Use of Main Components and Data Envelopment Analysis to Evaluate the Efficiency of Water Operators

José Sánchez
Uriel Rivera-Ortega
(UPAEP 2018)

Abstract

The purpose of this work is to analyze the efficiency of Water Operator Organizations of México from the premise that each operator organization is in different working conditions and evaluated from different performance indicators, so obtaining efficiency for each organism could ambiguously arise in this regard, this manuscript paper provides from analysis of generalized variables and their relationship with factors of operation those variables that influence the efficiency process. This document is based on a free database from the Program of Indicators of Management of Operative Organizations (PIGOO in spanish), which initially involves 139 municipal water systems throughout the Mexican Republic. The result highlights the influence of certain variables in the performance of the operation of water systems, according to this, the usefulness of this work is expected to contribute by defining elements that improve the competitiveness of the aforementioned organizations.

Keywords: Data Envelopment Analysis (DEA), Principal Components Analysis (ACP), Water Operators

0 Introduction

Efficiency is presented as an indicator of the competitiveness of organizations. There are various studies that have displayed methodologies by analyzing multiple criteria, in order to improve strategic competitiveness [1]. These techniques as an efficiency measurement method try to evaluate the performance of an entity with respect to an optimal value. Although it is possible to identify the best governmental practices by comparing economic elements, the optimization process of the organizations is guided by more than one measure of performance, which is a problem when there is not explicit relationship between performance measurements. The solution for this type of problems is not concentrated on obtaining a single optimal solution, but on creating a set of favorable solutions called efficient frontier [2]. To find this efficient frontier a nonparametric technique of interest is the Analysis Data Envelopment (DEA) which shows the maximum relation of the products (outputs) by giving their inputs (inputs).

In addition to the problem of conflict between measurements of performance, there is also the problem related to the amount of data associated with a number of variables performance; the number of variables in the DEA study should ensure that the number of alternatives is sufficiently discriminated to obtain the efficient frontier [3]; although for the first problem DEA technique resolves the conflict between
performance measurements. The second problem is approached from the technique called Principal Components Analysis (PCA), which aims to reduce data by grouping relatively homogeneous variables. This transformation of a wide set of variables with high correlation leads towards a reduced set of variables that explains most of the variations in the data [4].

1 Theoretical framework

The structured theory about the exploration and understanding of the competitiveness factors, as well as the relationship between the variables and the determination of the efficient frontier, are concentrated in the Principal Component Analysis and DEA techniques which will be below.

1.1 Principal Component Analysis

In the most investigations and study cases, the most frequent is to take as much information as possible and collect the largest number of variables involved and, consequently, a quantity of data of a different category. It is possible that in a collection of research data, these are interrelated from the different variables in which they have been included, this condition presages variability in a study, so it is necessary to reduce the number of variables under the theoretical justification that variables with strong correlation are actually measuring the same concepts but from different approaches.

The PCA consists of concentrating the information contained in an original set of variables to take it to another set of variables, always in smaller quantity than the original ones, therefore if there is a set of K original variables, the information is transformed into a set of W components being $W < K$ [4], each of these W components are translated into factors as a result of a linear combination of the K variables. The utility of the PCA is to take advantage of the fact that the resulting factors reflect the variability of the original groups. In PCA the first main component is the axis that passes through the center of the data and minimizes the distance of each data point to the same axis explaining the behavior of a group of data, the rest of the components arise as other axes that have the purpose of explaining what the first main component could not do, however one characteristic to be fulfilled by the following main components is that they must be orthogonal (independent) and pass through the centroid of the data. The variability explained by each axis gives place (originates) to the concept of eigenvalue which can also be expressed as a percentage of the total variation.

The mathematical model of the ACP is defined as follows [4]:

$$X_j = a_{j1}F_{1i} + a_{j2}F_{2i} + ... + a_{jk}F_{ki}$$  \hspace{1cm} (1.1)

Where $X_j$ is the value of the j-th variable in the i-th case resulting from the product of $a_{1j}$, $a_{2j}$, ..., $a_{kj}$, as vector of constants and each of the factors $F$.

