Data Envelopment Analysis for Cities Efficiency

F Ali, H Njadat and Q Yaseen
Jordan University of Science and Technology
Irbid, Jordan 22110
{faali14, qmyaseen, najadat}@just.edu.jo

Abstract. This paper discusses the envelopment analysis for selected cities. It uses the DEA (Data Envelopment Analysis) for both input and output oriented methods using a data set of 176 cities, and tests several classifiers in classifying the dataset using both cross-validation and percentage split.

1. Introduction

Data Envelopment Analysis (DEA) is used to estimate the production frontiers in the fields of economics and operations research. Empirically, it measures the decision making units DMUs according to the productive efficiency. Generally speaking, data envelopment analysis (DEA) is a data-driven way of measuring efficiency for each decision making unit (DMU) [1][2]. In addition to using DEA in production theory, it is an important tool for benchmarking in operations research.

DEA is used in many applications such as the electric utilities sector, where DEA is used as a tool to determine and specify individualized regulatory rates for firms given their comparative efficiency [3]. Moreover, DEA is used to assess the efficiency of various non-profit or public organizations such as hospitals [4], police forces [5], etc.

Measuring the cities efficiency is crucial for sustainable development of cities. Minimizing the resources consumptions and maintaining and improving the delivered services and the construction activities level is mandatory for cities efficiency. The aforementioned measuring process is performed using the DEA tool.

This work discusses the cities efficiency and uses the DEA for this purpose. It calculates the efficiency and shows how correctly the classifiers work to classify the cities in the used data set using many of the DEA models [6]. The rest of work is organized as follows. Next section discusses some related work. Section 3 introduces the approach and the results. Finally, section 4 concludes the work.
2. Related work

Some research about using DEA for cities efficiency in China were performed by Fang et. al [7] and Bian and Yang [8]. The authors investigated the DEA for the China’s Urban Agglomerations UAs using various sizes of populations and locations of about 35 central cities. They discussed the relationship among the elements the decomposed model and compares their efficiency performance. Moreover, the came up with important recommendations about improving the UAs efficiency performance.

Kuosmanen and Kortelainen [9] and Loikkanen and Susiluoto [10] used the DEA for cities efficiency in Finland. The authors used a large dataset of regional input and output variables and regional characteristics of about 83 NUTs 4-level regions in Finland during 1988-1999. The used the DEA to get the annual efficiency scores of all regions, and then, they explained the differences in inefficiency among regions using econometric approaches (logistic regression, Tobit Models).

Many other researchers investigated the using of DEA in many countries such as Raab and Lichty [11] who used it in USA cities, Bal and Orkcu [12] who used DEA in Turkey, and Suzuki et al. [13] who used the DEA to measure cities efficiency in Italy.

3. The Approach

The approach starts by obtaining the dataset that contains 176 DMU [14]. Each DMU represents a country with 14 attributes as input and 5 attributes as output. The attributes are { Real per capita GDP (Gross Domestic Product), Real per capita GDP growth, Population, Growth in population, Secondary school enrollment rates, IMF indicator of capital account restrictions, Institutional Investor Risk Ratings, Trade openness (exports + imports)/GDP, Financial openness (financial account inflows + financial account outflows)/GDP, Capital inflows (Inward FDI + inward portfolio + inward other)/GDP, Correlation of annual growth rates with OECD (Organization for Economic Cooperation and Development) average annual growth rates, Average annual growth rate of GDP deflator, Proportion of population speaking English, Proportion of population speaking major European language, Total ODA from all sources, Real per capita GNP, Oil export, Non-Oil Commodity Exporters, Ethnolinguistic fractionalization } and delete these fields from the dataset { Investment to GDP at PPP, Terms of trade shock (Growth in export deflator minus growth in import deflator, weighted by current LCU shares of exports and imports in GDP), Unweighted terms of trade shock }. The work is divided into three main parts, which are:

A. Data preprocessing: The dataset is not ready for DEA. Some processing has to be performed on the data to prepare it. First, some attributes are removed, while others are normalized to remove negative numbers since DEA does not deal with negative numbers. Moreover, the dataset contains missing values. In this case, it is filled by the attribute average. This preprocessing task prepares the dataset for efficiency calculation.

B. Efficiency calculation: The efficiency is calculated for the cities using input and output oriented and both CRS and VRS using EMS (Efficiency Measuring System). Figure 1 and Figure 2 show the calculation of input orient for both CRS and VRS. Figure 3 and Figure 4 show the efficiency calculation of output oriented for both CRS and VRS.
### Figure 1. Efficiency calculation for input oriented CRS.

![Table: Data for Efficiency Calculation](image)

### Figure 2. Efficiency calculation for input oriented VRS.

![Table: Data for Efficiency Calculation](image)
Figure 3. Efficiency calculation for output oriented CRS.

Figure 4. Efficiency calculation for output oriented VRS.
The figures show the number of efficient and inefficient DMUs as follows:

- The calculation for input oriented.
  - Number of efficient DUMs are 153.
  - Number of inefficient DUMs are 23.
  - For both CRS, VRS.
- The calculation for output oriented.
  - Number of efficient DUMs are 152, 154.
  - Number of inefficient DUMs are 24, 22.
  - For CRS and VRS.

