Universitat Autònoma de Barcelona

Four Essays on Data Visualization and Anomaly Detection of Data Envelopment Analysis Problems

Shahin Ashkiani

Thesis for the degree of Doctor of Philosophy

Four Essays on Data Visualization and Anomaly Detection of Data Envelopment Analysis Problems

Shahin Ashkiani

Supervisor:

Víctor M. Giménez García

Department of Business
Universitat Autònoma de Barcelona
Barcelona, Spain 2019

Four Essays on Data Visualization and Anomaly Detection of Data Envelopment Analysis Problems
Shahin Ashkiani

Department of Business
Autonomous University of Barcelona
Campus Bellaterra, 08193 Barcelona, Spain
Telephone +34 93 581 12 09

Typeset by the author using LaTeX.

to all who contribute to human knowledge,

and believe that

free access to knowledge is a basic human right

Abstract

Data visualization is a relatively neglected topic in the field of data envelopment analysis (DEA). In the comprehensive handbooks of DEA, there is hardly any chapter or section dedicated to data visualization methods, and in the applications of DEA, a very limited and peripheral role is usually assigned to data visualization. However, graphical representation of data can have definite benefits for the practitioners and researchers of the field, to such extent that the resulted insight to the DEA problems through visualization may not be gained using analytical methods. Data visualization, when it is applied correctly, it is able to reveal regularities and irregularities in the data. Regularities can be trends, or clusters, and irregularities are anything discordant, such as outliers. In some cases, data visualization helps to grasp the data much more quickly, as human brain is wired to absorb visual data more efficiently than grasping digits, and data visualization can summarize loads of digits into one chart. On the other hand, some patterns are become visible when all the variables and their relations are retained by the investigation method. High-dimensional data visualization is composed of such methods, and it is in the center of this thesis, to find regularities and irregularities in the various DEA datasets.

Despite the relative neglect, DEA data visualization toolbox is not empty, and in fact it has several useful tools. The first essay of this thesis is a visual survey of these available tools. Since there is no such survey in DEA literature, it is important to gather all the visualization tools in a toolbox, and identify and illustrate the important ones in order to help practitioners to pick the proper tools, and to help researchers to craft novel tools.

The second essay of this thesis suggests a new tool for this toolbox. This new tool is a visualization method for DEA cross-evaluation methodology, and can be used for various purposes including detection of outliers or uncommon decision making units (DMU). One type of these uncommon DMUs is called “maverick units”, and the third essay of this thesis is focused on this sort of DMUs. Maverick units are the subject of the second essay, and a new visual method, based on the preceding essay, is suggested to detect such DMUs, and a new index is devised to numerically identify them. It is shown that the new maverick index is theoretically and practically more justified and robust than the well-known maverick indexes of DEA literature.

The forth and last essay is an introduction to DEA-Viz, a new visualization software developed by the author of this thesis. DEA-Viz includes the implementation of the suggested cross-evaluation visualization method of essay one, as well as a selection of previously suggested visualization methods. Moreover, the DEA-Viz has novel visualization features in order to investigate maverick units in further details,

following the second essay. The importance of DEA-Viz lies in the facts that there is not any DEA software with the same functionality as DEA-Viz, or any DEA software with similar features of DEA-Viz. Thus, DEA-Viz can have an unparalleled role in analysis of DEA problems, and promotion DEA visualization.

Following the enhancement of this thesis, an R package including all the DEA-Viz tools, as well as some new methods is developed by the author. The package, can be found in author's online code repository, makes the code available to every interested user, and expands the current DEA visualization tools from static data, to panel data.

Acknowledgments

This thesis is the result of an almost four years of effort. Beside my family and friends, many people have been directly or indirectly affected on my work, and it is impossible to name them one by one. Nevertheless, there are some names that should not be overlooked.

I would like to thank Cecilio Mar Molinero and Víctor M. Giménez García for their comments on the project, and their logistical help regarding the project and my life in this period in general. Unfortunately, due to bureaucratic reasons, I was not allowed to mention Cecilio's name as one of my supervisors, however he has kindly supervised this project, specially at its early stage.

There is one scholar who encouraged me by his comments, and by showing genuine attention and care about my work and my ideas, even though he was not directly linked to this doctorate: many thanks Konstantinos Triantis. Besides, Peter Bogetoft helped me in the early stages of this path with his incisive critical comments, and Jan de Leeuw kindly replied to my questions regarding some technical methods used in the thesis.

A doctorate project is not only about scholars. Bearing these years was not possible without the presence of my dear friends: Friedrich Nietzsche, Jane Goodall, Thomas Mann, Leo Tolstoy, Haruki Murakami, Wayne Dyer, Emile Zola, and Romain Rolland, among others. Thank you for all you shared with me.

Shahin Ashkiani
Barcelona, June 2019

List of Publications

This thesis is based on the following appended articles:

Article 1. Shahin Ashkiani. *Visualization of Data Envelopment Analysis Problems: A Visual Survey* .

Article 2. Shahin Ashkiani and Cecilio Mar Molinero. *Visualization of Cross-Efficiency Matrix Using Multidimensional Unfolding.* Proceedings of the 15th International Conference of DEA, June 2017, University of Economics, Prague, Czech Republic, ISBN: 978 1 85449 433 7.

Article 3. Shahin Ashkiani,. *Mavericks Revisited: A New Index to Identify Maverick Units in Data Envelopment Analysis* .

Article 4. Shahin Ashkiani. *DEA-Viz: A Software for Visualization of Data Envelopment Analysis Problems* .

List of Acronyms

BCC	–	Banker, Charnes, and Cooper DEA Model
CCR	–	Charnes, Cooper and Rhodes DEA Model
CEM	–	Cross-Efficiency Matrix
CII	–	Column Isolation Index
CRS	–	Constant Return to Scale
D&G	–	Doyle and Green [Maverick Index]
DEA	–	Data Envelopment Analysis
DMU	–	Decision Making Unit
EDI	–	Efficiency Disparity Index
FDH	–	Free Disposal Hull
FPI	–	False Positive Index
MDS	–	Multidimensional Scaling
MDU	–	Multidimensional Unfolding
MI	–	Maverick Index
NLM	–	Non-Linear Mapping
PCA	–	Principal Component Analysis
RII	–	Row Isolation Index
SOM	–	Self-Organizing Maps
SSA	–	Smallest Space Analysis
VRS	–	Variable Return to Scale

Contents

Part I

Introduction

Introduction

1 Importance of Data Visualization

This doctorate thesis orients around data visualization in the data envelopment analysis (DEA) domain. Data visualization has various definitions, but I prefer to use the definition of Ward et al. (2010): "[Visualization] is communication of information using graphical representations". In this definition "information" is a very broad term, but in this thesis, information is confined to quantitative information, since DEA data is quantitative.

The ultimate goal of data visualization is efficiently gaining insight into the data, and thus the problem. The "efficient" term has a key role here, since it implies that detection and identification of some features of the data through visualization is much easier than through alternative approaches, i.e. analytical methods. However, the benefits of visualization are not limited to this "efficiency". Through proper visualization, some features of the data may be revealed that otherwise would remain concealed using analytical approaches. The reason of such power lies in the fact that the "proper visualization method" has a "holistic approach" to the data. Using a holistic approach, the information about components of a system and relations among these components are retained, and therefore "emergent properties" may be exposed.

In order to illustrate the above points, the efficiency of visualization and the emergent properties, and underline the importance of visualization, let's have a look at the following simple yet powerful numerical examples.

The first example is the famous Anscombe's quartet (Anscombe 1973). Anscombe (1973) presents four fabricated datasets, shown in the below Table 1, whose statistical characteristics, listed below, are identical up to two decimals:

Number of observations = 11

Number of variables = 2

Mean of variable 1 = 9.0

Mean of variable 2 = 7.5

Variance of variable 1 = 11

Variance of variable 2 = 4.12

Multiple $R^2 = 0.667$

Equation of the regression line $= y = 3 + 0.5x$

Sum of squares of $x - \bar{x} = 110.0$

Regression sum of squares = 27.50 (1 degree of freedom)

Residual sum of squares of y = 13.75 (9 degrees of freedom)

Estimated standard error of intercept coefficient$(b_1) = 0.118$

According to the identical, up to two decimal places, statistic measures of these datasets, we may assume that the datasets are identical or at least very similar. Nevertheless, the numeric measures may not tell the whole story and over-trusting them, as unquestionable representatives of the data, may be misleading. Such assumption and over-trust in numerical tools, may mar our understanding and analysis, as these datasets are drastically different from each other. Figure 1 shows these four datasets, followed by the datasets in Table 1.

Table 1: Datasets of Anscombe's quartet

	x1	**y1**	**x2**	**y2**	**x3**	**y3**	**x4**	**y4**
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Variance	11	4.12	11	4.12	11	4.12	11	4.12

Anscombe's quartet clearly shows the importance of using visualization as a supplementary tool to analytical methods. Through visualization of the Anscombe's datasets, the structure of the data can be swiftly grasped, since human brain is evolved to be a very powerful pattern-recognition machine, and "we acquire more information through vision than through all of the other senses combined." (Ware 2012)

Beside the emphasize on the importance of visualization, the importance of "proper visualization" must be brought into attention. This thesis benefits from high-dimensional data visualization, since a holistic view to the data may be achieved through high-dimensional data visualization. High-dimensional data visualization

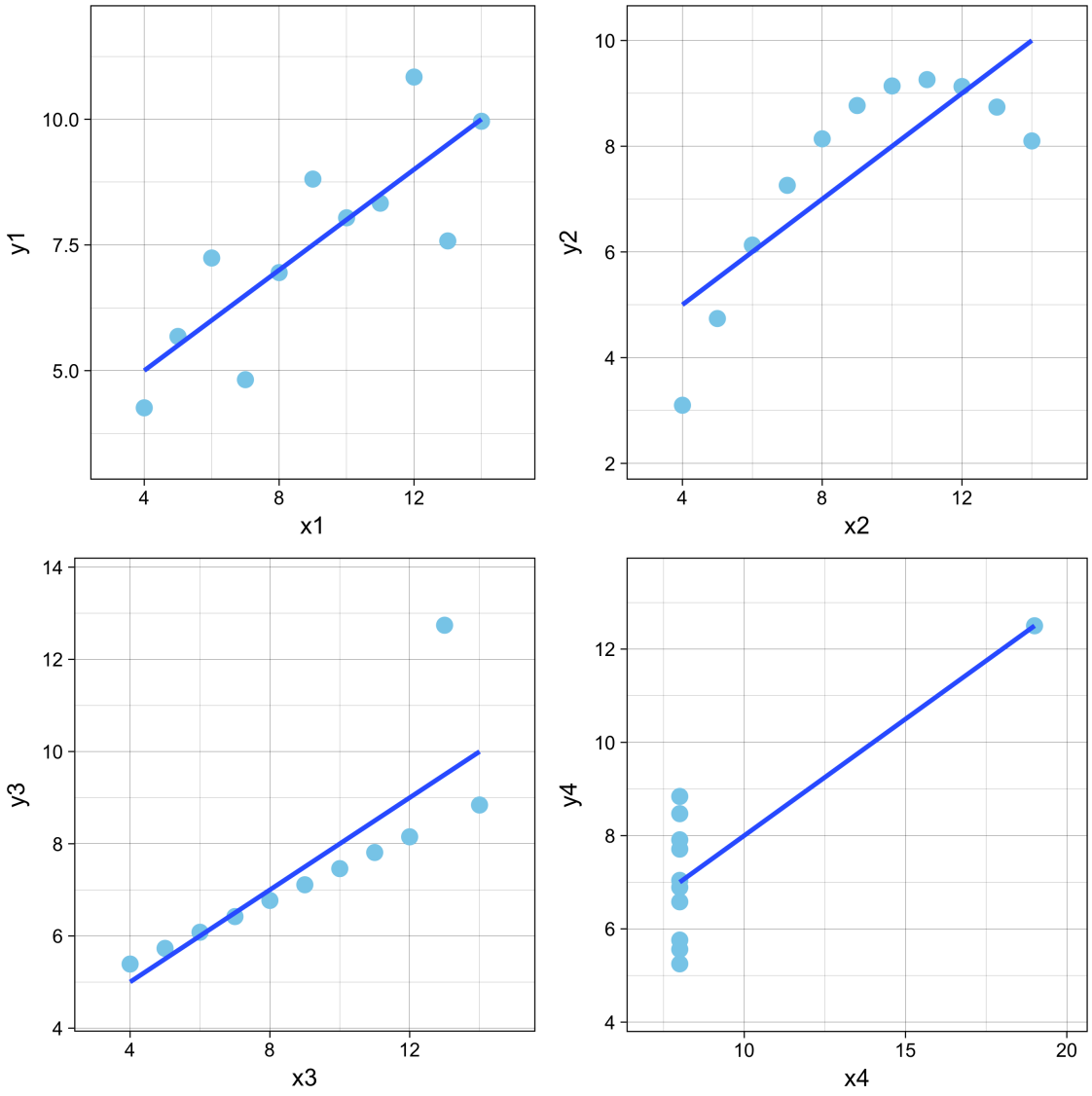

Figure 1: Scatterplots of the four fictitious datasets of Table 1

methods try to consider all the data, i.e. components and their relations, simultaneously, as much as possible.

If a dataset is considered a complex system, then trying to understand this system through understanding each component of it, separated from its relations and out of its parent system, may yield an incomplete and even misleading understanding, since the emergent properties would not be revealed. These properties emerge in the presence of all components and their relations, thus any approach that does not tend to study all the components and relations simultaneously, would fail to grasp such properties.

This point can be illustrated using a very simple dataset of two variables, x and y. Figure 2.10 shows the data points on a scatter-plot, whose coordinates are variables x and y on the top section. Since the dataset has only two variables, it can be visualized on a bi-dimensional space without loss of information. The structure of the data, formed a triangle, is clearly seen in this plot as well as a seemingly outlier, located in the center of the triangle.

Figure 2.10 has also two other plots in the bottom row, and they are the distributions of variable x and variable y. On each distribution, the corresponding positions of the outlier point are determined by the vertical dashed lines.

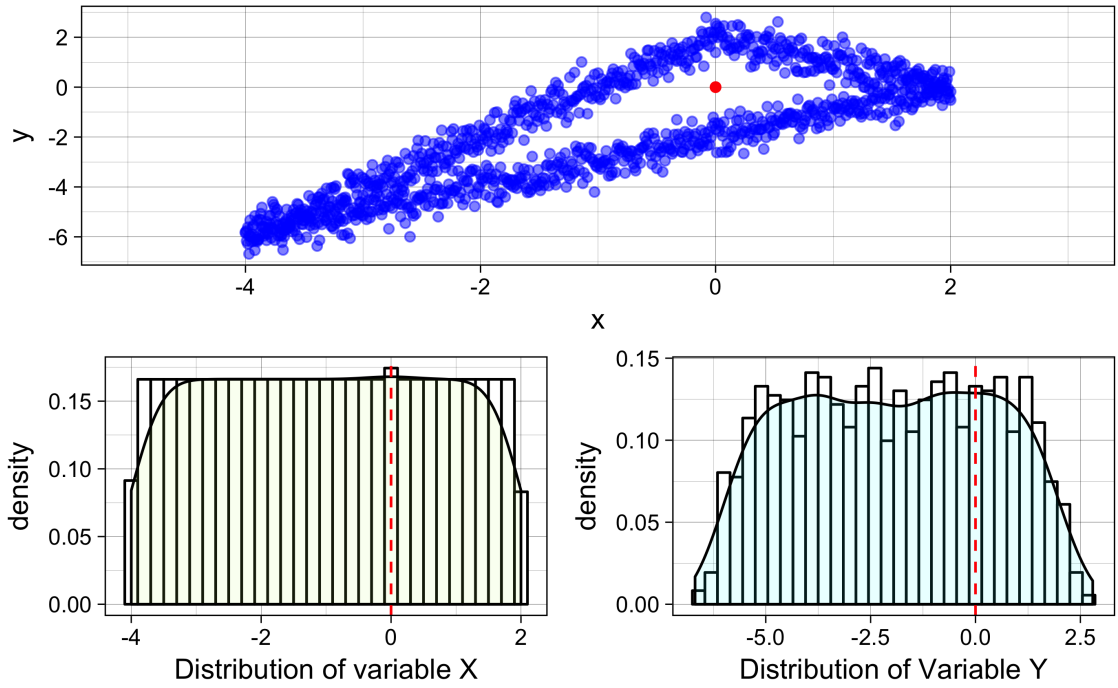

Figure 2: Importance of having a holistic view to the data

The top plot of Figure 2.10 is the holistic approach to the dataset, since it shows all variables and their relationships simultaneously. In contrast, the bottom plots are depiction of each variable in isolation, regardless of the other variable and thus regardless of their relationships.[1] In other words, the former is a holistic approach

[1]It is also possible to consider each point as a component of the system, and thus the relations would be between the coordinates of each point.

to the data, and the latter is a reductionist approach to the data. It is conspicuous that the reductionist approach fails to yield any comprehensive insight into the data, neither about the general structure of the data, nor about the irregularities of it. While both of these approaches are "data visualization", the proper way of visualization of a "dataset" is the holistic approach. Due to the explained reasons, this thesis is focused on high-dimensional data visualization of DEA problems. High-dimensional data visualization methods try to retain the information in the datasets, while reductionist approaches do not even try to do so.

Overall,Ware (2012) summarizes the advantages of data visualization as follows:

1. Representation of large amount of data: A plot can represent large number of quantitative data, which otherwise is difficult to comprehend using analytical methods, without loss of details. "A picture is worth a thousand words". This adage should change into "A plot is worth of a thousand digits" in our domain.

2. Emergent properties: Emergent property is a characteristic of the whole, i.e. components and their relations. Earlier, it was explained and illustrated that proper visualization, i.e. with holistic approach, can expose emergent properties of data.

3. Visibility of the data peculiarities: One of the most common goals of visualization is finding outliers and uncommon units, or any problem with the data.

4. Visibility of the data features: Characteristics of the data such as clusters of units become visible through visualization.

5. Raising new questions: Visualization is thought-provoking. New questions regarding the data, the methods, and even fundamental assumptions of analysis raise through scrutinizing the plots.

2 Data Visualization in DEA problems

Having all been said, relatively little attention has been paid to the visualization in DEA field. For instance, there is not even a single section about DEA visualization in the the prominent books such as Ramanathan (2003) Ray (2004) Thanassoulis (2001) Coelli et al. (2005) Cooper, Seiford, and Tone (2006). Even the handbook of data envelopment analysis Cooper, Seiford, and Zhu (2011) has no trace of DEA visualization. In these references, the visualization is restricted to scatter-plots of two variables, or two ratios, mainly for depiction of efficient frontier. Zhu and Cook (2007) is the only book with a chapter about DEA visulization, however the chapter is focused on promoting a specific method, and it is not about DEA visualization in general.

From the theoretical aspect, among several thousands DEA studies, there are fewer than twenty articles related to DEA visualization. This means that the topic is not totally dismissed from the theoretical aspect, however from the practical

aspect both practitioners and DEA software packages have vastly neglected the role of visualizations. Table 3 shows nine publicly available DEA softwares, and their visualization features. It can be seen that while almost half of these packages totally have overlooked the visualization features, the rest are focused on uni-variate or bi-variate visualizations, in the absence of high-dimensional data visualization. This lacking is more crucial when one considers the significance of high-dimensional visualization, and high-dimensionality of DEA problems.

The reasons behind such neglect may be one or some of the followings:

1. Multi-dimensionality of DEA problems: Adler and Raveh (2008) state that the difficulty of visualization of high-dimensional data is the reason of scarcity of graphical representation of DEA problems. Being so perhaps justifies the prevalent bi-variate plots in DEA studies, and rare multi-variate visualizations.

2. Doubt about the benefits of visualization: DEA visualization is ignored maybe because most of the DEA researchers and scholars are doubtful about its real benefits, the benefits that cannot be gained using analytical methods. When the analytical methods are considered exact and sophisticated, perhaps it is difficult to take visualization seriously.

3. Lack of high-dimensional visualization features in DEA software packages: A DEA practitioner, who believes in the power of visualization, would hardly find any high-dimensional data visualization in DEA software packages. So how can a practitioner visualize a DEA problem without programming the visualization methods from scratch?

All these possible reasons behind the relative dismiss of DEA visualization are addressed in this thesis. This thesis is composed of four articles, and while the articles are independent, they are thematically related and they emerge a cohesive whole around the broad topic of DEA visualization.

Having found data visualization an important topic in DEA research and practice, and diagnosed the possible reasons of visualization shortage, I wrote the current set of articles in order to improve the literature, and respond to shortages. The lack of multidimensional visualization tool is addressed in the first article by suggestion of a new tool, the third article is a survey in order to represent and evaluate the available DEA visualization tools altogether, and the forth article addresses the lack of a DEA visualization software package focused on high-dimensional data visualization. The second article is a new technique to detect maverick DMUs, a technique based on DEA visualization method of the first article. The importance of visualization is theoretically and practically highlight throughout the articles, specifically in the second article where visualization sheds light on the weakness of the current maverick detection methods, and helps to devise a new identification index for maverick units.

In the rest of this introduction, the articles are introduced briefly, and their contributions to DEA literature and community are highlighted.

No.	**Name**	**Web Address**	**Platform**	**Visualization Features**
1	Frontier Analyst	http://banxia.com/frontier/	Stand-alone software for MS Windows	Univariate plots (e.g. reference-set frequency, efficiency scores barchart, slack variables pie chart) Bi-variate plots (e.g. efficient frontier on scatterplot, scatterplot of variable correlation, multiple bar-chart of variable comparison between DMU pairs) Multi-variate (Radar-chart of unit variables)
2	PIM-DEAsoft	http://www.deasoftware.co.uk/	Stand-alone software for MS Windows	Uni-variate plots (e.g. efficiency scores bar-chart) Bi-variate plot (e.g. efficient frontier scatterplot efficiency scores time-series line-chart, Malmquist index time-serie)
3	DEAP	http://www.uq.edu.au/economics/cepa/deap.php	Stand-alone software for MS Windows	None
4	DEAFrontier	http://www.deafrontier.net/deasoftware.html	Ms Excel add-on for MS Windows	None.
5	Open Source DEA	http://opensourcedea.org/	Stand-alone open source software working on Ms Windows, OSx, and Linux	None
6	Benchmarking	https://cran.r-project.org/web/packages/Benchmarking/index.html	Package for R	Uni-variate plot (e.g. density plot of efficiency) Bi-variate plot (e.g. scatterplot of transformation curve, isoquant or production function)
7	DEA-Solver Pro	http://www.saitech-inc.com/products/prod-dsp.asp	Ms Excel add-on for Ms Windows	Uni-variate plot (efficiency scores bar-chart)
8	MaxDEA	http://www.maxdea.cn	Ms Windows and Ms Access	Bi-variate plot (e.g. scatterplot of efficiency frontier)
9	DEA Toolbox	http://www.deatoolbox.com/	MATLAB library	None.

Table 2: Visual Outputs of some prominent DEA softwares

3 A Brief Review of Thesis' Articles, and Their Relations

The four articles of this thesis are stand-alone papers, however they are more or less related to each other, and closely related to the broad DEA visualization topic. Following is a concise introduction to each article, and Figure 3 depicts the relations among these articles.

Article1, with the title of **Visualization of Data Envelopment Analysis Problems: A Visual Survey**, is a survey paper covering all DEA visualization efforts, including the first paper of this thesis. Up to the time of submission of this thesis, December 2017, there is no survey on this subject in DEA literature, and this paper would be the first one.

Article2, with the title of **Maverick Units Revisited: A New Index to Identify Maverick Units**, first critically evaluates the DEA literature regarding maverick units, then proposes a new method to find such units based on the visualization method suggested in paper1. This paper even goes further and suggests a new quantitative index to detect such units.

Article3,with the title of **Visualization of cross-efficiency matrices using multidimensional unfolding**, suggests a novel visualization method for graphically representation of cross-evaluation outcome. Through such visualization, the outliers and uncommon units can be identified.

Article4, with the title of **DEA-Viz: A Software for Visualization of Data Envelopment Analysis Problems**, is mainly about DEA-Viz, a software that has been developed by the author of this thesis. DEA-Viz includes most of high-dimensional DEA visualization methods, and is developed to facilitate and promote DEA visualization. A visualization case study is included in this paper. This software is unparalleled in DEA, and with further improvements can be a reference for DEA community.

Figure 3 depicts the structure of this thesis.

This thesis is framed in three parts. The current introduction is the first part, and part 2 is composed of the four essays, i.e. the main body of the thesis. The third part is an extensive conclusion, as a summary of the thesis in addition to possible improvements to the current work, and possible next steps in the DEA visualization fields.

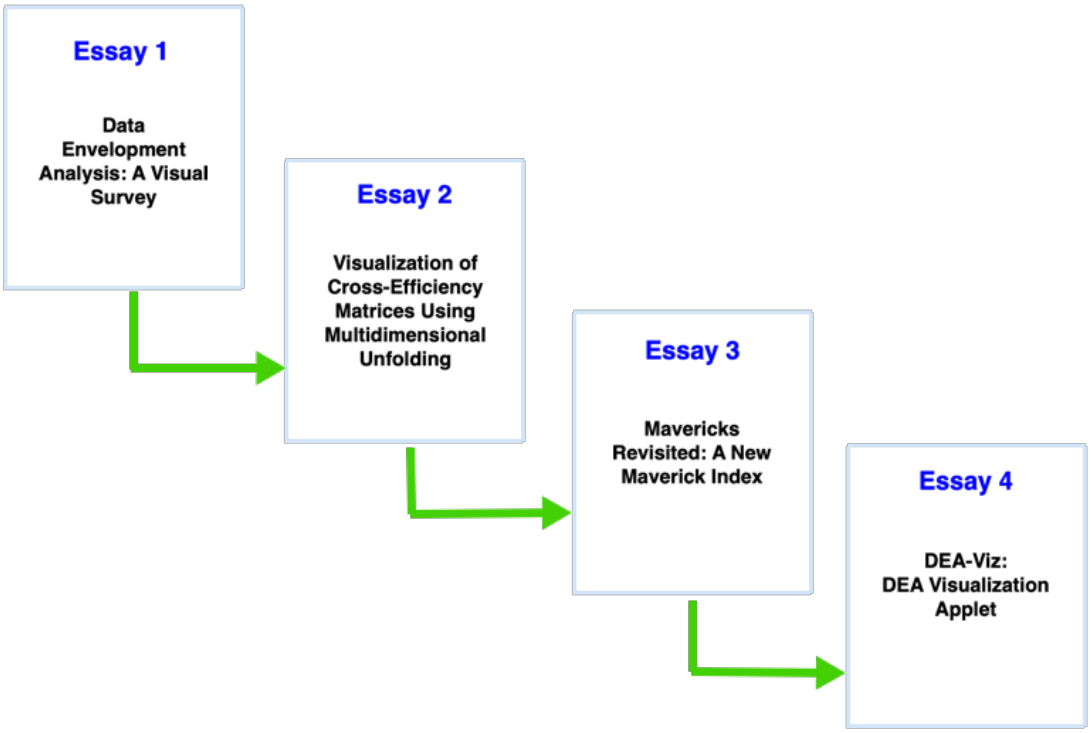

Figure 3: The articles and their relations, i.e. thesis structure

4 Contributions of the Essays

Concisely, the first paper adds a new tool to the DEA visualization toolbox, the second improves the DEA literature regarding the maverick units, based on critical appraisal of the literature and the visualization method of paper 1. Moreover, the second paper suggests a new index to detect maverick units, sort of outliers in DEA problems. It is illustrated that this new maverick detection method performs better than the available one in the literature, and the new method has more defend-able theoretical foundation than it.

The third paper is a survey, a comprehensive review of the available tools in the DEA visualization toolbox including the suggested tool of essay 1. This essay would be the first DEA visualization survey in the literature. The forth article is an introduction to a unique DEA software, which includes implementation of the visualization method of paper1 beside other main DEA visualization methods. This applet includes the visualization methods that are absent in other DEA software packages. The features of this applet is illustrated through a brief case-study in the forth article. This software hopefully promotes using of DEA visualization.

5 Introduction to the Thesis from Another Perspective

While this introduction could be finished at the end of the previous subsection, I decided to be more creative and invite the probable readers to have a look at this thesis from another perspective. To do so, I have done some text and citation analysis on the essays. All the computation is done by me in R Team (2016), using packages such as tidytext Silge and Robinson (2016), ggraph Pedersen (2017), and wordcloud Fellows (2014) among others. Hence, this section is an unconventional introduction to the thesis, and its essays. Nevertheless, being uncommon does not mean redundant or dull. On contrary, I found the findings of this text analysis very exciting, and due to its interesting results, I decided to add it to this introduction. This subsection was originally a personal project, but turned out to be informative enough to be part of the thesis.

5.1 Text Analysis Introduction to Essays

In this subsection, I would like to present the thesis from a new perspective, which is not conventional in similar theses. Using text analysis, a big-picture of the four essays, and the relations between the are presented. In this section, essays are distilled to the most frequent terms which capture the essence of each essay, then references of each essay are analyzed, and at the end, the relation between essays based on the main concepts of their references are depicted. The general goal is giving an overall understanding of the thesis, such as the main concepts and relations between papers, without going into the details.

Text analysis and text mining are very broad terms. They can be considered as part of information retrieval methods. The main idea of the text analysis of this thesis is finding the core concepts of each article, without reading them thoroughly. In the previous subsections, each article was introduced using a short description. However, it is possible to approach a text from analytical perspective, rather than subjective description of the author of the text.

In order to find the possible main concepts of each essay, I have used a simple method called bi-gram. The method is based on this assumption that frequently repeated words in a text can reveal the main topics of that text. In bi-gram, each pair of consecutive words is considered as a whole component, and such components with the highest frequency are found in a text.

The figure 4 is the sorted table of the most frequent pair of consecutive words in the article1.

As it is seen, row object, cross-efficiency, column object, row profile, and efficiency score are the top five frequent pair consecutive words, bi-grams, in this essay. It can be guessed that this essay is about cross-efficiency, and possibly the rows and columns of cross-efficiency matrix. Reviewing the rest of pairs gives this hint that "visualization" is an important concept here as well as anomaly detection.

Following are bi-gram bar-plots of articles 2, 3 and 4.

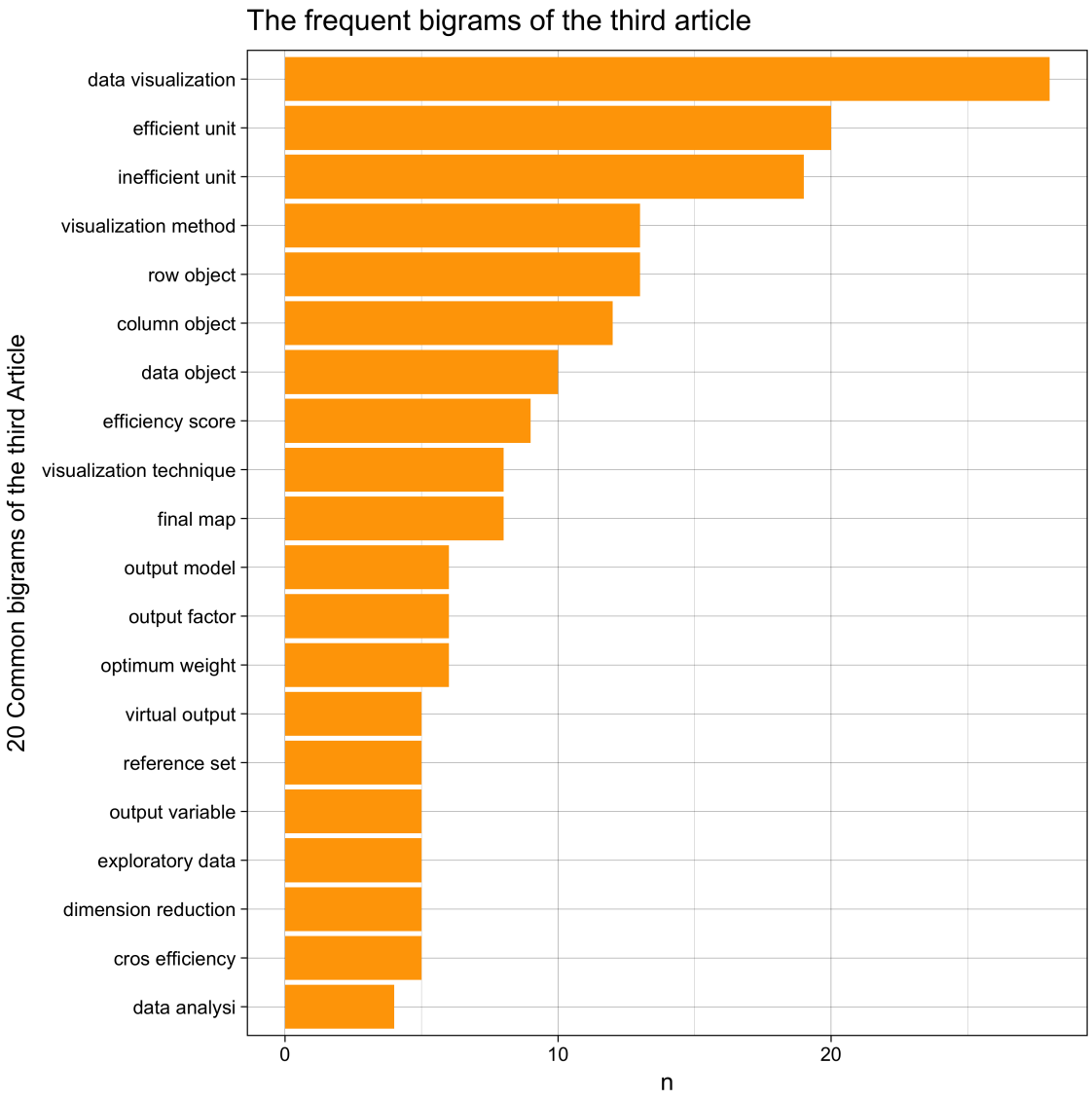

Figure 4: Frequent pairs of consecutive words in the article 3

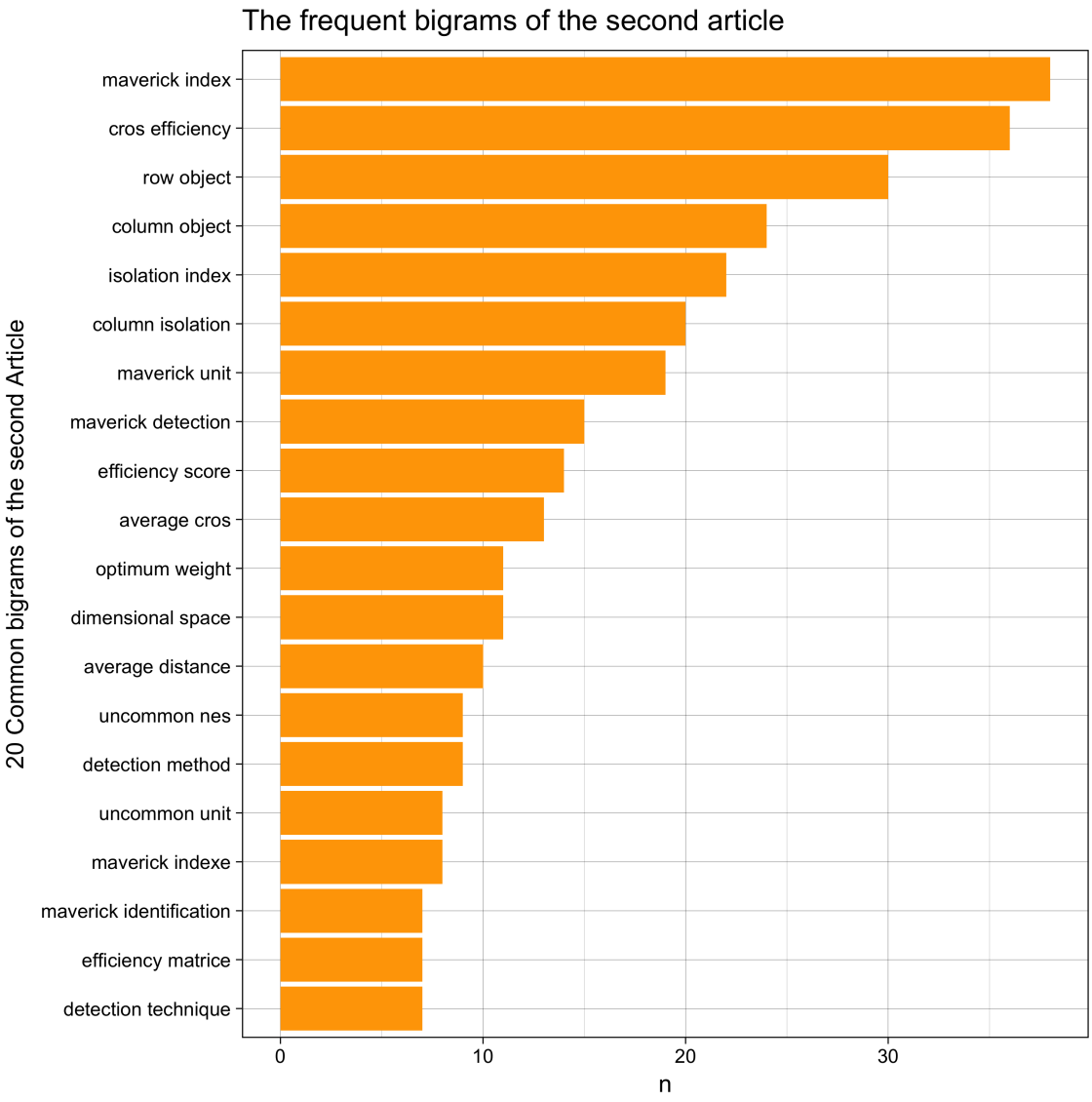

Figure 5: Frequent pairs of consecutive words in the article 2

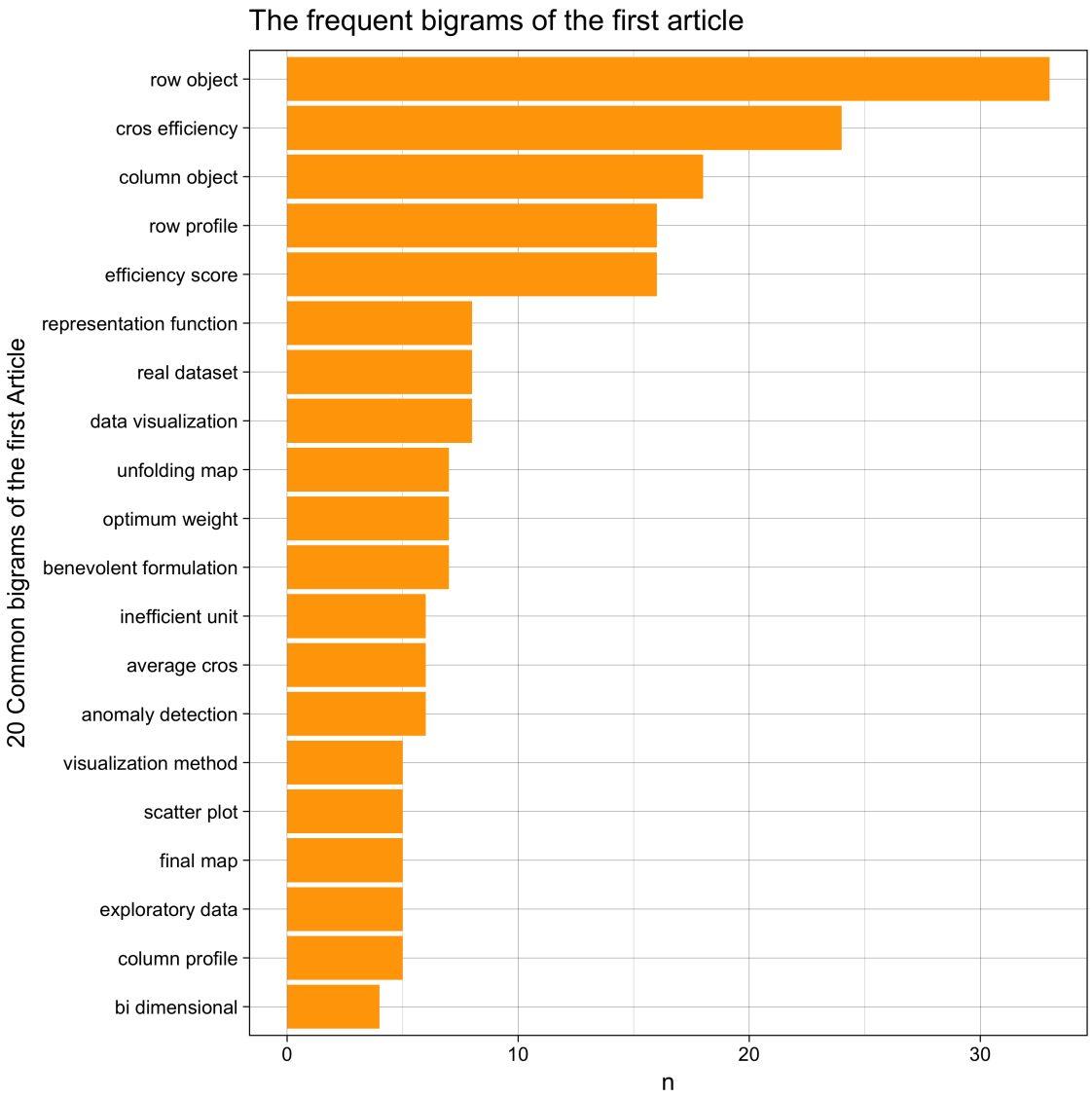

Figure 6: Frequent pairs of consecutive words in the article 3

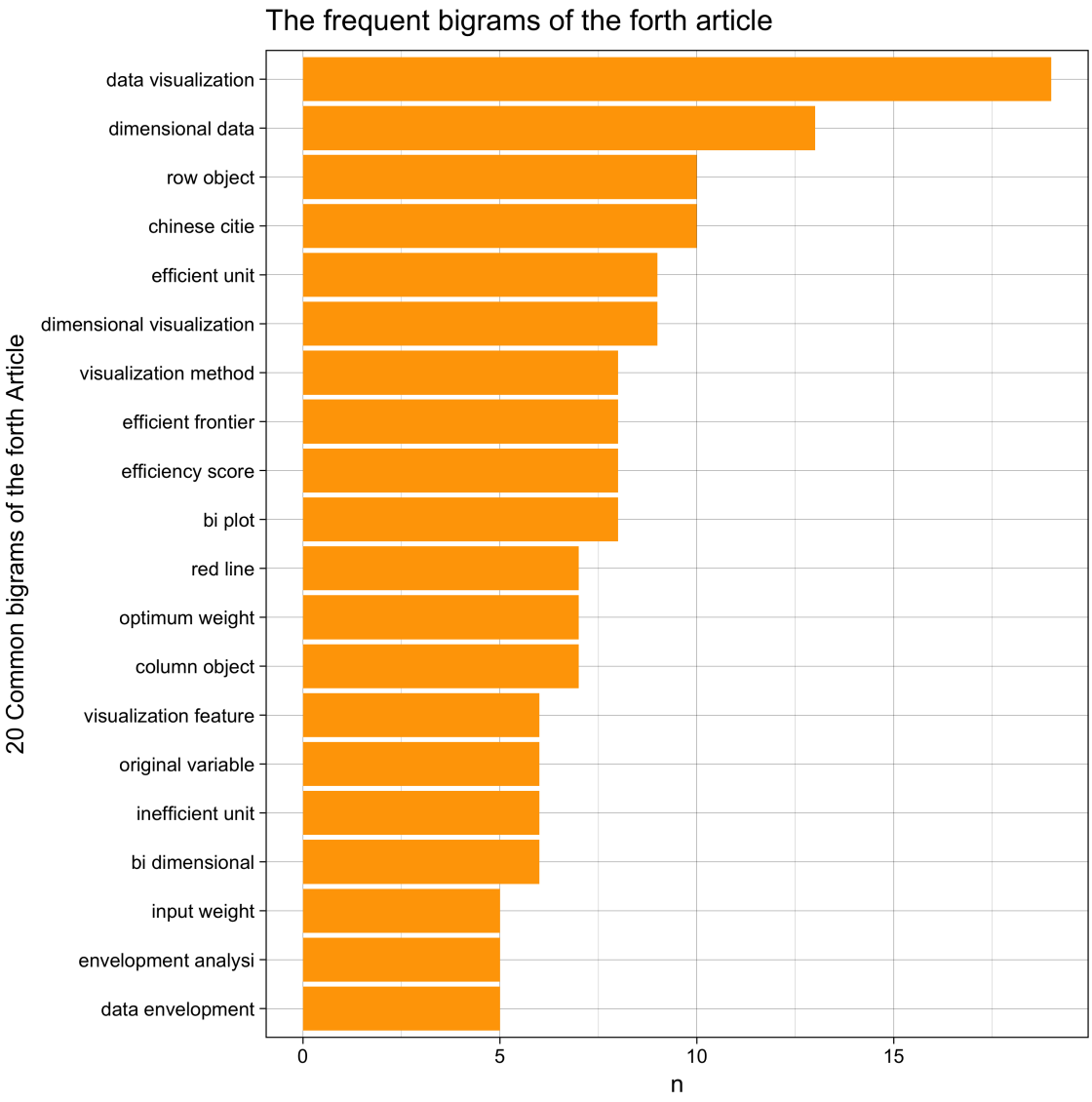

Figure 7: Frequent pairs of consecutive words in the article4

Beside the frequency bar-plots, it is possible to depict the bi-gram relations using graph representation. In these networks, the words are represented as nodes, and the relations as arrows. Hence, for a pair such as "A B", the arrow comes out of the first word, A, and points to the second word, B. The intensity of the arrows is proportional to the frequency of the corresponding bi-gram. Therefore, the more frequent pairs have more opaque and less pale arrows.

Figure 8 shows the bi-gram network of article1. Only pairs with more than 4 occurrences are chosen to be in the network, in order to make the network less crowded, and retain the most important bi-grams.

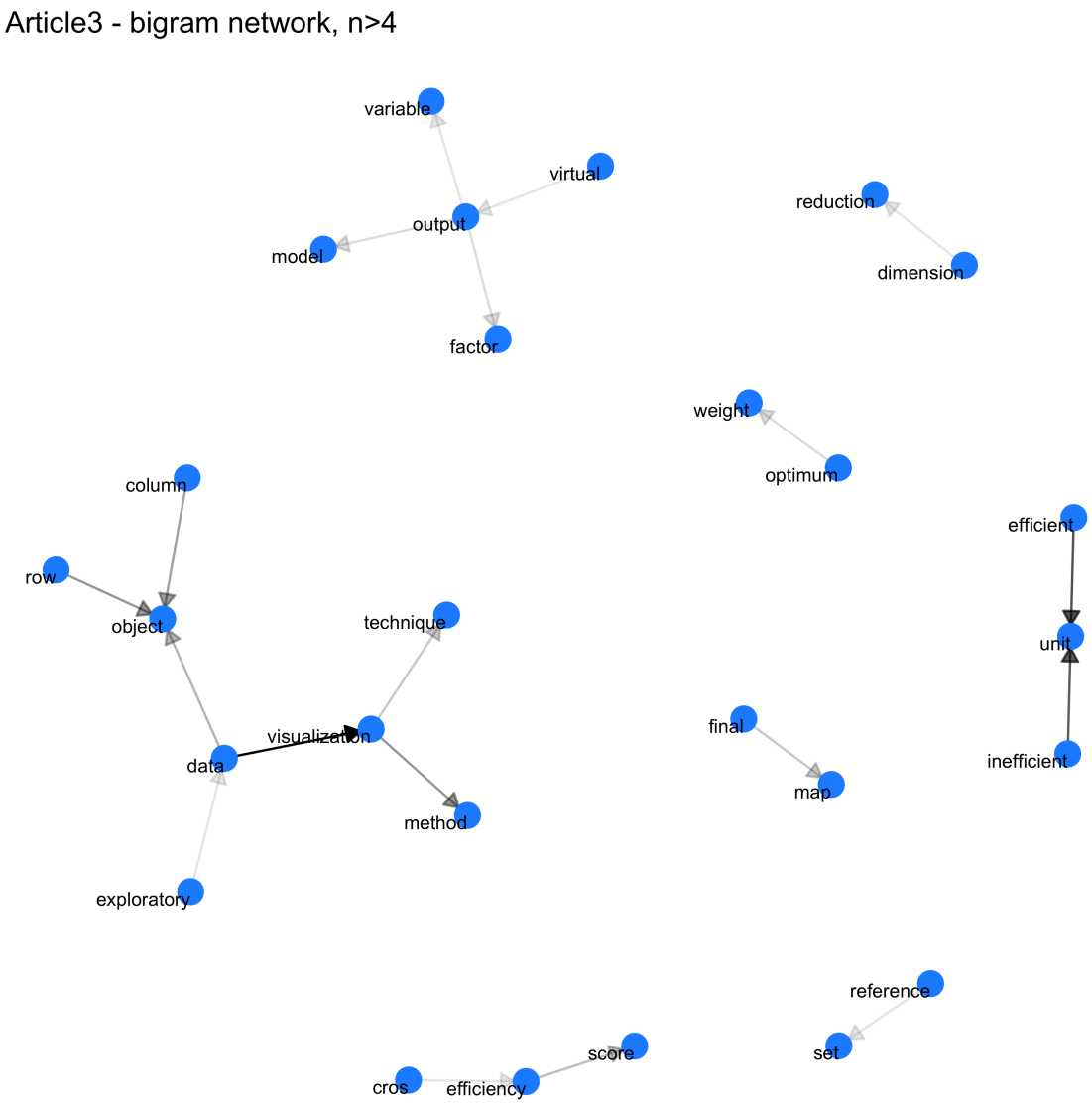

Figure 8: Bi-gram network of Article 1

It is seen that, for example, "Average-Cross- efficiency- Score", and "Exploratory-Data-Visualization-Method" are two frequent set of pairs. Considering other parts of the network, one can grasp the main concepts of the article1. Following figures are network representation of bi-gram terms of articles 2, 3 and 4.

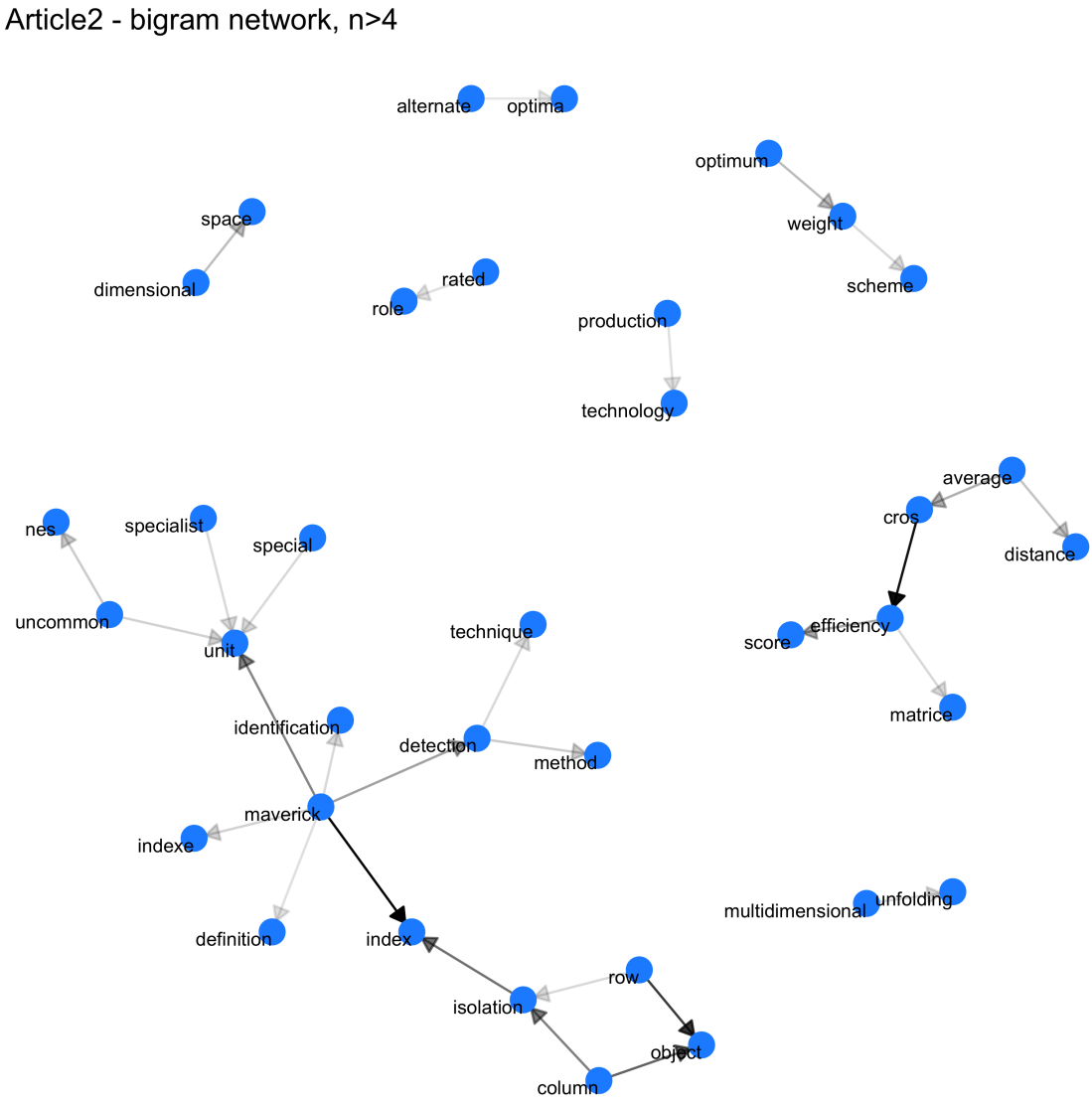

Figure 9: Bi-gram network of Article 2

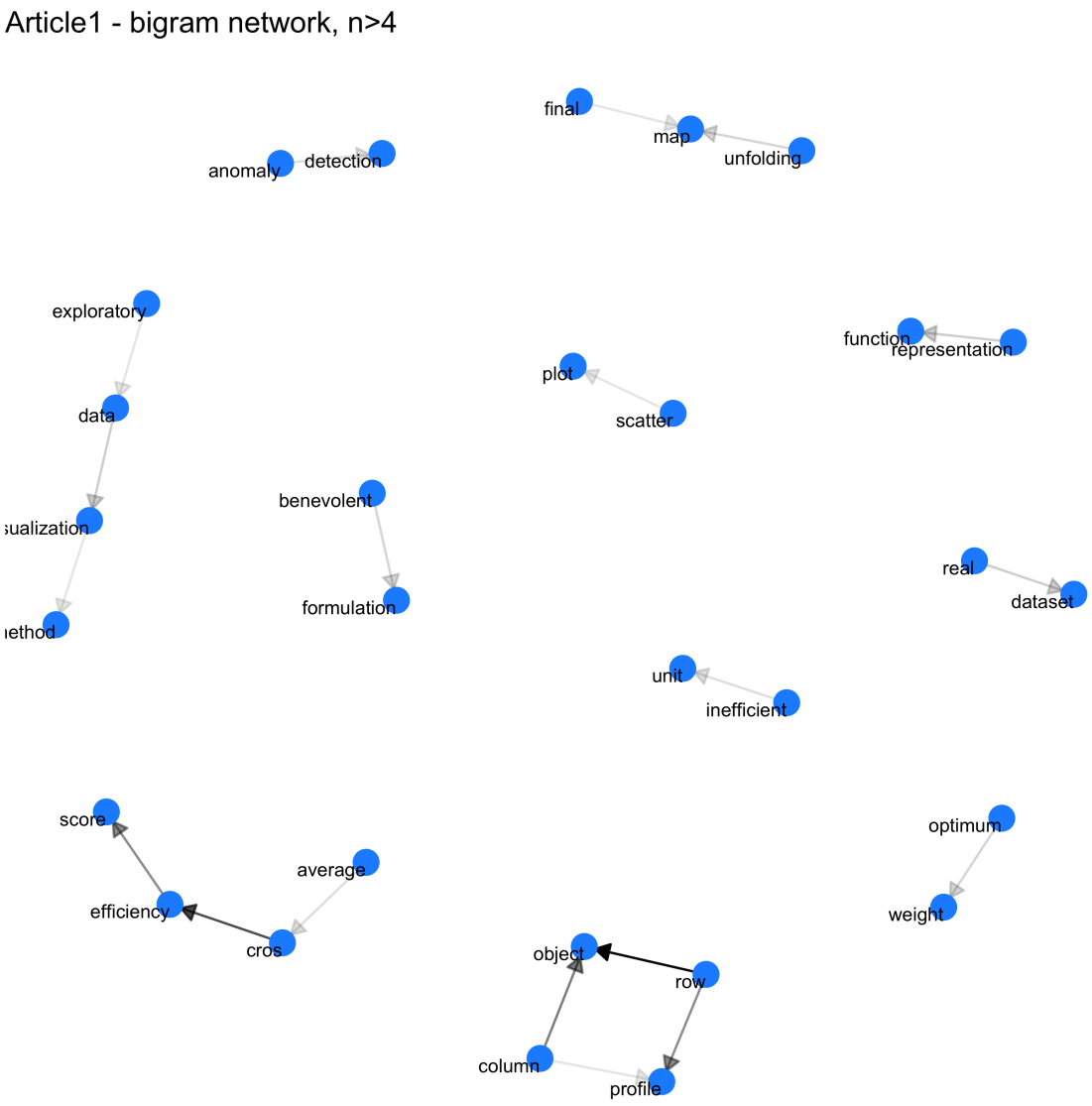

Figure 10: Bi-gram network of Article 3

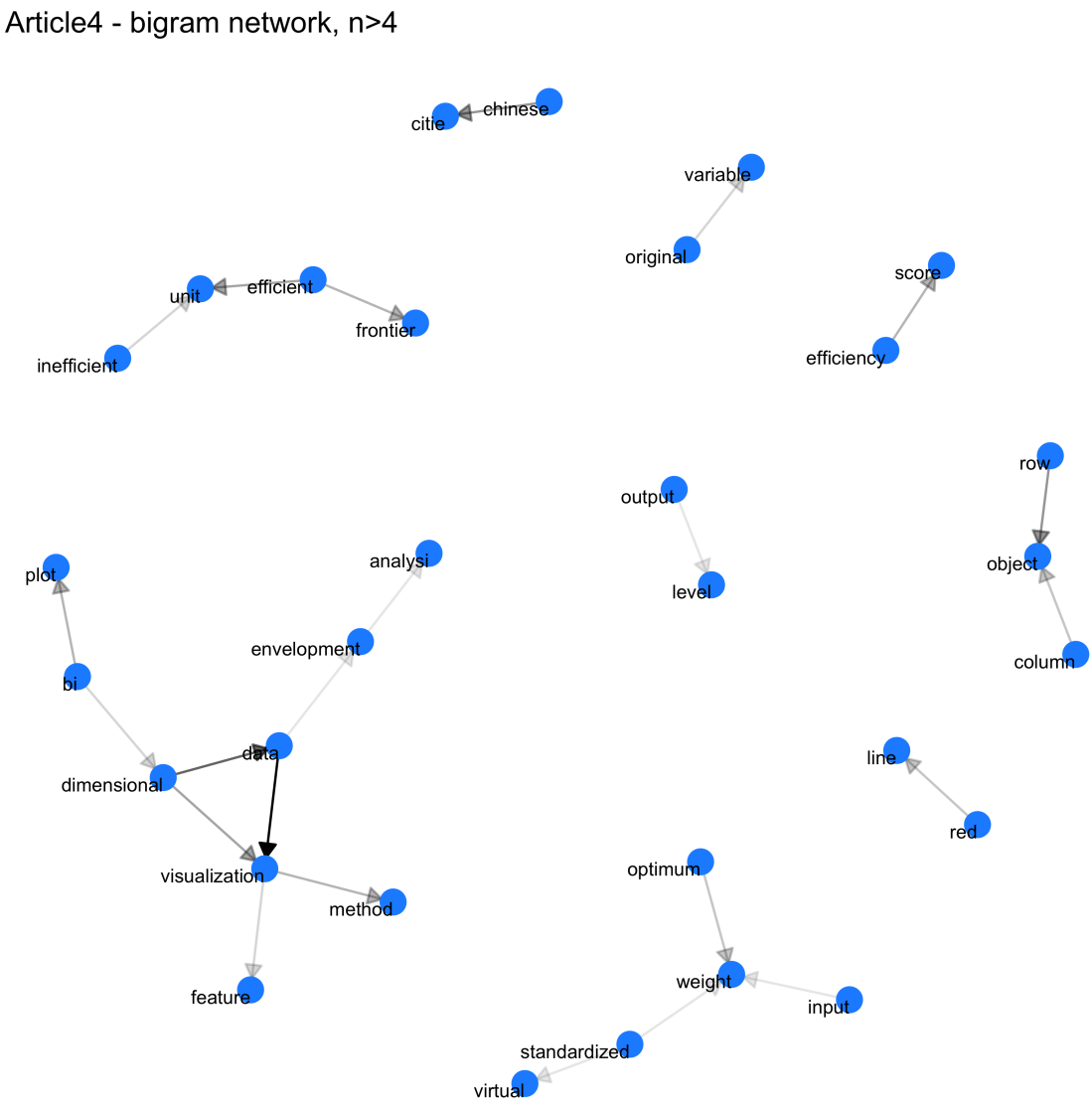

Figure 11: Bi-gram network of Article 4

5.2 Reference Analysis

In addition to the text analysis of the articles, analysis of the references of each article would be beneficial to get insight into the articles, and the thesis in general.