This model expresses that the information of the variables is explained entirely by the "K" factors. De la Garza & González [5] worked with the proportion of the variability of each variable by the factors, this is the reason that in PCA the initial value of all the variables is equal to 1. The components are chosen according to the highest variance, a form
to maximize the variance is increasing the coefficients \( a_{kj} \) maintaining the orthogonality of the transformation, which requires that the vector \( a_{i,j}, a_{2,j}, ..., a_{K,j} \) be equal to 1, that is [4]:

\[
\sum_{i=1}^{K} a_{ij}^2 = 1
\]  

(1.2)

1. 2 Data Envelopment Analysis (DEA)

The objective of a Data Envelopment Analysis is the formation of enveloping faces that define the efficient and inefficient units. In this case it is as important to know the "best" as to know the distance (actions) that separates the efficient Decision Measurement Units (DMU) of those who are not. This will be a guideline for the progress of the entities who wish to achieve the called benchmarks (optimal entities). This situation presents an implicit problem in relation to the "goals" that less efficient DMUs must reach because common sense explains the impossibility of competing with benchmarks for which the efficiency is too great to look for a comparison, in terms of González and Álvarez [6] DEA analyzes could be reconfigured to seek a single-stage procedure to achieve closer efficiency goals.

While a virtue of the DEA is obtaining efficiency in the use of multiple units in its function of inputs or products, also presents some criticism in that it does not contemplate influences on the process productive, which generates uncertainty in the final results, in studies by Drake and Howcroft [7] it is mentioned that the DEA is capable of working better if the number of observations is close to twice the sum of Inputs and Outputs, which would indicate that in studies with small samples should be added relatively too many categories, this results in a complexity to identify the optimal DMU, however the DEA models create from iterations a progress in the proposal to choose the efficient DMU [8].

2 Methodology

This work was carried out through an investigation of institutional references related to competitiveness factors that allude to the efficiency of municipal entities. From the database of PIGOOG, an explanatory study was documented, since the appropriate variables for the efficiency study were determined. Initially, 139 municipal water operators were considered from all over the republic. However, under a discrimination process, the study contemplated 51 organisms, the reason for excluding certain organisms consisted of not having complete information in all the study variables from 2010 to 2015; the research process is used for the integration of statistical techniques in order to model the influence of social variables in the performance of municipal water operators.

2.1 Selection of variables with ACP

The study begins with the collection of the information corresponding to the variables involved with the performance of the water operators in the different municipalities of the country; showing that isn’t competition in the service offered regarding to the distribution of the drinking water and drainage network, because each municipality and/or the metropolitan area is represented by a single water operator.

Under this principal condition, each municipality will be considered as a DMU object of efficiency study under the following indicators:
Table 1. List of indicators taken from the PIGOO database ($=\)mexican pesos$).

With this number of variables which in turn include data for 51 utilities for a period of 5 years, the ACP technique is used to determine whether more than one variable is strongly correlated with a different variables that are part of the study, that in technical terms one can consider multicollinearity. In these conditions it is convenient to consider the benefit of grouping these correlated variables considering that they are actually measuring the same issue.

Due to the number of variables used, it was not possible to plot the point cloud that would visually reflect the behavior of the data, however, for the ACP model, this requirement is not necessary in the development of the study. In this way, the first step of the ACP will be the choice of the main components under the premise that at least 90% of the variability of the data is explained. This value is statistically typical value in studies of this nature [9]

The separation of the variables was carried out in two types: Inputs and Outputs. This classification does not concern with the development of the ACP but with the efficiency process of the DEA. So the variables that will be worked in ACP are:

Table 2. List of indicators of inputs variables for the ACP study ($=\)mexican pesos$).
When the components have been selected, they are presented in the form of a correlation matrix. This step allows confirming the relationship between variables, in this case variables type input. The following table shows a triangular formation with its diagonal with values of 1 because the same concepts exist in the columns and rows axis, so the variable C in the row and column crossing will have correlation value of 1 since it is perfectly correlated to be the same concept. Values above 0.75 indicate strong correlation and because the correlation offers the same result regardless of the order of the even variables, the table presents a lower triangular matrix. To simplify the presentation, each variable has been represented by the variable number taken from table 2.

Table 4 shows that up to variable 15 that exceeds 90% of the total variability of the data collected.

The next step was to consider the data in the analysis of principal components. It was chosen to use the software Minitab due to its performance in statistical data analysis.

