Dr. Gürdal Ertek
gurdalertek.org
Working Papers
research.sabanciuniv.edu

Sabancı Üniversitesi

Akçay, A.E., Ertek, G., Büyüközkan, G. “Analyzing the solutions of DEA through information visualization and data mining techniques: SmartDEA framework”. “Expert Systems with Applications”. 39, pp. 7763–7775.

Note: This is the final draft version of this paper. Please cite this paper (or this final draft) as above. You can download this final draft from http://research.sabanciuniv.edu.

Analyzing the solutions of DEA through information visualization and data mining techniques: SmartDEA framework

Alp Eren Akçaya[a,1], Gürdal Ertek[a], Gülçin Büyüközkan[b]

[a]Sabancı University, Faculty of Engineering and Natural Sciences,
Orhanli, Tuzla, 34956 Istanbul, Turkey

[b]Department of Industrial Engineering, Galatasaray University, Çıragan Cad.
No: 36, 34257 Ortaköy, Istanbul, Turkey

[1]Present address: Carnegie Mellon University, Tepper School of Business, 5000

Forbes Avenue, Pittsburgh, 15213 PA, USA.

Analyzing the solutions of DEA through information visualization and data mining techniques: SmartDEA framework

Alp Eren Akçay$^{(a)1}$, $GürdalErtek^{(a)*}$, $GülçinBüyüközkan^{(b)}$

[a] *Sabancı University, Faculty of Engineering and Natural Sciences Orhanli, Tuzla, 34956, Istanbul, Turkey.*
[b] *Department of Industrial Engineering, Galatasaray University, Çırağan Cad No:36, 34257, Ortaköy, Istanbul, Turkey.*

Abstract

Data envelopment analysis (DEA) has proven to be a useful tool for assessing efficiency or productivity of organizations, which is of vital practical importance in managerial decision making. DEA provides a significant amount of information from which analysts and managers derive insights and guidelines to promote their existing performances. Regarding to this fact, effective and methodologic analysis and interpretation of DEA solutions are very critical. The main objective of this study is then to develop a general decision support system (DSS) framework to analyze the solutions of basic DEA models. The paper formally shows how the solutions of DEA models should be structured so that these solutions can be examined and interpreted by analysts through information visualization and data mining techniques effectively. An innovative and convenient DEA solver, SmartDEA, is designed and developed in accordance with the proposed analysis framework. The developed software provides a DEA solution which is consistent with the framework and is ready-to-analyze with data mining tools, through a table-based structure. The developed framework is tested and applied in a real world project for benchmarking the vendors of a leading Turkish automotive company. The results show the effectiveness and the efficacy of the proposed framework.

Keywords: Data envelopment analysis (DEA), information visualization, data mining, decision support system framework

[*] Corresponding author. ertekg@sabanciuniv.edu, Tel: +90(216)483-9568, Fax: +90(216)483-9550

[1] Present address: Carnegie Mellon University, Tepper School of Business, 5000 Forbes Avenue, Pittsburgh, 15213, PA, USA

1. Introduction

Data Envelopment Analysis (DEA) is a widely used method in performance evaluation and benchmarking of a set of entities. The popularity of DEA can be easily confirmed in the article of Emrouznejad et al. (2008) that has summarized previous DEA contributions during the past three decades. Its convenience in assessing the multiple input and output variables of these entities by not requiring congruity and an apriori relationship makes it a very popular management tool in many application areas. Another reason for its wide use is the managerial insights that come up with the solution of a DEA model. For instance, DEA assigns a peer group or reference set for an inefficient entity. The entity can take the entities in this reference set as role models in accordance with the assigned weights. Another important DEA result is the target values or projections for the input and output variables of an inefficient entity to achieve full efficiency. DEA thus provides significant amount of information from which analysts and managers derive insights and guidelines to enhance their existing performances. Regarding to this fact, effective and methodologic analysis and interpretation of DEA solutions is very critical (El-Mahgary and Lahdelma, 1995; Emrouznejad and De Witte, 2010; Yadav et al., 2010).

The main contribution in this study is the development of a general framework that enables DEA analysts to extract the most important and interesting insights in a systematic manner. In order to do so, a computer science and data mining perspective is adopted for designing the structure of the DEA results. Various data mining and information visualization techniques can be appropriate for the analysis of different types of DEA models (Lin et al, 2008; Seol et al., 2011). The paper provides a fundamental basis for the implementation of these techniques in the DEA solutions. A convenient and general notation is proposed for the DEA data included in the model, the other data and the results data generated by DEA solvers. The ultimate goal of the study is then to build a structure framework for the analysis of DEA results, enabling researchers and practitioners to make analytical benchmarking and performance evaluations.

In accordance with the proposed framework, a user-friendly and convenient DEA software, SmartDEA, is designed and developed. The software generates DEA solutions with a structure consistent with the framework. The analysis steps performed by a DEA analyst are importing the model data, analyzing the data by solving the appropriate DEA model, and making analytical inquires on the generated solution data to evaluate and benchmark the entities assessed in DEA model. The solution data generated by the software allows analysts to integrate the results and many of the data mining and information visualization techniques in a convenient and effective manner.

The rest of this paper is organized as follows: After a brief intoduction to the DEA, the theory of the basic DEA models are explained in Section 2, providing a fundamental background for DEA. Since it is critical to know when and where DEA is an appropriate method, advantages and drawbacks of DEA are also explained in the same section. An overview of DEA literature and existing DEA software is given in Section 3. Proposed framework, which is based on the integration of DEA results with data mining and information visualization techniques, is presented in Section 4. The developed DEA solver, SmartDEA, and its testing with the real world data of an automotive company are highlighted in Section 5. Concluding remarks are summarized in Section 6.

2. Data Envelopment Analysis (DEA)

2.1. Introduction to the DEA

DEA is a nonparametric performance evaluation technique with a wide range of application areas within various disciplines. In their breakthrough study, Charnes, Cooper, and Rhodes (1978) define DEA as "a mathematical programming model applied to observational data and a new way of obtaining empirical estimates of relations such as the production functions and/or efficient production possibility surfaces". Since its introduction in 1978, researchers in a multitude of disciplines distinguished and adopted DEA as a promising tool that is easily applicable to their fields for performance evaluation and benchmarking of entities and operational processes. A given set of entities, referred to as *Decision Making Units* (DMU) in DEA terminology, can be conveniently compared in terms of multiple inputs and multiple outputs, assuming neither a specific form of relationship between inputs and outputs, nor fixed weights for the inputs and outputs of a DMU.

DEA firstly provides an *efficiency score* between 0 and 1 for each DMU involved in the analysis. The efficiency score for a DMU is determined by computing the ratio of total weighted outputs to total weighted inputs for it. DEA enables variable weights, which are calculated in such a way that the efficiency score for the DMU is maximized. (Cooper et al., 2006). For a DEA model with *n* different DMUs, *n* different linear programming (LP) optimization models are solved to compute the efficiency scores of each of the DMUs.

A basic DEA model can provide important metrics and benchmarks for monitoring the comparative performances of entities in a group and take managerial actions to improve them. An *efficient frontier* or *envelopment surface*, which is drawn over the "best" DMUs, is the critical component of a DEA model. It is formed by the efficient DMUs which have efficiency scores of 1. The efficiency score of a DMU is basically the distance from each DMU to this efficient frontier. The efficiency scores of the inefficient DMUs are calculated in accordance with this distance represented as a Pareto ratio. Besides the efficiency scores, another result offered by DEA is the *reference sets*, or peer (sub)groups. For each inefficient unit, DEA identifies a set of corresponding efficient units that can serve as a benchmark peer group for the selected DMU. The solution of the LP formulation of the model results in the reference set for each DMU. Knowing that the DMUs in the reference set are efficient and have the same input and output structure, they can be regarded as "good" examples operating practices for the corresponding inefficient DMU (Boussofiane et al., 1991). The percentages of each reference set unit contributing to the composite unit (i.e. virtual producer - with respect to which the efficiency score of the inefficient DMU is found) are also computed by the DEA model.

One important shortcoming of DEA is the rank ordering of the efficiency scores for each DMU. Each inefficient DMU is labeled as such only in comparison to the given group, and this may be misleading: An inefficient DMU may have been labeled as efficient if it were part of an inferior group. Conversely, an efficient DMU may not neccessarily be efficient when compared as a part of another group. Also, theoretically, only the DMUs with the same reference sets can be strictly rank ordered (Avkıran, 1999). Yet another important result DEA provides is the *target inputs* and *target outputs*, which are referred to as *projections*. They represent up to which value an input should be decreased while keeping the outputs at the the same levels (i.e. input orientation) or how much an output should be increased while the input level remains unincreased (i.e. output orientation), respectively, so that the DMU becomes efficient.

2.2. Basic DEA Models

Since its introduction by Charnes, Cooper and Rhodes (1978) in their seminal work *Measuring Efficiency of Decision Making Units*, the CCR model served as the origin of many following ideas and models in DEA literature. Cooper et al. (2006) discusses the model in detail together with the classical alternative models. The theoretical formulations and summaries in this section are based on Cooper et al. (2006).

Let's suppose that there are n DMUs in the model: $\mathrm{DMU_1}, \mathrm{DMU_2}, \ldots, \mathrm{DMU_n}$. Suppose there are m inputs ans s outputs for each one of them. For $\mathrm{DMU_j}$ the inputs and outputs are represented by (x_{1j}, x_{2j}, x_{mj}) and (y_{1j}, y_{2j}, y_{sj}), respectively. As stated previously, for each $\mathrm{DMU_o}$, DEA tries to maximize the ratio

$$\frac{\text{Virtual output}}{\text{Virtual input}} \tag{1}$$

where

$$\text{Virtual input} = v_1 x_{1o} + \ldots + v_m x_{mo} \tag{2}$$

$$\text{Virtual output} = u_1 y_{1o} + \ldots + u_s y_{so} \tag{3}$$

and the weights v_i and u_r are not fixed in advance. Best weights are assigned according to the solution of the following fractional DEA model (**M1**):

$$\begin{aligned}
(FP_o) \quad & \max \quad \theta = \frac{u_1 y_{1o} + \ldots + u_s y_{so}}{v_1 x_{1o} + \ldots + v_m x_{mo}} \\
\text{s.t.} \quad & \frac{u_1 y_{1o} + \ldots + u_s y_{so}}{v_1 x_{1o} + \ldots + v_m x_{mo}} \leq 1 \\
& v_1, v_2, \ldots, v_m \geq 0 \\
& u_1, u_2, \ldots, u_s \geq 0
\end{aligned}$$

The fractional model can be transformed into the linear model shown below (**M2**):

$$\begin{aligned}
(LP_o) \quad & \max \quad \theta = u_1 y_{1o} + \ldots + u_s y_{so} \\
\text{s.t.} \quad & v_1 x_{1o} + \ldots + v_m x_{mo} = 1 \\
& u_1 y_{1j} + \ldots + u_s y_{sj} \leq v_1 x_{1j} + \ldots + v_m x_{mj} \\
& v_1, v_2, \ldots, v_m \geq 0 \\
& u_1, u_2, \ldots, u_s \geq 0
\end{aligned}$$

The fractional model (M1) is equivalent to linear model (M2), and *Unit Invariance Theorem* states that the optimal values of $max\ \theta = \theta^*$ in M1 and M2 are independent of the units in which the inputs and outputs are measured, under the requirement that these units are the same for every DMU.