In order to analysis the references, the references of each paper and the frequency of citation to each reference in each paper extracted using text mining techniques. Overall, around 140 unique references have been used in this thesis. The distribution of publication years of these reference for each article is presented in figure 12

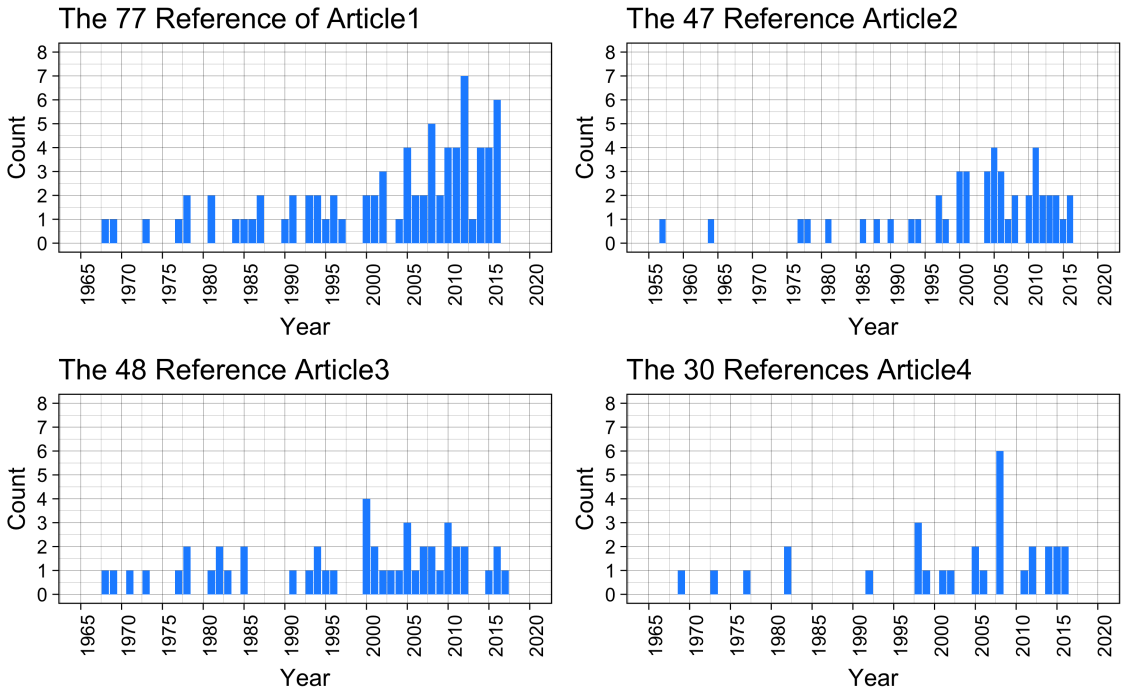

Figure 12: Publication Year Distributions of the References

The distributions are more or less left-skewed, and the modes are varied from 2000 in article 3 to 2012 in articles 1 and 2. Since the topic of this thesis is relatively niche in the field of DEA, there are relatively few papers are available in the literature on this topic.

In the next step on studying the references, I generated a network of the articles and their references. Some references are common between two or three articles of the thesis, and have been cited in more than one article. Some are uniquely related to the subject of one article and no more. I manually tagged the main subject of these references, so each of them belongs to one of the categories of : CEM(Cross-Efficiency), DEA-General(General DEA paper), DEA-Viz(Visualization of DEA), Viz(Data Visualization in general), Viz-Dr (Visualization by Dimension Reduction), and Software (Software packages that are cited). While the articles of the thesis are highlighted by red color, the references are color-mapped based on their categories. Figure 13 depicts the reference network.

The references are shown as nodes, while the links are arrows from a reference to the corresponding article, in which the reference is used. The intensity of the links is proportional to the frequency of citation of the reference in the corresponding article.

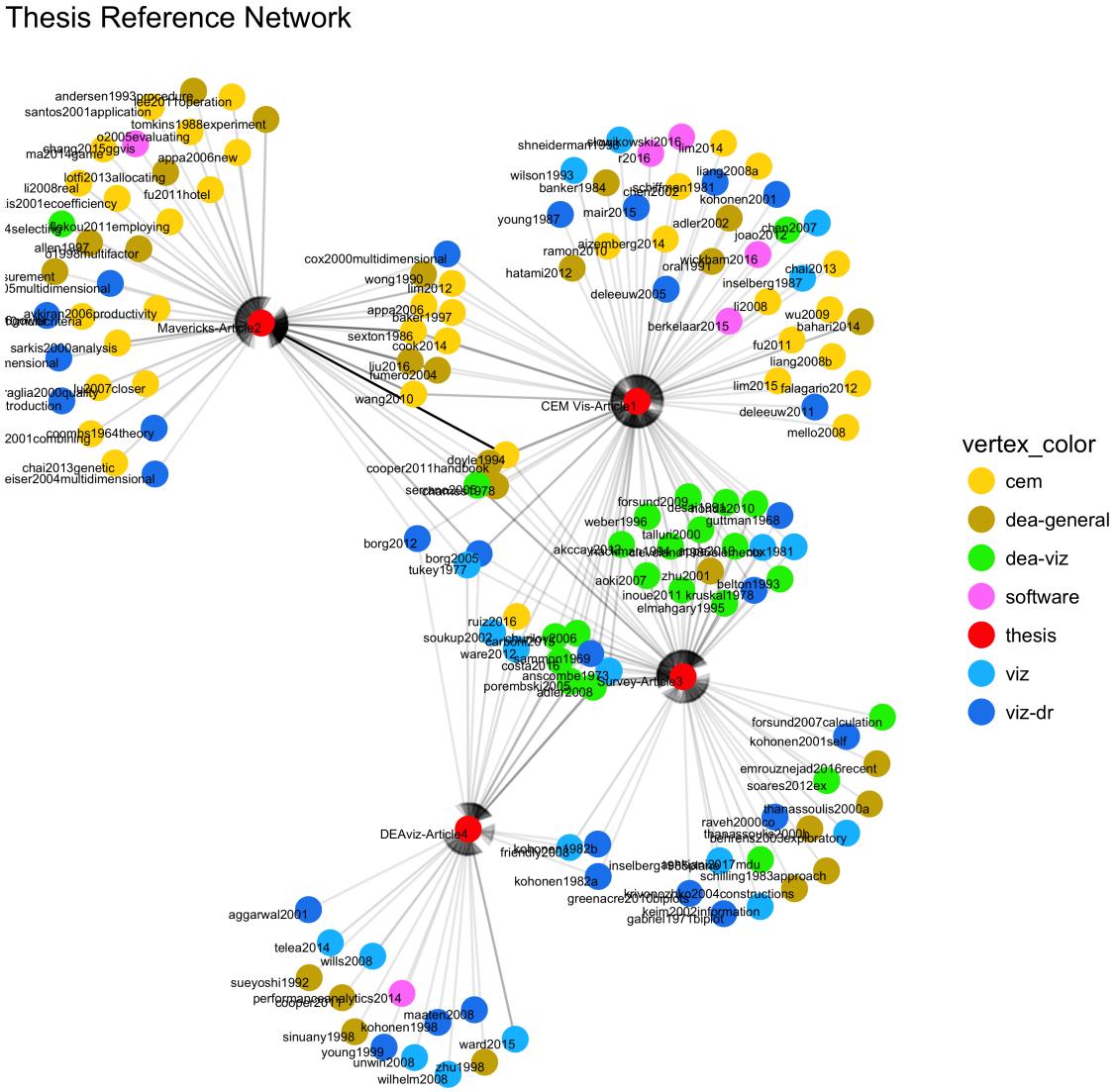

Figure 13: Complete Reference network of the thesis

This network helps in understanding the main categories, and general topics of each article. For instance, most of the references of the article 2 on top-left of the graph are yellowish, i.e. CEM and DEA, and blue, i.e. visualization through dimension reduction, and being so reveals that the article 2 is probably related to the intersection of these two categories. As noted earlier, article 2 is about the usage of dimension reduction of CEMs.

While the Figure 13 is helpful, it is too crowded to be fully comprehensible. In Figure 14, the references that have been cited only once in an article are removed from the network. Also the labels of the remaining references are made pale. Through this graph, it is much easier to identify the category of each article, i.e. the broad topics of each paper, and how the articles are related based on the common references.

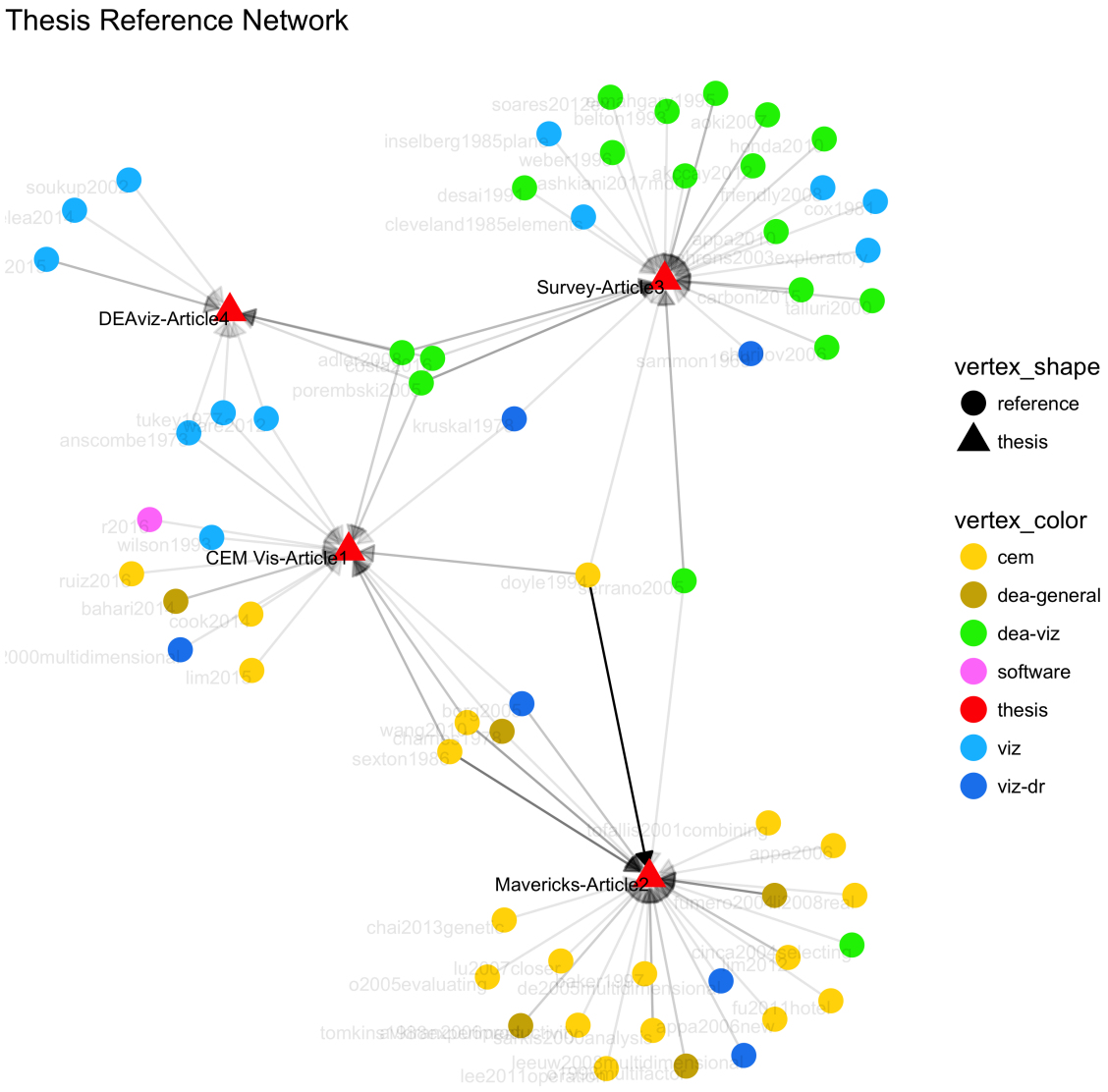

Figure 14: Filtered Reference network of the thesis

This subsection suggested a new perspective to gain insight into the thesis before reading the thesis. To the author's opinion, this perspective is more objective and

revealing than verbal introductions to each article. Moreover, a thesis on data visualization deserves a visual introduction such as this!

In the next part, the articles are presented, and then in the last part, the thesis is summarized.

Bibliography

Adler, Nicole and Adi Raveh (2008). "Presenting DEA graphically". In: *Omega* 36.5, pp. 715–729.

Anscombe, Francis J (1973). "Graphs in statistical analysis". In: *The American Statistician* 27.1, pp. 17–21.

Coelli, Timothy J, Dodla Sai Prasada Rao, Christopher J O'Donnell, and George Edward Battese (2005). *An introduction to efficiency and productivity analysis*. Springer Science & Business Media.

Cooper, William W, Lawrence M Seiford, and Kaoru Tone (2006). *Introduction to data envelopment analysis and its uses: with DEA-solver software and references*. Springer Science & Business Media.

Cooper, William W, Lawrence M Seiford, and Joe Zhu (2011). *Handbook on data envelopment analysis*. Vol. 164. Springer Science & Business Media.

Fellows, Ian (2014). *wordcloud: Word Clouds*. R package version 2.5. URL: https://CRAN.R-project.org/package=wordcloud.

Pedersen, Thomas Lin (2017). *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. R package version 1.0.0. URL: https://CRAN.R-project.org/package=ggraph.

Ramanathan, R (2003). *An Introduction to Data Envelopment Analysis: A Tool for Performance Measurement*. SAGE Publications Pvt. Ltd. ISBN: 0761997601.

Ray, Subhash C (2004). *Data envelopment analysis: theory and techniques for economics and operations research*. Cambridge university press.

Silge, Julia and David Robinson (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R". In: *JOSS* 1.3. DOI: 10.21105/joss.00037. URL: http://dx.doi.org/10.21105/joss.00037.

Team, R Core (2016). *R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014*. URL: https://www.R-project.org/.

Thanassoulis, Emmanuel (2001). *Introduction to the theory and application of data envelopment analysis*. Springer.

Ward, M.O., G. Grinstein, and D. Keim (2010). *Interactive Data Visualization: Foundations, Techniques, and Applications*. 360 Degree Business. CRC Press. ISBN: 9781439865545. URL: https://books.google.be/books?id=Kk7NBQAAQBAJ.

Ware, Colin (2012). *Information visualization: perception for design*. 3rd ed. Elsevier.

Zhu, Joe and Wade D Cook (2007). *Modeling data irregularities and structural complexities in data envelopment analysis*. Springer Science & Business Media.

Part II

Appended Essays

Article 1

Visualization of Data Envelopment Analysis Problems: A Visual Survey

Abstract

Exploratory data visualization is an essential step in analysis of quantitative problems among various fields. It can improve understanding of the researchers from a dataset, and lead to gain more profound insight into the problem. Graphical representations can encapsulate and represent large amount of data in a holistic manner, which makes researchers able to look at the big picture, and identify the possible emergent properties of the data. More specifically, data visualization can reveal patterns and regularities such as homogeneous groups, as well as irregularities and anomalies such as mavericks and outliers. However, the goal of exploration is not restricted to identification these items.

Nevertheless, very little attention is paid to data visualization in data envelopment analysis (DEA) field. This relative neglect is not mainly due to lack of proper visualization tools, but it is probably related to lack of visibility of the currently available DEA data visualization methods in addition to disregard of the importance of exploratory visualization.

This article is a comprehensive survey of available DEA visualization methods. It is tried to be as exhaustive as possible, and from this aspect it is the first of its kind in DEA literature. Currently, there is no similar survey in the DEA literature, and all literature reviews are selective to such extent that they have missed several important DEA visualization studies. Another advantage of the current article is its "visual approach" to visualization, in contrast to other brief "verbal" literature reviews. In other words, a dataset is graphically represented using eight significant visualization methods, in order to illustrate those methods. Finally, this survey, as a showcase of DEA visualization techniques, tends to promote DEA visualization in the DEA community.

Keywords: Data Envelopment Analysis, Data Visualization, Survey, Exploratory Data Analysis, Outlier Detection

1 Introduction

Data visualization has been around for centuries, and it is not a recent idea. However, since the second half of 20th century, information visualization has had a re-birth (Friendly 2008). Part of this rebirth is due to technological advances both from software aspect, e.g. classic Fortran to recent R statistical package, and hardware aspect, e.g. unprecedented powerful personal computers. Nonetheless, such significant

advances are mainly built upon theoretical ground of exploratory data analysis(EDA). EDA, not surprisingly, is about exploration of the data in order to gain better and deeper insight into the it, and it is "about looking at data to see what it seems to say", and making the data "more easily and effectively handle-able by minds." (Tukey 1977)

The motivations of data visualization can be summarized as follows:

1. **Providing a holistic view to the data** (Cleveland and Cleveland 1985, p. 10): It is very important to have an overall understanding of the data, i.e. looking at the data from a bird-eye view. Simultaneous consideration of observations and their relationships would yield a holistic view of the data. Examination of the big picture oftentimes results in discovering relations and structures which otherwise would remain concealed. This holistic view contrasts with reductionist analytical tools, which examine any observation regardless of the big picture, or summarize a set of data into a single digit in the cost of losing details.

2. **Improving the perception of analyst from the dataset and thus the problem** (Cleveland and Cleveland 1985, p. 9): Since data visualization is a tool to facilitate exploration of the data, such exploration and investigation would essentially award better understanding of the dataset. As an exploratory data analysis(EDA) approach, data visualization ultimately "... provides a sense of intimacy with the nuances of the data" (Behrens and Yu 2003)

3. **Discovering patterns and regularities, and thus irregularities in the data** (Keim 2002): "Pattern discovery" is considered as the main goal of exploratory data analysis. (Behrens and Yu 2003)

4. **Generation of further and more profound questions** (Cox and Jones,1981) (Cox and Jones 1981a): Exploratory data analysis usually leads to further questions, and bears new hypotheses about the data. Such questions necessitate further investigations, which in turn results in deeper understanding of the problem, or new insights from novel perspectives to the problem.

According to Friendly (2008), the current focus of data visualization, after its recent re-birth, is on high-dimensional, as well as dynamic and interactive data visualization. This point reminds us the fact that most of real-world DEA problems have several separate datasets which are high-dimensional each. In other words, beside the input and output variables dataset, which has usually more than three variables, each model of DEA generates further data that can be used in data visualization in order to have a holistic view, or to have a more profound insight into the data, or to have a look at the data from a specific perspective in the presence of one or more specific variables. The combination of such datasets in order to improve understanding of the problem increases the dimensionality of the DEA problem as well.

Considering the coincidence of the trend of high-dimensional data visualization, and emergence of related techniques, as well as the high-dimensionality of the DEA

problems from one side, and the evident benefits of data visualization from another side, it is expected to have data visualization as a routine and regular section of each DEA application. However, this is not the case in practice.

Seemingly, the importance of data visualization in DEA is vastly neglected by practitioners and researchers, to such extend that there is no visualization beyond uni-variate or bi-variate graphs in the DEA handbooks such as "the handbook of data envelopment analysis" (Cooper et al. 2011), or even in the recent proceedings of the main DEA conference, "The New applications of data envelopment analysis" (Emrouznejad et al. 2016), to just name a few. Since there are a dozen of published DEA visualization methods, it is safe to claim that such neglect is not due to lack of DEA visualization tools. The reason may lie in taking exploratory data visualization as a trivial task by DEA community, or may be due to invisibility of a few visualization methods among myriad DEA studies.

Although several techniques for graphical representation of high dimensional DEA problems are available, the current DEA visualization toolbox is not a perfect set of tools, or the best that we can have. Considering the various models of DEA, and accelerated advances in both data and information visualization, there is an unexploited potential for amelioration of such toolbox. This paper is a survey of the available DEA visualization tools, an incisive introduction to each of them, and illustration of the visual outcome of the significant ones. The main goal of this paper is presentation of the sporadic methods in a comprehensive yet succinct study in order to help DEA researchers and practitioners. The researchers can find opportunities to improve the current tools, or add new tools to this toolbox, while the practitioners can use the techniques to visually explore their DEA problems. Moreover, this survey intends to emphasize the importance of DEA data visualization through the depiction of the capabilities of visual configurations, and promote exploratory DEA visualization through making the benefits of visualization methods more visible.

In order to verbally introduce visualization methods, three aspects of each method are highlighted. The aspects are the dataset to be visualized by the method, the visualization technique, and main characteristics of the final graphical configuration. These information in addition to some more details about the methods are presented in Table 1.3 and Table 1.4 of this paper.

More importantly, a specific DEA dataset is visualized using a selection of methods throughout the paper. Doing so not only makes the readers able to compare the methods, but also relates the methods to each other, since these techniques are more complementary rather than competitive. Each of these techniques sheds light on a DEA problem from a different perspective, and while these perspectives have overlaps, there is no reason to restrict ourselves to using only one of them. Hence, there is no "the best choice" among these visualization method, and there is no need to have only one.

In order to achieve comprehensiveness, all proceeding papers as well as journal articles related to DEA visualization have been reviewed in this survey. However, these studies are categorized into two broad categories: the main and the peripheral methods. This categorization is based on two criteria: whether the method is published as a journal paper, and whether the method is based on general DEA

models. In other words, the methods that are found only in conference proceeding booklets, and the methods that are based on ad-hoc DEA models are assigned to the peripheral category. Doing so does not mean that the methods of the peripheral category are not useful or capable, but since this paper is aimed for the broadest audience in the DEA community, such ad-hoc or semi-baked methods are separated from the rest.

2 DEA Visualization Methods

2.1 The data set

The dataset, used in this paper to illustrate a selection of DEA data visualization techniques, is about 47 Japan prefectures with their 3 inputs and 3 outputs. The inputs are "area", "population" and "household income", and the outputs are "manufacturing shipment amount", "agricultural shipment amount" and "commercial sales amount". This dataset is published in Aoki et al. (2007), and originally is derived from Yomuri Shinbun year book of 2002. The dataset is presented in Table 1.1.

It is important to underscore that the dataset is chosen merely with the aim of illustration of the DEA visualization techniques, and thus any other dataset could be selected, regardless of its domain and validity. The Japan prefectures dataset is chosen due to the number of DMUs and the number of inputs and outputs, since too many DMUs would overcrowd some of the visual maps, and too few would understate the visualization. Similar rationale applies to the number of variables. While the purpose of this survey is illustration of the DEA visualization methods, it is important to pay enough attention to the scaling capabilities of the methods, i.e. how useful any specific method is in coping with larger datasets.

Table 1.2 includes the constant return to scale(CRS) and variable return to scale(VRS) efficiency scores of the 47 DMUs. Where ever a visualization method needs efficiency scores, CRS scores are chosen to be used, unless otherwise is stated. The reason of choosing CRS efficiency scores is related to the fewer number of efficient units under such assumption, and higher number of visualization techniques that work based on CRS scores, since some techniques are devised based on CRS models, and fewer can work with both VRS and CRS results.

Table 1.1: 47 Japan Prefectures with three inputs and three outputs

DMU No.	Prefecture	Area	Population	Household Income	Manufacturing shipment Amount	Agricultural Shipment Amount	Commercial Sales Amount
1	Hokkaido	83453	5707654	510910	57137	10551	223000
2	Aomori	9235	1472633	524671	13479	2648	41027
3	Iwate	15278	1413099	514243	23058	2849	40455
4	Miyagi	6861	2368591	466685	37492	2202	125793
5	Akita	11434	1183380	593805	16201	2058	35325
6	Yamagata	7394	1240877	596394	27451	2372	32899
7	Fukushima	13782	2124404	726739	53897	2651	54836
8	Ibaragi	6096	2991172	606735	105251	4147	78669
9	Tochigi	6408	2009064	539194	75784	2746	60558
10	Gunma	6363	2031732	423112	80682	2289	62677
11	Saitama	3767	6975947	587025	138134	2052	170111
12	Chiba	4996	5963514	537057	111173	4448	134275
13	Tokyo	2102	12166713	588143	180966	312	2031190
14	Kanagawa	2415	8561001	669167	213177	827	230377
15	Niigata	10939	2470837	671495	45952	3141	85106
16	Toyama	2802	1120320	727871	33527	800	38725
17	Ishikawa	4185	1180525	675826	24757	685	52128
18	Fukui	4189	828502	575075	18771	591	27799
19	Yamanashi	4201	889808	585536	23711	925	21648
20	Nagano	12598	2220208	515170	64803	2558	74064
21	Gifu	10209	2111893	630817	48699	1275	61251
22	Shizuoka	7329	3781677	622039	159122	2800	125139
23	Aichi	5117	7043300	577650	330531	3419	525132
24	Mie	5761	1862307	570017	76692	1334	44289
25	Shiga	3855	1342811	613583	61288	746	29394
26	Kyoto	4613	2644391	540202	54243	741	88244
27	Osaka	1893	8805081	481173	181207	377	766023
28	Hyogo	8392	5568305	456470	135787	1676	158703
29	Nara	3691	1140920	558440	23941	567	23602
30	Wakayama	4726	1066427	554173	21592	1174	22415
31	Tottori	3507	613097	489579	12194	770	16861
32	Shimane	6707	759693	527025	10925	685	18691
33	Okayama	7009	1950831	526963	63320	1362	64024
34	Hiroshima	8480	2877718	632261	68686	1160	142403
35	Yamaguchi	6110	1522749	640826	46736	835	43021
36	Tokushima	4145	822784	565567	15165	1242	21157
37	Kagawa	1862	1022827	611451	21571	845	52165
38	Ehime	5676	1489732	539538	34360	1452	43094
39	Kouchi	7105	812450	552280	6104	1096	19506
40	Fukuoka	4839	5028729	522150	75490	2388	266485
41	Saga	2439	875689	563563	15866	1455	21240
42	Nagasaki	4092	1516523	527037	13897	1369	37889
43	Kumamoto	6908	1859859	535312	24904	3358	48177
44	Oita	5804	1220061	592083	27758	1520	30555
45	Miyazaki	6684	1167904	555217	12863	3128	30229
46	Kagoshima	9132	1782954	576675	19801	4048	45750
47	Okinawa	2271	1327632	432973	6152	902	26751

Table 1.2: CRS and VRS efficiency scores of 47 Japan prefectures

DMU No.	Prefecture	CRS Efficiency	VRS Efficiency
1	Hokkaido	1	1
2	Aomori	0,79	0,96
3	Iwate	0,91	1
4	Miyagi	0,73	0,95
5	Akita	0,78	0,87
6	Yamagata	0,93	0,97
7	Fukushima	0,78	0,79
8	Ibaragi	1	1
9	Tochigi	1	1
10	Gunma	1	1
11	Saitama	0,73	0,86
12	Chiba	1	1
13	Tokyo	1	1
14	Kanagawa	1	1
15	Niigata	0,75	0,77
16	Toyama	0,72	0,96
17	Ishikawa	0,6	0,81
18	Fukui	0,61	0,9
19	Yamanashi	0,74	0,92
20	Nagano	0,84	0,88
21	Gifu	0,57	0,72
22	Shizuoka	0,96	0,96
23	Aichi	1	1
24	Mie	0,94	0,95
25	Shiga	0,99	1
26	Kyoto	0,46	0,81
27	Osaka	1	1
28	Hyogo	0,54	1
29	Nara	0,5	0,88
30	Wakayama	0,66	0,87
31	Tottori	0,72	1
32	Shimane	0,53	0,92
33	Okayama	0,76	0,86
34	Hiroshima	0,61	0,73
35	Yamaguchi	0,7	0,81
36	Tokushima	0,76	0,94
37	Kagawa	0,76	1
38	Ehime	0,68	0,85
39	Kouchi	0,58	0,9
40	Fukuoka	0,72	0,91
41	Saga	1	1
42	Nagasaki	0,59	0,87
43	Kumamoto	0,95	1
44	Oita	0,74	0,84
45	Miyazaki	1	1
46	Kagoshima	1	1
47	Okinawa	0,57	1

2.2 DEA Visualization toolbox

The order of the discussed papers in each category is mainly chronological. As explained in the previous section, the visualization methods are divided into two categories: main and peripheral. While all the conference proceedings and journal papers which have any contribution DEA visualization, i.e. visualization methods beyond simple efficiency histograms, are included in this paper, the dataset of Table 1.1 is visualized using only the main methods, i.e. the methods of the former category. The methods of the main category meet two criteria: first, the corresponding paper is a journal paper, and second the method is a general method based on frequently-used DEA models. Thus, it was decided that verbal review of conference proceeding papers, and relatively ad-hoc methods is sufficient. Therefore, the focus of the survey would be mainly on the visualization methods categorized in the former group.

It is of importance to emphasize that explanations of the methods include information about what data object the method visualizes, which algorithm, if any, is used in order to do so, and what characteristics the visual output has.

2.3 The main visualization methods

As stated earlier, in this survey, the DEA visualization methods are separated into two categories, called the main methods, and the peripheral methods. The methods of the first category are illustrated using visualization of the Table 1.1 dataset. These methods have some common characteristics, which put them in the main category. They are all published in peer reviewed journals, and they are either independent of DEA-models or based on very common DEA models such as (Charnes et al. 1978), (Banker et al. 1984) or Cross-efficiency(Sexton et al. 1986). Hence, it is anticipated that the visualization outcomes of these methods are important for a wide spectrum of audience in DEA community.

Information visualization(InfoVis) is the body of knowledge and techniques for transforming digital information into visual symbols. To do so, InfoVis uses "visual marks" to represent observations, and "visual channels" to represents corresponding variables of those observations. In other words, each observation is represented by a visual mark, such as a point or a line, and a collection of corresponding variables are represented by visual channels, such as visual marks' size, color, or shape. This translation is called "encoding" in the InfoVis jargon. Here, in order to explain the main DEA InfoVis techniques from a new perspective, visual marks and visual channels of each DEA-viz technique are elaborated as well as their limitations and potencies. Moreover, the main points that each technique intends to show or investigate would be highlighted. At the end, the prominent aspects of these techniques are summarized in Table 1.4.

2.3.1 Parallel Coordinates

The first method is suggested and developed by Desai and Walters (1991) and later Weber and Desai (1996). The authors benefit from parallel coordinates graphical

presentation in order to visualize DMUs based on the values of the inputs and output factors. Using parallel coordinates, each factor can be shown as an axis or coordinate, and consequently several factors can be shown as parallel axes or coordinates in one plot. Thus, factors of a DMU are determined by points on these axes, and every DMU is visualized by a continues piece-wise line that connects these points. Therefore, no dimension reduction technique has been used by the authors, and the high-dimensional data is tried to be precisely represented on the parallel coordinates plot. The outcome map can visually identify benchmarks and representation of the DEA efficient units and their variable ranges, inefficient units and paths of improvement for these inefficient units to the efficient surface. Additionally, the range of feasible levels of inputs and outputs according to the efficient units, within which other units must lie in order to be one on the efficient frontier, is depicted on the final plot. The path of improvement, suggested for inefficient DMUs on the plots, is based on the "value path" of (Schilling et al. 1983) and (Inselberg 1985). Figure 1.1 is the representation of efficient DMUs of Table 1.1 under constant return to scale (CRS) assumption as well as one arbitrary chosen inefficient unit, DMU2. The efficient units are depicted by dashed lines while the inefficient unit is shown by continues black line.

Parallel coordinates is mainly focused on circumventing the difficulties of representation of high-dimensional observations, i.e. observations with more than four variables. In contrast to Cartesian coordination system where there are two orthogonal coordinates each represents one variable, parallel coordinates system can theoretically represent unlimited number of variables as coordinates which are parallel to each other, rather than orthogonal. Each observation is represented as a line, i.e. the visual mark is the line, and the interception of the line with each coordinate is the corresponding variable value.

This method is suitable for detection of DMUs with extreme values, based on distribution of values over a single variable. For instance in the Figure 1.1, DMU23 has an extreme value on the coordinate related to manufacturing. Moreover, considering its whole line, we can gain an overall picture of a specific DMU in comparison to others based on inputs and outputs values. DMU2 is highlighted in the figure1, which shows this unit has commercial and manufacturing values, with an average agricultural output. While it has relatively high “area” input, its population and household income are among the lowest.

The map of Figure 1.1 is depiction of the main idea of Desai and Walters (1991), except for two differences. The authors have used original variables in contrast to normalized variables of Figure 1.1, and the feasibility range of Inselberg (1985) is absent in Figure 1.1. Nevertheless, the main idea, which is visualization of inputs and outputs using parallel coordinates, is preserved. Furthermore, the authors suggest to re-order the variables based on the managerial control on them. It seems that presentation of all units on one parallel coordinate graph, when the number of units is relatively large, can reduce the usefulness of the graph. Similarly, with increase of the number of inputs and outputs, the revealing feature of the map may depreciate.

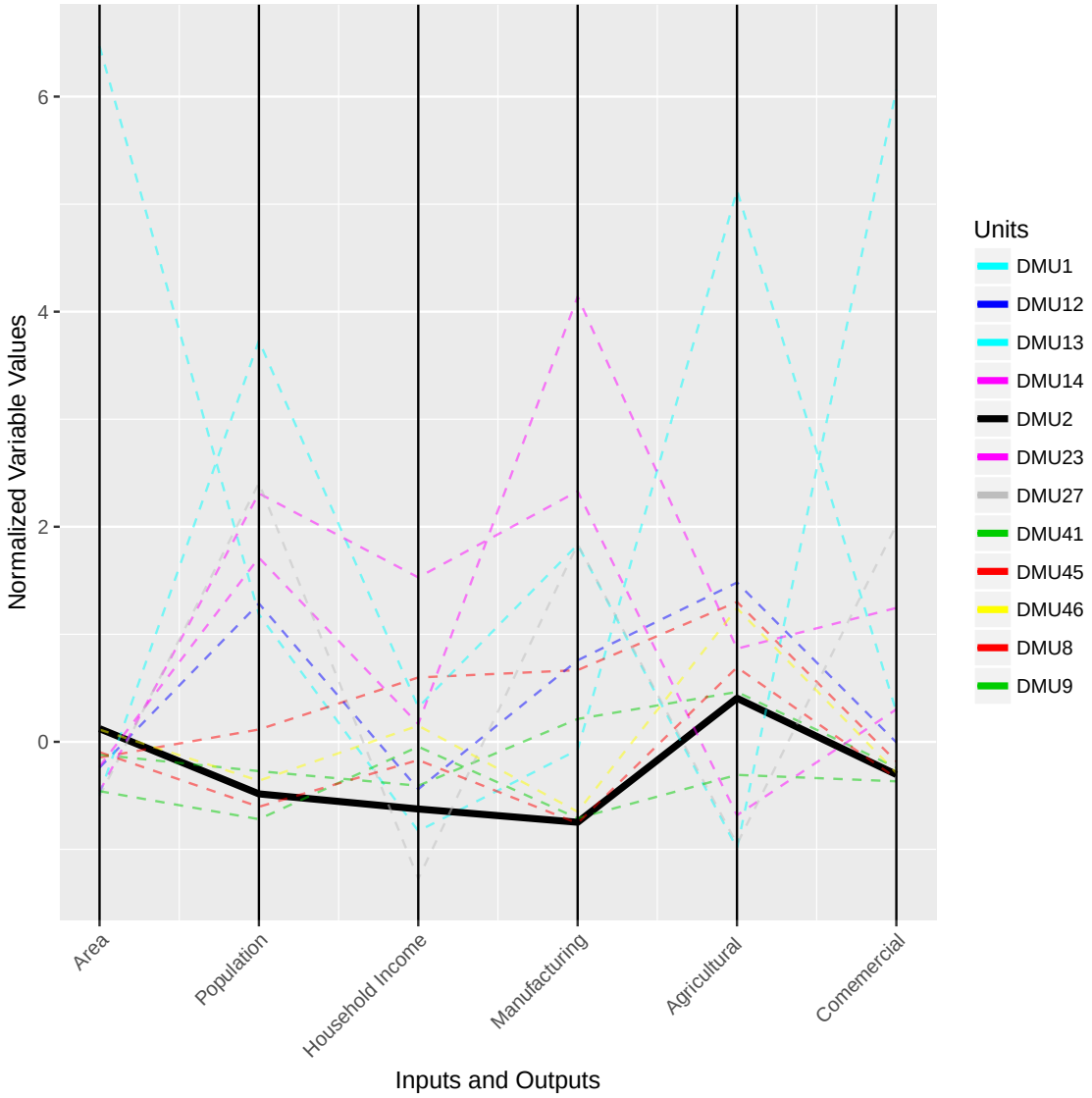

Figure 1.1: Weber and Desai (1996). Representation of 12 efficient DMUs and one inefficient DMU on Parallel Coordinates. The inefficient DMU2 has relatively high level of input1, while relatively low levels of input 2 and input 3, comparing to the efficient units. It can be seen that the output 1 and output 3 of this unit are its weak points, considering the efficient units' output ranges.

2.3.2 Scatter-plots, Bar-plots, Line-plots, ...

El-Mahgary and Lahdelma (1995) suggest a series of 2D plots in order to visualize DEA problems from different perspectives. The authors target "managerial community" and try to present a visual communication and presentation framework for them. To do so, the authors suggested a set of six plots.

Since the suggested plots in (El-Mahgary and Lahdelma, 1995) are very customizable and structurally familiar to common statistical graphs, it is sufficient and feasible to review only a few of them. These plots are mainly for representation of two variables regarding the two orthogonal coordinates and additionally one or two more variables encoded to shape or colors of visual marks. Figure 1.2, 1.3 and 1.4 are the visualization of Table 1.1 through the first, the second, and the third suggested plots of El-Mahgary and Lahdelma (1995), respectively.

In the line-plot of Figure 1.2, the DMUs are represented as point visual marks. The points are ordered horizontally based on a specific variable, here “population”, and a second variable is depicted as the vertical coordinate, here the CRS efficiency score. The ultimate goal is finding the probable association between these two variables, i.e. what a correlation scatterplot does. The line is supposed to provide "a global view on the effect of the unit’s variable on its efficiency" El-Mahgary and Lahdelma (1995).

Figure 1.3 is a scatterplot, each DMU is represented as a point. Two visual channels are the horizontal and vertical coordinations of these points, and the goal of the plot is detection of association between these two variables, i.e. visual channels. Moreover, this is a proper plot for identification of outliers, based on the two chosen variables.

In Figure 1.4, each DMU is represented by a bar. Three visual channels are showed, one on the vertical axes as the hight of bars, i.e. the position of upper edge of bars, one as the length of each color section, and the last visual channel is the color of these sections. The hight of bars is representation of “efficiency score”, the length of sections of each bar shows virtual outputs contribution to that efficiency score, and the color is for separation of these virtual outputs. The main goal of this plot is for evaluation of output contributions to each DMU’s efficiency score. Additionally, this plot can be used for detection of units that totally disregard one or a few outputs. Nevertheless, by increase of the number of outputs, reading the plot would become more difficult.

This plot, using the weight distributions of outputs, tries to highlight the outputs on which the efficient and inefficient units emphasize. Nevertheless, due to the alternate optima of the weights of efficient units, the plot should only be used for inefficient units.

From Figure 1.2, it can be seen that while efficient DMUs are scattered on the population range, the higher end of the population range includes more efficient units. This finding is supported by Figure 1.3, since the frequency of efficient units is considerably higher in the units with more than 5,000,000 population. Figure 1.4 is interpretable only for inefficient units, since the virtual outputs of efficient units can change due to alternate optima phenomenon of DEA. Nevertheless, there are interesting points in inefficient units such as unit four and five, which are totally

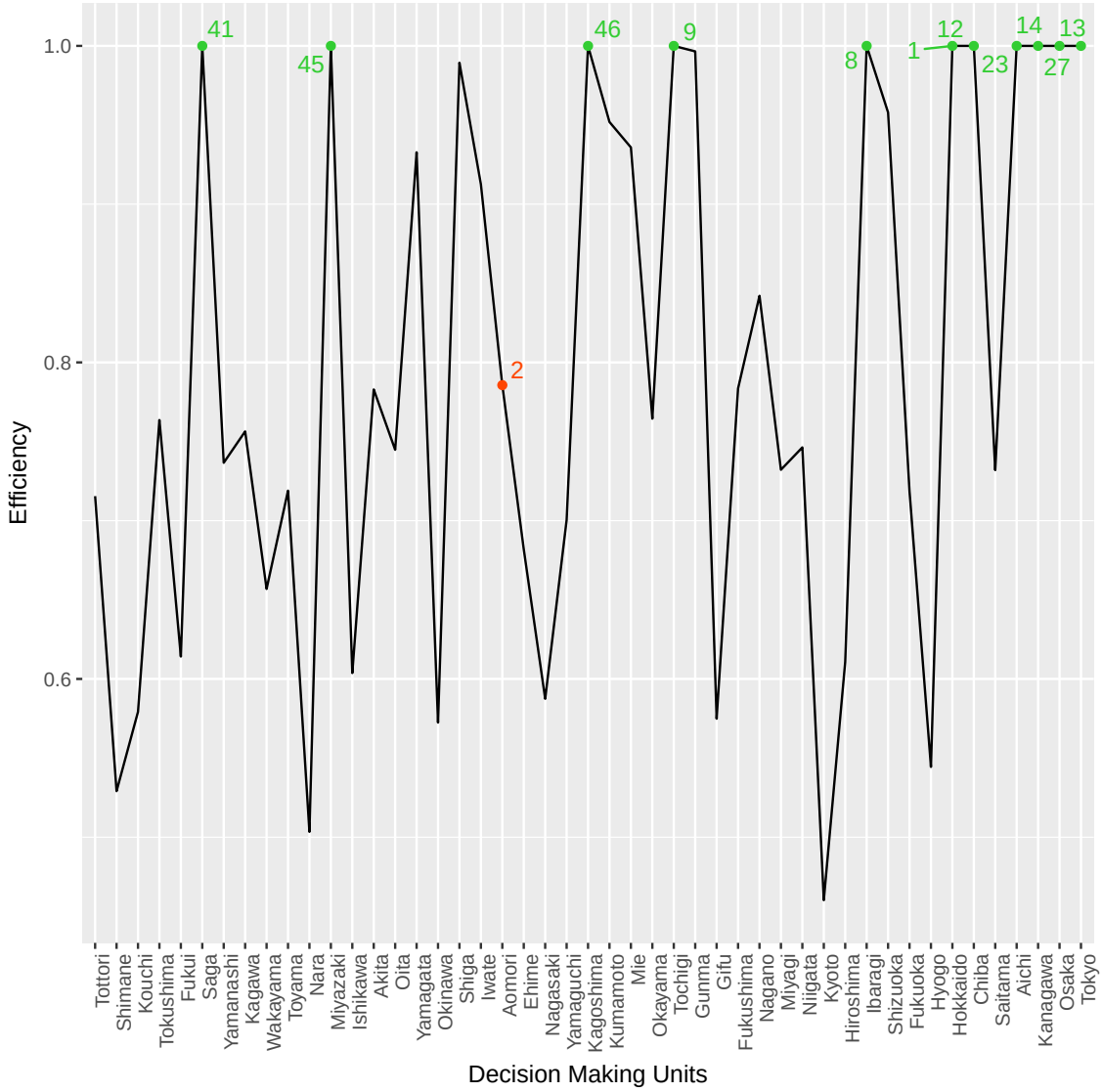

Figure 1.2: Efficiency of DMUs sorted by Population (input variable) (El-Mahgary and Lahdelma 1995). While efficient DMUs are seen all over the range of population, their frequency is higher in the higher end of the population range.

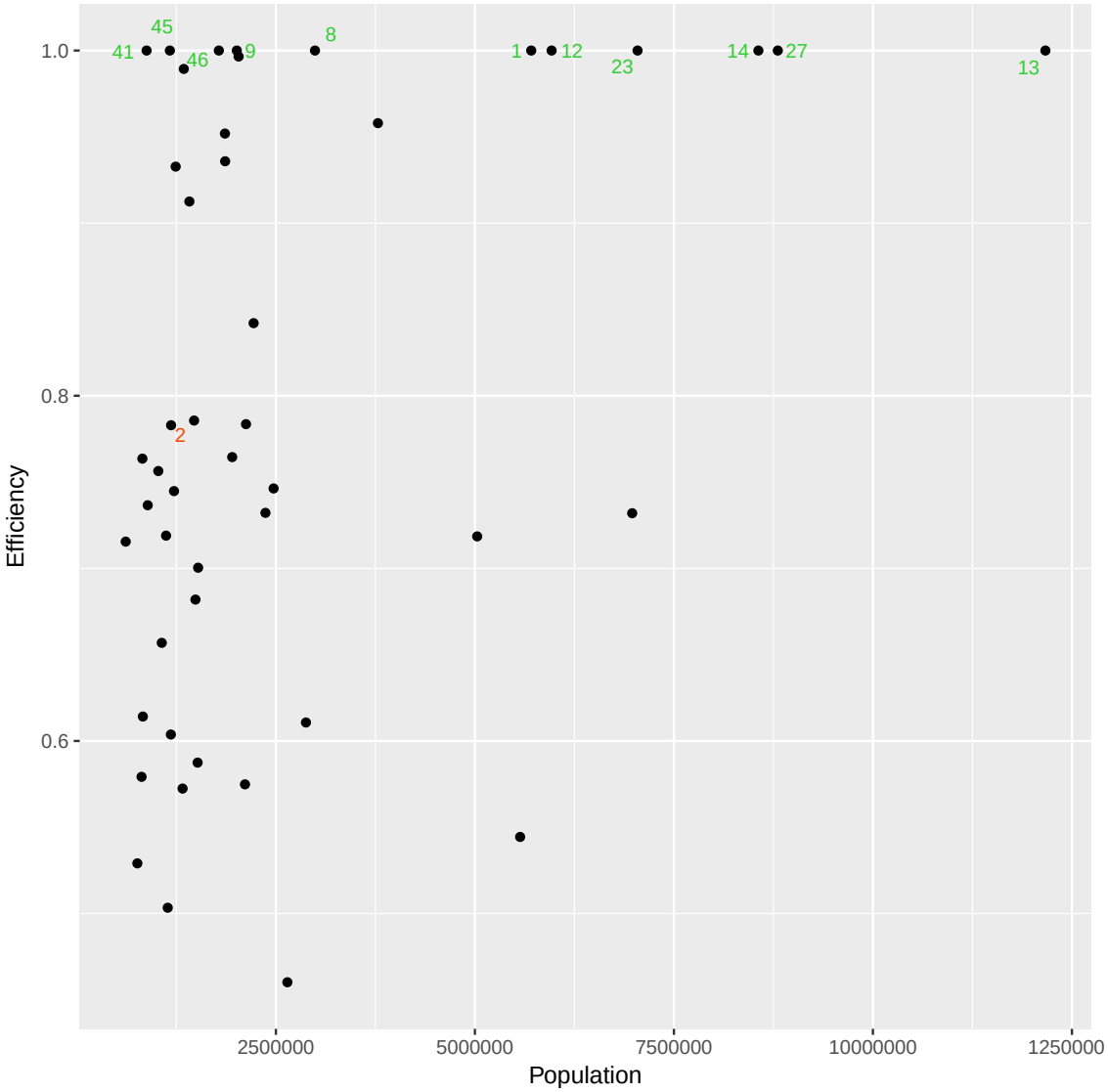

Figure 1.3: Scatterplot of Efficiency vs. Population shows that the frequency of efficient units is higher in the units with more than 5milions population (El-Mahgary and Lahdelma 1995).

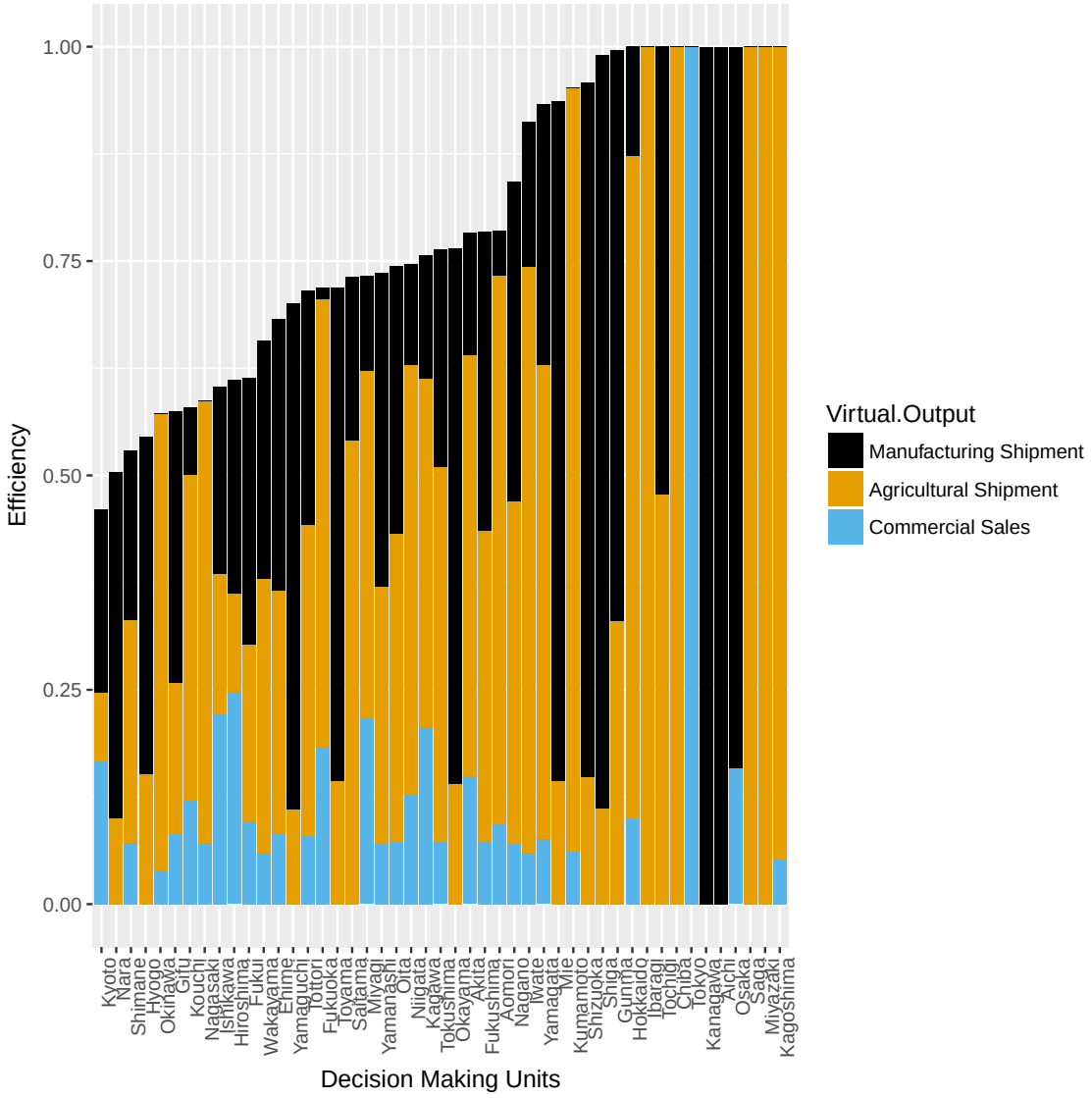

Figure 1.4: Segmented Virtual Outputs of each unit (El-Mahgary and Lahdelma 1995). Interesting points can be seen such as the case of units 4 (Hyogo), and unit5 (Okinawa) each of which has given zero weight to one of the outputs. Moreover, for efficient units, because of availability of alternate optima, this graph is not reliable.)

relied on two outputs, and disregard the third output. Higher number of DMUs or outputs make such plot difficult to read.

2.3.3 Cross-Efficiency Scatter-plot, and Bar-plot

Talluri et al. (2000) propose two DEA-visialization graphs as a part of their special-purpose DMU selection framework, which is a combination of aggressive cross-efficiency method with Friedman non-parametric statistical test. The first graph is a scatter-plot, such that each point represents a DMU with coordinates of the simple-efficiency score of Charnes et al. (1978) model and average cross-efficiency of Doyle and Green (1994) aggressive model. The second graph is composed of side-by-side box plots in a way that each box plot depicts the variation of cross-efficiency scores of a given unit in addition to highlight simple-efficiency. In other words, the box-plots are visualization of cross-efficiency matrix' columns. The suggested graphs of Talluri et al. (2000) can be considered as cross-efficiecy visualization.

The scatter-plot presented in Figure 1.5 is structurally the same as the Figure 1.3, thus there is no need to go through it deeply. The box-plot of Figure 1.6 is for representation of distribution of cross-efficiency scores of each DMU. Not only the spread by centrality measures of median, first and third quartiles are presented in each box-plot, but also the outlier scores are depicted. Nevertheless, this plot would become cumbersome by increase of the number of DMUs. From Figure 1.5 it can be seen that efficient units 14 and 27 have low average cross-efficiency. Thus, they are the top candidates of "maverick units", according to Doyle and Green (1994).

2.3.4 PCA Bi-Plot

The suggested method of Serrano-Cinca et al. (2005) is mainly an advisory tool to model selection in DEA. The visualization is a by product of this method, however being so does not reduce the importance and usefulness of this visualization. The method is devised in order to evaluate the efficiency of DMUs under all possible models. By models, the authors mean a specific combination of input(s) and output(s). Therefore, as an example, all possible models for a dataset of 2 inputs and 3 outputs include all one-input one-output models, one-input two-outputs models, one-input three-outputs models, two-inputs one-output models, two-inputs two-outputs models, and the-two inputs three-output model. The efficiency of each DMU under each combination is calculated as the element of the matrix to be visualized. Thus, the data object is a matrix of DMU vs. efficiency model, such that each row is the efficiency profile of the corresponding DMU and each $i-j$ element is the efficiency score of the $unit-i$ under $model-j$. Hence, each model profile can show the strong and weak points of units based on their input output levels. The dimensions of the data object are reduced by using Principal Component Analysis(PCA) into two components, and the result is visualized in a 2dimensional map. In the final map, the similar DMUs (according to their models profile) ideally locate close to each other, while the dissimilar DMUs far from each other. Moreover, the models are

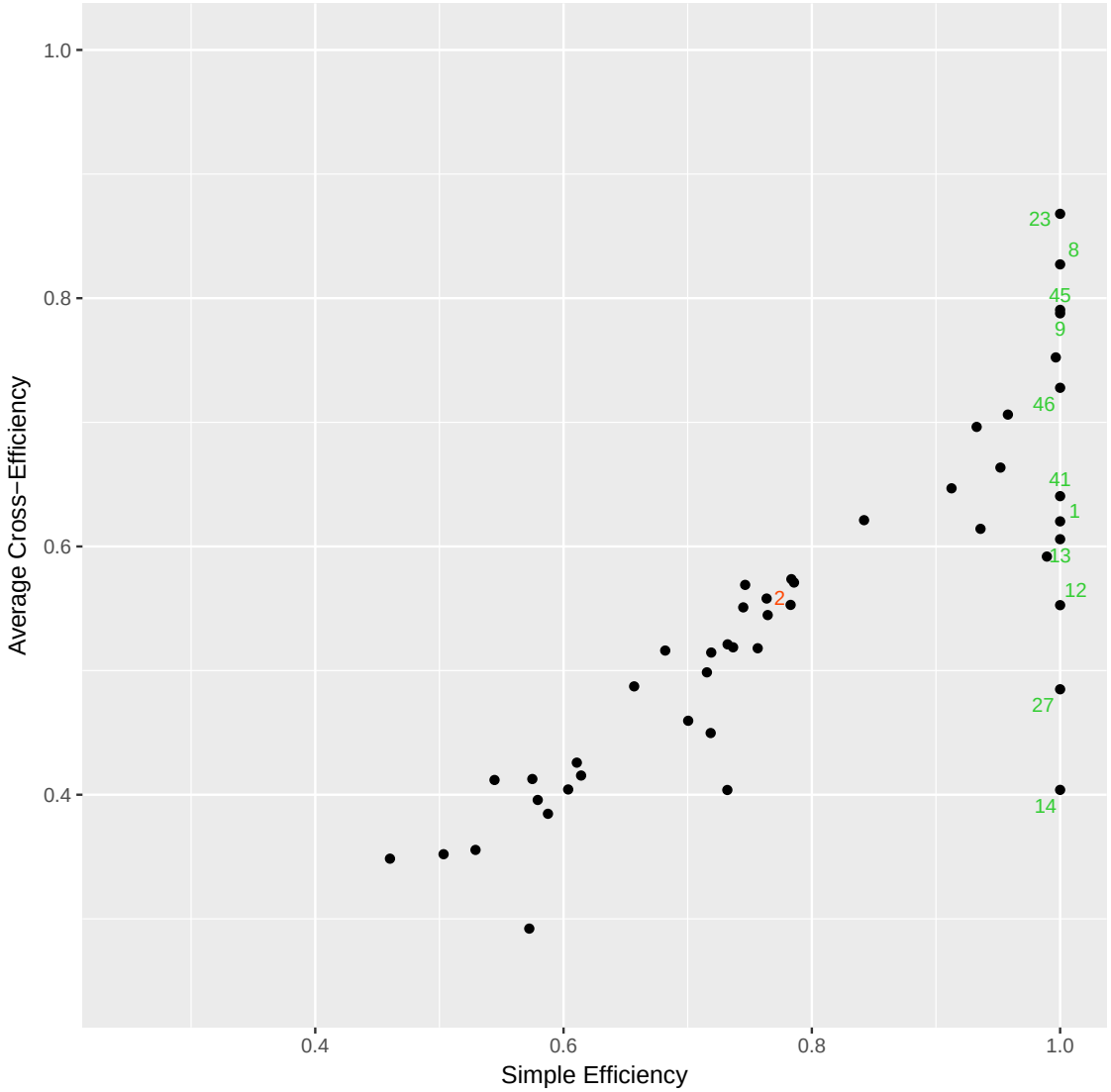

Figure 1.5: Scatterplot of simple efficiency vs. average aggressive cross-efficiency (Talluri et al. 2000). DMU14 has perfect simple efficiency while very low average cross-efficiency. On the other hand, DMU23 has a very high average cross-efficiency and perfect simple efficiency. They can be candidates of maverick unit and outlier unit respectively.

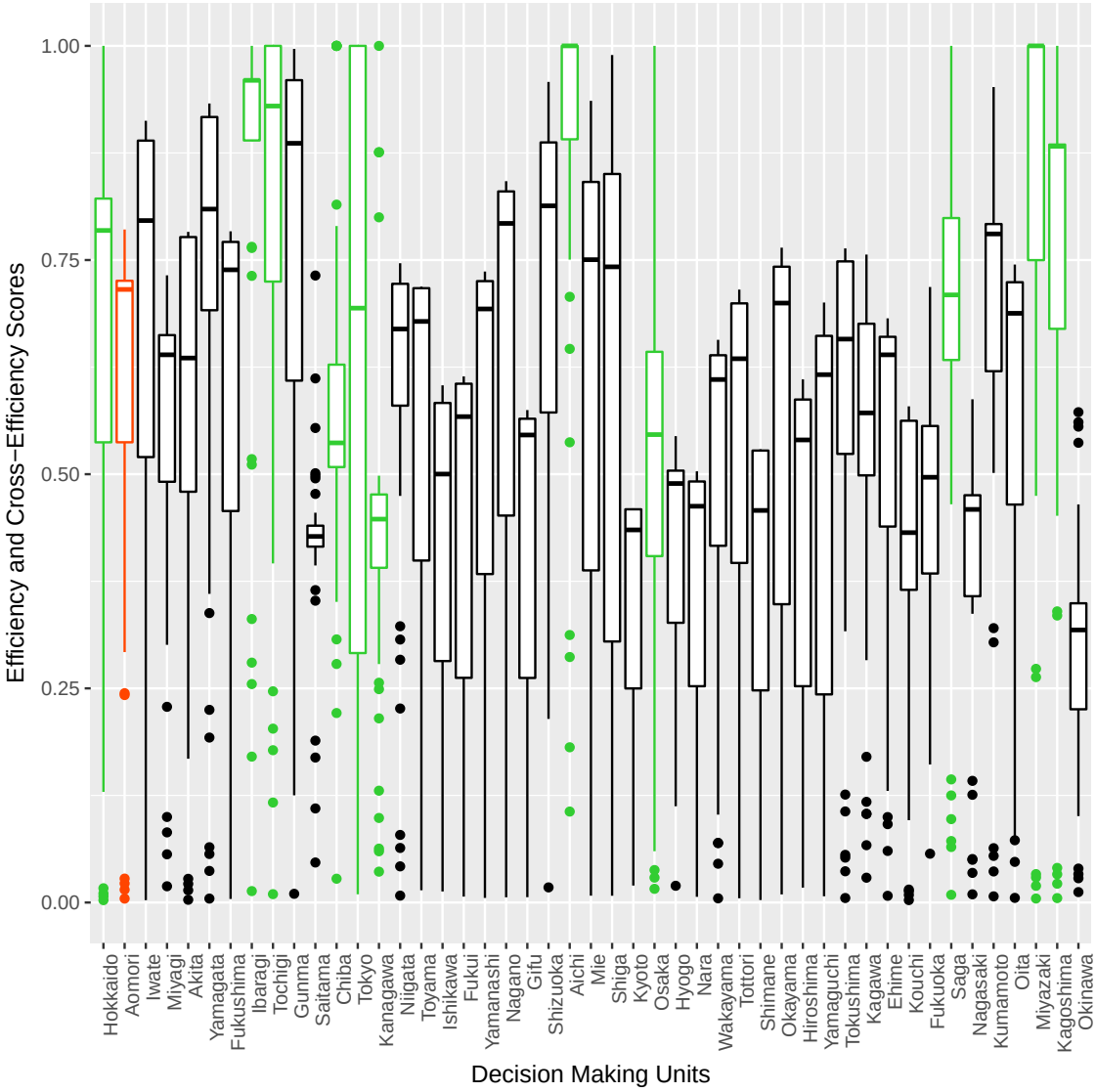

Figure 1.6: (Boxplots of cross-efficiency matrix columns (Talluri et al. 2000). It is interesting that prefectures Aichi (DMU23) and Miyazaki (DMU45) have such long-tail left skewed cross-efficiency distribution.

super-imposed to the map as vectors, using a multiple regression idea called property fitting (Kruskal and Wish 1978), in order to show the reason of difference among DMUs. Consequently, one can see not only which DMUs are (dis)similar, but why they are (dis)similar. Such maps are called biplots. (Gabriel 1971; Greenacre 2010).

Figure 1.7 is essentially a scatter-plot with add-on vectors, where vectors are essentially representation of their tip point. The visual marks are points, and depict DMUs. However, this plot is fundamentally different from previous ones due to the fact that the variables represented as visual channels, i.e. horizontal and vertical coordinations, are not single input or output variables, but the first two principal components(PCs) of PCA. The main goal of this plot is visualizing (dis)similarity of units, and this (dis)similarity is reflected in the distances between each pair of points, i.e. visual marks. As a result, an outlier unit can be detected when a point is far from the crowd. The vectors are representation of original variables, either the original variables are inputs or outputs, or efficiency scores of various models as suggested by (Serrano-Cinca et al, 2005). The main caveat of bi-plots are possible imprecision when the top two PCs do retrieve enough variance of the original dataset, and lack of alternative approach when such imprecision happens. Moreover, by increase of the number of variables, the vectors would become a mess, and difficult to comprehend, and points would be difficult to discern.

Figure 1.7 shows not only the units which are (dis)similar, but the reason of their (dis)similarity. Hence, the units 14,23,27 and 13 emerge one cluster which is different from the cluster composed of units 1, 41, 45, 46. According to Serrano-Cinca et al. (2005), the vectors can show the underlying difference of the clusters. For instance, the units that are located at 12 o'clock on the map are targeted by the vectors of models with output2 (Agriculture Shipment) as the sole output. It means that the strong point of those units is the output 2, i.e. agriculture shipment. In contrast, the cluster of units 14, 27, 23, and 13 are referred by vectors of models with emphasize on output1 and output 3. It is worth to mention that construing such bi-plot maps rapidly become more difficult by increase in the number of vectors, i.e. models. In order to overcome this caveat, it is tried to cluster the vectors and meaningfully label the clusters.(Serrano-Cinca et al. 2005)

2.3.5 Sammons Mapping

Porembski et al. (2005a) use a non-linear transformation of distances called"non-linear mapping" or NLM (Sammon 1969), which is a closely related to mutli-dimensional scaling(MDS) [1] techniques to according to Sammon (1969).[2]

[1]MDS is generally a class of non-linear dimension reduction techniques that try to represent the original inter-object distances in low dimensional space as precise as possible. While the MDS techniques are based on (dis)similarity between objects, the NLM is based on the distance between objects, or relative position of attribute vectors, in the original space. In contrast to PCA, in both MDS and NLM, the relation between dimensions in the destination space, and original dimensions is non-linear and thus less interpretable. Such loss can be seen as compensation of the non-linear dimension reduction. In order to read profoundly about MDS, interested readers are referred to Borg and Groenen (2005)

[2]The NLM is also famous as Sammon's mapping in the literature.

