparative Predictive Analysis of Neural Networks, Nonlinear Regression and Classification and Regression Tree (CART) Models", ***Expert Systems with Applications 29:1,*** 65-74.
23. Sengupta JK. (1996). "Data Envelopment analysis: a new tool for improving managerial efficiency", *International Journal of Systems Science*,Vol.27, No.12,pp.1205–1210.
24. Seol, H., Choi, J., Park, G. and Park, Y. (2006) **"**A Framework for Benchmarking Service Process using Data Envelopment Analysis and Decision Tree**",** ***Expert Systems with Applications*****,** in press.
25. Sohn, S. and Moon, T. (2004) "Decision Tree based on Data Envelopment Analysis for Effective Technology Commercialization", ***Expert Systems with Applications 26:2,*** 279-284.
26. Wallace, L. Keil, M. & Rai, A. (2004) "Understanding Software Project Risk: A Cluster Analysis", ***Information & Management 42***, 115–125.
27. Ward, J. (1963) "Hierarchical Grouping to Optimize An Objective Function", ***J. Am. Stat. Assoc. 58***, 236–244.
28. Wu, M.-C., Lin, S.-Y. and Lin, C.-H. (2006) "An Effective Application of Decision Tree to Stock Trading", ***Expert Systems with Applications 31:2,*** 270-274.

Appendix

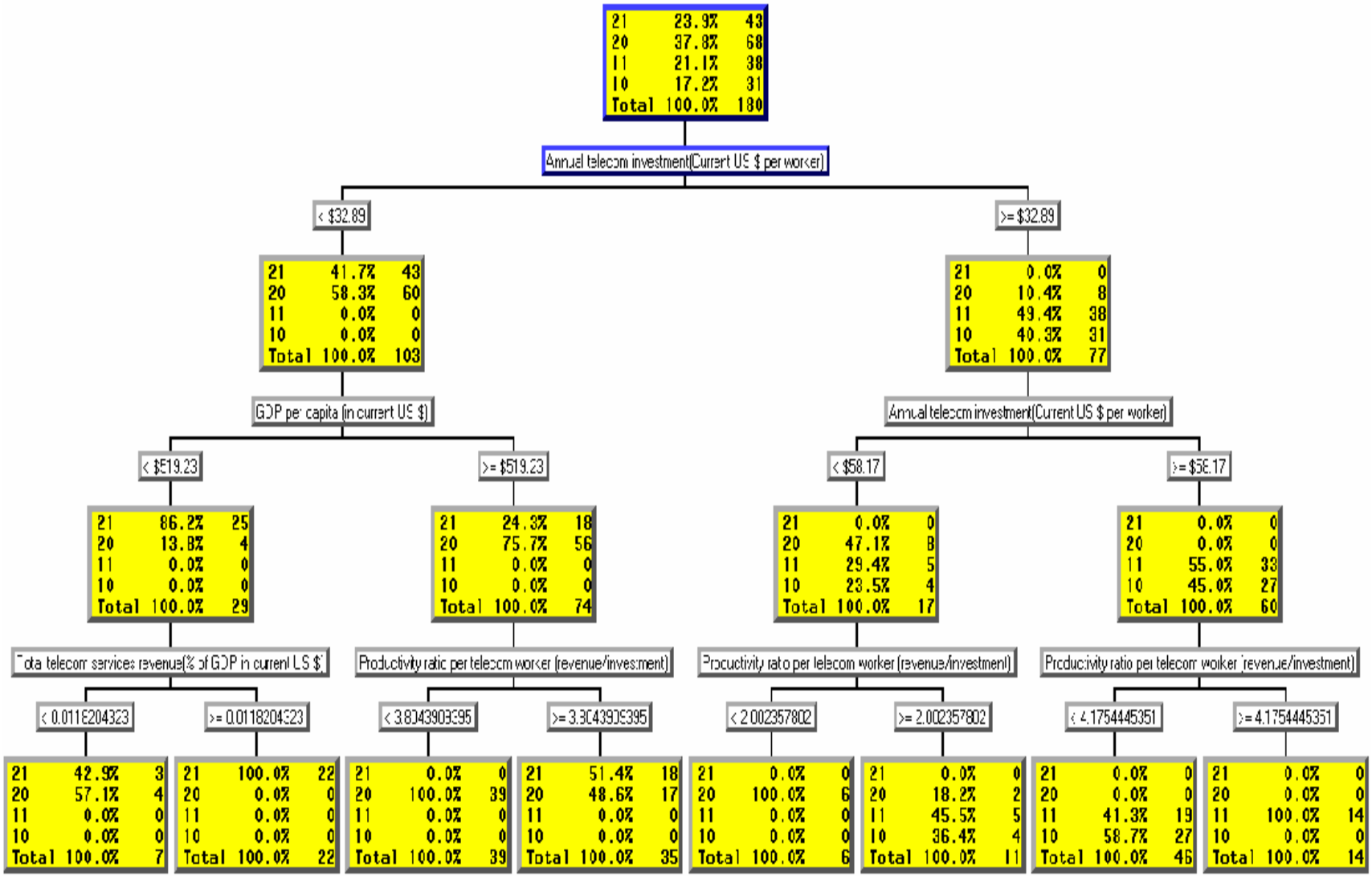
English Rules

```
IF  Total telecom services revenue(% of GDP in current US $) <
0.0118204323
AND GDP per capita (in current US $) <                  $519
AND Annual telecom investment(Current US $ per worker) <
$33
THEN
  NODE    :       8
  N       :       7
  21      :   42.9%
  20      :   57.1%
  11      :    0.0%
  10      :    0.0%

IF  0.0118204323 <= Total telecom services revenue(% of GDP in current
US $)
AND GDP per capita (in current US $) <                  $519
AND Annual telecom investment(Current US $ per worker) <
$33
THEN
  NODE    :       9
  N       :      22
  21      :  100.0%
  20      :    0.0%
  11      :    0.0%
  10      :    0.0%

IF  Productivity ratio per telecom worker (revenue/investment) <
3.8043909395
AND                  $519 <= GDP per capita (in current US $)
AND Annual telecom investment(Current US $ per worker) <
$33
THEN
  NODE    :      10
  N       :      39
  21      :    0.0%
  20      :  100.0%
  11      :    0.0%
  10      :    0.0%

IF  Productivity ratio per telecom worker (revenue/investment) <
2.002357802
AND                   $33 <= Annual telecom investment(Current US $ per
       worker) <                   $58
THEN
  NODE    :      12
  N       :       6
  21      :    0.0%
  20      :  100.0%
  11      :    0.0%
```

```
  10      :    0.0%

IF  4.1754445351 <= Productivity ratio per telecom worker
(revenue/investment)
AND                   $58 <= Annual telecom investment(Current US $ per
        worker)
THEN
  NODE    :      15
  N       :      14
