



Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Estadística

**Aplicación del ACP al análisis de índices de
eficiencia DEA del sistema bibliotecario**

Autor: César Ramos Matilla

**Tutores: Bonifacio Salvador González
Jesús Alberto Tapia García**

Año 2023

Índice general

1. Introducción	4
2. Análisis envolvente de datos	5
2.1. Unidades productivas	6
2.2. Productividad	6
2.3. Eficiencia relativa	7
2.4. Modelos DEA	8
2.4.1. Modelo BCC y CCR	10
2.4.2. Elección de inputs y outputs	14
2.4.3. Frontera eficiente	15
2.4.4. Elección del modelo	17
3. Análisis de componentes principales	18
3.1. Características generales de la técnica	18
3.2. Análisis normado	19
3.3. Inercia	19
3.4. Retención de componentes	20
3.5. Conceptos fundamentales para la interpretación del ACP	20
3.6. ACP y DEA	21
4. Análisis estadístico de la base de datos de bibliotecas públicas de titularidad estatal españolas	22
4.1. Análisis estadístico descriptivo univariante	22
4.2. Análisis estadístico descriptivo bivariante	25
4.3. Análisis Clúster	31
5. Escenarios y análisis de eficiencia DEA de las bibliotecas públicas de titularidad estatal españolas	33
5.1. Selección de inputs/outputs	33
5.2. Selección del modelo	35
5.3. Eficiencia DEA en los distintos escenarios	36
5.4. Eliminación de la influencia provocada por la elección de inputs/outputs utilizando el análisis de componentes principales	40
6. Conclusiones	51
7. Bibliografía	52
Anexos: I : Software y librerías utilizadas; II: Códigos	55;59

Índice de Figuras

3.1 - 3.10	<i>Histogramas y boxplots de las variables input/output</i>	23
3.11 - 3.20	<i>Diagramas de dispersión entre variables input/output</i>	26
3.21	<i>Diagrama de dispersión entre “Personal” y “UA” con outliers</i>	28
3.22	<i>Clústeres de las bibliotecas</i>	32
5.1	<i>Eficiencias DEA para los distintos escenarios</i>	37
5.2	<i>Mapa de provincias de España</i>	38
5.3	<i>Número de escenarios en los que una biblioteca es eficiente</i>	39
5.4	<i>Tabla de información sobre las inercias recogidas por los autovalores</i>	41
5.5	<i>Histograma inercias</i>	41
5.6	<i>Scree-plot inercias</i>	41
5.7	<i>Autovectores de las 2 primeras componentes principales</i>	42
5.8	<i>Orientación y dirección de crecimiento de las variables</i>	43
5.9	<i>Plot de las variables del ACP</i>	44
5.10	<i>Plot de las bibliotecas en las 2 primeras componentes principales</i>	46
5.11	<i>Representación conjunta de bibliotecas y variables</i>	47
5.12	<i>Contribuciones absolutas/relativas de las bibliotecas a la primera componente principal</i>	48
5.13	<i>Ranking de las eficiencias de las bibliotecas utilizando los resultados de la primera componente principal</i>	49

Resumen

En este trabajo fin de grado se analizarán las eficiencias del conjunto de bibliotecas españolas de titularidad estatal dados unos inputs y outputs, durante el año 2020. Este análisis de la eficiencia se realizará mediante el uso del DEA, análisis envolvente de datos.

Debido a que se medirá la eficiencia en cada centro para todos los escenarios, que son cada una de las posibles combinaciones de inputs y outputs, obteniendo diferentes resultados, se utilizará ACP, análisis de componentes principales, para obtener una medida global de eficiencia, obteniendo un ranking de las bibliotecas y eliminando el efecto que se genera a partir de la elección de unos determinados inputs/outputs.

Se estudiarán e interpretarán todos los resultados que se obtengan de estos análisis previos y se obtendrán las conclusiones de este proyecto.

Abstract

In this dissertation, the efficiencies of the set of Spanish state-owned libraries will be analyzed, given some inputs and outputs, during the year 2020. This efficiency analysis will be carried out using DEA, data envelopment analysis.

Because the efficiency will be measured in each center for all the scenarios, which are each one of the possible combinations of inputs and outputs, obtaining different results, PCA, principal component analysis, will be used to obtain a global measure of efficiency, obtaining a ranking of the libraries and eliminating the effect generated by the choice of certain inputs/outputs.

All the results obtained from these previous analyzes will be studied and interpreted and the conclusions of this project will be obtained.

Capítulo 1

Introducción

El análisis envolvente de datos (DEA) es una técnica no paramétrica que tiene como objetivo evaluar la eficiencia relativa de un conjunto de unidades de decisión (DMUs)

Uno de los principales condicionantes de esta técnica DEA es la elección de variables input (recursos) y output (resultados). Diferentes experimentadores, ante el mismo problema de medir la eficiencia DEA, pueden obtener resultados diferentes por la decisión en la elección de estas variables.

El objetivo de este trabajo fin de grado será utilizar técnicas estadísticas multivariantes que puedan conducir a una solución en la que se elimine la influencia de la elección de la información input/output a la hora de medir la eficiencia.

Para llevar a cabo este objetivo, se considerarán todos los escenarios que se obtienen al combinar de todas las maneras posibles los inputs y outputs previamente seleccionados. Por tanto, para cada unidad de decisión (DMU), se obtienen tantos índices de eficiencia como escenarios existan. Mediante la utilización del análisis de componentes principales (ACP) se trata de reducir el número de medidas de eficiencia a una sola, observando si esta medida única tiene interpretación y resume adecuadamente, desde el punto de vista de la eficiencia, todas las medidas de eficiencia de cada DMU.

Para ilustrar experimentalmente esta técnica, se utilizarán los datos de un conjunto de bibliotecas públicas de titularidad estatal españolas. Estos datos cuentan con gran variedad de medidas para cada uno de los centros, de entre las cuales tomaremos tres inputs y dos outputs. Esta información ha sido tomada de la página web del ministerio de cultura y deporte, la cual publica información sobre las bibliotecas de forma anual (en este caso los datos tomados pertenecen al año 2020) Estas estadísticas son públicas y de acceso gratuito para cualquier ciudadano mediante el siguiente enlace:

<http://mapabpe.mcu.es/mapabpe.cmd?command=GetAnexo&id=63>

Al realizar el estudio utilizando datos reales, se busca garantizar la utilidad práctica del método propuesto.

Se comenzará explicando cómo funciona la técnica no paramétrica DEA, junto con la definición algunos términos de relevancia para la comprensión de esta y la exposición de algunos de los distintos modelos que existen para esta metodología según las necesidades. Después, se comenta brevemente el funcionamiento y los objetivos de la técnica de ACP. Habrá una tercera parte en la que se elabora un estudio estadístico descriptivo de la base de datos de bibliotecas con la que se trabajará, con el fin de describir la información de las variables observadas y detectar posibles puntos de influencia o atípicos. En el último capítulo, se aplican la metodología DEA y el análisis de componentes principales en busca de lograr el objetivo planteado en este trabajo fin de grado.

Capítulo 2: Análisis envolvente de datos

El análisis envolvente de datos (DEA) es una metodología basada en modelos de programación lineal, propuesta por primera vez en 1978 por *Charnes, Cooper y Rhodes* para estudiar la eficiencia relativa de una serie de unidades de decisión.

Se sustenta en las nociones de eficiencia productiva y el cálculo de los índices de eficiencia técnica planteados por *Farrel* en 1957, cuya idea era visualizar la eficiencia desde una perspectiva no ideal, donde cada unidad de producción es evaluada en relación con otras tomadas de un grupo representativo y comparable. Así las medidas de eficiencia serían relativas y no absolutas, donde el valor alcanzado por una determinada unidad productiva correspondería a una expresión de la desviación observada respecto a aquellas consideradas como más eficientes dada la información disponible.¹

El análisis envolvente de datos es una técnica no paramétrica para la medición de la eficiencia relativa de un conjunto de entidades parecidas en situaciones en las que existen varios “inputs” y “outputs”. Estas entidades son las llamadas unidades de toma de decisiones (o DMUs, decision making units), las cuales deben ser comparables, es decir, tanto sus inputs como sus outputs deben ser medibles en unidades homogéneas para todas ellas.²

También podríamos definir este algoritmo como una técnica de programación matemática que partiendo de unas cantidades de “inputs” empleados y de “outputs” producidos por un conjunto de DMUs, nos sirve para determinar cuáles son las mejores opciones (las más eficientes), comparando cada DMU con todas las posibles combinaciones lineales de todas las unidades de la muestra.

El conjunto de DMUs eficientes forma la llamada frontera eficiente o superficie envolvente, midiéndose la eficiencia de cada unidad como la distancia a la misma, de forma que las unidades que forman la frontera serán las unidades eficientes y aquellas que no, las ineficientes. De esta manera se pueden clasificar las distintas DMUs en función de su eficiencia.

Cabe decir que ninguna DMU situada sobre la frontera eficiente puede, dado el nivel de inputs, mejorar uno de sus outputs sin empeorar otro.

Aunque este término se explicará con más detalle posteriormente, entendemos la eficiencia como una medida que pone en relación los medios empleados con los fines obtenidos, considerándose algo eficiente cuando para un determinado número de “inputs” es capaz de producir la máxima cantidad de “outputs”, o si para alcanzar cierto “output” emplea la menor cantidad de “inputs” (*Lovell, 1993*)³

¹ J. Escalona, Lenny. *Eficiencia técnica para las universidades públicas venezolanas a través del modelo DEA*.

² Vargas, Ángela María. *Análisis de la eficiencia de un hospital*. Escuela técnica superior de ingeniería. Sevilla.

³ Faura, Úrsula. *Comparación de rankings de eficiencia mediante ACP y DEA*. INE.

Uno de los principales problemas de la técnica DEA es el hecho de que clasifica una gran cantidad de unidades como eficientes, excepto si la suma del número de inputs y outputs es pequeña en relación con el número de observaciones (Como ocurrirá en nuestro caso, por lo que funcionará bastante bien)

Aun así, para resolver este problema, se han desarrollado diversos métodos para clasificar las DMUs, entre los que se encuentran técnicas multivariantes (correlación canónica, análisis discriminante etc.) y las que emplea la propia DEA (supereficiencia, eficiencia cruzada, etc.)

2.1 Unidades productivas

Dentro del algoritmo DEA las unidades productivas se refieren a las entidades o unidades económicas que se someten a evaluación para medir su eficiencia en la utilización de recursos y la generación de resultados. Estas unidades pueden ser empresas, organizaciones, instituciones, departamentos o cualquier otra entidad que realice actividades de producción o prestación de servicios.

En el análisis DEA, se considera que cada unidad productiva es un proceso de transformación que utiliza una combinación de insumos (inputs) para producir resultados o salidas (outputs). Los insumos pueden incluir recursos como mano de obra, capital, materiales, energía, entre otros, mientras que los resultados pueden ser ingresos, producción, calidad, eficiencia, entre otros indicadores relevantes.

El objetivo del algoritmo DEA es comparar y evaluar la eficiencia relativa de cada unidad productiva en relación con las demás unidades en el conjunto de datos analizado. Para ello, se establece una frontera de eficiencia que define el límite superior de desempeño alcanzado por las unidades más eficientes.

Las unidades productivas que se encuentran en la frontera de eficiencia se consideran eficientes, lo que implica que están utilizando sus recursos de manera óptima para generar resultados. Por otro lado, las unidades que se encuentran por debajo de la frontera eficiente se consideran ineficientes y se espera que realicen mejoras para acercarse a la eficiencia.

Al analizar y comparar las unidades productivas, el algoritmo DEA ofrece una evaluación objetiva de la eficiencia y puede ayudar a identificar áreas de mejora y oportunidades para optimizar el uso de los recursos disponibles.

2.2 Productividad

La productividad de cierta unidad se define como la relación existente entre los resultados que obtiene y los recursos empleados para su producción. Es una manera de cuantificar que tan bien se están utilizando dichos recursos. En el caso de un único “input” y un único “output” sería:

$$Productividad = \frac{Producto\ generado}{Recurso\ consumido} = \frac{Salida}{Entrada}$$

Para el caso general, en el que hay “e” inputs y “s” outputs, siendo x_{ij} la cantidad de recurso “i” utilizado por la unidad “j”, y y_{kj} siendo la cantidad de producto “k” generado por la misma unidad, u_{ij} y v_{kj} son los pesos que eliminan la dimensión de inputs y outputs, se tiene que:

$$Entrada_j = \sum_{i=1}^e u_{ij} x_{ij}$$

$$Salida_j = \sum_{k=1}^s v_{kj} y_{kj}$$

$$Productividad_j = \frac{\sum_{k=1}^s v_{kj} y_{kj}}{\sum_{i=1}^e u_{ij} x_{ij}} = \frac{Salida_j}{Entrada_j}$$

2.3 Eficiencia relativa

La podemos definir como la capacidad que tiene una unidad para obtener la salida máxima a partir de un grupo de inputs. Se consigue al comparar el valor observado de cada unidad con el valor de la unidad de mayor productividad:

$$Eficiencia_j = \frac{Productividad_j}{Productividad_{max}} = \frac{\frac{Salida_j}{Entrada_j}}{\frac{Salida_{max}}{Entrada_{max}}}$$

Siendo “j” el índice de la unidad a estudiar y “max” el de la unidad de máxima productividad.

Se hablará de eficiencia global cuando la unidad que se escoge como referencia con mayor productividad esta entre todas las que se disponen para el estudio.

La eficiencia técnica y la eficiencia relativa no son lo mismo en el contexto del algoritmo DEA.

La eficiencia técnica será cuando para la unidad escogida como referencia con mayor productividad solo se tienen en cuenta las unidades del mismo tamaño que la de la unidad que se quiere calcular.

Por otro lado, la eficiencia relativa, como he dicho antes, se refiere a la comparación de la eficiencia de una unidad productiva con respecto a otras unidades productivas del mismo conjunto. En otras palabras, mide la capacidad de una unidad productiva para producir la misma cantidad de outputs que otras unidades productivas utilizando menos insumos.

En el contexto del algoritmo, la eficiencia técnica se mide mediante el análisis de la distancia entre la frontera de producción (que representa la combinación de insumos y outputs que se consideran eficientes) y los datos de la unidad productiva analizada. Tras determinar la frontera, habrá varias maneras de obtenerla según el número de inputs y outputs que utilicemos. La eficiencia técnica de las DMUs vendrá dada por la desviación respecto a esta.

Si la distancia es cero, la unidad productiva se considera técnica y eficientemente productiva. Si la distancia es mayor que cero, la unidad productiva es técnicamente

2.4 Modelos DEA

La metodología DEA trata de construir una frontera de producción sobre la que se pueda decidir las unidades productivas (DMUs) que son eficientes y, sobre todo cuales no lo son y determinar cuánto pueden aumentar sus resultados con los recursos que se tienen (orientación output) o cuánto se pueden reducir sus recursos para obtener el mismo resultado (orientación input)⁴

Los modelos DEA pueden ser clasificados en función de:

1. Medida de eficiencia que obtenemos de ellos: modelos radiales y no radiales.
2. Orientación del modelo: input orientado, output orientado o input-output orientado.
3. Tipología de los rendimientos a escala que tiene la tecnología de producción, es decir, la forma en que los inputs son combinados para obtener un conjunto de outputs, de tal forma que esa combinación puede tener rendimientos a escala constantes o variables.

⁴ Martín, Pablo. Evaluación de la eficiencia de los jugadores de la NBA utilizando DEA. Universidad de Valladolid.

Medidas de eficiencia DEA

Un modelo radial en DEA es un tipo de perspectiva que utiliza una función de distancia para medir la eficiencia de una unidad productiva en relación con un conjunto de otras unidades productivas. En estos se asume que la eficiencia de una unidad está limitada por su mayor distancia a la frontera de eficiencia, es decir, la unidad más eficiente.

Por otro lado, un modelo no radial en DEA utiliza una función de distancia que no es necesariamente radial. Estos modelos pueden permitir diferentes formas de las funciones de distancia, lo que significa que la eficiencia de una unidad puede estar limitada por diferentes factores, como la distancia promedio a otras unidades o la variabilidad de la eficiencia de otras unidades.

Orientación del modelo DEA

La eficiencia puede ser representada con relación a dos orientaciones básicas, pudiendo hacer referencia a distintos modelos:

1. Input orientado: Buscan, dado el nivel de outputs, la mayor disminución proporcional en el conjunto de inputs mientras permanece en la frontera de posibilidades de producción (FPP es una representación gráfica de las cantidades máximas de producción que puede obtener una “economía”, en un periodo determinado, haciendo uso de todos los recursos que tiene disponibles).

Una unidad no es eficiente si es posible disminuir cualquier entrada sin alterar sus outputs.

2. Output orientado: En este caso, dado el nivel de inputs, buscan el máximo incremento proporcional de los outputs permaneciendo dentro de la frontera de posibilidades de producción.

En este sentido una unidad no puede ser caracterizada como eficiente si es posible incrementar cualquier salida sin incrementar ninguna entrada y sin disminuir ninguna otra salida.

Considerando las orientaciones definidas, una unidad será considerada eficiente únicamente si no es posible incrementar las cantidades de salida manteniendo fijas las cantidades de entrada usadas ni es posible disminuir las cantidades de inputs empleadas sin variar las cantidades de outputs obtenidas.

Tipos de rendimiento a escala

Los rendimientos a escala expresan cómo varía la cantidad producida por una unidad a medida que varía el uso de todos los factores que intervienen en el proceso de producción en la misma proporción.

1. Rendimientos constantes a escala: Cuando el incremento porcentual del Output es igual al incremento porcentual de los recursos productivos.
2. Rendimientos crecientes a escala: Se dice que la tecnología hace alarde de este tipo de rendimientos cuando el incremento porcentual del output es mayor que el incremento porcentual de los factores.
3. Rendimientos decrecientes a escala: Se dan si el incremento porcentual del output es menor que el incremento porcentual de los inputs.

2.4.1 Modelos BCC y CCR

Existen multitud de tipos de modelos de eficiencia DEA que se utilizan en función de los objetivos específicos de evaluación y de la naturaleza de los datos disponibles. Cada modelo tiene sus ventajas y limitaciones y la elección del modelo adecuado depende del contexto y los datos disponibles.

Pese a que hay varios, todos tienen el mismo objetivo: Encontrar un punto admisible de mayor productividad con el que puedan compararse las distintas DMUs del problema. De esta forma, dada cierta DMU_0 , se formula un modelo de programación lineal que busca una combinación lineal de las DMUs existentes, dando lugar a un grupo de puntos admisibles.

En el caso de que ningún punto fuese mejor que DMU_0 , entonces se le denomina unidad eficiente. Que una unidad sea mejor que otra significa que tiene menos inputs y/o más outputs que la unidad considerada. En el caso de que DMU_0 no es eficiente, el modelo la proyecta sobre la frontera eficiente y mide su eficiencia en términos de la disminución del consumo de las outputs totales e incremento en la producción de outputs que se tendría que realizar para que fuese eficiente.

Hay diferentes maneras de realizar la proyección y medición de la distancia entre la DMU_0 y el punto sobre el que se proyecta (como se ha visto anteriormente en el apartado de orientación del modelo)

Dentro de los modelos DEA con rendimientos constantes de escala, el más destacable es:

Modelo CCR (*Charnes, Cooper y Rhodes*)

Es un modelo desarrollado por (Charnes, Cooper, & Rhodes, 1978) que proporciona medidas de eficiencia radiales, Input-Output orientadas y supone una convexidad, fuerte eliminación sin costos de Inputs y Outputs y rendimientos constantes a escala.

Lo escribiremos utilizando la forma multiplicativa, aunque también se puede escribir de forma fraccional y envolvente.

El **modelo CCR Input orientado** en forma de cociente puede ser linealizado siguiendo la transformación lineal de Charnes y Cooper.

Realizando dicho cambio de variable se obtiene el modelo en forma multiplicativa:

$$\text{Max}_{\mu, \delta} \quad W_0 = \mu^T y_0$$

Sujeto a:

$$\delta^T x_0 = 1$$

$$\mu^T Y - \delta^T X \leq 0$$

$$\mu^T, \delta^T \geq I_\varepsilon$$

Donde:

Y es una matriz de outputs (númeroOutputs*n)

y_0 representa el vector de Outputs de la unidad que está siendo evaluada

X es una matriz de inputs (númeroInputs*n)

x_0 representa el vector de Inputs de la unidad que está siendo evaluada

μ es el vector (númeroOutputs*1) de pesos outputs y δ es el vector (númeroInputs*1) de pesos inputs

La unidad será clasificada como eficiente si $w_0^* = 1$ y existe al menos un óptimo ($\mu^* \delta^*$) con $\mu^* > 0$ y $\delta^* > 0$

Modelo BCC (*Banker, Charnes y Cooper*)

Si el modelo CCR consideraba rendimientos constantes a escala, el modelo BCC relaja este supuesto, que en gran parte de las ocasiones resulta excesivamente restrictivo y por tanto irreal, permitiendo que la tipología de rendimiento a escala que en un momento determinado caracterice la tecnología sea variable, esto es: constante, creciente o decreciente.

Los fundamentos del modelo BCC se encuentran en el modelo CCR, puesto que el primero es una extensión del segundo.

Modelo BCC input orientado

Al ser una extensión del anterior, su formulación es similar. El objetivo de este es hallar un hiperplano que, permaneciendo por encima de todas las unidades, minimice la distancia horizontal desde dicho hiperplano a la unidad.

La forma multiplicativa del modelo, obtenida de nuevo al realizar la transformación sobre el modelo fraccional, puede expresarse matricialmente de la siguiente forma:

$$\text{Max}_{\mu, \delta, k} \quad W_0 = \mu^T y_0 + k_0$$

Sujeto a:

$$\delta^T x_0 = 1$$

$$\mu^T Y + k_0 \leq \delta^T X$$

$$\mu^T, \delta^T \geq I_\varepsilon$$

k_0 no restringida

La unidad será clasificada como eficiente si $w_0^* = 1$ y existe al menos un óptimo $(\mu^* \delta^*)$ con $\mu^* > 0$ y $\delta^* > 0$. El valor k_A^* es usado para identificar el tipo de rendimiento a escala en el que opera localmente la unidad evaluada.

También es posible, a partir de los valores óptimos de los pesos o multiplicadores de inputs y outputs, determinar los porcentajes de contribución input/output.

La única diferencia entre la primera formulación y esta es que para el segundo caso se le suma un término constante k_0 (que en el supuesto de rendimientos constantes toma el valor cero)

De tal manera que si la solución óptima de este modelo para la DMU, que se supone eficiente:

- $k_0^* > 0$ para todas las soluciones óptimas, prevalecen rendimientos crecientes a escala.
- $k_0^* = 0$ para cualquier solución óptima, prevalecen rendimientos constantes a escala.
- $k_0^* < 0$ para todas las soluciones óptimas, prevalecen rendimientos decrecientes a escala.

