Método heurístico de selección de variables para el incremento del poder discriminatorio de DEA: caso de aplicación a las EPS Colombianas.

A heuristic variable selection method to increase the discriminatory power of DEA: A Case Study of Colombian EPS.

Andrés M. Gómez
angomezar@unal.edu.co
Escuela de Estadística, Universidad Nacional de Colombia, Medellín, Colombia.

Andrés M. Villegas
a.villegas@unsw.edu.au
Escuela de Riesgos y Estudios Actuariales, Universidad de Nueva Gales del Sur. Sídney, Australia.

Juan G. Villegas
juan.villegas@udea.edu.co
Departamento de Ingeniería Industrial, Universidad de Antioquia, Medellín, Colombia.

Resumen

El análisis envolvente de datos (DEA) es una técnica no paramétrica para medir la eficiencia relativa de un conjunto de unidades productoras homogéneas (denominadas DMU). Sin embargo, DEA no proporciona pautas claras para la selección de las variables, lo cual es crucial para la precisión y relevancia de los resultados, ya que cuanto mayor sea el número de variables de entradas y salidas en DEA, mayor será la dimensionalidad del espacio de solución y menos exigente será el análisis. Por consiguiente, mayor será la probabilidad de que algunas unidades ineficientes sean clasificadas como eficientes. En este trabajo presentamos una aproximación heurística en la selección de variables, tanto de entrada como de salida, para mejorar el poder discriminatorio de los modelos de DEA, basada en la eliminación iterativa de variables. La metodología propuesta elige en cada iteración la configuración de variables donde la métrica de impureza (índice Gini o entropía) es máxima. Esta propuesta pretende evitar los criterios ad hoc o juicios particulares en la elección de las variables, proporcionando un conjunto de modelos seleccionados por índices estadísticos de dispersión. Usando datos financieros para medir la eficiencia de EPS colombianas, ilustramos la metodología propuesta, con la cual encontramos una reducción del 71 % al 6 % en el número de DMUs clasificadas como eficientes, con tan solo la selección del 34 % de las variables iniciales y con una retención del 95 % de la varianza total de los datos.

Palabras claves: *Análisis envolvente de datos, Selección de variables, Problema de dimensionalidad, Medición del desempeño*

Abstract

Data Envelopment Analysis (DEA) is a non-parametric methodology based on mathematical programming to estimate the performance of a set of Decision Making Unit (DMUs) which use the same inputs to produce the same outputs. However, the DEA methodology does not provide clear guidelines for variable selection, which might prove problematic in the presence of a large number of input and output variables. In such a situation, the solution space

is high dimensional resulting in a significant probability that inefficient DMUs are deemed efficient and hindering the discriminatory power of the methodology. In this work, we present a heuristic approach to improve the discriminatory power of DEA models, based on the iterative elimination of variables. The proposed methodology chooses in each iteration the configuration of variables where the impurity metric is maximum. This proposal aims to avoid ad hoc criteria or subjective judgments in the choice of variables, by providing instead a set of models selected using statistical indices of dispersion. We illustrate the proposed methodology using financial data to measure the efficiency of Colombian Health Promoting Entities (EPS). In this case study, we find that the number of efficient DMUs is reduced from 71% in the DEA model with all variables to 6% in our method. This is achieved by selecting 34% of the initial variables while still explaining 95% of the variance of the data.

Keywords: *Data envelopment analysis, Variable selection, curse of dimensionality, Performance measurement*

1. Introducción

El análisis envolvente de datos (DEA, por las siglas de *Data Envelopment Analysis*) es una metodología no paramétrica basada en programación matemática desarrollada por Charnes, Cooper & Rhodes (1978) para medir la eficiencia relativa de un conjunto de unidades productoras homogéneas (denominadas DMU, *Decision Making Units*) que utilizan el mismo conjunto de insumos para producir el mismo conjunto de productos. DEA calcula la eficiencia para cada DMU con base en ponderaciones, en los insumos y productos, elegidos de tal manera que cada DMU alcance la máxima eficiencia posible. DEA hace posible la construcción de una frontera eficiente entre las diferentes unidades, con el fin de identificar las unidades eficientes que se usan como referentes del grupo evaluado.