The input and output data can be arranged in matrix notation X and Y, respectively:

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

$$Y=\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ y_{s1} & y_{s2} & \cdots & y_{sn} \end{pmatrix}$$

Based on the input and output matrices, the linear CCR model (**M2**) can be rewritten as follows with row vector $\boldsymbol{v}$ for input multipliers and row vector $\boldsymbol{u}$ for output multipliers. This form is called *multiplier form* where multipliers $\boldsymbol{u}$ and $\boldsymbol{v}$ are treated as variables (**M3**):

$$\begin{array}{lll} (LP_o) & \max & \boldsymbol{u y_o} \\ & \text{s.t.} & \boldsymbol{v x_o = 1} \\ & & -\boldsymbol{v}X + \boldsymbol{u}Y \leq 0 \\ & & \boldsymbol{v} \geq \boldsymbol{0} \\ & & \boldsymbol{u} \geq \boldsymbol{0} \end{array}$$

The dual of model M3 which is expressed with a real variable θ and a nonnegative vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$ of variables is named *envelopment form* (**M4**):

$$\begin{array}{lll} (DLP_o) & \min & \theta \\ & \text{s.t.} & \theta \boldsymbol{x_o} - X\boldsymbol{\lambda} \geq \boldsymbol{0} \\ & & Y\boldsymbol{\lambda} \geq \boldsymbol{y_o} \\ & & \boldsymbol{\lambda} \geq \boldsymbol{0} \end{array}$$

In model M4, the objective is to guarantee at least the output level $\boldsymbol{y_o}$ of DMU$_o$ in all dimensions while reducing the input vector $\boldsymbol{x_o}$ radially (i.e. proportionally) as much as possible. Thus, M4 is referred as *CCR-Input* model.

The *slack vector* is defined as:

$$\boldsymbol{s^-} = \theta \boldsymbol{x_o} - X\boldsymbol{\lambda} \tag{4}$$

$$\boldsymbol{s^+} = Y\boldsymbol{\lambda} - \boldsymbol{y_o} \tag{5}$$

where $s^- \in R^m$ are the input *excesses* and $s^+ \in R^s$ are the output *shortfalls*.

For an inefficient DMU$_o$, its *reference set* E_o is defined as follows:

$$E_o = \left\{ j \mid \lambda_j^* > 0 \right\} \quad (j \in \{1, \ldots, n\}) \tag{6}$$

An inefficient DMU can be improved either by increasing the output values by the output shortfalls in $\boldsymbol{s^+}$ or reducing the input values in proportional with θ^* and the input excesses in $\boldsymbol{s^-}$ are removed.

That is, following *CCR Projections* formulas are applied:

$$\widehat{\boldsymbol{x_o}} \Leftarrow \theta^* \boldsymbol{x_o} - \boldsymbol{s^-} \leq \boldsymbol{x_o} \tag{7}$$

$$\widehat{\boldsymbol{y_o}} \Leftarrow \boldsymbol{y_o} + \boldsymbol{s^+} \geq \boldsymbol{y_o} \tag{8}$$

DEA models discussed so far deal with reducing input variables while attaining at least the provided output levels. On the other hand, there exists another type of model which aims to maximize output levels while spending no more than existing resources or inputs. This kind of models are referred as *output-oriented* and *CCR-Output* model is given below (**M5**):

$$\begin{aligned} (DLPO_o) \quad \max \quad & \eta \\ \text{s.t.} \quad & \boldsymbol{x_o} - X\boldsymbol{\mu} \geq \boldsymbol{0} \\ & \eta \boldsymbol{y_o} - Y\boldsymbol{\mu} \leq \boldsymbol{0} \\ & \boldsymbol{\mu} \geq \boldsymbol{0} \end{aligned}$$

In DEA, θ^* represents the input reduction rate whereas η^* stands for the output enlargment rate (Cooper et al., 2006). It is clear that the less the η^* value, the more efficient the DMU, or vice versa. In order to obtain an efficiency score of between 0 and 1 and to relate the efficiency scores with the input oriented model, $1/\eta^*$ is used to express the efficiency score of the DMU in the output oriented model. Models M4 and M5 do not change the DMUs located on the efficient frontier; however target values of projections are calculated in a different way (Ulus et al., 2006).

The slack vector of the output-oriented model is defined by:

$$\begin{aligned} \boldsymbol{t^-} &= \boldsymbol{x_o} - X\boldsymbol{\mu} && (9) \\ \boldsymbol{t^+} &= Y\boldsymbol{\mu} - \eta\boldsymbol{y_o} && (10) \end{aligned}$$

For the output oriented model, the projections or target values for input and output variables are calculated as follows:

$$\begin{aligned} \widehat{\boldsymbol{x_o}} &\Leftarrow \boldsymbol{x_o} - \boldsymbol{t^-} && (11) \\ \widehat{\boldsymbol{y_o}} &\Leftarrow \eta^* \boldsymbol{y_o} + \boldsymbol{t^+} && (12) \end{aligned}$$

Input and output oriented CCR models which are discussed so far rely on the assumption of *constant returns to scale* (CRS), which typically is not a realistic assumption for real world problems. The reason is that it is difficult to expect an increase/decrease in output variables proportional to an augmentation/reduction in input variables in most of the cases. Another widely known DEA model developed by Banker, Charnes and Cooper (1984) enables the problem to be *variable returns to scale* (VRS). In BCC models, production frontiers are spanned by the convex hull of the existing DMUs and these frontiers have piecewise linear and concave characteristics (Cooper et al.,2006). The BCC models differs from the previously mentioned CCR models with a convexity condition in its constraints. The linear programming model of input-oriented BCC model (*BCC-Input*) is given below (**M6**) where $\boldsymbol{e}$ is the unit vector that has 1 at all indices:

$$\begin{aligned} (BCC_o) \quad \min \quad & \theta_B \\ \text{s.t.} \quad & \theta_B \boldsymbol{x_o} - X\boldsymbol{\lambda} \geq \mathbf{0} \\ & Y\boldsymbol{\lambda} \geq \boldsymbol{y_o} \\ & \boldsymbol{e}\boldsymbol{\lambda} = 1 \\ & \boldsymbol{\lambda} \geq \mathbf{0} \end{aligned}$$

For an inefficient DMU in BCC model, its reference set, E_o is defined based on an optimal solution $\boldsymbol{\lambda^*}$ as in Eq.6. The projection formulas to set the target input and output variables are Eq.7 and Eq.8.

The envelopment form of the output-oriented BCC model is formulated as

$$
\begin{array}{lll}
(BCC - O_o) & \max & \eta_B \\
 & \text{s.t.} & \boldsymbol{x_o} - X\boldsymbol{\lambda} \geq \mathbf{0} \\
 & & \eta_B \boldsymbol{y_o} - Y\boldsymbol{\lambda} \leq \mathbf{0} \\
 & & \boldsymbol{e\lambda} = 1 \\
 & & \boldsymbol{\lambda} \geq \mathbf{0}
\end{array}
$$

BCC versions of the Eq.11 and Eq.12 are used to calculate the projection values of the input and output variables of an inefficient DMU in the BCC-Output model.

In addition to the characteristics of each model and the preference of input or output orientation, the signs of input and output data are also taken into account in the selection of the model that will be used in an analysis. Table 1 summarizes the requirements on the signs of data that will be included in the model. *Semi-p* stands for semipositive data. Semipositive data means a set of nonnegative data points in which at least one of them is positive. For instance, consider a variable that is included as an input in the model. It may take a nonnegative value (i.e. either zero or a greater value), but at least for one DMU it has to take a positive value. *Free* allows the use of either negative, positive or zero data.

Table 1: Characteristics of cases handled by DSS

Model Data	CCR-I	CCR-O	BCC-I	BCC-O
X	Semi-p	Semi-p	Semi-p	Free
Y	Free	Free	Free	Semi-p

2.3. Advantages and Disadvantages of DEA

DEA provides significant opportunities in the performance evaluation and benchmarking of entities. Increasing popularity in the literature and the boom in the number of applied DEA studies in last decades show the extent of attention paid by the researchers. However, it is important to remember that DEA can offer reasonable solutions and guidelines only if it is used wisely under appropriate settings. Thus, disadvantages and drawbacks of DEA methodology should be clearly perceived before any prospective analysis.

Let's first summarize the advantages of DEA possesses. First of all, DEA can deal with multiple inputs and outputs without a priori relationship among them. In addition, the units of inputs and outputs do not need to be congruent. Cooper et al. (2006) discuss additional advantages of DEA such as its ability to indicate sources and amounts of inefficiency for each input and output variable belonging to a DMU. In terms of computational requirments, DEA models do not need high performance solvers unless the size of the problem is very large.

Besides the advantages, DEA has the following drawbacks that are common in many models. Availability and reliability of the data needed in the analysis may be a serious problem (Easton et. al., 2002) and missing data points in inputs or outputs of a DMU can lead to its exclusion from the model. DEA is strongly affected by possible errors and extreme points in data (Smith and Goddard, 2002). In addition, DEA can only measure *relative* efficiency. It can not lead to a rank ordering of DMUs by comparing their efficiencies with an absolute theoretical efficiency value. Moreover, even if a DMU emerges with a position on the efficient frontier, it may not perform well in real life. Since it is superior only in a specific observation set of DMUs, the best way to overcome this obstacle is to keep the number of DMUs contained in the analysis large. As a rule of thumb, the number of DMUs in the model should exceed the sum of number of inputs and outputs several times.