Modelo BCC output orientado

Si, por el contrario, se quiere evaluar la eficiencia de una unidad desde el punto de vista de la maximización de los outputs, dado el nivel de inputs, debe utilizarse este modelo.

Un cambio en la orientación del modelo prácticamente equivale a invertir el cociente entre el output virtual (output total) y el input virtual (input total). Linealizando el modelo BCC output orientado con forma de cociente se obtiene el modelo en forma multiplicativa:

$$\text{Max}_{\mu, \delta, k} \quad W_0 = \delta^T x_0 + k_0$$

Sujeto a:

$$\mu^T y_0 = 1$$

$$\delta^T X + k_0 \geq \mu^T Y$$

$$\mu^T, \delta^T \geq I_\varepsilon$$

k_0 no restringida

La eficiencia técnica Output pura de la unidad 0 vendrá dada por $\frac{1}{w_0^*}$ de tal forma que esta será eficiente si $w_0^* = 1$.

Como sucedía con el modelo anterior, el signo que tome k en la solución óptima indica el tipo de rendimiento a escala para una unidad. Sin embargo, a diferencia de ese, el término constante se encuentra asociado con el valor input y su signo está invertido, de tal forma que:

- $k_0^* > 0$ para todas las soluciones óptimas, prevalecen rendimientos decrecientes a escala.
- $k_0^* = 0$ para cualquier solución óptima, prevalecen rendimientos constantes a escala.
- $k_0^* < 0$ para todas las soluciones óptimas, prevalecen rendimientos crecientes a escala.

2.4.2 Elección de inputs y outputs

En el análisis envolvente de datos la elección de los inputs y outputs puede tener un impacto significativo en los resultados obtenidos. Los inputs son variables que miden los recursos de las unidades de producción, mientras que los outputs son variables que miden los resultados o productos generados por esas unidades.

La selección de los inputs y outputs puede influir en los resultados de DEA debido a dos razones principales:

Sensibilidad a la elección de los inputs y outputs: Diferentes combinaciones de inputs y outputs pueden llevar a diferentes resultados de eficiencia. Esto se debe a que ciertas variables pueden ser más informativas o relevantes para medir la eficiencia que otras.

Entonces podemos decir que la elección tiene un impacto directo en la evaluación de la eficiencia de la unidad de decisión. Si se seleccionan inputs y outputs relevantes y adecuadas, el resultado del algoritmo será más preciso y útil para la toma de decisiones. Por el contrario, si se eligen inputs y outputs incorrectos o irrelevantes, el resultado podría ser inexacto y conducir a ideas equivocadas.

En general, los variables elegidas deben ser representativas de las actividades y objetivos de la unidad de decisión. Deben ser relevantes para el análisis y medibles con precisión. Además, es importante que sean comparables entre las unidades de decisión que se evalúan.

Posible sesgo de selección: La elección de inputs y outputs puede estar influenciada por sesgos o preferencias subjetivas. Por ejemplo, si se seleccionan outputs que favorecen a ciertas unidades de decisión o se excluyen inputs que pueden afectar negativamente a algunas unidades, los resultados de DEA pueden verse distorsionados y sesgados.

2.4.3 Frontera eficiente

En el contexto del algoritmo DEA, la "frontera eficiente" se refiere a un conjunto de unidades o DMUs que alcanzan el nivel máximo de eficiencia en comparación con las demás unidades en el estudio.

El DEA es una técnica utilizada para evaluar la eficiencia relativa de diferentes unidades, que operan con múltiples inputs y outputs. El objetivo del DEA es identificar las unidades que están en la "frontera eficiente", lo que significa que están operando de manera óptima y no pueden mejorar su eficiencia sin disminuir la eficiencia de otras unidades.

La frontera eficiente se representa mediante un conjunto de puntos en un gráfico de dispersión bidimensional o tridimensional, donde cada punto representa una unidad del estudio. Las unidades que se encuentran en la frontera eficiente se consideran referencias de eficiencia y se utilizan como puntos de comparación para evaluar la eficiencia de las demás unidades.

El algoritmo DEA utiliza técnicas matemáticas para determinar la frontera eficiente y calcular las eficiencias relativas de las unidades en relación con esta frontera. Los puntos que se encuentran dentro de la frontera se consideran ineficientes y se puede calcular una medida de eficiencia para cada uno de ellos. El objetivo es que las unidades ineficientes realicen ajustes en sus procesos o recursos para acercarse a la frontera eficiente y mejorar su eficiencia relativa.

Para entender mejor cómo funciona, se va a realizar un ejemplo para un caso sencillo, en el cual solamente se cuenta con un input y un output:

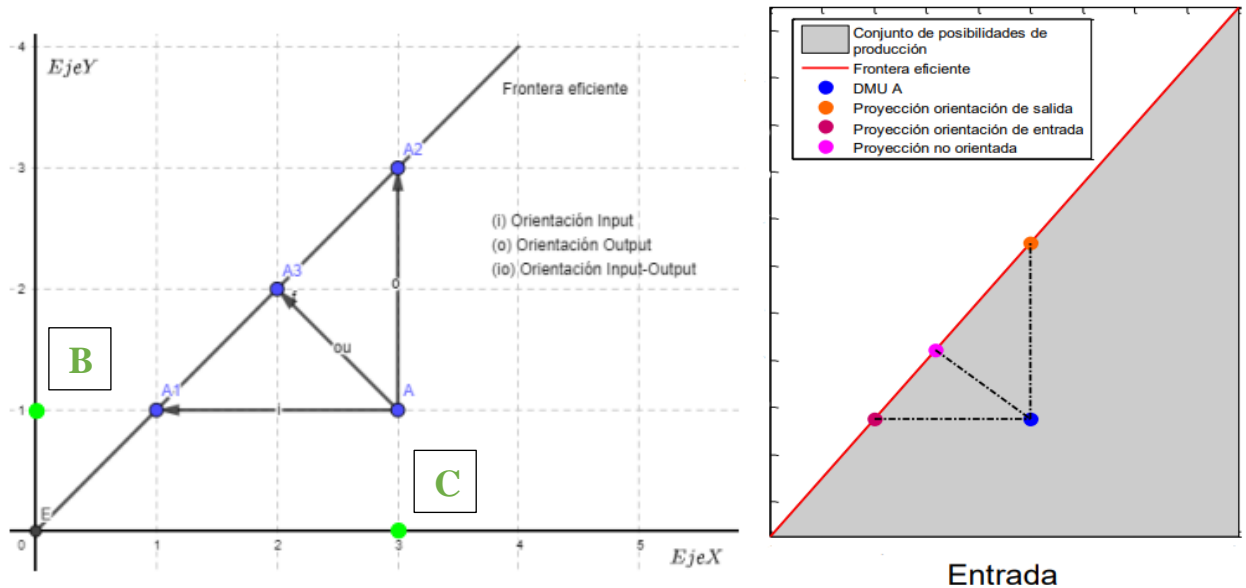


Fig.2.1 Ejemplificación de la obtención de la eficiencia

Fig.2.2 Ejemplo de frontera eficiente 1 output y

Aquí se ha representado, el caso de un único input y output, y en ella puede verse como la unidad A es ineficiente, ya que se sitúa por debajo de la frontera.

Desde el punto de vista del modelo Input orientado, la unidad A podría reducir la cantidad de input x (Los inputs son controlables) y seguir produciendo la misma cantidad de Outputs, es decir, la unidad A debería tomar como referencia la mejor práctica de la unidad A1. La eficiencia (técnica) de la unidad considerada vendría dada por:

$$Eficiencia\ Técnica(DMU\ A) = \frac{distanciaEuclídea(B, A1)}{distanciaEuclídea(B, A)}$$

De igual forma, al considerar la evaluación de la eficiencia a través de modelos Output Orientados (Outputs controlables), la unidad A sería clasificada como ineficiente. Esta unidad podría, consumiendo la misma cantidad de Input, producir una mayor cantidad de Output. En este caso, la eficiencia de la unidad A vendría dada por el cociente:

$$Eficiencia\ Técnica(DMU\ A) = \frac{distanciaEuclídea(C, A)}{distanciaEuclídea(C, A2)}$$

Bajo el supuesto de rendimientos constantes a escala, las medidas de eficiencia técnica coinciden.

Cabe la posibilidad de considerar una tercera opción, correspondiente a los denominados modelos no orientados (también conocidos como Input-Output orientados), en que tanto Inputs como Outputs son controlables, que buscan simultáneamente la reducción Input y expansión Output equiproporcional y que den lugar a medidas de eficiencia “hiperbólica”. En este caso se mide la “distancia hacia la frontera a lo largo de la hipérbola que pasa por el proceso productivo” que representa a la unidad evaluada.

No se nos puede pasar el comentar que la construcción de la frontera eficiente será diferente según se utilicen retornos de escala variable o constante. Aquí tenemos un ejemplo para cinco DMUs:

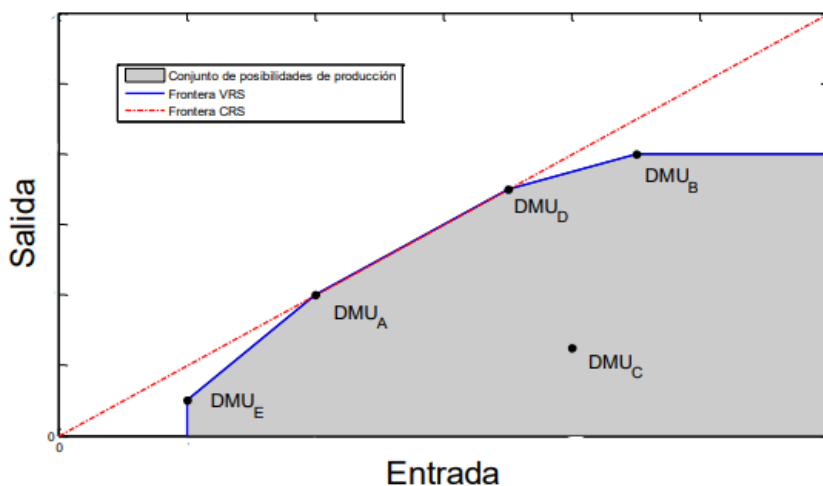


Fig.2.3 Ejemplo de las diferencias entre una frontera hecha con vrs y con crs

2.4.4 Elección de un modelo

En el desarrollo de los modelos se ha podido observar que los dos presentan similitudes, debido a su orientación a medir ratios de eficiencia, con la relación existente entre variables inputs y outputs.

La única diferencia que presenta el modelo BCC es la introducción de restricciones de convexidad. *Coll y Blasco (2006)* comentan que la restricción de convexidad asegura que la unidad combinada es de tamaño similar a las demás unidades y no se trata de una extrapolación de otra unidad combinada que opera en una escala de diferente tamaño. *Boussofiene, Dyson, y Thanassoulis (1991)* y *Pedraja y Salinas (1996)*, indican que el modelo BCC fue planteado con el propósito de estimar la eficiencia puramente técnica eliminando la influencia que pudiera tener la existencia de economías de escala en la evaluación de la ratio de eficiencia de las DMUs. ⁵

Además, el modelo CCR tiene como propiedad principal la proporcionalidad entre inputs y outputs en la frontera; es decir, el aumento (decremento) en la cantidad de los inputs, provocará crecimiento (reducción) proporcional en el valor de los outputs. En el modelo BCC, la DMU que tenga el menor valor de una determinada entrada o el menor valor de una determinada salida será eficiente, a esta DMU llamamos eficiente por defecto o eficiente al arranque (*Charnes, Cooper, Lewin, y Seiford, 1994*).

Generalmente la eficiencia calculada con el modelo BCC siempre será mayor que la calculada con el modelo CCR, ya que las unidades sobre las que se proyectan las unidades DMU_j analizadas son de menor productividad.

⁵ Campoverde, Jorge Arturo. *Evaluación de eficiencia de cooperativas de ahorro y crédito en Ecuador: aplicación del modelo DEA*. Universidad de Cuenca, Ecuador.

Capítulo 3: Análisis de componentes principales

Se trata de un procedimiento multivariante de interdependencia, lo cual quiere decir que no hay variable respuesta. Se quieren estudiar relaciones entre las variables y los individuos presentes en el conjunto de datos, no predecir una variable a partir de otras. El método está diseñado para variables continuas.

El ACP construye una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamaño del conjunto de datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente. Para construir esta transformación lineal debe construirse primero la matriz de covarianzas o matriz de coeficientes de correlación. Debido a la simetría de esta matriz existe una base completa de vectores propios de la misma. La transformación que lleva de las antiguas coordenadas a las coordenadas de la nueva base es precisamente la transformación lineal necesaria para reducir la dimensionalidad de datos. Además, las coordenadas en la nueva base dan la composición en factores subyacentes de los datos iniciales.

El ACP nos permitiría estudiar tanto proximidades entre los individuos como relaciones entre las variables.

3.1 Características generales de la técnica

El objetivo fundamental es reducir la dimensión del conjunto de datos minimizando la pérdida de información. Se construye un nuevo conjunto de variables (componentes) a partir de combinaciones lineales de las variables originales que recojan la mayor cantidad de información posible. Otros objetivos son:⁶

- Extraer la información fundamental contenida en un conjunto de datos.
- Permitir una descripción más sencilla del conjunto de datos (Construcción de índices que permitan “resumir” los datos)
- Analizar la estructura tanto de observaciones como de variables.
- Detección de outliers.

⁶ Fernández, Miguel A. *Análisis de componentes principales* (ANDA). Departamento de Estadística e I.O. Uva.

Disponemos de una matriz de datos (De dimensión: número bibliotecas * número de combinaciones), la matriz \mathbf{X} . Vamos a suponer que **centramos** los datos, es decir, la suma de todos los x_{ij} es 0) Llamaremos a esta matriz centrada \mathbf{X}^* . Si ahora calculamos $(\mathbf{X}^*)'\mathbf{X}^*$ obtenemos, salvo una constante, la matriz de covarianzas:

$$Cov(X) = \frac{1}{n-1} (X^*)'X^*$$

Si **diagonalizamos** la matriz de $Cov(X)$, obtenemos la matriz de autovectores, U , con sus autovalores asociados (*lambdas*)

Entonces tenemos la matriz \mathbf{X}^*U y la matriz de covarianzas de los nuevos datos es $Cov(XU) = Cov(X^*U)$

Ahora podríamos obtener un gráfico de las proyecciones sobre las dos primeras componentes (que son los que más información recogen) y una tabla con las inercias recogidas por cada componente.

3.2 Análisis normado

A partir de lo observado en el apartado anterior nos damos cuenta de que sería más razonable hacer un análisis de forma que lo que descubramos no sean las diferencias entre las varianzas de las variables.

Esto se resuelve haciendo un análisis normado que es lo habitual en la práctica. Lo que se hace es no solo centrar las variables sino **estandarizarlas** para que todas tengan también igual varianza.

Notamos que esto tiene la ventaja adicional de permitirnos combinar sin problemas variables medidas en unidades no compatibles ya que las variables estandarizadas no tienen unidades.

3.3 Inercia

Es una medida de la dispersión de la nube de puntos alrededor de un punto denominado centro de gravedad. Este punto en un ACP normado corresponde al vector nulo. Cuanto más lejos queden las proyecciones entre sí, mayor será la inercia y cuanto más inercia exista a la hora de proyectar menor será la pérdida de información.

Por esta razón, en la construcción del ACP, este se plantea como un problema de optimización en el que se busca un espacio de dimensión menor sobre el cual proyectar y en el que la inercia sea máxima.

3.4 Componentes a retener

Hay varios criterios para tomar una decisión en este sentido. La decisión puede depender también del uso que se vaya a hacer del resultado. Los más habituales son:

- Considerar el porcentaje de inercia acumulado y decidir retener un número de componentes s que haga que la inercia explica supere un determinado porcentaje.

$$100 \frac{\sum_{i=1}^s \lambda_i}{\sum_{i=1}^k \lambda_i}$$

- Extraer aquellas componentes cuya inercia explicada supere el promedio de los autovalores (En el caso del análisis normativo habitual, que es el que nos concierne, esto es equivalente a tomar las que tienen un valor superior a 1)
- Construir la denominada representación *scree plot*, en la que se representa en el eje de ordenadas el número del autovalor y en el de abscisas su valor. Buscaremos el codo del gráfico, es decir, el momento en el que comienza a descender más lentamente.

3.5 Conceptos fundamentales para la interpretación del ACP

A la hora de interpretar las proyecciones de los individuos sobre las componentes se tiene en cuenta:

- Distancia del individuo al origen: El origen es el centro de gravedad de la nube de individuos proyectados. De este modo, las bibliotecas más cerca del origen serán las observaciones más típicas.
- Contribuciones absolutas (a la inercia explicada por cada eje): Nos dice lo que ha contribuido cada biblioteca a la definición de un eje. Si hay puntos con una contribución absoluta mucho mayores que el resto, puede dudarse de la estabilidad del eje (ya que estaría excesivamente condicionado por estos) Para estos casos, realizaríamos el análisis si ellos para ver si los ejes se mantienen.

$$\text{contribucion absoluta}(\text{individuo}, \text{eje}) = \frac{(x'_{\text{individuo}} u_{\text{eje}})^2}{(n-1)\lambda_{\text{eje}}}$$

Recordamos que: $\sum_{i=1}^n \frac{(x'_i u_{\text{eje}})^2}{n-1} = \lambda_{\text{eje}}$

- Contribuciones relativas (Cosenos cuadrados): Nos indican como de cerca o lejos está cada punto de cada eje. Sirven para saber si un punto está mejor o peor representado en un eje o conjunto de ellos.

$$\begin{aligned} \text{contribucion relativa}(\text{individuo}, \text{eje}) &= \cos^2(\text{individuo}, \text{eje}) \\ &= \frac{(x'_{\text{individuo}} u_{\text{eje}})^2}{d^2(\text{individuo}, G)} \end{aligned}$$

Notar que ahora: $\sum_{\text{eje}=1}^k \cos^2(\text{individuo}, \text{eje}) = 1$ ⁷

3.6 ACP en DEA

Para mitigar la influencia de la elección de inputs y outputs al evaluar la eficiencia DEA, se puede utilizar el análisis de componentes principales de la siguiente manera:

Construir una base de datos en la que, con el modelo DEA elegido, se mida la eficiencia en todos los escenarios a los que la combinación de inputs/outputs da lugar.

ACP es una técnica estadística que permite reducir la dimensionalidad de los datos identificando las variables más relevantes. Al aplicarla a los datos de eficiencia, se obtienen los componentes principales que explican la mayor parte de la variabilidad en los datos originales.

El ACP permite seleccionar los inputs y outputs basándose en la importancia relativa de las componentes principales. Aquellas variables de los escenarios que tengan una mayor carga o contribución a las primeras componentes serán consideradas más importantes a la hora de medir la eficiencia.

⁷ Amat, Joaquín. *ACP (Principal component analysis, PCA) y t-SNE*. Ciencia de datos, 2017.

Capítulo 4: Análisis estadístico de la base de datos de bibliotecas públicas de titularidad estatal españolas

En este apartado se utilizarán técnicas estadísticas descriptivas univariantes y multivariantes para analizar la base de datos.

Las DMUs con las que se va a trabajar son las 53 de bibliotecas públicas de titularidad estatal que hay en España durante 2020.

Los motivos de haber seleccionado estas es variado, pero la razón principal es debido a su papel en la promoción de la cultura, la educación, la investigación y el acceso a la información. Además, se han seleccionado de esta categoría ya que cuentan con una gran variedad de artículos y obras en su catálogo de gran importancia y a las que puede acceder cualquier ciudadano español de forma gratuita.

4.1 Análisis estadístico descriptivo univariante

El **análisis descriptivo univariante** de una variable numérica es una técnica estadística que se utiliza para explorar y resumir las características y propiedades de una variable numérica en particular.

En este tipo de análisis, se estudia una sola variable a la vez, sin considerar la relación con otras variables. Este estudio incluye el cálculo de diversas medidas estadísticas, que permiten resumir la distribución de los valores de la variable y conocer la tendencia central, la variabilidad y la forma de la distribución.

También puede incluir la creación de gráficos para visualizar los datos, que pueden ayudar a identificar patrones o detectar valores atípicos entre otras cosas. A continuación, muestro los histogramas y boxplots para las cinco variables que estamos utilizando en nuestro proyecto.

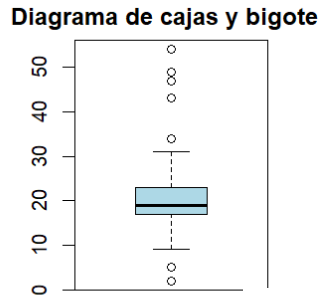
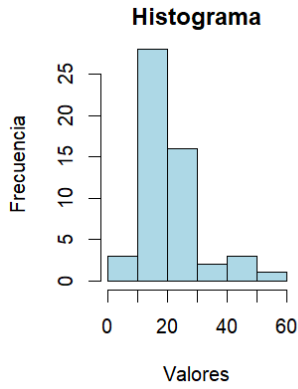


Fig.3.1 y Fig.3.2 Histograma y boxplot variable "Personal"

Fig.3.3 y Fig.3.4 Histograma y boxplot variable "UA"

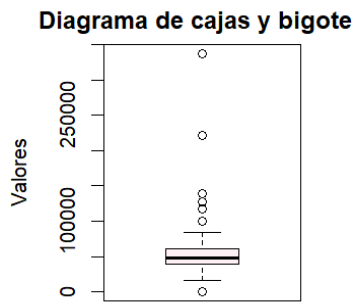
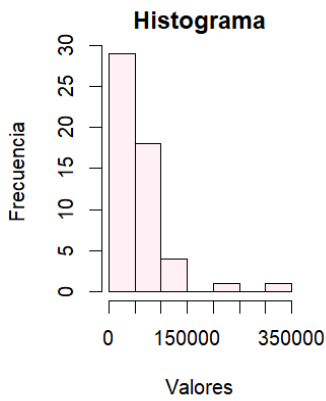
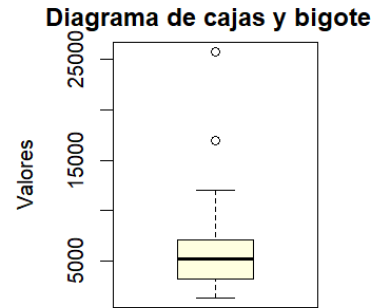
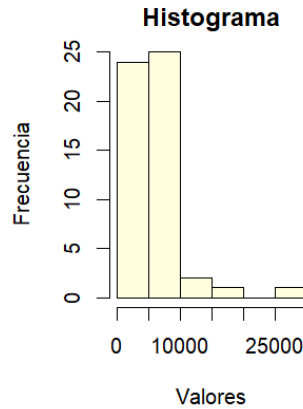


Fig.3.5 y Fig.3.6 Histograma y boxplot variable "Gastos"

Fig.3.7 y Fig.3.8 Histograma y boxplot variable "NU"

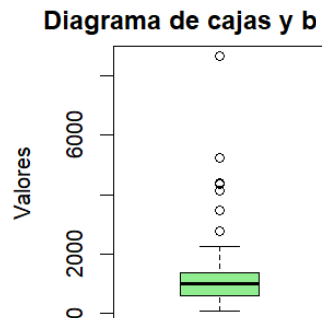
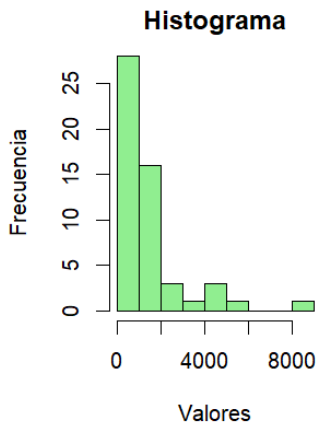
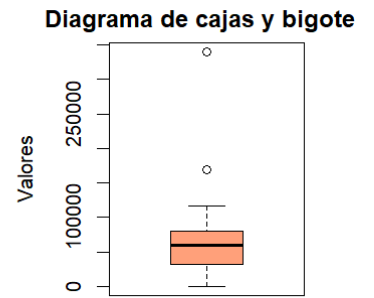
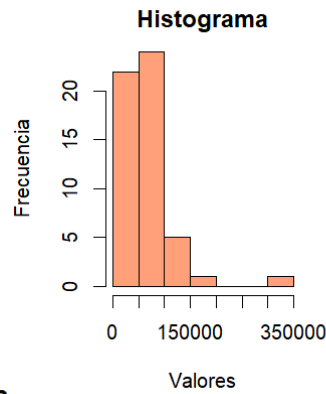


Fig.3.9 y Fig.3.10 Histograma y boxplot variable "Préstamos"

Personal: Observamos que la biblioteca con menor personal es la de Orihuela (municipio de la provincia de Alicante, Comunidad Valenciana) que cuenta únicamente con dos trabajadores, mientras que la que más empleados tiene es la de Murcia con cincuenta y cuatro. La media de personal en una biblioteca de estas características es de unas 22 personas, mientras que su desviación típica, medida que se utiliza para cuantificar la variación o la dispersión de un conjunto de datos numéricos, es de 9.6.