Esta metodología no requiere una formulación explícita de las relaciones funcionales de las entradas con las salidas del proceso, permitiendo la estimación de una medida general de eficiencia a partir de valores observados de múltiples entradas y múltiples salidas, sin requerir el uso de ponderaciones a priori. Aunque existen algunas reglas empiricias para determinar el número máximo de variables de entrada y de salida (Dyson, Allen, Camanho, Podinovski, Sarrico & Shale, 2001) (Cook, Tone & Zhu, 2014), la metodología DEA no proporciona pautas objetivas para la selección de las variables, lo cual es crucial para la precisión y relevancia de los resultados. Cuanto mayor sea el número de variables de entradas y salidas en DEA, mayor será la dimensionalidad del espacio de solución y menos exigente será el análisis. Y por consiguiente, mayor será la probabilidad de que algunas unidades ineficientes sean clasificadas como eficientes.

Entre los estudios que se han enfocado en mejorar el poder discriminatorio de DEA, existen algunos con la intención de aumentar el número de DMUs y conservar el mismo número de variables, usando datos de corte transversal y series de tiempo (Charles, Aparicio & Zhu, 2019). En cambio, otros han incorporado técnicas estadísticas como el análisis de covarianza parcial (Jenkins & Anderson, 2003), o análisis de componentes principales combinado con DEA (Adler & Golany, 2002), con el objetivo reducir el número de variables utilizadas y conservando el mismo número de DMUs.

En este trabajo se presenta una aproximación heurística para mejorar el poder

discriminatorio de los modelos de DEA, basada en la eliminación iterativa de variables, tanto de entrada como de salida. La metodología propuesta elige en cada iteración la configuración de variables donde la métrica de impureza (índice Gini o entropía) es máxima, proporcionando un conjunto de modelos seleccionados por índices estadísticos de dispersión. Esta propuesta pretende evitar los criterios ad hoc o juicios particulares en la elección de las variables, lo cual es común en la práctica (Dyson et al., 2001) (Cook et al., 2014).

Utilizando datos financieros de las entidades promotoras de salud (EPS) en Colombia, se pone en evidencia el problema de dimensionalidad, y se analizan las configuraciones arrojadas por el algoritmo de búsqueda iterativo contrastando el poder discriminatorio y la varianza explicada-retenida de cada modelo. Para el caso de estudio realizado se evidencia una reducción del 71 % al 6 % de DMUs clasificadas como eficientes, y lo anterior se logró seleccionando menos de la mitad las variables iniciales (34 %) con una varianza retenida total del 95 %. A continuación, se describe el procedimiento iterativo propuesto.

2. Metodología

La estructura básica del algoritmo heurístico propuesto se describe en la Tabla 1, el cual fue implementado en `R` (R Core Team, 2021) haciendo uso de la función `dea` de la librería `Benchmarking` (Bogetoft & Otto, 2020) para el cálculo de las eficiencias. Este método de solución está basado en el concepto de búsqueda iterativa donde se aumenta el poder discriminatorio de DEA, maximizando el índice Gini o la entropía de los puntajes de eficiencia. Los parámetros de inicialización son los valores de las variables de entrada y de salida (en las filas las DMS), y el criterio de parada que puede ser: el valor mínimo sugerido en las reglas empíricas (como $\frac{1}{2}$ o $\frac{1}{3}$ de la cantidad de DMUs) (Dyson et al., 2001) (Cook et al., 2014) o un valor mínimo (valor de épsilon=0.01 por defecto) en la variación entre la métrica de impureza máxima anterior y la actual.