Table 4 shows that up to variable 15 that exceeds 90% of the explained variability in the accumulated variable (highlighted in yellow on the table). Variable 1 explains this variability in 15.2%, however the other variables (up to 23 according to table 2) they contribute individually no more than 2%, so to achieve an accumulated 90% of the explanation of variability it is necessary to consider most of the variables, which is confirmed with fig. 1.
The projection of the original data in the main components created is the prelude to the application of the DEA technique. The DEA study consists of finding the organisms that are optimal in terms of efficiency, that is, a line of action for maximization is proposed of the efficiency of those entities that belong to the so-called efficient frontier [10].
Table 6 shows the list of entities that will work as DMUs for the DEA study, the inputs will be the main components obtained in the previous section and the outputs will be the indicators that are attached to what the clients of the operating agencies of the water perceived as the service offered.

<table>
<thead>
<tr>
<th>Water Operator</th>
<th>DMU</th>
<th>Water Operator</th>
<th>DMU</th>
<th>Water Operator</th>
<th>DMU</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aguascalientes</td>
<td>1</td>
<td>Guadalupe, Chihuahua</td>
<td>11</td>
<td>Huauchinango, Puebla</td>
<td>21</td>
<td>Naucalpan, México</td>
<td>Drinking water coverage reported (%)</td>
</tr>
<tr>
<td>Cancún, Quintana Roo</td>
<td>2</td>
<td>Culiacán, Sinaloa</td>
<td>12</td>
<td>Ixtapaluco, Puebla</td>
<td>22</td>
<td>Nicolás Romero, México</td>
<td>Sewage coverage reported (%)</td>
</tr>
<tr>
<td>Cd. Juárez, Chihuahua</td>
<td>3</td>
<td>Delicias, Chihuahua</td>
<td>13</td>
<td>La Piedad, Michoacán</td>
<td>23</td>
<td>Pachuca, Hidalgo</td>
<td>Consumption (kWh)</td>
</tr>
<tr>
<td>Cd. Mante, Tamaulipas</td>
<td>4</td>
<td>Durango, Durango</td>
<td>14</td>
<td>León, Guanajuato</td>
<td>24</td>
<td>Piedras Negras, Coahuila</td>
<td></td>
</tr>
<tr>
<td>Cd. Valles, San Luis Potosí</td>
<td>5</td>
<td>Ensenada, Baja California</td>
<td>15</td>
<td>Mochuelo, San Luis Potosí</td>
<td>25</td>
<td>Puebla, Puebla</td>
<td></td>
</tr>
<tr>
<td>Córdoba, Veracruz</td>
<td>10</td>
<td>Hermosillo, Sonora</td>
<td>20</td>
<td>Monterrey, Nuevo León</td>
<td>30</td>
<td>Sílao, Guanajuato</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. DMU’s, Inputs and Outputs.

To follow the case study, we proceeded with the study of the outputs. Knowing the premise that an omission of an important input or output results in biased conclusions for an AED analysis. The use of correlation analysis and main components prior to an efficiency study was vital for the validation of DEA results. Table 7 shows the results of the linear combination that produced the ACP and that are established as inputs, in the same table the output values that correspond to each DMU appear.
The DEA analysis was run in the Excel application and under the general mathematical model of DEA. Table 7 shows the results that in addition to the efficient DMUs include the slack values of the technique, the latter are particularly important when carrying out decisions on the future performance of water utilities. The values of 1.00 in the box of “Eff. Score” correspond to the DMUS whose frontier value in one of the indicators presents it as an efficient frontier, that is, the efficiency reaches the value of 1, for the case of the DMUSs with a score lower than 1 indicates "remote" that of a border value.

Table 7. Inputs and outputs for DEA.
Conclusions

In this manuscript it has been studied how the main components of the indicators of water operators contributed in the essence of the behavior of the input variables for the measurement of the efficiency, in fact the study was fulfilled in the sense of collating with greater knowledge the influence of variables that affect the global performance of the operators of water.

Due to the efficiency results shown in Table 8, where a total of 11 non-efficient DMUs are observed out of a total of 51 DMUs studied, it can be deduced that in the field of water operators, the existence of several performance indicators can position a water operator agency as efficient even when a particular variable does not comply. In this sense, this work should help organizations operator of water in subsequent studies in which critical variables are involved to standardize at a national level.

References