Otras medidas pueden ser, por ejemplo, la mediana, que es el valor que ocupa el lugar central de todos los datos cuando éstos están ordenados de menor a mayor. En este caso es 19. Nos sirve para medir la tendencia central, que es la ubicación del centro de un grupo de números en una distribución estadística.

Podemos observar de manera visual lo dicho anteriormente. Llama la atención que no parece haber mucha la variabilidad en los datos y también la cantidad de outliers que aparecen; Esto se debe a que hay determinadas bibliotecas muy grandes en comparación con las demás, y que, por tanto, para atender a un mayor número de personas y material, necesitarán disponer de un elevado número de empleados (Ocurre al contrario para las más pequeñas)

Usuarios activos: La cantidad de personas que participan en las bibliotecas son 5825 de media, la que mayor cantidad tiene es, de nuevo, la de Murcia con la exagerada cantidad de 25732 personas (Y digo exagerada que ya difiere mucho de la media), mientras que la que menos también vuelve a ser la de Orihuela. Esto tiene bastante sentido, y lo analizaremos en el análisis bivariante entre las variables usuarios activos y personal. La desviación típica es de 3981.658 usuarios, lo cual, como veremos gráficamente, es una variabilidad no demasiado alta.

La mediana toma el valor, como se observa claramente en el histograma o en el diagrama de cajas 5235, el cual, tiene un valor similar al de la media.

Gastos: Respecto al dinero utilizado para la adquisición bibliográfica, podemos observar que ronda los 61458 euros anuales, aunque puede diferir entre unas bibliotecas y otras con una desviación típica de 52163.94 euros. Vemos que, hasta ahora, esta es la variable que menor variabilidad. También podemos fijarnos que aparecen diversos outliers, los cuales existen para valores muy elevados, los cuales se corresponderán a las bibliotecas más grandes (Pertenecientes a ciudades importantes), las cuales cuentan con más recursos en comparación con las demás.

Nuevos usuarios: Las bibliotecas que más personas ha conseguido atraer para utilizar sus servicios son las de Cáceres, Tarragona, Las Palmas de Gran Canaria y Huelva, siendo Cáceres la que más con notable diferencia (Casi el doble respecto a las otras tres que más). Por el contrario, las que menos personas nuevas han adquirido este año han sido Madrid y Teruel.

Esto podría deberse a que ya son centros asentados y conocidos y a los cuales ya mucha gente de la zona esta “afiliada” y por lo tanto su crecimiento de nuevos usuarios es menor. La media de nuevas personas durante este año en este tipo de bibliotecas es de 1400 usuarios.

Prestamos: El centro que más material han prestado a lo largo del año de estudio es Murcia, con bastante diferencia respecto a los siguientes, ya que cuenta con la elevada cantidad de 336681 préstamos, mientras que la siguiente con mayor cantidad, Logroño, cuenta con 169786. Después de estas, hay un pequeño grupo de grandes bibliotecas que tienen alrededor de 100000 préstamos anuales, entre las que están Valladolid, Santander o A-coruña entre otras. En cambio, la que menor tasa tiene de esta variable es Madrid, con un sorprendente número de 186 préstamos únicamente pese a que sus otras variables toman valores bastante elevados.

Pese a que cuenta con dos outliers (Que serán como he dicho, Murcia y Logroño), esta variable es la que menor diferencia promedio tiene para cada centro, es decir, en general todas toman un valor dentro de unos márgenes similares.

En general, observamos que la mayoría de las bibliotecas siguen una misma dinámica, y esto se deberá, como comprobaremos a continuación, a la relación existente entre sus variables de estudio.

4.2 Análisis estadístico descriptivo bivalente

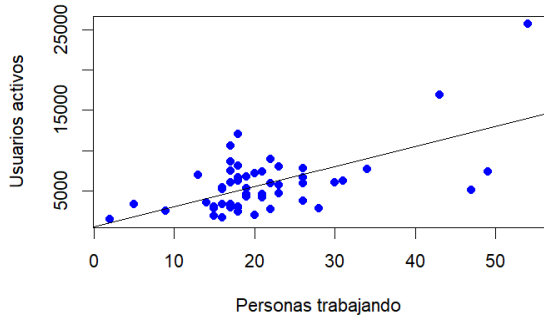
En este apartado se explorará y resumirá la relación entre las distintas variables numéricas. Especialmente se tendrá en cuenta la relación entre las variables de entrada “personal” y “usuarios activos” ya que gracias a los resultados obtenidos se llega a la conclusión de que, para poder hacer un estudio de la eficiencia correcto, se deberá tener en cuenta la influencia de ciertas bibliotecas con valores atípicos para estas variables ya que a la hora de homogeneizar las DMUs por tamaño se van a utilizar ambos campos como medidas.

En este tipo de análisis, se estudiará la asociación entre dos variables a la vez, en lugar de estudiarlas por separado. También se incluye a continuación la creación de gráficos de dispersión, que muestran la distribución conjunta de las dos variables.

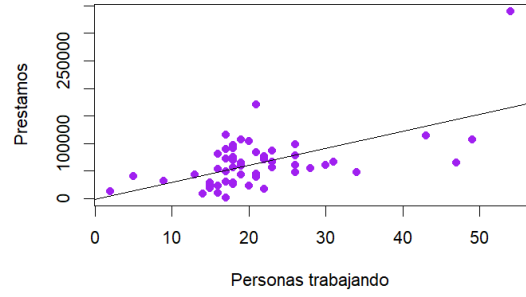
Además, se calcularán diversas medidas estadísticas, como la correlación y la covarianza, para cuantificar la relación entre las dos variables. En la primera combinación se va a explicar el procedimiento que posteriormente se aplicara en las demás combinaciones de forma más extensa.

Los siguientes 10 gráficos son los diagramas de dispersión para todas las combinaciones de variables Fig.3.11, Fig.3.12, Fig.3.13, Fig.3.14, Fig.3.15, Fig.3.16, Fig.3.17, Fig.3.18, Fig.3.19, Fig.3.20

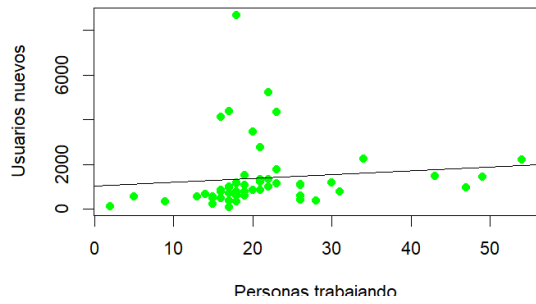
Relación entre Personal y UA



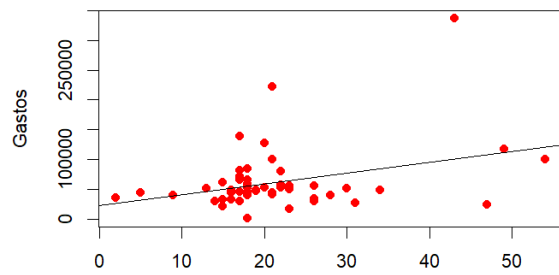
Relación entre Personal y Prestamos



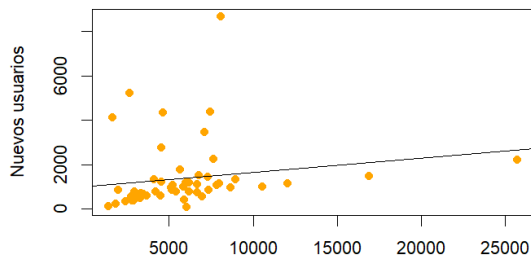
Relación entre Personal y NU



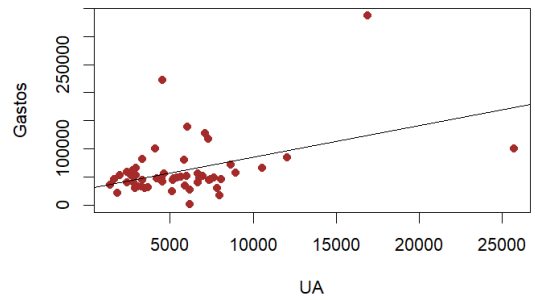
Relación entre Personal y Gastos



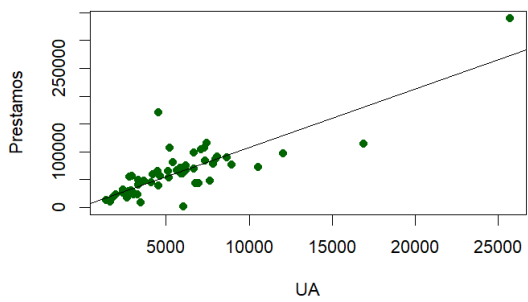
Relación UA y NU



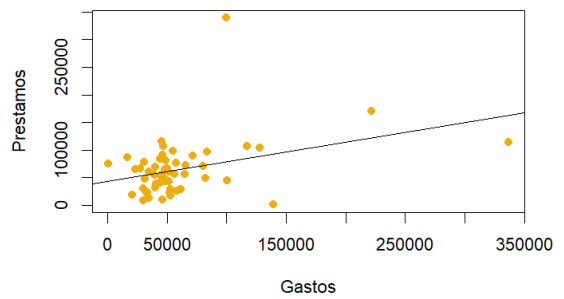
Relación entre UA y Gastos



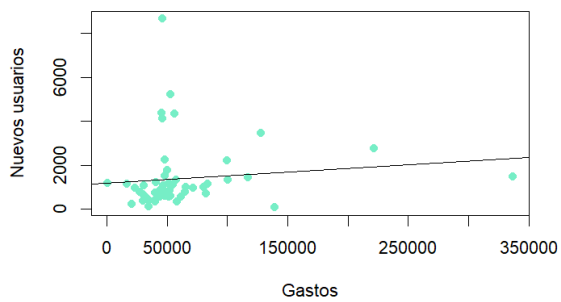
UA y prestamos



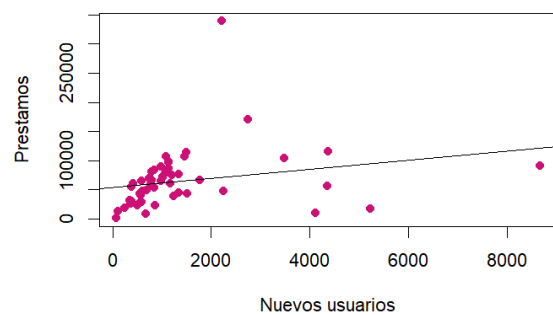
Gastos y Prestamos



Relación entre Gastos y NU



Relación entre Nuevos usuarios y Prestamos



Personal - Usuarios activos: Lo primero que se realizará será un diagrama de dispersión, el cual es un tipo de gráfico utilizado para visualizar la relación entre dos variables numéricas continuas. En este tipo de gráfico, cada punto representa una observación o medida en el eje horizontal y el eje vertical, lo que permite analizar si existe alguna relación entre ambas variables.

En este caso parece haber una relación lineal positiva bastante clara. Aunque no tiene mucha fuerza. Esto último quiere decir que los puntos no están situados demasiado cerca del patrón lineal que describe la forma de la relación.

Se comprobará la existencia de relación mediante el coeficiente de correlación de Pearson, el cual es una medida de dependencia lineal entre dos variables aleatorias cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Para este caso, como se está aplicando sobre los datos de una muestra, tomará la forma:

$$r_{X,Y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Donde n es el tamaño de la muestra (en este caso las 53 bibliotecas) y x_i e y_i son puntos muestrales individuales indexados con i.

El valor del índice de correlación varía en el intervalo [-1, 1], indicando el signo el sentido de la relación:

- Si r es 1, existe una correlación positiva perfecta.
- Si r es -1, existe una correlación negativa perfecta.
- Cuanto más cerca de 0 esté, menos relación habrá, siendo r igual a 0 la falta de asociación entre las variables (independencia)

Se considerará que existe una relación más o menos significativa a partir de 0.3 (Tomando en valor absoluto el resultado del coeficiente)

Para este primer caso toma un valor de 0.6, el cual es bastante elevado, por lo que teniendo en cuenta esto y lo visto gráficamente antes, se puede afirmar que existe una relación lineal entre ambas variables.

Se realizará un nuevo diagrama de dispersión en el que se verán e indicarán aquellas bibliotecas que tienen valores anormales para alguna de las variables mencionadas:

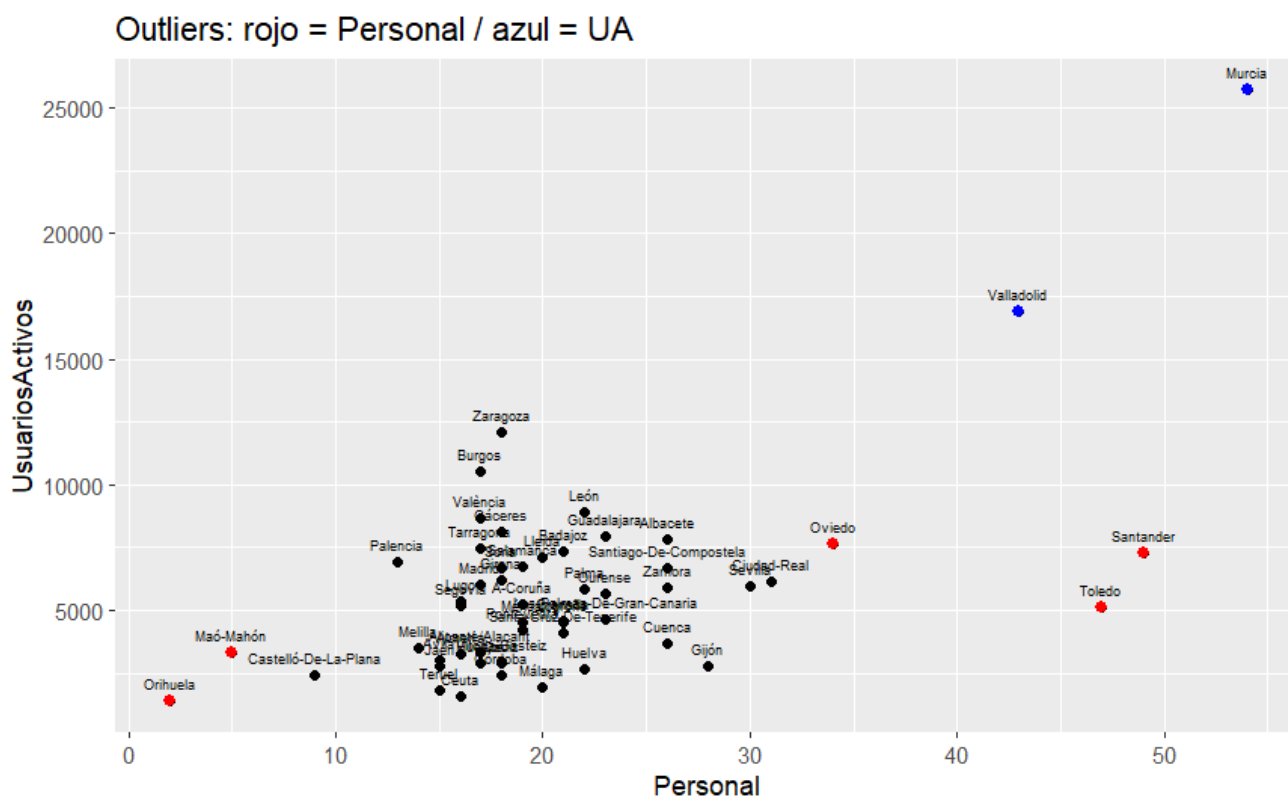


Fig.3.21 Diagrama de dispersión de las bibliotecas para las variables “Personal” y “UA” en el que se indican los outliers para cada uno de los ejes

Se aprecia en el gráfico que las bibliotecas con outliers en Personal son:

Orihuela: 2, Maó-Mahón: 5, Oviedo: 34, Santander: 49, Toledo: 47, Murcia: 54 y Valladolid: 43

Y las que tienen outliers en “usuarios activos” son:

Valladolid: 16930 y Murcia: 25732

Estos valores corresponden con los outliers en el análisis univariante de las dos medidas.

Personal – Gastos: En el gráfico se observa que la dirección de crecimiento es hacia arriba a la derecha, es decir, positiva (aunque no es muy clara) Se puede decir que la relación es débil ya que el conjunto de puntos no está cerca de una línea de tendencia y en cuanto a la forma, podríamos decir que parece ser lineal. Llama la atención la existencia de un dato atípico el cual tiene un valor demasiado alto de gastos en relación con el número de personas trabajando respecto al resto de bibliotecas (Siendo este punto la de Valladolid)

El coeficiente de correlación toma un valor de 0.334, valor para el cual ya se puede pensar que existirá cierta relación, la cual consiste en que cuanto más personal haya trabajado, mayor será el gasto en adquisición bibliográfica que hace la respectiva biblioteca. Esto tiene sentido, ya que se supone que cuanto más material se tenga, más personal hará falta para cuidarlo y gestionarlo. Cabe mencionar que dicha relación no será demasiado importante a la hora de estudiar las variables en conjunto.

Personal - Nuevos usuarios: En este caso, no es posible identificar la tendencia de la posible relación entre las variables, ya que aunque podría parecer lineal, cuando la variable personal toma valores de entre 20 y 30, aparecen dos caminos; uno en el que los usuarios activos crecen de distintas maneras para un mismo valor de la variable de trabajadores y otro que sigue la dinámica del principio y parece tener una forma lineal positiva (se dice positiva, aunque prácticamente es horizontal, es decir, el número de nuevos usuarios crece muy poco a poco en relación con la cantidad de personal de una biblioteca)

En esta ocasión, el coeficiente es 0.1095022, por lo que se descarta que haya una asociación lineal significativa entre ambas variables.

Personal – Prestamos: Se observa que esta dependencia va a ser una de las mayores de entre todas las combinaciones de variables que tenemos. La relación es lineal, positiva y con bastante fuerza. Esto quiere decir que la variable préstamos va a crecer de manera más o menos constante a la vez que crece el número de trabajadores en el centro. Esto tiene mucha coherencia ya que cuanto mayor sea el número de préstamos, más gente será necesaria para llevar a cabo esta tarea.

En este caso la correlación lineal alcanza casi el valor de 0.6, cifra bastante elevada que confirmará lo dicho anteriormente y que afianzará la idea de que existe una relación lineal importante entre estas variables.

Usuarios activos – Gastos: Tras la relación anterior, esta es otra de las mayores. Claramente es una relación creciente (positiva), con tendencia lineal y bastante fuerte. Fijándonos en el gráfico, llama la atención como se agrupan casi todos los datos entre 0 y 5000 usuarios activos, aunque esto engaña, ya que la salida de R tiende a representar todos los datos, por lo que, debido a dos bibliotecas que tienen valores atípicos en comparación con el resto de las bibliotecas, la dispersión queda así (Estas son la de Valladolid, Murcia y Logroño como se vio en el análisis univariante)

Esta relación cobra sentido ya a que, como es lógico, un mayor número de usuarios que acceden a los servicios de la biblioteca implica que los gastos de esta en adquisición de material para poder cumplir las demandas y expectativas de la gente deban crecer proporcionalmente.

Todo lo dicho anteriormente es respaldado por el valor del coeficiente de Pearson, el cual, en este caso toma un valor superior a 0.42.

Usuarios activos - Nuevos usuarios: Para esta combinación de variables, se ve que los datos están muy dispersos y que no parecen seguir ninguna tendencia en especial, por lo que no parece haber relación lineal. El valor del coeficiente de correlación también toma un valor bastante pequeño. No parece influir la cantidad de usuarios que ya están en una biblioteca con los que se “apuntan” nuevos.

Usuarios activos – Prestamos: Sin ninguna duda, esta es la relación lineal positiva más fuerte que aparecerá en todo el estudio. Gracias a lo observado en el gráfico y utilizando de apoyo también en el resultado del coeficiente de correlación, que toma el valor de 0.82, se puede afirmar que existe una relación lineal muy estrecha entre el número de usuarios en activo que tiene una biblioteca y el número de préstamos que se realizan. Este es un resultado que cabría esperar ya que cuantas más personas participan o forman parte de un centro, mayor será el número de préstamos que se pidan en este.

Gastos - Nuevos usuarios: Aunque podría parecer que, a mayor cantidad de nuevos usuarios, más se tendría que invertir en comprar material nuevo, esto no ocurre, de hecho, estas son de las variables que menos relación parecen tener, de ahí, el valor tan pequeño que toma la correlación lineal muestral. Hasta el gráfico de dispersión es ligeramente similar al plot nulo (que representa la ausencia de asociación lineal)

Gastos – Prestamos: Estas variables sí que parecen tener cierta relación lineal, aunque no se puede afirmar ya que, pese a que el coeficiente de correlación toma un valor de 0.364, gráficamente vemos que la mayoría de los datos se encuentran agrupados en forma de “burbuja” en una zona y que existen bastantes puntos atípicos, por lo que podría creerse que hay una ausencia de relación entre estas variables. Los outliers, en este caso, corresponden a las bibliotecas de Murcia, que tienen una cantidad de préstamos muy elevada pese a que tienen unos gastos de adquisición bastante en la media de las demás bibliotecas y la de Madrid, que cuenta con un ínfimo número de préstamos pese a que, al igual que antes, sus gastos de adquisición son bastante “normales”.