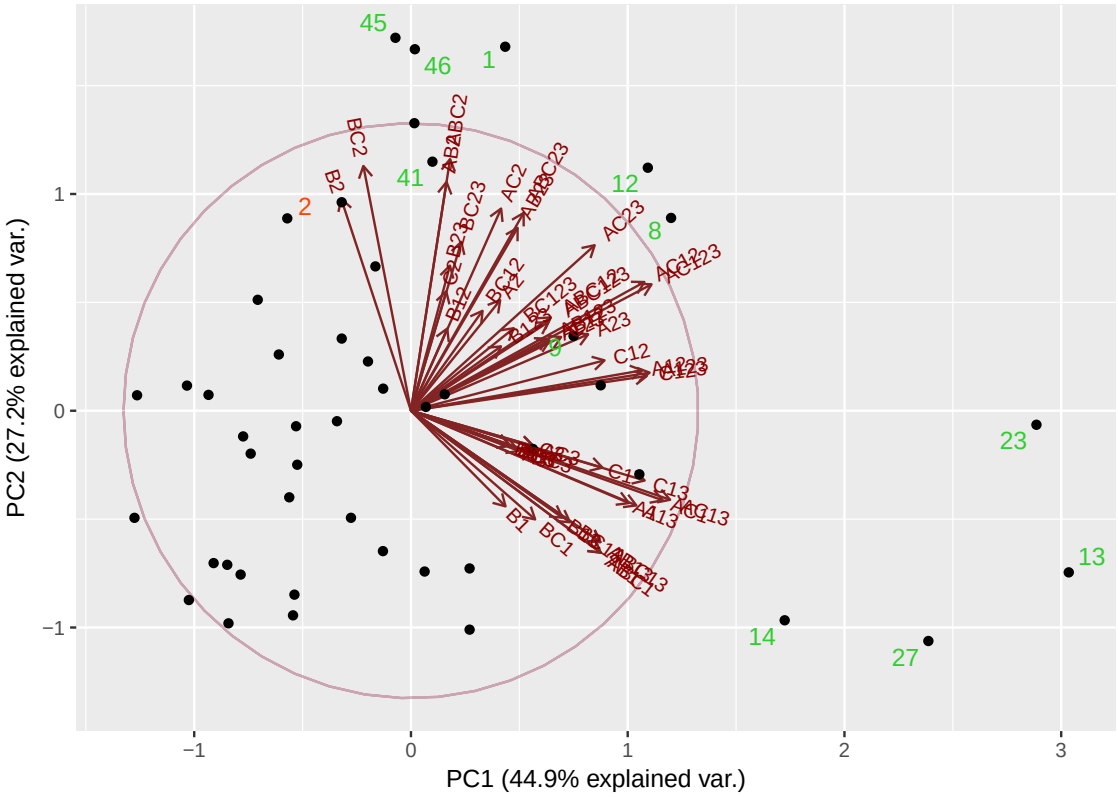

Figure 1.7: Similarity and dissimilarity of the DMUs based on their efficiency profiles are reflected in the relative position of the units on the map (Serrano-Cinca et al. 2005). Units 14 and 27 are similar, while units 45 and 14 are very different. The vectors show the models on which the units have high efficiency scores. For instance unit45 seems highly efficient under BC2 and ABC2 models, i.e. models with second and third inputs and the second output as well as all inputs and the second output, respectively. The total variance explained by the two components is a measure of the accuracy of the map.

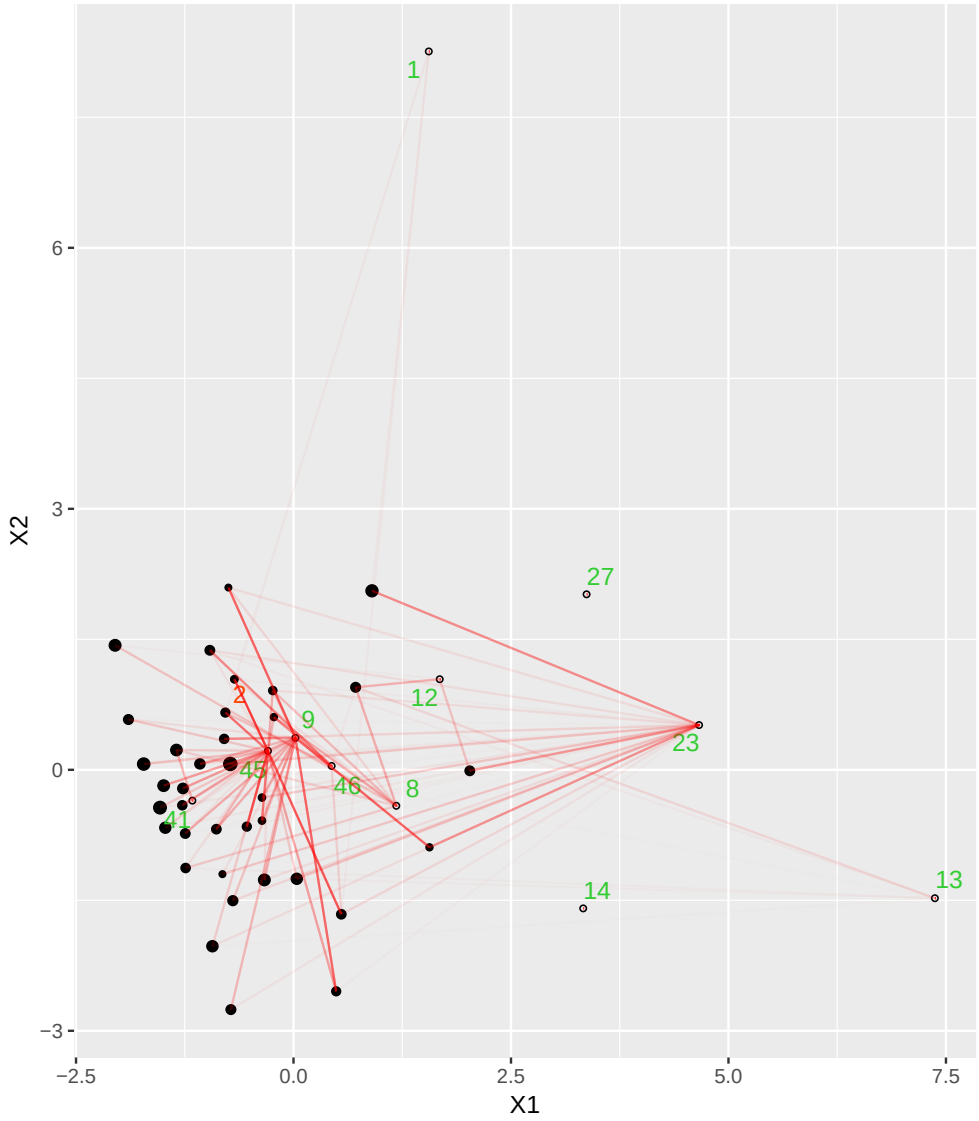

Figure 1.8: Similarity and dissimilarity of the DMUs based on their inputs and outputs profile (Porembski et al. 2005a). The inefficient units are shown in black dots, and their size is a direct function of their inefficiency. The red lines and their transparency are depiction of the relation between each inefficient unit and its references and the intensity of the relation.

The data object, fed in the NLM method, is the normalized input and output factors. The data object's dimensionality, which is originally equal to the total number of inputs and outputs, is reduced into two dimensions via Sammon's non-linear mapping, and new coordinates are mapped subsequently. The visual outcome of the method, is a two-dimensional map including the units as points such that the relative distance of the points are determined by their relative distance of their corresponding input output factors' profiles. The precision of the map is measured by the NLM stress value, which is 0.017 in the case of Figure 1.8. The lower the stress, the higher the precision of the map.

Since the map is based only on input and output variables without inclusion of any DEA model, it can be considered as a pre-DEA visualization, or sort of exploratory data analysis. However, the visualization method of Porembski et al. (2005a) does not stop at this level, and some DEA related information is added to the original map in subsequent steps. The authors suggest to enriched this visual configuration with DEA results, such as the reference sets of each inefficient unit, the influence of reference units on the inefficient units, as well as the magnitude of inefficiency of the units ,i.e. dual multipliers or the famous λ values. These separate sets of information are added to the Sammon's 2d graph in order to enrich it. Hence, the final map seems like a web, where the nodes are units and the edges are relations between inefficient units and their reference sets.

Figure 1.8 is structurally a scatter-plot, similar to Figure 1.7. The visual mark is point, representative of DMUs. The efficiency scores of units are encoded to point sizes, and corresponding lambda values of each pair of units are depicted as lines between the corresponding points. The magnitude of lambda is encoded to thickness of the lines. In contrast to Figure 1.7, this figure is produced by using Sammon-mapping algorithm, of MDS family, rather than PCA. PCA is more familiar method for reducing dimensionality, however it has some strong assumptions such as possibility of reducing the original space to a two orthogonal coordinates, while preserving the original topology. This linearity assumption is relaxed in MDS family, in the price of losing the meaning of coordinates.

In Figure 1.8, the location of each unit is determined by simultaneous consideration of all inputs and outputs. Thus, it is a very useful tool for detection of (dis)similarities between units, and finding outliers. The lambda values would help to detect target units and idols from the DEA model perspective. The caveat is reduction of readability of the plot by increasing the number of units. This problem can be alleviated by fading lines with lower lambda values than an arbitrary threshold. Moreover, MDS algorithms suffer from local optimums, which may cause the result to be imprecise. There is no alternative approach to circumvent such situation in MDS models. Additionally, the algorithm is not designed to cope with large datasets over several thousand observations.

The map of Figure 1.8 reveals several interesting points. Efficient units 27 and 14 are not referenced by any inefficient units, in contrast to units 9 and 45 are heavily referred by inefficient units. Thus, both of these two groups require further investigation since the first group seems like mavericks, and the second group seem like outliers.

2.3.6 MDS Co-plot

Adler and Raveh (2008) suggest a DEA visualization, called co-plot, based on non-metric MDS as the dimensionality reduction technique and a set of superimposed vectors. From the dimensionality-reduction aspect, the method is similar to Porembski et al. (2005a), except for the usage of non-metric MDS rather than metric MDS, and from the superimposed vectors idea, it is similar to Serrano-Cinca et al. (2005) or any other biplot. However, the main differentiating characteristic of the method lies in its data object which is visualized.

Co-plot, presented in Figure 1.9, is basically the same as bi-plot, but instead of PCA, here an MDS algorithm is used. So the units are points, their binary efficiency/inefficiency is encoded to their shape, either the point is filled or hollow. The pros and cons are similar to the previous method, however here the ratios of outputs to inputs are used as the original variables, and the interpretation of these ratios is less straight-forward than the interpretation of sole inputs or outputs.

Instead of input and output factors, Adler and Raveh (2008) benefit from ratio factors as their visualization data-object, such that each DMU has a profile of ratios of every output variable divided by every input variable. Since the DEA formulation is weighted sum of outputs over weighted sum of inputs, the authors use output over input ratios to bridge the visualization method to DEA efficiency formulation. Thus, for a problem with n DMUs with k inputs and l outputs, each DMU has a ratio profile including $l * k$ ratios and the final matrix is composed of the ratio-profiles of the n units. In order to visualize such matrix, Adler and Raveh (2008) applies Smallest Space analysis (SSA) of Guttman (1968), which can be categorized as a non-metric technique of multidimensional scaling (MDS). Therefore, in the final visual configuration, similar units locate closer to each other, and dissimilar units locate farther from each other in such way that the rank-order of the (dis)similarities are preserved as much as possible. Beforehand, the (dis)similarity of units is calculated based on (dis)similarity of their ratio profiles. Having generated a MDS map in 2dimensions, vectors of ratios are super-imposed on the map using Co-plot technique (Raveh 2000). The Co-plot technique is a biplot variation, similar to the property fitting technique used by Kruskal and Wish (1978). In conclusion, not only the (dis)similar unit are located accordingly on the map, but the source of (dis)similarity can be investigated using these vectors.

The proposed technique of Adler and Raveh (2008) from the aspect of dimension-reduction is similar to Porembski et al. (2005a) and from the aspect of super-imposed vectors is similar to Serrano-Cinca et al. (2005), however the use of ratio factors differentiates the Co-plot technique from the others. From this point of view that Porembski et al. (2005a) and Adler and Raveh (2008) use input and output variables (or ratio variables) as the visualization data-object, both methods can be seen as pre-DEA visualization. Finally, the usage of linear model vectors in order to explain a non-linear space is not an effective approach to every datasets.

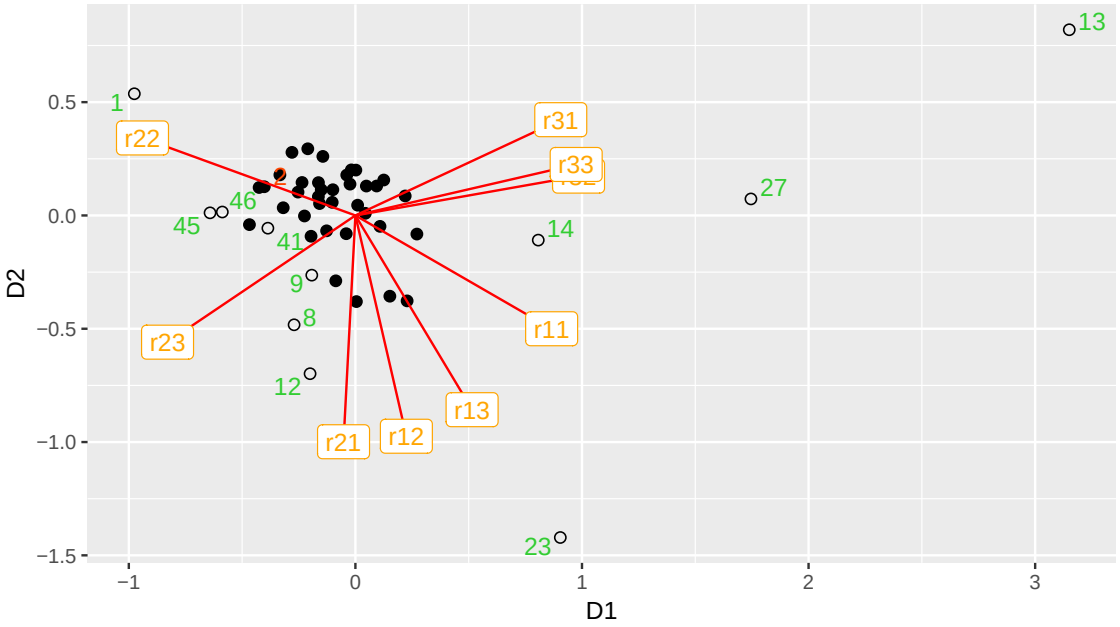

Figure 1.9: Similarity and dissimilarity of the DMUs based on their output to input ratio profiles (Adler and Raveh 2008). The inefficient units are shown in black dots, and efficient units are drawn as circles. It is important to note that efficient units tend to locate in the outer ring, since the efficiency increases in the direction of vectors, in general. Each vector represents the direction in which the corresponding ratio increases. The vectors' direction are set based on the highest correlation of the projection of points on them, and the original ratio values.

2.3.7 Frontier Scatter-plot

Figure 1.10 is structurally a scatter-plot in which the optimal frontier is presented as $x = y$ line. The units are presented as point marks, and their location is determined by aggregated inputs and outputs according to the suggested formulation of Costa et al. (2016). This map is useful for finding (dis)similarities between units, and seeing (in)efficiency based on the distance to the frontier line. The caveat is generally related to the data that is used for visualization, more specifically the aggregation formulations of Costa et al. (2016). The aggregation of all inputs, or outputs, into one single value makes it difficult to figure-out the roots of differences between two DMUs that stand far apart in the map. Besides, large number of DMUs would make them less discernible on the map.

The method suggested by Costa et al. (2016) is the generalized version of the method of Appa et al. (2010). This generalized version visualizes DMUs on a coordinate system of modified sum of virtual outputs and modified sum of virtual inputs for each DMU.

The modified virtual factors are derived from modified CCR or BCC models. Since there are only two variables for each unit, the modified virtual input and virtual output, a simple scatter-plot in 2dimensional space suffices for visualization. In the final map, the x-axis is the modified virtual inputs and y-axis is the modified virtual outputs, and the efficiency frontier is on $y = x$ line of the Cartesian coordinate space. Moreover, on the final map, all the efficient units are located on the $y = x$ line, while all the units on the right-hand side of this line are inefficient and their efficiency scores can be calculated according to the distance of the points to $y = x$ line. Nevertheless, the position of the efficient units are not unique, in other words the $y = x$ is only about the efficient units rather than their relation to each other or their size or how they have achieved the efficiency. Moreover, Costa et al. (2016) expand the basic map of Figure 1.10, by calculation of the range over which each efficient unit remains efficient, i.e. the alternate optima range.

2.3.8 Self-Organizing Map

Self-organizing maps (SOMs) belong to artificial neural network (ANN) field, and they have a totally different approach to dimensionality reduction and visualization of observations. In all previous DEA InfoVis techniques, observations were presented as visual marks, such as points, and some variables as visual channels such as point colors. Here in SOM, instead of visualization of observations, the algorithm visualizes the observations' feature space, in a 2d grid format. The original feature space of the observations can be presented perfectly in n-dimensional form, n equal to the number of variables defining each observation. However, when the number of dimensions is greater than three, visualization become difficult, if not impossible. SOM is a solution to this problem, by representation of the feature space in a two dimensional format, in which each node is a part of space which may or may not include any observation. Thus, contrary to previous methods, SOM has no visual mark for observations, but it has visual marks, i.e. nodes, for sections of the original

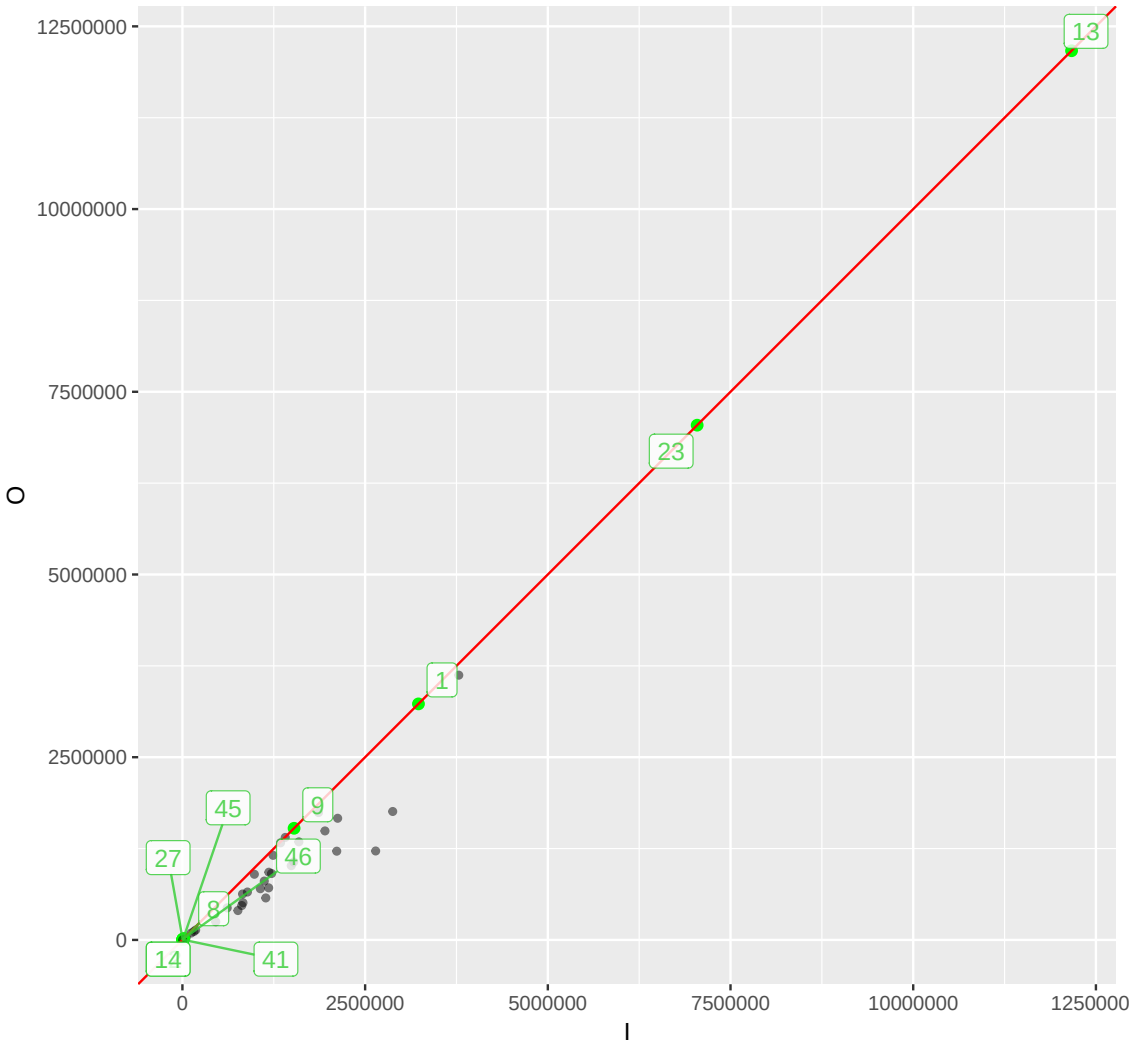

Figure 1.10: Costa et al. (2016) depicts the efficient frontier and the relevant position of the units to it. The efficient units, differentiated by green colour, reside on the frontier, shown by the red line, while inefficient units are located on the right hand side of the line. The vertical or horizontal distance of each inefficient unit to the frontier line, shows respectively the amount of total output or total input which unit needs to increase or decrease in order to achieve the efficiency.

feature space. At the end, a variable(feature) is encoded to these nodes using color visual channel, in order to show how that variable is distributed over the space. As we can see, the labels in each node are the observations located on that node. In Figure 1.11, CRS cross-efficiency score is encoded on the nodes, and Figure 1.12 is composed of same map with inputs and outputs encodings.

The positive side of SOM is its scalability. It can cope with thousands of observations and variables in a small computational cost, something that neither PCA nor MDS are designed for. When the number of observations increases, the two latter methods produce maps with overcrowded visual marks, however the SOM map would be intact, since it does not deal with observations, but the feature space. Moreover, in the cases where the first two principal components or dimensions do not sufficiently capture the variance of the original dataset, and therefore do not provide a precise visualization, SOM can easily cope with the problem by changing the size of the grid. The negative side of SOM is a compensation of the positive side, as the observations are not as tangible as previous methods, and the algorithm is not very familiar to users without ANN background.

To put it differently, SOM, a special artificial neural network method suggested by Kohonen (1982b) and Kohonen (1982a), is a non-linear topology-preserving projection method of representation of high-dimensional data in two-dimensional space, such that similar units reside close to each other, and dissimilar units locate far from each other. In contrast to MDS, SOM segments the original space by a finite number of models, i.e. nodes, which compose the final map in an orderly fashion, and each node may include one or some of the DMUs or remain empty. For a thorough introduction to SOM, readers are referred to Kohonen (2001).

SOM in DEA has been used mainly in order to identify homogeneous clusters of DMUs, regardless of any visualization concern. Nevertheless, there are some researches (Churilov and Flitman 2006; Soares de Mello et al. 2012; Carboni and Russu 2015, e.g.) which have relative emphasis on visual aspects of SOM-DEA.

Churilov and Flitman (2006) benefit SOM in order to identify homogeneous clusters of DMUs based on units' inputs, disregarding output variables. Soares de Mello et al. (2012) suggest identification of homogenous DMU clusters through application of SOM on normalized cross-evaluation, i.e. cross-efficiency profiles of the DMUs. Finally, Carboni and Russu (2015) use SOM to visualize Malmquist-DEA problem under VRS assumption.

Figure 1.11 is the visualization of the Japanese Prefectures dataset based on their inputs and outputs profile. Similar to Churilov and Flitman (2006) , the efficiency of the DMUs under VRS assumption have been mapped on the network. The grey nodes are the nodes in which no DMU is located. By reducing the size of the map, it is possible to fill the empty nodes, however in such case the precision of the map would decrease as well. In such SOM map, when a node has no filled neighbour node, probably it is different from the rest of the nodes according to its DMUs. For instance, DMU1 in the Figure 1.11 has no neighbour, and it seems that this DMU is different from the rest.

Figure 1.12 depicts the features space, i.e. inputs and outputs space, on the map. While the features are normalized in order to remove the effect of variables'

magnitude, and treat the variables with equal importance, the set of maps can reveal not only the non-linear distribution of the variables through the space, but also the relation among them. For instance, it is revealed that the upper right corner of the map is the location of DMU(s) with very high "area", medium "population", medium "household income", low "manufacturing and commercial shipments" and very high "agricultural shipment". Or it is possible to see that high "population" and high "manufacturing shipment" are correlated to some extent.

Since the notion of similarity is based on Euclidean distance in the SOM maps of this survey as well as the maps produced by Carboni and Russu (2015), the size of the DMUs heavily affect the measure of similarity. Hence, in Figure 1.11, some neighbour units, i.e. similar units, are very different in efficiency scores, e.g. units 10 and 28. Being so perhaps highlight the necessity of having a proximity measure specifically devised for measuring (dis)similarity of DMUs in DEA.

2.3.9 The peripheral methods

The second group of the DEA visualization methods is called peripheral, and they are explained verbally, without any visualization. It is decided to do so because either the DEA models, based on which the these methods are devised, are not common DEA models so the outcome of the method is not appealing for a wide range of DEA practitioners, or because of the fact that the method is still not published as a journal paper, or due to this reason that a more comprehensive version of the method is presented in the previous section. Regardless of these reasons, these methods are reviewed in order to keep this survey comprehensive.

Belton and Vickers (1993) is one of the first efforts in DEA visualization. Using inputs and outputs factors as the data visualization entity, the authors suggest a hierarchical multiple criteria value function, such that the function aggregates inputs and outputs into two scores, called aggregate measure of inputs and aggregate measure of outputs, respectively. Therefore, for each unit, the procedure generates a pair of co-ordinations, aggregate input and output scores, that can be plotted in a 2dimensional space. The aggregation is done through assigning pre-defined weights to each input and each output. The weight-sets can be calculated based on some specific criteria in this method, and the stability of the result can be evaluated through sensitivity analysis the weights. The final map depicts the position of efficient units, which emerge the frontier, as well as inefficient units. Thus the suggested method necessitates any researcher to evaluate the problem through this revised DEA model. Doing so means that the method cannot work with other DEA models such as CCR or BCC, specifically because of numerous alternate optima that any efficient unit has. Moreover, assessment of any single DMU in a scatter-plot of DMUs would become cumbersome and even impossible in problems with large number of units. This point has been mentioned in Belton and Vickers (1993).

Hackman et al. (1994) suggests a simplex-like algorithm in order to traverse the boundary of the production possibility set. The algorithm has two phases, and from its second phase, a pair of coordinates for each DMU is generated that can

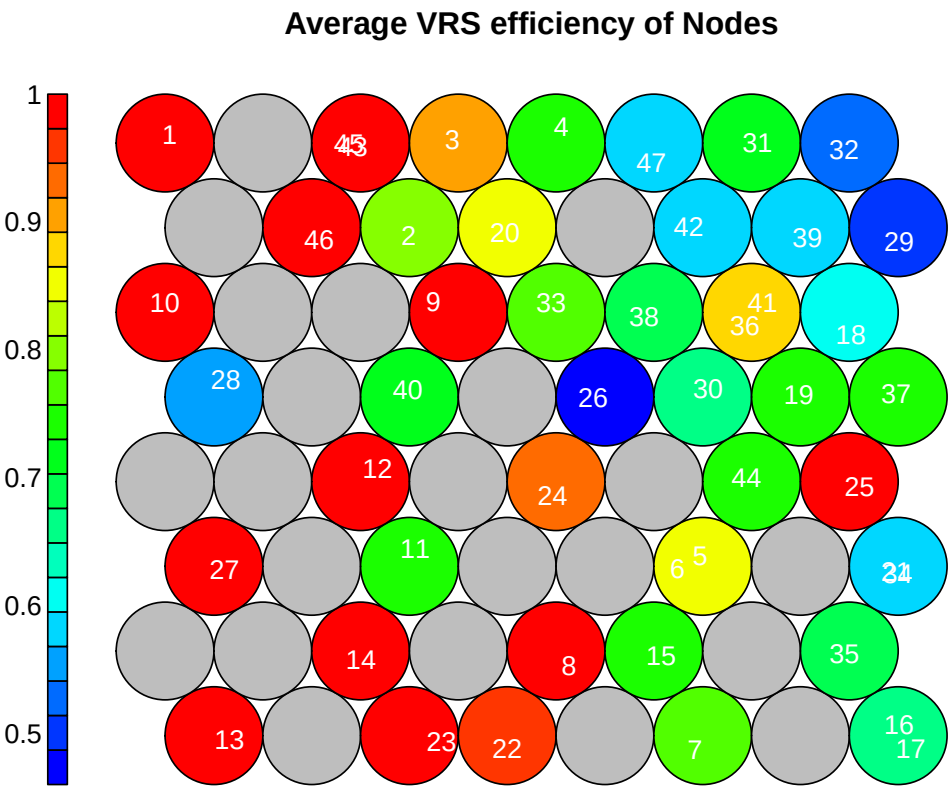

Figure 1.11: The average efficiency of each node in the self-organizing map (Carboni and Russu 2015). Since there is no standardization of DMUs, the magnitude of the DMUs heavily affected the position of the units on the map. Thus, the map has become inaccurate in some parts as DMUs which are neighbours, i.e. similar DMUs, have very different efficiency score, while it is expected that similar DMUs have similar efficiency scores.

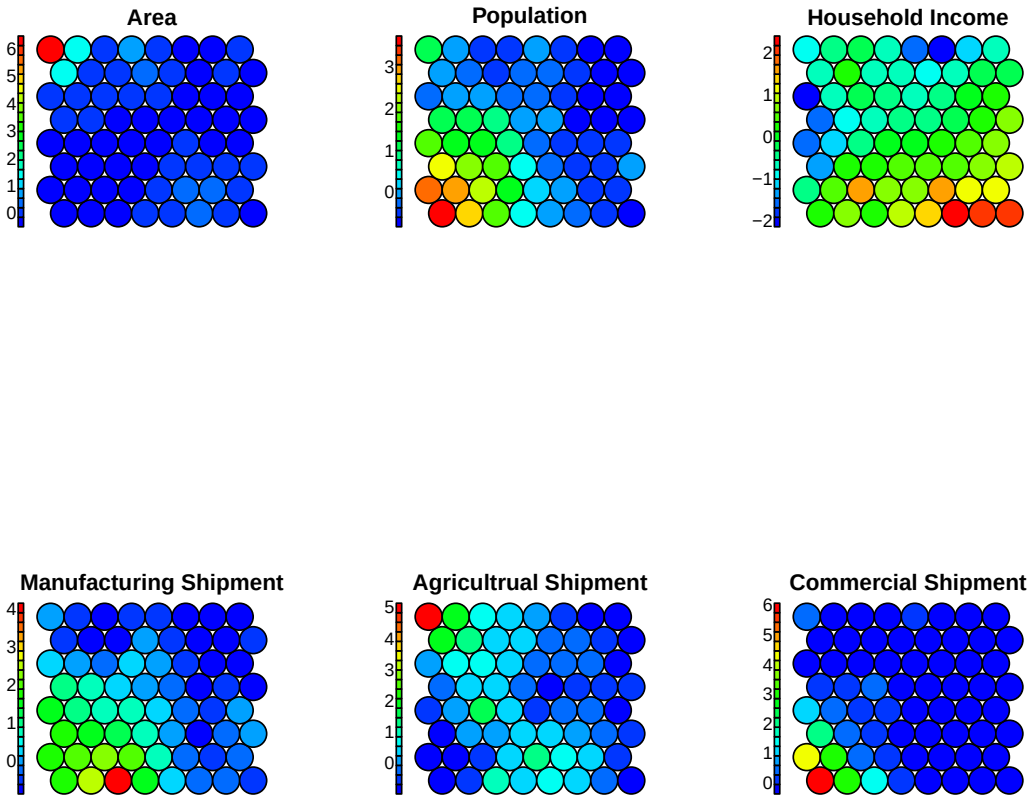

Figure 1.12: Carboni and Russu (2015) suggest SOM in order to visualize panel DEA, and identify homogeneous groups of DMUs. Each heatmap map shows the distribution of the corresponding normalized variable throughout the nodes. Such set of maps, in case not having too many variables, can reveal the correlation between variables.

be visualized in 2dimensional space, and can determine the technical and scale efficiency scores of the units. The input and output factors of DMUs are used as the visualization data-object in this method.

Aoki et al. (2007) suggest a DEA visualization method which is based on combination of BCC sensitivity analysis (a modified BCC-DEA model) and fuzzy correspondence analysis. The data entity that the method visualizes is a table of dual multiplier profiles, i.e. λ values, of each DMU. The dual multipliers, are relative positions between the corresponding unit and efficiency frontier. In other words, each unit has a profile that is composed of these λ values regarding each dominant unit in the dataset. The dominant units are the ones referred to by other units in the efficiency calculation process. Therefore, the visualization data object, for a problem with n DMUs ,of which m DMUs are dominant, has n rows and m<n columns, and each element of this matrix is corresponding λ value. The dimension-reduction technique, used by Aoki et al. (2007), is fuzzy correspondence analysis and the final map positions the units with similar lambda profiles close to each other and dissimilar units far from each other.

Førsund, Kittelsen, et al. (2009) suggest "contemporary methods of visualization of multi-input multi-output frontiers" using parametric optimization methods of Krivonozhko et al. (2004), and input output levels of production units. The visual outcomes include iso-quants and graphical development of scale-elasticity along ray frontiers, which have been defined based on ray production functions. (Førsund, Hjalmarsson, et al. 2007)

Appa et al. (2010) suggest a visualization method for problems with one input and several outputs. For each DMU, the value of the sole input and virtual outputs are computed in order to map in a 2d space. The aggregated value for virtual outputs is achieved through the modified CCR model of Thanassoulis (2000a) and Thanassoulis (2000b). Consequently, for each DMU there are only two variables, the visualization method is nothing but a scatter-plot, in which the coordinations are the input, and the sum of outputs, orderly. In the final map, efficient units are located on $y = x$ line, and inefficient units are positioned on the right-hand side of this line.

In a closely-related study to Aoki et al. (2007), Honda et al. (2010) suggest a DEA visualization method. The data-entity that Honda et al. (2010) visualize is a n by p matrix, where n is the number of DMUs in the problem and p is the number of DMUs which they are referred in super-DEA algorithm of Zhu (2001). In other words, the n by p matrix is consisted of dual multipliers vector of each n units under super-DEA model. A weight vector of generic unit-k is composed of the unit's weights regarding other units, such that if the generic unit-j is referred by the unit-k in the calculation of efficiency analysis, then related dual multiplier is non-negative, otherwise it is zero. From the aspect of using dual multipliers, the method is similar to Aoki et al. (2007), and from the aspect of using super-DEA model, it is different from it. The matrix is visualized into a 2d map by performing fuzzy PCA, set up in a way to "emphasize the mutual relations among efficient DMUs" as well as "to

clarify the relations between inefficient DMUs and their target units" Honda et al. (2010)

Inoue et al. (2011) suggest a DEA visualization method for presentation rather than exploration, i.e. the focus of the method is more inclined towards communication of the results, and not exploration of different aspects of the problem. Hence, it is different from the previously cited studies. The idea is based on suggested hierarchical structure for each unit. According to the suggested method, a unit can be broken down into smaller units such that the input and output variable set of each smaller unit is a proper subset of the original unit's variable set, the intersection of the variables of each two sub-units is the empty set, and the union of the subsets is the original unit's variable set. Each smaller unit is also possible to be broken down into sub-units accordingly. The efficiency of each sub-unit is then evaluated relative to similar sub-units of other DMUs. The criterion of choosing the subsets is subjectively based on the nature of the problem and expert's opinion.

Finally, for each unit and its sub-units, a 2d map can be built such that the vertical axes shows the efficiency score of the unit and each sub-units, and the horizontal axis depicts the difference between the each sub-unit efficiency score and average of efficiency scores of such sub-units over all the DMUs. The final map can convey the information regarding the relative and absolute strong and weak points of each DMU.

And finally, Akçay et al. (2012) suggest two visual maps in order to gain insight to the DEA problems, i.e as exploratory tools. Their first suggestion is mainly a presentation of DMUs in a scatter-plot such that coordinations are two variables (either inputs or outputs) by the choice of the user, and the size and colour of the DMU objects on the map are set based on another DEA variables, such as efficiency scores and a third variable, chosen by the user. Hence, up to four variables can be displayed in a 2d graph without loss of information. Previously, Porembski et al. (2005a) had used similar approach to loading map with extra DEA information. The second map ,suggested by Akçay et al. (2012), is a tile graph which can be divided based on a categorical variable, such as geographical regions of DMUs, into smaller tiles and each tile can visually include the related DMUs. Moreover, the size and colour of the DMUs in each tile can be adjusted by two other variables, chosen by the user. Finally, three to four variables can be shown on the tile map.

A summary of reviewed methods is presented in the next subsection.

2.4 Summary Table of DEA Visualization Methods

Table 1.3 includes a concise set of information about the reviewed methods of main and peripheral groups. It can be used before going into the details of the methods, or for recap of the methods.

Table 1.3: Summary of the reviewed DEA visualization methods

Number	Method	Data object to be visualized	Dimension reduction technique	High-dimensional Visualization?	DEA Model	Possible usage of the visual outcome	Availability in any DEA software
1	Belton and Vickers (1993)	Input and output variables	None	None	Hierarchical aggregation of Inputs and Outputs	Proximity of the units, efficient frontier	VIDEA
2	Desai and Walters (1991), Weber and Desai (1996)	Input and output variables	None	Yes, using parallel coordinates	Model Independent	Path of improvement for inefficient units, the range of variable feasibility based on efficient units, unit comparison	No
3	Hackman et al. (1994)	Input and output variables	None	None	Customized simplex algorithm for CRS and VRS methods	Visualization of efficient frontier	No
4	El-Mahgary and Lahdelma (1995)	Input and output variables, efficiency scores, Virtual inputs and virtual outputs, reference sets and lambda values	None	None	Model Independent	Comparison of inefficient units and their reference units, finding influential reference units, relations of variables and efficiency score	Most of DEA software has one or some of these plots
5	Talluri et al. (2000)	Cross-efficiency matrix	None	None	Cross-evaluation	Finding potential maverick units, finding potential efficient outlier	No
6	Serrano-Cinca et al. (2005)	Model-profile matrix	Principal Component Analysis	Yes, using Bi-plot	Model Independent	Finding potential maverick units, finding potential outliers, Model selection	No
7	Porembski et al. (2005b)	Input and Output Variables, lambda values(dual multipliers), efficiency scores	Sammon's Mapping	Yes	CRS or VRS	Identification of outliers, Clustering of the units	No
8	Aoki et al. (2007)	Dual multiplier profiles of inefficient units	Fuzzy Principal Component Analysis	Yes	Modified BCC model of sensitivity analysis	Identification of outliers, Clustering of the units	No
10	Adler and Raveh (2008)	Ratio variables of outputs over inputs	Smallest Space Analysis	Yes, Using Co-plot	Model Independent	Identification of outliers, Clustering of the units	Co-plot
11	Førsund, Kittelsen, et al. (2009)	Input and output variables	None	None	Parametric Optimization Method		No
12	Appa et al. (2010)	Dataset of sole input and aggregated outputs	None	None	a modified CCR model	Similarity of the units, relative efficiencies	No
13	Honda et al. (2010)	Dual multiplier profiles of inefficient units	Fuzzy Principal Component Analysis	Yes	Super-DEA algorithm	Clarification of relations of inefficient units, clarification of relations between inefficient units and their target units	No
14	Inoue et al. (2011)	Input and output variables	None	None	A Suggested Hierarchical Structure DEA model	Strong and Weak points of each DMU	No
15	Akçay et al. (2012)	Input and output variables	None	None	Model Independent	Relations between variables and efficiency scores	SmartDEA
16	Costa et al. (2016)	Input and Output variables	None	None	Aggregation of virtual standardized input and output weights of CCR or BCC models	Relative inefficiency of the units as the distance to the frontier	No
17	Churilov and Flitman (2006), Carboni and Russu (2015)	Input and Output Variables	Self-Organizing Maps	Yes	Model Independent	Identification of homogeneous clusters of units, Panel data visualization	No

Table 1.4: Main DEA Info-Viz Methods

Figure	Related Article	DEA-Viz Technique	Visual Mark	Visual Channel	Main Goal(s)	Pros	Cons
1	(Desai and Walters ,1991) Weber and Desai ,1996)	Parallel Coordinates	Line	Relative-position of the line Interception with each coordinates	Comparison of DMUs based on their variables(i.e. Inputs and outputs) Detection of the outliers based on Extreme variable values	Overall understanding of variable distributions, And relative-position of each single DMU among all.	The more the number of coordinates, the more the difficulty of comprehension of the map. Therefore it is more useful for the problems with few inputs and outputs No suggested algorithm For ordering the coordinates
2,3,4	(El-Mahgary and Lahdelma, 1995)	Scatter-plot, Bar-plot, Line-plot	Point, Bar, line	Vertical and horizontal location on Cartesian space for points, Color for stacked-barplot	Investigation of association between two variables, Detection of outliers based on two variables, Investigation of virtual outputs	Simple maps to understand	Not scalable to high number of observations, consideration of only two variables may be misleading for outlier detection, multiplicity solutions is reduces The usefullness of stacked-barplot
5,6	(Talluri et al. 2000)	Scatter-plot, Box-plot	Point, Box	Vertical and horizontal location On Cartesian space	Investigation of association between average cross-efficiency and simple efficiency, distribution of cross-efficiency Scores for each unit	Simple maps to understand	Not scalable to high number of units
7	(Serrano-Cinca et al. 2005)	PCA Bi-plot	Point, Vector	Vertical and horizontal location On Cartesian space	Investigation of (dis)similarity between units and outlier detection, Association between features	PCA is well-known and easy to interpret, Holistic visualization approach, Can be used for panel data	Not scalable to high number of observations,No solution if the first to PC don't capture high amount of variance, Difficulties in reading the feature vectors
8	(Porembski et al. 2005)	Sammons Mapping	Point, Line	Vertical and horizontal location on Cartesian space, size of points(~efficiency), thickness of lines(~lambda), Shape of points(~binary efficiency)	Investigation of (dis)similarity between units and outlier detection, analysis of target units, And mavericks	Unique perspective of consideration of Lambda values in visualization to investigate patterns And irregularities, holistic visualization approach	Not scalable to high number of units, over-crowded in presence of all links, the point size is not very discernible visual channel, No alternative if the precision of map is low
9	(Adler and Raveh 2008)	MDS Co-Plot	Point, Vector	Vertical and horizontal location on Cartesian space	Investigation of (dis)similarity between units and outlier detection, Association between output/input ratios	Easy to read, Holistic visualization approach	Not scalable to high number of observations, Difficulties in interpretation of ratios, No alternative if the precision of map is low
10	(Costa et al. ,2016)	Frontier Scatter-plot	Point, Line	horizontal location on Cartesian space	(dis)similarity between units and outlier detection, Distance of units from the frontier	Easy to read	Interpretation of aggregated inputs and outputs And contribution of each variable to themis cumbersome.
11,12	(Churilov and Flitman , 2006) (Soares de Mello et al, 2012) (Carboni and Russu ,2015)	Self-Organizing Map	Node	Horizontal and vertical location in final grid(~location in the feature space), Color of nodes(~average value of observations in the node)	(dis)similarity between units and outlier detection, Clustering, Association between Distribution of variables in feature space	Scalable to high number of observations, Can cope with low precision by increasing the grid size, can be used for panel data, Holistic visualization approach	Not very familiar to scholars without A background, Labeling all individual observations would lower readability

3 The conclusion

Data visualization is a necessary step in order to make sense of the data, since making sense of a data is a necessary step to make sense of the corresponding problem. Consequently, in order to know a DEA problem better and gain better understanding of it, DEA data visualization can be hired by both researchers and practitioners.

The suggested visual step is safely can be considered as a graphical exploratory data analysis of DEA. The goal of such exploration is manifold: from gaining insight into data to more specific goals such as searching for outliers or homogeneous groups. Such visual exploration of the data would supplement the subsequent quantitative analysis, since data visualization is essentially a holistic approach that "can retain the information in the data", while "numerical data analysis procedures ... are essentially data reduction techniques." Cleveland and Cleveland (1985, p. 9) For instance, Figure 1.11 is the visualization of a cross-efficiency matrix including 47 rows and 47 columns, i.e. 2209 digits, all in one plot which includes the observations and their relations. In contrast, the reductionist approach of average cross-efficiency is an attempt to reduce every column of such matrix, i.e. 47 digits, into one single digit. Such reduction has the cost of losing details and relations.

Moreover, "graphs reveal the major features of data, help in the production of ideas for further investigation, and are useful in checking assumptions" Cox and Jones (1981b)

Nevertheless, majority of DEA researches lack such data visualization step. The absence of DEA data visualization is tangible throughout DEA studies, from reference books such as "handbook of data envelopment analysis"Cooper et al. (2011) to stand-alone research papers. Conventionally, the data visualization in DEA is limited to uni-variate or bi-variate graphs, and multivariate visualization is dismissed. However, DEA problems are essentially multivariate, and any practical data visualization needs to deal with multi-variables in order to depict a holistic picture of the problem.

The relative neglect of multivariate data visualization is not due to lack of DEA data visualization methods. The toolbox of DEA data visualization, while far from a perfect toolbox, still includes variety of visualization techniques. However, these techniques have not been promoted and introduced enough in order to encourage researchers and practitioners to use them.

The current paper is a comprehensive survey of such visualization techniques; an incisive introduction of the available tools in DEA visualization toolbox. Considering the extend and depth of this survey, it is safe to claim that there is no similar research available in the literature.

In order to better present the visualization techniques, a dataset of 47 DMUs with 3 inputs and 3 outputs is visualized throughout the paper using a selection of the techniques.

As a result, the possible audience can compare the plots, i.e. the outcomes of the different techniques, to grasp a general idea of the capabilities of DEA visualization techniques. Having a comprehensive toolbox of the available DEA visualization tools is important for researchers who want to craft new tools for DEA visualization.

Additionally, visualization of the data set with different techniques makes it easier to choose proper tools form the visualization toolbox, even though, these techniques are more complementary, since they visually examine a problem from different perspectives.

It is worth to underline that visual analysis should go hand in hand with quantitative DEA models and methods in order to empower researchers and practitioners to deal with increasing complexity of the problems. Furthermore, the current toolbox of the DEA visualization can be ameliorated either through addition of new methods, or improvement of the available ones.

References

Adler, Nicole and Adi Raveh (2008). "Presenting DEA graphically". In: *Omega* 36.5, pp. 715–729.

Akçay, Alp Eren, Gürdal Ertek, and Gülçin Büyüközkan (2012). "Analyzing the solutions of DEA through information visualization and data mining techniques: SmartDEA framework". In: *Expert systems with applications* 39.9, pp. 7763–7775.

Aoki, Shingo, Kotaro Toyozumi, and Hiroshi Tsuji (2007). "Visualizing method for data envelopment analysis". In: *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. IEEE, pp. 474–479.

Appa, Gautam, Carlos A Bana e Costa, Manuel P Chagas, Fernando C Ferreira, and João O Soares (2010). "DEA in X-factor evaluation for the Brazilian Electricity Distribution Industry". In: *London School of Economics*.

Banker, Rajiv D, Abraham Charnes, and William Wager Cooper (1984). "Some models for estimating technical and scale inefficiencies in data envelopment analysis". In: *Management science* 30.9, pp. 1078–1092.

Behrens, John T and Chong-Ho Yu (2003). "Exploratory data analysis". In: *Handbook of psychology*.

Belton, Valerie and Stephen P Vickers (1993). "Demystifying DEA-a visual interactive approach based on multiple criteria analysis". In: *Journal of the Operational research Society*, pp. 883–896.

Borg, Ingwer and Patrick JF Groenen (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Carboni, Oliviero A and Paolo Russu (2015). "Assessing regional wellbeing in Italy: An application of Malmquist–DEA and self-organizing map neural clustering". In: *Social indicators research* 122.3, pp. 677–700.

Charnes, Abraham, William W Cooper, and Edwardo Rhodes (1978). "Measuring the efficiency of decision making units". In: *European journal of operational research* 2.6, pp. 429–444.

Churilov, Leonid and A Flitman (2006). "Towards fair ranking of Olympics achievements: The case of Sydney 2000". In: *Computers & Operations Research* 33.7, pp. 2057–2082.

Cleveland, William S and William S Cleveland (1985). *The elements of graphing data*. Wadsworth Advanced Books and Software Monterey, CA.

Cooper, William W, Lawrence M Seiford, and Joe Zhu (2011). *Handbook on data envelopment analysis*. Vol. 164. Springer Science & Business Media.

Costa, Carlos A Bana e, João Carlos CB Soares de Mello, and Lidia Angulo Meza (2016). "A new approach to the bi-dimensional representation of the DEA efficient frontier with multiple inputs and outputs". In: *European Journal of Operational Research* 255.1, pp. 175–186.

Cox, Nicholas J and Kelvyn Jones (1981a). "Exploratory data analysis". In: *Quantitative geography: A British view*, pp. 135–143.

Cox, Nicholas J and Kelvyn Jones (1981b). "Exploratory data analysis". In: *Quantitative geography: A British view*, pp. 135–143.

Desai, Anand and Lawrence C Walters (1991). "Graphical presentations of data envelopment analyses: management implications from parallel axes representations". In: *Decision Sciences* 22.2, pp. 335–353.

Doyle, John and Rodney Green (1994). "Efficiency and cross-efficiency in DEA: Derivations, meanings and uses". In: *Journal of the operational research society* 45.5, pp. 567–578.

Emrouznejad, A, R Banker, SC Ray, and L Chen (2016). "Recent Applications of Data Envelopment Analysis". In: *Jianghan University, Wuhan, China, ISBN* 978.1.

Førsund, Finn R, Lennart Hjalmarsson, Vladimir E Krivonozhko, and Oleg B Utkin (2007). "Calculation of scale elasticities in DEA models: direct and indirect approaches". In: *Journal of Productivity Analysis* 28.1, pp. 45–56.

Førsund, Finn R, Sverre AC Kittelsen, and Vladimir E Krivonozhko (2009). "Farrell revisited–Visualizing properties of DEA production frontiers". In: *Journal of the Operational Research Society* 60.11, pp. 1535–1545.

Friendly, Michael (2008). "Handbook of data visualization". In: *Handbook of data visualization*. Ed. by Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. Springer. Chap. A brief history of data visualization, pp. 15–56.

Gabriel, Karl Ruben (1971). "The biplot graphic display of matrices with application to principal component analysis". In: *Biometrika*, pp. 453–467.

Greenacre, Michael J (2010). *Biplots in practice*. Fundacion BBVA.

Guttman, Louis (1968). "A general nonmetric technique for finding the smallest coordinate space for a configuration of points". In: *Psychometrika* 33.4, pp. 469–506.

Hackman, Steven T, Ury Passy, and Loren K Platzman (1994). "Explicit representation of the two-dimensional section of a production possibility set". In: *Journal of Productivity Analysis* 5.2, pp. 161–170.

Honda, Katsuhiro, Shingo Aoki, Akira Notsu, and Hidetomo Ichihashi (2010). "Visual Assessment of DEA Efficiencies by Fuzzy PCA with Variable Selection". In: *SCIS & ISIS SCIS & ISIS 2010*. Japan Society for Fuzzy Theory and Intelligent Informatics, pp. 23–27.

Inoue, Kazushige, Takeo Ichinotsubo, and Shingo Aoki (2011). "DEA based hierarchical structure evaluation and visualization method". In: *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, pp. 1701–1704.

Inselberg, Alfred (1985). "The plane with parallel coordinates". In: *The visual computer* 1.2, pp. 69–91.

Keim, Daniel A (2002). "Information visualization and visual data mining". In: *IEEE transactions on Visualization and Computer Graphics* 8.1, pp. 1–8.

Kohonen, Teuvo (1982a). "Analysis of a simple self-organizing process". In: *Biological cybernetics* 44.2, pp. 135–140.

Kohonen, Teuvo (1982b). "Self-organized formation of topologically correct feature maps". In: *Biological cybernetics* 43.1, pp. 59–69.

Kohonen, Teuvo (2001). *Self-organizing maps, volume 30 of Series in information sciences*.

Krivonozhko, VE, OB Utkin, AV Volodin, IA Sablin, and M Patrin (2004). "Constructions of economic functions and calculations of marginal rates in DEA using parametric optimization methods". In: *Journal of the Operational Research Society* 55.10, pp. 1049–1058.

Kruskal, Joseph B and Myron Wish (1978). *Multidimensional scaling*. Vol. 11. Sage.

El-Mahgary, Sami and Risto Lahdelma (1995). "Data envelopment analysis: visualizing the results". In: *European Journal of Operational Research* 83.3, pp. 700–710.

Porembski, Marcus, Kristina Breitenstein, and Paul Alpar (2005a). "Visualizing efficiency and reference relations in data envelopment analysis with an application to the branches of a German bank". In: *Journal of Productivity Analysis* 23.2, pp. 203–221.

Porembski, Marcus, Kristina Breitenstein, and Paul Alpar (2005b). "Visualizing efficiency and reference relations in data envelopment analysis with an application to the branches of a German bank". In: *Journal of Productivity Analysis* 23.2, pp. 203–221.

Raveh, Adi (2000). "Co-plot: A graphic display method for geometrical representations of MCDM". In: *European Journal of Operational Research* 125.3, pp. 670–678.

Sammon, John W (1969). "A nonlinear mapping for data structure analysis". In: *IEEE Transactions on computers* 100.5, pp. 401–409.

Schilling, David A, Charles Revelle, and Jared Cohon (1983). "An approach to the display and analysis of multiobjective problems". In: *Socio-Economic Planning Sciences* 17.2, pp. 57–63.

Serrano-Cinca, Carlos, Yolanda Fuertes-Callén, and Cecilio Mar-Molinero (2005). "Measuring DEA efficiency in Internet companies". In: *Decision Support Systems* 38.4, pp. 557–573.

Sexton, Thomas R, Richard H Silkman, and Andrew J Hogan (1986). "Data envelopment analysis: Critique and extensions". In: *New Directions for Evaluation* 1986.32, pp. 73–105.

Soares de Mello, JCCB, EG Gomes, L Angulo-Meza, L Biondi Neto, UGP Abreu, TB Carvalho, and S Zen (2012). *Ex-post clustering of Brazilian beef cattle farms using SOMs and cross-evaluation DEA models, Applica-tions of self-organizing maps, Ed.*

Talluri, Srinivas, Mary M Whiteside, and Scott J Seipel (2000). "A nonparametric stochastic procedure for FMS evaluation". In: *European Journal of Operational Research* 124.3, pp. 529–538.

Thanassoulis, Emmanuel (2000a). "DEA and its use in the regulation of water companies". In: *European Journal of Operational Research* 127.1, pp. 1–13.

Thanassoulis, Emmanuel (2000b). "The use of data envelopment analysis in the regulation of UK water utilities: water distribution". In: *European Journal of Operational Research* 126.2, pp. 436–453.

Tukey, John Wilder (1977). *Exploratory Data Analysis*. Addison-Wesley. ISBN: 0201076160.

Weber, Charles A and Anand Desai (1996). "Determination of paths to vendor market efficiency using parallel coordinates representation: a negotiation tool for buyers". In: *European journal of operational research* 90.1, pp. 142–155.

Zhu, Joe (2001). "Super-efficiency and DEA sensitivity analysis". In: *European Journal of operational research* 129.2, pp. 443–455.

Article 2

Visualization of Cross-Efficiency Matrix Using Multidimensional Unfolding

Visualization of Cross-Efficiency Matrix Using Multidimensional Unfolding

Abstract

Visualization of data envelopment analysis (DEA) problems is a relatively neglected topic in the DEA literature. This may be partly due to multidimensionality of the DEA problems, and partly because of underestimation of the visualization usefulness. However, through information visualization, not only a vast amount of digits can be comprehensibly represented in a single map, but hidden patterns and structures of the data can be revealed through a bird's-eye view of the data object.

Any DEA visualization method must choose a DEA dataset to be visualized, and a technique to graphically present the chosen data set. Such visualization technique usually is a dimensionality reduction method, since the DEA datasets are oftentimes multidimensional. This study suggests a method to visualize DEA cross-efficiency matrix (CEM), using multidimensional unfolding (MDU) technique.

The suggested methodology is illustrated by means of two artificial datasets, and afterwards two real datasets have been visualized with the new method. The final maps can be used in anomaly detection, and since CEM is composed of the DMUs in both rating and rated aspects, the anomalies, such as maverick units or outliers, can be identified through a comprehensive approach. Nevertheless, the usage of this data exploration tool can go beyond anomaly detection, based on the goals of the researchers.

Keywords: Data Envelopment Analysis, Cross-Efficiency, Cross-Evaluation, Data Visualization, Anomaly Detection, Multidimensional Unfolding

1 Introduction

Since the seminal paper of Charnes et al. (1978), data envelopment analysis (DEA) has been widely accepted and studied to such an extent that the ISI Web of Science database has included 6,500 articles up to 2014, according to Liu et al. (2016). The number would be much higher if one considers all conference papers, unpublished dissertations, and working papers. Nevertheless, searching through this body of literature yields around a dozen of studies focused on the visualization of DEA problems. Hence, it is safe to conclude that visualization of DEA problems is a neglected topic, especially considering that DEA is a data-oriented approach.

This relative neglect may be due to the high-dimensionality of DEA problems, or may be due to considering visualization a trivial task with negligible added-value

to the quantitative analysis. While Adler and Raveh (2008) pointed to the former reason as the difficulty of DEA visualization, the latter reason can be linked to the common belief that visualization is a *low-level task, not appropriate for scientific attention*, as underlined by C.-h. Chen et al. (2007, p. 4). Whatever is the reason, in practice, DEA suffers from lack of a coherent set of visualization procedures, either as exploratory or explanatory tools.

Through exploratory data visualization, as will be discussed in this paper, not only can one literally see a huge amount of data in a single map, but also one can see the overall picture, which otherwise remains hidden. Thus, besides the advantage of the high rate of information communication, the regularities and patterns as well as irregularities and anomalies can be revealed through exploratory data visualization, and innovative ideas can be conceived by doing so. Hence, the exploratory DEA visualization can be a very useful tool for researchers.

"A picture is worth of a thousand words." This adage emphasizes on the amount of information that can be delivered by visualization, and how effectively this communication can be done. In order to assimilate it in our quantitative domain, let's write a slightly different version of it: *"A graph is worth of a thousand digits."* Possibly, one of the most conspicuous benefits of data visualization is the amount of information that can be communicated through a visual map, and the amount of data that can be replaced by a graph. For instance, in a cross-evaluation DEA example with 40 decision making units (DMU), the cross-efficiency matrix(CEM) is composed of 1600 cross-efficiency scores. Since it is not possible to evaluate these scores one-by-one, we use statistics such as average cross-efficiency, which reduces the 1600 scores to 40 column averages of the matrix. Through such reductionist approach, we lose a deluge of details and relations lie in the matrix. Moreover, in order to use a reductionist technique such as simple average, it is needed either to assume that would-be-lost details are trivial and not decisive, or to acquiesce that there is no other way to take all the scores into consideration at the same time.

Anscombe (1973) using a simple example, shows how dangerous it can be to unquestioningly trust analytic measures, and how influential the details can be, while they may seem insignificant.

Anscombe (1973) presents four fabricated datasets, known as Anscombe's quartet, shown in Table 2.1, that all have these characteristics in common:

Number of observations = 11

Number of variables = 2

Mean of variable 1 = 9.0

Mean of variable 2 = 7.5

Variance of variable 1 = 11

Variance of variable 2 = 4.12

Multiple $R^2 = 0.667$

Equation of the regression line = $y = 3 + 0.5x$

Sum of squares of $x - \bar{x} = 110.0$

Regression sum of squares = 27.50 (1 degree of freedom)

Residual sum of squares of y = 13.75 (9 degrees of freedom)

Estimated standard error of intercept coefficient (b_1) = 0.118

Table 2.1: Datasets of Anscombe's quartet

	x1	**y1**	**x2**	**y2**	**x3**	**y3**	**x4**	**y4**
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9	7.5	9	7.5	9	7.5	9	7.5
Variance	11	4.12	11	4.12	11	4.12	11	4.12

Although the identical, up to two decimal places, statistics of the datasets as well as the identical fitted regression models on the datasets may lead a researcher to conclude that the datasets are identical, or at least very similar, the scatter-plots can reveal something different. Figure 2.1 presents the scatterplots of four datasets of Table 2.1.

The surprising discrepancy between the perception of the datasets based on mere analytic measures, and scatter-plots, admonishes not to accept statistic measures unquestioningly, and suggests to use visualization as a complementary tool to analytic methods. This simple example trenchantly reminds the importance of details, and how a holistic approach to data can drastically change our understanding of the data through retaining the details and thus information, comparing to analytic reductionist methods that essentially simplify the problem through shaving some details.