3. Integration of DEA Results with Data Mining and Information Visualization

3.1. Extensions in the Analysis of DEA Results

Emrouznejad and De Witte (2010) underlined that in large and complicated datasets, a standard process could facilitate performance assessment and help to (1) translate the aim of the performance measurement to a series of small tasks, (2) select homogeneous DMUs and suggest an appropriate input/output selection, (3) detect a suitable model, (4) provide means for evaluating the effectiveness of the results, and (5) suggest a proper solution to improve the efficiency and productivity of entities (also called DMUs). As running a DEA model does not suffice for a meaningful analysis, in last steps of the study, the model and its results should be carefully reviewed according to the core objective of the study.

The meaning of DEA results needs interpretation for the transformation of mathematical terms into managerial insights to assess and improve the performances of inefficient DMUs. Significant amount of data that come by DEA results are open to further detailed analysis for the derivation of interesting insights and guidelines. Many of the data mining and information visualization techniques are very effective tools for this analysis. Representation of DEA results in accordance with on-line analytical processing (OLAP) technology can enable the managers and analysts to involve in faster and better decision making processes by performing multidimensional analysis of DEA results data.

DEA solutions include efficiency scores, reference sets for each DMU, and projection values for the input and output variables belonging to each DMU. They provide important in-depth information about the system; however, some hidden patterns or important insights in the DEA results data may remain undiscovered, unless state-of-the-art visualization and data mining techniques are employed.

Visualizing data helps decision makers to discover patterns and trends, improving the efficiency and effectiveness of the decision maker (Tagerden, 1999; Keim et al., 2005; Woo et al., 2005; Kim et al., 2008; Russell et al., 2008). Spence (2001), Keim (2002) and Chen (2004) discuss information visualization as one of the important growing fields of computer science, combining computer graphics, data mining activities and explanatory data analysis to better understand the data visually.

Ulus et al. (2006) discusses the data visualization as a fundamental concept of data analysis, helping the analyst with detecting outliers, discovering underlying patterns which are not possible to recognise with classical visualization techniques in statistics (i.e. histograms, quantile plots, box plots, symmetry plots etc.), and coming up with new insights and hypotheses. The authors use *colored scatter plot* to help building insights in a benchmarking study of industrial transportation companies traded in the New York Stock Exchange (NYSE). In the study, DEA is used as the primary methodology, and selected financial data of the companies are used as the input and output variables in the DEA model.

Ertek et al. (2007) discover hidden patterns in the benchmarking of Turkish apparel industry through information visualization techniques. In the paper, DEA is applied to determine the efficiency scores of the companies. Inputs, outputs, other related data and efficiency scores are visualized in *colored scatter plots* and *tile graphs* by means of Miner3D (Miner3D) and Visokio Omniscope (Omniscope) software, respectively. The former visualization technique is also referred as *starfield visualization* in information visualization terminology. The *x*-axis and *y*-axis of the starfield can represent various variables of the data on a two dimensional physical space. Similarly, color can reflect the values of another variable and interesting insights or underlying patterns can be discovered through visualization. The latter visualization type, tile graph, basically divides a bounded surface into rectangular regions hierarchically according to a given categorical variables, and allocates the appropriate area for each visualized entity according to a numerical variable. A carefully designed coloring scheme contributes further to the effectiveness of the tile graph.

Visual data mining has become an active research area (Fayyad et al., 1996; Liu and Salvendy, 2007). Fayyad and Grinstein (2002) suggest that without proper visualization techniques, decision making models, although a reduced form of original data, may not give the desired insight to help humans understand the phenomena of interest. Effective and creative data visualization can significantly reduce the time to understand the information and patterns, as well as contribute to the generation of actionable insights. Data visualization can also promote the formation of guidelines about the investigation of the data. Detailed reviews and information about data visualization terminology and applications can be found in the review papers by Keim (2002) and Hoffman and Grinstein (2002), as well as the books by Soukup and Davidson (2002) and Spence (2001).

3.2. Integrated Framework for DEA Results Based on Data Mining and Information Visualization Techniques

In this section, we propose a mathematical notation to formally represent DEA data model and results. This notation will then be used to describe the framework for the generation and representation of DEA results such that further data mining and information visualization processes can be readily conducted. The notation has been developed so as to conform with the widely adopted DEA notation available in the literature.

The framework exhibits the following four distinct methodological contributions: (1) The framework integrates data mining and information visualization with DEA; (2) It guarantees the generation of clean data for data mining, through data auditing at the DEA modeling stage; (3) It allows the incorporation of "other data" (data that does not enter the DEA models, but is related to the DMUs and the analysis) into the analysis process, which can be a great source of novel actionable insights; (4) Finally, the framework can accomodate multiple DEA models within the same analysis, which is much more general than representing the results of only a single model. This way, automation is enabled during the solution and analysis of DEA models, allowing for the benchmark of not only the DMUs, but also alternative DEA models.

In the first part of the framework the notation is given in terms of constants, indices, sets, DMU attributes, DMUs (i.e. representing the DMUs as objects), DEA results data and functions.

CONSTANTS

D	:	Number of distinct DEA models
N	:	Number of DMUs that exist in at least one of the DEA models
n_d	:	Number of DMUs in model d
K	:	Number of attribute groups
A_k	:	Number of atributes in attribute group k
A	:	Number of attributes
m_d	:	Number of inputs in model d
s_d	:	Number of outputs in model d
τ_d	:	Number of other data in model d

INDICES

d	:	Model
j	:	DMU index
h	:	DMU index in a reference set
k	:	Attribute group
l	:	Index in the attribute group k
a	:	Attribute index
i	:	Input index
r	:	Output index
t	:	Other data index

DMU ATTRIBUTES

α	:	The attribute object α
α_{kl}	:	l^{th} attribute of the k^{th} attribute group
α_a	:	a^{th} attribute

SETS

$\mathcal{D}$: Set of all DEA models

$\mathcal{N}$: Set of all DMUs

$\mathcal{N}_d$: Set of DMUs in model d

$\mathcal{I}$: Set of all inputs

$\mathcal{I}_d$: Set of inputs in model d

$\mathcal{R}$: Set of all outputs

$\mathcal{R}_d$: Set of outputs in model d

$\mathcal{T}$: Set of all other data

$\mathcal{T}_d$: Set of other data in model d

$\mathcal{A}$: Set of all attributes

$\mathcal{A}_k$: Set of all attributes in attribute group k

$\mathcal{A} = \cup_{\forall k}\mathcal{A}_k$, (according to their semantic meaning)

$\mathcal{A} = \mathcal{I} \cup \mathcal{R} \cup \mathcal{T}$, (according to their roles in the model)

DMUs

ψ_j : DMU j

Ψ_j : DMUs in the reference set of DMU j

FUNCTIONS

$$\Gamma_m(\alpha) \quad : \quad \begin{cases} 1 & \text{if} \quad \alpha \in \mathcal{I} \\ 2 & \text{if} \quad \alpha \in \mathcal{R} \\ 3 & \text{if} \quad \alpha \in \mathcal{T} \end{cases}$$

$i(\alpha)$: the input index of attribute α in the model given that the attribute is an input

$r(\alpha)$: the output index of attribute α in the model given that the attribute is an output

$$ir(\alpha) \quad : \quad \begin{cases} i(\alpha) & \text{if} \quad \Gamma(\alpha) = 1 \\ r(\alpha) & \text{if} \quad \Gamma(\alpha) = 2 \\ \text{Error} & \text{if} \quad \Gamma(\alpha) = 3 \end{cases}$$

DEA RESULTS DATA

θ_{jd} : Efficiency of j^{th} DMU in d^{th} model

E_{jd} : Reference set of the j^{th} DMU in d^{th} model

λ_{hjd} : Reference weight of the h^{th} DMU in the reference set E_j in model d

x_{ijd} : Original value of input i for j^{th} DMU in model d

$\boldsymbol{x_{jd}}$: Original input vector for j^{th} DMU in model d

$\widehat{x}_{ijd}$: Projection value of input i for j^{th} DMU in modeld

$\boldsymbol{\widehat{x}_{jd}}$: Projection input vector for j^{th} DMU in modeld

y_{rjd} : Original value of output r for j^{th} DMU in model d

$\boldsymbol{y_{jd}}$: Original output vector for j^{th} DMU in model d

$\widehat{y}_{rjd}$: Projection value of output r for j^{th} DMU in model d

$\boldsymbol{\widehat{y}_{jd}}$: Projection output vector for j^{th} DMU in model d

z_{tjd} : Value of other data t for j^{th} DMU in model d

$\boldsymbol{z_{jd}}$: Other data vector for j^{th} DMU in model d

If the model under consideration is known or there is only a single DEA model, the index d can be dropped from the formulas.

Table 2: DEA Results Table Structure 1: Efficiency Scores

Field Index	**Field Description**	**Notation in Framework**
1	DMU Name	ψ_j
2	Efficiency Score	θ_j
Next m fields		$\boldsymbol{x_j}$
Next s fields		$\boldsymbol{y_j}$
Next $\tau fields$		$\boldsymbol{z_j}$

Table 2 gives the structure of results data for *efficiency scores*, according to the proposed framework. This DEA table structure constitutes first part of the framework which is composed of three such data representation table. In designing and representing the structure, a database modeling perspective and related notation have been adopted. In Table 2, a record (row) exists

Table 3: DEA Results Table Structure 2: Reference Sets

Field Index	**Field Description**	**Notation in Framework**
1	DMU Name	ψ_j
2	Efficiency Score	θ_j
3	h^{th} DMU in the reference set E_j of DMU$_j$	Ψ_{hj}
4	Reference weight of h^{th} DMU in the reference set E_j of DMU$_j$	λ_{hj}

Table 4: DEA Results Table Structure 3: Projections

Field Index	**Field Description**	**Notation in Framework**
1	DMU Name	ψ_j
2	Efficiency Score	θ_j
3	Attribute Name	θ
4	Is input or output	$\Gamma(\alpha)$
5	Original value	$x_{ir(\alpha),j}$
6	Projection nalue	$\widehat{x}_{ir(\alpha),j}$
7	Difference	$x_{ir(\alpha),j} - \widehat{x}_{ir(\alpha),j}$
8	Percentage Difference	$100(x_{ir(\alpha),j} - \widehat{x}_{ir(\alpha),j})/x_{ir(\alpha),j}$

for each element of the set $\{j : j \in \mathcal{N}\}$ and the number of records is N.

Table 3 constitutes the second part of the framework. It describes how the *reference sets* are represented in the framework within a table structure.