Como se ejemplifica en la Figura 1 Para un ejemplo con dos entradas (I_1 I_2) y tres salidas (O_1 O_2 O_3), inicialmente se calculan los puntajes de eficiencia para el conjunto total de variables de entrada y salida (Iteración 0 - I_1 I_2 O_1 O_2 O_3), donde la métrica de impureza se espera que sea baja ya que muchas DMUs pueden ser clasificadas como eficientes erróneamente y la variancia retenida es del 100 %. Luego con las variables del modelo anterior, se crea un conjunto con todas las posibles combinaciones de variables, excluyendo solo una en cada configuración, tanto para las variables de entrada como de salida; a estos modelos candidatos se les calcula su eficiencia, sus métricas de impureza y su varianza explicada. Después, se elige la combinación de variables que maximiza la métrica de impureza, y se repite el proceso de forma iterativa hasta que se cumpla el criterio de parada. Siguiendo el ejemplo de la Figura 1, en la iteración 1 se elige el modelo con las variables I_1 I_2 O_2 O_3, las cuales serán el insumo de la iteración 2, para la cual se elige el modelo I_1 O_2 O_3, finalmente, el modelo que mejora el poder discriminatorio de DEA en la iteración 3 seria el modelo con las variables I_1 O_3.

Tabla 1: Metodología aplicada

Algoritmo de búsqueda iterativa

1: Parámetros: input_data, output_data, stop_criterion, épsilon (por defecto 0.0001)
2: dmu : número de unidades a evaluar
3: m: entradas
4: s: salidas
5: M: ${}_mC_m = \binom{m}{m}$
6: S: ${}_sC_s = \binom{s}{s}$
7: $Umbral \leftarrow$ stop_criterion $\in \{$épsilon, $\lfloor dmu/2 \rfloor, \lfloor dmu/3 \rfloor\}$
8: Calcular métricas: Eficiencia(M-S), Impureza(M-S), Varianza_Retenida(M-S)
9: $Bandera \leftarrow |m| + |s|$
10: **Mientras que** $Bandera > Umbral$ **haga**
11: M': ${}_mC_{m-1} = \binom{m}{m-1}$
12: S': ${}_sC_{s-1} = \binom{s}{s-1}$
13: **Si** $|m| > 1$ **entonces**
14: **Para** $i = 1$ hasta M' **haga**
15: Calcular métricas: Eficiencia(M'_i-S), Impureza(M'_i-S), Varianza_Retenida(M'_i-S)
16: $i = i + 1$
17: **Fin Para**
18: **Fin si**
19: **Si** $|s| > 1$ **entonces**
20: **Para** $j = 1$ hasta S' **haga**
21: Calcular métricas: Eficiencia(M-S'_j), Impureza(M-S'_j), Varianza_Retenida(M-S'_j)
22: $j = j + 1$
23: **Fin Para**
24: **Fin si**
25: Guardar y acumular proceso (modelos candidatos)
26: **Si** $arg\ max_{M'}(Impureza(M')) > arg\ max_{S'}(Impureza(S'))$ **entonces**
27: $M \leftarrow arg\ max_{M'}(Impureza(M'))$
28: $|m| \leftarrow |m| - 1$
29: **Si no**
30: $S \leftarrow arg\ max_{S'}(Impureza(S'))$
31: $|s| \leftarrow |s| - 1$
32: **Fin si**
33: Guardar y acumular ruta (modelos máximos heurísticamente)
34: **Si** stop_criterion = épsilon **entonces**
33: $Bandera \leftarrow \Delta Impureza(\text{Modelos}_{\text{máximos heurísticamente}})$
29: **Si no**
33: $Bandera \leftarrow |m| + |s|$
32: **Fin si**
35: **Fin Mientras**
36: Retornar el proceso (modelos candidatos) y la ruta (modelos máximos heurísticamente)