While Anscombe's quartet undermines the assumption of triviality of details, the assumption that we have no other way than accepting reductionist approaches is undermined by visualization. Through proper visualization huge amount of digits can be considered simultaneously, and thus we do not necessarily need to accept reductionist approaches as the only possible approach in order to make sense of large

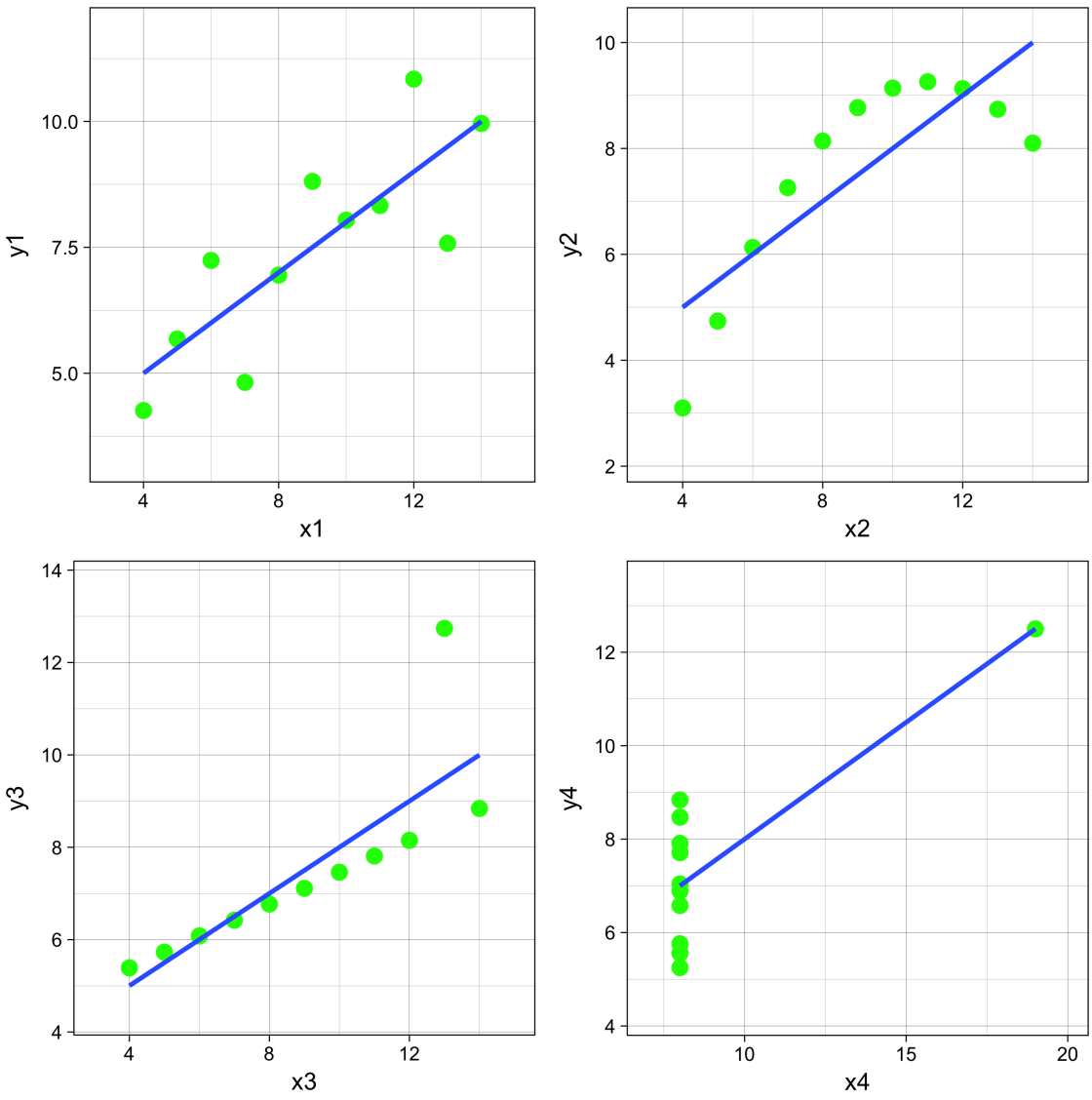

Figure 2.1: Scatterplots of the four fictitious datasets of Table 2.1

sets of data. In fact, the earlier mentioned matrix of 1600 cross-efficiency scores, can be approximately visualized in a single map in order to give researchers the opportunity of taking all details into account. The goal of this paper is to suggest a proper method to visualize such matrices, and to explain how the final maps can be read and explored.

The general goal of the exploratory data visualization is *looking at the data to see what it seems to say* Tukey (1977). Through vision, we gain more information than through all other senses together, and a considerable part of human brain is dedicated to analysis of visual information Ware (2012, p. 2). Human brain is wired for pattern-finding, and in fact, it is the most sophisticated pattern-recognition machine Soukup and Davidson (2002, p. 6). Hence, data visualization can be seen as an effort to get insight into data through using such natural capabilities, and helping brain's cognitive abilities.

Nevertheless, data visualization with all its power has to be considered a complementary tool to numerical techniques.(Cleveland and Cleveland 1985, p. 9) In the words of Anscombe: *"Both calculations and graphs should be studied; each will contribute to understanding"* (Anscombe 1973).

This study suggests a new method to visualize a DEA cross-efficiency matrix(CEM). Such matrix includes self-evaluation of the units as well as the peer-evaluation scores of each unit. Hence, the matrix is composed of n^2 cross-efficiency scores for a DEA problem of n decision-making units(DMUs). While such a matrix can easily be incomprehensible and overwhelming, its visual representation can be quickly and relatively easily understood. The method, as an exploratory data visualization tool, can be used for different purposes. However, in this study, the focus is on anomaly detection among the DMUs.

Before explanation of the new CEM visualization method, a brief review of DEA visualization studies is presented in Section 2, followed by an introduction to DEA CEM in Section 3, and an introduction to multidimensional scaling(MDS) and unfolding(MDU) in Section 4. In Section 5, the suggested CEM visualization method is illustrated using two artificial datasets and two real datasets. Finally, the conclusion of the study, and the next possible steps are presented in the last section.

2 DEA Visualization

Although in practice DEA visualization methods are not used frequently, the DEA visualization toolbox is not empty. The first efforts to visualize DEA problems were suggested almost a decade after the emergence of the first DEA models. The early visualization methods were focused on using uni-variate or bi-variate graphs to depict DEA-related data visually.

Desai and Walters (1991) and Weber and Desai (1996) suggested visualizing input and output levels of DMUs through parallel coordinates. Using the parallel coordinates(Inselberg and Dimsdale 1987), DMUs can be compared according to their inputs and outputs. The range of acceptable values for each input or output is depicted, and the path of improvement for inefficient units is shown. Nonetheless,

with the increase of the number of DMUs, the graph rapidly becomes too crowded and loses its comprehensibility.

Belton and Vickers (1993) suggested visualization in a scatter plot, where the coordinates of each unit are the aggregate value of inputs and the aggregate value of outputs of that unit. The aggregation can be done through a weighted sum of input and output levels, and the weights are achieved through a revised DEA model, suggested by the authors. Hackman et al. (1994) proposed a method to traverse the boundaries of the production possibility set from a given problem. As a result of the algorithm, a pair of coordinates is achieved for each DMU, and visualization can be done using these coordinates.

El-Mahgary and Lahdelma (1995) suggested visualizing different perspectives of DEA problems using a set of bi-dimensional plots. Among the plots, for instance, there is a scatter plot of DMUs where the coordinates are a chosen input and efficiency score to highlight the relationship between that specific input value and the efficiency measures.

Talluri et al. (2000) suggested the first cross-efficiency visualization method. The method was composed of two plots. The first one was a scatter-plot of simple efficiency and average cross-efficiency of each unit, while the second plot was a side-by-side boxplot, based on the column values of each unit in the CEM.

In the second decade of DEA visualization efforts, researchers started to use dimensionality-reduction techniques, such as principal component analysis(PCA) and MDS, to project multi-variate and high-dimensional DEA data objects into low-dimensional maps. To do so, Serrano-Cinca et al. (2005) developed a method based on the projection of efficiency vectors of DMUs into a bi-dimensional map using PCA. The efficiency vector of each unit is composed of the efficiency scores of the unit under different selection sets of input and output variables. The map is augmented by super-imposing the property vectors on it, such that each vector shows the direction in which the value of one property, here a specific DEA model, increases. The main goal of their method is variable selection, and visualization is a means to that goal. The method can be considered as a bi-plot variation.

Porembski et al. (2005) benefited from the *non-linear mapping* of Sammon (1969) to project the DMUs into a two-dimensional(2D) map using their input and output values. In this map, similar units would be located close to each other, while dissimilar units would reside far from each other. Porembski et al. (2005) further augmented the map by adding links between inefficient units and their reference units, such that the thickness of the links are a function of dual multipliers of the corresponding DEA model.

Aoki et al. (2007) used fuzzy correspondence analysis to visualize a matrix of dual weights of units, resulting from a modified BBC model of Banker et al. (1984). Hence, the final map is composed of neighborhoods of similar units based on the similarity of the dual-weight profile. Later, Honda et al. (2010) visualized the matrix of dual weights, resulting from the super-DEA algorithm by Zhu (2001) using fuzzy PCA. The goal of such a visualization is explained as emphasizing *the mutual relations among efficient DMUs* as well as *clarifying the relations between the inefficient DMUs and their target units.*

Adler and Raveh (2008) suggested a method to represent DEA problems graphically using ratio factors of DMUs and smallest space analysis(SSA), the non-metric MDS method of Guttman (1968). The ratio factors are calculated by division of each output to each input, such that, for a DMU with m inputs and s output, there would be $m \times s$ output/input ratios. Afterwards, the dimensionality of matrix of ratios is reduced to two dimensions using SSA, which can easily be plotted. A super-imposed vector map of properties, i.e. the ratios, is used to illustrate the differentiating factors of DMUs.

Visualization of the DEA frontier is the main goal that a group of studies has pursued. Førsund et al. (2009) focused their efforts on visualization of the properties of the frontier function. Appa, Costa, et al. (2010) suggested a visualization method able to depict the efficient frontier and the relative distance of DMUs from it. However, the method is limited to one-input DMUs. Costa et al. (2016) expanded this method to multi-input and multi-output DMUs.

Inoue et al. (2011) suggested an information visualization method for DEA. Their method is a visual representation of a hierarchical structure of sub-units, such that each sub-unit is composed of a subset of inputs and outputs of the original unit. Through the visual representation, the absolute and relative strengths and weaknesses of each unit can be highlighted.

Akçay et al. (2012) suggested a scatter-plot based visualization for DEA, such that two variables are shown as the coordinates, while the size and color of the points are reflections of two other variables. Hence, four variables of DMUs can be shown in a bi-dimensional map without dimensionality reduction. In addition, these authors proposed a tree-map, partitioned based on a categorical variable such as DMU geographical regions, to show the DEA-related information of each region in a subtle way. The partition color and size are also a reflection of their corresponding DMU variables.

In a technically different set of studies, Kohonen Self-organizing map(SOM) has been used to visualize different data objects of DEA problems. SOM is basically an artificial neural network which can be used to non-linearly project a high-dimensional set of objects into a usually two dimensional grid of nodes, such that the topology of the original data space is preserved as much as possible (Kohonen 2001). Visualization of DEA has been either the main goal or by-product of using SOM method in several studies such as Churilov and Flitman (2006), J. C. C. B. S. d. Mello et al. (2012), and Carboni and Russu (2015).

3 Cross-Efficiency

Consider the very basic model of DEA, suggested by Charnes et al. (1978) :

$$\text{Max} DMU_0 = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}} \tag{2.1}$$

Subject to:

$$\begin{aligned} \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} &\leq 1; \qquad j = 1, \cdots, n \\ v_i, u_r &\geq 0; \qquad r = 1, \cdots, s; i = 1, \cdots, m \end{aligned} \tag{2.2}$$

In this model, the left hand side of the constraints, can be interpreted as the efficiencies of other DMUs than DMU_0, evaluated by one optimum weight set of DMU_0. The efficiency of DMU_0 evaluated by DMU_0 is simple efficiency, and the efficiency of every other DMU, evaluated by DMU_0 is called cross-efficiency. Cross-efficiency of a generic DMU_k assessed by DMU_0 is thus expressed as below:

$$e_{0k} = \frac{\sum_{r=1}^{s} u_{r0} y_{rk}}{\sum_{i=1}^{m} v_{i0} x_{ik}} \tag{2.3}$$

If the cross-efficiency scores achieved by weight optimization of any DMU are arranged row-wise in a matrix, then we get a cross-efficiency matrix(CEM). In cross-efficiency matrix, the diagonal elements are simple efficiencies.

$$CEM_{n,n} = \begin{pmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,n} \\ e_{2,1} & e_{2,2} & \cdots & e_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n,1} & e_{n,2} & \cdots & e_{n,n} \end{pmatrix} \tag{2.4}$$

In the CEM, each DMU has two roles, a rating role and a rated role. In the rating role, a DMU evaluates all DMUs including itself, and in the rated role, a DMU is evaluated by all DMUs including itself. The rating role corresponds to the row profile, and the rated role corresponds to the column profile. In other words, the row profile is about how the corresponding DMU evaluates the DMU set, and the column profile is about how the corresponding DMU is evaluated by the DMU set.

This rather simple but interesting idea was presented by Sexton et al. (1986) to tackle one shortcoming of the original formulations of DEA. More specifically, Sexton et al. (1986) pointed to the problem of price efficiency in DEA, while appreciating the power of the DEA method in general. According to the authors, the problem of price efficiency stems from the ability of the DMUs to choose their input and output weights without constraint, a freedom that, in extreme cases, leads to technically efficient and price-inefficient DMUs. To improve this situation, the authors suggested *to subject each DMU to a range of input and output weights - not any range but rather the range of weights chosen by the other DMUs in analysis*, and named this approach cross-evaluation. Hence, a cross-efficiency score is the efficiency score resulted by evaluation of a particular DMU through the optimum weights of another DMU, and it contrasts with simple efficiency, i.e. the efficiency score resulting from self-evaluation. Therefore, the average of all cross-efficiencies (evaluation scores given to a DMU by other DMUs) can be used as an overall efficiency score of that specific DMU. The contrast between simple efficiency and average cross-efficiency can be an index of price inefficiency. However, Sexton et al. (1986) wisely emphasized that there is no guarantee that a unit with a high contrast, between self and average

cross-efficiency scores, is a price-inefficient unit, but such unit is a certain candidate for further investigation.

Doyle and Green (1994) extended the idea of cross efficiency and brought it to the next level. The authors coined self-appraisal and peer-appraisal terms for simple and cross-efficiency, where the peer of a DMU in this context simply means any other DMUs in that dataset. The authors suggested that cross-efficiency has the advantage of explicit weight restriction since the latter is rather arbitrary and authoritarian, while the former is a more democratic process. Prior to the price-efficiency issue, Doyle and Green (1994) suggested that the cross-efficiency approach can be used to establish a meaningful ranking among the set of DMUs with 100% simple efficiency. Hence, cross-efficiency was suggested to improve another DEA shortcoming, which is the lack of discrimination ability. This usage of cross-efficiency became the main characteristic of the approach to such an extent that cross-efficiency has become well-known as a DEA ranking method, for instance Adler, Friedman, et al. (2002) and Ruiz and Sirvent (2016).

Although cross-efficiency is an attractive method to improve the original formulation of DEA through increasing discrimination ability as well as finding possible mavericks (i.e., technically efficient but price-inefficient units), the method has some drawbacks. One of the main problems of cross-efficiency is the multiplicity of CEMs. A CEM, as the sole output of the method, is the source of further analysis and calculation of indices, such as the average cross-efficiency and maverick index. However, every DEA problem has multiple CEMs. The phenomenon of multiple CEMs stems from the well-known problem of alternate optima in DEA, i.e. the problem of multiplicity of optimum weights.[1] Without a unique CEM, the ranking of units would not be unique, and generally, the usefulness of other CEM-based indices would be undermined (Cook and Zhu 2014).

This problem has been addressed from the first studies of cross-evaluation. The general approach to this problem has been the definition of new criteria to choose an alternate optimum among all possible alternate optima. In other words, the researchers have tried to justify choosing one specific CEM among the numerous alternatives based on clear criteria. The choice of an optimum alternate in cross-evaluation methodology has received significant attention from researchers of the field.

To mitigate the CEM multiplicity problem, Sexton et al. (1986) and Doyle and Green (1994) suggested using a secondary objective, i.e., goal programming, as a potential remedy. Alternatively, the authors incorporated a secondary goal into the original optimization, such that each unit not only maximizes its simple efficiency (primary objective), but also minimizes other units' cross-efficiency scores(aggressive approach) or maximizes other units' cross efficiency scores(benevolent approach) as the secondary objective.

Subsequent to the classic benevolent and aggressive models, many new models have been proposed in the literature. For instance, Appa, Argyris, et al. (2006) suggested a new framework to generate a unique CEM using all hyperplanes that define the constant return to scale(CRS) production possibility set to evaluate DMUs

[1] Fumero (2004) illustrated this phenomenon in a numerical example.

across all weighting schemes. Liang et al. (2008a) extended the idea of secondary goals by introducing some alternative goals besides the classic aggressive and benevolent formulations. Liang et al. (2008b) proposed a game model of cross-efficiency. Wang and Chin (2010) developed a neutral model of CEM, in contrast to the aggressive and benevolent models. Ramón et al. (2010) suggested an approach to formulate CEM using the reasonable and realistic alternate optima. Lim (2012) suggested minimax and maximin formulations of cross efficiency. Cook and Zhu (2014) developed and proposed a multiplicative model, which ensures a unique CEM.

The proposed visualization method in this study is independent of the CEM formulation; in other words, any CEM can be visualized with this method, and the choice of CEM is based on the application and the goal of visualization. Nevertheless, in this study, the classic benevolent formulation is used, and the reason for doing so is explained in Section 3.2.

While cross-efficiency originally works under the CRS assumption, a variable return to scale (VRS) extension has recently been suggested by Lim and Zhu (2015). In addition, Chai et al. (2013) used cross efficiency with the free disposal hull(FDH).

The cross-efficiency methodology has found applications in various fields and sectors, for example, selection of R&D projects (Oral et al. 1991), technology selection (Baker and Talluri 1997), electricity distribution section efficiency analysis (T.-y. Chen 2002), real estate in many cities of China (Li 2008), ranking of the Athens Olympic games (S. d. Mello et al. 2008), benchmarking of the countries at the Summer Olympics (Wu et al. 2009), ranking basketball players (Cooper et al. 2011), hotel performance evaluation (Fu et al. 2011), public procurement tender benchmarking (Falagario et al. 2012), NBA team efficiency analysis (Aizemberg et al. 2014), and portfolio selection in the Korean stock market (Lim, Oh, et al. 2014) to name a few. For a comprehensive survey of the cross-efficiency methodology, readers are referred to Ruiz and Sirvent (2016).

4 Dimensionality Reduction of Cross-Efficiency Matrix

To visualize any cross-efficiency matrix(CEM), one must reduce the dimensionality of the matrix if the number of DMUs is above three. Doing so can be done using Multidimensional Unfolding(MDU), a special variety of Multidimensional Scaling(MDS). Hence, introduction to scaling is the prerequisite of the introduction to unfolding. This section consists of these two sequential steps.

4.1 Multidimensional Scaling

The MDS is composed of a set of techniques which can produce a spatial map of objects from the proximity matrix of those objects. These techniques attempt to represent, as precisely as possible, the proximity values of every pair of objects as spatial distances of the pairs, in a low-dimensional space (Kruskal and Wish 1978, p. 7). A proximity matrix of a dataset is a matrix of pairwise measurements

of (dis-)similarity, closeness, preference, or relatedness among the members of the dataset (Borg, Groenen, and Mair 2012, p. 3).

The low-dimensional space to which MDS projects a proximity matrix is usually a Euclidean bi-dimensional space. Hence, the final output of the MDS is a visual representation of the input, the proximity matrix, such that more similar objects locate closer to each other.

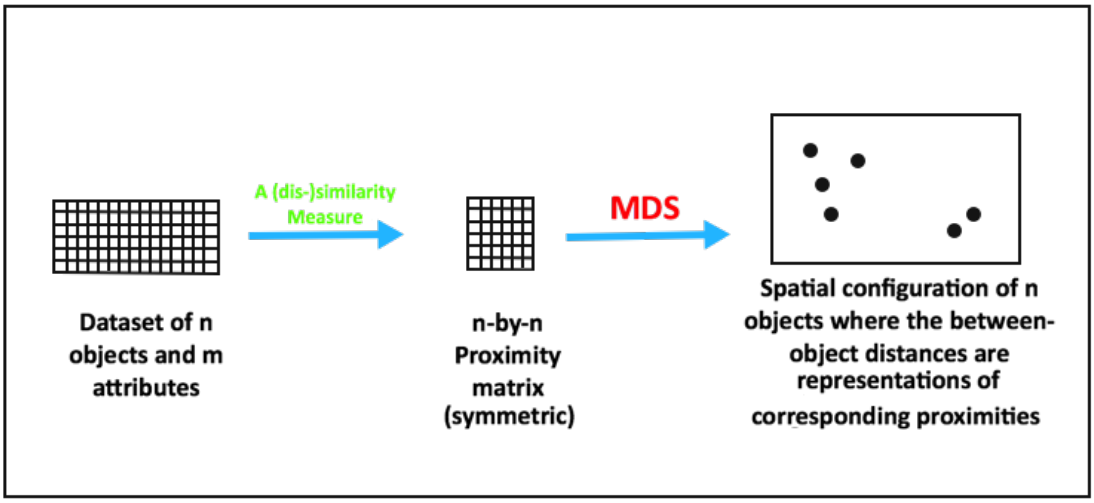

Figure 2.2: Dimensionality reduction through Multidimensional Scaling

The graphical display makes comprehension of the data more lucid, since it is much easier to look at a map of n objects, rather than analysis of a n-by-n symmetric matrix of digits. Moreover, a map enables researchers to see the big-picture, a perspective which may otherwise remain hidden.

Hence, MDS can be used as an exploratory tool in order to evaluate the structure of the data, i.e. the regularities and irregularities. Such visual exploration provides invaluable insight into data.

MDS includes different models based on how proximity values are mapped into distances in the final spatial configuration.(Borg and Groenen 2005, p. 37) Any specific model of MDS is defined by characteristics of such representation function, which transforms proximity values in order to be used as distance goals. In the mathematical notation it can be written:

$d_{ij} = f(p_{ij})$ [1]

Where d_{ij} is the distance between objects i and j in the final configuration, p_{ij} is the proximity value between objects i and j, and $f()$ is the representation function.

As it is stated earlier, MDS attempts to represent the proximity values by the spatial distances, as precise as possible. It means that distances in MDS configuration are not exactly match with their corresponding proximity values, since usually the original dimensionality of the objects is much higher than the dimensionality of the final map. Hence, the [1] should be re-written as [2] :

$d_{ij} \approx f(p_{ij})$ [2]

The values of the transformation are fitted "fitted distances", and MDS tries to minimize the difference between the original distances, and the fitted distances, a.k.a disparities. Kruskal and Wish (1978, p. 29)

$d_{ij} \approx \hat{d}_{ij} = f(p_{ij})$ [3]

The difference between the two sides of [2], the spatial distances and fitted distances, is known as the loss function in MDS, i.e. the error term:

$e_{ij} = f(p_{ij}) - d_{ij}$ [4]

In practice, MDS computation procedure tries to minimize such loss function through iteratively changing the position of objects on the final map (T. F. Cox and M. A. Cox 2000, p. 67). In order to evaluate the success of the procedure to find "as close as possible" fitted-values to the original distances, $Stress-1$ index is usually used as the loss function:

$Stress-1 = \sqrt{\frac{\sum[\hat{d}_{ij}-d_{ij}]^2}{\sum d_{ij}^2}}$ [5]

Therefore, the less $stress-1$ value, the less error on the final configuration, and the better goodness-of-fit in general.

There are many different MDS models based on different representation functions that can be used in transformation of proximity values. Accordingly, MDS models usually refer to general forms of the representation functions such as "ratio", "interval" or "ordinal".

In the ratio MDS, the representation function would be a linear function without intercept and with a fixed coefficient. Thus, ratio MDS would have the following form:

$f(p_{ij}) = b \cdot p_{ij}$,where b is constant [6]

If there is a non-zero intercept in such linear representation function, then we have the interval MDS formulation

$f(p_{ij}) = b \cdot p_{ij} + a$ [7]

Hence, there is a fixed-origin assumption for proximity values behind ratio MDS, which makes the model class different from interval MDS (Borg and Groenen 2005, p. 34).

If the representation function tries to optimally preserve only the rank-order of the proximity values, then it is labeled as "ordinal MDS". Such representation function must perform monotone transformation and satisfy the monotonicity constraint represented in [8] (Schiffman et al. 1981, p. 10).

$p_{ij} < p_{kl} \Rightarrow f(p_{ij}) \leq f(p_{kl})$ for every i, j, k, l in the possible range of the problem [8]

At the end, it is important to note that the ratio and interval MDS belong to metric MDS, and ordinal belongs to non-metric MDS.

4.2 Multidimensional Unfolding

A subclass of MDS is Multidimensional unfolding(MDU), which can cope with two-mode matrices. A matrix has one mode if the sets of items of the rows and columns are identical, and it has two modes if the sets of items of the rows and columns are different. In other words, the modes are the number of different item-sets represented in a matrix (Forrest W. 1987, p. 50).

Alternatively, in contrast to an nxn input matrix of MDS, the input matrix of MDU includes proximity values between n row objects and m column objects. The proximity of MDU input matrices are usually about preference, such that p_{ij} is a measure of how much $row - object_i$ prefers $column - object_j$ (Mair et al. 2015). The MDU tries to represent both sets of items in a joint space, where the row items are column items located such that the preference values (i.e., proximity values) are reflected in the corresponding distances between the row and column objects as precisely as possible (T. F. Cox and M. A. Cox 2000, p. 7).

All the mentioned characteristics make MDU an appropriate technique for dimensionality reduction and visualization of DEA CEMs. While it may seem that the sets of DMUs in the rows and columns of a CEM are identical, there is a subtle difference between their roles in the two sets. The DMUs in a CEM are in two rating and rated roles, i.e. evaluating and evaluated roles, and hence two distinct sets. Thus, CEM is a two-mode matrix. The CEMs are asymmetric since $e_{ij} \neq e_{ji}$ for generic DMUs i and j, and are unconditional matrices since all cross-efficiency scores are comparable between rows and columns (De Leeuw 2005).

The cross-efficiency scores can be interpreted as endorsement values of row DMUs to column DMUs, such that the higher the cross-efficiency score, the stronger the endorsement. Such interpretation is also closely related to preference, since the rating DMUs prefer rated DMUs with higher achieved cross-efficiency scores.

As a result, the unfolding 2D map of a CEM is composed of the DMUs as points in two distinct sets of rows and columns. Similar to MDS maps, the distances reflect the proximities, i.e. endorsements. Hence, the higher a cross-efficiency score given by row $unit_i$ to column $unit_j$, the closer the column $unit_j$ to the row $unit_i$. Consequently, unfavorably evaluated DMUs would be located far from the evaluating DMUs, while the highly praised evaluated DMUs would reside closer to the evaluating DMUs.

In the next section, the unfolding map of CEM is elaborated through two artificial examples. Afterwards, the method is applied on two real datasets, and the visual results are construed. All computations regarding DEA CEM are done using the R(Team 2016) code written by the corresponding author, and the rest of the computations and visualizations are done using lpSovle (Berkelaar et al. 2015), SMACOF (Jan De Leeuw and Mair 2011), ggplot2 (Wickham 2016), and ggrepel (Slowikowski n.d.) packages of R statistical software (Team 2016).

5 Illustration of cross-efficiency matrix unfolding

The first dataset to evaluate is a small CEM with three DMUs. The cross-efficiency scores have been manually assigned to make this CEM a good starting point to illustrate the CEM visualization. Hence, such a combination of scores may never be seen in a real dataset, and this data fabrication has the sole purpose of elaboration of CEM visualization.

As in Table 2.2, there are three units in this CEM. Both DMU_1 and DMU_2 have similar row profiles. This can be checked either based on a distance measure, such as Euclidean distance, of the profiles or just a rough evaluation through the order rank of the rated peers by these DMUs, (i.e., for both DMUs, the ranking of the peers

Table 2.2: A fabricated CEM

	DMU1	DMU2	DMU3
DMU1	1.0	0.4	0.1
DMU2	0.9	0.5	0.4
DMU3	0.2	0.4	0.9

is 1, 2, and 3.) In contrast, the ranking of DMU 3 is 3, 2, and 1. Hence, it has a different row profile. On a metric basis, such as Euclidean distance, these relations can be assessed according to Table 2.3.

Table 2.3: Row-wise Euclidean distances of Table 2.2

	DMU1	DMU2	DMU3
DMU1	0.0000000	0.3316625	1.1313708
DMU2	0.3316625	0.0000000	0.8660254
DMU3	1.1313708	0.8660254	0.0000000

Similarly, we can get an idea of column profiles. For example, DMU_1 is strongly endorsed by itself as well as DMU_2, while DMU_3 is strongly endorsed only by itself, and DMU_2 is a rather inefficient unit receiving similar low cross-efficiency scores.

Nevertheless, we will not analyze the similarities of the profiles quantitatively, since this is the purpose of the final map to visually represent such relations as precise as possible. In this example, such an assessment is done to have a benchmark for map validation.

According to (dis)similarity of the row profiles, it is expected to have row objects of DMU_1 and DMU_2 close to each other, while row object of DMU_3 should be located far from them. From the other side, it is anticipated that the column object of DMU_1 is positioned close to the row objects of DMU_1 and DMU_2, while it should be far from the row object of DMU_3. Following the same pattern, the relations of the row and column objects can be portrayed verbally; however, doing so becomes protracted and frustrating by increasing the size of the CEM, and the big-picture of the CEM cannot be grasped in this manner. Hence, CEM visualization is a handy tool to depict all these relations quickly and efficiently.

Before venturing on to use MDU on CEM 1, it should be underscored that MDS and MDU work with dissimilarity values; however, cross-efficiency scores are measures of similarity and preference. Hence, it is necessary to convert the CEM into a dissimilarity matrix. Doing so can be done by subtraction of the CEM from a same-size matrix of ones.

Figure 2.3 represents the unfolding map of CEM1.

From now on, the column/row profile and column/row object are used interchangeably. The row objects are shown with solid blue circles, and the column objects are portrayed in red triangles. Row objects of DMU_1 and DMU_2 reside close to each other, which means that these two units have similar row profiles. In

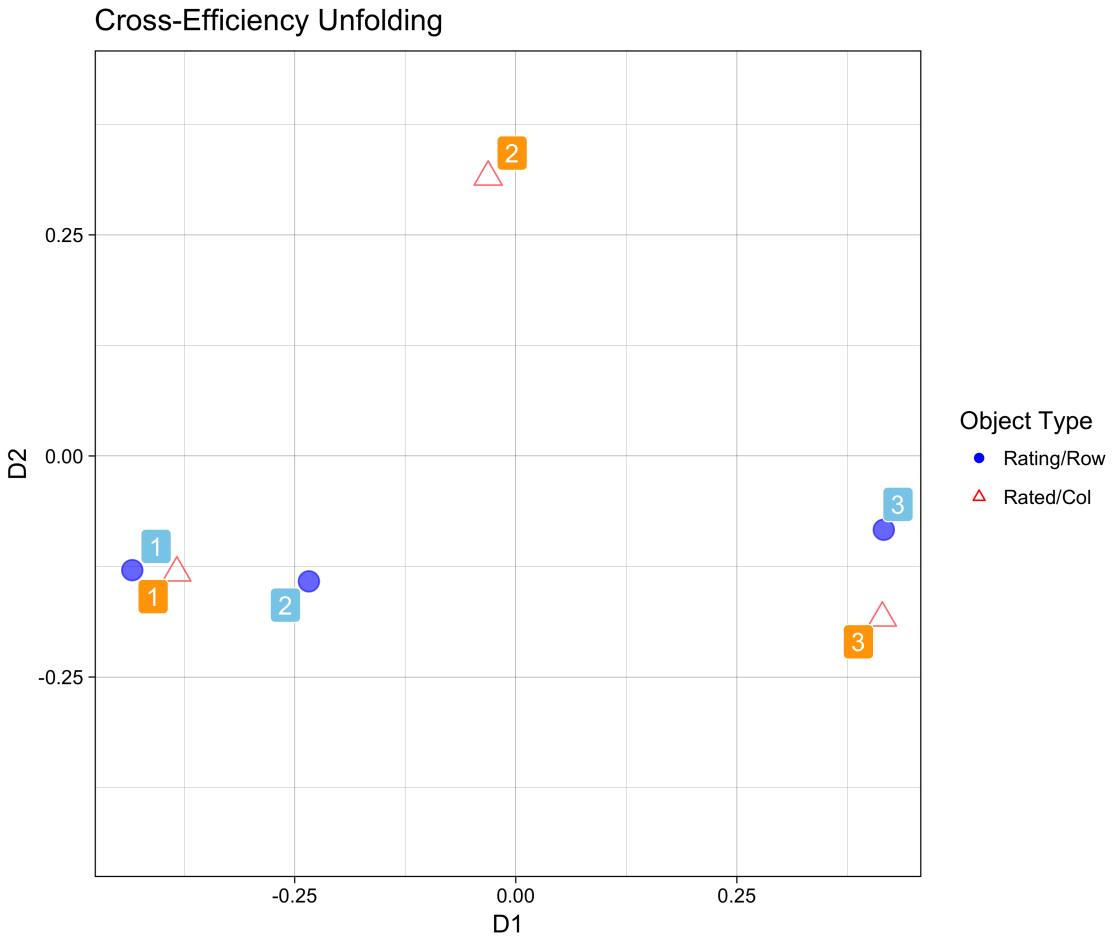

Figure 2.3: Unfolding map of Table 2.2

contrast, DMU_3 is very distant from its peers; hence, it has a very different row profile. Therefore, two clusters of row profiles are discernible.

Consideration of column objects makes the interpretations more interesting. Column Object 1 is in the center of attention of row Objects 1 and 2, as it is located between these two row objects and relatively close to them. This is because both DMU_1 and DMU_2 endorse DMU_1 as the most efficient unit. Column Object 2 is far from every row object since it is considered a very inefficient unit. Column Object 2 is almost equidistant from all row objects; however, it is closer to row Object 2. On the other hand, column Object 3 is far from row Objects 1 and 2, but close to its corresponding row object. It means that DMU_3 considers itself as an efficient unit, while peers disagree with it. From the point of view of anomaly detection, DMU_3 is the most interesting unit since it is not only uncommon from the row profile perspective, but it is also discordant from the column profile. While DMU_2 may be another candidate of anomaly units, DMU_3 has two subtle differences from DMU_2. Row-wise, the row profile of DMU_2 is similar to DMU_1; however, the row profile of DMU_3 is unique, and column-wise, DMU_2 does not consider itself an efficient unit, while DMU_3 has 100% simple efficiency.

It is of immense importance to emphasize that the row profiles are functions of the corresponding unit's optimum weight. Hence, a discordant row object reflects discordant optimum weights. On the other hand, the column profiles are roughly a reflection of the corresponding units' input and output levels.

In the CEM unfolding maps, such as Figure 2.3, the dimensions have no inherent meaning. As stated before, the MDU attempts to represent only the distances; thus, the dimensions do not hold any intrinsic meanings. This means that, while we know that row Object 3 is different from row Objects 1 and 2, the map cannot clarify how these units are different.

Before venturing on visualization of the second dataset, it should be underscored that the choice of CEM approach changes the cross-efficiency scores, and consequently the unfolding configuration. In this example as well as the two real datasets, benevolent CEM is used for two reasons. First, the benevolent CEM is a well-established formulation in the literature, with a simple concept behind it. The formulation forces every DMU to choose an optimum weight set among its alternate optima, such that the chosen weight set is the most favorable on average for other DMUs. Under an ideal benevolent formulation, DMUs would choose the optimum weight set, which is the most favorable and thus the most compatible with the structure of other DMUs. Hence, under such a circumstance, if a DMU shows uncommon behavior, then the DMU probably has an uncommon nature. This is the second reason for choosing a benevolent formulation since the suggested CEM visualization method in this study can be used in the detection of uncommon units,i.e. anomalies.

Nonetheless, there is one serious issue regarding the benevolent formulation. The benevolent formulation of Doyle and Green (1994) does not find the most benevolent weights, but an approximation of it. In fact, it attempts to find the benevolent weights based on an imaginary aggregated DMU,and not based on considering every single DMU. While the authors are aware of this caveat, this formulation is still used

in this study due to the stated rationale.

It is noteworthy that any other CEM formulation can be used in this visualization method, without the least necessity of change in the procedure. Hence, the choice of benevolent formulation is not a critical topic, and the CEM formulation can be changed based on different rationales.

The second artificial dataset, used to illustrate the CEM visualization, is regarding seven departments in a university (Wong and Beasley 1990; Wang and Chin 2010). Table 2.4 shows the input and output levels of these departments.

Table 2.4: Input/Output levels of seven Academic departments (Wang and Chin 2010)

	i1	i2	i3	o1	o2	o3
DMU1	12	400	20	60	35	17
DMU2	19	750	70	139	41	40
DMU3	42	1500	70	225	68	75
DMU4	15	600	100	90	12	17
DMU5	45	2000	250	253	145	130
DMU6	19	730	50	132	45	45
DMU7	41	2350	600	305	159	97

The inputs are $I1$: number of academic staff, $I2$: academic staff salaries in thousands of pounds, and $I3$: support staff salaries in thousands of pounds. The outputs include $O1$: number of undergraduate students, $O2$: number of postgraduate students, and $O3$: number of research papers.

The benevolent CEM of Table 2.4, computed by Wang and Chin (2010), is presented in Table 2.5.

Table 2.5: Benevolent Cross-efficiency of seven academic departments (Wang and Chin 2010)

	DMU1	DMU2	DMU3	DMU4	DMU5	DMU6	DMU7
DMU1	1.000	0.981	0.769	0.641	0.938	1.000	1.000
DMU2	0.922	1.000	0.772	0.701	0.899	1.000	1.000
DMU3	1.000	0.851	1.000	0.454	0.495	1.000	0.294
DMU4	0.688	1.000	0.735	0.820	0.765	0.951	1.000
DMU5	1.000	0.846	0.665	0.414	1.000	0.910	1.000
DMU6	1.000	0.981	0.769	0.641	0.938	1.000	1.000
DMU7	1.000	0.981	0.769	0.641	0.938	1.000	1.000

Figure 2.4 depicts the unfolding configuration of Table 2.5.

While it is possible to assess every little relationships in the CEM-unfolding map, we focus on detection of uncommon units. Hence, the striking points of the map are related to DMU_3 and DMU_4, both row and column objects. In addition, DMU_3 is distant from the crowd in both aspects, while its row and column objects

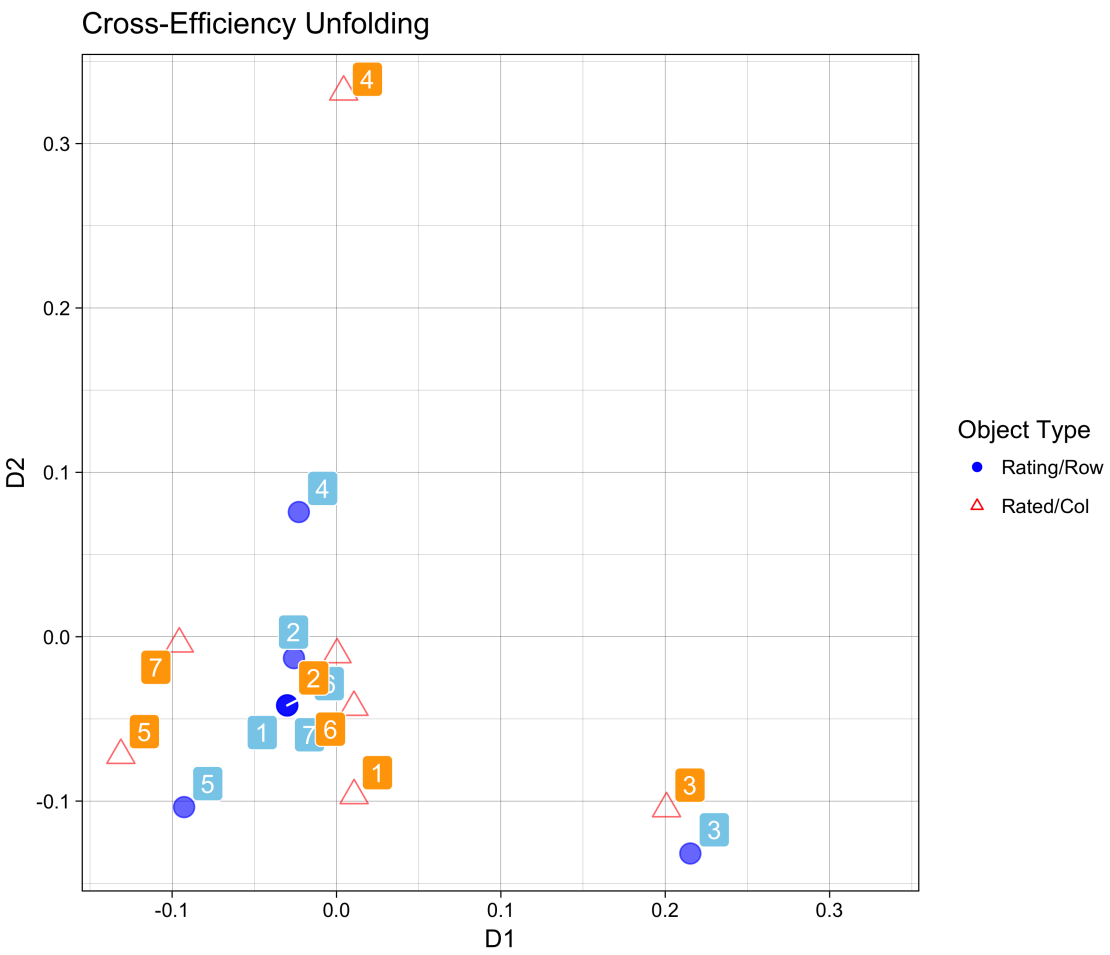

Figure 2.4: Unfolding map of Table 2.5

are very close to each other. It means that DMU_3 considers itself an efficient unit, and its row profile is relatively different from the others. In addition, the isolation of column Object 3 is due to the relatively low cross-efficiency scores that it has received. While the column object of DMU_4 is also isolated, which reflects low average cross-efficiency scores of the DMU_4, the row object of DMU_4 is closer to other row objects, compared to the location of the row object of DMU_3. Being so means that the row profile of the DMU_4 is more common, than the row profile of DMU_3. According to the visual evidences, DMU_3 is a unit with an uncommon structure, while DMU_4 is an inefficient unit with a relatively similar structure to the other units. The situation of other units can be construed following the same pattern. For instance, it seems that DMU_5 has an uncommon structure but in a different way from DMU_3, and with less significance. It should be kept in mind that all the maps have noise, due to the reduction of the CEM dimensionality to two dimensions. Therefore, the maps should not be considered an exact tool, but an exploratory tool with a degree of error to gain approximate insight into the data and find interesting behaviors through examination of evidences.

The first real dataset is from Bahari and Emrouznejad (2014), adapted from Hatami-Marbini et al. (2012). The dataset, presented in Table 2.6, is about 38 hospitals as DMUs with three inputs and two outputs. The inputs are $I1$: number of beds, $I2$: labor-related expenses in dollars, and $I3$: patient care supplies and other expenses in dollars. The outputs are $O1$: number of outpatient department visits and $O2$: number of inpatient department admissions.

The rationale of choice of this dataset is the possibility of assessment of visualization evidences through comparison with the outlier detection results of Bahari and Emrouznejad (2014). The benevolent CEM is not presented in this paper due to space limitations, but its visualization is presented in Figure 2.5. To avoid overcrowding, the row object labels have been omitted from Figure 2.5, and separately expanded and represented in Figure 2.6.

In Figure 2.5 and Figure 2.6, the row objects are quite homogeneous and thus overlapped, compared to previous examples. However, Units 19, 20, and 23 are slightly far from the crowd. Hence, the position of the majority of the column objects is more informative since most of them are located far from the row objects in two opposite corners of the map. This means that the majority don't have the overall endorsement by the most of the row objects, so they are repelled by the row objects. In contrast, five column objects are located close to the row objects, which means that they are preferred units by most evaluators, i.e. DMUs. These column objects with relatively high average cross-efficiency scores are the candidates of anomalies in the dataset. All of them seem to have high endorsement from the row objects, so they may be considered efficient outliers.

The findings of the unfolding map is completely compatible with the result of Wilson's outlier-detection method (Wilson 1993), implemented in Benchmarking package (Bogetoft and Otto 2015) of R (Team 2016). According to Wilson's method, DMUs 3, 7, 8, 12, and 33 are top 5 outliers. Similarly, Figure 2.5 suggests DMUs 3, 7, 12, 33, and 8 as potential anomalies, and more specifically potential efficient outliers based on their location relative to the majority of the row objects.

Table 2.6: Inputs and Outputs levels of 38 Hospitals (Bahari and Emrouznejad 2014)

DMU Number	i1	i2	i3	o1	o2
1	82	5714634	3206439	38267	4191
2	78	6269841	3587936	40809	5146
3	54	5088442	2188030	45900	6638
4	79	5588288	2967667	41238	5640
5	75	6398933	2495584	40608	5442
6	85	5661388	2491011	47600	5731
7	58	4197778	2056911	47142	5371
8	69	5931560	2372624	62250	5127
9	76	6474930	2935223	52283	4098
10	80	5612130	3773095	42067	3396
11	74	6503709	3336595	38130	5261
12	57	3868281	1972823	51414	5078
13	78	6206204	2987544	51122	4273
14	69	6923132	3122178	40357	4501
15	78	5508713	3093310	44322	4028
16	81	6011040	2824636	41160	5056
17	77	5981179	3175478	37518	5246
18	82	6260171	3885466	41545	3204
19	49	5237457	2304481	40817	4177
20	64	6001559	2400623	34825	5269
21	85	5534937	2380023	38374	4583
22	84	6424688	3377135	34170	5960
23	80	5385786	3490456	54930	4027
24	78	6520290	3296928	39664	4351
25	89	6963796	3828260	55381	5059
26	79	5916156	2906462	51016	4384
27	85	6308766	3226419	51689	5391
28	70	6070424	2357057	41160	5649
29	81	6256229	3442054	37960	4585
30	78	6156335	3057755	52204	5334
31	78	5918535	3207840	37362	5285
32	85	6489480	3697026	41591	5252
33	55	6058943	2250349	62364	5542
34	78	6364389	3135457	34825	4391
35	80	5671696	3231573	47174	5259
36	79	6494440	2708815	39187	4124
37	81	6845886	3845247	42930	4881
38	85	6260432	3740583	45698	4456

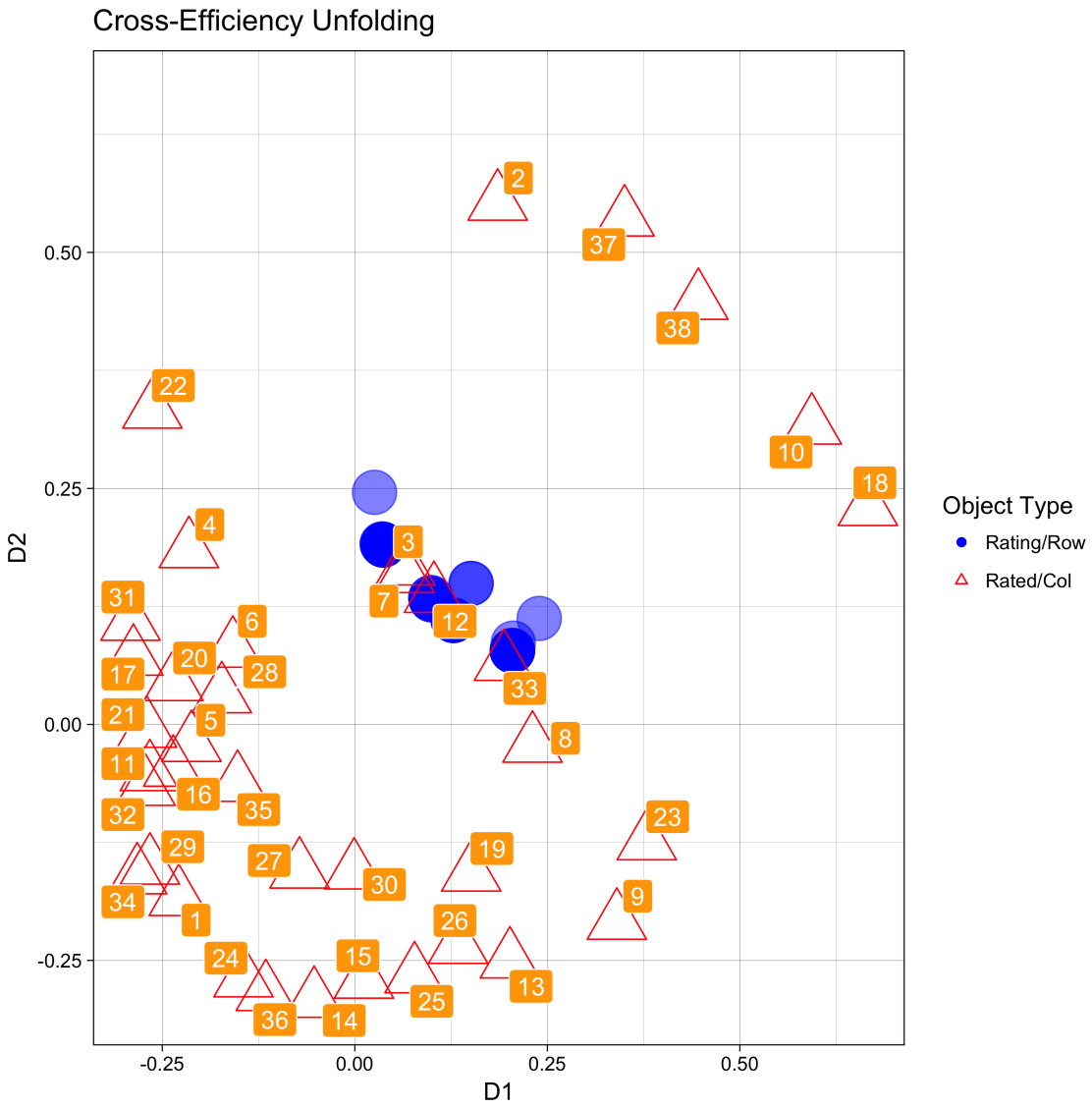

Figure 2.5: Unfolding map of benevolent cross-efficiency matrix of Table 2.6

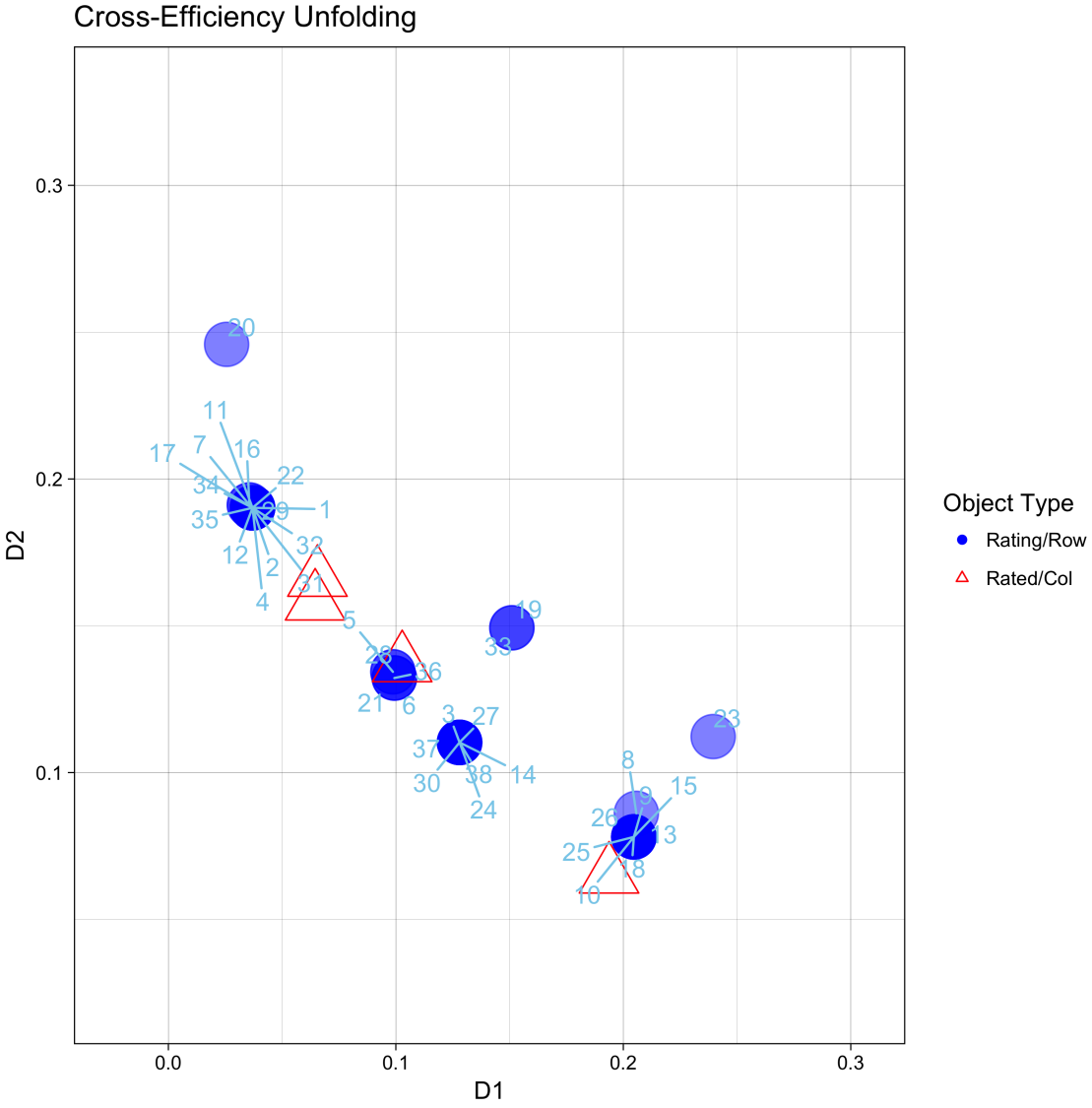

Figure 2.6: Row objects of unfolding map of benevolent cross-efficiency matrix of Table 2.6

In addition, according to Bahari and Emrouznejad (2014), DMUs 3, 8, 12, 19 and 33 are potential outliers of the dataset. This result is highly compatible with the Figure 2.5 findings, however it is noteworthy that Bahari and Emrouznejad (2014) use variable return to scale DEA, while in this study all CEMs are computed under constant return to scale assumption. At last, the stress-1 index, the badness of fit measure, for configuration of Figure 2.5 is 0.102, which is a totally acceptable degree of noise.

The second real dataset, is the famous dataset of 35 Chinese cities with 3 inputs and 3 outputs that has been used in Adler and Raveh (2008) and Costa et al. (2016), two other DEA visualization studies. The input and output levels are presented in the Table 2.7.

The CEM of Table 2.7 is not presented in this paper due to lack of space. The MDU visualization of the Table 2.7 is presented in Figure 2.7, without labels, and Figure 2.8, with labels.

The row objects of Figure 2.7 are more diverse than the row objects of Figure 2.5. According to the experience of authors, most of the unfolding maps are similar to Figure 2.7 rather than Figure 2.5. Being so diverse makes interpretation of Figure 2.7 less straightforward. While there are some column objects close to the crowd of row objects, and they may be possible outliers, we avoid the crowded parts of the map and focus our attention on the far-from-the-crowd row object on the bottom of the map.

The most telling unit in Figure 2.8 seems to be Unit 24, which is far from the crowd by both row and column aspects. This unit is similar to Unit 3 in Figure 2.3 and unit3 in Figure 2.4, a unit that has a different optimum weight distribution and thus a different structure or strategy. Moreover, the position of the column object 24, which is far from the majority of row objects, suggests that DMU 24 has a low average cross efficiency, while the relative closeness of Unit 24's row and column objects suggests that the simple efficiency is relatively high. Having high self-efficiency and low average cross-efficiency, this unit is a potential maverick according to Sexton et al. (1986) and Doyle and Green (1994). Unit 35 has the same behavior but with less intensity. These finding are compatible with the Doyle&Green Maverick Index(MI) of the Table 2.7 CEM. The Top 5 units based on the MI score are presented in Table 2.8.

Nevertheless, the badness-of-fit of configuration of Figure 2.7 is relatively high ($stress - 1 = 0.19$), which necessitates practitioners to cautiously evaluate the map, since the position of some objects may be heavily affected by noise.

Till now, we used visualization mainly as an alternative way of finding the instances, such as mavericks and outliers, that we could find with quantitative methods as well, and we validated our visual findings with the numerical methods. While visualization helps to find such cases more quickly and efficiently, a good visualization leads researchers to formulate new questions about the data, and even the well-established methods. N. J. Cox and Jones (1981) states that *displays reveal the major features of data, help in the production of ideas for further investigation*, and Ware (2012, p. 3) underlines *visualization facilitates hypothesis formation.*

By further exploration of unfolding maps such as Figure 2.7, we may question

Table 2.7: 35 major cities of China, from Adler and Raveh (2008)

DMU No.	DMU Name	Inputs			Outputs		
		Industrial Labour Force	**Working Funds**	**Investments**	**Gross Industrial Output**	**Profit and Tax**	**Retail Sales**
1	Beijing	110.22	794509	724255	2374342	680119	12790
2	Changchum	31.34	183319	101556	473369	118062	3460
3	Changsha	18.12	99307	83395	255540	50355	2652
4	Chengdu	46.86	304726	173655	734613	150853	4381
5	Chongqing	77.39	443862	210947	1037584	189878	5233
6	Dalian	37.96	282373	198278	753961	194512	3708
7	Fuzhou	16.03	96623	103560	222634	43984	2222
8	Guangzhou	50.92	389641	354879	1154147	275588	8362
9	Guiyang	17.52	101368	76476	257718	72917	1118
10	Hangzhou	34.32	212524	120028	726172	159354	4106
11	Harbin	48.2	356752	138972	672427	124508	3856
12	Hefei	17.02	95076	56690	270087	62387	1486
13	Hohot	10.15	56096	42493	127132	33069	852
14	Jinan	26.39	152034	78312	441724	109039	2441
15	Kunming	27.59	168224	112871	439756	117719	2148
16	Lanzhou	35.89	235416	107328	580669	140557	2151
17	Lhasa	0.44	1908	7394	1665	286	398
18	Nanchang	23.3	129132	42700	317158	61472	1663
19	Nanjing	42.18	269246	222623	836544	208006	3779
20	Nanning	11.33	59166	36627	176140	139399	1253
21	Ningbo	13.52	69895	72845	320516	74492	2866
22	Shanghai	206.73	1577603	959226	6743346	1880041	18316
23	Shenyang	68.62	419358	198494	1017454	195987	5072
24	Shenzhen	3.41	35478	278230	78313	11461	2778
25	Shijiazhuang	24.88	138931	68661	453445	94216	1745
26	Taiyuan	38.34	221065	170776	513907	84812	1896
27	Tianjin	98.38	628243	541587	2252611	538202	6895
28	Urumqi	15.97	94622	130771	204232	38294	1323
29	Wuhan	64.87	442813	183811	1218527	295199	5090
30	Xiamen	6.48	46821	97627	130646	37698	1109
31	Xian	46.2	294539	140906	635575	101261	3292
32	Xining	10.54	74188	45629	100509	18627	858
33	Yinchuan	5.06	30959	46014	59757	11458	524
34	Zhengzhou	28.09	144141	86791	413025	105784	2359
35	Zhuhai	1.26	12504	86457	35760	6667	1046

Table 2.8: Top maverick units based on Doyle&Green Index

Rank	**Units**	**D&G MI**
1	24	2.004
2	35	1.485
3	30	0.715
4	17	0.379
5	28	0.363

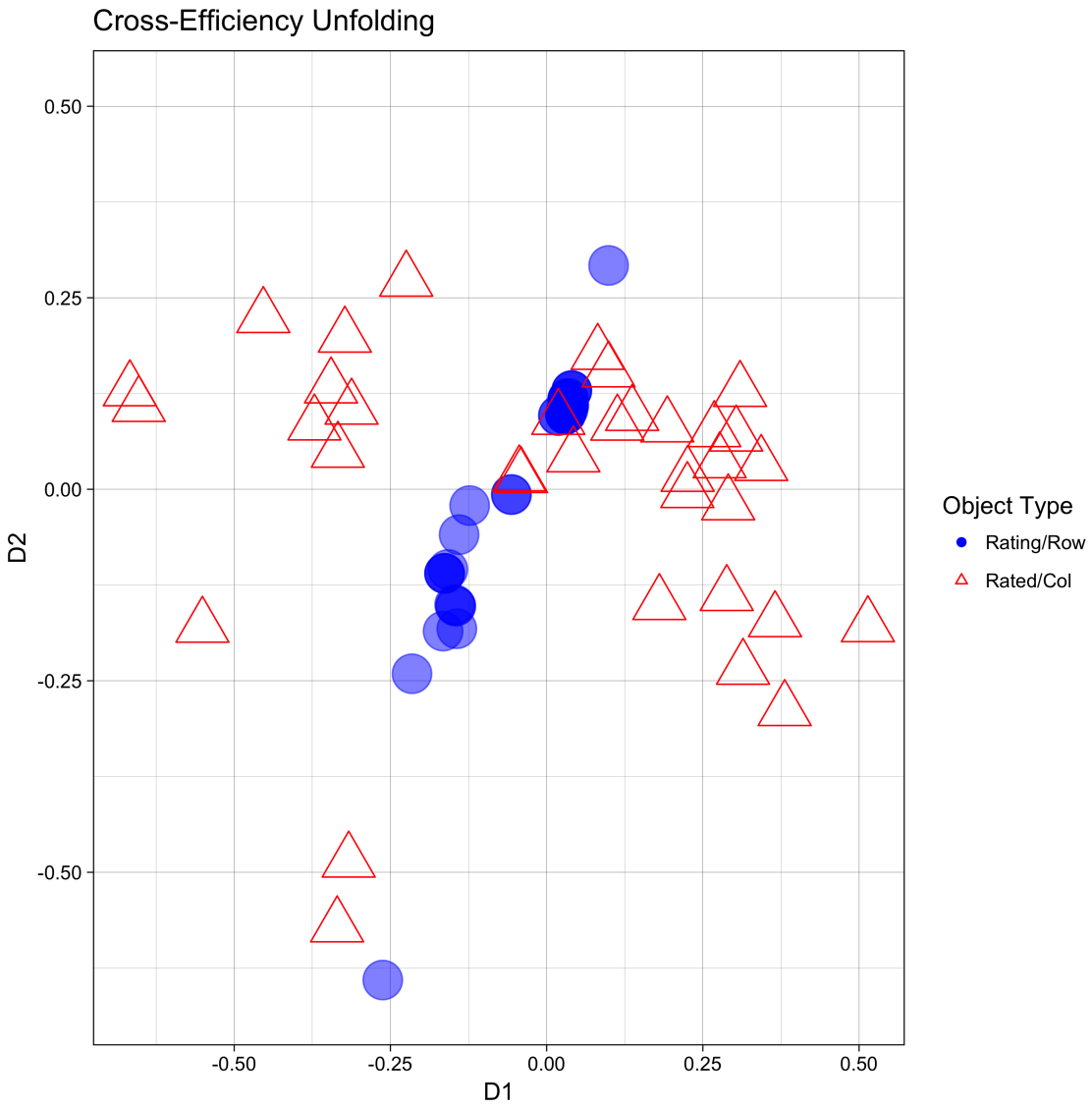

Figure 2.7: Unfolding map of benevolent cross-efficiency matrix of Table 2.6

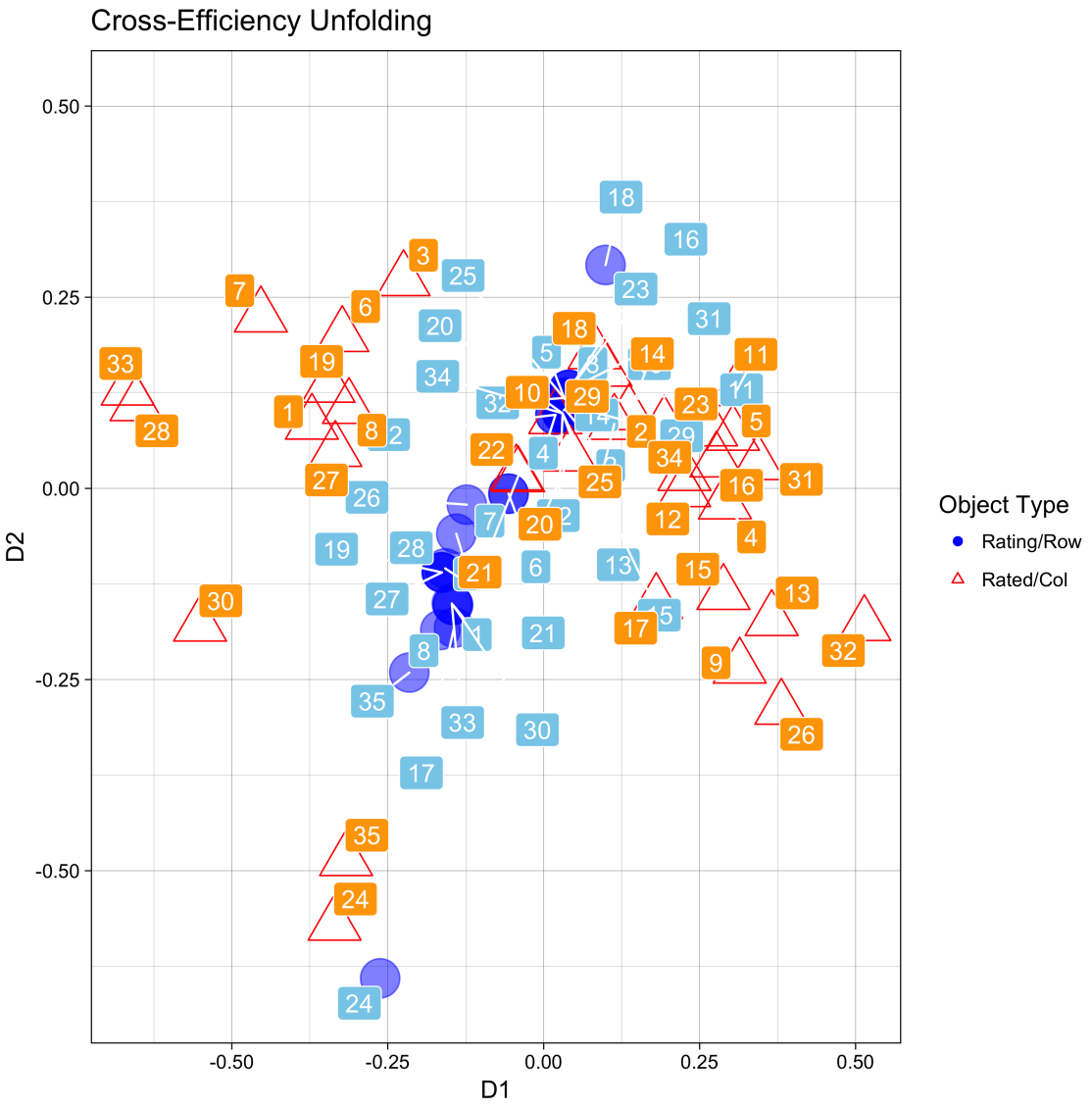

Figure 2.8: Unfolding map of benevolent cross-efficiency matrix of Table 2.6

our assumptions and understandings both about the data and about the methods. Following are two questions, one about the methodology and one about the data, regarding the findings of Figure 2.7, that demand further investigations and research:

1. What is measured by Doyle&Green Maverick Index(MI)? Maverick units are supposedly technically efficient units that are price inefficient (Sexton et al. 1986). Later, Doyle and Green (1994) portray mavericks as the units with unbalanced, unacceptable optimum weights, units with non-zero weights on a few inputs and outputs, disregarding the rest. Nevertheless, the MI measures uncommon-ness of a unit, its distance from the crowd of units, as it is seen in the Figure=2.7, while MI is supposed to measure unbalance-ness of optimum weights. The index is composed of average cross-efficiency, the average distance of a column unit from all row units, and self-efficiency, the distance of a column unit from its corresponding row unit. Therefore, MI does not consider the distance of the row unit from other row units, while this distance is the gauge of uncommon-ness of optimum weights. Moreover, it has never clearly stated which CEM should be used for MI? Changing the CEM formulation, for instance from benevolent to aggressive, changes the position of points on the unfolding map and thus MI scores.