In Table 3, a record (row) exists for each element of the set $\left\{(j, h) : j \in \mathcal{N}, R \in E_j\right\}$ and the number of records is $\sum_{j \in \mathcal{N}} |E_j|$.

Table 4 constitutes the final part of the framework. It describes how the *projection* variables and their values are represented in the framework within a table structure. In Table 4, a record (row) exists for each element of the set $\{(j, \alpha) : j \in \mathcal{N}, \alpha \in (\mathcal{N} \cup \mathcal{R})\}$ and the number of records is $N \times R$. In the next section, the implementation of this framework is discussed in a real world application by using a novel DEA solver.

4. SmartDEA: The Developed Software

4.1. An Overview of DEA Software

As a technique, DEA mainly relies on the solution techniques of LP models. Availability of a wide range of LP solvers make DEA an affordable and convenient method to quantitatively measure efficiencies of entities. In addition, as the application areas of DEA have expanded and the methodology drew more attention from practitioners, as well as academicans, the extent and quality of DEA software available in the market prominently increased.

Hollingsworth (1997) and Bowlin (1998) review three DEA softare packages: Warwick DEA for Windows, IDEAS, and Frontier Analyst. The former author discussed these three software in detail in terms of the intended use and area of application, ease of use, and package facilities. The most comprehensive DEA software review paper is by Barr (2004). The author surveys approximately 20 DEA software packages, and reviews eight of them in detail according to a comparison scheme. Barr (2004) evaluates the eight DEA software (DEA Solver Pro, Frontier Analyst, OnFront, Warwick DEA, DEA Excel Solver, DEAP, EMS, and Pioneer) according to eight primary criteria:

1. Available models the software offers
2. Key DEA features and capabilities such as non-discretionary or categorical factors, priorities on variables and sensitivity analysis
3. Platform and interoperability
4. Existence of user interface
5. Reporting such as seperate worksheets, customized reports, graphs and charts,
6. Documentation and support such as availability of technical reference manual with modeling details, availability of user guide, technical customer support
7. Test performance

8. Availability (i.e. academic and commercial licencing costs, maintenance cost, and availability of free demo)

Barr divides the eight software into two subgroups: Commercial (EA Solver Pro, Frontier Analyst, OnFront, Warwick DEA) and Noncommercial (DEA Excel Solver, DEAP, EMS, Pioneer). Then, the author details the characteristics of the packages in a comparison chart that is completed according to the fulfillment of the criteria cited above.

An important off-the-shelf DEA package worth mentioning (which was not reviewed by Barr (2004)) is PIM-DEA (Emrouznejad and Thanassoulis, 2010). Emrouznejad and Thanassoulis (2010) suggest several reasons to select and use PIM-DEA: (1) It is an up-to-date DEA software from academics with decades experience in teaching, researching and applying DEA; (2) It has enhanced functionality features reflecting the cumulative feedback from its users since the 1980s; (3) It has extensive data handling facilities for very large data sets (e.g. automatic selection of subsets of units by category for batch runs) and it permits import and export of data to MS Excel and the export of results to MS Excel; (4) Its graphical facilities including illustration and real time updating of the Production Possibility Set and its frontier as the model specification is modified, histogram of efficiencies, trend of efficiencies over time, trend of Malmquist indices over time, trend of efficiency change boundary shift and scale efficiency change over time.

None of the four major contributions of the framework in our paper have been mentioned in any of the earlier reviews (except the automation in PIM-DEA), but these contributions are all crucial for data mining.

4.2. Motivation for SmartDEA

After the advantages and disadvantages of the various state-of-the-art DEA Solvers are carefully reviewed, it is decided to build a completely new DEA software considering the following facts with respect to methodology and application:

1. The developed software must be designed based on a solid framework. The SmartDEA software is based on the framework proposed and presented in this paper. In order to reduce the efforts and save time in analyzing the results of the DEA, an effective and ready-to-import results table structure is needed, as suggested in our framework. The open source and commercial DEA software in the market lack an effective model data and solution representation in their results files. Thus, most of the time further analysis using

these files can be a cumbersome task to arrange the files for information visulization and data mining.

2. The generation of clean data for data mining is extremely important, since data cleaning is typically the most time consuming task in data mining. Dasu and Johnson (2003) report that "exploratory data mining and data cleaning constitute 80% of the effort that determines 80% of the value of the data mining results". The best way to achieve clean results is through data auditing at the DEA modeling stage.
3. Incorporation of "other data" (data that does not enter the DEA models, but is related to the DMUs and the analysis) into the analysis process is crucial. As was observed in earlier studies (Ulus et al., 2006; Ertek et al. 2007), such additional data can be a great source of novel actionable insights.
4. The software and the underlying framework should accomodate multiple DEA models within the same analysis, which is much more general than representing the results of only a single model. This way, automation is enabled during the solution and analysis of DEA models, allowing for the benchmark of not only the DMUs, but also alternative DEA models.
5. Commercial DEA software in the market provides high performance but with a serious licencing cost[2].
6. Freely available solvers exist but with a limited number of DMUs which restricts the analysis we plan to conduct. It is a fact that the limit on DMUs with the free software can be too limiting for a real life application which can have hundreds DMUs.
7. The Graphical User Interface (GUI) design and development is commonly neglected in these software while it is a critical feature for the use of the sofware in industrial applications by practitioners who are not very comfortable with the underlying theory.
8. An innovative, effective and user-friendly software can lead to advancements in future DEA research with different data in other problem scenarios.

4.3. Modeling Process in SmartDEA

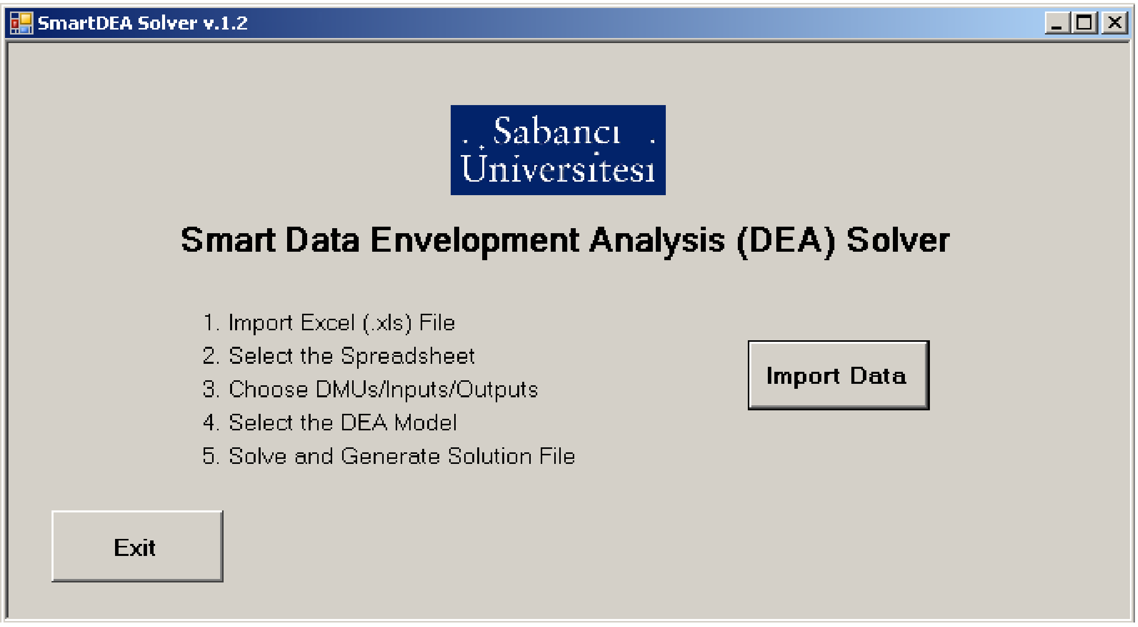
The coding of the SmartDEA decision support system (DSS) was completed in the C# language under MS Visual Studio.NET 2005. As the mathematical programming model solver, the

[2] ranging from £800 to £2500

open source lp_solve dynamic link library (.dll) file is embedded in the software and called whenever the LP model is solved. Thus, the optimization tasks within the software are outsourced to the lp_solve library (LP_Solve). Since the integration of the DEA results with information visualization and data mining tools is the main objective of the framework, special attention was paid in designing the format of the results files. We selected the results file format as MS Excel (.XLS) spreadsheet file format. The widespread use of MS Office (which includes MS Excel) on Windows platforms in industry was the main motivation behind this technology selection. The fact that any user can conveniently create, modify and store data in this file format will let the data import and export with SmartDEA solver be very rapid and in a user-friendly environment. The structure of the results files is designed to visualize and mine the data effectively without further data cleaning or preperation.

There are five main stages in the DEA solution process which is summarized on the first window that appears on the screen after initialization of the software (Figure 1). User selects the data file in MS Excel format to import the DEA data into SmartDEA. After the file is chosen in the file dialog box, a second window appears on the screen, asking the spreadsheet that will be used to construct the DEA model. That is, the developed software supports MS Excel files with multiple spreadsheets; users can place different data into different spreadsheets and they can be separately loaded into the software with this mechanism. In addition, there exists a control about the consistency of the data in the selected spreadsheet. The data should be in the required format, with each of the DMU Name, inputs, outputs and other data fields are arranged in columns. Each row represents the variables for a DMU and no space should be added between any columns on the spreadsheet. If the selected spreadsheet is not in the required format controlled by the consistency check, an error message appears and user is asked to select any other spreadsheet. Figure 2 shows the spreadsheet selection window. In SmartDEA, data cleaning on the results file at the end of the DEA process is eliminated through data auditing at the model construction phase, in the beginning of the DEA process.

After a spreadsheet is selected, if the data on the spreadsheet is in the required format, the third window appears on the screen (Figure 3) in which the user contructs the DEA model. There are three buttons in the model selection group box, representing the DMU names, inputs and outputs. After the user selects the appropriate field in the listbox, it is selected as a DMU name field, input field or output field by clicking on these buttons. There is no need to follow any



Smart Data Envelopment Analysis (DEA) Solver

1. Import Excel (.xls) File
2. Select the Spreadsheet
3. Choose DMUs/Inputs/Outputs
4. Select the DEA Model
5. Solve and Generate Solution File

Figure 1: First window that appears after the initialization of the software

order in model construction but at least one input and one output is required as well as exactly one DMU name field. The reset operation can be performed on any previously selected field by using either Clear Selected Field and Clear All buttons. Any selected field is distinguished from other fields by a coloring scheme. A selected DMU name is colored light purple, while yellow-green and orange colors are reserved for inputs and outputs in the model. Other data that come within the spreadsheet are not colored, distinguishing them from the model data.