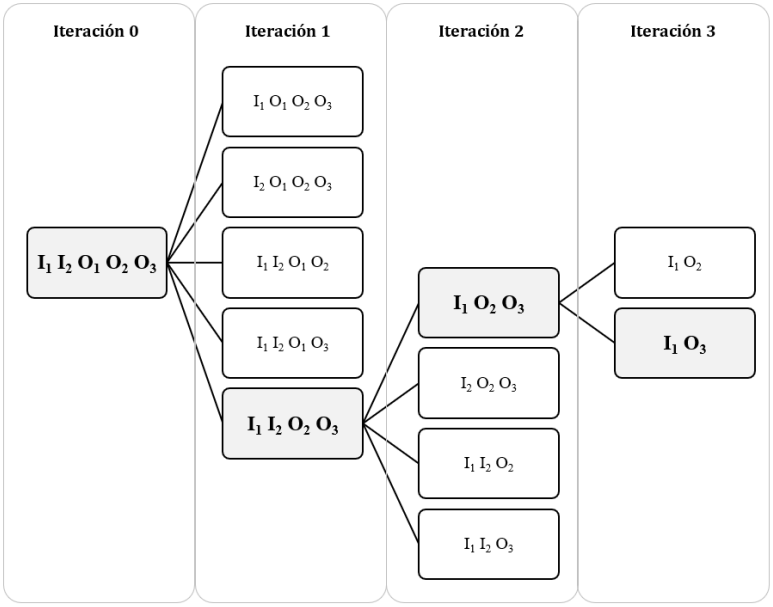

Figura 1: Ejemplificación del proceso iterativo

3. Caso de estudio

Los datos usados en este trabajo fueron tomados de Fontalvo Herrera (2017), quien analizó la eficiencia del sector salud Colombiano. La información está compuesta por indicadores financieros del 2011, reportados en la Superintendencia de Salud para 17 entidades promotoras de salud (EPS) que hacen parte del régimen contributivo, con la formulación CCR-O, es decir, rendimientos constantes a escala - orientado a las salidas, donde la eficiencia técnica $= 1/\eta^*$ (para la formulación del modelo CCR-O, véase la Ecuación 1) (Cooper, Seiford & Tone, 2007). Vale aclarar que este estudio no cumple la regla empírica que establece que la cantidad de DMUs deber ser tres veces la cantidad total de las variables, para este caso en particular se debieron haber evaluado como mínimo 21 DMUs dadas las 7 variables que fueron utilizadas en el modelo. La Tabla 2 resume el valor de las variables de entradas y salidas.

$$\begin{aligned} \max_{\eta,\mu} \quad & \eta \\ \text{subject to} \quad & x_o \geq X\mu \\ & \eta y_o \leq Y\mu \\ & \mu \geq 0 \end{aligned} \tag{1}$$

Dichas Variables están compuestas cuatro entradas de entrada: los activos totales (I_1), las inversiones (I_2), las cuentas por cobrar FOSYGA (I_3), y los activos fijos (I_4) como la propiedad, planta y equipo con los que cuentan las EPS. Y tres variables de salidas, que son : los ingresos por la unidad de pago por captación de usuarios (O_1), los ingresos por recobros