Nuevos usuarios – Prestamos: En este caso, podría parecer que hay cierta relación lineal positiva muy fuerte al principio, pero debido a la enorme cantidad de outliers que aparecen (entre el que destaca la biblioteca de Murcia nuevamente, que cuenta con una cantidad de nuevos usuarios similar a los demás centros, pero que tiene un elevadísimo número de préstamos) y el bajo valor que toma el coeficiente de Pearson, entre otras cosas, debido a estos, no se puede afirmar la existencia de relación lineal entre los nuevos usuarios y los préstamos (En verdad los nuevos usuarios no tienen por qué afectar notablemente al incremento de préstamos, ya que al final, la variable que sí influirá sobre los préstamos y como se ha visto anteriormente, es la cantidad de usuarios activos total, que en proporción a los nuevos es una cantidad significativamente mayor)

Atendiendo a los resultados anteriores podemos decir que están más relacionadas entre sí son:

- Personal y usuarios activos.
- Personal y préstamos.
- Usuarios activos y gastos.
- Usuarios activos y préstamos.

4.3 Análisis clúster

La homogeneidad de las DMUs en DEA implica que todas las unidades bajo estudio deben ser comparables. Por ejemplo, una biblioteca de un municipio pequeño donde casi no hay personal que la atienda o en la que apenas pueda haber nuevos usuarios es difícilmente comparable con una biblioteca municipal de una capital. Si las DMUs no son homogéneas se vuelve difícil realizar una comparación justa y precisa. Uno de los métodos para poder filtrar DMUs es mediante un análisis clúster, que se encarga de agrupar bibliotecas según qué tan parecidas sean utilizando los valores que toman sus variables.

Existen varios motivos por los cuales se busca la homogeneidad de las DMU en DEA:

Comparabilidad: La homogeneidad permite que las unidades sean comparables entre sí. Al utilizar los mismos insumos y producir los mismos resultados, se puede medir y comparar la eficiencia relativa de las DMU de manera más precisa.

Benchmarking: Al establecer una base de referencia común, las DMU pueden identificar sus fortalezas y debilidades en relación con otras unidades similares. Esto facilita el proceso de benchmarking y proporciona información valiosa para mejorar la eficiencia y el desempeño de las DMU.

Análisis de mejores prácticas: Al tener unidades homogéneas, es posible identificar y analizar las mejores prácticas utilizadas por las DMUs más eficientes. Esto permite a las unidades menos eficientes aprender de las estrategias exitosas y adoptar medidas correctivas para mejorar su rendimiento.

Eliminación de variables no controlables: Al mantener la homogeneidad de las DMU, se pueden eliminar las diferencias atribuibles a factores externos o no controlables. Esto permite evaluar y comparar la eficiencia relativa de las unidades basándose principalmente en su gestión interna y decisiones operativas.

Es importante destacar que la homogeneidad en DEA es una suposición simplificada para facilitar el análisis comparativo. En la práctica, es posible que las DMU difieran en ciertos aspectos, pero se busca minimizar estas diferencias y enfocarse en los factores que son controlables y relevantes para la eficiencia operativa.⁸

Tras realizar el análisis clúster con R se obtienen el gráfico Fig.3.22 en el que se muestran los principales grupos de bibliotecas según los valores que toman sus variables de estudio.

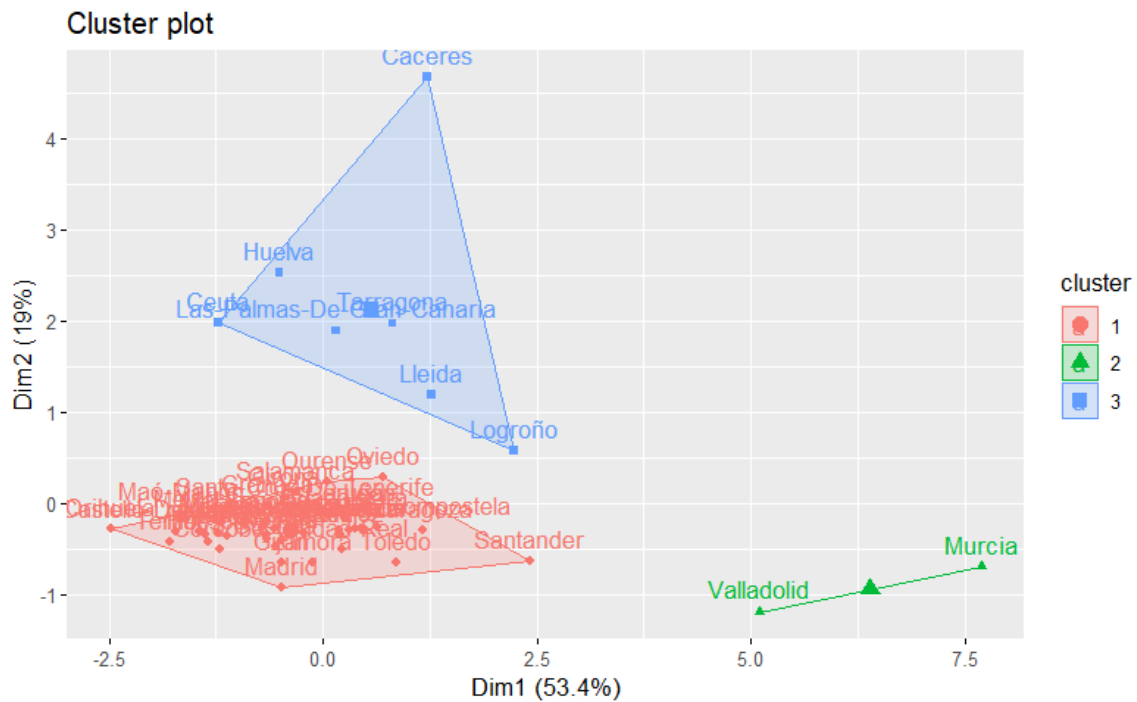


Fig.3.22 Salidas Análisis Clúster

Se han utilizado diversos métodos para hallar el número de clústeres necesarios (“wss”, “silhouette”...) y en la mayoría se indicaban que eran necesarios tres grupos.

Se observan dos grupos que engloban muchas bibliotecas y uno con muy pocas. El rojo, especialmente agrupa la mayor parte de todas (83%), siendo este el que engloba las que toman valores “normales” y similares en sus variables. El grupo azul cuenta con 7 centros (13%) los cuales ya se distancian un poco de la norma general tomando valores atípicos en algunos de sus inputs/outputs (En este caso para un elevado valor de “nuevos usuarios” y un valor bajo de “préstamos”, ambas outputs, ya que las proporciones de sus inputs son similares a las del grupo anterior) En tercer lugar se encuentra el grupo verde, que está bastante alejado de los otros dos, y es que estas dos bibliotecas (Valladolid y Murcia) toman valores muy altos para todas sus variables.

⁸ Romero, L. Ramirez, M. *Análisis cluster (R)*. RPubs.

La razón por la cual se decide no eliminar outliers como Valladolid o Murcia es debido a la pérdida de información que esto podría generar ya que pueden contener información valiosa sobre las variaciones extremas en los datos y su eliminación puede conducir a una pérdida de información importante. Además de lo anterior, estas observaciones atípicas pueden revelar ineficiencias o patrones inesperados en el conjunto de datos, que pueden ser útiles para comprender mejor el rendimiento y la eficiencia de las unidades analizadas. También hay que tener en cuenta que la eliminación de outliers puede tener un impacto significativo en los resultados del análisis de eficiencia DEA ya que los outliers a menudo influyen en los puntajes de eficiencia al alterar la frontera eficiente. Otra razón por la que se deben seguir utilizando es debido a que la supresión de estos puntos puede provocar un sesgo de selección, es decir, su eliminación puede introducir sesgos en el análisis debido a la selección subjetiva de qué valores considerar como outliers, dificultando la comparabilidad y la interpretación de los resultados.

En resumen, se deberían utilizar los outliers en un análisis de eficiencia DEA y considerar su impacto en los resultados en lugar de eliminarlos de manera indiscriminada, y realizar un análisis más detallado de estos valores para comprender su origen y su efecto en la eficiencia general del conjunto de datos.

Capítulo 5: Escenarios y análisis de la eficiencia DEA en las bibliotecas públicas de titularidad estatal españolas

En este último apartado se tratarán de alcanzar los objetivos del TFG, es decir, la obtención e interpretación de las eficiencias y el estudio de posterior de estas tras aplicarle el ACP, observando si se logra eliminar el efecto de la elección de la información input y output.

En primer lugar, se van a explicar las razones por las cuales se han seleccionado las cinco variables que se van a utilizar para medir la eficiencia DEA de las bibliotecas, y también se va a exponer el modelo DEA a utilizar.

A continuación, se estudiará detalladamente la tabla que contiene las eficiencias para cada centro en los diversos **escenarios**, los cuales son todas las combinaciones posibles de inputs/outputs que se pueden utilizar para obtener las eficiencias mediante la técnica DEA (Cabe mencionar que como mínimo es necesario al menos 1 input y 1 output), ya que se busca observar una posible variación del comportamiento en términos de eficiencia de las bibliotecas producido por la elección de unas u otras variables.

En último lugar se tratará de resumir toda la información que aporta la tabla anterior en una sola componente mediante la aplicación del ACP, obteniendo así una medida fácilmente interpretable y que asegure un resultado fiable acerca de la posible eficiencia o falta de ella en una determinada biblioteca.

5.1 Selección de inputs/outputs

Para la **elección de los Inputs** se debe elegir los recursos que necesita o se invierten en una biblioteca a lo largo de un año regular para producir el mayor número posible de outputs.

Se van a considerar inputs aquellas variables que representen recursos tangibles e intangibles utilizados en el proceso de producción o en la realización de una actividad. Pueden incluir materias primas, mano de obra, capital... Estos recursos deben ser controlados en cierta medida por las DMUs y además deben ser medibles y comparables para permitir la obtención de una medida de eficiencia relativa al identificar las unidades que logran mejores resultados utilizando una cantidad dada de insumos.

La decisión de los inputs elegidos en este trabajo es:

Personal y auxiliares de biblioteca

El personal bibliotecario es un grupo de profesionales que trabajan en bibliotecas y se encargan de gestionar y facilitar el acceso a los recursos de información para los usuarios. Este personal puede incluir bibliotecarios, archivistas, especialistas en información y otros profesionales relacionados con la biblioteconomía.

Entre las tareas del personal bibliotecario se encuentran la adquisición y catalogación de materiales bibliográficos, la organización de los recursos de información en la biblioteca, la atención al público y la promoción de los servicios de la biblioteca. También pueden estar encargados de la gestión de programas educativos y culturales, la administración del presupuesto y la planificación estratégica de la biblioteca.

Un auxiliar de biblioteca es un profesional que trabaja en una biblioteca y se encarga de llevar a cabo una variedad de tareas operativas y administrativas para apoyar el funcionamiento diario de esta. Este personal es responsable de tareas como el mantenimiento de los materiales bibliográficos, la organización de eventos y actividades y la gestión de préstamos y devoluciones de materiales bibliográficos. Los auxiliares de biblioteca también pueden ser responsables de la limpieza y mantenimiento de la estancia.

Usuarios activos totales

Son aquellas personas que han utilizado los servicios o recursos de información de la biblioteca en el último año. Estos servicios pueden incluir la consulta de libros, revistas, periódicos, material audiovisual, el uso de servicios de internet, la asistencia a actividades culturales y educativas, entre otros.

La definición de usuario activo puede variar según la biblioteca, pero en general se utiliza como una medida para evaluar la utilización de los recursos y servicios de la biblioteca por parte de la comunidad a la que se dirige.

La cifra de usuarios activos es un indicador importante para la evaluación del impacto y la eficacia de las bibliotecas públicas estatales en la sociedad, y también se utiliza para planificar y mejorar los servicios que se ofrecen a los usuarios. En España, la Red de Bibliotecas Públicas de titularidad estatal está gestionada por el Ministerio de Cultura y Deporte.

Gastos de adquisición bibliográfica

Son los costos asociados a la compra de nuevos materiales bibliográficos para enriquecer y mantener actualizada la colección de la biblioteca.

Estos gastos incluyen la adquisición de libros, revistas, periódicos, materiales audiovisuales y otros recursos de información que son adquiridos por la biblioteca para satisfacer las necesidades de sus usuarios y proporcionar acceso a la información actualizada y relevante.

En el caso de las bibliotecas públicas de titularidad estatal en España, los gastos de adquisición bibliográfica pueden estar cubiertos en parte o en su totalidad por el

Ministerio de Cultura y Deporte, a través del programa de subvenciones destinadas a la adquisición de fondos bibliográficos y documentales para bibliotecas.

Además de la compra de nuevos materiales bibliográficos, los gastos de adquisición bibliográfica también pueden incluir la suscripción a bases de datos y recursos digitales, la adquisición de licencias de software para la gestión de la biblioteca, y otros costos asociados a la gestión y mantenimiento de la colección y servicios de la biblioteca.

Para la **elección de los outputs** hay que seleccionar las características, productos o servicios que es capaz de producir una biblioteca a lo largo de un año y que mejor representen a dicho centro.

Se van a considerar outputs aquellas variables que tengan resultados medibles (que se pueden cuantificar y medir en términos de cantidad, calidad o valor), que reflejen los objetivos y metas específicos de las unidades analizadas y que sean útiles para evaluar el desempeño relativo de las DMUs al comparar cómo utilizan sus inputs para generar diferentes resultados.

La elección de outputs en este trabajo es:

Nuevos usuarios totales

Son aquellos que han registrado su pertenencia a la biblioteca y han utilizado alguno de sus servicios o recursos de información por primera vez en el último año.

Esto significa que un usuario que ha utilizado los servicios de la biblioteca en años anteriores no se considerará un nuevo usuario total, sino que se contará como un usuario activo. Los nuevos usuarios totales son una medida importante para evaluar la capacidad de la biblioteca para atraer a nuevos miembros de la comunidad y para alcanzar a aquellos que no han utilizado previamente los servicios de la biblioteca.

La cifra de nuevos usuarios totales es un indicador clave para medir el alcance y la efectividad de las actividades de promoción y publicidad de la biblioteca, y para evaluar si está llegando a nuevos segmentos de la población. En España, las bibliotecas públicas de titularidad estatal suelen llevar un registro de los nuevos usuarios totales como parte de su gestión y estadísticas, que pueden ser consultadas por los responsables de la biblioteca y el público en general.

Préstamos totales

Son la suma de todos los préstamos de materiales bibliográficos que se han realizado en un período determinado, en este caso un año. Estos préstamos pueden incluir la consulta y el retiro de libros, revistas, periódicos, materiales audiovisuales y otros recursos de información de la biblioteca.

5.2 Selección del modelo

Para la realización del trabajo se ha utilizado el modelo **DEA BCC Output Orientado**.

Este enfoque permite identificar las mejores prácticas y proporciona una visión de cómo las bibliotecas pueden utilizar sus recursos de manera más eficiente para lograr mejores resultados.

Los outputs seleccionados, es decir, nuevos usuarios y préstamos totales, son indicadores clave del desempeño de una biblioteca. Estas variables reflejan el éxito en atraer y retener usuarios, así como el nivel de uso de los recursos bibliotecarios.

En el contexto de las bibliotecas, los gastos de adquisición bibliográfica y el personal son dos variables clave que representan los recursos financieros y humanos utilizados. Al utilizar un modelo DEA BCC output orientado, se puede evaluar cómo las bibliotecas están utilizando estos recursos para lograr los resultados deseados, es decir, nuevos usuarios y préstamos.

Comparación y benchmarking: El modelo DEA BCC output orientado permite comparar el desempeño relativo de diferentes bibliotecas y establecer un benchmark (estudio profundizado sobre los competidores para entender las estrategias y mejores prácticas utilizadas por ellos) para identificar las mejores prácticas.

Al analizar cómo las bibliotecas utilizan sus recursos y generan resultados, es posible identificar aquellas que están utilizando de manera más eficiente sus insumos y obtener ideas para mejorar el desempeño de otras bibliotecas.

5.3 Eficiencia DEA en los distintos escenarios

Como se ha explicado al inicio del capítulo, se considera escenario a cada una de las posibles combinaciones de inputs/outputs para las cuales se puede aplicar DEA en cada biblioteca y obtener sus valores de eficiencia a partir de ellos. Se obtienen los resultados de la eficiencia DEA mediante la aplicación de diversas funciones que se incluyen en el paquete “deaR” de R.

Una vez realizado el análisis DEA en cada escenario, se obtiene una tabla de 21 filas (Todas las posibles combinaciones inputs/outputs) y 53 columnas (Todas las DMUs)

Debido a las elevadas dimensiones de dicha tabla, únicamente se va a mostrar un pequeño fragmento (Fig. 5.1) que contiene las eficiencias en cada uno de los escenarios para ocho de las bibliotecas, aunque el análisis se va a realizar es utilizando la matriz de datos completa.

A partir de ahora se utilizará, para facilitar la comprensión, la nomenclatura: **A** = Personal, **B** = Usuarios activos, **C** = Gastos adquisición, **1** = Préstamos y **2** = Nuevos usuarios.

	A_1	A_2	A_12	B_1	B_2	B_12	C_1	C_2	C_12	AB_1	AB_2	AB_12	AC_1	AC_2	AC_12	BC_1	BC_2	BC_12	ABC_1	ABC_2	ABC_12
A-Coruña	0.70	0.12	0.70	0.61	0.16	0.61	0.54	0.12	0.55	0.70	0.16	0.70	0.85	0.12	0.85	1.00	0.16	1.00	1.00	0.16	1.00
Albacete	0.39	0.12	0.40	0.39	0.13	0.39	0.50	0.18	0.52	0.39	0.13	0.40	0.54	0.18	0.54	0.66	0.18	0.67	0.66	0.18	0.67
Alicante/Alacant	0.36	0.09	0.36	0.44	0.12	0.44	0.17	0.08	0.18	0.44	0.13	0.44	0.40	0.09	0.40	0.59	0.12	0.59	0.61	0.13	0.61
Almería	0.17	0.07	0.18	0.21	0.09	0.22	0.14	0.08	0.16	0.21	0.10	0.22	0.23	0.09	0.23	0.40	0.14	0.43	0.40	0.15	0.44
Ávila	0.19	0.07	0.20	0.24	0.09	0.25	0.14	0.08	0.17	0.24	0.11	0.25	0.26	0.10	0.26	0.44	0.15	0.47	0.45	0.16	0.49
Badajoz	0.49	0.10	0.49	0.43	0.10	0.43	0.43	0.10	0.44	0.49	0.10	0.49	0.63	0.10	0.63	0.67	0.11	0.67	0.68	0.11	0.68
Burgos	0.52	0.12	0.52	0.33	0.12	0.34	0.29	0.12	0.32	0.52	0.12	0.52	0.60	0.12	0.60	0.42	0.12	0.43	0.60	0.12	0.60
Cáceres	0.62	1.00	1.00	0.46	1.00	1.00	0.46	1.00	1.00	0.62	1.00	1.00	0.75	1.00	1.00	0.68	1.00	1.00	0.75	1.00	1.00
Cádiz	0.20	0.07	0.20	0.32	0.11	0.33	0.13	0.07	0.16	0.32	0.12	0.33	0.23	0.07	0.23	0.46	0.11	0.49	0.48	0.12	0.51
Castelló-De-La-Plana	0.44	0.09	0.44	0.49	0.07	0.49	0.18	0.04	0.18	0.49	0.13	0.49	0.52	0.09	0.52	0.73	0.09	0.74	0.80	0.13	0.81
Ceuta	0.07	0.54	0.54	0.38	1.00	1.00	0.05	0.48	0.48	0.38	1.00	1.00	0.08	0.54	0.54	0.40	1.00	1.00	0.40	1.00	1.00
Ciudad-Real	0.30	0.09	0.30	0.36	0.11	0.36	0.45	0.14	0.46	0.36	0.11	0.36	0.45	0.14	0.46	0.67	0.17	0.67	0.67	0.17	0.67
Córdoba	0.17	0.04	0.17	0.40	0.07	0.40	0.11	0.04	0.12	0.40	0.07	0.40	0.20	0.04	0.20	0.51	0.07	0.51	0.51	0.07	0.51
Cuenca	0.24	0.07	0.24	0.37	0.10	0.37	0.30	0.10	0.32	0.37	0.10	0.37	0.33	0.10	0.33	0.75	0.17	0.76	0.75	0.17	0.76
Gijón	0.26	0.05	0.26	0.65	0.07	0.65	0.30	0.05	0.30	0.65	0.07	0.65	0.34	0.05	0.34	1.00	0.10	1.00	1.00	0.10	1.00
Girona	0.51	0.14	0.51	0.41	0.16	0.42	1.00	1.00	1.00	0.51	0.16	0.51	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.10	1.00
Granada	0.23	0.14	0.26	0.23	0.19	0.31	0.21	0.16	0.28	0.23	0.19	0.31	0.30	0.16	0.31	0.43	0.22	0.51	0.43	0.22	0.51
Guadalajara	0.48	0.13	0.48	0.44	0.13	0.44	0.73	0.30	0.74	0.48	0.13	0.48	0.75	0.30	0.75	0.82	0.30	0.82	0.82	0.30	0.82
Huelva	0.09	0.60	0.60	0.21	1.00	1.00	0.08	0.60	0.60	0.21	1.00	1.00	0.11	0.60	0.60	0.29	1.00	1.00	0.30	1.00	1.00
Huesca	0.22	0.05	0.22	0.34	0.07	0.34	0.20	0.06	0.20	0.34	0.08	0.34	0.30	0.07	0.30	0.64	0.14	0.65	0.64	0.14	0.65
Jaén	0.23	0.08	0.24	0.35	0.11	0.35	0.12	0.07	0.14	0.35	0.13	0.35	0.27	0.08	0.27	0.46	0.11	0.48	0.48	0.13	0.50
Las-Palmas-De-Gran-Canaria	0.31	0.50	0.53	0.33	0.67	0.77	0.25	0.50	0.53	0.33	0.67	0.77	0.36	0.50	0.53	0.56	0.67	0.87	0.56	0.67	0.87
León	0.43	0.16	0.44	0.37	0.16	0.40	0.33	0.16	0.38	0.43	0.16	0.44	0.52	0.16	0.52	0.51	0.16	0.53	0.54	0.16	0.54
Lleida	0.64	0.40	0.73	0.55	0.43	0.70	0.31	0.40	0.56	0.64	0.43	0.73	0.70	0.40	0.74	0.65	0.43	0.73	0.70	0.43	0.74
Logroño	1.00	0.32	1.00	1.00	0.43	1.00	0.50	0.32	0.65	1.00	0.43	1.00	1.00	0.32	1.00	1.00	0.43	1.00	1.00	0.43	1.00
Lugo	0.62	0.10	0.62	0.45	0.11	0.45	0.39	0.09	0.40	0.62	0.12	0.62	0.73	0.10	0.73	0.73	0.11	0.73	0.82	0.12	0.82
Madrid	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01	0.01