  21      :    0.0%
  20      :    0.0%
  11      :  100.0%
  10      :    0.0%

IF  Full-time telecommunication staff(% of total labor force) <
0.0031414015
AND 3.8043909395 <= Productivity ratio per telecom worker
(revenue/investment)
AND                  $519 <= GDP per capita (in current US $)
AND Annual telecom investment(Current US $ per worker) <
$33
THEN
  NODE    :      16
  N       :       8
  21      :  100.0%
  20      :    0.0%
  11      :    0.0%
  10      :    0.0%

IF  Annual telecom investment(Current US $ per telecom worker)
    <                $9,450
AND  2.002357802 <= Productivity ratio per telecom worker
(revenue/investment)
AND                   $33 <= Annual telecom investment(Current US $ per
        worker) <                   $58
THEN
  NODE    :      18
  N       :       5
  21      :    0.0%
  20      :   40.0%
  11      :   60.0%
  10      :    0.0%

IF                $9,450 <= Annual telecom investment(Current US $ per
        telecom worker)
AND  2.002357802 <= Productivity ratio per telecom worker
(revenue/investment)
AND                   $33 <= Annual telecom investment(Current US $ per
        worker) <                   $58
THEN
  NODE    :      19
  N       :       6
  21      :    0.0%
  20      :    0.0%
  11      :   33.3%
  10      :   66.7%
```

```
IF  Full-time telecommunication staff(% of total labor force) <
0.0039016912
AND Productivity ratio per telecom worker (revenue/investment) <
4.1754445351
AND                   $58 <= Annual telecom investment(Current US $ per
        worker)
THEN
  NODE    :      20
  N       :      11
  21      :    0.0%
  20      :    0.0%
  11      :    0.0%
  10      :  100.0%

IF  0.0031414015 <= Full-time telecommunication staff(% of total labor
force)
    < 0.0054371357
AND 3.8043909395 <= Productivity ratio per telecom worker
(revenue/investment)
AND                  $519 <= GDP per capita (in current US $)
AND Annual telecom investment(Current US $ per worker) <
$33
THEN
  NODE    :      22
  N       :      12
  21      :    8.3%
  20      :   91.7%
  11      :    0.0%
  10      :    0.0%

IF  Total telecom services revenue(Current US $ per person)
    <                   $210
AND 0.0039016912 <= Full-time telecommunication staff(% of total labor
force)
AND Productivity ratio per telecom worker (revenue/investment) <
4.1754445351
AND                   $58 <= Annual telecom investment(Current US $ per
        worker)
THEN
  NODE    :      24
  N       :      30
  21      :    0.0%
  20      :    0.0%
  11      :   63.3%
  10      :   36.7%

IF                   $210 <= Total telecom services revenue(Current US
$ per
        person)
AND 0.0039016912 <= Full-time telecommunication staff(% of total labor
force)