2. What is special about DMU24 It is seen that this unit stands far from the crowd from both rating and rated aspects. Is this discordant behaviour because due to the units nature and how it performs or it is just due to error in input and output levels? Is such unit considered as an outlier which distorts the problem, and thus should be removed? Using (Wilson 1993) outlier detection the top five outliers of this dataset are 1, 8, 20, 22, and 27.

6 Conclusions

Visualization plays a key role in exploration of large datasets, not only through representation of the quantitative data in a more comprehensible manner but also through revealing the overall picture of the dataset, which otherwise would remain uncovered. Moreover, DEA, a data-oriented method, is not an exception for the possibility of benefiting from data visualization.

In this study, a method for visualization of cross-efficiency matrices (CEM) has been suggested. Any CEM includes large amount of information about both simple efficiency scores of DMUs and peer-evaluation scores of the units. Hence, the visualization of such matrix enables researchers and practitioners to explore how each unit evaluates itself as well as how each unit is evaluated by the others. The CEM, as a two-way, two-mode asymmetric matrix, can be visualized using multidimensional unfolding (MDU).

Using two artificial and two real datasets, the method has been elaborated with focus on anomaly detection, including mavericks and outliers. While CEM visualization, as a data exploratory tool, can be used for different purposes, here the emphasis has been on anomaly detection due to importance of identification such

exceptional units. Following the visual information seeking mantra Shneiderman (1996), i.e. "Overview first, Zoom and filter, then details-on-demand", a researcher can use the current method for the "overview first", then choosing the interesting points in order to "zoom and filter" and possible further investigations on details.

All that said, the current visualization framework can be improved in several aspects. First, it is well known that benevolent formulation of CEMs cannot find the most benevolent optimum weights, and thus the result is not unique. Hence, without any change to the current procedure, the benevolent CEM can be replaced by other formulations according the study goals, as far as the new formulation is justified. In addition, cross-efficiency is traditionally based on CRS assumption, which makes the evaluation results less realistic and justifiable in most cases. This problem can be resolved by the VRS version of cross-evaluation, suggested by Lim and Zhu (2015). Moreover, in the age of interactive visualization, using still graphs for representation of multidimensional data is neither appropriate nor sufficient. Hence, using new visualization tools to make the CEM unfolding configuration interactive benefits from the latest tools as input to produce the most comprehensible and informative output.

Finally, the findings of the suggestion data exploratory tool should be considered as suspicious units which require further investigation from other perspectives, and by other means. Hence, the CEM-MDU visualization method proposes a way to gain insight into the data for a graphical detection work, since *"Exploratory data analysis is detective work–in the purest sense–finding and revealing the clues. Confirmatory data analysis ... goes further, assessing the strengths of the evidence."* (Tukey 1977, p. 21)

Appendix A

In this appendix, three points regarding the essay 2 are clarified. Since the essay is too long in the current format, and in order to keep the main body incisive, these further elaborations are presented as an appendix.

On The Chosen Benevolent Approach to CEM

Multiplicity of optimum weights of efficient DMUs in DEA problems is a well-known problem. As far as we do not want to use the weights directly, no problem arises due to this phenomenon. In other words, if we want to use the weights for calculation of efficiency score, then the multiplicity of weights won't cause any problem. However, this is not the case in cross-efficiency and cross- evaluation. In cross-evaluation, different set of optimum weights of an efficient unit, will possibly yield different cross-efficiency scores for other units, i.e. the DMUs that are evaluated by the optimum weights of that efficient unit. Hence, with myriad number of different CEMs a problem would have. In order to rectify this issue, and based on some justifications such as the competitive nature of the relations of DMUs in DEA problems, two specific approaches to CEM were proposed for the first time by Sexton et al. (1986). The authors suggested to use a goal programming formulation in the cross- efficiency formulation, by which one of the alternate optima is chosen in a non-arbitrary fashion. To do so, a secondary goal is defined that either ideally chooses the optimum alternate which minimizes the other DMUs cross-efficiency scores (aggressive formulation) or maximizes the other DMUs cross-efficiency scores (benevolent formulation). Nevertheless, in practice these two attractive approaches turns out to be problematic due to their non-linear fractional optimization programming. In order to alleviate this issue, Sexton et al. (1986) and later Doyle and Green (1994) suggest "adequate surrogate" formulations. Nonetheless, these formulations are just "surrogates" and approximations. To put it differently, the suggested surrogates do not necessarily find the optimum answers for the secondary goals, either it is the minimization of other cross-efficiency scores, or the maximization. Having all said, in this research the benevolent approach has been used, and the rationale is twofold. Firstly, the focus of the current work is on visualization of CEM, regardless of the formulation of the CEM. Using a more modern formulation might cause distraction, as the benevolent approach is the oldest and possibly the most established formulation of CEM. Secondly, the CEM visualization is supposed to be used as a tool for evaluation of maverick units, sort of outliers with incongruent behavior. In order to find such units, the DMUs should be examined under the most possible congruent behaviour. As the behaviour of the DMUs are determined by their optimum weights, such congruent behaviour is achieved only under the optimum weights that are the most similar to other DMU's optimum weights. If in such scenario, a DMU is still showing dissident behaviour, then such DMU is possibly different by nature, and not by the choice of a specific optimum weight set. Nonetheless, the current formulation can be improved or replaced by other up-to-date CEM formulations. Ideally, the new formulation should be a benevolent approach

without any approximation or surrogates, such that it finds the most benevolent optimum weights to satisfy the maverick-detection goal of this thesis. However, there is no such formulation in the literature to the best of our knowledge. Regardless of this fact, there are several interesting formulations that can be used to generate the CEM. Liang et al. (2008a) suggest game cross-efficiency formulations where each DMU tries to maximize its own efficiency score in a way not to deteriorate other DMUs' cross-efficiency. Liang et al. (2008b) suggested alternative secondary goal similar to the benevolent approach of Sexton et al. (1986) with a different formulation to maximize the worst performer. Ramón et al. (2010) suggest a CEM formulation which supposedly avoid unrealistic weights, i.e. profiles of weights with zeros. This approach is also attractive in the framework of this thesis, since the maverick units are often-times associated with the profiles with weights of zeros. Relaxing the benevolent idea, there are several other CEM approaches that should be considered in different circumstances than finding maverick units. For instance, Cook and Zhu (2014) present "Maximum Log Cross-efficiency" formulation which tends to find the unique optimum weights, in contrast to the formulations that do not try to find unique optima. While all these formulations seem interesting enough to be examined in the visualization framework, there is no open implementation of their algorithms, and thus the any user has to program their algorithms from scratch in order to do a reproducible and verifiable research. This is a burdensome task, that should be done in the future steps following this thesis.

On the Effect of the Secondary Goal on CEM

Change of the secondary goal will possibly change the CEM scores, and thus the final configuration map. The CEM scores in the final map are ideally reflected in intra-item distances in reverse fashion, i.e. the higher the cross-efficiency score between DMUi and DMUj, the lower the distances between corresponding items of these two DMUs on the map. This is because the "closeness" on the map is conventionally interpreted as "similarity", and when a DMU can achieve high efficiency score using another DMUs optimum weights, these two are deemed similar. In order to predict the effect of the secondary goal choice on the visualization map, one has to predict the possible changes in the CEM scores by change of the secondary goal. However, there is no straightforward relationship between the outcome of all different secondary goals in general. For instance, Liang et al. (2008a) suggested three different secondary goals for CEM, but without any definite conclusion whether cross-efficiency scores of these formulations are "always" greater or less than the counterparts. Nevertheless, by restricting the scope of this question to the two benevolent and aggressive approaches of Doyle and Green (1994), which are used in this thesis, the answer to this question would be easier, since there is a theoretical relationship between the CEM scores of these two formulations. The secondary goal of benevolent formulation always try to maximize the average cross-efficiency scores of peers, while the aggressive formulation always try to minimize the average cross- efficiency of other DMUs. Hence, overall the CEM scores of the benevolent are greater than the CEM scores of the aggressive approach. The greater cross-efficiency scores ideally are presented in shorter intra-items distances on the map. In other words, the DMU items on the

visualization map of benevolent CEM would be closer to each other in comparison to the aggressive CEM configuration. Below the two figures are presented to depict such relationships. The data is from Wang and Chin (2010)

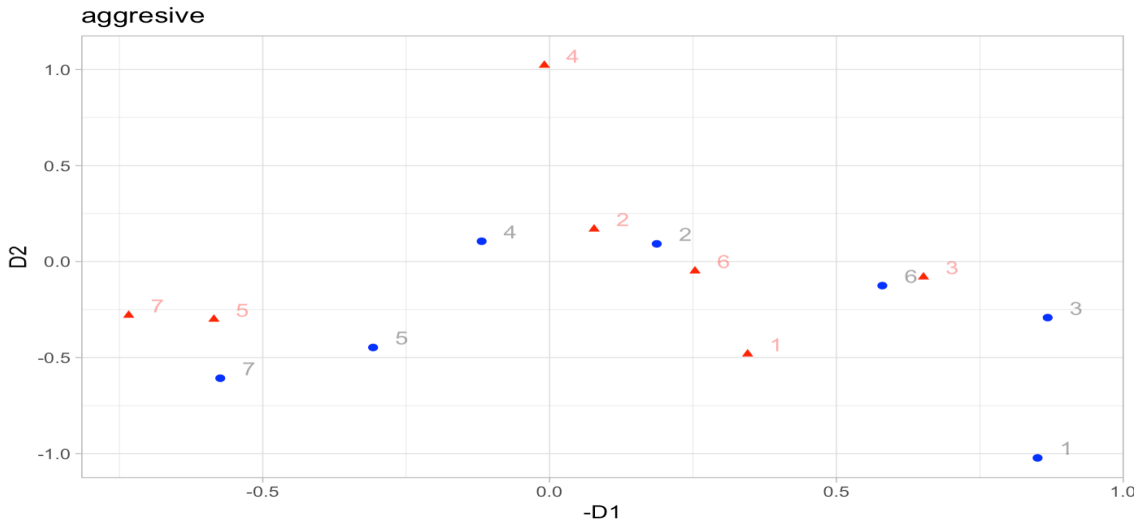

Figure 2.9: In aggressive CEM, the units tend to lower peers' average cross-efficiency, therefore they reside far from each other, comparing to the benevolent formulation.

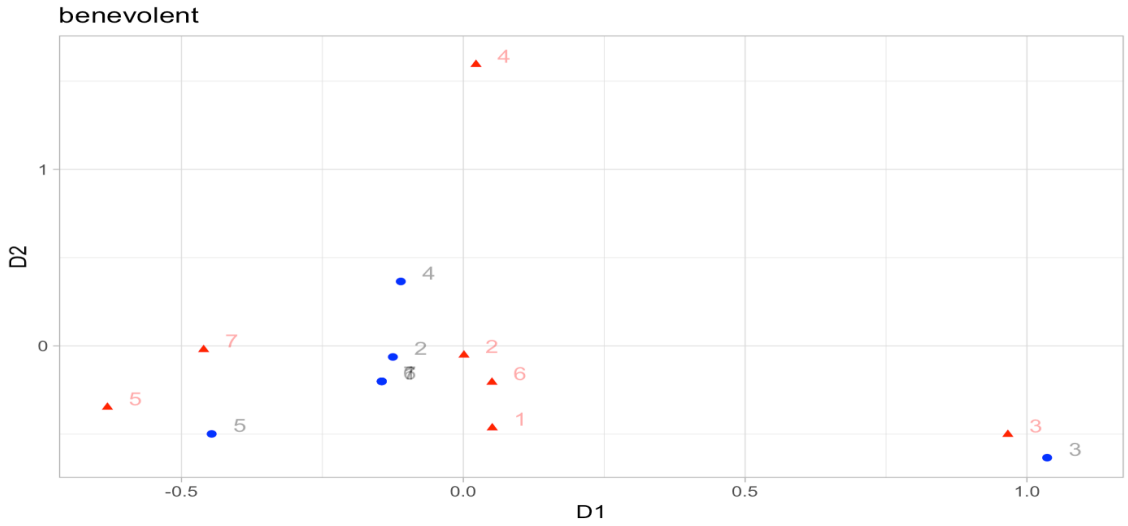

Figure 2.10: It can be seen that regardless of some items such as DMU4 and DMU3, how benevolent approach has caused the items to locate in a smaller area, and closer to each-other.

On the Reductionist vs Holistic Approaches

In order to understand cross-efficiency matrix(CEM), two steps are conventionally used: First, decomposition, and then, statistical summarization. In the decomposition, a CEM is broken down into its components, i.e. columns which are corresponding to rated DMUs, and in the statistical summarization step, each component is evaluated by means of "arithmetic average" or similar measures. In doing so, there are two fundamental assumptions taken for granted. First, there is no influential and important relationship between components, and second, there is no harm in condensing set of cross-efficiency scores of a column, into a single number. In both steps, some "details" are discarded, without the least effort to retain them, under this implicit assumption that those details are not decisive. The idea of possibility of comprehensively knowing a whole by knowing its parts, is called "reductionism" in the philosophy of science.Honderich (2005)(p793)

Subsequently, the approaches that are taken towards CEM are called "analytical", since they tend to analysis the under-study phenomenon.1 Lars Skyttner (2005)(p15), formulates the reductionism as the conceptual foundation of analytical method that works in three steps:

1. Dissect conceptually/physically.
2. Learn the properties/behaviour of the separate parts.
3. From the properties of the parts, deduce the properties/behaviour of the whole.

Nevertheless, not all the phenomena are possible to be understood with reductionism. The reason lies in the fact that the relations between components of a system play an important role in the behaviour of the system, and by reducing a system to its parts, the relations are neglected. In other words, in the presence of all components and their relationships, a system tends to show some collective properties, called "emergent properties", that do not appear in any decomposition level. In the words of (Ackoff 1979): "...systems are wholes which lose their essential properties when taken apart. Therefore, they are wholes that cannot be understood by analysis." Hence, my effort in this thesis is to take a "holistic/system thinking/expansionist" approach to know CEMs, as a system of DMUs. Doing so is not due to the belief of uselessness of analytical methods, but due to possibility of improving our understanding of CEMs by changing the assessment tools, as Meadows (2008)(p6, p83-84) articulates: "I don't think the systems way of seeing is better than the reductionist way of thinking. I think it's complementary, and therefore revealing. [...] the reductionist dissection of regular science teaches us a lot. However, one should not lose sight of the important relationships that bind each subsystem to the others and to the higher levels of the hierarchy, or one will be in for surprises." 1 The neglect of the relationships between DMUs in the CEM analysis is to such level that by shuffling cross-efficiency scores in each column, neither the "average cross-efficiency score", nor the "maverick index" changes.

In practice, I chose high-dimensional information visualization as the holistic method of knowing CEMs. Using High-D InfoVis, I try to preserve the details,

including the relations between DMUs, as much as possible. This possibility is determined by the dimension-reduction technique, such as principal component analysis or multidimensional scaling, as they tend to lose some information as the compensation of presentation of the data in the low dimensional space. More precisely, the visualization of CEMs as suggested in this thesis, represents all CEM DMUs as visual objects whose distances are ideally reflections of their corresponding cross-efficiency scores. Therefore, the current method's endeavor is to provide us "the big picture" that is composed of all system components and their relations, something that mentioned analytical methods do not try to preserve and present. At last, I should underscore that not all information visualization approaches are "holistic". For instance, it is possible to visualize one variable, such as one input in a DEA problem or a single column in a CEM, while disregarding all other variables and the relations. This is why using "high-D InfoVis" is critical, rather than any "InfoVis" approach.

References:

References

Ackoff, Russell L (1979). "The future of operational research is past". In: *Journal of the operational research society* 30.2, pp. 93–104.

Adler, Nicole, Lea Friedman, and Zilla Sinuany-Stern (2002). "Review of ranking methods in the data envelopment analysis context". In: *European journal of operational research* 140.2, pp. 249–265.

Adler, Nicole and Adi Raveh (2008). "Presenting DEA graphically". In: *Omega* 36.5, pp. 715–729.

Aizemberg, Luiz, Marcos Costa Roboredo, Thiago Graça Ramos, João Carlos CB Soares de Mello, Lidia Angulo Meza, and Alessandro Martins Alves (2014). "Measuring the NBA Teams' Cross-Efficiency by DEA Game". In: *American Journal of Operations Research* 4.03, p. 101.

Akçay, Alp Eren, Gürdal Ertek, and Gülçin Büyüközkan (2012). "Analyzing the solutions of DEA through information visualization and data mining techniques: SmartDEA framework". In: *Expert systems with applications* 39.9, pp. 7763–7775.

Anscombe, Francis J (1973). "Graphs in statistical analysis". In: *The American Statistician* 27.1, pp. 17–21.

Aoki, Shingo, Kotaro Toyozumi, and Hiroshi Tsuji (2007). "Visualizing method for data envelopment analysis". In: *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. IEEE, pp. 474–479.

Appa, Gautam, N Argyris, and H Paul Williams (2006). "A methodology for cross-evaluation in DEA". In: URL: `http://eprints.lse.ac.uk/22714/1/06081.pdf`.

Appa, Gautam, Carlos A Bana e Costa, Manuel P Chagas, Fernando C Ferreira, and João O Soares (2010). "DEA in X-factor evaluation for the Brazilian Electricity Distribution Industry". In: *London School of Economics*.

Bahari, Ali Reza and Ali Emrouznejad (2014). "Influential DMUs and outlier detection in data envelopment analysis with an application to health care". In: *Annals of Operations Research* 223.1, pp. 95–108.

Baker, RC and Srinivas Talluri (1997). "A closer look at the use of data envelopment analysis for technology selection". In: *Computers & Industrial Engineering* 32.1, pp. 101–108.

Banker, Rajiv D, Abraham Charnes, and William Wager Cooper (1984). "Some models for estimating technical and scale inefficiencies in data envelopment analysis". In: *Management science* 30.9, pp. 1078–1092.

Belton, Valerie and Stephen P Vickers (1993). "Demystifying DEA-a visual interactive approach based on multiple criteria analysis". In: *Journal of the Operational research Society*, pp. 883–896.

Berkelaar, Michel et al. (2015). *Package 'lpSolve'*.

Bogetoft, Peter and Lars Otto (2015). *Benchmarking with DEA and SFA*. R package version 0.26.

Borg, Ingwer and Patrick JF Groenen (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Borg, Ingwer, Patrick JF Groenen, and Patrick Mair (2012). *Applied multidimensional scaling*. Springer Science & Business Media, p. 3.

Carboni, Oliviero A and Paolo Russu (2015). "Assessing regional wellbeing in Italy: An application of Malmquist–DEA and self-organizing map neural clustering". In: *Social indicators research* 122.3, pp. 677–700.

Chai, Song, Yubai Li, Jian Wang, and Chang Wu (2013). "A genetic algorithm for task scheduling on NoC using FDH cross efficiency". In: *Mathematical Problems in Engineering* 2013.

Charnes, Abraham, William W Cooper, and Edwardo Rhodes (1978). "Measuring the efficiency of decision making units". In: *European journal of operational research* 2.6, pp. 429–444.

Chen, Chun-houh, Wolfgang Karl Härdle, and Antony Unwin (2007). *Handbook of data visualization*. Springer Science & Business Media.

Chen, Tser-yieth (2002). "An assessment of technical efficiency and cross-efficiency in Taiwan's electricity distribution sector". In: *European Journal of Operational Research* 137.2, pp. 421–433.

Churilov, Leonid and A Flitman (2006). "Towards fair ranking of Olympics achievements: The case of Sydney 2000". In: *Computers & Operations Research* 33.7, pp. 2057–2082.

Cleveland, William S and William S Cleveland (1985). *The elements of graphing data*. Wadsworth Advanced Books and Software Monterey, CA.

Cook, Wade D and Joe Zhu (2014). "DEA Cobb–Douglas frontier and cross-efficiency". In: *Journal of the Operational Research Society* 65.2, pp. 265–268.

Cooper, William W, Lawrence M Seiford, and Joe Zhu (2011). *Handbook on data envelopment analysis*. Vol. 164. Springer Science & Business Media.

Costa, Carlos A Bana e, João Carlos CB Soares de Mello, and Lidia Angulo Meza (2016). "A new approach to the bi-dimensional representation of the DEA efficient

frontier with multiple inputs and outputs". In: *European Journal of Operational Research* 255.1, pp. 175–186.

Cox, Nicholas J and Kelvyn Jones (1981). "Exploratory data analysis". In: *Quantitative geography: A British view*, pp. 135–143.

Cox, Trevor F and Michael AA Cox (2000). *Multidimensional scaling*. CRC press.

De Leeuw, J (2005). *Multidimensional unfolding. Entry in the encyclopedia of statistics in behavioural science*.

De Leeuw, Jan and Patrick Mair (2011). "Multidimensional scaling using majorization: SMACOF in R". In: *Department of Statistics, UCLA*.

Desai, Anand and Lawrence C Walters (1991). "Graphical presentations of data envelopment analyses: management implications from parallel axes representations". In: *Decision Sciences* 22.2, pp. 335–353.

Doyle, John and Rodney Green (1994). "Efficiency and cross-efficiency in DEA: Derivations, meanings and uses". In: *Journal of the operational research society* 45.5, pp. 567–578.

Falagario, Marco, Fabio Sciancalepore, Nicola Costantino, and Roberto Pietroforte (2012). "Using a DEA-cross efficiency approach in public procurement tenders". In: *European Journal of Operational Research* 218.2, pp. 523–529.

Forrest W., Young (1987). *Multidimensional Scaling: History, Theory, and Applications*. Lawrence Erlbaum Associates. ISBN: 0898596637.

Førsund, Finn R, Sverre AC Kittelsen, and Vladimir E Krivonozhko (2009). "Farrell revisited–Visualizing properties of DEA production frontiers". In: *Journal of the Operational Research Society* 60.11, pp. 1535–1545.

Fu, Yan, Dongdong Li, and Ning Li (2011). "Hotel Performance Evaluation Based on Cross-efficiency DEA Models". In: *Management and Service Science (MASS), 2011 International Conference on*. IEEE, pp. 1–4.

Fumero, Francesca (2004). "Multiple solutions identification in data envelopment analysis". In: *Central European Journal of Operations Research* 12.3, p. 307.

Guttman, Louis (1968). "A general nonmetric technique for finding the smallest coordinate space for a configuration of points". In: *Psychometrika* 33.4, pp. 469–506.

Hackman, Steven T, Ury Passy, and Loren K Platzman (1994). "Explicit representation of the two-dimensional section of a production possibility set". In: *Journal of Productivity Analysis* 5.2, pp. 161–170.

Hatami-Marbini, Adel, Madjid Tavana, and Ali Emrouznejad (2012). "Productivity growth and efficiency measurements in fuzzy environments with an application to health care". In: *International Journal of Fuzzy System Applications (IJFSA)* 2.2, pp. 1–35.

Honda, Katsuhiro, Shingo Aoki, Akira Notsu, and Hidetomo Ichihashi (2010). "Visual Assessment of DEA Efficiencies by Fuzzy PCA with Variable Selection". In: *SCIS & ISIS SCIS & ISIS 2010*. Japan Society for Fuzzy Theory and Intelligent Informatics, pp. 23–27.

Honderich, Ted (2005). *The Oxford companion to philosophy*. OUP Oxford.

Inoue, Kazushige, Takeo Ichinotsubo, and Shingo Aoki (2011). "DEA based hierarchical structure evaluation and visualization method". In: *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, pp. 1701–1704.

Inselberg, Alfred and Bernard Dimsdale (1987). "Parallel coordinates for visualizing multi-dimensional geometry". In: *Computer Graphics 1987*. Springer, pp. 25–44.

Kohonen, Teuvo (2001). *Self-Organising Maps*. 3rd ed. Berlin, Heidelberg, Newyork: Springer-Verlag New York. Inc.

Kruskal, Joseph B and Myron Wish (1978). *Multidimensional scaling*. Vol. 11. Sage.

Li, Ning (2008). "Real estate cross-efficiency measurement based on peer appraisal DEA model and method in main cities of China". In: *Management Science and Engineering, 2008. ICMSE 2008. 15th Annual Conference Proceedings., International Conference on*. IEEE, pp. 1667–1673.

Liang, Liang, Jie Wu, Wade D Cook, and Joe Zhu (2008a). "Alternative secondary goals in DEA cross-efficiency evaluation". In: *International Journal of Production Economics* 113.2, pp. 1025–1030.

Liang, Liang, Jie Wu, Wade D Cook, and Joe Zhu (2008b). "The DEA game cross-efficiency model and its Nash equilibrium". In: *Operations Research* 56.5, pp. 1278–1288.

Lim, Sungmook (2012). "Minimax and maximin formulations of cross-efficiency in DEA". In: *Computers & Industrial Engineering* 62.3, pp. 726–731.

Lim, Sungmook, Kwang Wuk Oh, and Joe Zhu (2014). "Use of DEA cross-efficiency evaluation in portfolio selection: An application to Korean stock market". In: *European Journal of Operational Research* 236.1, pp. 361–368.

Lim, Sungmook and Joe Zhu (2015). "DEA cross-efficiency evaluation under variable returns to scale". In: *Journal of the Operational Research Society* 66.3, pp. 476–487.

Liu, John S, Louis YY Lu, and Wen-Min Lu (2016). "Research fronts in data envelopment analysis". In: *Omega* 58, pp. 33–45.

El-Mahgary, Sami and Risto Lahdelma (1995). "Data envelopment analysis: visualizing the results". In: *European Journal of Operational Research* 83.3, pp. 700–710.

Mair, Patrick, Jan De Leeuw, and Marcus Wurzer (2015). "Multidimensional unfolding". In: *Wiley StatsRef: Statistics Reference Online*.

Meadows, Donella H (2008). *Thinking in systems: A primer*. chelsea green publishing.

Mello, João Carlos Correia Baptista Soares de, Eliane Gonçalves Gomes, Lidia Angulo Meza, Luiz Biondi Neto, Sergio de Zen, Thiago Bernardino de Carvalho, and Urbano Gomes Pinto de Abreu (2012). *Ex-Post Clustering of Brazilian Beef Cattle Farms Using Soms and Cross-Evaluation Dea Models*. INTECH Open Access Publisher.

Mello, Soares de, João Carlos Correia Baptista, Lidia Angulo Meza, and Brenda Branco da Silva (2008). "Some rankings for the Athens Olympic Games using DEA models with a constant input". In: *Investigação Operacional* 28.1, pp. 77–89.

Oral, Muhittin, Ossama Kettani, and Pascal Lang (1991). "A methodology for collective evaluation and selection of industrial R&D projects". In: *Management Science* 37.7, pp. 871–885.

Porembski, Marcus, Kristina Breitenstein, and Paul Alpar (2005). "Visualizing efficiency and reference relations in data envelopment analysis with an application to the branches of a German bank". In: *Journal of Productivity Analysis* 23.2, pp. 203–221.

Ramón, Nuria, José L Ruiz, and Inmaculada Sirvent (2010). "On the choice of weights profiles in cross-efficiency evaluations". In: *European Journal of Operational Research* 207.3, pp. 1564–1572.

Ruiz, José L and Inmaculada Sirvent (2016). "Ranking Decision Making Units: The Cross-Efficiency Evaluation". In: *Handbook of Operations Analytics Using Data Envelopment Analysis*. Springer, pp. 1–29.

Sammon, John W (1969). "A nonlinear mapping for data structure analysis". In: *IEEE Transactions on computers* 100.5, pp. 401–409.

Schiffman, Susan S, Forrest W Young, and M Lance Reynolds (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*.

Serrano-Cinca, Carlos, Yolanda Fuertes-Callén, and Cecilio Mar-Molinero (2005). "Measuring DEA efficiency in Internet companies". In: *Decision Support Systems* 38.4, pp. 557–573.

Sexton, Thomas R, Richard H Silkman, and Andrew J Hogan (1986). "Data envelopment analysis: Critique and extensions". In: *New Directions for Evaluation* 1986.32, pp. 73–105.

Shneiderman, Ben (1996). "The eyes have it: A task by data type taxonomy for information visualizations". In: *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, pp. 336–343.

Skyttner, Lars (2005). *General systems theory: Problems, perspectives, practice*. World scientific.

Slowikowski, Kamil (n.d.). "ggrepel: Repulsive Text and Label Geoms for 'ggplot2', 2016". In: *R package version 0.5*.

Soukup, Tom and Ian Davidson (2002). *Visual data mining: Techniques and tools for data visualization and mining*. John Wiley & Sons.

Talluri, Srinivas, Mary M Whiteside, and Scott J Seipel (2000). "A nonparametric stochastic procedure for FMS evaluation". In: *European Journal of Operational Research* 124.3, pp. 529–538.

Team, R Core (2016). *R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014*. URL: `https://www.R-project.org/`.

Tukey, John Wilder (1977). *Exploratory Data Analysis*. Addison-Wesley. ISBN: 0201076160.

Wang, Ying-Ming and Kwai-Sang Chin (2010). "A neutral DEA model for cross-efficiency evaluation and its extension". In: *Expert Systems with Applications* 37.5, pp. 3666–3675.

Ware, Colin (2012). *Information visualization: perception for design*. 3rd ed. Elsevier.

Weber, Charles A and Anand Desai (1996). "Determination of paths to vendor market efficiency using parallel coordinates representation: a negotiation tool for buyers". In: *European journal of operational research* 90.1, pp. 142–155.

Wickham, Hadley (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wilson, Paul W (1993). "Detecting outliers in deterministic nonparametric frontier models with multiple outputs". In: *Journal of Business & Economic Statistics* 11.3, pp. 319–323.

Wong, Y-HB and JE Beasley (1990). "Restricting weight flexibility in data envelopment analysis". In: *Journal of the Operational Research Society*, pp. 829–835.

Wu, Jie, Liang Liang, and Feng Yang (2009). "Achievement and benchmarking of countries at the Summer Olympics using cross efficiency evaluation method". In: *European Journal of Operational Research* 197.2, pp. 722–730.

Zhu, Joe (2001). "Super-efficiency and DEA sensitivity analysis". In: *European Journal of operational research* 129.2, pp. 443–455.

Article 3

Mavericks Revisited: A New Index to Identify Maverick Units in Data Envelopment Analysis

Mavericks Revisited: A New Index to Identify Maverick Units in Data Envelopment Analysis

Abstract

In recent times data envelopment analysis (DEA) has been extensively used in order to assess efficiency levels in organisations. DEA has many advantages: it can deal with several inputs and outputs; inputs and outputs do not have to be measured in the same units; there is no need to specify a production function. But DEA has shortcomings. The technique is based on multiple comparisons, and the assessed units can choose the basis of the comparison by setting a set of weights. Unfortunately, some weights can take zero values with the consequence that a particular input or output is excluded from the comparison. Therefore, some units can achieve technical efficiency while being price inefficient.

In order to tackle some DEA shortcomings, including the problem of technical and price efficiency, cross-efficiency method was suggested about three decades ago. As one of its goals, the method tries to identify the problematic units which achieve their efficiency in an unacceptable way. These units are called mavericks in the literature, and in order to identify them some maverick indexes have been proposed. However, the literature does not agree on exact characteristics of maverick units, and the commonly used maverick identification methods do not accurately deliver what they are supposed to deliver.

The purpose of this paper is to address these issues in order to improve the highlighted drawbacks in both theoretical and practical aspects of maverick literature. To do so, a new maverick detection method is suggested. From theoretical perspective, the method is based on some adjustments and refinements of maverick assumptions and definitions, and from the practical perspective, it is based on multidimensional unfolding of cross-efficiency matrices in high dimensional space. The paper starts with a critical appraisal of the maverick literature, following by some refinement to address the drawbacks. Afterwards, the new maverick index is explained, and its application on a real data set is presented.

Keywords: Data envelopment analysis, cross-efficiency, cross-evaluation, Maverick, Multidimensional unfolding, Data visualization, Anomaly Detection

1 Introduction

Data envelopment analysis is a non-parametric method in order to measure relative efficiency of a set of decision making units (DMUs). To do so, the method uses linear

programming, and relevant inputs and outputs levels of the DMUs. While a DMU has a generic, flexible and broad definition (Cooper et al. 2011) , it is usually an entity, such as an organization, which converts input(s) to output(s).

In order to address some shortcoming of the initial DEA model presented by Charnes et al. (1978), Sexton et al. (1986) suggested the concept and method of cross-efficiency evaluation. More specifically, cross-efficiency was suggested to address the issue of technical efficiency and price inefficiency.[1] According to the authors, while DEA is "a very powerful data analysis technique ... [it is] not without its shortcomings." The authors found DEA a very powerful tool to assess technical efficiency of the DMUs, but not equivalently capable of evaluation of price efficiency of the units. Thus, a unit can be technically efficient while price inefficient. Alternatively, a unit can adapt unacceptable weights, shadow prices, and achieve technical efficiency.

In order to overcome the problem, Sexton et al. (1986) suggested two approaches: weight restrictions and cross-efficiency. Cross-efficiency was proposed to highlight the units which have adopt uncommon optimum weights considering the range of optimum weights chosen by the majority of the DMUs in the dataset. These highlighted units are the ones that have very high self-appraisal, and simultaneously very low peer-appraisal. However, such a DMU, found through cross-efficiency method, should be treated cautiously, since "it is not possible to say with certainty that this DMU is price inefficient, the fact that it uses input mix so different from the majority of other DMUs, warrants some attention." (Sexton et al. 1986) The idea of cross-evaluation was expanded significantly by Doyle and Green (1994), since the authors found cross-efficiency a more justified approach to deal with unacceptable weights, comparing to imposed weight restrictions approach. Moreover, the authors coined the term of "maverick" to call the problematic units, discussed by Sexton et al. (1986), and suggested a maverick index (MI) in order to measure the maverick-ness of the units. Ever since, Doyle and Green MI has been used in the literature in order to detect these problematic units which supposedly are specialist units, i.e. units with extremely focused production strategy, and with technically highly efficient units through unacceptable weights.

While more than twenty years ago, cross-efficiency was called "a neglected aspect of data envelopment analysis" by (Doyle and Green 1994), the concept has remained relatively neglected until the recent attention. According to a citation network cluster analysis on DEA papers published between 2000 and 2014, (**liu2016**) state that "cross-efficiency and ranking" is one of the top four research fronts in DEA. Therefore, it is fair to call cross-efficiency a late-bloomer subject in DEA. This paper suggests a new maverick index. Having found some caveats and shortcoming in the maverick operational definition and maverick detection methods literature, the author proposes a new maverick index, based on a minimal definition of maverick units. The new maverick index stems from cross-efficiency in general, and more specifically it is rooted in visualization of cross-efficiency matrices. The new index does not suffer from the hidden problems of well-established Doyle and Green MI (Doyle and Green 1994).

[1] For explanation of the price and technical efficiency see Farrell (1957)

The rest of the paper is structured in three sections. In the next section the maverick literature is reviewed and critically evaluated in order to highlight the caveats and shortcomings. Afterwards, the new maverick index is explained through a fabricated data set, and then applied to a real dataset. At last, the summary of the current study is presented in addition to the improvement aspects, i.e. future possible research topics.[2]

2 What is a Maverick? Maverick definition through literature

As stated before, mavericks are the units with technical efficiency and price inefficiency. However, this definition has not seemed operational enough to researchers, and thus efforts have been made in order to characterize these units such that the characteristics can be measured in DEA. In this subsection, such efforts are reviewed.[3]

The concept of the maverick units in DEA appeared in Sexton et al. (1986) for the first time, as a by-product of devising cross-evaluation method. The authors characterize such units as the units which "operate far from the crowd", with perfect self-appraisal ("often perfectly efficient") and simultaneously low peer-appraisal. Moreover, these units have drastically different optimum weight distribution comparing to other units'. Nevertheless, the "maverick" term was coined later on by Doyle and Green (1994)

Doyle and Green (1994) portray mavericks, "in their extreme case", as 100 efficient units that "achieve efficiency by weighting a single input and a single output: the rest are accorded to zero weights." Therefore, these units are "best at one thing (but nevermind the rest)". However, this distribution of inputs and outputs' weights is not acceptable, and the unit becomes like "a car with the best engine but no wheels". Moreover, the mavericks are those which "enjoy the greatest relative increment when shifting from peer-appraisal to self-appraisal". At last, maverick units generally are "rejected as a part of the paragon set" while they achieve 100 efficiency. Additionally, Doyle and Green (1994) suggest the "Maverick index" or MI in order to detect the maverick units. It is noteworthy that the maverick definition of Doyle and Green (1994) is for defining the mavericks in "their extreme case", and there is no clue about non-extreme cases and the threshold of this extremity. The more recent papers in the literature have used different versions of the above definitions with few additions and subtractions. Therefore, they are mentioned very briefly in the following paragraphs. Baker and Talluri (1997) use the "false positive" term in order to label the units which have "several low efficiencies" in their peer appraisal set. In contrast to "false positives", there are "good overall performers" which have "several

[2]All the analysis has been done in R statistical software(Team 2016) through the programs written by the author.

[3]Sexton et al. (1986) suggest two approaches to deal with the problem: Weight restriction and cross-efficiency. In the first approach, there is no need for further definition of mavericks, hence such units cannot emerge. In contrast, cross-efficiency lets such units emerge, and then compromise their unacceptable efficiency through peer-appraisals.

high efficiencies". The false positive units are essentially identical to maverick units of Sexton et al. (1986) and Doyle and Green (1994), specially when one considers the false positive index, suggested detection technique of false positive units, and compares it to the Maverick Index (MI). (see the section 2.1) (O'Neill 1998) defines the maverick units as technically efficient units, with zero weights in all but one input and output variables. They also locate on the extreme portions of the best-practice frontier. Besides, O'Neill (1998) considers the "specialist units" and "maverick units" as equivalent terms.

Sarkis (2000) and Sarkis (2001) delineate mavericks as the units which show themselves falsely efficient by weighing heavily on a single input or output. Therefore, mavericks in this definition do not necessarily have zero-weights in their optimum weight scheme. Tofallis (2001) characterizes usually efficient units which use unusual/extreme optimum weights. Moreover, efficient mavericks are not "referenced by any inefficient unit", and thus they don't appear in the reference list. Tofallis (2001) suggests a method to detect and downgrade these self-promoted units. Santos and Themido (2001) define mavericks as units "which are only efficient because of a very unbalanced choice of weights". According to the authors, this unbalanced choice of weights is also "unfair and unrealistic", and the maverick units indeed misuse the DEA freedom of choosing weights. Serrano-Cinca et al. (2005) suggest a variable selection method based on an application of principal component analysis (PCA) on DEA models. The final outcome of the method is a PCA visual map of units and super-imposed vectors of models. The authors define maverick units as the units standing at extreme points [on the map] which is because of using "an unusual mix of inputs and outputs to achieve efficiency." Therefore, mavericks are highlighted through their "discordant behavior".

Appa and Williams (2006) and Appa, Argyris, et al. (2006) suggest a new framework to solve DEA models and a new formulation of cross-efficiency matrices. Having considered the multiplicity of optimum weights and alternate optima of each unit, the authors assign "maverick" characteristic to some weight sets rather than units. Thus, a maverick unit is a unit with an "unrealistic weight set", which consequently causes "unrealistic behavior" comparing to the behaviors of the [majority of] peers. Moreover, these units achieve their efficiency through "efficient use of only a subset of inputs and outputs assign zero to all others", they are "specialize in a subset of inputs and outputs", and they are in contrast to all-round performers.

In a widely under-rated paper in the DEA maverick literature, Fumero (2004) considers maverick units as the units with "specialized behavior", however their behavior must be evaluated based on their optimum weights distribution and not the behavior of the peers. Therefore, Fumero (2004) attacks the maverick problem directly from the optimum weight rather than proxies such as cross-efficiency scores. While there is no word about necessity of presence of zero-weights in maverick alternate optimum, Fumero (2004) measures the amount of maverick-ness through the variance of "virtual outputs" in that alternate optimum. The author considers specialists and mavericks as equivalent DMUs which use "focused strategy" to achieve their maximum efficiency.

O'Neill and Dexter (2005) characterize mavericks as the units which use "an

unusual production technology", considering the production technologies of rest of DMUs. Avkiran (2006) defines a maverick unit as unit whose "efficiency score is based only on one or a small number of all the available inputs/outputs", and to do so "can lead to assigning of zero weights" to majority of the input/output variables. Lu and Lo (2007) underscore "heavily weighting a few favorable inputs and outputs" as the differentiation characteristic of mavericks. The authors call mavericks "false positive" units. Li (2008) follows the well-established maverick definition of Doyle and Green (1994), and emphasizes that efficient mavericks achieve their efficiency "just by using inappropriate weights." Tofallis (2010) emphasizes on the location of the mavericks "on the perimeter of the observed frontier", and describes these units as the ones with significantly unbalanced scores of different criteria. The author names balanced units as "good all-rounders". Flokou et al. (2011) refer to mavericks as units with over-estimated efficiency scores, false positive units that must be separated from true positive ones. Wang and Chin (2010), Lee and Pai (2011), Fu et al. (2011), and Ma et al. (2014) follow the practical definition of Doyle and Green (1994) by considering the mavericks as the units with the most discrepancy between self-appraisal and peer-appraisal efficiency scores. However, Wang and Chin (2010) suggest a new CEM formulation which yields different results from Doyle and Green (1994), not because of the different maverick definition, but due to different CEM approach. Similar to Wang and Chin (2010), Lim (2012) devises a new unique CEM formulation while defining mavericks as the units with unrealistic optimum weights. Chai et al. (2013) describe the mavericks as the units with unbalanced metrics, the units that "cheat" in order to achieve high efficiency scores through "weighting a single input and a single output and setting the rest weight coefficients close to 0". (Lotfi et al. (2013) state that a DMU which puts "a huge weight on one or a few factors and assigning a zero or very small weights to other factors" is a maverick unit. Majority of these more recent papers have focused on suggestion of new CEM formulations rather than refinement of the maverick definition.

As one can evaluate, the various definitions, if not to some extent incompatible, are at best similar to the metaphoric story of "elephant in the dark". Moreover, since the definition is not the goal, but it is a concept based on which a detection technique is devised, the detection techniques have received more attention than the philosophy behind them. Hence, researchers have been more focused on maverick identification tools. However, in the next section, through review of detection techniques, it is shown that the maverick detection techniques are incompatible with their corresponding maverick definitions, based on which the techniques have been developed. Put it in other words, what maverick detection techniques try to detect is not what they are supposed to detect according to the maverick definitions.

2.1 How to detect mavericks? Maverick detection indices through literature

In contrast to various operational definitions of mavericks, there are a few maverick indexes in order to detect the individual maverick units in each problem.

We review these indexes in this section with emphasis on the most frequently used, and thus the most important one: Maverick index by Doyle and Green (1994)

The first maverick index was suggested by (Sexton et al. 1986). Based on the characteristic of maverick regarding "perfect efficiency, but low peer-evaluation scores", the authors suggested that the gap between self-evaluation score and average peer-evaluation score can be used in order to detect maverick units. This idea becomes mature and mathematically formulated in (Doyle and Green 1994). Following emphasis on this characteristic that mavericks "enjoy the greatest relative increment when shifting from peer-appraisal to self-appraisal", (Doyle and Green 1994) suggest the first maverick index which tries to capture this "relative increment" numerically.

$$M_k = \frac{E_{kk} - e_k}{e_k} \tag{3.1}$$

Where

$$e_k = \frac{1}{n-1} \sum_{s \neq k} E_{sk} \tag{3.2}$$

In the formulation of Maverick index(MI), E_{kk} is the simple-efficiency or self-appraisal of the generic DMU_k, and e_k is the average peer-appraisal of that unit. Therefore, the higher the MI, the more maverick the DMU. Hence, there is no "to be or not to be" maverick threshold for a unit, and in practice units with the highest amount of MI are the objects of interest.

Consequently, Doyle and Green (1994) convert the qualitative maverick index idea of Sexton et al. (1986) into mathematical formula, which essentially captures the incompatibility of self-evaluation and peer-evaluation of any unit. In other word, a unit is highly maverick when how it perceives itself is highly different from how it is perceived by other units, on average.

According to this index, every unit in a DEA problem has a degree of maverick-ness. Although some new indexes have been suggested by successor researchers, the Maverick index (MI) of Doyle and Green (1994) is the most popular and the most commonly used maverick-detection index in the literature. Those less common indexes are reviewed in the rest of this section briefly.

Baker and Talluri (1997) suggest False Positive Index (FPI) in order to detect False Positive units. FPI structurally is almost identical to MI with a very negligible difference. The formulation is as follows:

$$FPI_k = \frac{E_{kk} - e_k}{e_k * m} \tag{3.3}$$

Where:

E_{kk} is the simple efficiency of unit k

e_k is the mean score of robot k obtained from the CEM

The only difference between FPI and MI, is in the formulation of "average cross-efficiency score". In MI, self-evaluation score is not included in the average cross-efficiency score of a unit, therefore the average cross-efficiency is calculated

based on $m-1$ units rather than all m units. Generally, FPI is the second common maverick index through the literature, however it is hardly different from MI.

Having stated that mavericks cause difficulties in super-efficiency models, O'Neill (1998) defines its maverick index in the area of super-efficiency models (Andersen and Petersen 1993) based on the author's novel performance measure, multifactor analysis. The detection technique is repeated in O'Neill and Dexter (2005) and beside the formulation complexity, it has one specific discrimination from MI or FPI: the index has a binary nature. Therefore, based on this index, a unit is a maverick or not, there is no middle ground. For more detailed explanation, one can refer to O'Neill (1998).

Sarkis (2000) uses a maverick index which is essentially the same as Doyle and Green (1994) with the following formulation:

$$M_k = \frac{E^*_{kk}}{e_k} \tag{3.4}$$

Where E^*_{kk} is the self-evaluation score or simple efficiency of the unit k. This formulation can be achieved with simple algebraic manipulation of MI. However, the differentiation point of maverick index of Sarkis (2000) is not the index formulation but the definition of a threshold for either being or not being a maverick. Sarkis (2000) defines this threshold as

$$MI_{k\in E} = \overline{MI} + \sigma\rho \tag{3.5}$$

Where $\overline{MI}$ is the average of all unit's maverick indexes, σ is the standard deviation of maverick indices of the units, and ρ is the "false positivity factor", an arbitrary chosen number, which is equal to unity in the Sarkis (2000). Therefore, units with perfect CCR efficiency (Charnes et al. 1978) and MI greater than the threshold are considered as mavericks.

While Sarkis (2000) targets this MI flaw, which leaves one ambivalent about whether a unit should be considered as a maverick due to lack of a decisive rule, and tries to resolve the problem with setting a threshold, it seems the arbitrariness of the "false positivity factor" defeats the threshold's purpose.

Serrano-Cinca et al. (2005) suggest a totally a new method in order to detect mavericks. While the main purpose of their PCA-DEA method is variable and model selection in DEA, as a by-product the authors suggest a visual technique in order to detect mavericks. The visualization is based on 2D PCA and property fitting.

Fumero (2004) is the first who criticizes the MI and tries to tackle the maverick by direct evaluation of optimum weights rather than proxies. The author's method is based on virtual outputs (or virtual inputs), and based on this concept that a maverick/specialist unit has focused strategy. However, the virtual outputs or virtual inputs also suffer from multiplicity of weights, and they are not unique for an efficient DMU. Fumero (2004) argues that since the evaluation of all alternate optima is not possible, the only possible way is evaluation of the alternates with the most extreme values. Having computed these alternates, the author suggests aggregation of them through simple average. If the final weight scheme, achieved through averaging the

extreme alternates, has high variance among its virtual outputs, then the unit is labeled as maverick. This variance is measured by "equilibrium index".

Appa and Williams (2006) and Appa, Argyris, et al. (2006) suggest a new approach to formulate and generate CEMs. According to this method, the authors also suggest a new maverick index. Although the authors' CEM benefits from "all hyperplanes that define the CRS production possibility set" instead of choosing one alternate per unit of (Doyle and Green 1994), or arbitrary synthesized alternate optimum of Fumero (2004), their maverick index is structurally very similar to (Doyle and Green 1994) MI, derived of benevolent CEM. Hence, their contribution seems more on CEM formulation rather than maverick index.

Wang and Chin (2010) first formulate CEM with their "neutral approach", and then define "Efficiency Disparity Index(EDI)" for a generic DMU_k as:

$$EDI_k = \frac{E_{kk} - \overline{\theta}_k}{\overline{\theta}_k} \times 100\%, k = 1, \ldots, m \tag{3.6}$$

where E_{kk} is the CCR-efficiency of DMU_k, and $\overline{\theta}_k$ is its efficiency computed by using the average set of cross-weights. As one notices, the EDI is structurally very similar to MI except for the percentage expression.

(Lim 2012) suggests a new CEM formulation, minimax aggressive and maximin benevolent, and shows that these formulations can improve the accuracy of maverick index. Nevertheless, the maverick index that (Lim 2012) uses is the the original MI.

To the author's best knowledge, MI (and FPI) have been the most frequently used indexes by significant margin among all the indexes. For instance, besides the already mentioned papers in this section, many other researches (Braglia and Petroni 2000; Avkiran 2006; Lu and Lo 2007; Li 2008; Lee and Pai 2011; Fu et al. 2011; Chai et al. 2013) have used MI (or FPI) in their studies. For this reason, in this research MI is at the center of attention, and considered as the benchmark of the new index.

2.2 Critical Appraisal of the Maverick Literature

The very first point in the review of the maverick literature is lack of a categorical operational definition of maverick units. This shortage has caused discrepancies among the various definitions. For instance, while some have mentioned uncommon-ness as the main characteristic of maverick unit, the other have emphasized on specialist unit, i.e. units with focused strategy. Similarly, while some definitions consider maverick units as efficient units, some sufficed to state that maverick units are over-rated because of the unbalanced metrics.

Nevertheless, there is a common ground among the definitions. All the definitions emphasize on heavily unbalanced weight distribution in alternate optima of a maverick unit. In its extreme case, this heavily unbalanced weight scheme becomes a weight scheme with non-zero weights in only one input and one output. Moreover, as shown in the section 1.2, almost all of the papers have used MI of Doyle and Green (1994), or a structurally similar index in order to detect the mavericks.

Several problems arise here considering the intersection of the definitions, and the D&G maverick index. These problems stem from tacit assumptions behind maverick studies. The very first problem is the assumption of "maverick nature" for some units. A nature that does not change, therefore a unit is either a maverick or not, and a method is needed to detect such unit. However, this assumption generally does not hold, since some units are able to show both specialist and all-round performer behaviors. In other words, some units can achieve their optimum efficiency through adaptation of a very focused, or very balanced optimum weight sets.[4] Such units can be called units with "dual-roles".

The following numeric example can depict this point better. Table 3.1 consists of inputs' and outputs' levels of 20 DMUs.

Table 3.1: Input and output levels of 20 DMUs. According to (Fumero 2004), This dataset is artificially fabricated for sake of the numeric example, and the inputs/outputs do not bear any meaning

Unit	**Input 1**	**Input 2**	**Output 1**	**Output 2**	**Output 3**	**Output 4**
DMU1	1204,651	4,542	1707	330	0,143	0,587
DMU2	349,531	4,966	776	107	0,167	0,718
DMU3	504,882	2,983	860	115	0,154	0,662
DMU4	179,618	3,445	492	52	0,167	0,717
DMU5	196,747	3,66	265	50	0,167	0,593
DMU6	457,718	4,727	881	105	0,154	0,68
DMU7	338,626	5,28	722	91	0,154	0,537
DMU8	207,752	1,796	377	51	0,143	0,701
DMU9	71,724	3,162	227	11	0,2	0,739
DMU10	82,839	5,941	225	10	0,2	1,018
DMU11	56,176	7,349	33	2	0,143	0,767
DMU12	467,688	2,563	724	156	0,133	0,681
DMU13	209,132	2,701	364	70	0,167	0,704
DMU14	105,861	1,718	190	11	0,154	0,629
DMU15	129,407	4,551	293	17	0,167	0,72
DMU16	50,129	2,551	140	6	0,182	0,543
DMU17	53,018	3,247	211	9	0,182	0,938
DMU18	90,132	16,429	119	2	0,143	0,696
DMU19	111,031	5,706	88	2	0,133	0,524
DMU20	44,482	50,277	19	0	0,182	0,365

Table 3.2 consists of six alternate optima of DMU_{17} from a DEA dataset with 20 DMUs. Table 3.1, Table 3.2 and the example have been borrowed from Fumero (2004)

Table 3.2 clearly shows that DMU_{17} can achieve its optimum efficiency through adaptation of a very focused weight set such as weight set 1, in which all weights are

[4]Such phenomenon is caused by the multiplicity of weights in DEA, which subsequently stems from the degeneracy of its linear programming

zero except one input and one output, OR adaptation of a well balanced weight set such as weight set 6, in which there is no zero weights or high variance among values.

Table 3.2: Some alternate optima of DMU_{17} (Fumero 2004)

Unit	Efficiency	V.I. 1	V.I. 2	V.O. 1	V.O. 2	V.O. 3	V.O. 4
DMU17	1	0	1	1	0	0	0
DMU17	1	0,1	0,9	0	0	0	1
DMU17	1	0,99	0,01	0	0	0,869	0,131
DMU17	1	1	0	0	0,45	0	0,55
DMU17	1	0,55	0,45	1	0	0	0
DMU17	1	0,66	0,34	0,25	0,11	0,22	0,42

These dual-role units can exist in any DEA problem. Therefore, a perfectly efficient unit which is detected as a maverick, may be able to show non-maverick behavior through adaptation of another optimum alternate. Hence, assuming that there are some units that are mavericks by nature, is not generally true.

The second problem is related to the MI and what it actually delivers and what it is supposed to deliver. Considering either the "zero-weight" or "heavily weighted" definition of maverick units, maverick index is supposed to detect the units with such characteristics. However, what maverick index (as well as FPI) actually delivers is a measure of "uncommon-ness", a distance measure between self-perception and peer-perception.[5]

An uncommon unit is not necessarily a unit with "heavily unbalanced weight scheme"‘ or worse, a unit with "zero-weights in all but one input and output". An uncommon unit can be a unit with unbalanced metrics only if the majority of units are all-round performers. However, there is generally no guarantee for holding this assumption of balanced majority.[6] The problem becomes worse as Maverick Index even does not measure this "uncommon-ness" accurately and comprehensively. This point will be explained the section 2.1.

Before articulating any suggestion to overcome such problems, it is worthy to come back to the roots and read Sexton et al. (1986) once again: "it is not possible to say with certainty that this DMU is price inefficient, the fact that it uses input mix so different from the majority of other DMUs, warrants some attention." As one can see, suggested usage of cross-efficiency to identify mavericks has not been supposed to be a certain tool for detection of opportunistic units, or specialist units. Moreover, it is advised that the method should be used with care, keeping in mind that the findings are not necessarily maverick units.

In the author's opinion, part of these confusion about maverick units is caused by the maverick term and its usage in maverick index. In order to overcome this confusion, it is better to enrich DEA maverick literature with some data mining

[5]This problem has been pointed to in Fumero (2004)

[6]In fact, according to empirical experience of the author, it seems in most datasets, there are several clusters of units rather than a significant majority. This situation is called heterogeneity of DMUs in DEA literature

terminology. Hence, maverick indexes can generally be considered as anomaly detection techniques. Tools to highlight novelties and exceptions. Therefore, the maverick identification techniques do not identify mavericks, but they highlight exceptions in the set of DMUs.[7] These exceptions are not necessarily mavericks, i.e. units with technical efficiency and price inefficiency. However, they are the most suspicious units to be mavericks, and hence they are worth of further investigation.

In an equivalent terminology, while by operational definitions, i.e. zero weights or focused weights characteristics of mavericks, a maverick unit is a "specialist unit", maverick index tries to find "special unit" in a set. More gravely, while MI tries to find "special units" rather than "specialist units", it even fails to find the "special units" due to the index' partial evaluation of uncommon-ness. "Special units" can be "specialist units" where the majority of the units are "generalists" or "all-round performers". However, there is no guarantee that majority of the units in a set are generalists, and even there is no guarantee to have a significant "majority". It is totally anticipated to have several separate clusters with similar number of members in a set, such that each cluster has its own focused strategy.[8]

The foundation of the new maverick detection method, suggested in the second part of this paper, is based on the answers to the mentioned problems and shortcoming in the maverick DEA literature. To do so, it is necessary to emphasize that maverick detection methods based on cross-efficiency data are supposed to identify the uncommon units, i.e. exceptions and anomalies, which may or may not be maverick. Thus, it is of importance to clarify the relations among uncommon units, specialist units, and zero-weights phenomenon: a special unit is specialist, if the majority of the units in the dataset are all-round performers. A specialist unit may have zero-weights in all its alternate optima, but such feature, similar to all-round characteristic, must be investigated directly from the optimum weight schemes, and not any proxies.

In order to circumvent the dual-role phenomenon, benevolent formulation of CEM can be hired. In the benevolent formulation, a unit chooses the most favourable optimum alternate from the point of view of other units. In other words, a unit chooses the most globally compatible and congruous alternate optimum among all its alternate optima. Therefore, if a unit is still incongruous under benevolent condition, then the unit is probably an exception, an uncommon unit, by its nature. It is worth to mention that aggressive formulation seems to be be more popular in the literature, e.g. (Lim 2012), mainly because of the notion of competition among units. However, under aggressive formulation, the units are forced to choose their most incongruous alternate optima. Consequently, some units become maverick while they can show non-maverick behaviour if they are permitted. In other words, the aggressive formulation makes the dual-role problem worse, by forcing all dual-role units to show maverick behaviour.

[7]Here the usage of "outlier" term is deliberately and intentionally avoided, since outlier in DEA has its own meaning, definition and literature. The relation between outliers and cross-efficiency anomalies is something to be investigated separately.

[8]Generally, the problem emerges when the maverick indexes try to measure specialist-ness based on other units' behaviour rather than the distribution of unit's optimum weights

In order to rectify the problem of incompatibility between maverick definitions and maverick indexes, one approach can be change of the name of the maverick indexes, to anomaly indexes or equivalent terms. In the current situation, it is confusing that a maverick index does not find mavericks. However, this approach may be difficult to implement since the literature is well-established. Therefore, in order not to misuse maverick indexes, i.e. not to mislead by incompatibility between maverick definition and maverick detection techniques, we should come back to the roots, the Sexton et al. (1986) purpose of cross-efficiency, and keep in mind the aim of the CEM-based maverick identification tools. These tools detect and underscore uncommon units, and maverick indexes are practically measures uncommon-ness.