	A_1	A_2	A_12	B_1	B_2	B_12	C_1	C_2	C_12	AB_1	AB_2	AB_12	AC_1	AC_2	AC_12	BC_1	BC_2	BC_12	ABC_1	ABC_2	ABC_12
Málaga	0.14	0.10	0.16	0.54	0.19	0.59	0.10	0.10	0.15	0.54	0.19	0.59	0.17	0.10	0.17	0.64	0.19	0.70	0.64	0.19	0.70
Maó-Mahón	1.00	0.34	1.00	0.36	0.10	0.36	0.21	0.07	0.23	1.00	0.34	1.00	1.00	0.34	1.00	0.59	0.12	0.61	1.00	0.34	1.00
Melilla	0.07	0.10	0.11	0.06	0.12	0.14	0.05	0.11	0.11	0.07	0.14	0.16	0.09	0.16	0.16	0.13	0.21	0.21	0.14	0.24	0.24
Mérida	0.42	0.07	0.42	0.39	0.09	0.39	0.32	0.07	0.33	0.42	0.09	0.42	0.51	0.07	0.51	0.70	0.10	0.70	0.71	0.10	0.71
Murcia	1.00	0.26	1.00	1.00	0.26	1.00	1.00	0.26	1.00	1.00	0.26	1.00	1.00	0.26	1.00	1.00	0.26	1.00	1.00	0.26	1.00
Orihuela	1.00	1.00	1.00	1.00	1.00	1.00	0.07	0.02	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Ourense	0.37	0.20	0.41	0.37	0.25	0.43	0.32	0.20	0.40	0.37	0.25	0.43	0.45	0.20	0.45	0.59	0.25	0.65	0.59	0.25	0.65
Oviedo	0.20	0.26	0.31	0.24	0.27	0.36	0.23	0.26	0.36	0.24	0.27	0.36	0.24	0.26	0.36	0.36	0.27	0.42	0.36	0.27	0.42
Palencia	0.41	0.09	0.41	0.23	0.07	0.23	0.20	0.06	0.22	0.41	0.09	0.41	0.47	0.09	0.47	0.34	0.07	0.34	0.47	0.09	0.47
Palma	0.41	0.12	0.41	0.39	0.14	0.39	0.25	0.12	0.27	0.41	0.14	0.41	0.47	0.12	0.47	0.56	0.14	0.57	0.56	0.14	0.57
Pontevedra	0.39	0.09	0.39	0.39	0.12	0.39	0.30	0.09	0.32	0.39	0.13	0.39	0.47	0.09	0.47	0.69	0.13	0.70	0.70	0.13	0.71
Salamanca	0.28	0.17	0.32	0.23	0.19	0.30	0.21	0.17	0.29	0.28	0.19	0.32	0.33	0.17	0.34	0.35	0.19	0.39	0.36	0.19	0.39
Santa-Cruz-De-Tenerife	0.26	0.16	0.29	0.30	0.22	0.36	0.13	0.16	0.23	0.30	0.22	0.36	0.30	0.16	0.31	0.41	0.22	0.46	0.41	0.22	0.46
Santander	0.34	0.17	0.38	0.55	0.18	0.55	0.31	0.17	0.38	0.55	0.18	0.55	0.34	0.17	0.38	0.68	0.18	0.68	0.68	0.18	0.68
Santiago-De-Compostela	0.50	0.13	0.50	0.53	0.15	0.53	0.45	0.13	0.47	0.53	0.15	0.53	0.59	0.13	0.59	0.79	0.15	0.79	0.79	0.15	0.79
Segovia	0.42	0.11	0.42	0.31	0.12	0.31	0.28	0.10	0.31	0.42	0.13	0.42	0.49	0.11	0.49	0.52	0.13	0.54	0.57	0.14	0.57
Sevilla	0.28	0.14	0.30	0.33	0.16	0.35	0.29	0.14	0.33	0.33	0.16	0.35	0.33	0.14	0.34	0.52	0.16	0.54	0.52	0.16	0.54
Soria	0.48	0.09	0.48	0.37	0.10	0.37	0.39	0.10	0.39	0.48	0.10	0.48	0.60	0.10	0.60	0.61	0.11	0.61	0.64	0.11	0.64
Tarragona	0.84	0.54	0.98	0.60	0.53	0.81	0.60	0.51	0.83	0.84	0.55	0.98	1.00	0.54	1.00	0.92	0.54	1.00	1.00	0.55	1.00
Teruel	0.15	0.03	0.15	0.55	0.06	0.55	0.14	0.05	0.15	0.55	0.06	0.55	0.24	0.07	0.24	1.00	1.00	1.00	1.00	1.00	1.00
Toledo	0.21	0.11	0.24	0.37	0.15	0.37	0.47	0.20	0.50	0.37	0.15	0.37	0.47	0.20	0.50	0.81	0.28	0.82	0.81	0.28	0.82
València	0.65	0.12	0.65	0.44	0.11	0.44	0.34	0.11	0.35	0.65	0.12	0.65	0.74	0.12	0.74	0.58	0.11	0.59	0.74	0.12	0.74
Valladolid	0.40	0.17	0.42	0.43	0.17	0.45	0.34	0.17	0.40	0.43	0.17	0.45	0.40	0.17	0.42	0.43	0.17	0.45	0.43	0.17	0.45
Vitoria-Gasteiz	0.39	0.09	0.39	0.62	0.14	0.62	0.23	0.09	0.25	0.62	0.15	0.62	0.45	0.09	0.45	0.83	0.14	0.84	0.86	0.15	0.87
Zamora	0.31	0.05	0.31	0.33	0.06	0.33	0.36	0.06	0.36	0.33	0.06	0.33	0.41	0.06	0.41	0.58	0.08	0.58	0.58	0.08	0.58
Zaragoza	0.66	0.13	0.66	0.42	0.13	0.42	0.32	0.13	0.35	0.66	0.13	0.66	0.75	0.13	0.75	0.50	0.13	0.51	0.75	0.13	0.75

Fig.5.1 Eficiencias DEA para los distintos escenarios

Se observa rápidamente que el mismo algoritmo, aplicado sobre diferentes combinaciones de variables, proporciona medidas de eficiencia muy diversas.

En todos los escenarios se aprecia una tendencia general, la cual es que cuantas más variables input/output se utilizan para aplicar DEA, mayor suele ser el valor de la eficiencia para esas bibliotecas. Aunque hay excepciones, como por ejemplo “Cáceres”, la cual aparece como eficiente en muchos escenarios, pero resulta que, para algunos, como en el que se utilizan todas las variables excepto “nuevos usuarios”, toma un valor menor. Esto es indicador de que los valores para su output “Préstamos” no son demasiado buenos, por lo que únicamente funciona realmente bien (es eficiente) en el caso en el que se toma únicamente “nuevos usuarios” como output o las dos variables. Esto es un claro ejemplo de cómo afecta la elección de una u otra variable al análisis de la eficiencia DEA y es justamente lo que se tratará de mitigar aplicando ACP.

Trabajando con la tabla completa, se llega a ver que existen ciertas combinaciones de recursos y resultados para las cuales hay un notable número de bibliotecas que son eficientes en comparación con los demás escenarios. Estas son:

- La combinación que usa todos los inputs y outputs para la cual salen 12 bibliotecas eficientes (Toman el valor 1)
- El escenario que toma todos los inputs y únicamente “préstamos” como output, con 11 eficientes.
- La combinación que utiliza todos los inputs excepto “personal” y todos los outputs, con 8 eficientes.

Por el contrario, hay situaciones en las cuales aparecen muy pocos centros como eficientes, como es el caso de:

- Las combinaciones: “personal”-“préstamos”, “gastos adquisición”-“préstamos” y “gastos adquisición”-“nuevos usuarios”, las cuales solamente tienen 2 bibliotecas eficientes.

Esto apoya lo dicho previamente sobre que los valores de eficiencia son mucho más bajos para aquellas combinaciones que utilizan pocas variables al realizar el análisis.

También va a tener importancia ver que bibliotecas aparecen como eficientes en más casos, ya que si un centro aparece con valores de eficiencia muy altos (llegando a ser 1, es decir, eficientes, en varios escenarios) esto significará que, en general, esa biblioteca está funcionando de una forma correcta en el sentido de la eficiencia y viceversa.



Fig.5.2

A continuación, se adjunta una tabla con aquellos centros que más veces aparecen como eficientes. Se estudiará y se tratarán de sacar conclusiones del porqué de estos resultados.

Como se observa en la Fig.5.3 las cinco que más aparecen: Orihuela (provincia de Alicante), Murcia, Cáceres, Logroño (provincia de La Rioja) y Girona. La gran mayoría pertenecen a comunidades autónomas costeras o bastante transitadas (como es Logroño, que no es un pueblo costero, pero es parada importante para los peregrinos en el camino de Santiago) a excepción de Cáceres. Esto podría deberse a que, al ser zonas de turismo elevado, muchos de los visitantes extranjeros las visitan o utilizan durante sus estancias en España.

Pese a que una biblioteca únicamente se considera eficiente cuando toma el valor 1, hay muchas que obtienen valores de eficiencia muy elevados para muchas de sus combinaciones que se deben tener en cuenta, como Tarragona, Maó-Mahón (Menorca), A-Coruña o Gijón (Asturias). Es posible observar que ocurre lo mismo que anteriormente, es decir, que las que suelen aparecer como más eficientes para varias combinaciones son bibliotecas situadas en zonas de veraneo.

Especialmente, cabe mencionar que bastantes de las bibliotecas a priori más eficientes pertenecen a las zonas cercanas al mar mediterráneo (Situadas al este del país, como se observa en la Fig.5.2)

Biblios	vecesEf	Biblios	vecesEf	Biblios	vecesEf
A-Coruña	4	Huelva	8	Zaragoza	0
Albacete	0	Huesca	0	Zamora	0
Alicante/Alacant	0	Jaén	0	Vitoria-Gasteiz	0
Almería	0	Las-Palmas-De-Gran-Canaria	0	Valladolid	0
Ávila	0	León	0	València	0
Badajoz	0	Lleida	0	Toledo	0
Burgos	0	Logroño	12	Teruel	6
Cáceres	14	Lugo	0	Tarragona	5
Cádiz	0	Madrid	0	Soria	0
Castelló-De-La-Plana	0	Málaga	0	Sevilla	0
Ceuta	8	Maó-Mahón	8	Segovia	0
Ciudad-Real	0	Melilla	0	Santiago-De-Compostela	0
Córdoba	0	Mérida	0	Santander	0
Cuenca	0	Murcia	14	Santa-Cruz-De-Tenerife	0
Gijón	4	Orihuela	18	Salamanca	0
Girona	12	Ourense	0	Pontevedra	0
Granada	0	Oviedo	0	Palma	0
Guadalajara	0			Palencia	0

Fig.5.3 Número de escenarios en los que una biblioteca es eficiente

5.4 Eliminación de la influencia provocada por la elección de inputs/outputs utilizando el análisis de componentes principales

Como se ha dicho anteriormente, en este apartado se tratará de eliminar el efecto de la elección de inputs y outputs que se produce a la hora de medir la eficiencia DEA en las bibliotecas mediante la utilización del análisis de componentes principales, reduciendo el número de medidas de eficiencia a una sola y observando si esta medida única resume adecuadamente, desde el punto de vista de la eficiencia, todas las medidas de eficiencia de cada DMU.

Se definen como escenarios a todas aquellas combinaciones posibles de inputs/outputs utilizados para medir la eficiencia DEA en cada una de las bibliotecas. Para cada uno de ellos se obtendrán resultados distintos, los cuales se tratarán de resumir mediante los valores de la primera componente principal.

Metodología

Para la realización del ACP sobre la tabla de eficiencias DEA es fundamental apoyarse en algún software, es este caso R, aunque también se han contrastado diversas partes haciendo uso de la herramienta SAS.

En primer lugar, se leerán los datos obtenidos en el apartado anterior (5.3), es decir, se usará la tabla de eficiencias de las distintas combinaciones de inputs/outputs para cada una de las bibliotecas. Se centrarán estos datos (aunque no es estrictamente necesario realizar un análisis normado ya que las variables tienen las mismas unidades de medida) y posteriormente se realizará el código correspondiente (Se encuentra en el anexo II) y se obtendrán las salidas necesarias para construir un ranking de eficiencia de las bibliotecas.

Aplicación del ACP

Se llevará a cabo, en primer lugar, un análisis de las bibliotecas. En este caso se hará un análisis normado, por lo que se trabajará con la matriz $Corr(TablaEficiencias)$, siendo $TablaEficiencias$ la tabla del apartado anterior. Previamente se estandarizarán los datos tras leerlos (Variables con media 0 y varianza 1)

Por este motivo, las componentes son autovectores de la matriz de correlaciones y son distintos de los de la matriz de covarianzas. Si se actúa así, se les da igual importancia a todas las variables originales. En la matriz de correlaciones todos los elementos de la diagonal son iguales a uno. Si las variables originales están tipificadas, esto implica que su matriz de covarianzas es igual a la de correlaciones, con lo que su variabilidad total (traza) es igual al número total de variables que hay.

Lo primero será la obtención e interpretación de los autovalores y autovectores. Después se calcularán los “scores” para las 21 componentes y se comprobará que tienen media cero (es decir, que el gráfico de las bibliotecas estará centrado en cero) y varianza igual al autovalor correspondiente (más concretamente, la suma de los cuadrados de los scores en cada componente divididos por $n-1$, que es la varianza, son iguales a los autovectores) También se comprueba que la inercia total suma 21 al hacer un análisis normado.

Se construirá una tabla (Fig.5.4) que contenga las inercias, los porcentajes de inercia explicada, la inercia acumulada y el porcentaje de inercia acumulada por cada uno de los primeros autovalores. A partir de esta tabla y las Fig.5.5 y Fig.5.6, se tomará una decisión acerca del número adecuado de componentes que se deberían retener para perder la menor cantidad de información posible.

	Autovalor	Inercia	% Inercia	Inercia acumulada	% Inercia Acumulada
1	1	12.0786	57.5174	12.0786	57.5174
2	2	4.3607	20.7654	16.4394	78.2827
3	3	1.6885	8.0403	18.1278	86.3230
4	4	1.4981	7.1339	19.6260	93.4569
5	5	0.5834	2.7779	20.2093	96.2348
6	6	0.3377	1.6083	20.5471	97.8431

Fig.5.4 Tabla con información de la inercia recogida por cada autovalor

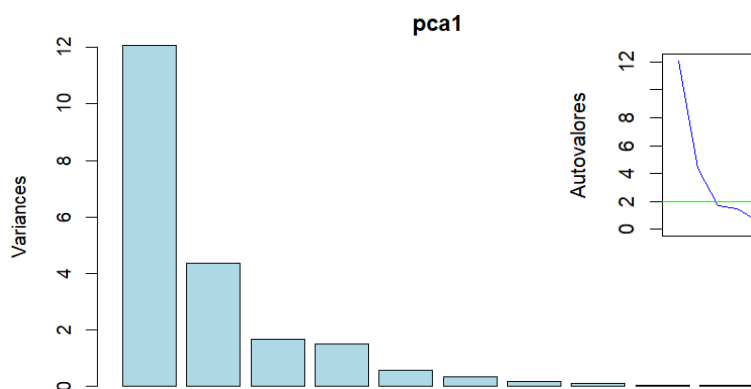


Fig.5.5 Histograma de la cantidad de variación recogida por las principales componentes

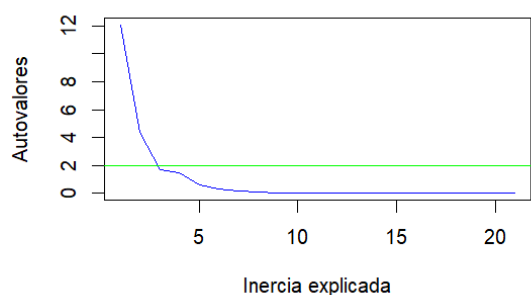


Fig.5.6 Scree-plot de los autovalores junto a su cantidad de inercia explicada

Tras la observación de la tabla y los gráficos, parece que lo ideal sería utilizar las cuatro primeras componentes, ya que estas acumulan más del 93% de la inercia total. Además, su inercia explicada supera el promedio de los autovalores, que en este caso es uno. Pero examinando el scree-plot, se puede ver que el codo de la gráfica se encuentra situado en el autovalor dos, por lo que también parece razonable tomar únicamente los dos primeros autovalores. A esta misma conclusión se llega si se estudia el histograma.

Pese a todo esto, lo fundamental en este apartado será ver si la primera componente es capaz de resumir por sí misma toda la información que nos aporta la tabla con las eficiencias DEA en los distintos escenarios para cada una de las bibliotecas. Esta componente recoge el 57.5% de la variabilidad, valor que se va a considerar suficientemente alto como para poder lograr este objetivo.

También es imprescindible estudiar los pesos (autovectores) asignados a las primeras componentes para los distintos escenarios:

	PC1	PC2
A_1	0.2086	0.2401
A_2	0.2149	-0.2567
A_12	0.2538	0.0671
B_1	0.2092	0.2070
B_2	0.1952	-0.3272
B_12	0.2506	-0.1006
C_1	0.1510	0.2391
C_2	0.1801	-0.2184
C_12	0.2075	0.0207
AB_1	0.2246	0.2373
AB_2	0.1995	-0.3189
AB_12	0.2626	-0.0338
AC_1	0.2131	0.2539
AC_2	0.2235	-0.2262
AC_12	0.2559	0.1153
BC_1	0.1955	0.2579
BC_2	0.1968	-0.2887
BC_12	0.2454	0.0263
ABC_1	0.2038	0.2886
ABC_2	0.2004	-0.2832
ABC_12	0.2532	0.0664

Fig.5.7 autovectores de las primeras componentes

Los pesos asignados en la primera componente en todas las variables son aproximadamente iguales entre ellos, esto significa que todas las variables contribuyen por igual en la dirección y la variabilidad de esa componente. Esto puede ocurrir cuando las variables tienen una correlación similar o cuando todas las variables son igualmente importantes en el contexto del análisis (Esta situación se dará debido a la previa estandarización de los datos)

Una cuestión interesante en lo que se refiere a la interpretación de las variables es notar que, a diferencia de la nube de puntos, la nube de las variables no está centrada con lo que puede ocurrir que todas las coordenadas de las variables en un eje tengan el mismo signo, es decir que, por ejemplo, $G_{variable\ eje} > 0$, para algún eje.

Cuando eso ocurre en el primer eje, es decir $G_{variable\ eje1} > 0$ (como es el caso) o $G_{variable\ eje1} < 0$, para todas las variables se dice que el primer eje es un **factor de tamaño**.

El motivo es que cuando nos movamos en la dirección del primer eje todas las variables crecerán (si nos movemos en la dirección en la que apuntan los $G_{variable\ eje1}$) o decrecerán (si nos movemos en la contraria) a la vez y las observaciones serán más grandes (o pequeñas) en todas las variables a la vez con lo que tienen mayor “tamaño”.

Se han realizado unas representaciones (Fig.5.8 y Fig.5.9) que darán una idea de que variables crecen simultáneamente y cuáles no, observando más claramente el factor tamaño que se acaba de explicar.

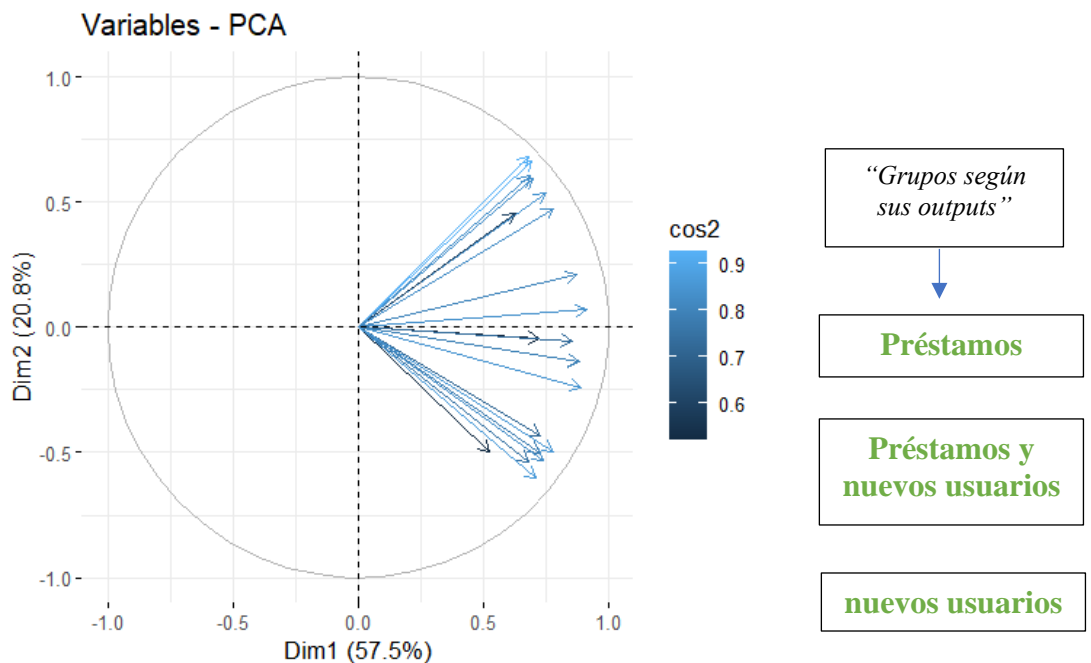


Fig.5.8 Orientación y direcciones de crecimiento de las variables

Cada variable mide una característica observada sobre las bibliotecas y, por tanto, se puede ver qué variables miden cosas parecidas en estas. Se puede observar como las variables se agrupan en **tres grupos** distintos. Estos grupos parecen formarse **en función de los outputs** que toman los distintos escenarios.

El grupo más relacionado con la primera dimensión es el que cuentas con todas las variables de salida del estudio.