C. Classification: After calculating the efficiency for the DMUs, the approach classifies the DMUs using the DMUs efficiency as a class using the classifiers: Decision tree (J48), K nearest neighbor (KNN), Support vector machine (SVM) and NaïveBayse using both cross-validation and percentage-split for the dataset. The classification result is shown in table 1.

| Classifier | Accuracy Cross Validation | Accuracy Percentage Split |
|------------|---------------------------|---------------------------|
| 1          | J48                       | 85.2273 %                 | 75%                       |
| 2          | Naive Bayes               | 74.4318 %                 | 78.3333 %                 |
| 3          | SVM                       | 86.9318 %                 | 75%                       |
| 4          | KNN-9-1                   | 86.9318 %                 | 75%                       |

As shown in the table, the best classifiers are the SVM and KNN with K=9 which achieved an accuracy of 86.9318 %. The decision tree was the next and achieved an accuracy of 85.2273 %. Since the class is imbalanced in the dataset with 153 as efficient country and 23 as inefficient country, the SMOTE filter is used in Weka to balance data. Table 2 shows the results after balancing the data.

| Classifier | Accuracy Cross Validation | Accuracy Percentage Split |
|------------|---------------------------|---------------------------|
| 1          | J48                       | 87.2131 %                 | 89.4231 %                 |
| 2          | Naive Bayes               | 74.4262 %                 | 75 %                      |
| 3          | SVM                       | 79.6721 %                 | 75%                       |
| 4          | KNN-4                     | 83.2787 %                 | 79.8077 %                 |

The table shows that the best classifier is the decision tree with accuracy of 89.4231 % followed by KNN with K=4 and accuracy = 83.2787 %.

D. Clustering: After the classification process, the K-mean cluster is used for the clustering process. The results are shown in Table 3, where the clustering accuracy is 54.7541%.
Table 3. The clustering accuracy.

| Clusterer  | Accuracy                                      |
|------------|-----------------------------------------------|
| 1 K-mean   | Incorrectly clustered instances: 45.2459%     |
|            | Correctly clustered instances: 54.7541%      |

4. Conclusion

This paper has discussed the envelopment analysis for a number of cities using a number of classifiers. The paper has shown that both input and output oriented DEA achieves almost the same result. The best classifiers were the support vector machine and k nearest neighbor in the unsmoothed data with accuracy of 86.9318%. In the smoothed data, the best classifier was the decision tree with accuracy of 89.4231%. However, the clustering shows a low accuracy of 54.7541% using the K-means.

References

[1]. Adler N, Lea Friedman L and Sinuany-Stern Z 2002 Review of ranking methods in the data envelopment analysis context European Journal of Operational RESEARCH
[2]. Kaoru Tone K and Miki Tsutsui M An epsilon-based measure of efficiency in DEA revisited-A third pole of technical efficiency Tokyo, Japan.
[3]. Berg S. 2010. Water Utility Benchmarking: Measurement, Methodology, and Performance Incentives. International Water Association.
[4]. Vera A.; Kuntz L. 2007. Process-based organization design and hospital efficiency. Health Care Management Review. 32 (1), pp. 55–65.
[5]. Thanassoulis E 1995. Assessing police forces in England and Wales using data envelopment analysis. European Journal of Operational Research. 87 (3), pp. 641–657.
[6]. Jacobs I and Bean C 2004 Data Envelopment Analysis, Theory Techniques for Economic and Operation Research, Cambridge University.
[7]. Fang, C., et al., 2013. Input output efficiency of urban agglomerations in China: an application of data envelopment analysis (DEA). Urban Study, 50 (13), pp. 2766–2790.
[8]. Bian, Y., Yang, F., 2010. Resource and environment efficiency analysis of provinces in China: a DEA approach based on Shannon’s entropy. Energy Policy, 38, pp.1909–1917
[9]. Kuosmanen, T., Kortelainen, M., 2005. Measuring eco-efficiency of production with data envelopment analysis. J. Ind. Ecol. 9 (4), pp.59–72
[10]. Loikkanen, H., Susiluoto, I., 2002. An evaluation of economic efficiency of Finnish regions by DEA and TOBIT models. In: 42nd Congress of the European Regional Science Association, Dortmund.
[11]. Raab, R., Lichty, R., 1997. An efficiency analysis of Minnesota counties: a data envelopment analysis using 1993 IMPLAN input–output analysis. J. Reg. Anal. Policy. 27 (1), pp. 75–93.
[12]. Bal, H., Orkcü, H.H., 2011. A new approach to cross efficiency in data envelopment analysis and performance evaluation of Turkey cities. Gazi Univ. J. Sci., 25 (1), pp. 107–117
[13]. Suzuki, S., Nijikamp, P., Rietveld, P., 2011. Regional efficiency improvement by means of data envelopment analysis through Euclidian distance minimization including fixed input factors: an application to tourist regions in Italy. Reg. Sci., 90 (1), pp. 67–89.
[14]. Easterly W and kraay A 2000. Small States, Small Problems? Income, Growth, and Volatility in Small States, The World Bank.