Although following the above suggestions, using benevolent CEM and emphasizing on the aim of maverick identification tools, improves the anomaly detection procedure, the indexes such as the D&G's MI, and FPI, are still in-comprehensive, and thus inaccurate even under such conditions.

Mainly, what D&G MI measures is the distance between self-perception and average peer-perception. MI gauges this distance based on the rated role of units in CEM, i.e. the MI value for each unit is calculated based on the unit's column profile in CEM. Hence, MI neglects the rating role of units in the evaluation of units' uncommon-ness, while the rating role is direct reflection of the DMU's optimum weight and it augments the column profile. In other words, every unit in CEM has two aspects, a rating aspect (row) and a rated aspect (column) and MI only considers one of them (the rated aspect, the column) in the measuring of maverick-ness.

In summary, according to Sexton et al. (1986), the mavericks are the units with simultaneous technical efficiency and price inefficiency. The authors suggest cross-efficiency in order to find "suspicious" units, the units that demand further investigation. Thus, the findings of cross-efficiency cannot be considered as a certain maverick unit, but a suspicious unit, a candidate. "it is not possible to say with certainty that this DMU is price inefficient, the fact that it uses input mix so different from the majority of other DMUs, warrants some attention". Therefore, maverick detection methods, such to D&G MI, are supposed to find uncommon units, the candidates. Considering this point, the current study is an effort to come back to the roots, and it suggests an index to identify the possible candidates. The new index evaluates the units with a more holistic approach comparing to the previous indexes.

In the next section, it is shown through visualization that why MI index does not measure the uncommon-ness comprehensively, and subsequently this shortcoming is rectified by a new maverick index.

2.3 The Motivation of Maverick Detection

At last, we are at the point of talking about the motivations behind maverick-detection. This sub-section is located at the end of the literature review part due to predecessor nature of the critical evaluation. In other words, as the maverick definitions and indexes are revisited, the motivation behind maverick-detection must be revisited as well.

In general, there is a consensus about "unrealistic efficiency score" of the units with zero-weights. There are two broad approaches to the problem of these ubiquitous units in the literature. The first approach tries to prevent DMUs to choose zero-weights (or more generally unrealistic weights) through methods such as weight restriction and value judgment (Allen et al. 1997). In contrast, cross-evaluation method (Sexton et al. 1986; Doyle and Green 1994) permits the DMUs to choose their weight and then tries to yield a more realistic efficiency score of the units through cross-efficiency score. That is to say, the cross-evaluation method permits the peer units audit each other, while in weight restriction this audition is done by the user. Nevertheless, both approaches seek one goal: to achieve more realistic efficiency scores for the units, including the problematic units with unrealistic weight schemes, including zero-weights.

While this common goal successfully can be achieved with either of two approaches (although each approach causes its own second-order problems), some researchers put one step forward in order to detect the problematic units individually. To do so, several maverick indexes have been suggested, among which the most common is D&G maverick index. However, the outcome of such index does not necessarily highlight the maverick units, i.e. the units with focused strategy or zero weights or focused strategy. The maverick indexes based on CEM must be seen as anomaly detection method, which measures the amount of in-congruence of units. If the majority of the units are all-round performers, then extreme discordant units are mavericks. Consequently, maverick indexes shed light on special units, and not necessarily specialist units. Moreover, not every specialist unit should be evaluated as an unrealistic unit. It seems the specialist units can also be categorized into true and false specialist units, and mavericks should point to the latter category. Without knowing the "reality" of a DMU, it is difficult to talk about "unrealistic" behaviour of it.

Finally, the special units, as the outcome of maverick identification indexes, are worth of further investigation since their dissonant behaviour possibly stems from their uncommon and radically different structure, production technology, or strategy.

3 The New MI: Origins and Motivations

The visualization method, presented in the section 2 and briefly explained in Appendix A, is an exploratory tool which leaves the evaluation of units' uncommon-ness to the user. The user can detect the suspicious units based on the guidelines which consider both aspects of units in the CEM. However, since the visualization is based on a dimension-reduction technique, the final map is not exact and precise. This inaccuracy increases with the increment of the stress (noise) in the final map, which may consequently affect the maverick detection process. The current study suggests a numeric index which is almost exact and perfectly precise, and has no noise or inaccuracy. This new numeric maverick index can be used with the visual tool simultaneously, in order to have the advantages of quantitative and qualitative worlds at the same time. Section 2 is dedicated to explanation of the new MI.

The new maverick index can be explained more lucidly through CEM visualization, since the new method can be interpreted as the extrapolation of the CEM visualization in high dimensional space. Moreover, the idea of the new MI has been emerged from the visual detection method, therefore it is rational to start the new MI explanation from its origin.[9]

Let's consider the following dataset of seven academic departments, Table 3.3, from Wang and Chin (2010), and previously Wong and Beasley (1990). This fabricated dataset is chosen in order to facilitate the explanation of the method. Later on, the new method is applied on a real and larger dataset.

Table 3.3: The input and output levels of 7 academic departments (Wang and Chin 2010)

Unit	**Input1**	**Input2**	**Input3**	**Output1**	**Output2**	**Output3**
DMU1	12	400	20	60	35	17
DMU2	19	750	70	139	41	40
DMU3	42	1500	70	225	68	75
DMU4	15	600	100	90	12	17
DMU5	45	2000	250	253	145	130
DMU6	19	730	50	132	45	45
DMU7	41	2350	600	305	159	97

The three inputs are Number of academic staff, Academic staff salaries in thousands of pounds, and Support staff salaries in thousands of pounds. Subsequently, the three outputs are the Number of undergraduate students, Number of postgraduate students, and Number of research papers. From the resource table of Table 3.3, Wang and Chin (2010) presents the benevolent CEM, Table 3.4, as follows:

Table 3.4: Benevolent CEM of 7 Academic departments (Wang and Chin 2010)

Unit	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1	0,981	0,769	0,641	0,938	1	1
2	0,922	1	0,772	0,701	0,899	1	1
3	1	0,851	1	0,454	0,495	1	0,294
4	0,688	1	0,735	0,82	0,765	0,951	1
5	1	0,846	0,665	0,414	1	0,91	1
6	1	0,981	0,769	0,641	0,938	1	1
7	1	0,981	0,769	0,641	0,938	1	1

The visualization of this CEM is presented in Figure 3.1. The technical procedure of visualization using multidimensional unfolding has been briefly explained in Appendix A. Here, I suffice to explain how to read the map.

There are two types of objects on the map, compatible with the two modes of the CEM: the row and column objects. The round shapes in blue colour are the

[9] A brief explanation of visualization method is presented in Appendix A

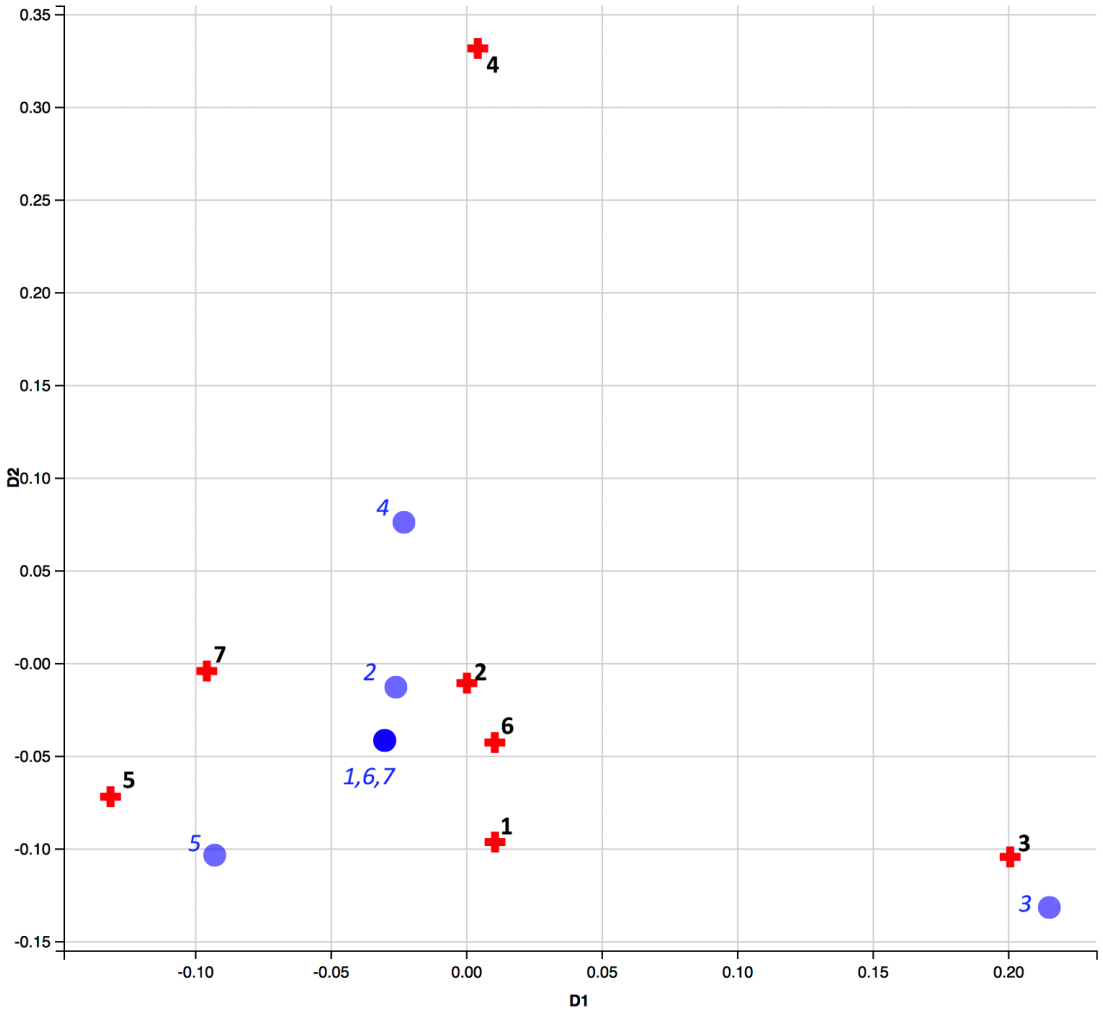

Figure 3.1: The unfolding map of benevolent CEM of 7 Academic department

DMUs in their rating roles, row objects of the CEM. Subsequently, the cross-shape objects in red, are the DMUs in their rated roles, column objects of the CEM. The distances between any row-column (blue-red) pair of object on the map ideally is proportional to the corresponding cross-efficiency score in the CEM, such that higher cross-efficiency is shown by closer distance rather than lower cross-efficiencies. Indeed, the map is a sort of "preference map", in which the row objects locate closer to the column-objects which the row-objects prefer, comparing to the column-objects which the row-objects don't prefer. Hence, a 100% cross-efficiency given by DMU_i to DMU_j is depicted by a closer distance between DMU_i (row object) and DMU_j (column object) comparing the distance between DMU_i (row-object) and DMU_k (column-object) which corresponding cross-efficiency score is, for instance, 50% .

The last necessary guideline in order to make one able to interpret the map is the meaning of the distance between two objects from the same type. The distance between any two row-object approximately shows the closeness and similarity of those two DMUs in their rating roles and row profiles. Similarly, the closeness between two column objects is the sign of similarity of their column profile, rated roles of

those DMUs.

It's of great importance to emphasize that the distances between row and column objects on the map are reflection of inverse of corresponding cross-efficiency scores. Put it differently, the higher a cross-efficiency score, the shorter the corresponding distance on the map.

Moreover, DMU_i row object to DMU_i column object distance is the reflection of the self-evaluation, and the average distance of DMU_i column object to all row objects excluding DMU_i row object is the average cross-efficiency of DMU_i.

Therefore, according to the Doyle and Green's MI formulation, the D&G's MI can be interpreted on the map as the difference between self-distance of DMU_i and the average column object to row objects distance of DMU_i, divided by the average column object to row objects distance of DMU_i.

Figure 3.2 depicts the D&G MI visually for the DMU_5. The row-column distances are presented in orange lines for all pairs composed of $column_5$, except Column5-Row5. These orange lines are cross-efficiency scores which have been given to DMU_5. Column5-Row5 distance is in green, in order to differentiate it from the other distances, since it is "simple-efficiency" of the DMU_5.

One important point which is revealed in visualization of D&G MI is the absence of DMU_i row-object to other row-objects distances. In other words, the D&G MI only considers the column aspect in order to measure the uncommon-ness of a unit, while any unit in cross-efficiency matrix has two aspects: row and column.

Table 3.5 shows the D&G MI values for the seven academic departments as well as the average cross-efficiency scores and the CEM.

Table 3.5: Average cross-efficiency, MI and MI ranking of the Benevolent CEM of 7 academic departments

Unit	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	1	0,981	0,769	0,641	0,938	1	1
2	0,922	1	0,772	0,701	0,899	1	1
3	1	0,851	1	0,454	0,495	1	0,294
4	0,688	1	0,735	0,82	0,765	0,951	1
5	1	0,846	0,665	0,414	1	0,91	1
6	1	0,981	0,769	0,641	0,938	1	1
7	1	0,981	0,769	0,641	0,938	1	1
Avg. cr-ef	0,935	0,94	0,746	0,582	0,829	0,977	0,882
MI	0,07	0,064	0,34	0,408	0,206	0,024	0,133
MI - Ranking	5	6	2	1	3	7	4

According to D&G MI, DMU_4 has the highest MI ranking, in other words DMU_4 is the most incongruous unit. However, one can see that DMU_4 in its rating role (row-object in blue) is close (similar) to the majority of the units. In other words, the optimum weights of DMU_4, which is the reflection of its production technology, is similar to majority of its peers. It can be concluded that DMU_4 from the rating aspect is not an uncommon unit.

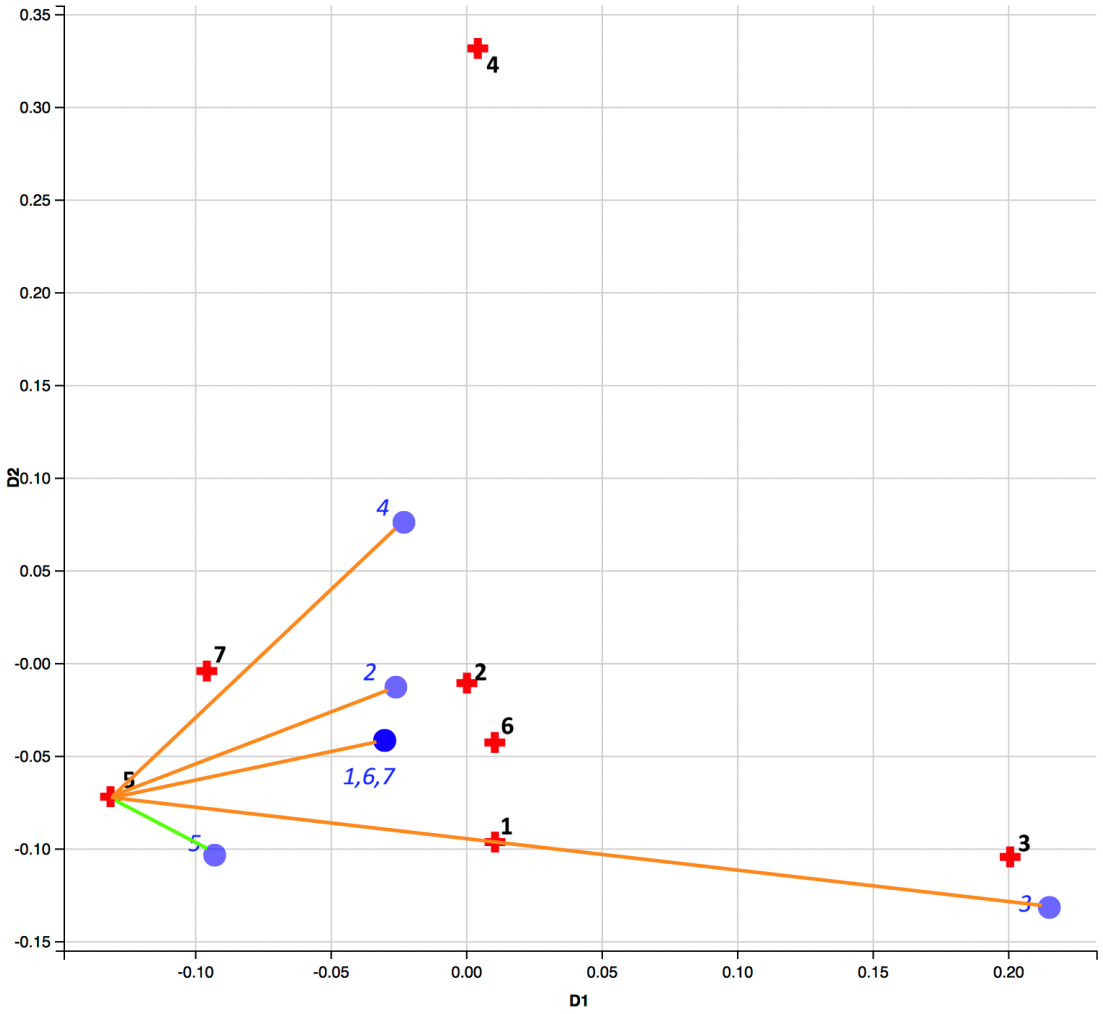

Figure 3.2: D&G MI of DMU5, represented as distances on the unfolding map

The idea of the new maverick index comes from this concept that incongruous-ness of a unit should be comprehensively evaluated by considering both rating and rated roles of each unit, and not only the rated (column) role, since the two roles are complementary. Therefore, the accuracy of the D&G MI, can be improved through a more comprehensive evaluation.

Hence, the author suggests to consider both row and column objects of a DMU, when one wants to evaluate the uncommon-ness and maverick-ness. The CEM visualization, in the form of the Figure 3.1, can help to explore many relations among DMUs in the CEM, namely uncommon-ness. While one should keep in mind that the visual map is an exploratory tool, one can detect DMU_3 as the most suspicious unit in maverick-ness, and designate it for further investigation.

The visualization method, presented in the section 2 and briefly explained in Appendix A, is an exploratory tool which leaves the evaluation of units' uncommon-ness to the user. The user can detect the suspicious units through eye-balling, and based on the guidelines which consider both aspects of units in the CEM. However, since this visualization is based on a dimension-reduction technique, the final map is

not exact and precise. The final map is visual unfolding of the corresponding CEM in bi-dimensional space. Any CEM is an asymmetric matrix with 2m different objects, m row objects and m column objects, therefore a bi-dimensional map is a result of reduction of an 2m-dimensional space into 2-dimensional space. The resulting noise of a multidimensional unfolding map can be measured by STRESS-1 value: the higher the stress, the more inaccurate the map. The stress of the Figure 3.1 map is around 0.4.[10] This inaccuracy increases with the increment of the stress (noise) in the final map, which may consequently affect the maverick detection through eye-balling. Besides, the visually identification of anomalies may become difficult when there are close candidates. In such cases, having a quantitative and numeric index is more practical to rank order the uncommon units.

Therefore, even though the visualization of CEM can drastically help in identification of anomalies, still two issues should be addressed. The first issue is regarding the inevitable imprecision of the map, and the second is about rank order of the units based on their uncommon-ness. In the rest of this study, a numeric maverick index is suggested, an index which is almost exact and perfectly precise, and has the least noise or inaccuracy. This new numeric maverick index can be used with the visual tool simultaneously, in order to have the advantages of quantitative and qualitative worlds at the same time. This new index is explained in the next section.

3.1 The new MI

As mentioned in the previous section, the idea of the new maverick index was conceived in pursue of tackling the two problems of visual maverick-detection method, while preserving the comprehensive characteristic of the visual method.

Therefore, the new maverick index is closely knitted to the visual maverick-detection method. Indeed, the new maverick index is the exact and quantitative version of the visual method.

In order to preserve the comprehensive evaluation of the uncommon-ness, which subsequently means considering both rating and rated (row and column) aspects of each DMU in the evaluation of its maverick-ness, the new maverick index benefits from quantitative equivalence of rating(row) and rated(column) measure of in-congruence of each unit. Since, the distances on the map are dissimilarities between objects -such that the shorter the distance, the higher the similarity/preference, and the higher the correspondent cross-efficiency score-, the average distances could be used as a measure of dissonance. In other words, when a unit in its rating(row) or rated(column) role is on average far from the crowd, it means that the unit is different from the crowd based on that specific aspect. This average distance can be used as a measure of uncommon-ness between two or among several units as well.

Consequently, two average distances are calculated for each object on the map. These are the two indices that are going to compose the final new maverick index.

It was explained in section 2.1, that the D&G MI considers only one aspect, which is the column or rated aspect, in evaluation of uncommon-ness. In other words, the D&G MI takes the average cross-efficiency into account in calculation

[10] More about stress in Borg and Groenen (2005) book or J. d. Leeuw and Mair (2008)

of the maverick index as well as the self-efficiency. The average cross-efficiency of DMU_i is the average distance of DMU_i-Column object from all row objects, except DMU_i-Row object. The simple-efficiency of DMU_i is the distance of column object of DMU_i from the row object of DMU_i. Therefore, the only data entity that D&G MI benefits from is the column profile of each unit.

In contrast, the new MI tries to benefits from both row and column profiles of each DMU, since both of them include information about the unit. They are two complementary aspects of a unit. To do so, first we define "column isolation index". Column isolation index is the average distance of column object-i to all row objects, excluding row object-i. In other words, column isolation index is equivalent to average cross-efficiency. In problems with m DMUs, Mathematical presentation/formulation of the column isolation index(CII) is as follows:

$$CII_i = \frac{\sum_{j=1}^{j=m, j\neq i} crd_{ij}}{m-1} \tag{3.7}$$

Where: crd_{ij} = column-$object_i$ to row-$object_j$ Euclidean distance in R^n , $n \in \{2, \ldots, 2m\}$, m is the number of DMUs, and n is the dimensions of the space in which the CII is calculated. At 2 dimensions, we are calculating Euclidean distance on the MDU map, and compromising for the possible imprecision of it. The higher the n, the lower this imprecision. The maximum number of n is equal to the total number of row and column objects, i.e. $2m$, since the space is depiction of the CEM with $2m$ objects.

The idea of average distance can be extrapolated for the row objects too. Consequently, the average distance of DMU_i-Row object to all row objects, gauges the similarity of the row aspect (rating role) of DMU_i to other units.

Similar to column isolation index (CII), this average distance of row object-i to other row objects is called "row isolation index" (RII). Therefore, every column object has a corresponding "column isolation index", and every row object has a corresponding "row isolation index". Naturally, every DMU has a pair of row and column isolation index.

In problems with m DMUs, Mathematical presentation/formulation of the row isolation index (RII) is as follows:

$$RII_i = \frac{\sum_{j=1}^{j=m, j\neq i} rrd_{ij}}{m-1} \tag{3.8}$$

Where: rrd_{ij} = row-$object_i$ to row-$object_j$ Euclidean distance in R^n , $n \in \{2, \ldots, 2m\}$, m is the number of DMUs, and n is the dimensions of the space in which the CII is calculated. At 2 dimensions, we are calculating Euclidean distance on the MDU map, and compromising for the possible imprecision of it. The higher the n, the lower this imprecision. The maximum number of n is equal to the total number of row and column objects, i.e. $2m$, since the space is depiction of the CEM with $2m$ objects.

As an example in 2d space, Figure 3.3 shows the distances (orange lines) from DMU_5-Column object (rated role) to all DMUs-Row Objects (rating roles), except

row object 5. The average of these distances is the column isolation index (CII) of unit 5, equivalent to average cross-efficiency of this unit. The column isolation index is the measure of uncommon-ness of DMU_5's rated role (column profile).

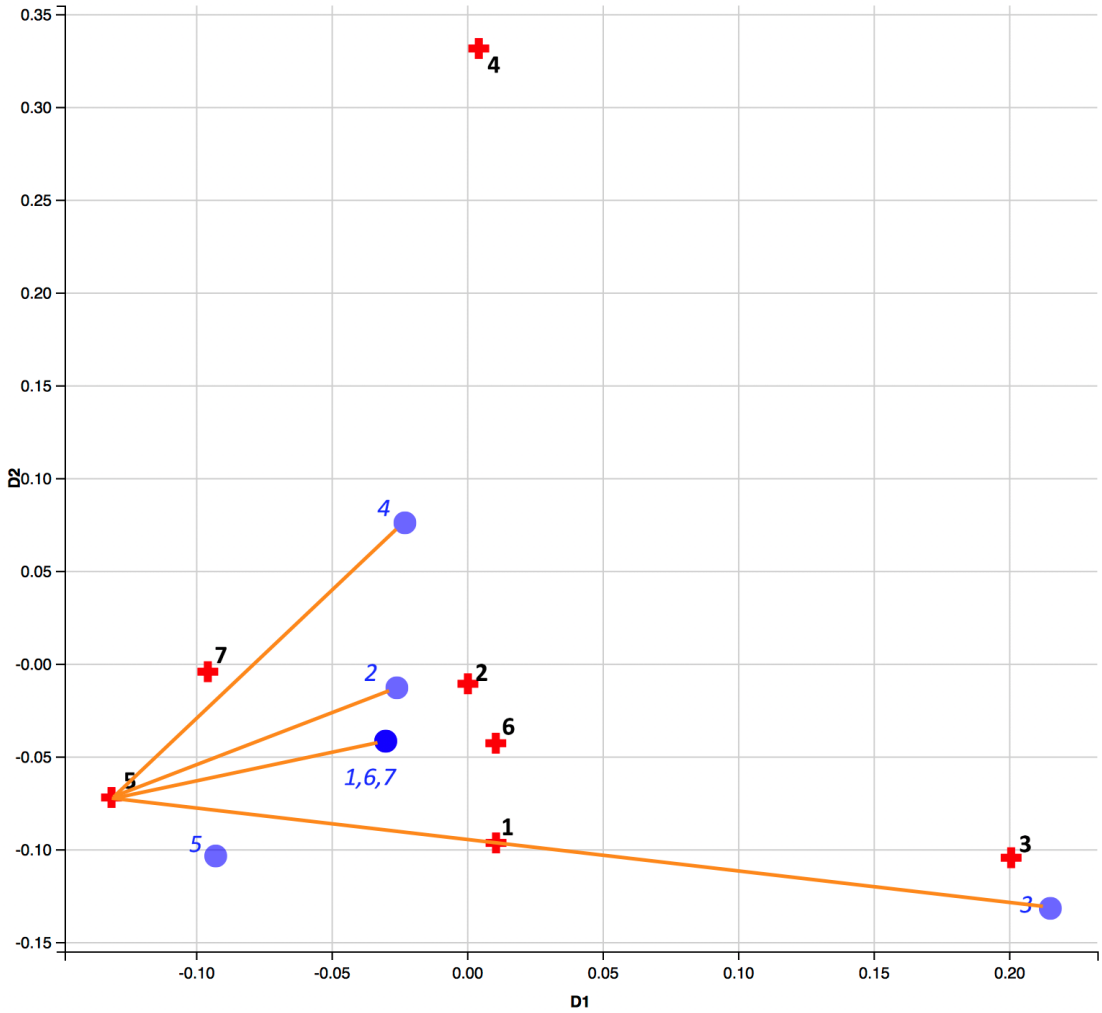

Figure 3.3: Column to row distances of column-object-5

Generally, each column-$object_i$ to row-$object_j$ distance is a function of a cross-efficiency score, the e_{ji} element of the corresponding CEM in Table 3.6. More accurately, every column-$object_i$ row-$object_j$ on the map is the approximately equal to (1-e_{ji}).[11]

Therefore, there is an indirect relationship between the column isolation index-i, the average column-to-rows distance value of generic DMU_i, and average cross-efficiency of DMU_i. It means the lower the average cross-efficiency, the higher column isolation index. In other words, the more isolated the column object from the row objects, the lower the average cross-efficiency of the column object. Low

[11]The rationale stems from the mechanism of multidimensional unfolding. Unfolding works with dissimilarity values, and CEM is a similarity matrix. One way of conversion of CEM into a dissimilarity matrix is 1-CEM, and this is the chosen way in this study.

average cross-efficiency means that the rated unit (column object) is evaluated as a relatively inefficient unit by majority of the rating units. This happens when most of the colleagues do not endorse the under-evaluation unit as an efficient unit. It is worth to pay attention to the point that the distance between column-$object_i$ and row-$object_i$ is not included in the formulation and computation of the column isolation index(CII). Column isolation index formulation follows the same pattern of average cross-efficiency formulation suggested by Doyle and Green (1994)

Consequently, high average distance between a column profile and row profiles means low average cross-efficiency, which subsequently means incompatibility of peers' optimum weights and the unit's input-output distribution. This situation happens either because the unit is severely under-performance from self-appraisal point of view, or it performs differently while being efficient from self-appraisal point of view. (inefficient while structurally common unit vs. a unit with significantly different production technology). Whether the unit is using a different production technology, or it suffers from inefficiency becomes clear by considering the location of its row object on the map regarding the rest of row objects.

If the unit's row object is close to the crowd, it means that its optimum weight scheme is similar to others, therefore the isolation of its rated role (column object) is because of inefficiency. If its row object is also far from the row objects crowd, it means that its optimum weight scheme is also different from the crowd, therefore the isolation of its rated role (column object) is because of different production technology, and not necessarily because of inefficiency.

Hence, the second average distance that one should consider is the average distance of each row profile from other row profiles, or briefly "row isolation index". Row isolation index (RII) shows the incongruence of row-object with the crowd of row-objects. This dis-harmony is due to different and special optimum weight schema of the DMU whose row-object is isolated. As an example, Figure 3.4 shows the distances from DMU5-Row object (rating role) from all other DMUs-Row objects (rating roles). Cyan lines depict the row-profile to row-profile distances.

In summary, in order to measure incongruence and discordant behaviour of a DMU, one should consider both aspects which have been explained above. If a unit's column object is far from the crowd of row objects, and simultaneously the unit's row object is also far from the crowd of row objects, then not only the output/input level distributions of the unit are different and unacceptable from the point of view of the majority of units, but also its optimum weight distribution, i.e. production technology, is different from what majority use. It seems more certain to determine a maverick unit considering both of these features instead of only one, as it happens in D&G MI.[12]

[12]Here one point must be clarified. It may seem appealing to use column profile to column profile distances as the measure of column profile isolation, similar to row profile isolation distance. However, we have used average cross-efficiency instead of such profile distances. There are two reasons behind this decision. First, the column profile distance does not mean low endorsement. Two profiles are close when their profiles' Euclidean distance are low, and two profile are far when their Euclidean distance is high. Hence, a column object may locate isolated from other column objects when the cross-efficiency scores of its profile is much lower than the other column profiles, or perhaps because the cross-efficiency scores of its profile is much higher than the other column

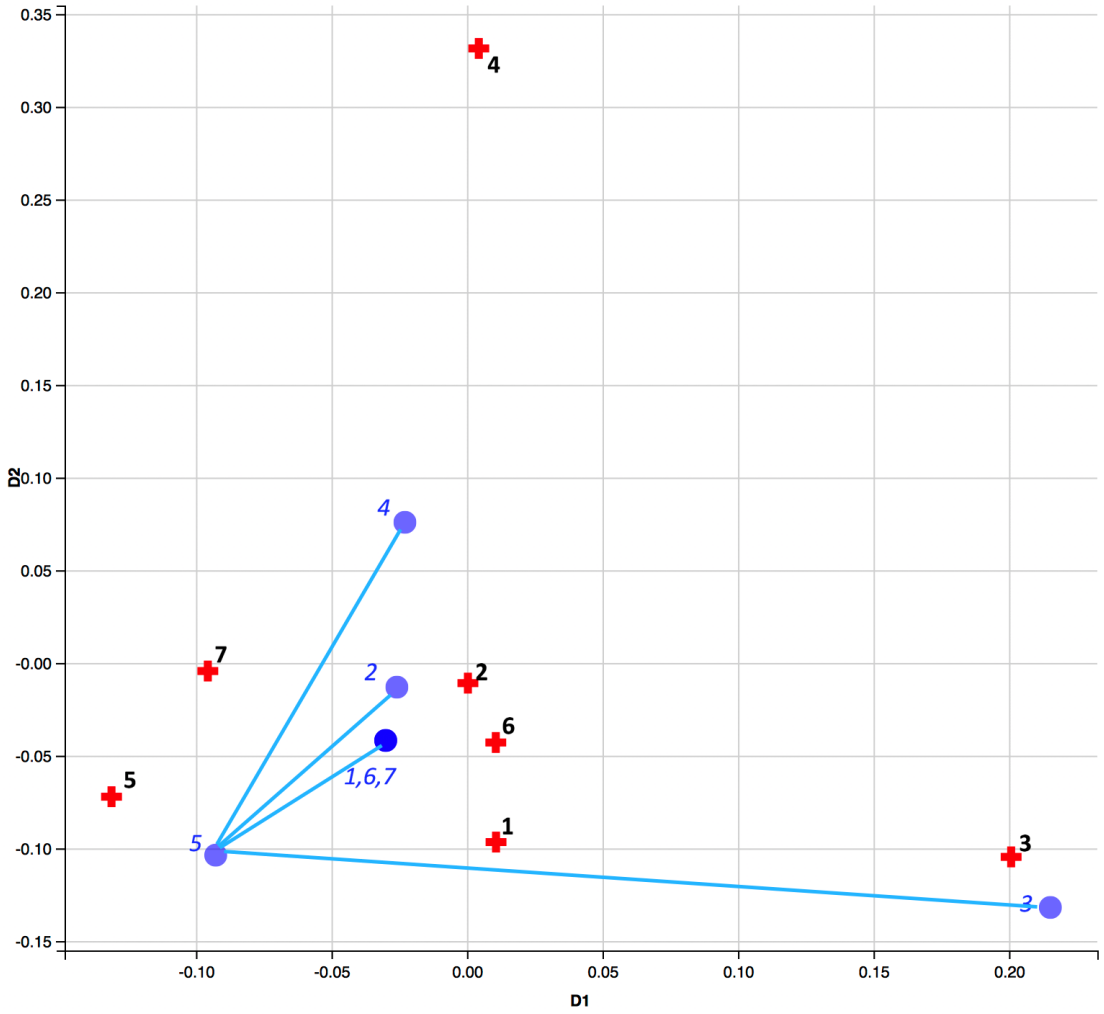

Figure 3.4: Row-to-row distance of row-object-5

Having been defined the two measures of row and column isolation indexes, one problem of the visual maverick-detection method is resolved. The new measures are quantitative, and they can augment the qualitative visual method.

The second issue, which is the more important of the two, is related to the stress in the 2dimensional map. Since the visual configuration is presented in 2 dimensions and any CEM of m DMUs has 2m dimensions[13], the dimension-reduction through multidimensional unfolding technique inevitably causes stress and noise in the 2d map. The higher the stress, the lower the precision of the map. Lower map precision

profiles. Hence, the interpretation of the location of column profiles in absence of locations of the row profiles is not helpful. However, not only the average cross-efficiency is well-established in the literature, but also its interpretation is fairly straightforward considering the relative location of corresponding row profile, as explained in this section.

[13]Any CEM is a two-mode matrix, because the row objects(units in their rating role) are not identical to the column objects(units in their rated role). Moreover, any CEM with m objects on rows (equivalently m objects on columns), can be transformed into a 2m*2m symmetric matrix. For more details: Borg and Groenen (2005)

is equivalent to inaccuracies in the locations of the row and column objects, and hence the relations among them, comparing to the CEM, the main data source. Even though, in some cases the noise in 2d map may be relatively low, there is no guarantee that such cases are achieved in visualization of every CEM.

It has been underscored that the visual maverick-detection method must be deemed as an exploratory tool. A tool in order to find the suspicious units. In other words, this visual tool is for finding evidence in order to direct further investigations rather than holding a trail and convicting any unit. The user of an exploratory tool is a detective and not a judge (Tukey 1977). Keeping all these cautions in mind reduces the probability of misleading by the inaccuracy of the map, or using the the exploratory tool in an inappropriate way. However, the probability of making mistakes due to inherent inaccuracy is still present. As a consequence, an index with high precision can augment the visual method.

Considering the two row and column isolation indexes, qualitative nature of visualization tool has been augmented with these quantitative tools. However, as explained above, in 2 dimensional space, these indexes may be imprecise. In order to significantly increase the precision of these two quantitative indexes, a simple yet effective solution is the computation of the row and column isolation indexes not in a space with reduced dimensionality, such as a 2d space, but in high dimensional space. In other words, unfolding the CEM in high dimensional space rather than low dimensional space, since to do so drastically reduces the imprecision.[14]

In high-dimensional space, the notions of individual distance, average column-isolation distances, and average row-isolation distances are identical to what were shown in Figure 3.3 and Figure 3.4 as well as what were explained correspondingly. The only difference is that the coordination of objects in, for instance, $2m-1$ dimensional space has $2m-1$ elements, hence it is not possible to effectively visualize the space and objects.

In summary, the author suggests row and column isolation indexes which both together can quantitatively measure the maverick-ness of a unit considering both rating(row) and rated(column) roles of the unit. In addition, a solution is suggested to significantly improve the precision of these measures, based on computation of the numeric indexes in high-dimensional space rather than 2D map. The last step is aggregation of those two measures. While it is possible to use the two measures simultaneously and to take them into account separately, it seems that a sole index would be more acceptable. This final aggregated index is what can be coined as "the new maverick index".

$$New\ Maverick\ Index_i = RII_i \times CII_i \tag{3.9}$$

The multiplication operation can be substitute with any other operation which aggregates the two indexes into one, based on the characteristic of the under-study problem. In the next section, the whole procedure of the New MI is applied to a real

[14]Any asymmetric matrix, such as CEM, with 2m objects (m row objects and m column objects) can be mapped in a 2m-1 dimensional space with close to zero stress, using multidimensional scaling (Borg and Groenen 2005). However, in the context of ratio unfolding, it is not always possible to reduce the stress through increment of dimensionality. This point is explained in the Appendix A.

dataset with 20 DMUs. The results are analyzed and compared to CEM visualization map as well as D&G MI.

3.2 Application of the New MI to a real dataset

In order to explain, both practically and conceptually, the procedure of generation of the new maverick index, the dataset from Tomkins and Green (1988) is used. The input/output levels and related explanations are presented in Table 3.6, where the DMUs are 20 university departments of accounting.

In order to venture on computation of the new maverick index, it is more helpful to start with 2d visualization of the benevolent cross-efficency(CEM) matrix of these 20 university departments of accounting. Figure 3.5 is this visual configuration.

Table 3.6: Inputs and outputs of 20 university departments of accounting (Tomkins and Green 1988). I1: Average full-time academic staff number, I2: Academic salaries, I3: Non-Salaries Expenditure, O1:Average number of Undergraduate studets, O2: Number of Research Postgraduates, O3: Number of Taught Postgraduates, O4: Research Council Income, O5: Other Research Income, O6: Other Income, O7: Number of Publications

Unit	**I1**	**I2**	**I3**	**O1**	**O2**	**O3**	**O4**	**O5**	**O6**	**O7**
D1	8.33	164	12	128.87	4	6	0	0	20752	35
D2	6,00,	127	10	83.35	0	0	0	0	0	29
D3	9.67	228	103	92.77	10	10	2162	96063	168256	36
D4	5.44	111	13	80.3	4	0	0	0	0	17
D5	8.17	181	15	119.85	3	14	0	4371	0	41
D6	13.00	292	41	112.69	4	25	150	4513	0	88
D7	10.00	154	58	120.58	7	69	0	2139	9219	79
D8	14.67	313	21	197.26	4	10	2182	583	0	48
D9	5.27	120	13	86.2	0	11	16486	20443	2881	73
D10	4.00	92	8	50.25	1	2	0	4637	0	22
D11	8.00	166	14	83.25	1	0	0	0	1102	59
D12	7.33	153	9	93.74	0	28	0	0	0	14
D13	7.00	143	12	147.9	4	0	0	385	0	13
D14	8.33	146	24	151.81	1	22	0	0	0	46
D15	13.00	221	31	208.5	16	31	1347	32514	9680	53
D16	13.00	266	47	134.5	1	41	0	8.000	0	28
D17	11.00	341	30	138.1	8	50	0	0	0	31
D18	14.33	288	29	156.37	5	54	0	1361	0	88
D19	8.00	149	20	206.7	2	39	0	7	0	18
D20	4.11	364	29	50.17	1	65	0	0	0	17

As it was explained in the second essay of this thesis, a CEM can be visualized using multidimensional unfolding. The result is presented in the Figure 3.5, where

the column objects are shown in red crosses and the row objects are shown in blue dots.[15]

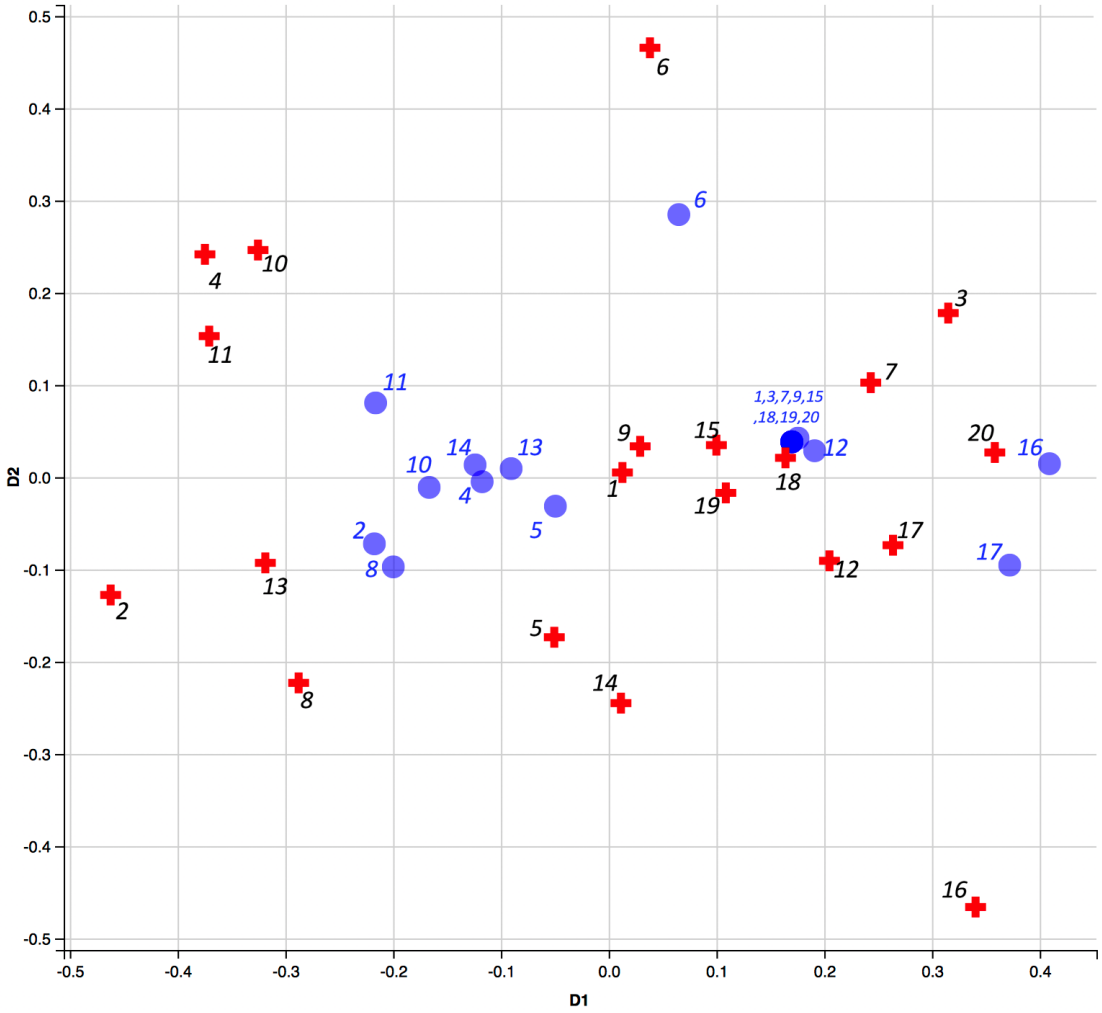

Figure 3.5: Unfolding map of benevolent CEM of 20 accounting departments

Through eye-balling the map, one can notice some of the objects have been located far from the crowd. The crowd here means the majority of the row objects. Although roughly four cluster of row objects can be distinguished at the first glance, two of them have more members than the other two, therefore closeness of a column object to either of them is a sign of crowd endorsement.

The most prominent discordant column objects seem to be column object 16, column object 6, column object 2 and to a less-extent a cluster of column objects including 4, 10, and 11. The isolation of a column object is reflection of the low average cross-efficiency score of it, and as it has been explained in the previous section, the mere isolation of a column object does not guarantee detection of a maverick. Every DMU has two aspects in a CEM, rating/rated or row/column, and

[15]Computational part of this visualization has been done by SMACOF package in R (J. d. Leeuw and Mair 2008) and the graphical part by GGVIS package of R (Chang and Wickham 2015).

both should be taken into consideration in order to evaluate the incongruence of a DMU.

Therefore, the next exploratory step is to check whether the corresponding row objects of these column objects also have been locate far from the crowd. To do so, one can notice that row object 6 as well as 16 and 17 are far from the crowds, in two different ways. The row object 6 is equi-distant from the two major clusters, while far from them. In contrast, the row objects 16 and 17 are very far from one major cluster and relatively far from the other, while located close to each other. It seems that while the optimum weight distribution of the unit 6 has a little and simultaneously equal similarity with the units of the two major clusters, the optimum weight distribution of the units 16 and 17 are very different from one of the two major clusters (the more sporadic one, the one on the left), relatively more similar to the other major cluster (the more dense one, the one on the right). The close adjacency of row units 16 and 17 is a sign of similarity of their optimum weight profiles. In summary, row objects 6, 16 and 17 seem discordant units among all 20 DMUs. Since the units in intersection of the uncommon row and column object sets should be considered as possible mavericks, the DMU 16 and DMU 6 are the suspicious units to high degree of maverick-ness, our top candidates from the visual tool.

As stated before, the visual map should be considered as an exploratory tool, a tool in order to find evidence rather than firm judgments and convictions. The reason is not only the qualitative nature of the map, but also the intrinsic inaccuracy of the object locations. The inaccuracy is the price that we pay in order to reduce the high-dimensionality of CEM to the familiar 2dimensions space.[16]

The new maverick index has been devised in order to improve this situation, based on this idea that the inaccuracy will be almost eliminated in the higher dimensional spaces. Therefore, the location of the objects would be as precise as possible, even though the price of this high precision is impossibility of effective visualization. While the objects' locations have been in the center of attention, indeed what we need to measure in order to compute the degree of incongruence is the distances. It is also easily possible to calculate the Euclidean distances in the high dimensional space.

Concisely, the idea of the new maverick index is based on measurement of the Euclidean distance of the objects in high dimensional space in order to calculate the average distance of any column object to row objects (column isolation index), as well as any row object to the rest of row objects(row isolation index). Afterwards, these two measures can be aggregated in different ways such as simple multiplication. Obviously, the multiplication of the two measures, in order to combine them into one index, is just one possible way of aggregation among many.

Therefore, the total steps for every DEA problem in order to achieve the new MI are as follows:

[16]The Stress index of Figure 3.5 is 0.34

1. Generation of benevolent CEM from the input/output data
2. Computation of object coordinates in high dimensional space, for instance in the space with $2m - 1$ dimensions where m is the number of DMUs
3. Check the stress index in order to be certain that the process is on the right track and the the units' locations in high dimensional space is highly precise.
4. Computation of inter-objects [Euclidean] distances in the high dimensional space
5. Computation of each isolation index based on CII (formulation 2.7) and RII (formulation 2.8). [a]
6. Computation of the new MI by multiplication of CII into RII

[a]The isolation index for a column object-i, is the average Euclidean distance of the object to all row objects, except to row object-i . Very similarly, the isolation index for a row object is the average Euclidean distance of the row object to other row objects. The isolation index for each column object is equivalent to the average cross-efficiency of that DMU.

For each DMU, Multiplication of the corresponding pair of isolation indexes. The result is the new maverick index.

Table 3.7 contains the row and column isolation index values as well as the amount of the new maverick index for each DMU. The table also has been sorted by the new MI in decreasing order. The highest new MI scores correspond to DMUs 16,6,2 and 11. These are the units with the most overall uncommon behavior. The results of the new MI confirm our educated guess based on the visual map. While DMUs 16 and 6 have been labeled as possible maverick units as a result of exploration of the visual map, the new MI shows high amount of maverick-ness for unit 2 and unit 11. These recent units are surprises, since their maverick-ness is not conspicuous on the map, specially considering their row objects.

According to the RII, DMUs 17, 6, 11 and 16 show the highest degree of row object uncommon-ness and oddity among all units. This uncommon behavior of the row object (rating aspect), is compatible with our findings from exploration of the CEM visual map of Figure 3.5, since row objects 17,16 and 6 are clearly far from the crowd, and row object 11 is also a bit isolated in the map. One should consider the inherent inaccuracy of the map, the cost of visualization, while comparing the numeric indices with the visual findings. In other words, the visual findings are estimations, and hopefully very good estimations in our method.

Based on the Table 3.7, DMUs 16, 2, 4 and 6 have the highest column isolation index. Previously we had labeled the units 16, 6, 2 as the isolated column objects based on eye-balling the CEM visual map in Figure 3.5. Moreover, we had pointed to a cluster of relatively isolated column units including unit 4. The new MI results of column isolation index confirms our previous educated guesses.

Table 3.8 presents the Doyle and Green's MI for the 20 units, ordered decreasingly by the MI score. Therefore, the highest maverick units according to the D&G MI are DMUs 2, 13, 4, and 11. Two units, 2 and 11, are in common with the findings of

Table 3.7: RII,CII and New MI of 20 Academic accounting departments

Unit	Row Isolation Index	Col Isolation Index	The New MI
16	1,3859	1,6459	2,281
6	1,4373	1,5379	2,2105
2	1,3297	1,6175	2,1509
11	1,3883	1,5309	2,1252
8	1,3297	1,5097	2,0074
4	1,2657	1,5698	1,9869
10	1,2725	1,5289	1,9454
17	1,4459	1,3388	1,9358
14	1,367	1,411	1,9288
13	1,264	1,48	1,8706
5	1,2191	1,3556	1,6525
3	1,1415	1,4169	1,6173
20	1,1415	1,3782	1,5732
7	1,1448	1,2955	1,4831
12	1,1341	1,2828	1,4548
18	1,1415	1,2159	1,3879
1	1,1415	1,1813	1,3484
19	1,1415	1,1632	1,3277
15	1,1415	1,115	1,2727
9	1,1415	1,1071	1,2637

the new MI while the other two, 13 and 4, are not in the top 4 maverick units of the new MI.

The new maverick indices for the 20 university departments of accounting as well as the corresponding Doyle and Green's MI are presented in the Table 3.9.

While the new MI confirms our CEM visual map explorations about DMU_{16} and DMU_6, maverick-ness of DMU_2 are DMU_{11} were not conspicuous. Interestingly, the Doyle and Green MI does not nominate $DMU_1 6$ and DMU_6 in top four maverick units, and instead emphasizes on units 13 and 4.

Although the row isolation index has no equivalent concept in the literature, the column isolation index values have the equivalent meaning to average cross-efficiency scores. The column isolation index of DMUi is the average distance of the DMU_i in its rated role (column profile) to all the DMUs in their rating roles (row profiles). Therefore, each individual distance between DMU_i (rated role) and DMUj (rating role) is equivalent to e_{ji} in CEM. Consequently, a DMU with the lowest average cross-efficiency has the most isolated and the farthest column object. The comparison between average cross-efficiency scores and column isolation index values is presented in Table 3.10.

The identical order of the two indexes, average cross-efficiency and column isolation index, is the sign of perfect compatibility of the conceptual and practical aspects of the column isolation index.

Table 3.8: The D&G MI of 20 Accounting Departments

Unit	**D&G MI**
2	0,9822
13	0,746
4	0,6573
11	0,6355
8	0,6141
3	0,5664
16	0,5482
20	0,4733
14	0,471
10	0,3852
17	0,3472
7	0,3065
6	0,304
12	0,2846
5	0,2736
18	0,1796
1	0,1297
19	0,1068
15	0,0495
9	0,0372

Having detected the top uncommon units, we can go further and examine different aspects of these units on their own and in contrast to other units. Since the outcome of the maverick indexes is uncommon units, our method is expected to shed light on the special units with discordant behavior. Furthermore, it was emphasized that the motivation of the maverick-detection is finding the units with in-congruent behavior in order to scrutinize the causes of their disharmonious behavior. To do so, one suggestion is investigation these units further based on the optimum weights which they have chosen under CEM benevolent formulation. Additionally, another complementary step can be evaluation of the units' "reality" and the compatibility of the findings of data investigation and that "reality". This reality is reflected in the levels of inputs and outputs of the units in addition to the strategy of the units as well as the production technology of them. While this step is of great importance, we have unfortunately no access to the background and other extra information about these 20 University departments of accounting in the dataset of Tomkins and Green (1988).

Table 3.9: Comparison of D&G MI and the New MI for 20 accounting departments

Unit	D&G MI	The New MI	D&G MI Rank	New MI Rank
1	0,1297	1,3484	17	17
2	0,9822	2,1509	1	3
3	0,5664	1,6173	6	12
4	0,6573	1,9869	3	6
5	0,2736	1,6525	15	11
6	0,304	2,2105	13	2
7	0,3065	1,4831	12	14
8	0,6141	2,0074	5	5
9	0,0372	1,2637	20	20
10	0,3852	1,9454	10	7
11	0,6355	2,1252	4	4
12	0,2846	1,4548	14	15
13	0,746	1,8706	2	10
14	0,471	1,9288	9	9
15	0,0495	1,2727	19	19
16	0,5482	2,281	7	1
17	0,3472	1,9358	11	8
18	0,1796	1,3879	16	16
19	0,1068	1,3277	18	18
20	0,4733	1,5732	8	13

Table 3.10: Comparison of average cross-efficiency scores and Column isolation index scores for 20 accounting departments

Unit	**Average Cross-Efficiency**	**Unit**	**CII**
16	0,3988	16	1,6459
2	0,4282	2	1,6175
4	0,4785	4	1,5698
6	0,5122	6	1,5379
11	0,5192	11	1,5309
10	0,5212	10	1,5289
8	0,5413	8	1,5097
13	0,5727	13	1,48
3	0,6384	3	1,4169
14	0,6444	14	1,411
20	0,6787	20	1,3782
5	0,7018	5	1,3556
17	0,7198	17	1,3388
7	0,7654	7	1,2955
12	0,7785	12	1,2828
18	0,8477	18	1,2159
1	0,8852	1	1,1813
19	0,9035	19	1,1632
15	0,9528	15	1,115
9	0,9642	9	1,1071

4 Conclusion

This study aimed to improve the shortcomings of DEA maverick literature and the most commonly used maverick index, Doyle and Green's MI. To do so, some theoretical adjustments and clarification have been suggested in order to circumvent the dual-role DMUs phenomenon, causing by multiplicity of weights, and in order to avoid misinterpretation of the outcome of maverick identification indexes, as these indexes are anomaly detection tools rather than maverick identification.

Furthermore, a new maverick index was proposed which detects exceptional units more comprehensively than common indexes such as D&G MI. The new maverick index evaluates DMUs from both rating and rated roles, i.e. row and column cross-efficiency matrix profiles, in contrast to D&G MI which only considers column profiles. In cross-efficiency matrices, a unit's row profile is a function of the unit's chosen optimum weight set, and therefore an uncommon row profile is a sign of an uncommon optimum weight set, and hence an in-congruent DMU from the rating aspect. Consequently, the new maverick index identifies the DMUs which are not only far from the crowd from the column aspect, but also far from the crowd from the row aspect. Alternatively, the new maverick index measures the uncommon-ness of each unit from two perspectives: how the unit evaluates other units, and how other units evaluate that unit. A unit which assesses others very differently, i.e. very different optimum weight set, and simultaneously it is assessed as a very unit, i.e. very unacceptable input and output levels, will receive a relatively high new maverick score. In addition, using benevolent cross-efficiency formulation forces DMUs to choose their most benevolent, i.e. congruent, optimum weights. Therefore, if a DMU under such condition shows in-congruent behaviour, then such DMU warrants more attention.

The main concept of the new maverick index has been stem from visualization of cross-efficiency matrices. However, unlike the visual map, which has inevitable noise due to dimension reduction of cross-efficiency matrices, the new index does not suffer from imprecision or noise, and thus it is very precise. Briefly, the new maverick index is based on multidimensional unfolding of the cross-efficiency matrices in high dimensional space.

At the end of this study, after explanation of the new maverick index side-by-side to well-established D&G index, the two indexes have applied on a real data set of 20 accounting academic departments from Tomkins and Green (1988).

Finally, the current study can be improved in several aspects. One of them is through replacement of the specific benevolent CEM formulation which has been used in this study and borrowed from Doyle and Green (1994) with a more appropriate one such as Maximum log cross-efficiency, suggested in Cook and Zhu (2014). Even though the benevolent CEM formulation of this study is just an approximation of the most benevolent weights for each unit, it has been used due to one reasons: the formulation is well-established in the literature. The other formulations which can be used in order to improve the current framework are the ones based on game-theory, such as Lim (2012). Moreover, it is very important to clarify the relation of outliers and mavericks, as two important anomalies in DEA. Hence, defining a framework

in order to discern the outliers based on the CEM unfolding map would be a very important progress in the field. In addition, Euclidean distance function can be replaced by any other distance functions such as Manhattan distance, in order to avoid the Euclidean's surprising behaviour in high-dimensional space.(Aggarwal et al. 2001)

Appendix A

The new maverick index stems from visualization of cross-efficiency matrices (CEMs). The visualization is done using multidimensional unfolding, a variant of multidimensional scaling family which can cope with asymmetric matrices. CEMs are not only asymmetric, but rectangular with two ways and two modes. For more about asymmetry, ways and modes, please see Schiffman et al. (1981) and T. F. Cox and M. A. Cox (2000). Moreover, the conditionality of the matrix must be considered in the unfolding process (Coombs 1964; De Leeuw 2005) Multidimensional Unfolding in is essentially similar to its parent, multidimensional scaling as both have metric and non-metric subversions, different transformation functions and a loss function to minimize in order to find the best possible configuration. The main difference lies in the fact that unfolding techniques are able to cope with the matrices in which the modes are not identical even though their difference is as subtle as difference in the roles of identical objects. These modes are usually individuals (on rows) and objects of interest (on columns). Therefore, the arrays of the proximity matrix are the measure of confirmation given by the individuals to the objects (Heiser and Busing 2004). Typically in MDU this measure of confirmations or endorsements, the proximity data, are preference rank- orders of the items given by individuals such that the unfolding models are categorized for preference and choice (De Leeuw 2005). The final configuration is a Euclidean joint map of both modes in which the individuals' locations are ideal points of them and more preferred items locate closer the judges (Borg, Groenen, and Mair 2012, p. 45).

The CEM can be seen as a proximity matrix, in such a way that the higher the cross-efficiency of i-j, the higher the endorsement of rating uni_i to rated $unit_j$. On the map, the higher the cross-efficiency of i-j, the closer the rating unit-i to the rated unit-j. Hence, the distances on the map ideally are reverse proportional to the cross-efficiency scores. However, similar to every other modelling, the noise or error should not be neglected. The error is usually measured through stress index. For in depth knowledge about unfolding, please see Borg and Groenen (2005)

Although there are some other statistical software packages such as SPSS with capability of performing Multidimensional unfolding, the computations of this study have been done by smacof package (J. d. Leeuw and Mair 2008) and the visualization has been done by ggvis package, both as packages of R statistical software.

The 2d unfolding maps, such as Figure 3.1 and Figure 3.5, have been generated under "ratio" Multidimensional Unfolding. (In order to know in detail about the ratio and interval metric MDS, please look at Borg and Groenen (2005) in general and its chapter 9, specifically). Since the 2d map has inevitable noise, in order to increase the precision of the Row Isolation Index, and Column Isolation Index, it seemed a good idea to increase the the number of dimensions and perform unfolding in high-dimensional space. However, there is no guarantee that increment of number of dimensions necessarily reduces the stress of unfolding outcome. This phenomenon is explained by the Gower rank of the matrices (J. d. Leeuw 2016). Thus, I have done all unfolding/asymmetric MDS computations in high dimensional space under interval setting. The non-reducible Stress phenomenon has been circumvented with

this subtle change, while no harm has been done to the study, theoretically and practically. Interval and ratio MDS are two variations of metric MDS, and closely related to each other.

The R markdown files and all codes are available by request. Please send an email to my address: Contact@Shahin-Ashkiani.com

References

Aggarwal, Charu C, Alexander Hinneburg, and Daniel A Keim (2001). "On the surprising behavior of distance metrics in high dimensional spaces". In: *ICDT*. Vol. 1. Springer, pp. 420–434.

Allen, Robert, Antreas Athanassopoulos, Robert G Dyson, and Emmanuel Thanassoulis (1997). "Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions". In: *Annals of Operations Research* 73, pp. 13–34.

Andersen, Per and Niels Christian Petersen (1993). "A procedure for ranking efficient units in data envelopment analysis". In: *Management science* 39.10, pp. 1261–1264.

Appa, Gautam, N Argyris, and H Paul Williams (2006). "A methodology for cross-evaluation in DEA". In: URL: `http://eprints.lse.ac.uk/22714/1/06081.pdf`.

Appa, Gautam and H Paul Williams (2006). "A new framework for the solution of DEA models". In: *European Journal of Operational Research* 172.2, pp. 604–615.

Avkiran, Necmi (2006). "Productivity analysis in the service sector with data envelopment analysis". In: *Available at SSRN 2627576*.

Baker, RC and Srinivas Talluri (1997). "A closer look at the use of data envelopment analysis for technology selection". In: *Computers & Industrial Engineering* 32.1, pp. 101–108.

Borg, Ingwer and Patrick JF Groenen (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Borg, Ingwer, Patrick JF Groenen, and Patrick Mair (2012). *Applied multidimensional scaling*. Springer Science & Business Media, p. 3.

Braglia, Marcello and Alberto Petroni (2000). "A quality assurance-oriented methodology for handling trade-offs in supplier selection". In: *International Journal of Physical Distribution & Logistics Management* 30.2, pp. 96–112.

Chai, Song, Yubai Li, Jian Wang, and Chang Wu (2013). "A genetic algorithm for task scheduling on NoC using FDH cross efficiency". In: *Mathematical Problems in Engineering* 2013.

Chang, W and H Wickham (2015). "ggvis: Interactive Grammar of Graphics". In: *R package version 0.4* 1.

Charnes, Abraham, William W Cooper, and Edwardo Rhodes (1978). "Measuring the efficiency of decision making units". In: *European journal of operational research* 2.6, pp. 429–444.

Cook, Wade D and Joe Zhu (2014). "DEA Cobb–Douglas frontier and cross-efficiency". In: *Journal of the Operational Research Society* 65.2, pp. 265–268.