Se va a realizar un gráfico en el que se observan claramente estas agrupaciones. El análisis será similar al que se hará de la nube de las bibliotecas, pero sobre la matriz traspuesta de \mathbf{X} , de manera que las filas de esta son las variables de los datos. Esto va a permitir observar relaciones entre las variables, establecer relaciones de las variables con los ejes y también relacionar las variables con los individuos como se podrá observar en la Fig.5.9

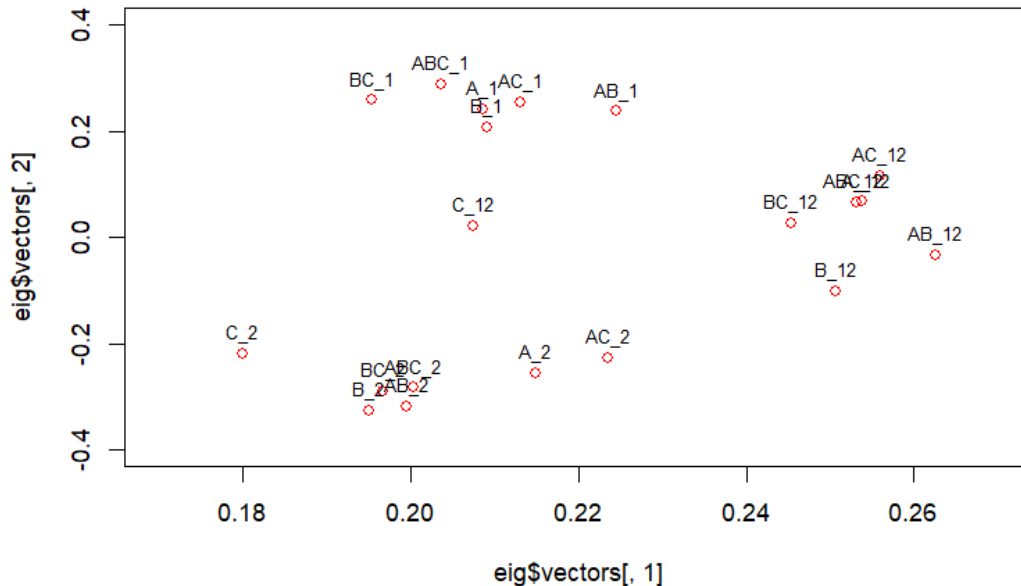


Fig.5.9 Plot de las variables del ACP

Ahora se procederá a interpretar las primeras componentes:

La **primera componente**, que como se ha dicho, recoge el 57.5% de la información total, representa la “eficiencia”, es decir, mide un tamaño de eficiencia de la biblioteca, siendo las más eficientes aquellas que tomen valores elevados para dicha componente. Las bibliotecas con un valor elevado para este primer eje son aquellas que obtienen muy buenos resultados (en relación con la cantidad de préstamos que realizan y el número de nuevos usuarios que registran) en relación con recursos que utilizan (siendo estos, en este caso, la cantidad de personas que trabajan en ella, la cantidad de usuarios que son activos y los gastos de adquisición bibliográfica que tienen) mientras que las que tomen un valor bajo en esta componente representarán a aquellos centros en los cuales, pese a tener muchos recursos, apenas obtienen resultados.

La **segunda componente** principal, recoge el 20.8% de la inercia. En esta componente hay escenarios con pesos positivos y escenarios con pesos negativos. Podemos ver como los escenarios con pesos positivos son los que contemplan el output 1 en su definición (“préstamos”) mientras que los escenarios con pesos negativos y valores absolutos grandes son los que contemplan solo el output 2 (“nuevos usuarios”).

Entonces la segunda componente principal separa escenarios con solo el output 1 de aquellos que solo tienen el output 2 y puede interpretarse como orientación al output 1 en contra de la orientación al output 2.

Como se observará en la Fig.5.10, hay bibliotecas como Murcia que tomarán valores positivos en este eje, lo cual es indicativo de que su eficiencia es mayor cuando únicamente se utilice como output a “préstamos” mientras que tendrá una eficiencia bastante mala cuando únicamente se utiliza a “nuevos usuarios” como variable output. Con la biblioteca de Cáceres ocurrirá justo, al contrario.

Se van a representar las bibliotecas sobre las dos primeras componentes principales (Fig.5.10). Estas componentes, también conocidas como PC1 y PC2, son ejes ortogonales que capturan la mayor variabilidad en los datos después de aplicar la transformación ACP, como se ha visto anteriormente:

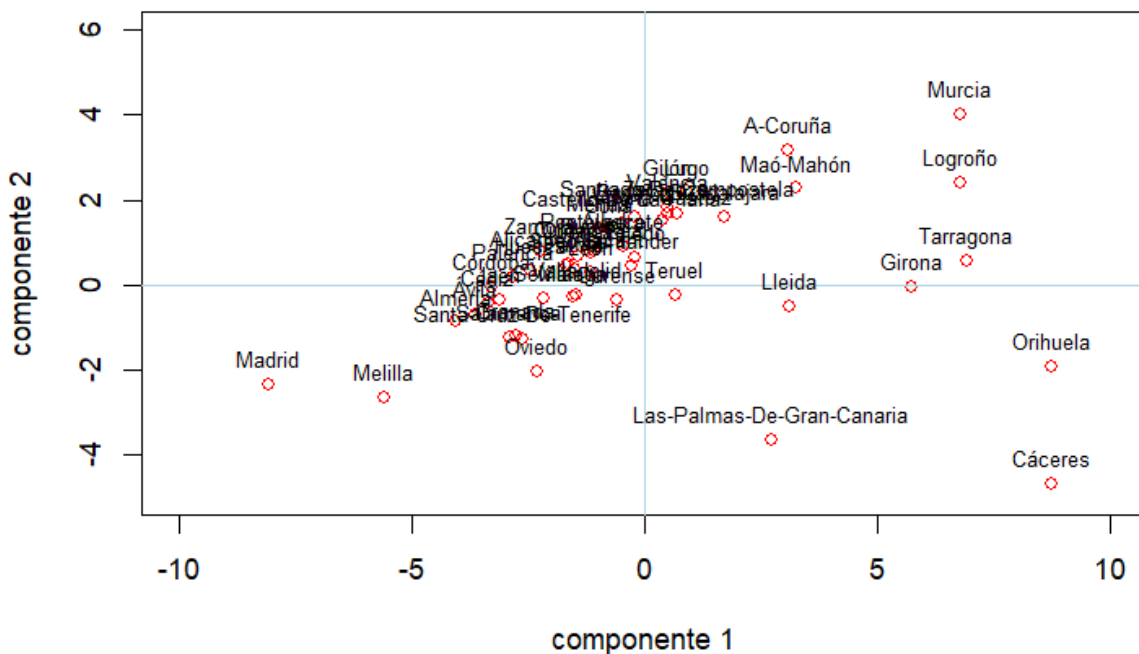


Fig.5.10 Plot de las bibliotecas en las dos primeras componentes principales

Bibliotecas que se encuentran próximas indica que estas toman valores parecidos respecto a su eficiencia DEA en la mayor parte de los escenarios.

Existe una gran aglomeración de centros situados alrededor del cero, estos representarán el caso general, es decir, las que tienen un comportamiento “normal” (Tendrán proporciones parecidas respecto sus inputs/outputs y por tanto obtendrán valores de eficiencia muy similares para todos los escenarios), lo son, entre otras, las bibliotecas de Santander, Teruel o Tenerife.

A pesar de que muchas se reúnen en torno a este primer grupo, hay un número considerable de ellas que son especialmente más eficientes que las demás debido a que cuentan con un elevado número de outputs en proporción a los pocos inputs que han utilizado. Estas son aquellas situadas a la derecha de la gráfica, como Orihuela, Murcia o Tarragona. Por el contrario, las situadas en la zona contraria, tendrán valores de eficiencia ínfimos, como son los centros de Madrid o Melilla (aquellos que obtienen pocos resultados a pesar de haber utilizado muchos recursos)

Para una interpretación más visual de la segunda componente y observación de cómo se relacionan las bibliotecas con las distintas variables se va a construir un biplot que muestre las representaciones simultáneas de ambas (Fig.5.11)

1	A-Coruña	28	Málaga
2	Albacete	29	Maó-Mahón
3	Alicante/Alacant	30	Melilla
4	Almería	31	Mérida
5	Ávila	32	Murcia
6	Badajoz	33	Orihuela
7	Burgos	34	Ourense
8	Cáceres	35	Oviedo
9	Cádiz	36	Palencia
10	Castelló-De-La-Plana	37	Palma
11	Ceuta	38	Pontevedra
12	Ciudad-Real	39	Salamanca
13	Córdoba	40	Santa-Cruz-De-Tenerife
14	Cuenca	41	Santander
15	Gijón	42	Santiago-De-Compostela
16	Girona	43	Segovia
17	Granada	44	Sevilla
18	Guadalajara	45	Soria
19	Huelva	46	Tarragona
20	Huesca	47	Teruel
21	Jaén	48	Toledo
22	Las-Palmas-De-Gran-Canaria	49	València
23	León	50	Valladolid
24	Lleida	51	Vitoria-Gasteiz
25	Logroño	52	Zamora
26	Lugo	53	Zaragoza
27	Madrid		

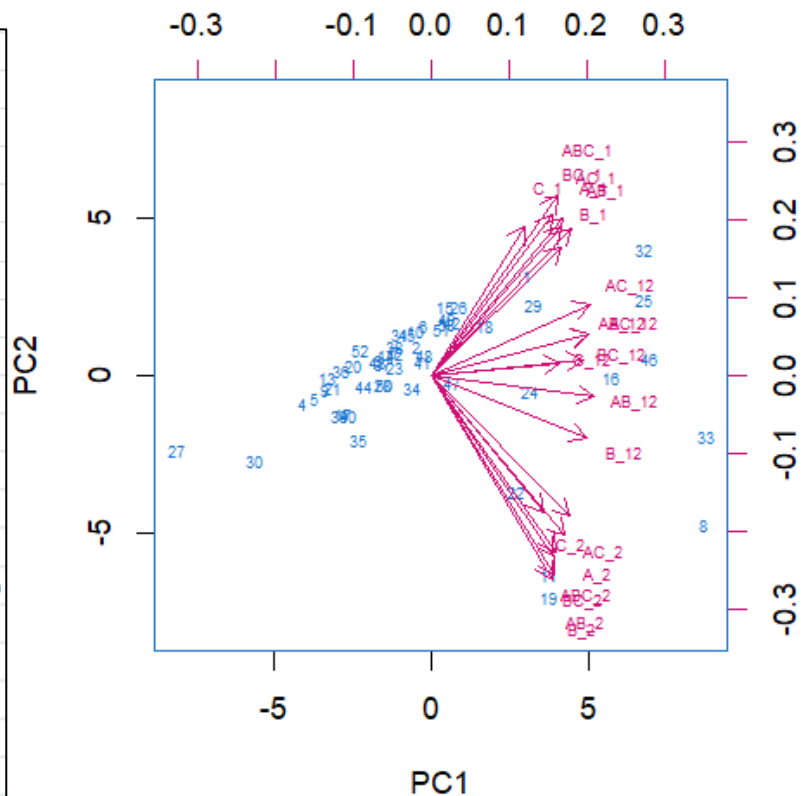


Fig.5.11 Representación conjunta de las bibliotecas y las variables

Esta es una buena representación ya que en ACP se tienen dos nubes de puntos (individuos y variables como se ha visto anteriormente) que están en espacios con orígenes y bases distintas. Sin embargo, se pueden representar las direcciones definidas por las variables activas sobre la base de los ejes factoriales de la nube de los puntos de los individuos. Para analizar la Fig.5.11 se debe tener en cuenta que: Sólo la dirección de las variables cuenta en la interpretación conjunta con los individuos. Estas proyecciones de las variables estarán dentro del círculo unidad y cuanto más cerca esté el punto de la circunferencia, mejor representada estará la dirección de crecimiento en el espacio en el que se trabaja. Los individuos cercanos al centro de la representación tomarán valores cercanos a la media de las variables y los que estén lejos del centro en la dirección de crecimiento de la variable tomarán valores altos en esa variable.

Se tienen bibliotecas que son más eficiente cuando se las aplica DEA con una combinación de inputs/outputs que tenga como output únicamente “préstamos” (1) como puede ser Maó-Mahón o Murcia, otras que lo son cuando el output es “nuevos usuarios” (2) como es el caso de Huelva o Las Palmas de Gran Canaria. Por último, las bibliotecas del tercer grupo (12), que son más eficientes cuando se toman ambos outputs son, entre otras, las de Tarragona, Girona o Lleida. Cabe destacar que gran parte de estos centros son más eficientes cuando se utilizan combinaciones con todos los outputs disponibles.

Conviene también estudiar las **contribuciones absolutas** de las bibliotecas a cada uno de los ejes, esencialmente al primero, ya que esto nos indica lo que ha contribuido cada una a la definición de un eje y especialmente habrá que prestar atención a que no haya ninguna con una contribución absoluta mucho mayor que el resto ya que entonces puede dudarse de la estabilidad de esa componente. En este caso, y fijándonos en la Fig.5.10 observamos que todas tienen una **contribución similar**. En cuanto a la **contribución relativa**, viendo los resultados en la tabla Fig.5.10, podemos observar si las bibliotecas objeto de estudio han sido representadas de forma correcta respecto a cada uno de los ejes. Aquí sí que se obtienen **resultados más dispares** ya que, fijándonos en la primera componente se ve que centros como el de Sevilla o Jaén están mejor representados que otros como por ejemplo Albacete o Santander.

Biblioteca	CA	CR	Biblioteca	CA	CR	Biblioteca	CA	CR
A-Coruña	0.0150	0.4563	Huelva	0.0225	0.2082	Zaragoza	0.0004	0.0351
Albacete	0.0003	0.0726	Huesca	0.0096	0.8069	Zamora	0.0079	0.8082
Alicante/Alacant	0.0043	0.7106	Jaén	0.0153	0.9321	Vitoria-Gasteiz	0.0002	0.0193
Almería	0.0264	0.9450	Las-Palmas-De-Gran-Canaria	0.0117	0.3388	Valladolid	0.0035	0.4744
Ávila	0.0211	0.9538	León	0.0021	0.5435	València	0.0004	0.0397
Badajoz	0.0001	0.0129	Lleida	0.0155	0.8380	Toledo	0.0001	0.0070
Burgos	0.0019	0.2504	Logroño	0.0734	0.8245	Teruel	0.0007	0.0136
Cáceres	0.1210	0.7279	Lugo	0.0013	0.1229	Tarragona	0.0764	0.9572
Cádiz	0.0180	0.9525	Madrid	0.1041	0.8817	Soria	0.0009	0.1947
Castelló-De-La-Plana	0.0004	0.0514	Málaga	0.0037	0.2460	Sevilla	0.0074	0.9259
Ceuta	0.0226	0.2391	Maó-Mahón	0.0170	0.3075	Segovia	0.0044	0.7696
Ciudad-Real	0.0021	0.3666	Melilla	0.0499	0.7805	Santiago-De-Compostela	0.0008	0.1261
Córdoba	0.0171	0.8587	Mérida	0.0015	0.2977	Santander	0.0001	0.0364
Cuenca	0.0032	0.3801	Murcia	0.0734	0.6570	Santa-Cruz-De-Tenerife	0.0110	0.7760
Gijón	0.0004	0.0136	Orihuela	0.1214	0.6801	Salamanca	0.0134	0.7383
Girona	0.0522	0.4356	Ourense	0.0006	0.4487	Pontevedra	0.0020	0.4674
Granada	0.0122	0.8138	Oviedo	0.0085	0.4919	Palma	0.0037	0.8299
Guadalajara	0.0046	0.2476	Palencia	0.0127	0.7314			

Fig.5.12
Contribuciones
relativas y
absolutas a la
primera
componente

Ranking

Una vez decidido que la primera componente resume “el tamaño de la eficiencia DEA” para todos los escenarios propuestos, se considera como nueva medida de eficiencia global al resultado de la primera componente principal en cada biblioteca (Fig.5.13). Esta medida permite crear un ranking de las bibliotecas (En la tabla también se va a añadir el resultado que obtiene cada una cuando se le ha aplicado DEA utilizando todos los inputs/outputs para comparar)

Bibliotecas	Ranking	ABC_12
Orihuela	8.732	1.00
Cáceres	8.717	1.00
Tarragona	6.925	1.00
Logroño	6.792	1.00
Murcia	6.788	1.00
Girona	5.723	1.00
Ceuta	3.765	1.00
Huelva	3.762	1.00
Maó-Mahón	3.264	1.00
Lleida	3.121	0.74
A-Coruña	3.071	1.00
Las-Palmas-De-Gran-Canaria	2.706	0.87
Guadalajara	1.700	0.82
Lugo	0.888	0.82
Santiago-De-Compostela	0.695	0.79
Teruel	0.651	1.00
Zaragoza	0.495	0.75
València	0.489	0.74
Gijón	0.478	1.00
Vitoria-Gasteiz	0.365	0.87
Toledo	-0.197	0.82
Badajoz	-0.200	0.68
Santander	-0.269	0.68
Albacete	-0.450	0.67
Castelló-De-La-Plana	-0.490	0.81
Ourense	-0.613	0.65
Soria	-0.764	0.64
Mérida	-0.980	0.71

Fig.5.13 Ranking de eficiencias de las bibliotecas utilizando la primera componente

Bibliotecas	Ranking	ABC_12
Burgos	-1.081	0.60
Pontevedra	-1.121	0.71
Ciudad-Real	-1.142	0.67
León	-1.151	0.54
Cuenca	-1.427	0.76
Valladolid	-1.480	0.45
Palma	-1.515	0.57
Málaga	-1.524	0.70
Alicante/Alacant	-1.653	0.61
Segovia	-1.665	0.57
Sevilla	-2.153	0.54
Zamora	-2.234	0.58
Oviedo	-2.312	0.42
Huesca	-2.451	0.65
Santa-Cruz-De-Tenerife	-2.627	0.46
Granada	-2.769	0.51
Palencia	-2.827	0.47
Salamanca	-2.899	0.39
Jaén	-3.097	0.50
Córdoba	-3.281	0.51
Cádiz	-3.358	0.51
Ávila	-3.645	0.49
Almería	-4.072	0.44
Melilla	-5.596	0.24
Madrid	-8.085	0.01

La medida global de eficiencia toma valores positivos y negativos diferentes para cada una de las bibliotecas. Es interesante observar que ninguna de las bibliotecas que toman valores negativos en la primera componente llegan a ser eficientes en ningún escenario.

Observando la Fig.5.13 se va a comparar los resultados del ranking con las medidas de eficiencia que se obtienen mediante DEA utilizando las cinco variables input/output. Se observa como las bibliotecas que están en la parte superior del ranking(mejores resultados basado en la medida global que se ha obtenido) son eficientes para el escenario completo (toman valor 1) Pero solo se da este caso hasta la novena posición (correspondiente a Maó-Mahón), ya que a partir de aquí, existen bibliotecas como la de A-Coruña, Teruel o Gijón que pese a ser eficientes para el escenario tratado, no se encuentran en las primeras posiciones del ranking, sino que estas están por debajo de otras como Lleida o Las Palmas de Gran Canaria que tienen valores de eficiencia de 0.74 y 87 respectivamente. Esto se debe a que a la hora de realizar el ranking mediante ACP, se han tenido en cuenta los resultados para todas las combinaciones de inputs/outputs, por lo que esto habrá afectado en cierta medida a que se de este fenómeno en la tabla. A partir de aquí esta situación se da para otros casos a lo largo del ranking, habiendo bibliotecas situadas en mejor posición que otras, pero cuyos resultados de eficiencia son menores que las que van después. En el caso de las últimas bibliotecas del ranking, también se cumple, pero solo con las dos del final, cuyos valores de eficiencia son los más bajos en relación con el resto, y esta son la de Madrid (la que peores resultados obtiene, situada en la última posición del ranking) y la de Melilla, ya que, a partir de aquí, bibliotecas con eficiencias muy bajas en este escenario como la de Oviedo (0.42) están situadas bastante arriba en la lista.

A partir de la Fig.5.13 podemos decir que las tres bibliotecas públicas de titularidad estatal españolas más eficientes son la de: Orihuela, Cáceres y Tarragona, mientras que las más ineficientes (de las cuales habría que revisar su funcionamiento ya que, al menos para los inputs/outputs utilizados, funcionan muy mal) son las de: Madrid y Melilla.

Este ranking ordena las bibliotecas por el valor que toman en su primera componente, y se puede decir que sus resultados son bastantes consistentes en relación con el número de veces que estas bibliotecas aparecen como eficientes para las distintas combinaciones de inputs/outputs. Debido a esto, se puede decir que los resultados del algoritmo DEA para todas las posibles situaciones, parece estar representado de una forma bastante fiel por la primera componente obtenida con el ACP, la cual, como se ha visto, recoge el 57.5% de la información de la tabla de eficiencias que se había usado en su construcción.

Relacionando el número de veces que una biblioteca es eficiente para los escenarios con la medida de eficiencia global, se observa que solamente el 22.64% de las bibliotecas son eficientes en uno o más casos, por lo que es más común que un centro no sea eficiente (pese a que existen algunas que puede tomar un valor de eficiencia bastante alto, sin llegar a 1 para casi todas las combinaciones). Además, no hay ninguna biblioteca que únicamente sea eficiente para una combinación, ya que cuando una resulta tomar valor 1 en alguno de los escenarios siempre lo hace al menos un par de casos más.

Llama la atención observar que incluso las bibliotecas más eficientes toman, para alguno de los 21 escenarios unos valores de eficiencia muy bajos. Esto indica que, aunque en muchas situaciones, sus combinaciones de variables arrojan resultados muy positivos, puede ser que, para una en concreto, tenga un resultado “malo”, como podría ser el caso de la biblioteca de Huelva, la cual es eficiente en 8 situaciones, que son bastantes, como se ha ido viendo, pero que en un caso concreto tiene una eficiencia con valor de 0.07 (casi cero), lo cual es indicador de una eficiencia pésima.

Tras todo lo visto a lo largo del trabajo, se puede concluir diciendo que la elección de variables input/output tienen una gran influencia a la hora de obtener los resultados de eficiencia mediante DEA y también, se puede afirmar que la técnica de ACP es muy útil a la hora de resumir los resultados de todos los posibles escenarios que se generan a partir de los inputs y outputs elegidos, eliminando la influencia de estos.

Capítulo 6: Conclusiones

El objetivo planteado en este trabajo fin de grado se ha logrado ya que se ha conseguido eliminar el efecto que produce la elección de unos determinados inputs/outputs a la hora de medir la eficiencia DEA, en este caso, sobre el conjunto de 53 bibliotecas públicas de titularidad estatal españolas para los 21 escenarios que se han generado debido a los recursos y resultados elegidos.

Esto se ha conseguido mediante la aplicación del análisis de componentes principales sobre la tabla de las eficiencias DEA de cada una de las bibliotecas para todos los escenarios posibles, logrando así una nueva medida para la eficiencia, la primera componente, que se puede utilizar para la construcción de un ranking de eficiencias para los centros. Cabe mencionar que la primera componente, la cual es una componente de tamaño y como se ha dicho, representará la eficiencia, recoge aproximadamente el 57.5% de la información de los datos originales pertenecientes a la tabla de eficiencias adquirida mediante DEA.