The next stage in the SmartDEA solver is the selection of the model type. Classical DEA models, CCR-Input, CCR-Output, BCC-Input and BCC-Output, are available in the solver. The user chooses one of these model types by selecting it in the listbox and clicking on the OK button (Figure 4). Depending on the size of the model, but generally in the order of miliseconds, SmartDEA solver returns the efficiency scores, reference set and projection values for each DMU in the model. DEA solution summary is given on the left side of the solution display window of the software (Figure 5). The efficiency score of each DMU is provided with a self-explanatory coloring scheme which displays efficient DMUs light sky blue and an inefficient DMU in blue. DMU detail group box on the right side of the window provides all information about the se-

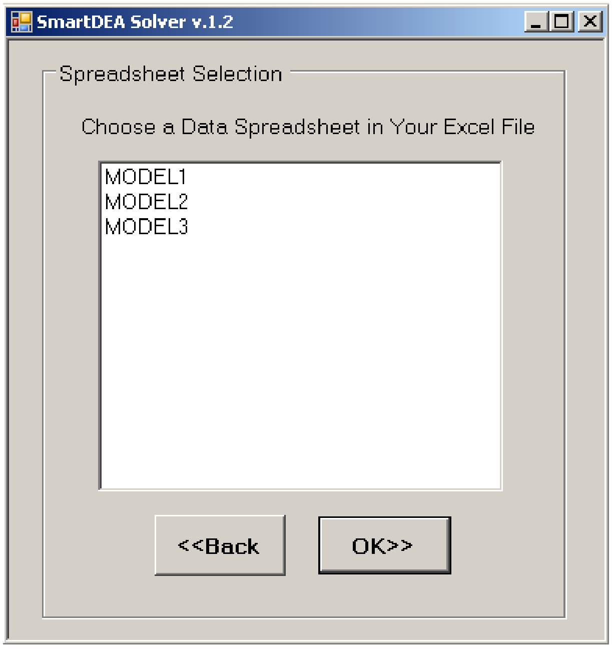
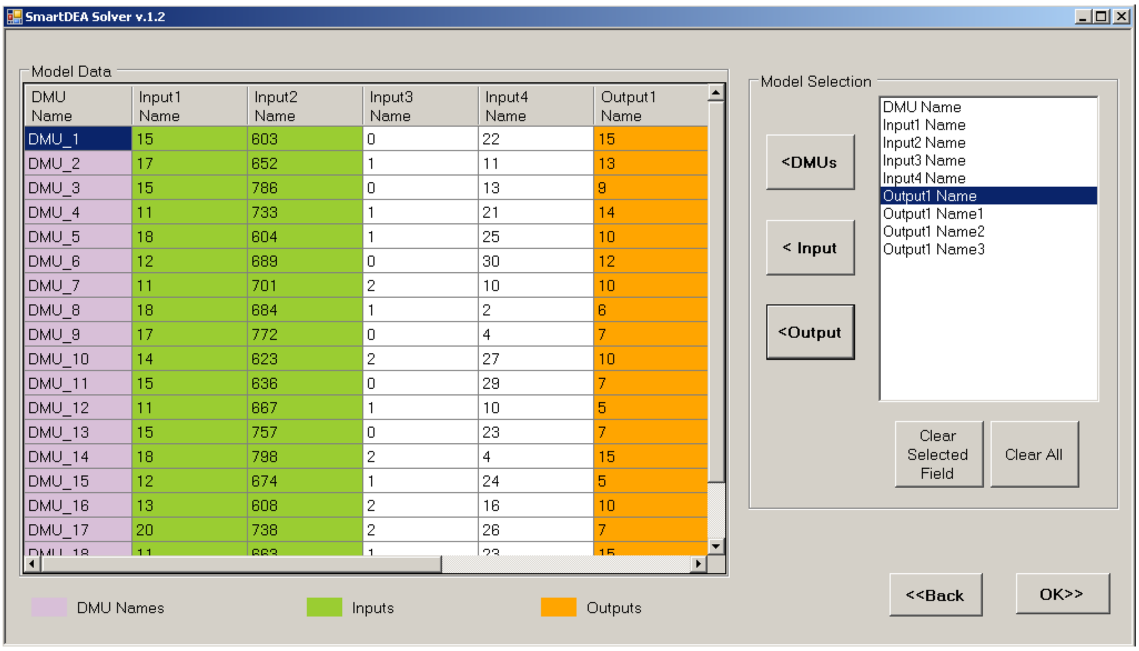


Figure 2: Spreadsheet selection window



DMU Name	Input1 Name	Input2 Name	Input3 Name	Input4 Name	Output1 Name
DMU_1	15	603	0	22	15
DMU_2	17	652	1	11	13
DMU_3	15	786	0	13	9
DMU_4	11	733	1	21	14
DMU_5	18	604	1	25	10
DMU_6	12	689	0	30	12
DMU_7	11	701	2	10	10
DMU_8	18	684	1	2	6
DMU_9	17	772	0	4	7
DMU_10	14	623	2	27	10
DMU_11	15	636	0	29	7
DMU_12	11	667	1	10	5
DMU_13	15	757	0	23	7
DMU_14	18	798	2	4	15
DMU_15	12	674	1	24	5
DMU_16	13	608	2	16	10
DMU_17	20	738	2	26	7
DMU_18	11	663	1	23	15

🟪 DMU Names 🟩 Inputs 🟧 Outputs

Figure 3: Model construction window

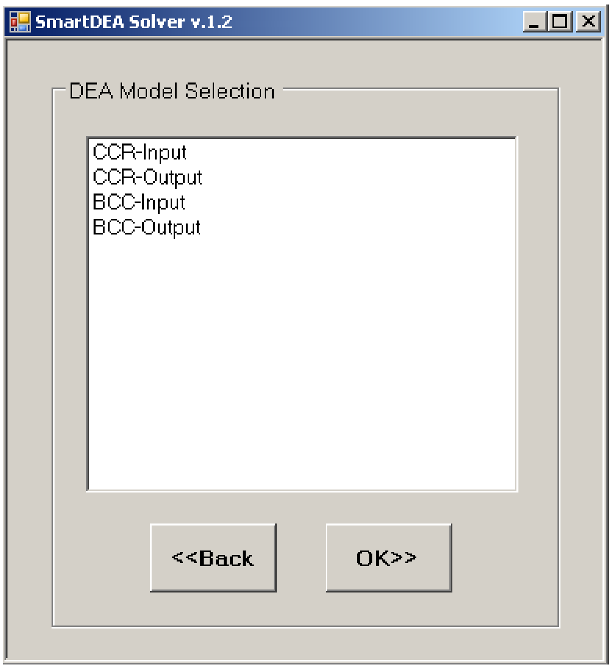
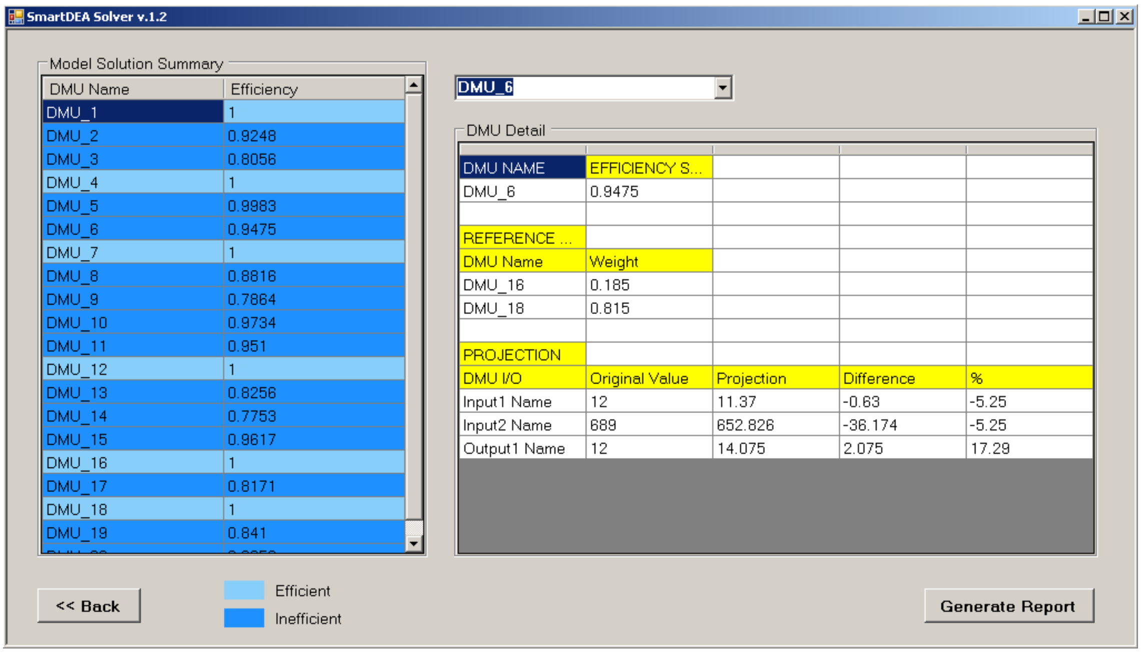


Figure 4: DEA model type selection window



Model Solution Summary

DMU Name	Efficiency
DMU_1	1
DMU_2	0.9248
DMU_3	0.8056
DMU_4	1
DMU_5	0.9983
DMU_6	0.9475
DMU_7	1
DMU_8	0.8816
DMU_9	0.7864
DMU_10	0.9734
DMU_11	0.951
DMU_12	1
DMU_13	0.8256
DMU_14	0.7753
DMU_15	0.9617
DMU_16	1
DMU_17	0.8171
DMU_18	1
DMU_19	0.841

🟦 Efficient

🟦 Inefficient

DMU_6

DMU Detail

DMU NAME	EFFICIENCY S...			
DMU_6	0.9475			
REFERENCE ...				
DMU Name	Weight			
DMU_16	0.185			
DMU_18	0.815			
PROJECTION				
DMU I/O	Original Value	Projection	Difference	%
Input1 Name	12	11.37	-0.63	-5.25
Input2 Name	689	652.826	-36.174	-5.25
Output1 Name	12	14.075	2.075	17.29

Figure 5: Solution display window of the software

Table 5: Data structure of “Efficiency Scores” spreadsheet

Field No	Field Description
1	DMU Names
2	Efficiency Scores
Next m+s+τ	Model Data (Inputs, outputs and other data are highlighted)

Table 6: Data structure of “Reference Sets” spreadsheet

Field No	Field Description
1	DMU Names
2	Efficiency Scores
3	Reference DMU$_i$
4	λ_i

lected DMU. In addition to the reference set and reference weight information, target values or projections of each input and output variable are given at the end of the grid view. Absolute and percentage differences between the original and target values of input and output variables are also displayed on the window. The user can access all this information about a DMU either by selecting it in the combobox and displaying them on the window or by generating a report file in the MS Excel format.