Tabla 2: Entradas y salidas de 17 EPS

	DMU	I 1	I 2	I 3	I 4	O 1	O 2	O 3
DMU 1	Aliansalud	74,659,289	19,712,425	43,573,089	506,726	196,384,549	81,430,596	193,387,216
DMU 2	Comfenalco Valle	73,192,236	810,002	36,472,629	15,382,635	154,675,269	39,994,397	138,684,343
DMU 3	Compensar	147,183,633	15,127,388	75,399,989	4,243,481	408,946,118	108,062,571	425,911,666
DMU 4	Coomeva	552,127,504	12,757,631	403,192,739	12,549,111	1,412,872,026	384,858,018	1,464,245,760
DMU 5	EPM	11,497,792	8,997,307	0	0	11,851,157	0	8,712,201
DMU 6	Comfenalco Antioquia	55,003,905	4,500,000	43,468,103	1,489,166	177,047,201	50,863,177	177,934,783
DMU 7	Sura	247,430,381	78,407,570	101,337,602	17,131,377	696,932,272	198,384,339	713,300,668
DMU 8	Famisanar	254,676,457	21,000	176,767,704	8,508,024	669,227,215	124,616,077	666,839,628
DMU 9	Fondo de Pasivo Ferrocarriles	39,370,561	16,297,183	3,739,882	229,932	54,289,171	0	104,784,504
DMU 10	Golden Group	8,873,885	0	1,003,682	686,940	24,376,955	522,175	25,567,417
DMU 11	Nueva EPS	1,058,955,212	3,191,139	725,870,151	2,942,152	1,903,254,182	788,690,698	1,811,061,005
DMU 12	Colpatria	29,626,063	18,584,149	982,376	6,142	44,524,895	6,509,171	43,522,988
DMU 13	Salud Total	365,396,570	198,238,511	75,315,458	35,242,132	787,461,401	109,833,981	817,409,281
DMU 14	Saludvida	68,974,504	4,785,992	7,520,665	7,904,296	42,480,528	6,890,123	29,513,895
DMU 15	Sanitas	379,950,832	25,376,938	242,987,341	3,000,112	589,479,187	227,092,530	582,135,254
DMU 16	S.O.S.	118,171,938	1,319,225	86,320,782	3,688,966	376,769,004	89,411,705	359,358,087
DMU 17	Solsalud	106,869,758	2,613,716	46,573,124	3,457,365	87,208,246	17,644,963	87,656,995

Tabla 3: Modelos seleccionados por el heurístico

	I_1	I_2	I_3	I_4	O_1	O_2	O_3	Indice Gini	DMUs Eficientes	% DMUs Eficientes	Varianza Retenida		
											Entradas	Salidas	Total
$Modelo_1$	✓	✓	✓	✓	✓	✓	✓	0.0778	12	70.6 %	1.0000	1.0000	1.0000
$Modelo_2$	✓	✓		✓	✓	✓	✓	0.1284	9	52.9 %	0.9986	1.0000	0.9997
$Modelo_3$		✓		✓	✓	✓	✓	0.4452	5	29.4 %	0.5249	1.0000	0.9946
$Modelo_4$				✓	✓	✓	✓	0.7732	1	5.9 %	0.4380	1.0000	0.9543
$Modelo_5$				✓	✓	✓		0.7732	1	5.9 %	0.4380	0.9995	0.9509

al FOSYGA (O_2), es decir, ingresos de cuentas por conceptos no incluidos por el plan básico de salud (PBS) pero autorizados por el comité técnico o por fallos de tutelas, y por ultimo los ingresos operacionales (O_3).

4. Resultados

La Tabla 3 se resume los resultados de la aplicación de la metodología propuesta al caso de estudio. Al implementar el heurístico propuesto vemos que para el primer modelo, 12 DMUs fueron clasificadas como eficientes, lo que representa un 70.59 % del total de DMUs, este modelo usa la configuración de variables propuestas por Fontalvo Herrera (2017) (se replican sus mismos resultados). Para los cuatro modelos restantes, que fueron seleccionados por la metodología por tener el mayor índice Gini sobre los puntajes de eficacia en cada iteración, se observa como la cantidad de variables (de entrada o salida) disminuyen al igual que la cantidad de DMUs clasificadas como eficientes. Cabe resaltar que en este caso de estudio, el número de DMUs ubicadas en la frontera se reducen del 71.59 % en el modelo DEA con todas las variables al 5.88 % en el $Modelo_5$, que solo usa el 34 % de las variables iniciales y retiene el 95.09 % de la varianza de los datos.

Cabe resaltar que la DMU_{12} fue clasificada como eficiente en los cinco modelos, posicionándola como un referente dominante del grupo evaluado. Adicionalmente, en el $Modelo_3$ donde se logra explicar un 52.49 % y 100 % de la varianza de las entradas y salidas respectivamente (y en conjunto un 99.46 %), tenemos que la DMU_5, DMU_8, DMU_{10}, DMU_{11}, y DMU_{12} se ubican en la frontera eficiente (véase la Tabla 4).