La mayoría de las posiciones de las bibliotecas en el ranking, tanto las mejores como las peores en el sentido de sus eficiencias, coinciden con aquellas que aparecían más veces como eficientes e ineficientes respectivamente al aplicar el algoritmo DEA sobre distintas combinaciones de inputs/outputs. Debido a que estos resultados son consistentes se puede afirmar que, los efectos de la elección de ciertos inputs y outputs han sido eliminados.

Sin embargo, es importante recordar que este proceso implica tomar decisiones basadas en la importancia de los componentes principales y supone que los componentes seleccionados capturan adecuadamente la información relevante de las salidas originales.

Podemos decir que las bibliotecas más eficientes en muchos escenarios son las de Orihuela, Murcia y Cáceres entre otras, mientras que las que nunca llegan a serlo y que peores resultados en cuanto a la eficiencia arrojan son, por ejemplo, las de Madrid, Melilla o Almería.

Se ha concluido que únicamente el 22% de las bibliotecas son eficientes para algunos de los escenarios, por lo que se puede decir que, en general, el funcionamiento de estos centros y su gestión de los recursos podría ser bastante mejor en el sentido de la eficiencia. Al menos respecto a las variables que se han utilizado en el estudio. También llama la atención que dentro del grupo de bibliotecas más eficientes la gran mayoría pertenezcan a zonas de turismo de playa en la costa mediterránea, estas son, por ejemplo, las comunidades de Murcia, Valencia o Cataluña.

Capítulo 7

Bibliografía

1. Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE (Joaquín Amat Rodrigo)
https://www.cienciadedatos.net/documentos/35_principal_component_analysis
2. DATA ENVELOPMENT ANALYSIS - History, Models and Interpretations (William W. Cooper, Lawrence M. Seiford and Joe Zhu)
https://link.springer.com/chapter/10.1007/1-4020-7798-X_1
3. An algorithm for Data Envelopment Analysis (DEA) (J. H. Dulá, Department of Management and Marketing. The University of Mississippi)
<http://www.people.vcu.edu/~jdula/WORKINGPAPERS/NewDEAalgo.pdf>
4. El método DEA y su aplicación al estudio del sector energético y las emisiones de CO2 en América Latina y el Caribe (Andrés Ricardo Schuschny)
<https://www.cepal.org/es/publicaciones/4752-metodo-dea-su-aplicacion-al-estudio-sector-energetico-emisiones-co2-america>
5. Data Science and Productivity Analytics (Vincent Charles, Juan Aparicio, Joe Zhu)
<https://link.springer.com/book/10.1007/978-3-030-43384-0>
6. Evaluación de la eficiencia mediante el análisis envolvente de datos (Vicente Coll Serrano, Olga M.^a Blasco Blasco. Universidad de Valencia)
https://www.uv.es/vcoll/libros/2006_evaluacion_eficiencia_DEA.pdf
7. Preparing your data for DEA (Joseph Sarkis)
https://www.researchgate.net/publication/226132954_Preparing_your_data_for_DEA
8. Implementation of various DEA approaches in MATLAB (Yarpiz) (Esto solo lo he tomado como idea para implementar mi código)
<https://es.mathworks.com/matlabcentral/fileexchange/53145-data-envelopment-analysis-dea>
9. TECNICAS ECONOMETRICAS EN DETALLE. EL ANÁLISIS ENVOLVENTE DE DATOS (Repositorio Institucional de la Universidad de Alicante)
<https://rua.ua.es/dspace/bitstream/10045/19658/6/Materiales.Teoría.Bloque.III.pdf>

10. Aplicación de la metodología DEA para el estudio del desempeño de futbolistas de La Liga (Velázquez Martínez, José Luis)
<https://biblus.us.es/bibing/proyectos/abreproy/92440/>
11. Análisis por Envoltura de Datos (DEA) (Repositorio de bibliotecas de la universidad de Sevilla)
<https://biblus.us.es/bibing/proyectos/abreproy/4559/fichero/Cap%C3%ADtulo+I.pdf>
12. He utilizado versiones anteriores de mi campus virtual para buscar apuntes de otras asignaturas como pueden ser Análisis de datos o Análisis multivariante.
https://cursoanterior5.campusvirtual.uva.es/2021_2022/my/
13. El coeficiente de correlación (Instituto de ciencias de la educación de la universidad de Zaragoza)
https://ice.unizar.es/sites/ice.unizar.es/files/users/leteo/materiales/01._document_o_1_correlaciones.pdf
14. Modelos DEA de metafrontera (Juan Manuel Saborido Bermejo)
<https://biblus.us.es/bibing/proyectos/abreproy/5291/fichero/MODELOS+DEA+DE+METAFRONTERA.pdf>
15. Metodología para medir la eficiencia mediante la técnica DEA (Chediak P. Francisco y Valencia A. Luz Stella)
http://vector.ucaldas.edu.co/downloads/Vector3_7.pdf
16. Metodología de análisis envolvente de datos (DEA), procesos administrativos y operacionales de las políticas gubernamentales en los países latinoamericanos (Idana Beroska Rincón Soto; León Julio Arango Buelvas; Oscar Javier Torres Yarzagaray)
<https://dialnet.unirioja.es/servlet/articulo?codigo=7286075>
17. MÉTODO DE SELECCIÓN DE VARIABLES PARA MEJORAR LA DISCRIMINACIÓN EN EL ANÁLISIS DE EFICIENCIA APLICANDO MODELOS DEA (Marcela C. González-Araya, Nelson G. Valdés Valenzuela)
<https://dialnet.unirioja.es/descarga/articulo/3998825.pdf>
18. Tutorial Data Envelopment Analysis with dear, version 1.0 (Español) (Noviembre 2018) (Vicente Coll-Serrano, Rafael Benítez, Vicente J. Bolós)
https://www.uv.es/dearshiny/Tutoriales_deaR/Tutorial_deaR_espa%C3%B1ol.pdf

19. Comparación de rankings de eficiencia mediante análisis de componentes principales y DEA (INE, Estadística española, Volumen 54) (Úrsula Faura Martínez, Juan Cándido Gómez Gallego, María Concepción Pérez Cárceles, Juan Gómez García)

https://www.ine.es/ss/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3B+filename%3DArticulo_4.pdf&blobkey=urldata&blobtable=MungoBlobs&blobwhere=211%2F249%2FArticulo+4%2C0.pdf&ssbinary=true

También he utilizado referencias de algunos libros de escritores ingleses, ya que la mayoría de la información reciente sobre este temario proviene de las comunidades angloparlantes, aunque solo los he ojeado buscando lo que me importaba.

ANEXO I

Software y librerías utilizadas

Pese a que el trabajo ha sido realizado mediante el apoyo de determinadas librerías, muchas partes se han ido haciendo mediante código a mano, ya que es una mejor forma de entender el procedimiento que se está realizando.

- **R y Rstudio**

R es un ambiente de programación formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones. Además, es gratuito y de código abierto (software libre) Esto convierte a R software en una herramienta estadística estable, confiable y a la vanguardia, ya que está sometida a una actualización permanente.

R un lenguaje orientado a objetos: Esto significa que las variables, datos, funciones, resultados, etc., se guardan en la memoria activa del computador en forma de objetos con un nombre específico. Esta característica permite aplicar cálculos a un conjunto de valores a la vez sin la necesidad de utilizar un algoritmo más sofisticado como una función bucle.

R un lenguaje interpretado (como Java) y no compilado (como Fortran o Pascal). Es decir, los comandos escritos en el teclado son ejecutados directamente sin necesidad de construir un ejecutable. Esto facilita mucho nuestro trabajo con el análisis de datos complejos.

- **library(deaR)**

Para la aplicación del algoritmo DEA a los datos, se utilizará el paquete de R **deaR** que permite ejecutar un amplio y variado número de modelos basados en el Análisis envolvente de datos. La aplicación del algoritmo consta de los siguientes pasos:

1. Se cargan los datos y el paquete.
2. Se adecuan los datos al formato que utiliza deaR mediante la función *read.data()* o *make_datadea()* que tienen como argumentos:

- data: Se refiere al conjunto de datos a analizar (tiene que ser una dataframe)
- dmus: Indicar el número de la columna donde se encuentran las DMUs. Por defecto deaR considera que las DMUs se encuentran en la primera columna.
- ni: Es el número de inputs.
- no: Es el número de outputs.
- inputs: En lugar de indicar el número de inputs se puede indicar el número de las columnas donde se encuentran los inputs.
- outputs: En lugar de indicar el número de outputs se puede indicar el número de las columnas donde se encuentran los outputs.
- nc-inputs: Si entre los inputs hay inputs no controlables se puede indicar qué input es no controlable.
- nc-outputs: Si entre los outputs hay outputs no controlables se puede indicar qué output es no controlable.

3. Se aplica la función *model_basic()*, que tiene como argumentos principales:

- data: Conjunto de datos en el formato de lectura de deaR.
- orientation: Orientación del modelo: input, output o direccional.
- rts: Rendimientos a escala del modelo (constantes, variables, no-crecientes, no- decrecientes, generalizados).
- dmu-ref: Selección de un subconjunto de DMUs
- dmu-eval: DMUs a evaluar del subconjunto seleccionado.
- dir-input: Vector de dirección input en modelos direccionales.
- dir-output: Vector de dirección output en modelos direccionales.

Existen más argumentos, aunque no serán utilizados

4. Extracción de los resultados:

Todas las salidas se encuentran almacenados en los objetos *result_pft()* y *result_nft()*. Estos objetos son una lista componentes.

Para extraer los resultados del análisis DEA, deaR cuenta con una serie de funciones específicas. Estas son:

- *eficiencias ()*: Extrae las puntuaciones de eficiencia.
- *slacks ()*: Extrae las holguras.
- *targets ()*: Extrae los valores objetivo.
- *lambdas ()*: Extrae las intensidades.
- *references ()*: Extrae el conjunto de referencia de las DMUs ineficientes.
- *rts ()*: Extrae los rendimientos a escala que caracterizan a una DMU.
- *multipliers ()*: Extrae los multiplicadores (o pesos) del modelo DEA en forma multiplicativa.

Para obtener un resumen de todos los resultados obtenidos al ejecutar el modelo DEA BCC output-orientado se utilizará la función *summary()*. En este caso, para que los valores de la eficiencia estén entre 0 y 1, siendo 1 la que es eficiente, se utilizarán como resultados $1/\text{eficiencias}$

- **library(prcomp)**

Por defecto, la función *prcomp()* centra las variables para que tengan media de 0. Con el argumento *scale = TRUE* indicamos que queremos escalar las variables para que tengan desviación estándar igual a 1. Hay que pasarla un dataframe con los datos de los individuos y las variables.

En cuanto a las salidas, los elementos *center* y *scale* se corresponden con las medias y las desviaciones estándar originales de las variables previo escalado e implementación del ACP. La matriz *rotation* proporciona los vectores de cargas de las componentes principales (cada columna contiene el vector de cada componente principal).

La función los denomina la matriz de rotación ya que si multiplicáramos la matriz de datos por *datos\$rotation*, obtendríamos las coordenadas de los datos en el nuevo sistema rotado de coordenadas. Estas coordenadas se corresponden con los scores de las componentes principales.

- **library(factoextra)**

Paquete de R que facilita la extracción y visualización de los resultados de los análisis exploratorios de datos multivariados, que incluyen entre otros el análisis de componentes principales.

- **library(ggplot2)**

La librería *ggplot2* de R es un sistema organizado de visualización de datos. Forma parte del conjunto de librerías llamado *tidyverse*. En este documento se introduce su uso, principalmente a través de ejemplos. La primera parte muestra las características generales del sistema utilizando como ejemplo los diagramas de dispersión. En la segunda parte se detalla cómo representar algunos de los gráficos más conocidos. Los elementos necesarios para representar un gráfico con *ggplot2* son los siguientes:

- Un *data frame* que contiene los datos que se quieren visualizar.
- Los *aesthetics*, es decir, una lista de relaciones entre las variables del fichero de datos y determinados aspectos del gráfico (como por ejemplo coordenadas, formas o colores).
- Los *geoms*, que especifican los elementos geométricos (puntos, líneas, círculos, etc) que se van a representar.

Normalmente estos elementos se van añadiendo de forma consecutiva en distintas capas. Para añadir una nueva capa se usa el signo +.

El comando *ggplot* se usa para generar el sistema de coordenadas (por defecto, rectangulares) y posteriormente vamos añadiendo los geoms con sus correspondientes aesthetics. En principio los aesthetics se pueden asignar individualmente para cada geom.

ANEXO II:

Código TFG: Análisis / DEA / ACP

Cesar Ramos

ANALISIS UNIVARIANTE

```
datos <- read.table("C:/Users/cesar/Downloads/2º CUATRIMESTRE ( 4º )/TFG/DATOS.txt",header=TRUE)
```

```
nombres <- datos$Biblioteca
```

```
attach(datos)
```

Estudio:

- Resumen de distintos estadísticos
- Histograma
- Boxlot

```
par(mfrow=c(1,2))
```

```
# Personal:
```

```
summary(Personal)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.00  17.00   19.00   21.42  23.00   54.00
```

```
sd(Personal)
```

```
## [1] 9.604554
```

```
hist(Personal, col = "lightblue",main="Histograma", xlab="Valores", ylab="Frecuencia")
```

```
boxplot(Personal, col = "lightblue",main="Diagrama de cajas y bigotes", ylab="Valores")
```

```
# Usuarios activos:
```

```
summary(UsuariosActivos)
```

```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1379   3264   5235   5825   7133  25732
```

```
sd(UsuariosActivos)
```

```
## [1] 3981.658
```

```
hist(UsuariosActivos, col = "lightyellow",main="Histograma", xlab="Valores", ylab="Frecuencia")
```

```
boxplot(UsuariosActivos, col = "lightyellow",main="Diagrama de cajas y bigotes", ylab="Valores")
```

```
# Gastos adquisición bibliográfica:
```

```
summary(GastosAdqBibliografica)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      532  39549   48023   61458   61349   336681
```

```
sd(GastosAdqBibliografica)
```

```
## [1] 52163.94
```

```
hist(GastosAdqBibliografica, col = "#FFF0F5",main="Histograma", xlab="Valores", ylab="Frecuencia")
```

```
boxplot(GastosAdqBibliografica, col = "#FFF0F5",main="Diagrama de cajas y bigotes", ylab="Valores")
```

```
# Prestamos totales:
```

```
summary(Prestamos)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      186  31918   59397   64331   80206   339648
```

```
sd(Prestamos)
```

```
## [1] 50993.51
```

```
hist(Prestamos, col = "lightsalmon",main="Histograma", xlab="Valores", ylab="Frecuencia")
```

```
boxplot(Prestamos, col = "lightsalmon",main="Diagrama de cajas y bigotes", ylab="Valores")
```

```
# Nuevos usuarios:
```

```
summary(NuevosUsuarios)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       73    596    975    1400    1351    8668
```

```
sd(NuevosUsuarios)
```

```
## [1] 1523.876
```

```
hist(NuevosUsuarios, col = "#90EE90",main="Histograma", xlab="Valores", ylab="Frecuencia")
```

```
boxplot(NuevosUsuarios, col = "#90EE90",main="Diagrama de cajas y bigotes", ylab="Valores")
```

ANALISIS BIVARIANTE

Estudiamos:

- Diagrama de dispersión
- Correlaciones
- Regresión lineal

```
plot(Personal, UsuariosActivos, main = "Relación entre Personal y UA", xlab = "Personas trabajando", ylab = "Usuarios activos", col="blue",pch = 19)
cor(Personal, UsuariosActivos)
```

```
## [1] 0.6023782
```

```
abline(lm(UsuariosActivos ~ Personal))
```

```
plot(Personal, GastosAdqBibliografica, main = "Relación entre Personal y Gastos", xlab = "Personas trabajando", ylab = "Gastos", col="red", pch = 19)
cor(Personal, GastosAdqBibliografica)
```

```
## [1] 0.3340551
```

```
abline(lm(GastosAdqBibliografica ~ Personal))
```

```
plot(Personal, NuevosUsuarios, main = "Relación entre Personal y NU", xlab = "Personas trabajando", ylab = "Usuarios nuevos", col="green",pch = 19)
cor(Personal, NuevosUsuarios)
```

```
## [1] 0.1095022
```

```
abline(lm(NuevosUsuarios ~ Personal))
```

```
plot(Personal, Prestamos, main = "Relación entre Personal y Prestamos", xlab = "Personas trabajando", ylab = "Prestamos", col="purple",pch = 19)
cor(Personal, Prestamos)
```

```
## [1] 0.582092
```

```
abline(lm(Prestamos ~ Personal))
```

```
plot(UsuariosActivos, GastosAdqBibliografica, main = "Relación entre UA y Gastos", xlab = "UA", ylab = "Gastos", col="brown",pch = 19)
cor(UsuariosActivos, GastosAdqBibliografica)
```

```
## [1] 0.4282427
```

```
abline(lm(GastosAdqBibliografica ~ UsuariosActivos))
```

```
plot(UsuariosActivos, NuevosUsuarios, main = "Relación UA y NU", xlab = "Usuarios activos", ylab = "Nuevos usuarios", col="orange",pch = 19)
cor(UsuariosActivos, NuevosUsuarios)
```



```
## [1] 0.1668588
```

```
abline(lm(NuevosUsuarios ~ UsuariosActivos))
```

```
plot(UsuariosActivos, Prestamos, main = "UA y Prestamos", xlab = "UA", ylab = "Prestamos", col="darkgreen",pch = 19)  
cor(UsuariosActivos, Prestamos)
```

```
## [1] 0.8214071
```

```
abline(lm(Prestamos ~ UsuariosActivos))
```

```
plot(GastosAdqBibliografica, NuevosUsuarios, main = "Relación entre Gastos y NU", xlab = "Gastos", ylab = "Nuevos usuarios", col="aquamarine2",pch = 19)  
cor(GastosAdqBibliografica, NuevosUsuarios)
```

```
## [1] 0.113543
```

```
abline(lm(NuevosUsuarios ~ GastosAdqBibliografica))
```

```
plot(GastosAdqBibliografica, Prestamos, main = "Gastos y Prestamos", xlab = "Gastos", ylab = "Prestamos", col="#EEAD0E",pch = 19)  
cor(GastosAdqBibliografica, Prestamos)
```

```
## [1] 0.3645629
```

```
abline(lm(Prestamos ~ GastosAdqBibliografica))
```

```
plot(NuevosUsuarios, Prestamos, main = "Relación entre Nuevos usuarios y Prestamos", xlab = "Nuevos usuarios", ylab = "Prestamos", col="#CD1076",pch = 19)  
cor(NuevosUsuarios, Prestamos)
```

```
## [1] 0.234367
```

```
abline(lm(Prestamos ~ NuevosUsuarios))
```

ANALISIS CLUSTER

```
library(tidyverse)
```

```
library(cluster)
```

```
library(factoextra)
```

```
library(NbClust)
```

```
datosC <- datos
```

```
rownames(datosC) <- datos$Biblioteca
```

```
datosC <- datosC[,-1]
```

```

# Normalizamos
datosC <- scale(datosC)

# Calcular matriz de distancias
m.distancias <- get_dist(datosC, method = "euclidean")
fviz_dist(m.distancias, gradient = list(low="blue",mid="white",high="red"))

# Estimar el numero de clueters (Hay varios metodos)
fviz_nbclust(datosC, kmeans, method = "wss")

fviz_nbclust(datosC, kmeans, method = "silhouette")

fviz_nbclust(datosC, kmeans, method = "gap_stat")

resNumClust <- NbClust(datosC, distance="euclidean", min.nc = 2,max.nc=5, method = "k
means", index = "alllong")

# Realmente querría hacer 3 (Por arriba, por abajo y los del medio)
# Aunque estos metodos me dicen que la mejor opción sería utilizar 2

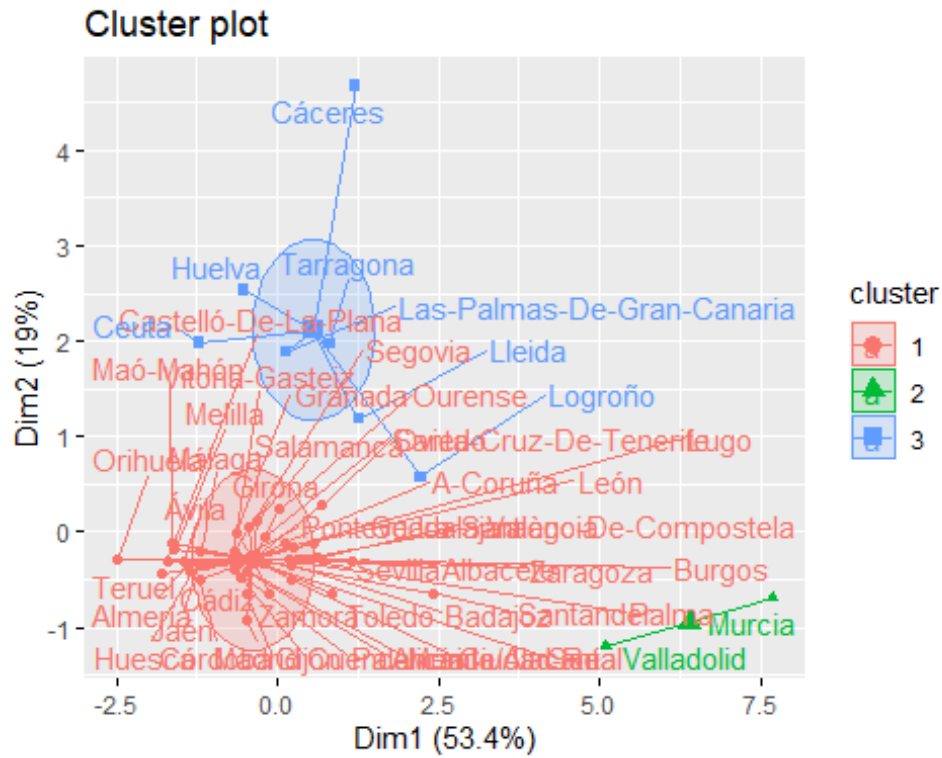
k3 <- kmeans(datosC, centers = 3, nstart = 25)
k3

# Graficamos
fviz_cluster(k3, data=datosC)

fviz_cluster(k3, data=datosC, ellipse.type = "euclid", repel = TRUE, star.plot = TRUE
)

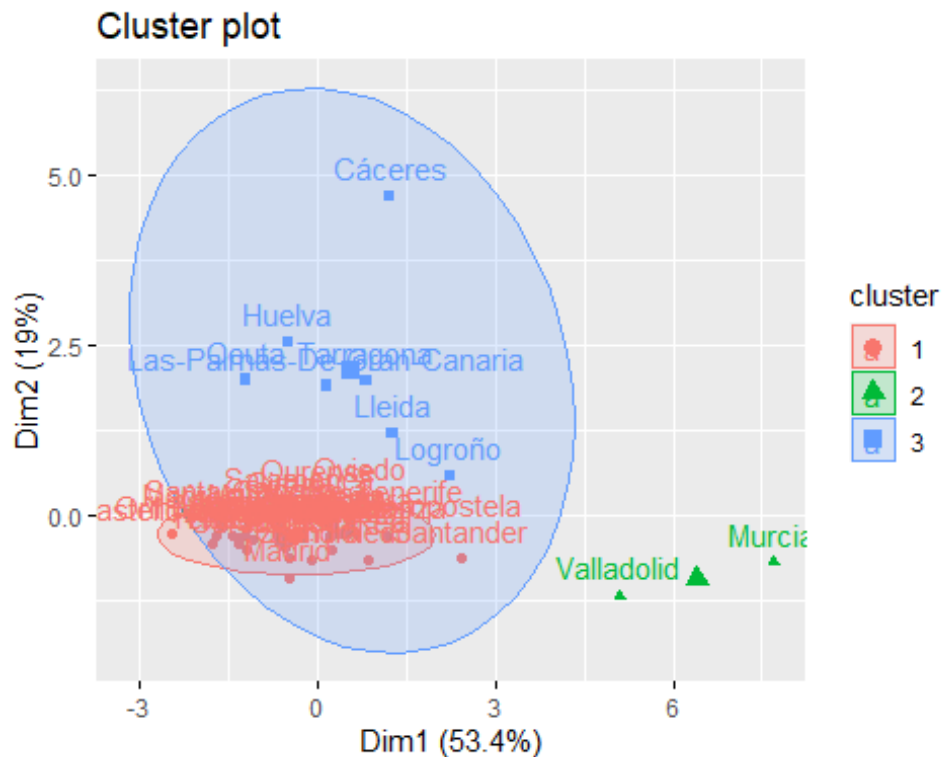
## Too few points to calculate an ellipse