Coombs, Clyde H (1964). "A theory of data." In:

Cooper, William W, Lawrence M Seiford, and Joe Zhu (2011). *Handbook on data envelopment analysis*. Vol. 164. Springer Science & Business Media.

Cox, Trevor F and Michael AA Cox (2000). *Multidimensional scaling*. CRC press.

De Leeuw, J (2005). *Multidimensional unfolding. Entry in the encyclopedia of statistics in behavioural science*.

Doyle, John and Rodney Green (1994). "Efficiency and cross-efficiency in DEA: Derivations, meanings and uses". In: *Journal of the operational research society* 45.5, pp. 567–578.

Farrell, Michael James (1957). "The measurement of productive efficiency". In: *Journal of the Royal Statistical Society. Series A (General)* 120.3, pp. 253–290.

Flokou, Angeliki, Nick Kontodimopoulos, and Dimitris Niakas (2011). "Employing post-DEA cross-evaluation and cluster analysis in a sample of Greek NHS hospitals". In: *Journal of medical systems* 35.5, pp. 1001–1014.

Fu, Yan, Dongdong Li, and Ning Li (2011). "Hotel Performance Evaluation Based on Cross-Efficiency DEA Models". In: *Management and Service Science (MASS), 2011 International Conference on*. IEEE, pp. 1–4.

Fumero, Francesca (2004). "Multiple solutions identification in data envelopment analysis". In: *Central European Journal of Operations Research* 12.3, p. 307.

Heiser, WJ and FMTA Busing (2004). *Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations*.

Lee, Zon-Yau and Chung-Che Pai (2011). "Operation analysis and performance assessment for TFT-LCD manufacturers using improved DEA". In: *Expert Systems with Applications* 38.4, pp. 4014–4024.

Leeuw, Jan de and Patrick Mair (2008). "Multidimensional scaling using majorization: SMACOF in R". In:

Leeuw, Jan de (2016). "Gower Rank". [Online; Accessed 10th Dec 2016]. URL: `http://gifi.stat.ucla.edu/gower/gower.pdf`.

Li, Ning (2008). "Real estate cross-efficiency measurement based on peer appraisal DEA model and method in main cities of China". In: *2008 International Conference on Management Science and Engineering 15th Annual Conference Proceedings*, pp. 1667–1673.

Lim, Sungmook (2012). "Minimax and maximin formulations of cross-efficiency in DEA". In: *Computers & Industrial Engineering* 62.3, pp. 726–731.

Lotfi, Farhad Hosseinzadeh, Adel Hatami-Marbini, Per J Agrell, Nazila Aghayi, and Kobra Gholami (2013). "Allocating fixed resources and setting targets using a common-weights DEA approach". In: *Computers & Industrial Engineering* 64.2, pp. 631–640.

Lu, Wen-Min and Shih-Fang Lo (2007). "A closer look at the economic-environmental disparities for regional development in China". In: *European Journal of Operational Research* 183.2, pp. 882–894.

Ma, Chaoqun, Debin Liu, Zhongbao Zhou, Wei Zhao, and Wenbin Liu (2014). "Game cross efficiency for systems with two-stage structures". In: *Journal of Applied Mathematics* 2014.

O'Neill, Liam (1998). "Multifactor efficiency in data envelopment analysis with an application to urban hospitals". In: *Health Care Management Science* 1.1, pp. 19–27.

O'Neill, Liam and Franklin Dexter (2005). "Evaluating the efficiency of hospitals? perioperative services using DEA". In: *Operations research and health care*. Springer, pp. 147–168.

Santos, Jorge and Isabel Themido (2001). "An application of recent developments of data envelopment analysis to the evaluation of secondary schools in Portugal". In: *International Journal of Services Technology and Management* 2.1-2, pp. 142–160.

Sarkis, Joseph (2000). "An analysis of the operational efficiency of major airports in the United States". In: *Journal of Operations management* 18.3, pp. 335–351.

Sarkis, Joseph (2001). "Ecoefficiency: How data envelopment analysis can be used by managers and researchers". In: *Intelligent Systems and Smart Manufacturing*. International Society for Optics and Photonics, pp. 194–203.

Schiffman, Susan S, Forrest W Young, and M Lance Reynolds (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*.

Serrano-Cinca, Carlos, Yolanda Fuertes-Callén, and Cecilio Mar-Molinero (2005). "Measuring DEA efficiency in Internet companies". In: *Decision Support Systems* 38.4, pp. 557–573.

Sexton, Thomas R, Richard H Silkman, and Andrew J Hogan (1986). "Data envelopment analysis: Critique and extensions". In: *New Directions for Evaluation* 1986.32, pp. 73–105.

Team, R Core (2016). *R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014*. URL: `https://www.R-project.org/`.

Tofallis, Chris (2001). "Combining two approaches to efficiency assessment". In: *Journal of the Operational Research Society* 52.11, pp. 1225–1231.

Tofallis, Chris (2010). "Multicriteria ranking using weights which minimize the score range". In: *New Developments in Multiple Objective and Goal Programming*. Springer, pp. 133–140.

Tomkins, Cyril and Rodney Green (1988). "An experiment in the use of data envelopment analysis for evaluating the efficiency of UK university departments of accounting". In: *Financial Accountability & Management* 4.2, pp. 147–164.

Tukey, John Wilder (1977). *Exploratory Data Analysis*. Addison-Wesley. ISBN: 0201076160.

Wang, Ying-Ming and Kwai-Sang Chin (2010). "A neutral DEA model for cross-efficiency evaluation and its extension". In: *Expert Systems with Applications* 37.5, pp. 3666–3675.

Wong, Y-HB and JE Beasley (1990). "Restricting weight flexibility in data envelopment analysis". In: *Journal of the Operational Research Society*, pp. 829–835.

Article 4

DEA-Viz: A Software for Visualization of Data Envelopment Analysis Problems

DEA-Viz: A Software for Visualization of Data Envelopment Analysis Problems

Abstract

Visualization of data envelopment analysis (DEA) problems can be a very informative step in order to get insight into the problems and their characteristics. Through the various visualization methods, one can literally look at the DEA problem and the related data from different perspectives. Doing so would help to efficiently identify possible regularities and patterns, as well as irregularities and anomalies in the data.

Despite its importance and benefits, visualization is vastly neglected in the DEA applications. Even though there are several suggested methods for DEA visualization in the literature, the graphical investigation step is absent in the majority of the studies. The reason may lie in the fact that the DEA software packages severely lack any high-dimensional visualization features.

This paper is an introduction to DEA-Viz, a new DEA software which is mainly focused on visualization of the DEA problems. The software is a free-to-use cloud-based applet that can shed light onto a DEA problem using various visualization methods, and thus from various aspects. The intuitive design of the DEA-Viz helps researchers to generate diverse plots of their datasets, in order to gain insight into the problems and augment the quantitative analysis. The software can be found here: https://ashkiani.shinyapps.io/dea-viz/

Keywords: Data Envelopment Analysis, Data Visualization, Anomaly Detection, Software

1 Introduction

The goal of this paper is introduction of DEA-Viz, a data envelopment analysis (DEA) software specialized in visualization. In order to do so, it seems necessary to begin with an introduction to data visualization, and its necessity in general. Therefore, this section reviews data visualization, and tries to address its necessity, benefits and purposes.

Ward et al. (2015) define visualization as "communication of information using graphical representations". In this sense, visualization has been with human-being for millenniums. Even if we confine visualization to quantitative data visualization, and define it as graphical representation of the numerical data, it has been around for centuries. Figure 4.1 shows "the time distribution of events considered milestones in the history of data visualization" (Friendly 2008).

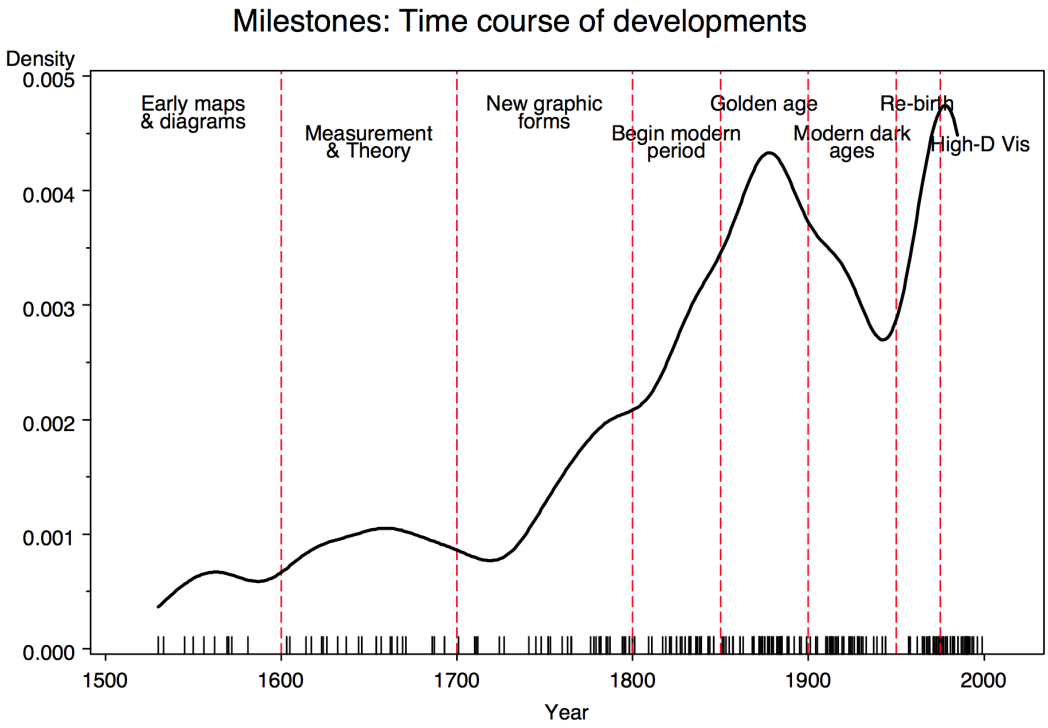

Figure 4.1: Distribution of data visualization milestone events over time

Nonetheless, the topic has received much more attention in recent decades. Chen et al. (2008) argue that unprecedented computation power of early computers that was used in quantitative modelling had attracted all the attentions, but recently part of this attention is re-directed towards graphics.

Data visualization makes us able to literally look at the data, and it makes the data "more easily and effectively handle-able by minds" (Tukey 1977). In fact, "we acquire more information through vision than through all the other senses combined", and "20 billion or so neurons of the brain devoted to analyzing visual information provide a pattern-finding mechanism that is a fundamental component in much of our cognitive activity." (Ware 2012, p. 2). Hence, Visualization, as a cognitive tool, sharpens our thinking ability, and improves problem perception.

While some of the findings of a proper visualization are also possible to be discovered through quantitative analytic methods, there may be some "emergent properties" appeared through visualization which otherwise couldn't be found using analytical approaches. In the introduction of this thesis, two simple examples of such situations, including Anscombe's quartet Anscombe (1973), are presented. Nevertheless, the emphasis on visualization should not overshadow the importance of quantitative analysis, since these are complementary approaches to the reality of the data.

Overall, Ware (2012) enumerates some of the benefits of visualization as follows:

1. Representation of large amount of data: Visualization enables to relatively easily and quickly understand huge amount of data.

2. Emergent properties: Emergent property is a characteristic of the whole, i.e. components and their relations. Proper visualization, specially high-dimensional visualization, tries to retain data, and keep both components and their relations. Therefore, it is very probable to see the emergent properties of the data through visualization, while reductionist and analytical approaches may not able to reveal such characteristics, since they do not even try to preserve all the information.

3. Visibility of the data peculiarities: If there is any problem, such as error or artifact, with the data, it can be quickly revealed using appropriate visualization.

4. Visibility of the data features: Visualization helps to detect and understand the patterns and features of the data at both large, and small scales. In other words, we can see the big picture and delve into the details using various visualizations.

5. Raising of new questions: A good visualization makes us able to think more profoundly about the data, and the related problem. Such deep thinking often includes re-evaluation of data and the problem assumptions, which had been taken for granted. Moreover, new questions and hypotheses raise by detection of the features and occurrences in the visual map that we don't have any answer for them.

Based on its purpose, visualization can be categorized into the following four categories (Ward et al. 2015, p. 46):

- Exploration: The visual map is prepared to help its users to investigate a dataset. The investigation is either for a general goal such as familiarity with features of the data, or more specific goal, such as finding outliers.

- Confirmation: The visualization is used as a consequent to quantitative analysis, and the goal is verification of some already shaped conclusions.

- Presentation: The visualization is for conveying message to its audience. Hence, the important findings are determined, and the goal is presentation of them to possible audience.

- Interactive Presentation: The visualization is for presentation but in an interactive mode, so the users can be involved in the process. Doing so helps to engage the audience, and encourage them to explore.

Considering above categorization, DEA-Viz is a mixture of the first and the last categories. DEA-Viz is for exploration of the data, to get insight into it, look at it from various perspectives, and find subjectively interesting points in it. DEA-Viz is an interactive applet, such that the users can zoom-in and get further information about visualization components and explore in the data interactively. Finally, the findings can be downloaded and presented to probable audience, although the main purpose of the applet is exploration.

As stated above, the general goal of using a visualization tool, such as DEA-Viz, is gaining insight into the data through interactive exploration. Such insight is gained by answering various kind of questions. Telea (2014, pp. 4-7) explains two types of questions that can be asked to gain insight:

- concrete questions: It is when our questions are clearly formulated before visualization process. For instance, when we want to know more about a specific observation, or we want to detect some specific outliers, we can use visualization to efficiently find the answers.

- open-end exploration: It is when we are not very familiar with the data, and we have no single precise question about it. What we want to do is looking at the data to know it better, to investigate it, and to hunt interesting points in it. The interesting points can be about the distribution, the anomalies, the clusters, the structure of the data, or more specific domain-related issues. Telea (2014, p. 6) adds that "This role of visualization closely matches the perspective of a researcher who is interested in studying a phenomenon in order to find out novel facts and establish unexpected correlations."

DEA-Viz is designed to answer both of these question types, however our approach would be closer to the "open-end exploration".

All being said, it is of great importance to emphasize that visualization is a complementary tool to quantitative methods, and not a rival to them. In the words of Tukey (1977): "Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone–as the first step."

2 Visualization of DEA Problems

While visualization is an indispensable step in quantitative studies nowadays, its absence in data envelopment analysis studies is conspicuous. For instance, there is no visualization chapter or section in the DEA handbook (Cooper et al. 2011), despite the handbook's comprehensiveness. The reasons of such absence perhaps are three-fold:

1. Multidimensionality of DEA problems: In order to use visualization in DEA problems, one has to decide which data set is intended to be visualized. The data sets of DEA problems, such as the inputs and outputs dataset or the the dual multipliers profiles of the DMUs, are oftentimes high-dimensional. Adler and Raveh (2008) argues that the reason of visualization absence in DEA problems is such high-dimensionality, which causes difficulty in graphical presentation of the problems.

2. Doubt about benefits of visualization: Another possible reason of not using data visualization in DEA is considering the visualization a superfluous step, since the wide variety of DEA quantitative methods generate the expected results. In other words, the added-value of the visualization is not clear. It

seems that researchers and users are suspicious to qualitative visual maps, comparing to exact precise numbers.

3. Lack of high-dimensional visualization features in DEA software packages: even if a researcher wants to use visualization methods, non of the current DEA software packages can be helpful in case of high-dimensional data visualization, so the researcher must code the visualization methods from scratch in a programming language. Statistical packages such as SPSS have some features for dimension-reduction and data visualization, however these features are limited to the pre-defined ones, and expansion of them requires programming as well.

This paper and DEA-Viz applet intend to address all these three issues. Firstly, we are not helpless in the face of multidimensionality of DEA problems, since several methods are already available in DEA literature to cope with this difficulty, using dimension-reduction techniques such as principal component analysis(PCA), multidimensional scaling(MDS) and Self-Organizing Maps(SOM) or techniques such as parallel coordinates. After about two decades of delay from the beginning of high-dimensional data visualization trends, DEA researchers have adopted and assimilated these techniques into DEA realm.

In the previous section the benefits of visualization have been enumerated. Nevertheless, in the rest of this paper, in order to emphasize on and clarify the added-values of the visualization, a dataset is visualized, and the graphs are explored to find what is either hidden from DEA quantitative methods, or easier to identify using visual analysis. Through visual exploration, the trends, clusters, relations, as well as anomalies, which could remain uncovered otherwise, can be detected. Also, new questions would be generated through assessment of the assumptions about the data or the DEA models.

The third possible cause of unpopularity of visualization in DEA may lie in the fact that the current DEA packages, to the best of author's search and knowledge, are limited to uni-variate or bi-variate visualizations. In other words, the visualization features of the current software packages lacks any high-dimensional data visualization capability, while DEA problems with more than two DMUs or more than two input/output variables are inevitably high-dimensional. Table 4.1 includes the list of some of the current DEA packages, and their visualization outcomes.

As one can see, beside the software packages without any visualization output, some have very limited visualization features, specially due to lack of high-dimensional data visualization methods. High-dimensional visualization methods tends to retain the relations among variables, and "holistically" visualize a given dataset. Visualization of a multi-variable dataset in a reductionist manner, similar to understanding of a complex system through examination of its components regardless of their relations, is prone to yielding incomplete understanding. If we consider each variable as a component, then the relation between components are preserved by high-dimensional visualization of the dataset. However, visualization of each variable separately is conspicuously ignoring the relationship between the components, and thus its related information.

Table 4.1: Visual Outputs of some prominent DEA softwares

No.	Name	Web Address	Platform	Visualization Features
1	Frontier Analyst	http://banxia.com/frontier/	Stand-alone software for MS Windows	- Univariate plots (e.g. reference-set frequency, efficiency scores barchart, slack variables pie chart) - Bi-variate plots (e.g. efficient frontier on scatterplot, scatterplot of variable correlation, multiple bar-chart of variable comparison between DMU pairs) - Multi-variate (Radar-chart of unit variables)
2	PIM-DEAsoft	http://www.deasoftware.co.uk/	Stand-alone software for MS Windows	- Uni-variate plots (e.g. efficiency scores bar-chart) - Bi-variate plot (e.g. efficient frontier scatterplot efficiency scores time-series line-chart, Malmquist index time-serie)
3	DEAP	http://www.uq.edu.au/economics/cepa/deap.php	Stand-alone software for MS Windows	None
4	DEAFrontier	http://www.deafrontier.net/deasoftware.html	Ms Excel add-on for MS Windows	None
5	Open Source DEA	http://opensourcedea.org/	Stand-alone open source software working on Ms Windows, OSx, and Linux	None
6	Benchmarking	https://cran.r-project.org/web/packages/Benchmarking/index.hml	Package for R	- Uni-variate plot (e.g. density plot of efficiency) - Bi-variate plot (e.g. scatterplot of transformation curve, isoquant or production function)
7	DEA-Solver Pro	http://www.saitech-inc.com/products/prod-dsp.asp	Ms Excel add-on for Ms Windows	Uni-variate plot (efficiency scores bar-chart)
8	MaxDEA	http://www.maxdea.cn	Ms Windows and Ms Access	Bi-variate plot (e.g. scatterplot of efficiency frontier)
9	DEA Toolbox	http://www.deatoolbox.com/	MATLAB library	None

In DEA, it is very common to plot the efficient frontier in a graph of two chosen variables, inputs or outputs, from the set of all variables. Doing so is a reductionist approach to a high-dimensional dataset, since it does not retain all the components, and the relations among all them.

Since in almost all cases of DEA problems, the number of variables, such as inputs and outputs, are more than two, dimension-reduction techniques are used in order to reduce their dimensionality into lower dimensions, usually 2 dimensions. The cost of such dimension-reduction is losing some information in the data, however these approaches at least try to preserve the information, while reductionist approaches do not try to do so.

Despite this crucial importance of high-dimensional visualization, none of the current DEA software packages have proper high-dimensional visualization features. Hence, if a researcher believes in the power of high-dimensional visualization, none of the current software packages can help to do so, and the only available option would be coding the visualization process from scratch. Doing so not only is time-consuming, but also demands coding knowledge. Possibly this obstacle, beside the other stated reasons, dissuade researchers from visualization of the DEA problems.

This paper, and the pertaining software, DEA-Viz, intend to address all three possible causes of visualization neglect. Multidimensionality of DEA problems is not an obstacle anymore, since the DEA-Viz software includes several data visualizations for high-dimensional DEA data. Additionally, The DEA-Viz does not demand its users to know any programming language, and its intuitive graphical user-interface makes working with the package easy. Furthermore, through the rest of this section as well as the illustration of the software with a real dataset, the author is hopeful to motivate DEA researchers to benefit from data visualization in their studies. Thus, promotion of data visualization in DEA field is another goal of this paper and DEA-Viz.

3 DEA-Viz

In this section, DEA-Viz and its capabilities are introduced using a real dataset. However, before going to show the some features and outcomes of the applet, a concise introduction is presented in the following subsection.

3.1 Introduction

Ward et al. (2015, p. 6) state that "In virtually any domain, visualization can be, and is becoming, an effective tool to assist in analysis and communication." DEA-Viz is born to assist analytic exploration of the DEA problems, and communication of the possible results.

DEA-Viz is developed to ease and promote using visual exploration in DEA, and its focus is on high-dimensional data visualization. As it is shown in Table 4.1, high-dimensional data visualization in particular is totally ignored in the current DEA softwares packages. The available features to visual high-dimensional data in the current packages are limited to radar charts, a sort of chart with a very

limited capability in visualization of multiple DMUs. Nevertheless, the role of the high-dimensional data visualization is crucial in the DEA problems, which have multi-inputs and multi-outputs as well as many outcome variables such as multiplier weights, efficiency scores, optimum weights, slack values and so on.

High-dimensional visualization methods based on dimension-reduction techniques, such as principal component analysis(PCA), multidimensional scaling(MDS), and self-organizing maps(SOM), are the foundation of DEA-Viz plots. These visualization methods have been presented in the DEA literature, but remained unimplemented, and thus vastly unavailable to the users. Through these various methods, DEA practitioner can get insight into its problem, evaluate the structure of its data, and detect anomalies. Soukup and Davidson (2002) states that "visualizations help reduce your time-to-insight-the time it takes you to discover and understand previously unknown trends, behaviors, and anomalies" comparing with other analytic tools. DEA-Viz is developed partly to decrease such time-to-insight, and partly to yield insights that other tools fail to yield.

Nevertheless, the toolbox of DEA-Viz is not confined to high-dimensional methods. Simple distributions of inputs and outputs variables as well as efficiency scores are presented beside the scatter-plots of every pair of these variables. It is important to emphasize that these visualization tools, either uni-dimensional, bi-dimensional or high-dimensional, are mostly complementary methods. In other words, these various methods attack the data from complementary perspectives, and while they have overlaps of method and data, they still enable us to look at the problem from different perspectives. Hence, there is no "the best plot" in general, and all plots can be useful in their position, and can contribute to the understanding of the problem and gaining insight.

DEA-Viz plots are mostly interactive, so the user can get further information about selected parts of a map, zoom-in into the over-crowded areas of a map, manipulate the transparency and size of the graph objects and so on. All these features are added in order to overcome the limits of classic bi-dimensional visualization. In this sense, DEA-Viz can be considered as a modern visualization tool (Ward et al. 2015, p. 30).

DEA-Viz is a platform free software, it runs on cloud and does not need any particular operating system to run in. Working with DEA-Viz does not demand any programming language, but if one knows R, one can assess and manipulate the source code of DEA-Viz, since the package is presented under AGPL 3.0 license.

Ward et al. (2015, p. 6) state that "In virtually any domain, visualization can be, and is becoming, an effective tool to assist in analysis and communication.", and DEA-Viz is developed to be an effective DEA visualization tool in analytic data exploration and communication of the findings of this exploration.

3.2 Exploration of a Real Dataset

In order to illustrate the features of DEA-Viz, a real dataset is visualized and explored in this section. The dataset is the well-known dataset of 35 Chinese cities, which have been used in some DEA visualization studies (e.g. Costa et al. 2016; Adler and

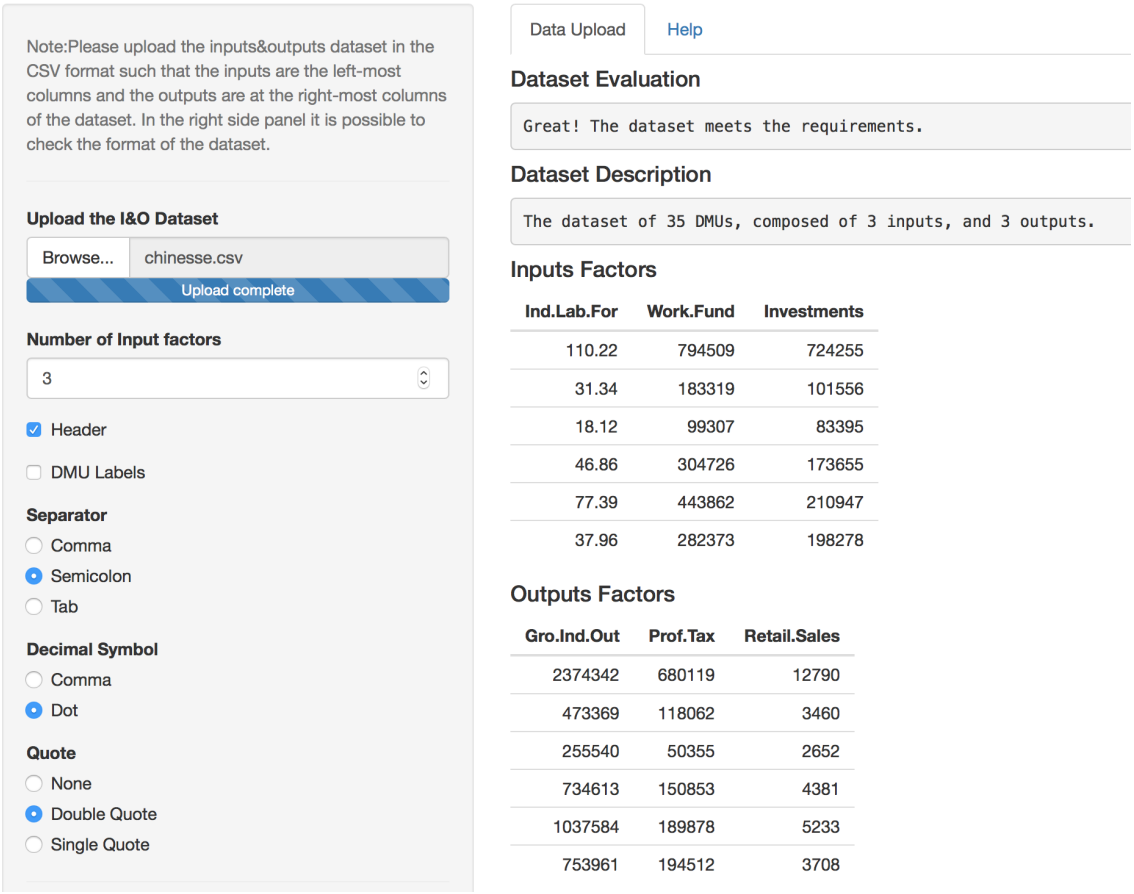

Figure 4.2: Importation of Chinese Cities Dataset into DEA-Viz

Raveh 2008), and some non-visualization DEA studies (e.g. Sueyoshi 1992). Since this dataset has been used in previous DEA visualization studies, it connects this study to the previous ones, and makes evaluation, comparison, and validation easier.

The dataset, composed of 35 DMUs with three inputs and three outputs, is presented in Table 4.2.

3.2.1 Data Importation

The DEA-Viz has two main sections: Upload Data and Plots. The former is dedicated to data importation, and it has some limited data manipulation features. The data must be in the "comma separated values (CSV)" format. In addition the values must be all numerical, and non-negative. Missing values are not allowed as well. Figure 4.2 shows importation of the 35 Chinese Cities dataset into the applet.

The structure of each subsection of DEA-Viz is similar to Figure 4.2. Each subsection has a side-panel, which includes all the features related to that subsection, and a main panel, which includes the outcomes such as tables and visual maps. Moreover, some subsections have main panels with several tabs, including a "Help" tab. The help tab includes necessary manual of its corresponding subsection.

Table 4.2: 35 major cities of China, from Adler and Raveh (2008)

DMU No.	DMU Name	Inputs			Outputs		
		Industrial Labour Force	Working Funds	Investments	Gross Industrial Output	Profit and Tax	Retail Sales
1	Beijing	110.22	794509	724255	2374342	680119	12790
2	Changchum	31.34	183319	101556	473369	118062	3460
3	Changsha	18.12	99307	83395	255540	50355	2652
4	Chengdu	46.86	304726	173655	734613	150853	4381
5	Chongqing	77.39	443862	210947	1037584	189878	5233
6	Dalian	37.96	282373	198278	753961	194512	3708
7	Fuzhou	16.03	96623	103560	222634	43984	2222
8	Guangzhou	50.92	389641	354879	1154147	275588	8362
9	Guiyang	17.52	101368	76476	257718	72917	1118
10	Hangzhou	34.32	212524	120028	726172	159354	4106
11	Harbin	48.2	356752	138972	672427	124508	3856
12	Hefei	17.02	95076	56690	270087	62387	1486
13	Hohot	10.15	56096	42493	127132	33069	852
14	Jinan	26.39	152034	78312	441724	109039	2441
15	Kunming	27.59	168224	112871	439756	117719	2148
16	Lanzhou	35.89	235416	107328	580669	140557	2151
17	Lhasa	0.44	1908	7394	1665	286	398
18	Nanchang	23.3	129132	42700	317158	61472	1663
19	Nanjing	42.18	269246	222623	836544	208006	3779
20	Nanning	11.33	59166	36627	176140	139399	1253
21	Ningbo	13.52	69895	72845	320516	74492	2866
22	Shanghai	206.73	1577603	959226	6743346	1880041	18316
23	Shenyang	68.62	419358	198494	1017454	195987	5072
24	Shenzhen	3.41	35478	278230	78313	11461	2778
25	Shijiazhuang	24.88	138931	68661	453445	94216	1745
26	Taiyuan	38.34	221065	170776	513907	84812	1896
27	Tianjin	98.38	628243	541587	2252611	538202	6895
28	Urumqi	15.97	94622	130771	204232	38294	1323
29	Wuhan	64.87	442813	183811	1218527	295199	5090
30	Xiamen	6.48	46821	97627	130646	37698	1109
31	Xian	46.2	294539	140906	635575	101261	3292
32	Xining	10.54	74188	45629	100509	18627	858
33	Yinchuan	5.06	30959	46014	59757	11458	524
34	Zhengzhou	28.09	144141	86791	413025	105784	2359
35	Zhuhai	1.26	12504	86457	35760	6667	1046

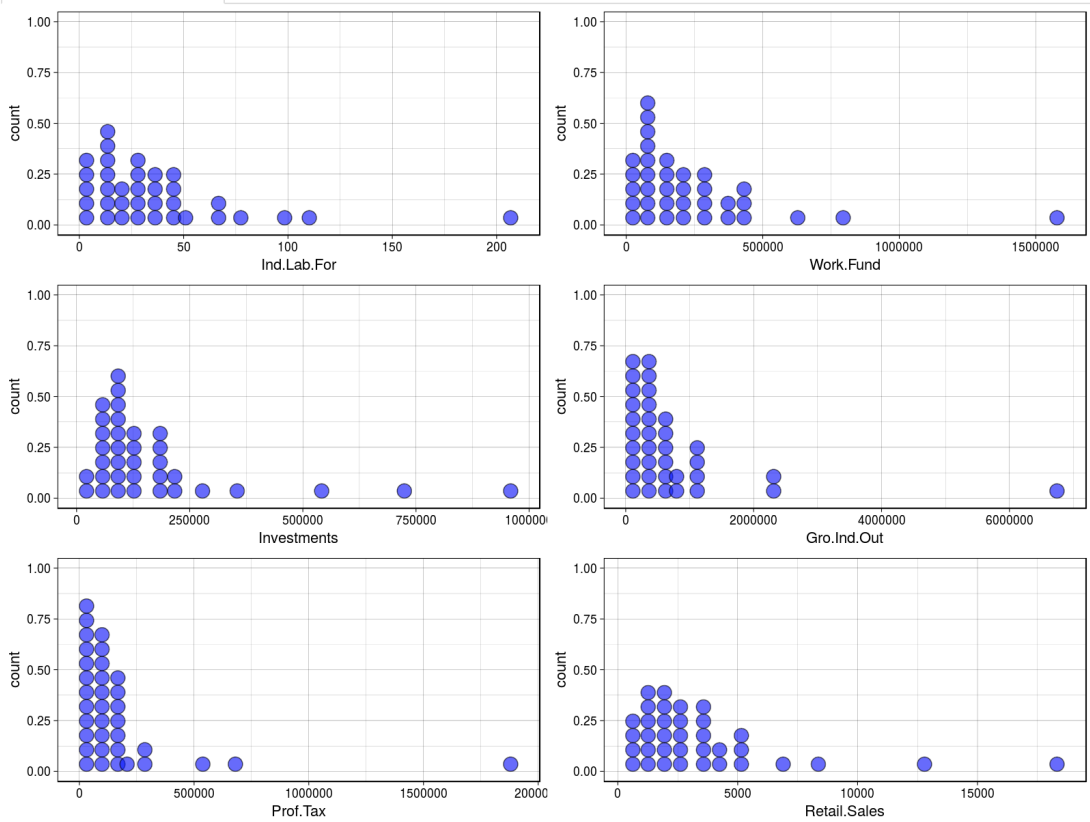

Figure 4.3: Distributions of inputs and outputs

3.2.2 Distributions

The first subsection of the "Plots" section is "distributions". Distributions include Inputs and Outputs distribution dotplots as well as the dotplots of various efficiency scores such as constant return to scale (CRS), variable return to scale (VRS), free disposal hull (FDH). While these plots are uni-variate plots, they are very good starting points to get insight into the data.

Figure 4.3 is composed of the distributions of the inputs and outputs and Figure 4.4 is composed of CRS, VRS, and FDH distributions. In these dotplots, each dot represents one DMU.

In Figure 4.3, it is seen that most of distributions are right skewed. There are some DMUs with much higher input or output values comparing to the rest, but most of the units have approximately similar values for their parameters. The magnitude of these variables define the scale of the units.

It is possible in the "distribution section" to highlight a specific DMU from the rest. However, currently there is no need to do so, since we have not designated any specific DMU to focus on. However, we can come back to this distribution section when we have such unit of interest.

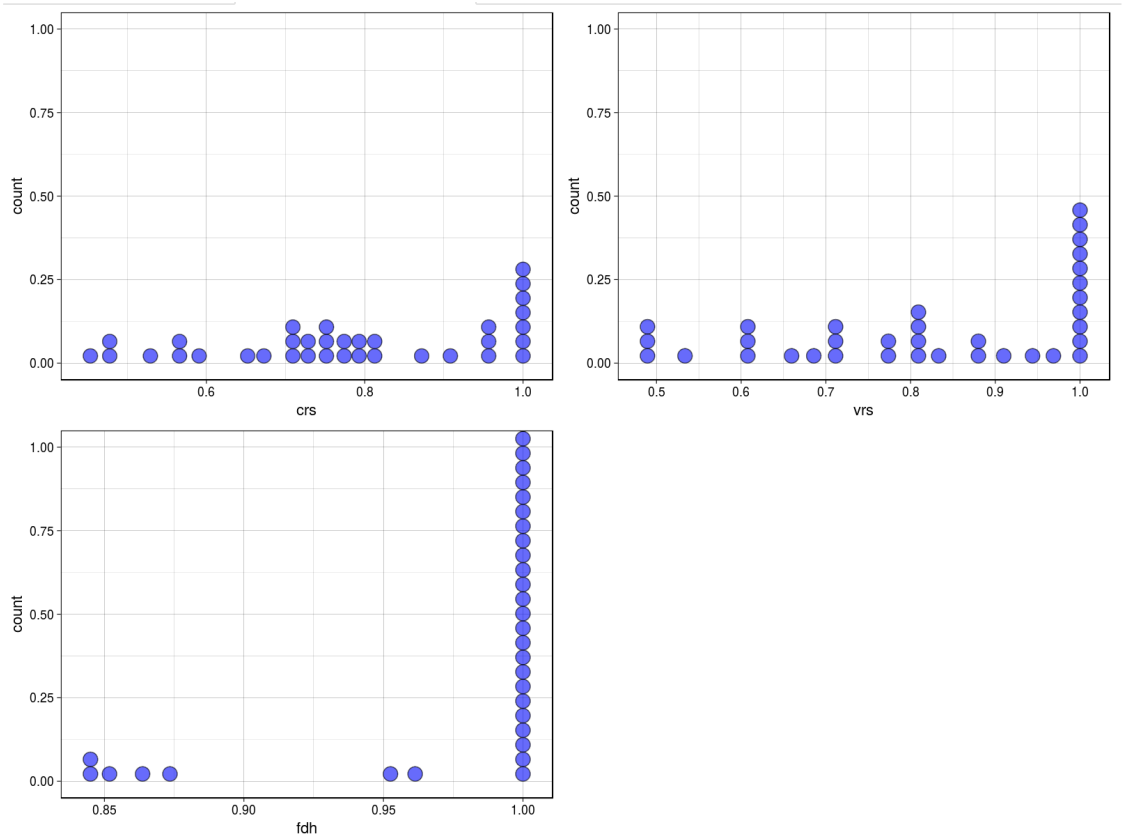

Figure 4.4: Distributions of Efficiency Scores

3.2.3 Correlations

The second preliminary set of plots is the correlation plots. The correlations between each pair of inputs and/or outputs, as well as efficiency scores of a specific model such as CRS. There are different approaches to visualization of correlations, but for low number of inputs and outputs, the comprehensive plot of "Performance Analytics" package (Peterson and Carl 2014). The lower triangle of the matrix of Figure 4.5 includes the scatter plots between variables including CRS scores. The distributions of the variables are presented in the matrix diagonal, and the upper triangle of the matrix is composed of correlation coefficients. Statistically significant correlation coefficients are marked by red asterisks.

3.2.4 Cross-Efficiency Unfolding

The first high-dimensional DEA visualization method that we use to visualize the Chinese cities dataset is Cross-efficiency unfolding. Suggested by (Ashkiani and Molinero 2017), this method visualizes the cross-efficiency matrix (CEM) of the dataset. The cross-efficiency evaluation is a method to allow DMUs to evaluate each other, and its outcome is a matrix n-by-n for a dataset of n DMUs. Any CEM has two sets of objects that appear in the row and column of the matrix. The row objects are the DMUs in their rating role, and the column objects are the same DMUs in

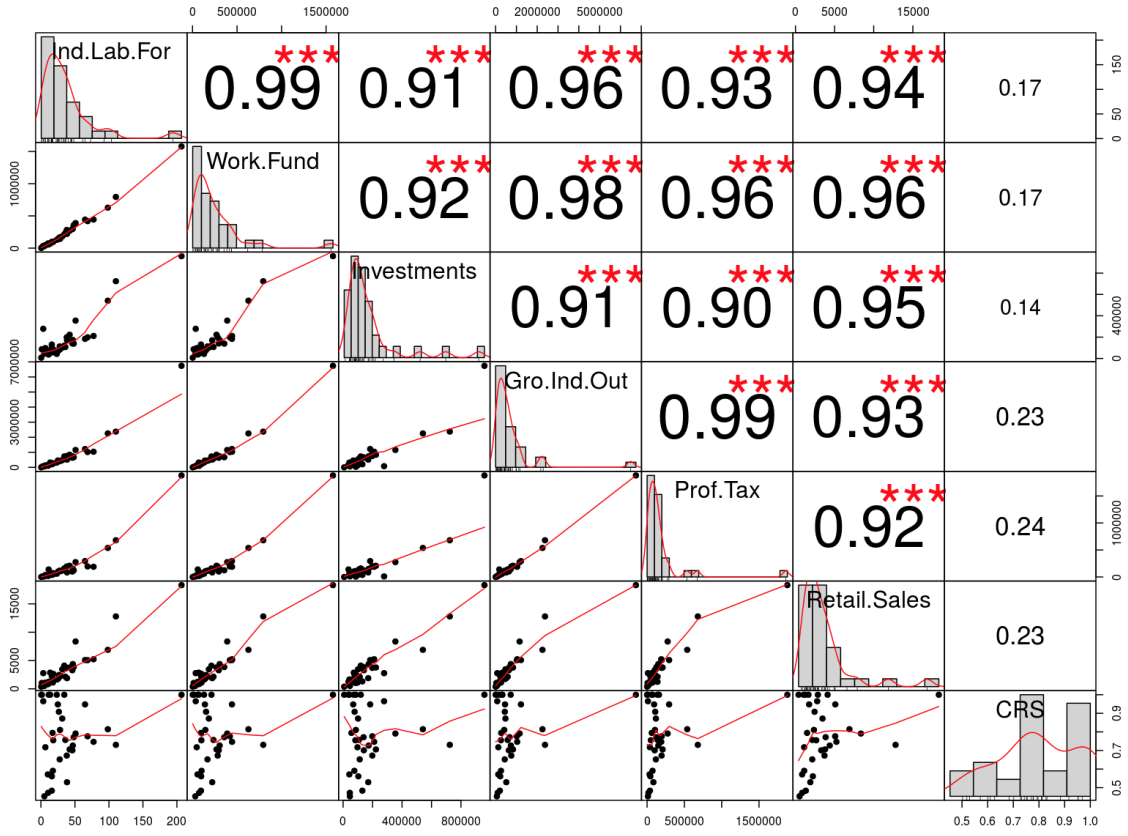

Figure 4.5: Correlation Plots

their rated role. In order to keep this paper as concise as possible, interested readers are referred to (Ashkiani and Molinero 2017), or Ruiz and Sirvent (2016).

Figure 4.6 is the unfolding map of CEM of 35 Chinese cities. The rating and rated objects are shown in blue dots and red triangles respectively. Ideally, the distance of every pair of same-class object is reflection of the dissimilarity of the pair. Thus, two close row objects probably have very low dissimilarity, i.e. high similarity. The distance between row object and a column object is ideally determined by its corresponding cross-efficiency score, and it is a measure of "preference". Therefore, if a column object is close to some row objects, that column object is preferred by the row objects, i.e. the row objects give high cross-efficiency score to that column object.

It is possible to explore Figure 4.6 from different aspects such as the clusters of row objects, and clusters of column objects. However, here we focus on units which are far from the crowd in both row and column objects. Unit 24 has such characteristic. Its row object is isolated, which means that its optimum weights profile is uncommon among row objects. Moreover, its column object is isolated, which means that it is not preferred by majority of the row objects, and it has low efficiency scores on average. In the third essay of this thesis, such units are explained as possible mavericks.

From now on, we have a special unit that we can investigate further on its details. For instance, it is claimed in the previous paragraph that DMU24 has

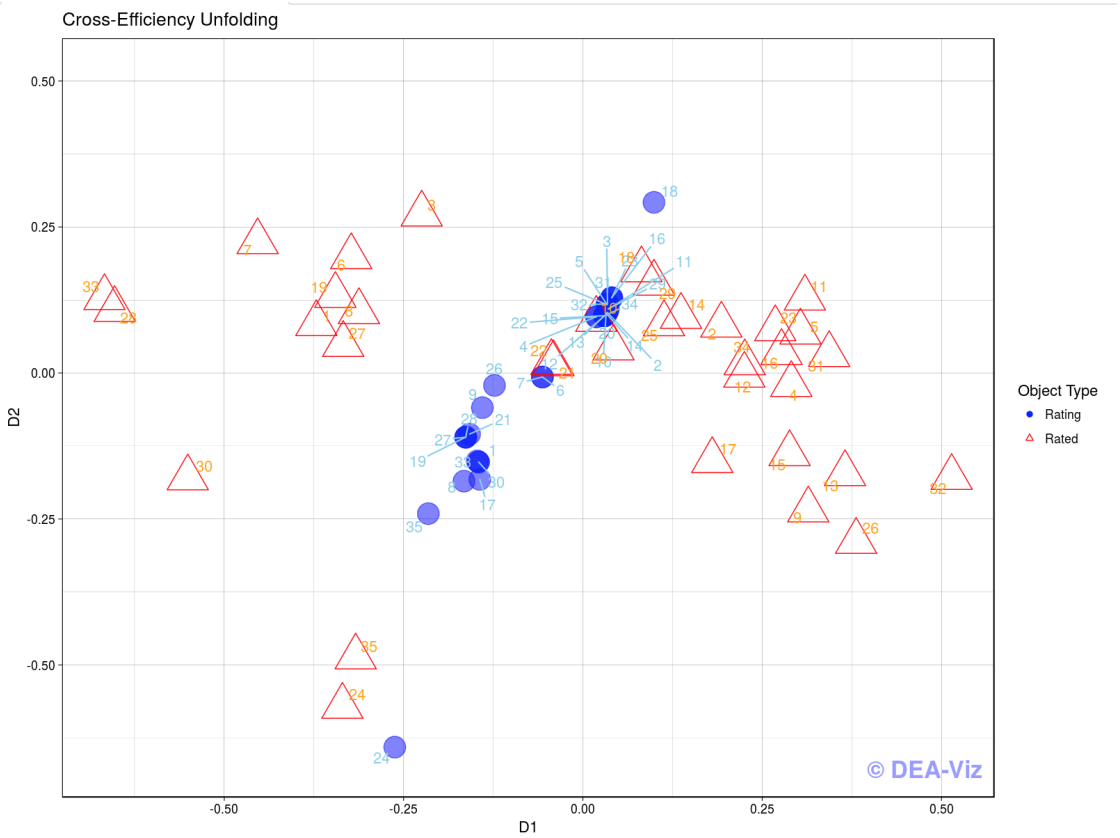

Figure 4.6: Unfolding Map of Benevolent CEM of the Dataset

uncommon optimum weight set. Figure 4.7 includes the distribution of standardized optimum weights of the units, chosen under benevolent cross-evaluation. For the standardization, the method suggested by Costa et al. (2016) is used. Briefly, the method standardized the input weights by the summation of the input weights, so the range of each standardized weight is between 0 and 1, and the summation of the standardized input weights is 1. The same is true for the output weights. This standardization makes removes the scale effect from the DMUs, and makes comparision of weights among the DMUs possible. Moreover, it is now clear that how each DMU allocates its weights to the inputs, and to the outputs. In other words, how each DMU emphasises on each input and output is clarified by these standardized weights. Figure 4.7 and Figure 4.8 enables us to explore the standardized input and output weights of DMU24.

As it is seen in the Figure 4.7, DMU24 has put all its input weights on the input1 and disregarded the input2 and input3. Its emphasize on input1 is totally uncommon according to the distribution of input1 weights. From Figure 4.8, it can be seen that the DMU24 is totally disregarded output2, while equally emphasizing on output1 and output3. In both latter cases, DMU24 shows uncommon behaviour.

Although the Figure 4.7 and Figure 4.8 reveal further details about the behaviour of DMU24, the standardized weight data is a high-dimensional data, and should be visualized using a high-dimensional data visualization technique. Figure 4.9 is principal component analysis(PCA) bi-plot of the weights dataset. The bi-plot is

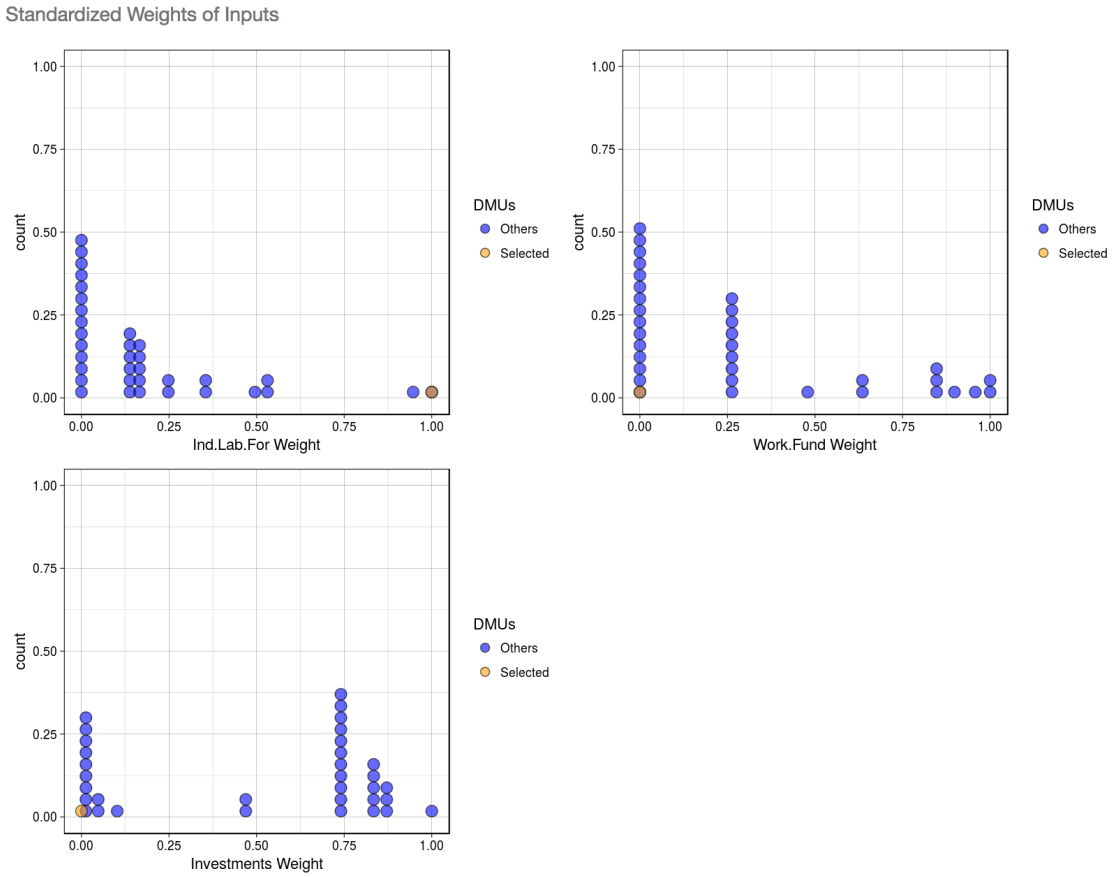

Figure 4.7: Standardized Input Weights - DMU24 in Yellow

essentially a PCA plot, so the more similar the DMUs, the closer they are on the map, and super-imposed vectors of variables. Here, the variables are standardized weights. The vectors points towards the direction "Awhich is most like the variable represented by the vector", i.e. "which has the highest squared multiple correlation with the principal components.", and the size of the vector is "proportional to the squared multiple correlation between fitted values for the variable, and the variable itself." (F. 1999).

As we can see in the Figure 4.9, DMUs 24, 9 and 35 are shaping a cluster which is far from other DMUs. The prominent characteristic of this cluster is high value of "Ind.Lab.For Weight", i.e. standardized weight of input1.

Using Figures 4.7 to 4.9, one can investigate further the weights of any specific DMU, such as DMU24. Doing so yields deeper insight into the CEM. Besides, Figures 4.7 to 4.9 can be used for validation of educated guesses based on unfolding map of the Figure 4.6.

3.2.5 Porembski Graph

Porembski et al. (2005) suggest the first high-dimensional data visualization method for DEA. Their method is based on Sammon's mapping (Sammon 1969), which is a non-linear dimension reduction technique from the class of multidimensional

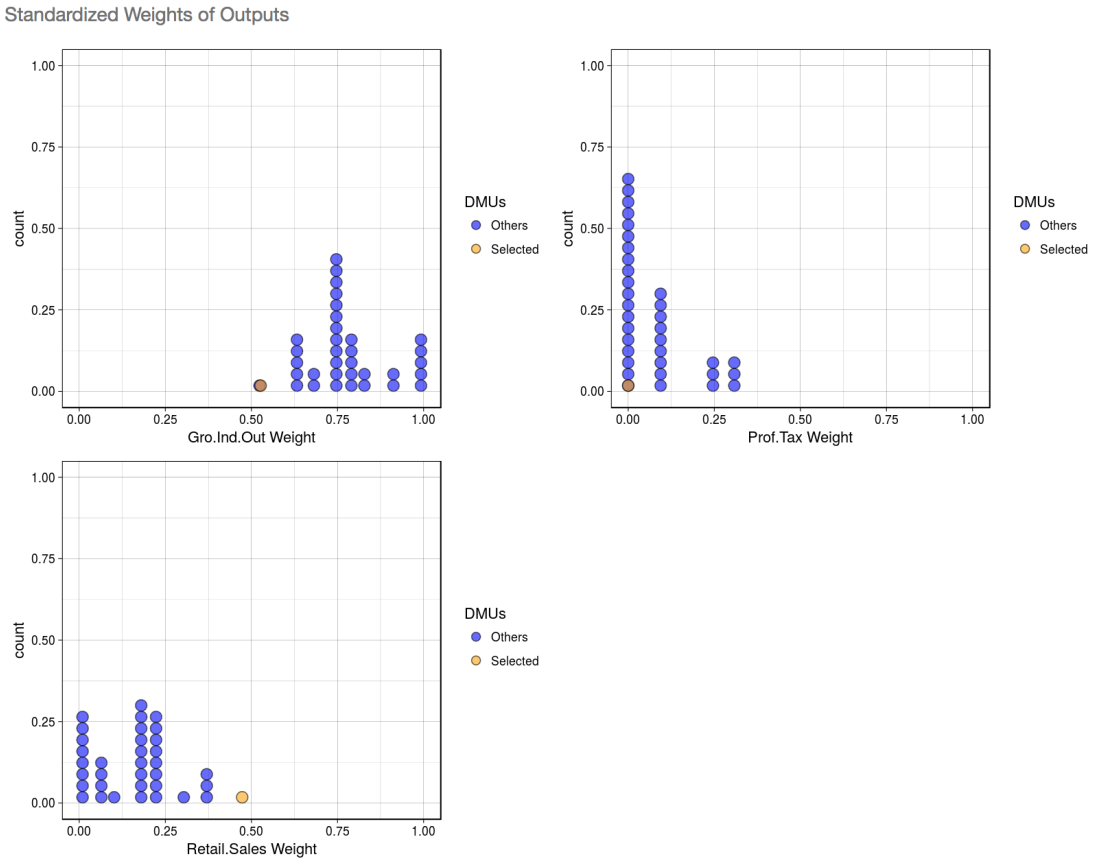

Figure 4.8: Standardized Output Weights - DMU24 in Yellow

scaling (MDS). Using this dimension reduction technique, the dimensionality of the dataset of inputs and outputs is reduced to two dimensions, i.e. coordinates of DMUs in the final map. The main difference between PCA and a MDS technique such as Sammon's mapping lies in the fact that MDS tries to retain the inter-point distances of the original space, in the final map. Such distance is usually measured by Euclidean distance function, and considered as proximity measure. Porembski et al. (2005) suggest to augment the bi-dimensional map using multiplier weights, i.e. lambda values. The lambda values are added to the map as edges or links between efficient units and inefficient units, such that the intensity of the edges is proportional to the magnitude of the lambdas. Doing so enables the practitioners to detect the efficient units that are in many reference sets, i.e. receive links from many inefficient DMUs. Moreonver, the efficient units which are not in any reference set can be detected easily. Figure 4.10 is the visualization of the Chinese cities using this method. Figure 4.11 is identical to Figure 4.10, but the DMU labels are added to the plot.

As we can see, DMU22 is far from the crowd. Being so suggests that DMU22 is drastically different from others considering the input and output levels.

In order to ease investigation of overlapping units, DEA-Viz has zoom-in feature. Figure 4.12 is the zoomed-in view of the down-left crowded part of Figure 4.11.

As it is seen in Figure 4.12, DMU24 is relatively far from the crowd, and thus it

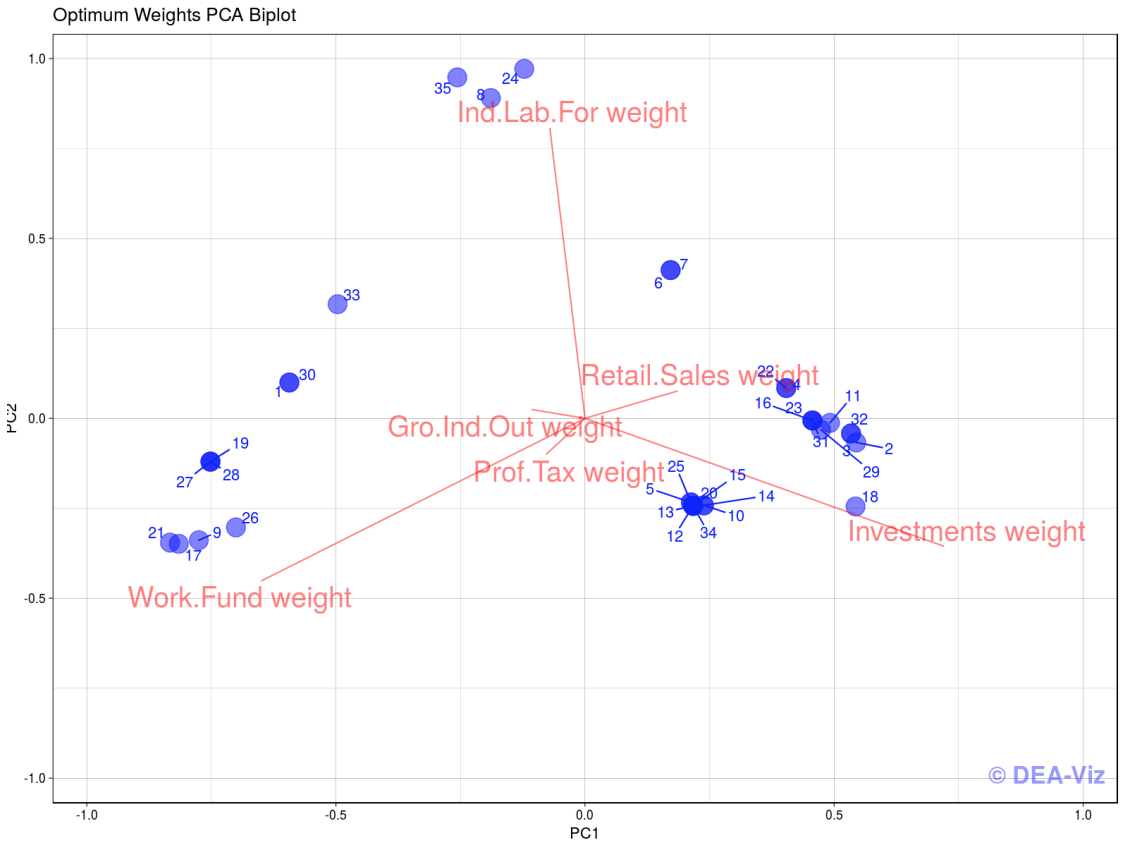

Figure 4.9: Standardized Weights Biplot

has relatively different input and output levels. It is interesting that DMU35 and DMU17 are the target unit of DMU24. Previously in Figure 4.9, we had seen that the DMUs 24,35 and 8 emerge a cluster based on their cross-efficiency optimum weights.

Here, we can come back to the distributions of inputs and outputs in order to know what makes DMU24 relatively different from the rest of the units. Figure 4.13, is the distributions of the inputs and outputs with DMU24 in yellow color.

As it is seen in Figure 4.13, the DMU24 has nothing outstanding from the mainstream DMUs, except for the "investment" variable, i.e. input3.

Another interesting point of Figure 4.12 is the efficient units which receive many links from inefficient units, i.e. efficient units such as DMU17 and DMU10. These efficient DMUs are in the reference sets of several inefficient units, and thus are influential in our CRS model.

3.2.6 PCA Biplot

Previously in Figure 4.9, the biplots are introduced. Briefly, high-dimensional data is reduced using PCA, and then vectors of original variables are superimposed on the PCA map. Using the direction of the vectors, one can investigate the differences between objects or clusters of the objects on the map.

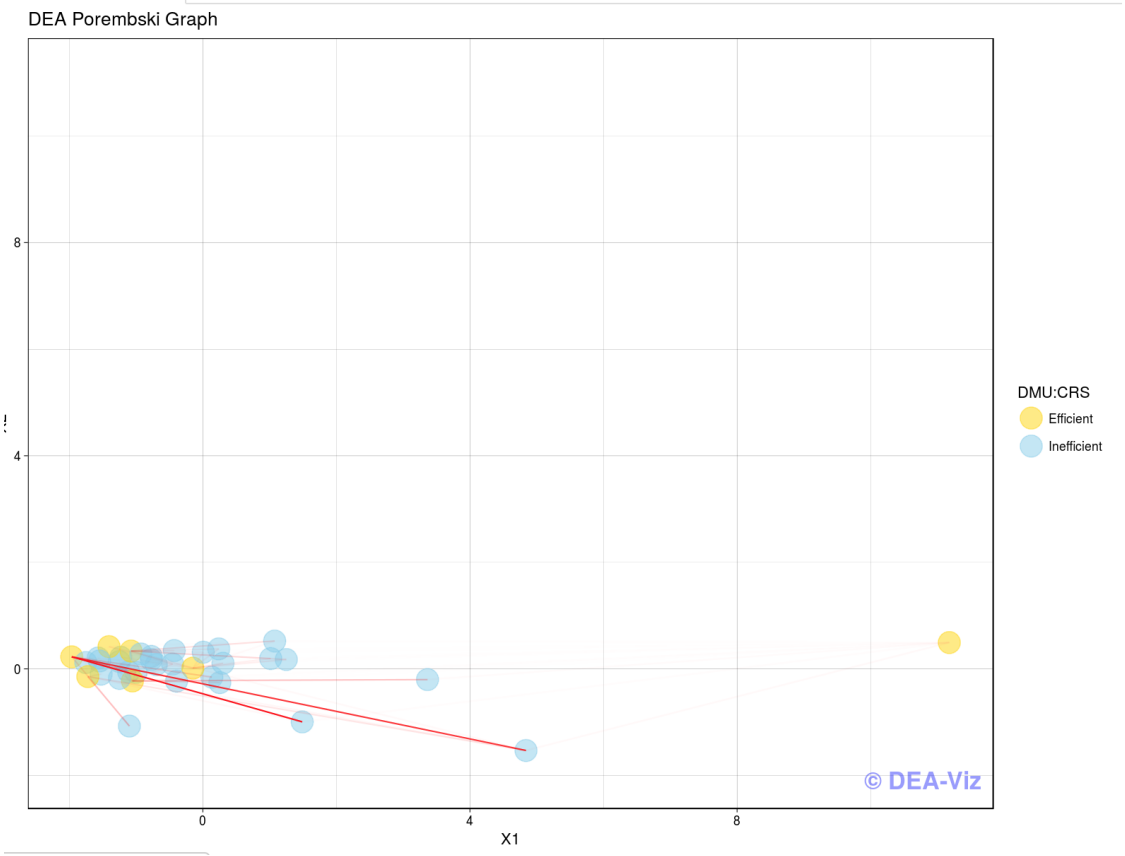

Figure 4.10: Porembski Graph

Here, the data which is fed into bi-plot method is the standardized inputs and outputs. The standardization is done in order to remove the effect of different scales of the variables. Figure 4.14 shows the bi-plot of the 35 Chinese cities.