Once the user clicks on the Generate Report button, an .XLS file is formed with three distinct spreadsheets named as *Efficiency Scores*, *Reference Sets* and *Projections*. Each one of them stores the output data of the SmartDEA solver in a database format, that is each variable (field) is in a different column with each row represents a unique information about the solution. The data structure of each generated spreadsheet is given in Table 5, Table 6 and Table 7, respectively. Variable names and their descriptions of them are provided in these tables. These table structures enable the user to quickly and conveniently import the results data into visualization and data mining tools and perform exlanatory, descriptive, and predictive data mining.

As noticed in the stages of solution process described above, users move to the next stage progressively after the previous one is completed appropriately. It is also available for users to go

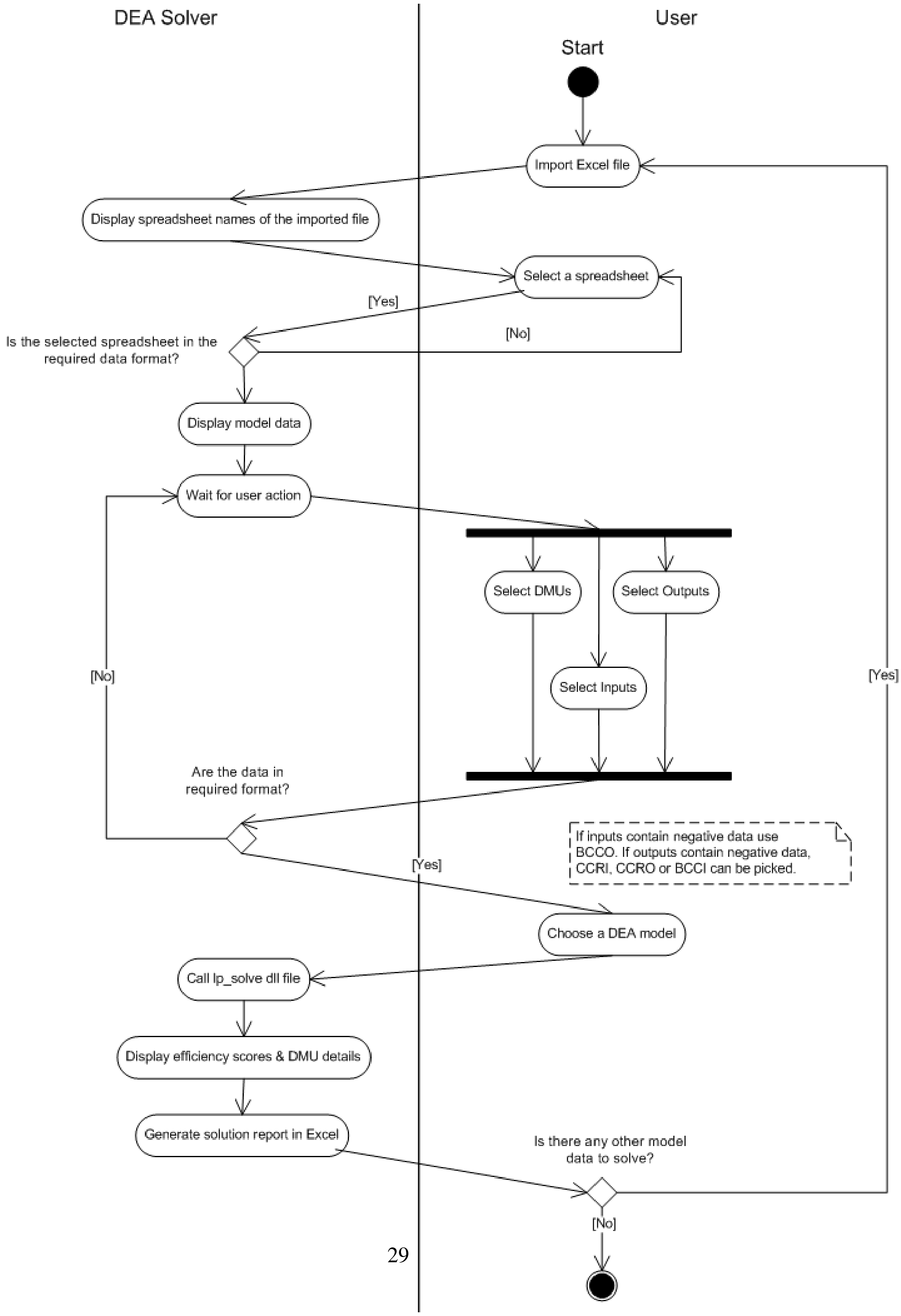
Table 7: Data structure of “Projections” spreadsheet

Field No	Field Description
1	DMU Names
2	Efficiency Scores
3	Variable Names
4	Input or Output
5	Original Value
7	Difference
8	Percentage Difference

back one step and do the current activities again with different settings. This capability saves time and brings convenience especially when users work with same model data but different input or output settings or different models. For instance, let a user work with the transportation activities data of a logistics company. But he or she is not sure about the correct selection of inputs and outputs or about the DEA model such as BCC-Input or CCR-Input. In such a case, the structure of the software enables the user carrying out different analysis without loading the model data again and again.

Figure 6 gives the Unified Modelling Language (UML) activity diagram of the SmartDEA DSS, documenting what happens once the software is executed. The UML activity diagram is extensively used in computer sciences for process modeling and representation, as well as workflow specification (Miles and Hamilton, 2006). UML activity diagram given in Figure 6 displays the process flow independent of the programming language and it helps other developers undestand the logic and the structure, leading to a continuous development with multiple developers.

Compared to other DEA solvers in the market, the advantages of SmartDEA is its effective results file table structure especially suitable for visualization tools such as Miner3D (Miner3D) and Visokio Omniscope (Omniscope). A further step in the development of the software can be the integration with any visualization and data mining tool as endorsed by the suggested software framework.



DEA Solver
User
Start
Import Excel file
Display spreadsheet names of the imported file
Select a spreadsheet
[Yes]
[No]
Is the selected spreadsheet in the required data format?
Display model data
Wait for user action
Select DMUs
Select Outputs
Select Inputs
[No]
[Yes]
Are the data in required format?
If inputs contain negative data use BCCO. If outputs contain negative data, CCRI, CCRO or BCCI can be picked.
[Yes]
Choose a DEA model
Call lp_solve dll file
Display efficiency scores & DMU details
Generate solution report in Excel
Is there any other model data to solve?
[No]

Figure 6: UML activity diagram of the SmartDEA solver

5. Test (Application) of the SmartDEA software: Benchmarking Vendors of a Turkish Automotive Company

The continuous performance evaluation is a critical management objective in today's highly competitive markets. In order to be succesful in the fierce market conditions, companies need to measure the performances of themselves and their business partners to make any improvement in the existing conditions. In this section, the test of the SmartDEA is made by using two different DEA models for benchmarking the vendors of a leading Turkish automotive company, ABC[3], with respect to after-sale services.

5.1. An overview of DEA methodology on Vendor and Supplier Evaluation

DEA has been widely used as a strong performance evaluation tool in the literature. Gattoufi et al. (2004a) present a classification scheme for the DEA literature, based on the following criteria: (1) Data source, (2) Type of the implemented envelopment, (3) Analysis, and (4) Nature of the paper. For the content analysis of DEA literature and its comparison with operations research and management science fields, the readers can refer to the paper by Gattoufi et al. (2004b). Considering the wide range of applications that are addressed by DEA, only some of the studies related our application area, namely vendor and supplier evaluation problems, are discussed here.

In recent years, DEA has drawn attention as a method for vendor and supplier evaluation and benchmarking processes. Exemplary research in this area includes the following. Weber (1996) shows how the application of DEA for vendor performance measurement can lead to significant financial benefits. In the study, a DEA model is formulated to measure vendor efficiencies and the implementation of the technique in a baby food manufacturer is provided as a case study. Weber et al. (2000) integrates multi-objective programming and DEA to evaluate the number of vendors to employ. The authors initially solve for the number of vendors and then evaluate their efficiencies. In the study, a case study is provided for a Fortune 500 company in a just-in-time (JIT) manufacturing environment. Ross et al. (2002) provide an integrated benchmarking approach to distribution center performances using DEA modeling. Approximately 100 distribution centers are benchmarked and their productivities are examined by using extensive tools

[3] In order to keep the name of the company confidential and for convenience, the ABC company will be used from now on to refer the automotive company.

of DEA such as facet analysis and window analysis. After the role model distribution centers are determined, strategic managerial insights are obtained. The authors also discuss the strengths of DEA in environments where explicit knowledge about the relationship between the inputs and outputs is obscure. Sun (2004) examines the opportunities in using the DEA as a fundamental tool for evaluation of the joint maintenance shops in the Taiwanese Army and their continuous improvement. The author observes DEA to be a valuable benchmarking tool for the managers of the maintenance shops. The use of DEA leads to more efficient use of scarce resources. Talluri et al. (2006) evaluates the performance of vendors using DEA but with a significant variation in the model. In order to get a precise assessment of vendor performances, the authors attempt to consider variability in vendor attributes by suggesting a chance-constrained data envelopment analysis approach. The paper discusses the approach in detail with a case study of a pharmaceutical company. The assessment of organizational units is another important area addressed by the DEA.