```



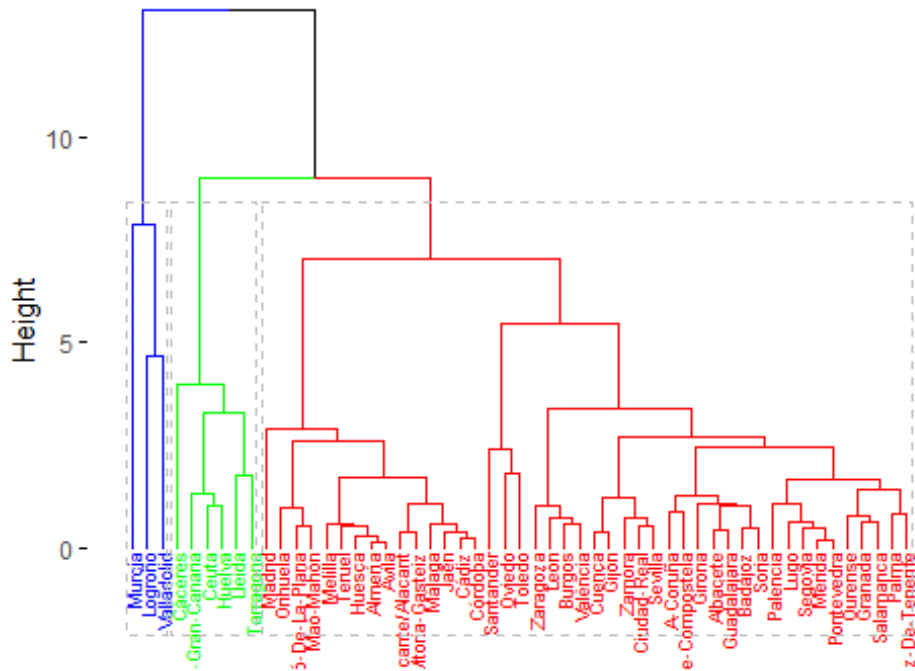
```
fviz_cluster(k3, data=datosC, ellipse.type = "norm")
```

```
## Too few points to calculate an ellipse
```



```
res3 <- hcut(datosC, k = 3, stand = TRUE)
fviz_dend(res3, rect = TRUE, cex = 0.5,
           k_colors = c("blue", "green", "red"))
```

Cluster Dendrogram



```
clus <- as.vector(k3$cluster)
```

```
datos$cluster <- clus
```

```
library(ggplot2)
library(magrittr)
```

```
library(dplyr)
```

```
ggplot(datos, aes(x = Personal, y = UsuariosActivos)) +
  geom_point()
```

```
variable1_outliers <- datos[datos$Personal %in% boxplot(datos$Personal, plot = FALSE)
$out, ]
```

```
variable2_outliers <- datos[datos$UsuariosActivos %in% boxplot(datos$UsuariosActivos,
plot = FALSE)$out, ]
```

```
ggplot(datos, aes(x = Personal, y = UsuariosActivos)) +
  geom_point() +
  geom_point(data = variable1_outliers, aes(x = Personal, y = UsuariosActivos), color
= "red", size = 2) +
  geom_point(data = variable2_outliers, aes(x = Personal, y = UsuariosActivos), color
= "blue", size = 2) +
  geom_text(aes(label = Biblioteca), vjust = -1, size = 2)+
  labs(title = "Outliers: rojo = Personal / azul = UA")
```

ANALISIS ENVOLVENTE DE DATOS (DEA)

Primero realizo una ejemplo de como será el análisis, utilizando todas las entradas y salidas:

```
library(deaR)

## Warning: package 'deaR' was built under R version 4.2.3

datosB <- datos[,-1]

# Transformo Los datos para que pueda aplicarse sobre ellos el algoritmo

datosDEA <- make_deadata(datosB,
                        inputs = 1:3,
                        outputs = 4:5)

# BCC output - orientado

modeloDEA <- model_basic(datosDEA,
                        orientation = "oo",
                        rts = "vrs")

eff <- eficiencias(modeloDEA)

eff_Tecnica <- round(1/eff,4) # entre 0 y 1

# Comprobamos cuantas DEAs son eficientes y cuantas no:

ef <- 0
inef <- 0

for(i in 1:length(eff_Tecnica)){

  if(eff_Tecnica[i]==1){
    ef <- ef+1
  }else{
    inef <- inef+1
  }

}

x <- c(rep("ineficiente",inef),rep("eficiente",ef))
x <- table(x)
```

Ahora generalizaré para todas las combinaciones lo visto:

```
# par(mfrow=c(3,7))

datosT <- list()

datosT[[1]] <- make_deadata(datosB,
                        inputs = 1,
                        outputs = 4)

datosT[[2]] <- make_deadata(datosB,
                        inputs = 1,
                        outputs = 5)

datosT[[3]] <- make_deadata(datosB,
                        inputs = 1,
                        outputs = 4:5)

datosT[[4]] <- make_deadata(datosB,
                        inputs = 2,
```

```

        outputs = 4)
datosT[[5]] <- make_deadata(datosB,
        inputs = 2,
        outputs = 5)
datosT[[6]] <- make_deadata(datosB,
        inputs = 2,
        outputs = 4:5)
datosT[[7]] <- make_deadata(datosB,
        inputs = 3,
        outputs = 4)
datosT[[8]] <- make_deadata(datosB,
        inputs = 3,
        outputs = 5)
datosT[[9]] <- make_deadata(datosB,
        inputs = 3,
        outputs = 4:5)
datosT[[10]] <- make_deadata(datosB,
        inputs = 1:2,
        outputs = 4)
datosT[[11]] <- make_deadata(datosB,
        inputs = 1:2,
        outputs = 5)
datosT[[12]] <- make_deadata(datosB,
        inputs = 1:2,
        outputs = 4:5)
datosT[[13]] <- make_deadata(datosB,
        inputs = c(1,3),
        outputs = 4)
datosT[[14]] <- make_deadata(datosB,
        inputs = c(1,3),
        outputs = 5)
datosT[[15]] <- make_deadata(datosB,
        inputs = c(1,3),
        outputs = 4:5)
datosT[[16]] <- make_deadata(datosB,
        inputs = 2:3,
        outputs = 4)
datosT[[17]] <- make_deadata(datosB,
        inputs = 2:3,
        outputs = 5)
datosT[[18]] <- make_deadata(datosB,
        inputs = 2:3,
        outputs = 4:5)
datosT[[19]] <- make_deadata(datosB,
        inputs = 1:3,
        outputs = 4)
datosT[[20]] <- make_deadata(datosB,
        inputs = 1:3,
        outputs = 5)

```

```

datosT[[21]] <- make_deadata(datosB,
                            inputs = 1:3,
                            outputs = 4:5)

info <- list()

for(j in 1:21){

  modeloDEAT <- list()

  modeloDEAT[[j]] <- model_basic(datosT[[j]],
                                orientation = "oo",
                                rts = "vrs")

  eff <- eficiencias(modeloDEAT[[j]])

  salidas <- round(1/eff,4) # entre 0 y 1

  info[[j]] <- salidas

  ef <- 0
  inef <- 0

  for(i in 1:length(salidas)){

    if(salidas[i]==1){
      ef <- ef+1
    }else{
      inef <- inef+1
    }

  }

  x <- c(rep("ineficiente",inef),rep("eficiente",ef))
  x <- table(x)

  # barp <- barplot(x, main = "Frecuencia absoluta",col = c("#FFF8DC","lightblue"), y
  lim = c(0,55))
  # text(barp, x + 2, labels = x)

  # etiquetas <- paste0(c("eficiente","ineficiente"), " = " ,c(ef,inef))
  # pi <- pie(x, col = c("#FFF8DC","lightblue"),labels = etiquetas)

}

```

Ahora vamos a obtener una tabla con los resultados:

```

TablaRes <- data.frame(row.names = datos$Biblioteca,
                      A_1 = info[[1]], A_2 = info[[2]], A_12 = info[[3]],
                      B_1 = info[[4]], B_2 = info[[5]], B_12 = info[[6]],
                      C_1 = info[[7]], C_2 = info[[8]], C_12 = info[[9]],
                      AB_1 = info[[10]], AB_2 = info[[11]], AB_12 = info[[12]],
                      AC_1 = info[[13]], AC_2 = info[[14]], AC_12 = info[[15]],
                      BC_1 = info[[16]], BC_2 = info[[17]], BC_12 = info[[18]],
                      ABC_1 = info[[19]], ABC_2 = info[[20]], ABC_12 = info[[21]])

TablaResRed <- round(t(TablaRes),2)

# write.table(TablaRes, "TablaResCEntender.txt")

v <- c()
for(j in 1:53){

  cont = 0
  for(i in 1:21){

```

```

    if(TablaRes[j,i]==1){
      cont = cont+1
    }
  }
  v <- c(v,cont)
}

vecesEfic <- data.frame("Biblios"=nombres,"vecesEf"=v)

minimo <- c()
for(j in 1:53){

  minimo <- c(minimo,min(TablaRes[j,]))

}
# minimo

maximo <- c()
for(j in 1:53){

  maximo <- c(maximo,max(TablaRes[j,]))

}
# maximo

```

ANÁLISIS DE COMPONENTES PRINCIPALES (ACP)

```

An <- read.csv("C:/Users/cesar/Downloads/ENTREGA TFG/TablaResCEntender.txt",header=TRUE)

# Estandarizamos Los datos:
A <- An[,2:22]

X0 <- scale(A)

```

ANÁLISIS DE LAS FILAS:

- Manualmente:

```

# Analisis de las covarianzas / correlaciones:

Y0 <- (t(X0)%*%X0)/(nrow(A)-1)

# Autovalores y autovectores:

eig <- eigen(Y0)
round(eig$values,3)

## [1] 12.079 4.361 1.688 1.498 0.583 0.338 0.191 0.124 0.052 0.049
## [11] 0.012 0.008 0.006 0.005 0.002 0.001 0.001 0.001 0.000 0.000
## [21] 0.000

# round(eig$vectors,3)
eig$vectors <- eig$vectors*-1

# Calculamos Los scores en Las 4 componentes y comprobar que tienen media 0 y varianz
a igual al autovalor correspondiente:

sco <- X0%*%eig$vectors

```



```

# Las medias de Los scores son 0, es decir que el grafico de Los puntos esta centrado
en 0

mediassco <- colMeans(sco)
mediassco <- apply(sco,2,mean)

# La suma de Los cuadrados de Los scores en cada componente divididos por n-1 (varianza)
son iguales a Los autovalores:

varianzassco <- apply(sco,2,var)

scofr <- as.data.frame(sco)

# La inercia total:

inerciatotal <- sum(diag(Y0))
sumaautoval <- sum(eig$values)

# Construir una matriz que tenga en La primera columna Los autovalores, en La segunda
el porcentaje de inercia explicado por cada autovalor, en La tercera Los autovalores a
cumulados hasta La fila correspondiente y en La cuarta el porcentaje acumulado:

a<-as.vector(1:nrow(Y0))

autoval<-eig$values

cumsumaut<-cumsum(eig$values)

autvperc<-100*eig$values/sum(eig$values)

cumperc<-cumsum(autvperc)

tabla<-cbind(a,autoval,autvperc,cumsumaut,cumperc)

colnames(tabla) <- c("Autovalor","Inercia","% Inercia","Inercia acumulada","% Inercia
Acumulada")

tabla <- round(tabla,4)

# Grafico scree plot con Los autovalores:

plot(a,autoval,type="l",col="blue",xlab="Inercia explicada", ylab="Autovalores")
abline(h=2,col="green")

# Grafico nombres:

plot(sco[,1],sco[,2], ylim = c(-5,6), col="red", xlim = c(-10,10),xlab="componente 1"
,ylab="componente 2")
abline(h=0, col="lightblue")
abline(v=0, col="lightblue")
text(sco[,1],sco[,2],labels=An$Bibliotecas,cex= 0.7, pos=3, col="black")

# Tabla con posiciones

posicionesXY <- data.frame("X-Componente1" = round(sco[,1],2),
                          "Y-Componente2" = round(sco[,2],2))
rownames(posicionesXY) <- nombres

datosC <- scale(posicionesXY)

# Grafico variables:
plot(eig$vectors[,1],eig$vectors[,2], col="red",xlim = c(0.17,0.27),ylim=c(-0.4,0.4))

```

```
text(eig$vector[1],eig$vector[2],labels=colnames(A^-1),cex= 0.7, pos=3, col="black")
```

```
# Contribuciones Absolutas:
```

```
ca<-t(t(sco^2)/((nrow(A)-1)*eig$values))
ca
```

```
casum<-colSums(ca)
casum
```

```
# Contribuciones relativas:
```

```
cr<-sco^2/rowSums(sco^2)
cr
```

```
crsum<-rowSums(cr)
crsum
```

```
cr12<-cr[,1]+cr[,2]
cr12
```

```
TablaContr <- data.frame("Biblioteca" = nombres,
  "CA"=round(ca[,1],4),
  "CR"=round(cr[,1],4))
```

```
cam <- cbind(round(ca[,1:2],3),round(cr[,1:2],3))
rownames(cam) <- datos$Biblioteca
colnames(cam) <- c("ca1","ca2","cr1","cr2")
cam <- as.data.frame(cam)
```

```
cam$sumcr <- cam$cr1 + cam$cr2
```

```
cam <- as.data.frame(cam[,-c(1:4)])
rownames(cam) <-datos$Biblioteca
colnames(cam) <- c("Suma CR")
```

```
library(ggplot2)
```

```
# Grafico de burbujas de Las contribuciones absolutas al primer eje
```

```
ggplot(data = scofr, aes(x = V1, y = V2, label = An$Bibliotecas)) +
  geom_hline(yintercept = 0, colour = "gray65") +
  geom_vline(xintercept = 0, colour = "gray65") +
  geom_point(aes(size = cr[,1]), alpha = 0.75, shape = 21, colour= "blue")+
  geom_text(colour = "black", alpha = 0.8, size = 3) +
  ggtitle("ACP Grafico de las Bibliotecas y ca al primer eje")
```

ACP Grafico de las Bibliotecas y ca al primer eje

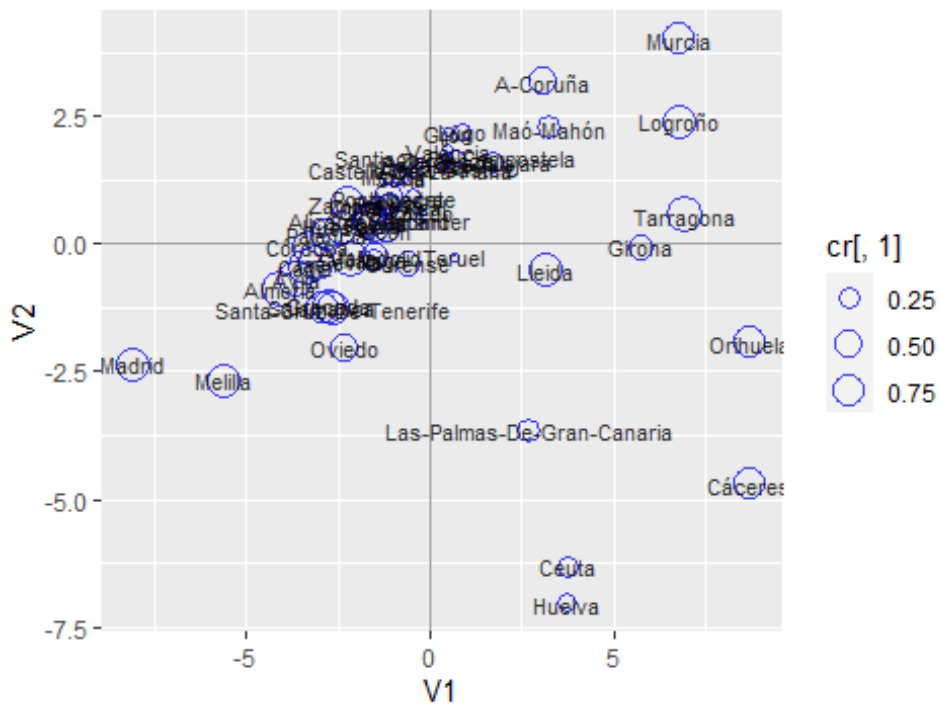
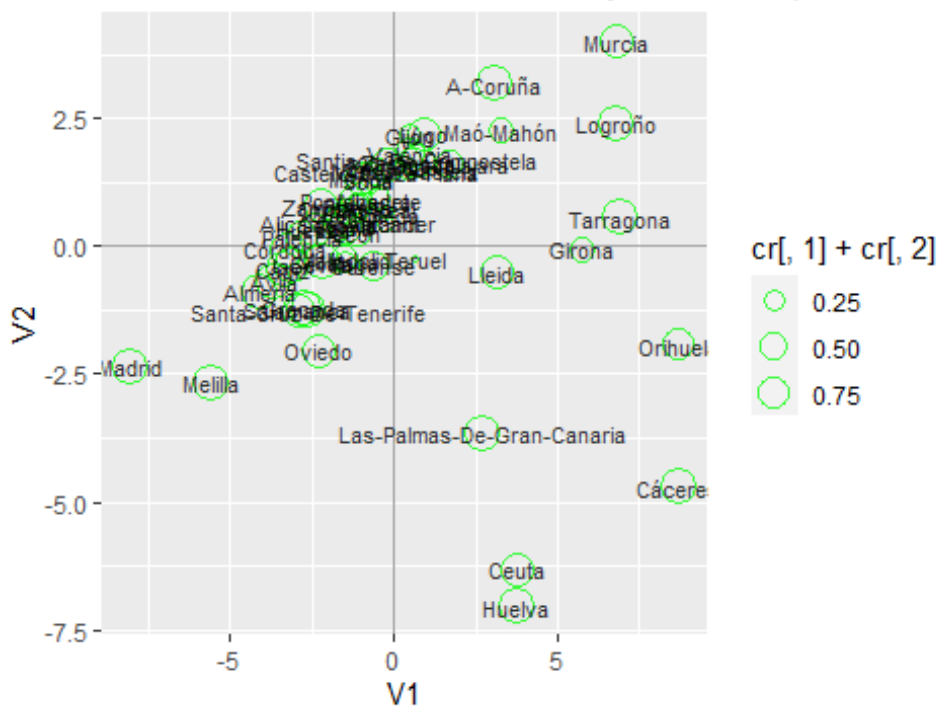


Grafico de burbujas de suma de Las contribuciones relativas a Los dos primeros ejes

```
ggplot(data = scofr, aes(x = V1, y = V2, label = An$Bibliotecas)) +
  geom_hline(yintercept = 0, colour = "gray65") +
  geom_vline(xintercept = 0, colour = "gray65") +
  geom_point(aes(size = cr[,1]+cr[,2]), alpha = 0.75, shape = 21, colour= "green")+
  geom_text(colour = "black", alpha = 0.8, size = 3) +
  ggtitle("ACP Grafico de las Bibliotecas y cr a los 2 primeros ejes")
```

ACP Grafico de las Bibliotecas y cr a los 2 primeros ej



- Prcomp:

```

library(ggplot2)
library(factoextra)

A <- scale(A)

pca1=prcomp(A, scale.=TRUE)

pca1$x[,1]<-pca1$x[,1]*-1
pca1$rotation[,1]<- pca1$rotation[,1]*-1

#Autovalores
pca1$sdev^2

#Autovectores
pca1$rotation

#Graficos
screepplot(pca1,col="lightblue")

tablaACP <- round(pca1$rotation,4)

# Ajustar Los valores mínimos y máximos a 0 y 1 respectivamente
# componente_principal_ajustada <- (PrimComp - min(PrimComp)) / (max(PrimComp) - min(
PrimComp))

# Resultado
# componente_principal_ajustada

biplot(pca1, scale = 0, cex = 0.5, col = c("dodgerblue3", "deeppink3"))

fviz_pca_var(pca1, col.var = "cos2",
             geom.var = "arrow",
             labelsize = 2,
             repel = FALSE)

# Los pesos de Las variables para Las dos primeras componentes:

pca1$rotation[,1]

# Funcion para crear un circulo
circle <- function(center = c(0, 0), npoints = 100) {
  r = 1
  tt = seq(0, 2*3.1416, length = npoints)
  xx = center[1] + r * cos(tt)
  yy = center[1] + r * sin(tt)
  return(data.frame(x = xx, y = yy))
}
corcir = circle(c(0, 0), npoints = 100)

library(FactoMineR)

## Warning: package 'FactoMineR' was built under R version 4.2.3

xx <-data(A)

```

```
## Warning in data(A): data set 'A' not found
PCA1 <- PCA(A[,1:10], scale.unit=TRUE, ncp=5, graph=T)
```

```
dataMH <- datos[c(33,27),]
```

- FactoMineR: Esta también se podría utilizar.

Realizaremos un tabla para sacar conclusiones:

```
TablaConclusiones <- data.frame("Bibliotecas" = nombres,
                               "Ranking" = pca1$x[,1])
TablaConclusiones[,2] <- round(TablaConclusiones[,2],3)
TablaConclusiones <- TablaConclusiones[order(TablaConclusiones$Ranking, decreasing =
TRUE), ]

indice <- data.frame("bibliotecas" = nombres,
                    "numero" = 1:53)
```