Figure 4.14 shows the considerable difference of DMU22 from the other DMUs, based on their locations on the map. It seems that this DMU has very high values of the variables, since the projection of its position on the variable vectors is much higher than other units.

In order to focus on the crowd of units in Figure 4.14, we can using zooming feature of DEA-Viz. Figure 4.15 includes all units of Figure 4.14 except DMU22. Here in Figure 4.15 it can be seen that DMUs 1 and 27, while being relatively different from the others and having relatively higher scales, are not perfectly efficient. Moreover, the location of DMU24 shows that this unit is high on "Retail Sales" , and "Investment" while relatively low on the other variables. This finding is compatible with the findings of the Figure 4.13.

3.2.7 Multidimesional Scaling Color-Plots

Multidimensional Scaling (MDS) is a class of non-linear dimension-reduction techniques. Comparing to PCA, it has fewer assumptions about data. For instance, PCA assumes that the topology of the data can be reduced in a linear fashion, or

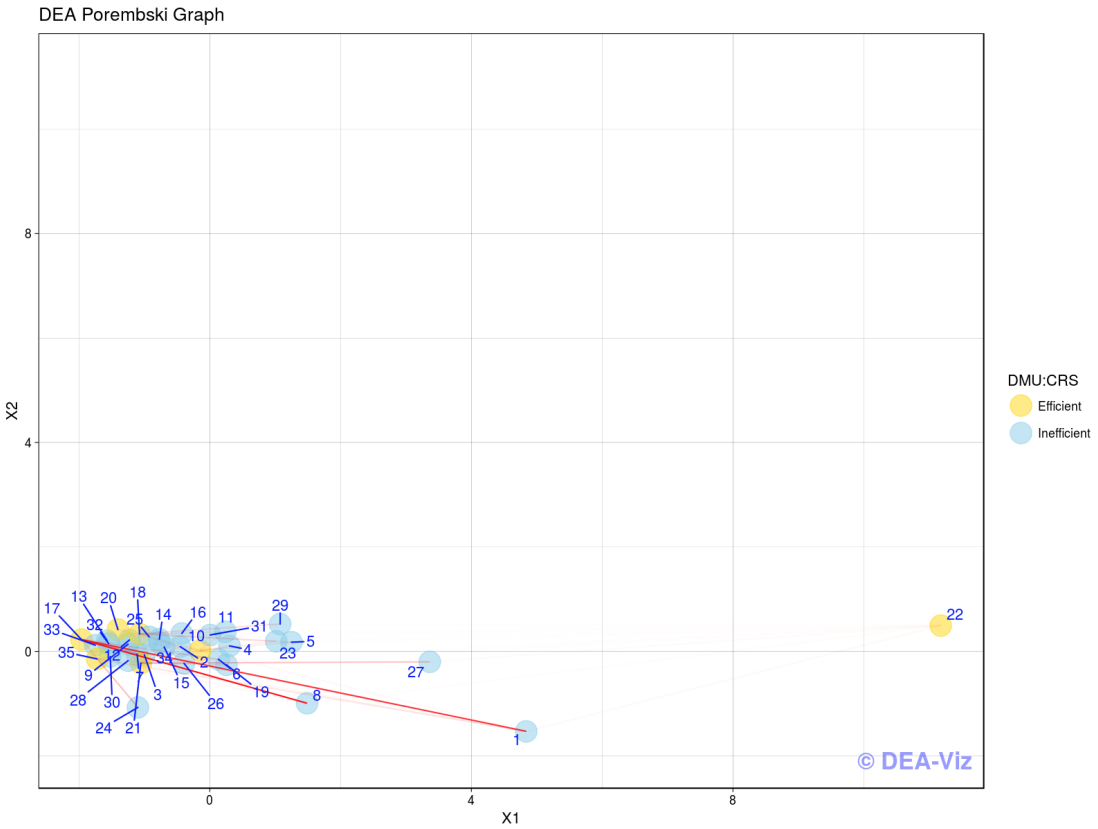

Figure 4.11: Porembski Graph with labels

the higher variance of a variable, the higher load of information it bears, and the higher importance. However, these assumptions in practice may be problematic. MDS tries to retain the original inter-object distances and represent them in the low dimensional space. For detailed information about MDS, excellent sources such as Borg and Groenen (2005) and Borg, Groenen, and Mair (2012) are available.

As explained before, Porembski et al. (2005) suggest using MDS to visualize DEA problems. Later on, Adler and Raveh (2008) also use a MDS technique to visualize DEA data. However, Adler and Raveh (2008) visualize ratios instead of original inputs and outputs. Ratios are calculated by division of every output to every input, and according to the authors the idea of using ratios is borrowed from Zhu (1998) and Sinuany-Stern and Friedman (1998). The method of Adler and Raveh (2008) is called co-plot, and it has the variable vectors super-imposed on the final map, similar to Bi-plots.

The MDS section of DEA-Viz is inspired by the Co-plot, except for the vectors. Since the MDS is a class of non-linear data transformation, the vectors, which are essentially linear, cannot show the non-linearity of the variables on the objects. Hence, continuous spectrum of colors, from blue to red, is used in DEA-Viz to show the values of a selected variable on the MDS map. It is possible to use original variables, or ratios as the input of the MDS in DEA-Viz, and using metric or non-metric MDS with Euclidean or Manhattan distance functions.

Figure 4.16 shows MDS map of ratio variables, and the CRS efficiency scores of

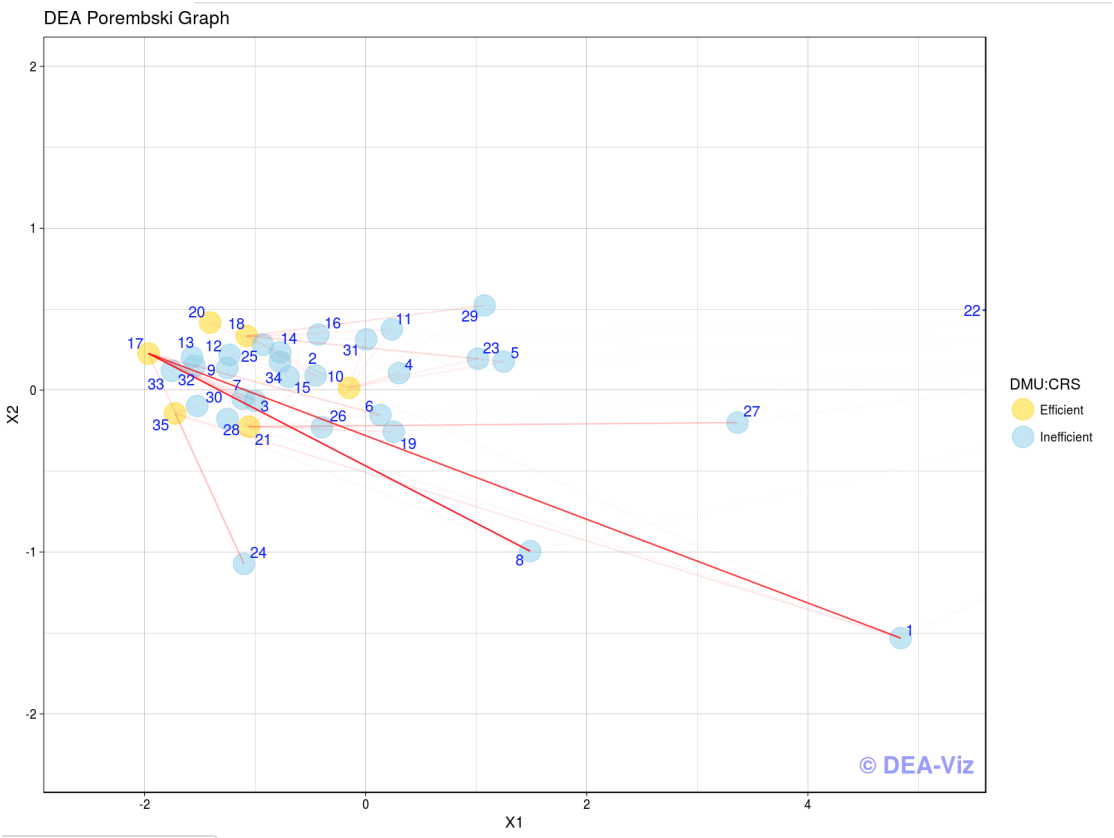

Figure 4.12: Zoomed view of the overlapping units of Porembski Graph

the DMUs are mapped on them as their colors.

It is seen that DMUs are plotted from left to right, and the most distance, i.e. dissimilarity, is between DMU17 and DMU22. It seems that the DMUs in the center of the map have lower CRS efficiency score in general. DMUs 24 and 35 are located relatively far from the crowd. The position of DMU20 is interesting, and this unit may require further investigation in order to figure out how it is different from the others.

One interesting point is revealed by comparing the position of DMUs in the bi-plot of Figure 4.14 and Figure 4.16. The bi-plot of Figure 4.14 is based on original variables, while Figure 4.16 is based on ratios of original variables. Using ratios removes the influence of scale of the units. Thus, DMUs 1 and 27, which are distinct from the rest in Figure 4.15, are not isolated in the Figure 4.16. Being so means that these units have higher scale than others, but from the distribution of ratios they are not special. In contrast, DMU20 separates itself from the rest when ratios are used, as seen in Figure 4.16. DMU20 possibly does not have specially high or low scale, but it has special distribution of ratios. On the other hand, DMU24 is relatively isolated in both Figure 4.15 and Figure 4.16, and its location suggests that this DMU is relatively special regardless of using ratio variables, and the original variables.

In bi-plots, practitioner could use vectors in order to figure out the difference of two units or two clusters of the units. Here such question can be investigated

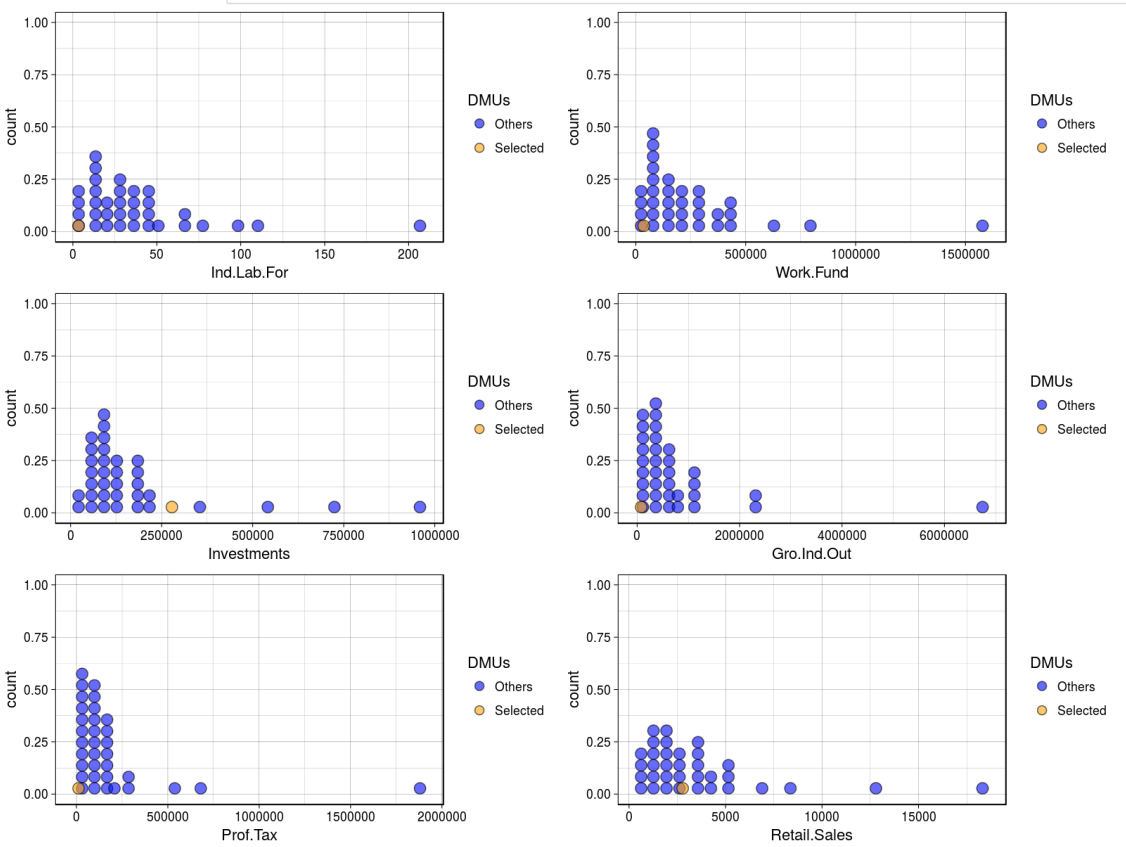

Figure 4.13: Distributions of Inputs and Outputs, DMU24 in yellow

using different maps. For instance, Figure 4.17 shows the distribution of "retail sales(O3)/investment(I3)" over the units, and Figure 4.18 depicts the distribution of "Profit and Tax(O2)/Industrial Labour Force(I1)".

Considering Figure 4.17 and Figure 4.18, one can see that the while the DMU17 is high in the ratio of "retail sales/investment", it is low in the ratio of "Profit and Tax/Industrial Labour Force". This situation is approximately opposite for the DMU22 on the other side of the map. Meanwhile, it seems that DMU20 is not very low in any of these ratios, and indeed it has the highest value of the "Profit and Tax/Industrial Labour Force". Using these maps, one can get insight into the different ways that the units such as 20, 22,and 17 achieve their perfect efficiency scores.

3.2.8 Self-Organizing Map

Self-organizing map is an application of artificial neural network, which tries to preserve the topology of the dataset in a space projection method to

Self-Organizing Map (SOM), suggested by Kohonen (1982b) and Kohonen (1982a) is an unsupervised learning method in order to represent a high-dimensional dataset into a two dimensional grid of units, such that the topology of the dataset in the original space is preserved. The main goal of the SOM is division of the original data space into a pre-defined number of disjoint sub-spaces, and therefore segmentation of

Figure 4.14: Distributions of Inputs and Outputs, DMU24 in yellow

the space rather than the observations. In the final SOM settings, similar observations would reside close to each other, either in one unit or a neighbourhood of units, and dissimilar observations would locate in distant units. For detailed information about SOM, redears are referred to Kohonen (1998)

DEA-Viz has a section for visualization of inputs and outputs variables of DMUs, using SOM. The visualization of DEA using SOM is suggested by Churilov and Flitman (2006), and later Carboni and Russu (2015)

Figure 4.19 is visualization of 35 Chinese cities inputs and outputs data using 7-by-7 SOM.

The nodes of SOM in Figure 4.19 is colored by CRS efficiency scores of the units. The gray nodes are empty, while some nodes have more than one DMU. The neighbour nodes are ideally similar, so for instance DMU18 and DMU25 are similar according to Figure 4.19. Besides, DMUs such as 24 and 22 have no DMU in their adjacent nodes, so we can conclude that there are no closely similar DMUs to them in the dataset. It is noteworthy that the similarity of the nodes are measured using Euclidean distance function.

Figure 4.20 shows the distribution of the original variables on the SOM map of Figure 4.19, so we can figure out how two nodes are different, and investigate the correlation of variables.

While both methods are non-linear projection of high-dimensional data into low

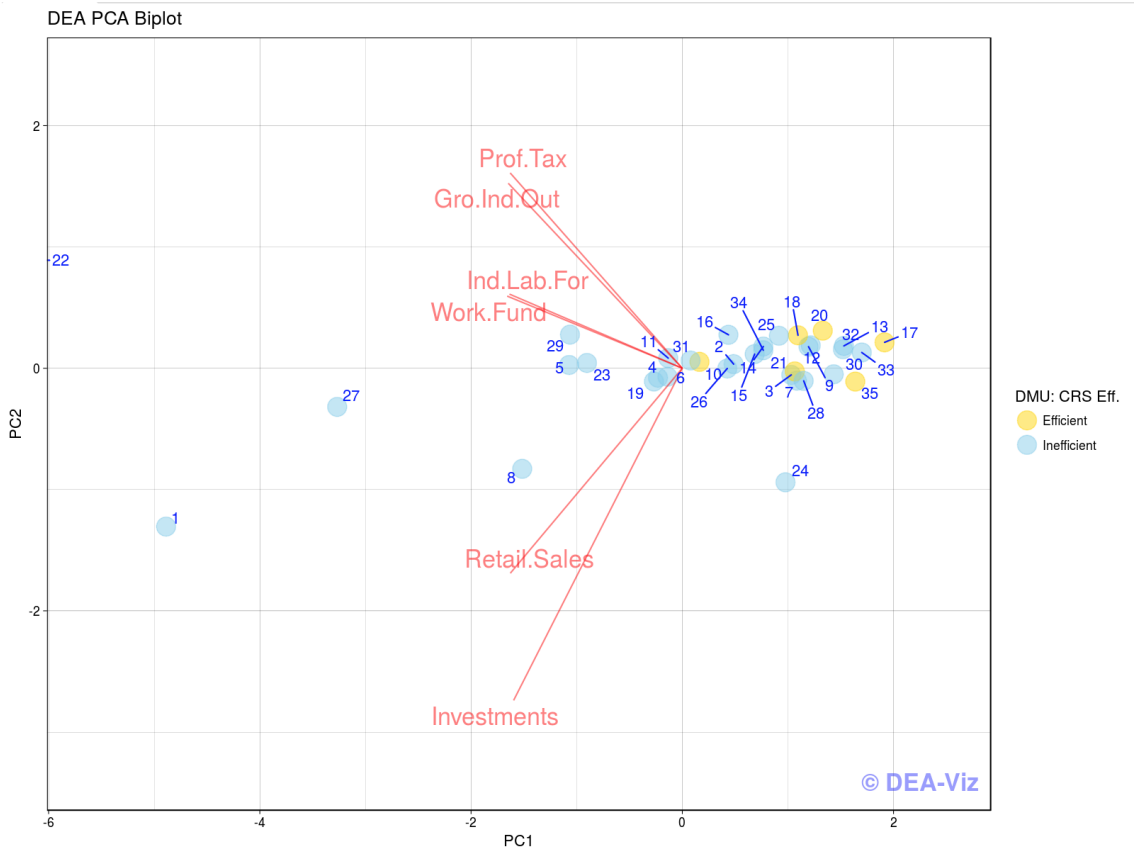

Figure 4.15: Distributions of Inputs and Outputs, DMU24 in yellow

dimensional space, SOM is computationally much more efficient than MDS, and it has one level of abstraction by-default which makes datasets with high number of DMUs more tractable.

3.2.9 Frontier Visualization

Perhaps the most frequent plot in DEA textbooks and papers is graphical depiction of the efficient frontier in a bi-dimensional space, where the dimensions are two variables from the set of inputs and outputs. However, as it is explained in section2, trying to understand a high-dimensional dataset by selection of a subset of variables and ignoring the rest is a reductionist approach which cannot yield the true understanding of the dataset, and even doing so may mislead the practitioners by the distorted perception of the dataset. This does not mean that the frequent plot of efficient frontier is not useful, it is indeed useful for educational purposes, however for real problems a more holistic approach is needed.

Costa et al. (2016) suggest a method for graphical depiction of the efficient frontier and the DMUs. The method is based on calculation of summation standardized virtual inputs, and summation of standardized virtual outputs. These two values compose coordinates of each unit, and thus enable us to present the DMUs and efficient frontier in a bi-dimensional map. For the details about the steps of standardization please refer to the original paper of the authors.

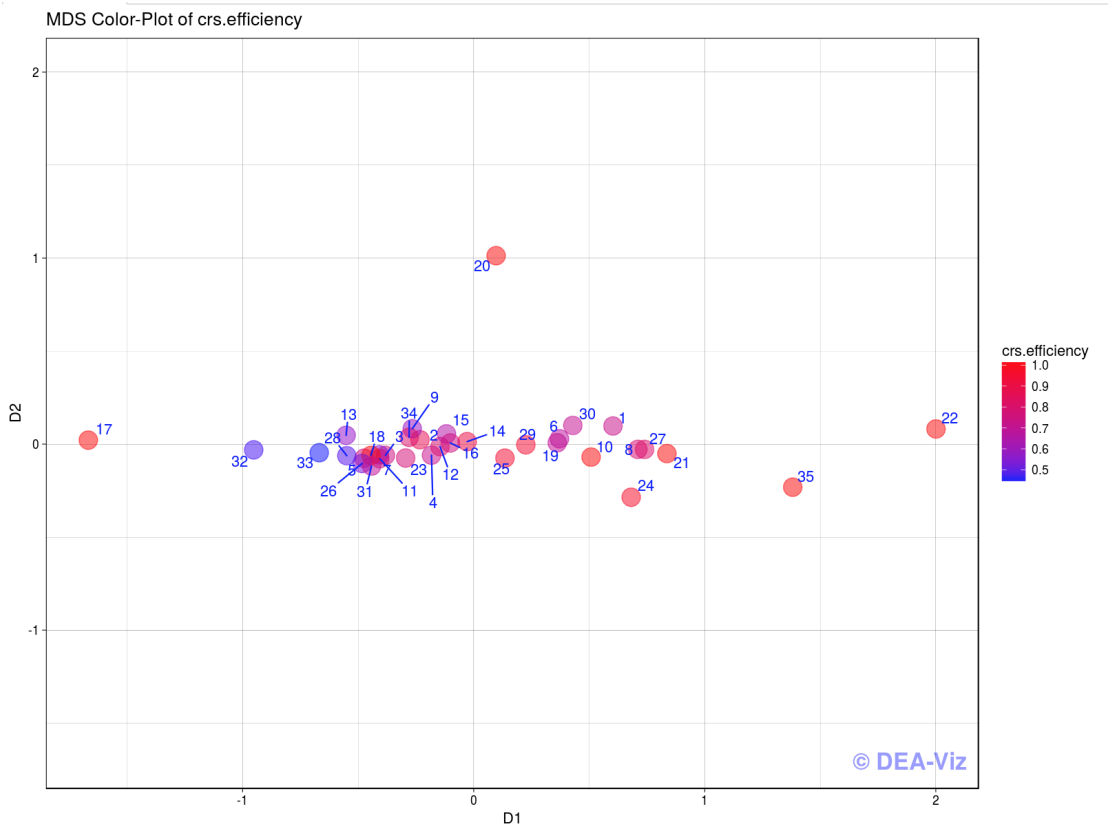

Figure 4.16: MDS map of ratio variables

Figure 4.21 is the frontier visualization of CRS input-oriented model of the 35 Chinese cities. Figure 4.22 is identical to Figure 4.21, but with added DMU labels.

The coordinates of the Figures 4.21 and 4.22 are the summation of standardized virtual inputs(horizontal axis) and summation of standardized virtual outputs(vertical axis). The efficient frontier is shown by the red line, where each point on it has equal summation of standardized virtual inputs and summation of standardized virtual outputs. Accordingly, the perfectly efficient units lie on the red line, while the inefficient units would locate on the right hand side of the red line of efficient frontier.

The distance of the inefficient units from the red line of efficient frontier is proportional to their degree of inefficiency. In other words, the more inefficient, the far from the red line. Since the magnitude of the summation of standardized inputs or outputs is not merely due to the scale of the unit, but it is due to the mixture of the magnitude of the inputs and outputs levels and the optimal weights, it is not possible to claim with certainty that units such as 5 and 26 have relatively higher scale. However, it is worth of it to investigate these units further. Figures 4.23 and 4.24 shows the distribution of inputs and outputs of the DMUs 5 and 26, respectively.

Figures 4.23 and 4.24 show that the input and output levels of the DMU5 and DMU26 are not exceptionally high. Consequently, the position of the inefficient units on the Figures 4.21 and 4.22 is not merely due to their inputs and outputs levels, but it is about how they use these inputs and outputs. The position of the efficient

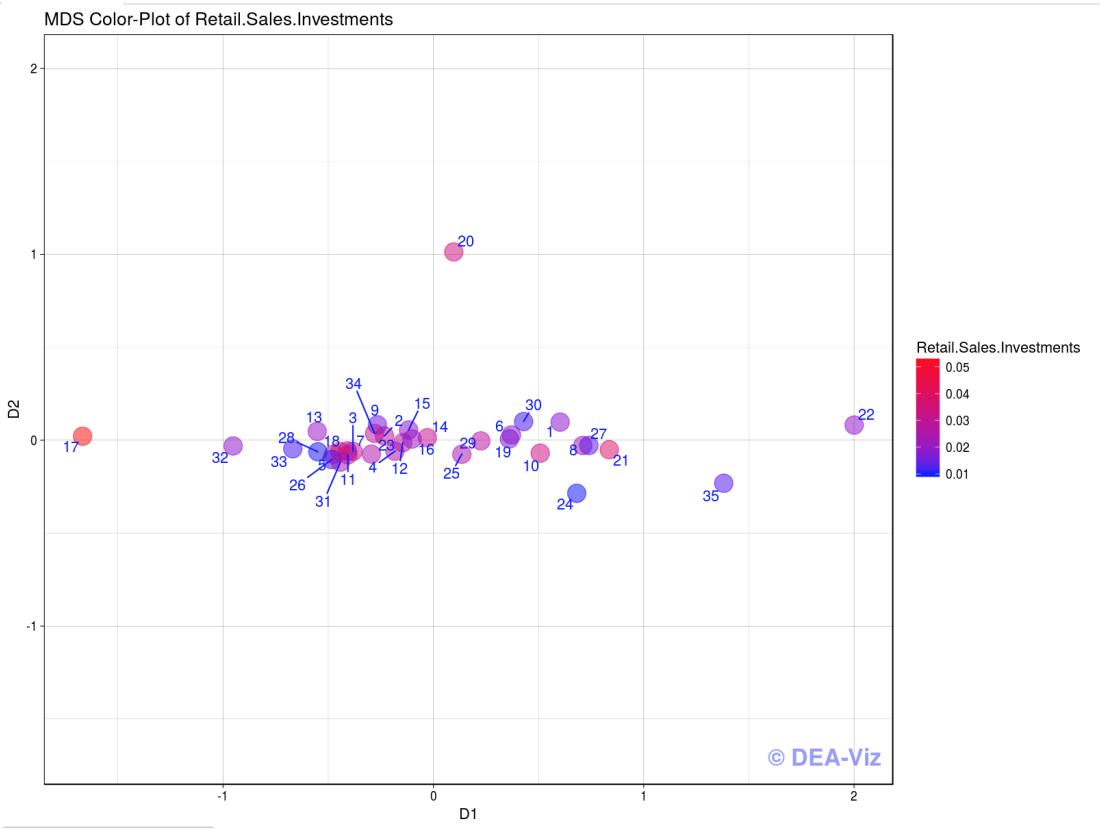

Figure 4.17: The distribution of retail sales(O3)/investment(I3) on MDS map

units is not fixed due to multiplicity of weights, and by adapting alternate optimum weights the positions would change on the red line.

In summary, the interpretation of the positions of the DMUs on the frontier visualization of Costa et al. (2016) is not straightforward, and for evaluation of the scale of the DMUs, it is suggested to use PCA bi-plots and MDS of the inputs and outputs.

3.2.10 Findings

In previous sections, the dataset of 35 Chinese cities has been visualized using several methods available in the version 1.0 of the DEA-Viz. The findings of each visualization are mentioned in the corresponding section, but they are gathered in this section in order to sum up the understandings yield by visualization.

- Maverick Units: Using unfolding map of cross-efficiency matrix (CEM), Figure 4.6, DMU24 was nominated as the top maverick unit, i.e. a unit which works differently from the rest. In the third paper of this thesis, it is explained what mavericks are and how they can be detected using unfolding maps of CEM.

- Influential Efficient Units: The efficient units, such as DMU10 or 17, that are available in the reference sets of many inefficient units, are detected in

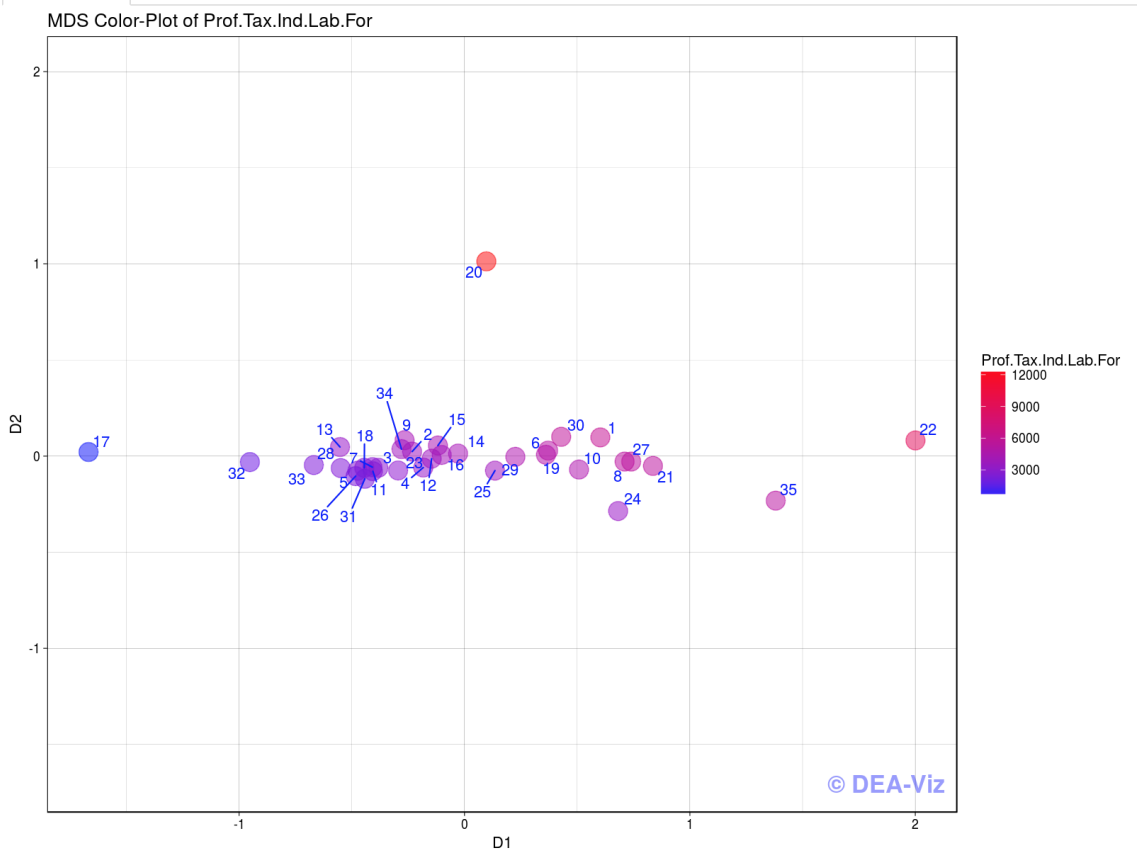

Figure 4.18: The distribution of retail Profit and Tax(O2)/Industrial Labour Force(I1) on MDS map

Figure 4.12. Their presence or absence can drastically influence on the problem.

- Units with high scale: The units that have high levels of inputs or outputs detected through Figure 4.14. These units, such as DMU22 or DMU1, stand far from the crowd of DMUs and it is possible to distinguish the extra ordinary high inputs or outputs of them using the vectors of the bi-plots.

- Units with ratio variables: If we remove the magnitude and scale from the input and output levels using variable ratios, then the units with different distributions of inputs and outputs, i.e. ratio values, would reside far from their neighbours. Figure 4.16 shows that DMUs such as 20,17 and 22 are such units.

Beside the above items, the plots can be used for detection of the clusters of DMUs, or the correlated input and output variables.

As it is stated before, exploratory data visualization usually leads to further and deeper questions both about the data and the methods. For instance, each of the units with interesting behaviour, such as maverick or influential efficient units, can be the subject of further investigation. Is the behaviour due to error in data or the reason is intrinsic? What should be done to the units in either case? How can

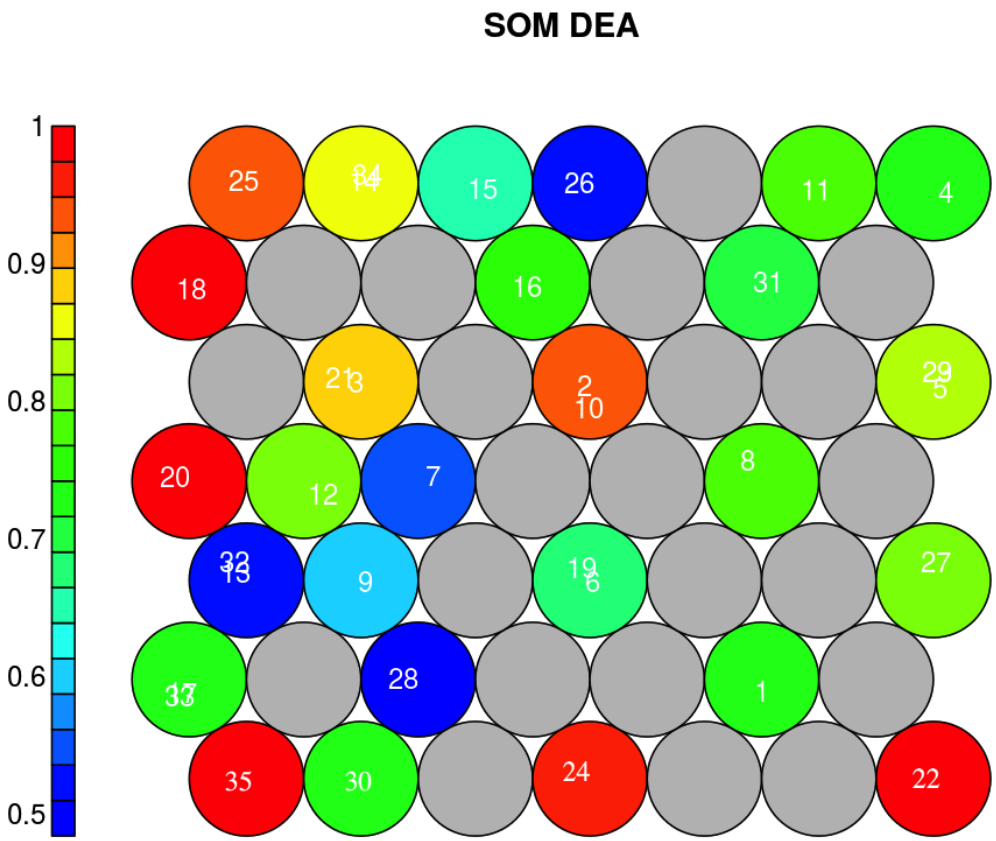

Figure 4.19: Self-Organizing Map of Chinese cities dataset

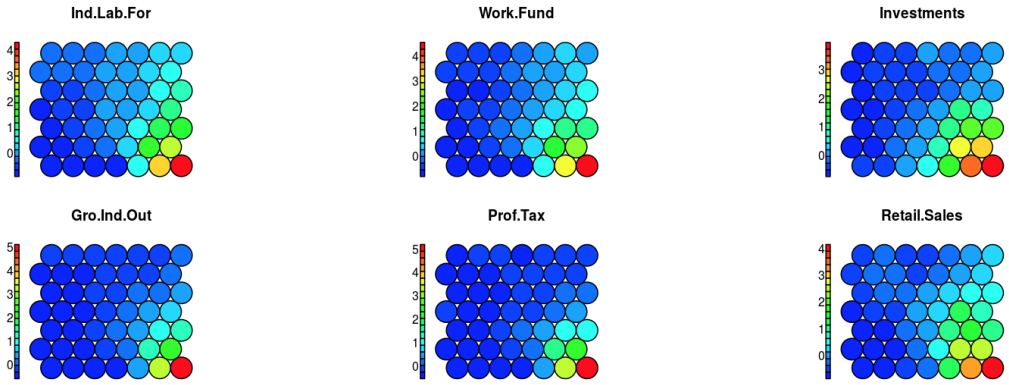

Figure 4.20: Variable distribution on Self-Organizing Map

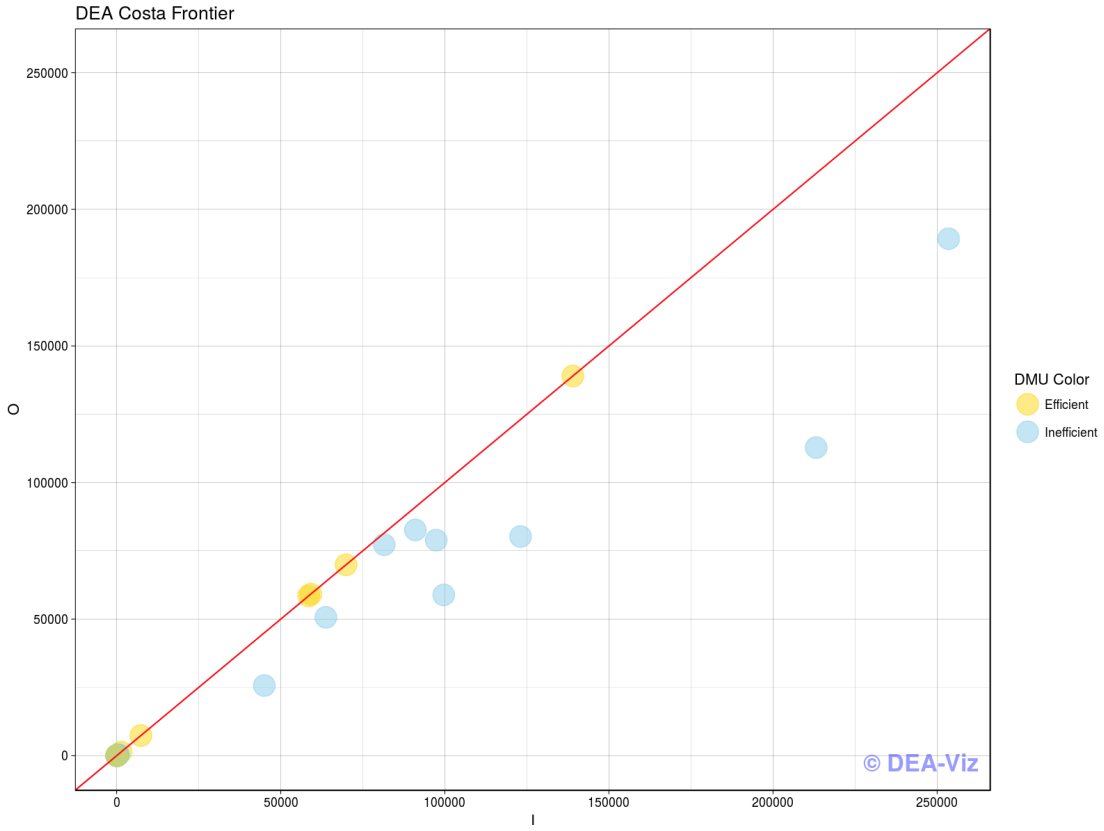

Figure 4.21: Frontier Visualization of 35 Chinese Cities

we deal with multiplicity of weights in visualization? How can we incorporate new variables such as slacks or Malmquist index in the visualizations?

Additionally, more fundamental questions can be asked after these visualizations. For instance, the most conventional proximity measure in methods such as SOM and MDS is Euclidean distance, however it is shown that Euclidean distance behaves strangely in high dimensional spaces, and generally it is not a good measure for such problems (Aggarwal et al. 2001). Therefore, shouldn't we use a better proximity measure, i.e. specifically devised for DEA problems? What about using dual multipliers or lambda values in such DEA proximity measure?

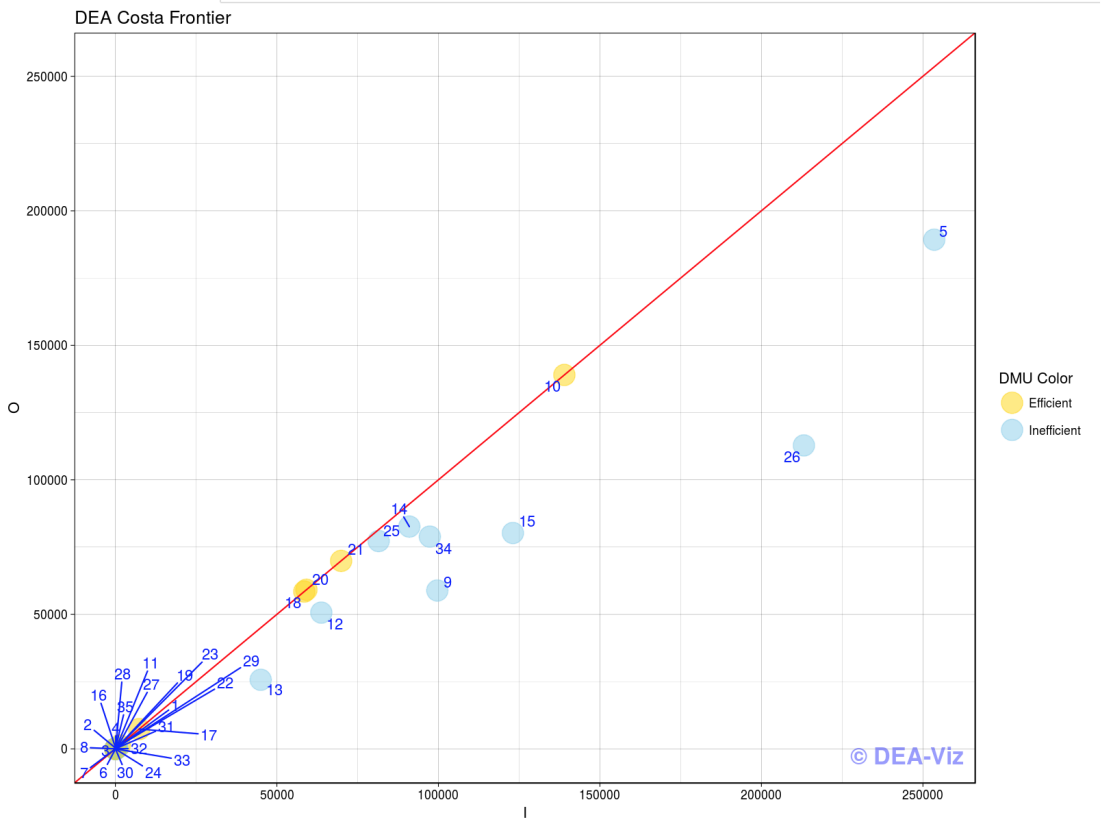

Figure 4.22: Frontier Visualization of 35 Chinese Cities with added labels

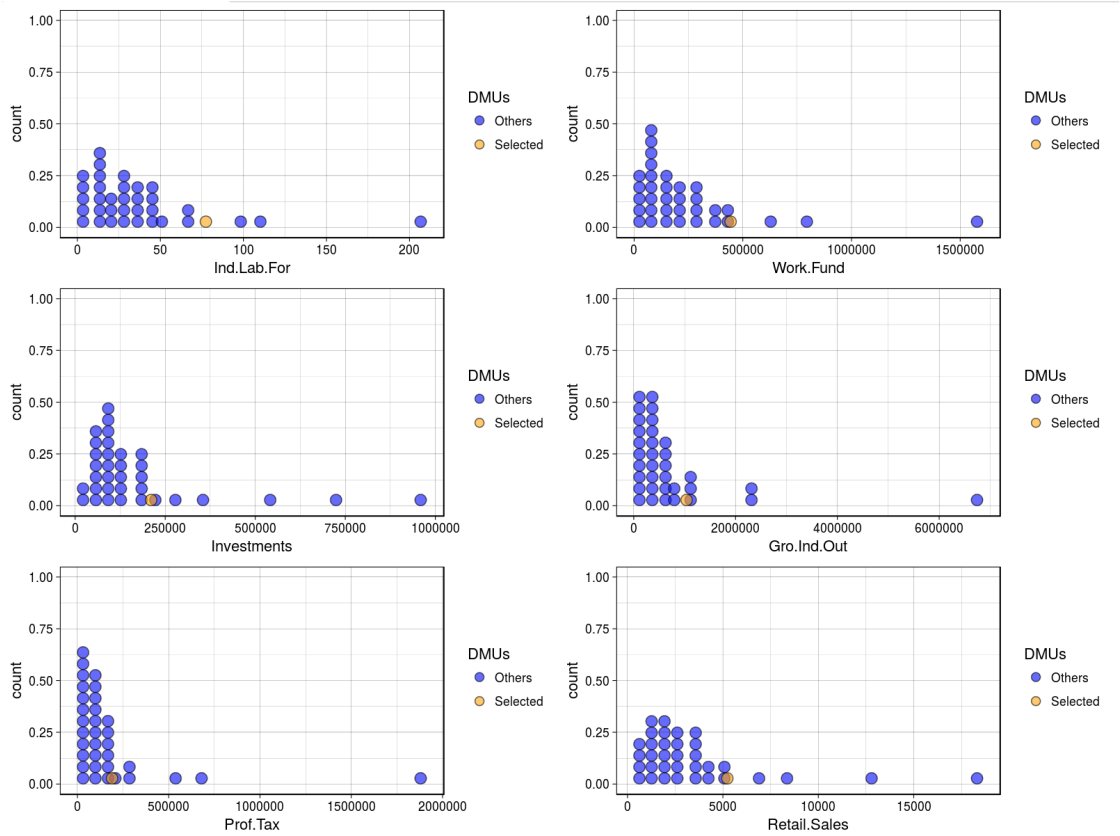

Figure 4.23: Inputs and Outputs Distributions of DMU5

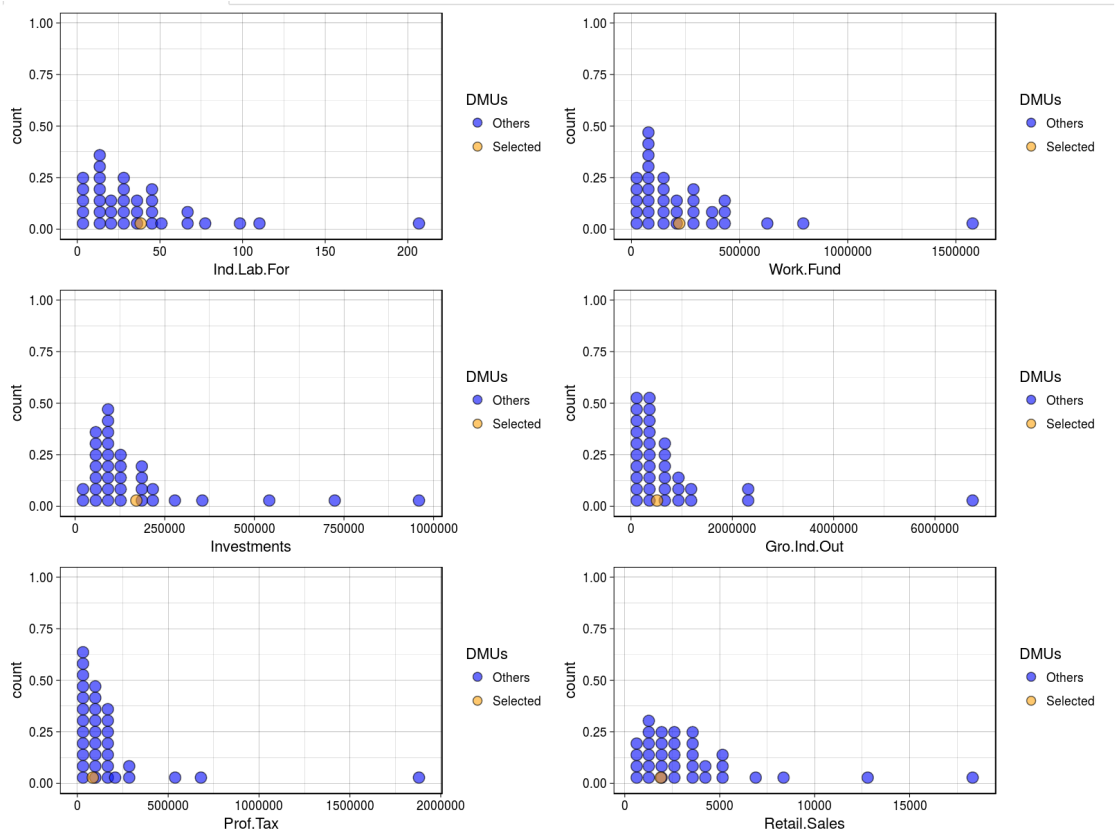

Figure 4.24: Inputs and Outputs Distributions of DMU26

4 Conclusion

It is known that human brain is wired for pattern recognition, and it is specifically powerful at doing so. The information visualization methods try to aid this ability by transformation of large amount of digits into visual elements (Soukup and Davidson 2002, p. 6).

Ward et al. (2015) state that visualization is a foundation of new knowledge discovery tools, through providing "alternate views of the data, and help to describe some structures, patterns and anomalies" in the data.

Nevertheless, data visualization is relatively neglected in Data envelopment analysis(DEA), even though DEA is a data-driven approach. The role of data visualization becomes more significant when the quantitative data is high-dimensional, since having more variables increases the complexity of the dataset by increasing the number of components and the relations. Similar to other complex systems, these datasets are not possible to be fully comprehended using reductionist approaches.

DEA data-sets oftentimes are high-dimensional, and while several high-dimensional data visualization methods are suggested in the literature, the visualization features of DEA software packages are limited to uni-variate or bi-variate data. Therefore, these features are not very helpful to DEA practitioners in order to get insight to DEA problems. Beside DEA software packages, DEA textbooks and handbooks are also vastly neglected DEA visualization.

The reasons behind such neglect are probably manifold, but disregarding visualization is not anything new in quantitative fields. Anscombe (1973) enumerates some of the beliefs behind inclination towards quantitative analysis and disregard of visualizations:

- Numerical calculations are exact, but graphs are rough
- For any particular kind of statistical data, there is just one set of calculations constituting a correct statistical analysis
- Performing intricate calculations is virtuous, whereas actually looking at the data is cheating

This paper is an introduction to DEA-Viz, a DEA software developed for visualization of DEA problems. DEA-Viz intends to fill the shortage of visualization features of the current DEA packages. Its specific focus is on high-dimensional visualization, and it benefits from the main suggested DEA visualization methods of the literature. Moreover, DEA-Viz is a cloud-based applet, available at `https://ashkiani.shinyapps.io/dea-viz/` , and can be run on any device with internet connection. Through approaching the DEA data from various aspects, and providing multiple plots, DEA-Viz can help to reveal some hidden characteristics of the data. These characteristics may make or mar the analysis, and are difficult to detect using quantitative methods.

In order to illustrate the capabilities of DEA-Viz, the dataset of 35 Chinese cities is visualized using various methods, and the findings of these visualizations gathered at the end of the corresponding section. Maverick units, influential efficient

units, uncommonly high-scale units, different units based on their ratio variables were found through exploration of the plots. Moreover, the general structure of the dataset, such as the clusters of units, could be examined. At the end, the exploration raised further questions about the data, and the methods, the question points which required further investigations.

Nevertheless, various aspects of DEA-Viz can be improved. For instance, new high-dimensional data visualization such as t-sne of Maaten and Hinton (2008) could be added to the features. Morevoer, new methods based on graph and network visualization, or non-Euclidean proximity measures can enrich the current toolbox. DEA-Viz also would benefit greatly from data linked-views (Wilhelm 2008; Wills 2008).

References

Adler, Nicole and Adi Raveh (2008). "Presenting DEA graphically". In: *Omega* 36.5, pp. 715–729.

Aggarwal, Charu C, Alexander Hinneburg, and Daniel A Keim (2001). "On the surprising behavior of distance metrics in high dimensional spaces". In: *ICDT*. Vol. 1. Springer, pp. 420–434.

Anscombe, Francis J (1973). "Graphs in statistical analysis". In: *The American Statistician* 27.1, pp. 17–21.

Visualization of Cross-Efficiency Matrices Using Multidimensional Unfolding (2017).

Borg, Ingwer and Patrick JF Groenen (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Borg, Ingwer, Patrick JF Groenen, and Patrick Mair (2012). *Applied multidimensional scaling*. Springer Science & Business Media, p. 3.

Carboni, Oliviero A and Paolo Russu (2015). "Assessing regional wellbeing in Italy: An application of Malmquist–DEA and self-organizing map neural clustering". In: *Social indicators research* 122.3, pp. 677–700.

Chen, Chun-houh, Wolfgang Karl Härdle, and Antony Unwin, eds. (2008). *Handbook of data visualization*. Springer.

Churilov, Leonid and A Flitman (2006). "Towards fair ranking of Olympics achievements: The case of Sydney 2000". In: *Computers & Operations Research* 33.7, pp. 2057–2082.

Cooper, William W, Nuria Ramón, José L Ruiz, and Inmaculada Sirvent (2011). "Avoiding large differences in weights in cross-efficiency evaluations: application to the ranking of basketball players". In:

Costa, Carlos A Bana e, João Carlos CB Soares de Mello, and Lidia Angulo Meza (2016). "A new approach to the bi-dimensional representation of the DEA efficient frontier with multiple inputs and outputs". In: *European Journal of Operational Research* 255.1, pp. 175–186.

F., Young (1999). *Principal Components:BiPlots*. URL: `http://forrest.psych.unc.edu/research/vista-frames/help/lecturenotes/lecture13/biplot.html` (visited on 11/12/2017).

Friendly, Michael (2008). "Handbook of data visualization". In: *Handbook of data visualization*. Ed. by Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. Springer. Chap. A brief history of data visualization, pp. 15–56.

Kohonen, Teuvo (1982a). "Analysis of a simple self-organizing process". In: *Biological cybernetics* 44.2, pp. 135–140.

Kohonen, Teuvo (1982b). "Self-organized formation of topologically correct feature maps". In: *Biological cybernetics* 43.1, pp. 59–69.

Kohonen, Teuvo (1998). "The self-organizing map". In: *Neurocomputing* 21.1, pp. 1–6.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov, pp. 2579–2605.

Peterson, Brian G. and Peter Carl (2014). *PerformanceAnalytics: Econometric tools for performance and risk analysis*. R package version 1.4.3541. URL: `https://CRAN.R-project.org/package=PerformanceAnalytics`.

Porembski, Marcus, Kristina Breitenstein, and Paul Alpar (2005). "Visualizing efficiency and reference relations in data envelopment analysis with an application to the branches of a German bank". In: *Journal of Productivity Analysis* 23.2, pp. 203–221.

Ruiz, José L and Inmaculada Sirvent (2016). "Ranking Decision Making Units: The Cross-Efficiency Evaluation". In: *Handbook of Operations Analytics Using Data Envelopment Analysis*. Springer, pp. 1–29.

Sammon, John W (1969). "A nonlinear mapping for data structure analysis". In: *IEEE Transactions on computers* 100.5, pp. 401–409.

Sinuany-Stern, Zilla and Lea Friedman (1998). "DEA and the discriminant analysis of ratios for ranking units". In: *European Journal of Operational Research* 111.3, pp. 470–478.

Soukup, Tom and Ian Davidson (2002). *Visual data mining: Techniques and tools for data visualization and mining*. John Wiley & Sons.

Sueyoshi, Toshiyuki (1992). "Measuring the industrial performance of Chinese cities by data envelopment analysis". In: *Socio-Economic Planning Sciences* 26.2, pp. 75–88.

Telea, Alexandru C (2014). *Data visualization: principles and practice*. CRC Press.

Tukey, John Wilder (1977). *Exploratory Data Analysis*. Addison-Wesley. ISBN: 0201076160.

Ward, Matthew O, Georges Grinstein, and Daniel Keim (2015). *Interactive data visualization: foundations, techniques, and applications*. 2nd ed. CRC Press.

Ware, Colin (2012). *Information visualization: perception for design*. 3rd ed. Elsevier.

Wilhelm, Adalbert (2008). "Handbook of data visualization". In: *Handbook of data visualization*. Ed. by Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. Springer. Chap. Linked Views for Visual Exploration, pp. 200–216.

Wills, Graham (2008). "Handbook of data visualization". In: *Handbook of data visualization*. Ed. by Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. Springer. Chap. Linked Data Views, pp. 218–239.

Zhu, Joe (1998). "Data envelopment analysis vs. principal component analysis: An illustrative study of economic performance of Chinese cities". In: *European journal of operational research* 111.1, pp. 50–61.

Overall Conclusion

Data visualization is an indispensable part of analytical studies in almost any quantitative fields, and is used for exploration and presentation purposes. However, in data envelopment analysis(DEA), data visualization is a relatively neglected topic. Although a dozen of DEA data visualization methods have been suggested in the literature, DEA researchers and practitioners rarely use these methods in their studies. Moreover, the usage of visualization in DEA is mainly limited to presentation of the results, rather than exploration of the data. The neglect of data visualization is to such extent that there is hardly any high-dimensional visualization method in DEA software packages.

Nevertheless, data visualization in general, and high-dimensional data visualization in particular can play an important role in gaining insight into the data and the problem, detection of structures, and identifications of anomalies, to just name a few. Visualization can present large amount of data, reveals emergent properties and various features of the data through retaining the information as much as possible. High-dimensional data visualization is the latest wave of developed techniques in the visualization research field, and this thesis is written based on it.

The thesis is composed of four independent yet inter-related articles. In other words, while each article is a stand-alone paper, each succeeding article has some references to the proceeding ones. All four articles are focused on DEA data visualization, and each plays a unique role.

The first article suggests a visualization framework for DEA cross-evaluation method. Cross-evaluation is one of the top 4 research fronts of DEA according to Liu et al. (2016), so this new tool is inline with the current DEA hot topics. However, the main reason behind choosing cross-evaluation has been the cross-efficiency matrix(CEM), which is a source of huge amount of data about the decision making units(DMU) and their relations. The current approaches to understand CEM, such as average cross-efficiency, do not even try to preserve the details of CEM, while visualization tries to graphically represent all the details, i.e. the DMUs and their relations. A guide to interpretation of the CEM plots is provided in the first article, followed by visualization of two real datasets. The visualizations are used to explore clusters of DMUs, and different types of outliers such as efficient influential units as well as maverick units. Nevertheless, the CEM visualization usage should not be limited to identification of these types of outliers.

The second article is a detailed investigation on the "maverick units". Following the visual information seeking mantra Shneiderman (1996), "Overview first, zoom and filter, then details-on-demand", after the overview of the CEM map on the

first article, I decided to "zoom and filter" in an interesting type of units called the mavericks, and "demanding further details" about them. The second article starts by a critical appraisal of the maverick literature in DEA, finding some caveats in the current literature and proposing possible improvements. Afterwards, a new visual technique based on CEM visualization is suggested to detect such units. However, visualization should be a supplement to numerical methods, and the second article also includes a new numerical index to identify maverick units. The idea of the new numerical index stems from the visual identification method.

The third article is a survey about visualization methods in DEA. It is the third article, because it includes the suggested method of the first article, and the maverick notion of the second article. This survey is important because there is no equivalent literature review on the DEA visualization topic, so the practitioners can evaluate the techniques and choose the ones that suits their goals, and researchers can assess the set, find the unavailable necessary tools in order to develop them. A real dataset is visualized through the survey using various DEA visualization techniques, and doing so has made the article a "visual survey", which facilitates its reading.

The forth article is an introduction to DEA-Viz, a new DEA software focused on visualization, and it includes a case study to illustrate the features of the software. I developed DEA-Viz in order to implement CEM visualization method, and facilitate using it. Then I added further DEA visualization techniques to promote DEA visualization. DEA-Viz is an unparalleled software, as none of the current DEA packages has similar high-dimensional visualization features as it has.

In conclusion, this thesis has several contributions to DEA literature. In the first article, a new method for visualization of cross-evaluation methodology is suggested, and based on this visualization method, a new index for detection of maverick units is proposed in the second article. Moreover, in the second article, the DEA literature related to maverick units is critically appraised, some caveats are highlighted, and improvements are recommended. A visual survey of current visualization methods of DEA is presented in the third article, and the forth article is an introduction to DEA-Viz, a unique applet for visualization of DEA problems. A brief case-study is also included in the forth paper.

The current set of papers could be improved on various points. Throughout this thesis and everywhere that is needed, the (dis)similarity between a pair of DMUs is measured by Euclidean distance, as Euclidean distance is the most common proximity measure. However, it is shown that Euclidean distance in high-dimensional space does not behave as expected, and in general when the number of dimensions is high, the points become approximately equi-distant, so the concepts of neighbourhood and proximity are undermined.Aggarwal et al. (2001) Hence, it seems that it is safer to avoid Euclidean distance in DEA high-dimensional problems. Moreover, the concept of proximity of two DMUs in DEA is not defined clearly, and doing so may lead us to development of a DEA specific proximity measure.

In addition to improvement of the chosen proximity measure, the visualization method of cross-evaluation can be enhanced. In that method, benevolent cross-efficiency matrixDoyle and Green (1994) has been used, and this usage was shown as an improvement from the previous usages. Nonetheless, the benevolent cross-

efficiency matrix is an approximation of the most benevolent optimum weight, and not the most benevolent optimum set. Also such optimum alternative is not unique due to the problem of multiplicity of alternate optima in DEA. Both of these caveats should be improved in the next research steps, and one suggestion to do so perhaps is using multiplicative DEA model in cross-efficiency which can resolve the mentioned shortages. (Cook and Zhu 2014)

The suggested maverick detection technique in this thesis, is basically a variety of the nearest neighbour based anomaly detection technique.(Chandola et al. 2009) However, other techniques, such as clustering based, may be computationally more efficient, and may yield new perspectives to the data.

DEA-Viz is currently a prototype and at its earliest versions. Hence, many new features can enrich its toolbox. For instance, using linked data views can ease understanding of the data, by simultaneously looking at the data from two or more different perspectives, i.e. linked-views.Wills (2008)

Beside the possible improvements in the already available concepts and features, the current framework can be expanded from several aspects. For instance, new variables, such as slacks, can be added to the visualization datasets in order to have new perspectives into the DEA problems. Moreover, visualization of panel data would be a significant progress in DEA visualization.

At last, dimension-reduction techniques, such as t-SNE Maaten and Hinton (2008), or approaches such as network visualization can enhance DEA visualization toolbox.

Bibliography

Aggarwal, Charu C, Alexander Hinneburg, and Daniel A Keim (2001). "On the surprising behavior of distance metrics in high dimensional spaces". In: *ICDT*. Vol. 1. Springer, pp. 420–434.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3, p. 15.

Cook, Wade D and Joe Zhu (2014). "DEA Cobb–Douglas frontier and cross-efficiency". In: *Journal of the Operational Research Society* 65.2, pp. 265–268.

Doyle, John and Rodney Green (1994). "Efficiency and cross-efficiency in DEA: Derivations, meanings and uses". In: *Journal of the operational research society* 45.5, pp. 567–578.

Liu, John S, Louis YY Lu, and Wen-Min Lu (2016). "Research fronts in data envelopment analysis". In: *Omega* 58, pp. 33–45.

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.Nov, pp. 2579–2605.

Shneiderman, Ben (1996). "The eyes have it: A task by data type taxonomy for information visualizations". In: *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, pp. 336–343.

Wills, Graham (2008). "Handbook of data visualization". In: *Handbook of data visualization*. Ed. by Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. Springer. Chap. Linked Data Views, pp. 218–239.