AND Productivity ratio per telecom worker (revenue/investment) <
4.1754445351
AND                   $58 <= Annual telecom investment(Current US $ per
        worker)
THEN
  NODE    :      25
```

```
  N       :       5
  21      :    0.0%
  20      :    0.0%
  11      :    0.0%
  10      :  100.0%

IF  Total telecom services revenue(Current US $ per worker)
    <                   $43
AND 0.0054371357 <= Full-time telecommunication staff(% of total labor
force)
AND 3.8043909395 <= Productivity ratio per telecom worker
(revenue/investment)
AND                  $519 <= GDP per capita (in current US $)
AND Annual telecom investment(Current US $ per worker) <
$33
THEN
  NODE    :      28
  N       :       5
  21      :   20.0%
  20      :   80.0%
  11      :    0.0%
  10      :    0.0%

IF                    $43 <= Total telecom services revenue(Current US
$ per
        worker)
AND 0.0054371357 <= Full-time telecommunication staff(% of total labor
force)
AND 3.8043909395 <= Productivity ratio per telecom worker
(revenue/investment)
AND                  $519 <= GDP per capita (in current US $)
AND Annual telecom investment(Current US $ per worker) <
$33
THEN
  NODE    :      29
  N       :      10
  21      :   80.0%
  20      :   20.0%
  11      :    0.0%
  10      :    0.0%
```


21 23.9% 43
20 37.8% 68
11 21.1% 38
10 17.2% 31
Total 100.0% 180
Annual telecom investment(Current US $ per worker)
< $32.89
>= $32.89
21 41.7% 43
20 58.3% 60
11 0.0% 0
10 0.0% 0
Total 100.0% 103
21 0.0% 0
20 10.4% 8
11 49.4% 38
10 40.3% 31
Total 100.0% 77
GDP per capita (in current US $)
Annual telecom investment(Current US $ per worker)
< $519.23
>= $519.23
< $58.17
>= $58.17
21 86.2% 25
20 13.8% 4
11 0.0% 0
10 0.0% 0
Total 100.0% 29
21 24.3% 18
20 75.7% 56
11 0.0% 0
10 0.0% 0
Total 100.0% 74
21 0.0% 0
20 47.1% 8
11 29.4% 5
10 23.5% 4
Total 100.0% 17
21 0.0% 0
20 0.0% 0
11 55.0% 33
10 45.0% 27
Total 100.0% 60
Total telecom services revenue(% of GDP in current US $)
Productivity ratio per telecom worker (revenue/investment)
Productivity ratio per telecom worker (revenue/investment)
Productivity ratio per telecom worker (revenue/investment)
< 0.0118204323
>= 0.0118204323
< 3.8043909395
>= 3.8043909395
< 2.002357802
>= 2.002357802
< 4.1754445351
>= 4.1754445351
21 42.9% 3
20 57.1% 4
11 0.0% 0
10 0.0% 0
Total 100.0% 7
21 100.0% 22
20 0.0% 0
11 0.0% 0
10 0.0% 0
Total 100.0% 22
21 0.0% 0
20 100.0% 39
11 0.0% 0
10 0.0% 0
Total 100.0% 39
21 51.4% 18
20 48.6% 17
11 0.0% 0
10 0.0% 0
Total 100.0% 35
21 0.0% 0
20 100.0% 6
11 0.0% 0
10 0.0% 0
Total 100.0% 6
21 0.0% 0
20 18.2% 2
11 45.5% 5
10 36.4% 4
Total 100.0% 11
21 0.0% 0
20 0.0% 0
11 41.3% 19
10 58.7% 27
Total 100.0% 46
21 0.0% 0
20 0.0% 0
11 100.0% 14
10 0.0% 0
Total 100.0% 14