Tabla 4: Puntajes de eficiencia (CCR-O) de los modelos seleccionados

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
DMU 1	1.0000	1.0000	0.5781	0.1516	0.1516
DMU 2	1.0000	0.7211	0.0985	0.0025	0.0025
DMU 3	1.0000	0.8945	0.1617	0.0240	0.0240
DMU 4	0.9745	0.9745	0.2006	0.0289	0.0289
DMU 5	1.0000	1.0000	1.0000	0.0000	0.0000
DMU 6	1.0000	1.0000	0.1928	0.0322	0.0322
DMU 7	1.0000	0.8912	0.0668	0.0109	0.0109
DMU 8	1.0000	1.0000	1.0000	0.0138	0.0138
DMU 9	1.0000	1.0000	0.5958	0.0000	0.0000
DMU 10	1.0000	1.0000	1.0000	0.0007	0.0007
DMU 11	1.0000	1.0000	1.0000	0.2529	0.2529
DMU 12	1.0000	1.0000	1.0000	1.0000	1.0000
DMU 13	0.8481	0.6915	0.0371	0.0029	0.0029
DMU 14	0.5460	0.1917	0.0136	0.0008	0.0008
DMU 15	0.7410	0.7123	0.3073	0.0714	0.0714
DMU 16	1.0000	1.0000	0.3852	0.0229	0.0229
DMU 17	0.3459	0.2665	0.0560	0.0048	0.0048

Conforme al objetivo de este trabajo, se puede evidenciar como el número de DMUs con una puntuación de eficiencia igual a 100 % disminuye en los últimos modelos sugeridos (véase la Tabla 4).

5. Conclusiones

En este trabajo se ilustra el problema de dimensionalidad presente en DEA ante la presencia de muchas variables de entrada y salida, y pocas DMUs, donde es alta la probabilidad de clasificar una DMU como eficiente cuando no lo es.

Los resultados de este estudio permiten establecer una configuración de variables con la cual se maximiza heurísticamente la entropía o grado de desigualdad de los puntajes de eficiencia (índice Gini). De igual forma, esta metodología brinda un conjunto de modelos con los cuales se garantiza mejorar la discriminación de eficiencia, cada uno de ellos con diferente cantidad de variables.

Con el propósito guiar la selección de variables relevantes para la clasificación de eficiencia, se indica la varianza explicada para los modelos elegidos en cada iteración, lo cual resulta especialmente útil para los tomadores de decisiones, porque les permite dimensionar la varianza retenida por las variables de entradas, de salidas y en conjunto.

La metodología presentada en este artículo se puede complementar con la aleatorización en la selección del modelo en cada iteración, ya que permite diversificar el espectro de solución. Esta noción es muy común en las propuestas heurísticas y se ha evidenciado que tiene buenos resultados en la práctica.

Referencias

Adler, N. & Golany, B. (2002). Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operational Research Society*, *53*(9), 985–991.

Bogetoft, P. & Otto, L. (2020). *Benchmarking with DEA and SFA*. R package version 0.29.

Charles, V., Aparicio, J., & Zhu, J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. *European Journal of Operational Research*, *279*(3), 929–940.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research*, *2*(6), 429–444.

Cook, W. D., Tone, K., & Zhu, J. (2014). Data envelopment analysis: Prior to choosing a model. *Omega*, *44*, 1–4.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). *Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software*, volume 2. Springer.

Dyson, R. G., Allen, R., Camanho, A. S., Podinovski, V. V., Sarrico, C. S., & Shale, E. A. (2001). Pitfalls and protocols in dea. *European Journal of operational research*, *132*(2), 245–259.

Fontalvo Herrera, T. J. (2017). Eficiencia de las entidades prestadoras de salud (eps) en colombia por medio de análisis envolvente de datos. *Ingeniare. Revista chilena de ingeniería*, *25*(4), 681–692.

Jenkins, L. & Anderson, M. (2003). A multivariate statistical approach to reducing the number of variables in data envelopment analysis. *European Journal of Operational Research*, *147*(1), 51–61.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.