Liu et al. (2000) present an application of DEA to assess the overall performances of suppliers for a manufacturing company. Using a simplified model of DEA, the general performance of the suppliers are benchmarked with the strategic aim of reducing the number of suppliers effectively and providing the suppliers targets to improve their positions. Narasimhan et al. (2001) propose a supplier evaluation method using DEA combined with a weighted model to categorise suppliers into four performance clusters: HE (high performance and efficient), HI (high performance and inefficient), LE (low performance and efficient), and LI (low performance and inefficient). Easton et al. (2002) apply DEA in purchasing management. Their study implements DEA to help managers improve the efficiency of the purchasing operations. The comparison of purchasing efficiencies of different companies operating in the petroleum industry is presented using a DEA model. Celebi and Bayraktar (2008) explore an integrated neural networks (NNs) and DEA model for evaluation of suppliers under incomplete information of evaluation criteria. First step of the study is the identification of evaluation criteria that need to be considered in the problem. Then, by use of performance history data and opinions of the experts or decision makers, back propagated neural networks are trained to reduce the set of the attributes into predefined set of major performance measures. Final evaluation of suppliers is done by DEA which uses the results of NNs. Wu (2009) presents also a hybrid model using DEA, decision trees (DT) and NNs to assess supplier performance. The model consists of two modules: Module 1 applies DEA and classifies suppliers into efficient and inefficient clusters based on the resulting efficiency scores. Module 2 utilizes firm performance-related data to train DT, NNs model and apply the trained decision tree model to new suppliers. Azadeh and Alem (2010) present a flexible decision making scheme based on DEA methodology for choosing appropriate method for supplier selection under certainty, uncertainty and probabilistic conditions. They uses three DEA models. DEA is applied when the input data is crisp. It would also use Fuzzy DEA or Chance Constraint DEA when data is not crisp. This is an important issue in real situations, we may face with deterministic, probabilistic or fuzzy data and the flexible approach is capable of handling such situations for managers and decision makers. Recently, Zeydan et al. (2011) used DEA analysis to rank efficient and inefficient suppliers in an automotive factory of Turkey.

5.2. DEA Model

In the DEA models of this study the vendors are benchmarked firstly in terms of their overall operations and then in terms of their purchases from ABC, respectively. The relationships of the ABC's spare parts warehouse and its vendors are analyzed by considering three inputs and one output. For both of the models, total *spare parts area* of vendors, total vendor *expenses* and *spare parts employees* of vendors are taken as inputs. However, while *total revenue* of vendors is taken as output in the first model, the amount of vendors' *purchase from ABC* is considered as the output variable in the second. The first model reflects a world view where the vendors are considered as an integral part of ABC's ecosystem. The second model reflects an alternative world view, where vendors are considered external to ABC, as customers.

In this benchmarking study, BCC-Input is selected as the DEA model that will be applied in the problem. There are two primary reasons behind this choice. (1) BCC model imposes more flexibility in such a way that the solution is not restricted with constant returns to scale. As previously stated, BCC model implies variable returns to scale and it is preferable in our problem. (2) Input oriented models set targets for input variables, up to which value they can be reduced while yielding at least the same amount of outputs. We believe that building short term road maps to reduce resource consumption of DMUs is set as one of the objectives and input oriented approach enables us to get insights by looking at the projected input variables. The data used in the analysis meets the sign requirements imposed by the BCC-I model in Table 1. Due to confidentiality issues, in this paper we can not present the full analysis, results and

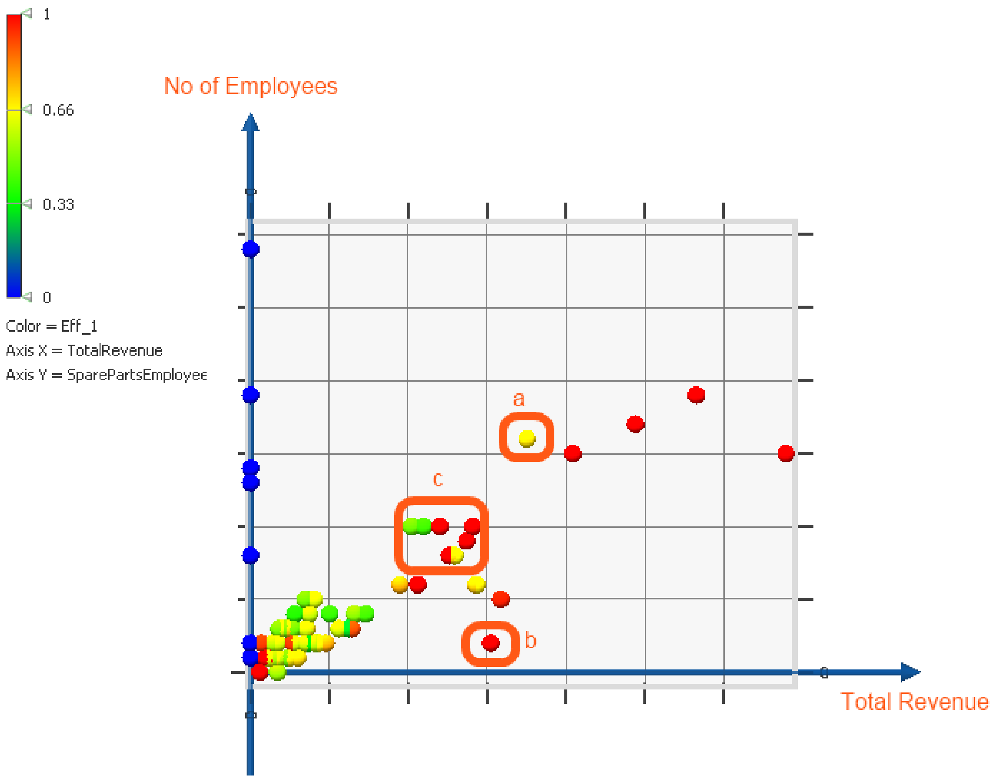


Figure 7: An example of starfield visualization

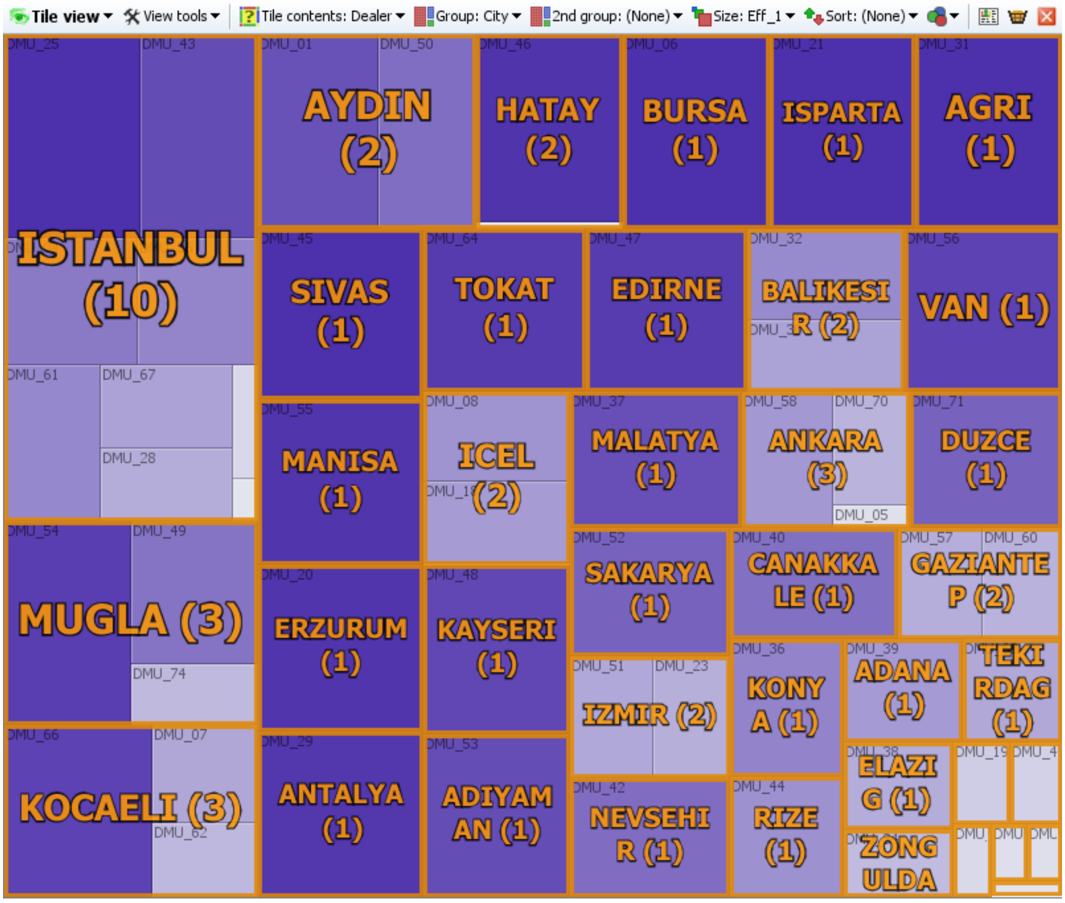
insights. Rather the developed DEA solver and its framework are focused as well as the data mining perspective and integration of DEA solutions into this framework.

5.3. Analysis Examples

In this section, applications of some of the possible data mining and information visualization techniques are exemplified through the DEA models built for ABC. The proposed software framework provides the structure of the data used in these techniques. In order to familiarize the reader with these techniques, sample applications are discussed below. The visualizations belong to the solution of the first DEA model that is used in the benchmarking of ABC's vendors.

Starfield visualization is illustrated in Figure 7 which is generated in the information visualization software, Miner3D (Miner3D). It is fundamentally a two-dimensional scatter plot with a capability of visualization high-dimensional data. In this visualization sample, x-axis and y-axis show the total revenue and the number of spare parts employees for the vendors. The corresponding x and y values are omitted from the graph due to confidentiality reasons. The color stands for the efficiency scores of the vendors. In addition to these variables, size of the DMUs can also be assigned any other data. Important patterns can be revealed by a clever choice of four visualization parameters: *x-axis*, *y-axis*, *color* and *size*. In Figure 7, three actionable insights can be observed immediately: (a) The vendor highlighted here has generated a great amount of revenue, but is not efficient. Since it also employs a large number of employees, the inefficiency can be explained through the high value for number of employees. Still, ABC managers should focus on this vendor, since it is one of the largest employers in ABC's after-sales vendor network. (b) The vendor highlighted here has generated a very large revenue with very few employees. (c) The cluster of vendors highlighted here operate with approximately the same number of employees and generate approximately the same amount of revenue. However, their efficiencies show high variance. It is worth investigating what causes such a variance in efficiencies, and identifying what each of these vendors can learn from each other.

Another effective visualizing scheme for DEA problems are tile graphs. Figure 8 gives an example of tile graph generated by Visokio Omniscope software (Omniscope). It allows analysts to cluster the data according to a criterion and effectively investigate the patterns hidden in the DEA results. Similarly, analysts need to choose variables among DEA model data, other data and DEA results to visualize on the graph. A systematic framework again makes it very convenient to derive insights and generate guidelines to be followed in the management of DMUs for better performance. In Figure 8, tile contents are the DMUs to be benchmarked. In this sample illustration, they are grouped according to their geographical region while size and color show the efficiency values that come up with the DEA results. While the vendor network (number and locations of the vendors) is reflected as is, due to confidentiality, randomly generated efficiency scores are used in the graph. This visualization allows the visual identification of efficient and inefficient vendors in each city, as well as the overall distribution of efficiency scores in a city in comparison to other cities. The analysts choose among the DEA model data, other data and the DEA results to be represented by the *tile content, first clustering criteria, second clustering criteria, size* and *color*.



Tile view ▾ View tools ▾ Tile contents: Dealer ▾ Group: City ▾ 2nd group: (None) ▾ Size: Eff_1 ▾ Sort: (None) ▾

Figure 8: An example of tile graph

As illustrated with the two examples, the analytic framework presented in our paper makes it possible to investigate these visulization options and discover significant actionable insights systematically.

6. Concluding Remarks

In this study, a software framework is proposed to represent the results of any DEA study in a formal manner, enabling the analysts to make analytical benchmarking between different DMUs. In accordance with this framework, an innovative DEA solver, SmartDEA, is developed and tested in a real world project for benchmarking the vendors of a Turkish automotive company. The framework allows analysts to identify hidden patterns and derive managerial insights by integrating the results of any DEA study with various types of information visualization, data mining and OLAP technologies in the implementations of DEA studies. To summarize, DEA results are examined in a computer science oriented perspective and data mining point of view.

This study provides a strong fundamental background to start any kind of analytical benchmarking in DEA by developing a formal way of representation for DEA results. One of the future work can be to devise and develop an analytical analysis framework, taking the proposed framework in this paper as a basis. Associated with this analysis framework, the ABC's DEA results can be analyzed by integrating them with data mining and information visualization techniques analytically. In addition, the software framework can be extended to allow many sub-categories of these techniques. As noticed, the proposed framework can be applied in any application area. Another future work can be focusing on any specific area and building domain specific frameworks to analyze DEA results. The developed software can be improved by adding help, information or tips for those not related with DEA concepts.

Acknowledgements

The authors express their sincere gratitude towards to the case company's managers and assisting staff for their unlimited support in the validation of the proposed framework. The authors also thank Sabancı University alumni Tuna Çaprak for his help in the case study, in association with his Leaders for Industry (LFI) project at the company.

References

Avkıran, N. (1999). An Application Reference for Data Envelopment Analysis in Branch Banking: Helping the Novice Researcher. International Journal of Bank Marketing. 17(5), 206-220.

Azadeh, A., Alem, S.M. (2010). A flexible deterministic, stochastic and fuzzy Data Envelopment Analysis approach for supply chain risk and vendor selection problem: Simulation analysis. Expert Systems with Applications, 37(12), 7438-7448.

Banker, R.D., Charnes, A., and Cooper, W.W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science, 30(9), 1078-1092.

Barr, R.S. (2004). DEA software tools and technology: A state-of-the-art survey, Handbook on Data Envelopment Analysis. Hingham, MA, USA: Kluwer Academic Publishers, 539.

Boussofiane, A., Dyson, R.G. and Thanassoulis, E. (1991). Applied data envelopment analysis. European Journal of Operational Research, 52, 1-15.

Bowlin, W.F. (1998). Measuring performance: an introduction to data envelopment analysis (DEA). Journal of Cost Analysis, 3(1), 328.

Charnes, A., Cooper, W.W., Rhodes, E. (1978). Measuring the efficiency of decision making units. European Journal of Operational Research, 2, 429-444. (correction by the same authors. Short Communication: 1979. Measuring the Efficiency of Decision Making Units. European Journal of Operational Research, 3.

Celebi, D., Bayraktar, D. (2008). An integrated neural network and data envelopment analysis for supplier evaluation under incomplete information. Expert Systems with Applications, 35(4), 1698-1710.

Chen, C. (2004). Information visualization: Beyond the horizon. Berlin: Springer.

Cooper, W.W., Seiford, L.M., Tone, K. (2006). Introduction to Data Envelopment Analysis and Its Uses: With DEA Solver Software and References. Springer, New York.

Dasu, T. and Johnson, T. (2003). Exploratory Data Mining and Data Cleaning. Wiley-IEEE.

Easton L., Murphy D.J., Pearson J.N. (2002). Purchasing Performance Evaluation: with Data Envelopment Analysis. European Journal of Purchasing & Supply Management, 8 , 123-134.

El-Mahgary, S., Lahdelma, R. (1995). Data envelopment analysis: Visualizing the results. European Journal of Operational Research, 83(3), 700-710.

Emrouznejad, A., De Witte, K. (2010). COOPER-framework: A unified process for non-parametric projects. European Journal of Operational Research, 207(3), 1573-1586.

Emrouznejad, A., Parker, B., Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. Journal of Socio-Economic Planning Sciences, 42(3), 151-157.

Emrouznejad, A., Thanassoulis, E. (2010). Performance Improvement Management Software (PIMsoft): A user guide, www.DEAsoftware.co.uk.

Ertek, G., Can, M.A., and Ulus, F. (2007). Benchmarking the Turkish apparel retail industry through data envelopment analysis (DEA) and data visualization. EUROMA 2007, Ankara, Turkey.

Fayyad, U., Grinstein, G.G. (2002). Introduction. In U.M. Fayyad, G.G. Grinstein, & A. Wierse (Eds.), Information visualization in data mining and knowledge discovery, (pp. 47-82).

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery: an overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, M. Park (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 1-36). AAAI Press, California, CA.

Gattoufi, S., Oral, M., Kumar, A., Reisman, A. (2004a). Content analysis of data envelopment analysis literature and its comparison with that of other OR/MS fields. Journal of the Operational Research Society, 55, 911935.

Gattoufi, S., Oral, M., Reisman, A. (2004b). A taxonomy for data envelopment analysis. Socio Economic Planning Sciences, 34(23), 141158.

Hoffman, P.E., Grinstein, G. G. (2002). A survey of visualizations for high-dimensional data mining. Chapter 2. In U.M. Fayyad, G.G. Grinstein, & A. Wierse (Eds.), Information visualization in data mining and knowledge discovery, (pp. 47-82).

Hollingsworth, B. (1997). Review: A Review of Data Envelopment Analysis Software. The Economic Journal, 107(443), 1268-1270.

Keim, D.A. (2002). Information visualization and data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1), pp. 1-8.

Keim, D.A., Sips, M., Ankerst, M. (2005). Visual Data-Mining Techniques. Visualization Handbook, 2005, 831-843.

Kim, Y.G., Suh, J.H., Park, S.C. (2008). Visualization of patent analysis for emerging technology. Expert Systems with Applications, 34(3), 1804-1812.

Lin, C., Lin, C-M., Li, S-T., Kuo, S-C. (2008). Intelligent physician segmentation and management based on KDD approach. Expert Systems with Applications, 34(3), 1963-1973.

Liu, J., Ding, F.Y., Lall, V. (2000). Using data envelopment analysis to compare suppliers for supplier selection and performance improvement. Supply Chain Management: An International Journal, 5(3), 143150.

Liu, Y., Salvendy, G. (2007). Design and evaluation of visualization support to facilitate decision trees classification. International Journal of Human-Computer Studies, 65, 95-110.

lp_solve:
http://lpsolve.sourceforge.net/5.5/

Miles, R., Hamilton, K. (2006). Learning UML 2.0. O'Reilly Media.

Miner3D:
http://www.miner3d.com

Narasimhan, R., Talluri, S. and Mendez, D. (2001). Supplier evaluation and rationalization via data envelopment analysis: an empirical examination. Journal of Supply Chain Management, 37(3), 28-37.

Omniscope:
http://www.visokio.com/omniscope

Ross, A., Droge, C. (2002). An integrated benchmarking approach to distribution center performance using DEA modeling. Journal of Operations Management, 20(1), 19-44.

Russell, S., Gangopadhyay, A., Yoon, V. (2008). Assisting decision making in the event-driven enterprise using wavelets. Decision Support Systems, 46(1), 14-28.

Seol, H., Lee, S., Kim, C. (2011). Identifying new business areas using patent information: A DEA and text mining approach. Expert Systems with Applications, 38(4), 2933-2941.

Smith, P.C., Goddard M. (2002). Performance Management and Operational Research: A Marriage Made in Heaven?. Journal of the Operational Research Society, 53(3), 247-255.

Spence, R. (2001). Information Visualization. ACM Press, Essex, England.

Soukup, T., Davidson, I. (2002). Visual Data Mining. New York, NY: John Wiley & Sons.

Sun, S. (2004). Assessing Joint Maintenance Shops in The Taiwanese Army Using Data Envelopment Analysis. Journal of Operations Management, 22, 233-245.

Talluri, S., Narasimhan, R. and Nair, A. (2006). Vendor performance with supply risk: a chance-constrained DEA approach. International Journal of Production Economics, 100(2), 212-222.

Tegarden, D.P. (1999). Business information visualization. Communications of the AIS Archive 1 (1).

Ulus, F., Kose, O., Ertek, G., Sen, S. (2006). Financial benchmarking of transportation companies in the New York Stock Exchange (NYSE) through data envelopment analysis (DEA) and visualization. 4th International Logistics and Supply Chain Congress, İzmir, Turkey.

Weber, C.A. (1996). A data envelopment analysis approach to measuring vendor performance. Supply Chain Management, 1(1), 2830.

Weber, C.A., Current, J.R., Desai, A. (2000). An optimization approach to determining the number of vendors to employ. Supply Chain Management: An International Journal, 2(5), 90-98.

Woo, J.Y., Bae, S.M., Park, S.C. (2005). Visualization method for customer targeting using customer map. Expert Systems with Applications, 28(4), 763-772.

Wu, D. (2009). Supplier selection: A hybrid model using DEA, decision tree and neural network. Expert Systems with Applications, 36(5), 9105-9112.

Yadav, V.K., Padhy, N.P., Gupta, H.O. (2010). A micro level study of an Indian electric utility for efficiency enhancement. Energy, 35, 4053-4063.

Zeydan, M., Colpan, C., Cobanoglu, C. (2011). A combined methodology for supplier selection and performance evaluation. Expert Systems with Applications, 38(3), 2741-2751.