UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICA

En una encrucijada de Análisis Envolvente de Datos (AED) y Análisis de Componentes Principales (ACP)

Trabajo Especial de Grado presentado ante la ilustre Universidad Central de Venezuela por la **Br. Valeska La Torre** para optar al título de Licenciado en Matemática.

Tutor: Ricardo Ríos

Caracas, Venezuela

Junio, 2010

Nosotros, los abajo firmantes, designados por la Universidad Central de Venezuela como integrantes del Jurado Examinador del Trabajo Especial de Grado titulado "**En una Encrucijada de Análisis Envolvente de Datos (AED) y Análisis de Componentes Principales (ACP)**", presentado por la **Br. Valeska La Torre**, titular de la Cédula de Identidad **V-17.125.596**, certificamos que este trabajo cumple con los requisitos exigidos por nuestra Magna Casa de Estudios para optar al título de **Licenciado en Matemática**.

Ricardo Ríos
Tutor

Mairene Colina
Jurado

Angie Pineda
Jurado

Dedicatoria y Agradecimiento

Es importante agradecer todas y cada una de las cosas que nos suceden en la vida, desde las más simples y cotidianas hasta las más complejas y únicas.

Cada una de ellas deja una huella y un recuerdo en nosotros que nos hace crecer y ser mejores cada día.

También es importante agradecer a los que hacen posible que esas cosas sucedan. Por eso agradezco y dedico a:

Dios y San Antonio por darme fuerza espiritual, salud y paciencia.

Mi Familia por apoyarme y ayudarme en todo momento.

La Universidad Central de Venezuela y especialmente a la Escuela de Matemáticas, por darme la oportunidad de alcanzar una de mis metas y por enseñarme el sentido de compromiso y responsabilidad.

Los Amigos y Compañeros que me ayudaron, apoyaron y siempre estaban ahí para darme ánimos.

Mairene y Angie por sus valiosas correcciones que ayudaron a mejorar este trabajo.

Además quiero hacer un agradecimiento especial por su colaboración, asesoría, paciencia, orientación, ayuda y apoyo al profesor Ricardo Ríos.

Índice general

Introducción

Este trabajo centra su atención en el estudio de la cooperación que pueda existir entre el Análisis Envolvente de Datos (AED) y el Análisis de Componentes Principales (ACP).

En situaciones de la vida real, existen entidades productoras o unidades de producción o de servicio, llamadas Unidades de Toma de Decisión (UTD) tales como empresas, universidades, hospitales, industrias, personas, ciudades, países, etc., cada una de las cuales usan un conjunto de recursos (entradas) para generar un conjunto de resultados (salidas). Se considerará que todas las UTD a analizar, transforman el mismo tipo de entradas en salidas de la misma naturaleza, aunque por supuesto varía, de una UTD a otra, la cantidad utilizada de cada tipo de entrada y la cantidad en que se transforma para cada salida. Este conjunto de UTD forma un Conjunto de Referencia.

Resulta conveniente medir la eficiencia de las UTD, es decir, analizar cómo conocer el desempeño global de las UTD a través de sus entradas y salidas. Para ello, sería deseable poder resumir cada conjunto de entradas y salidas en una medida. A tal fin se adopta como medida de la eficiencia el cociente de la suma ponderada de salidas sobre la suma ponderada de entradas. Para obtener esta medida única, es necesario obtener pesos o ponderaciones que hagan sumables las entradas y salidas entre sí.

Existen varias formas posibles de determinar estos pesos o ponderaciones, entre las que se encuentran: Análisis de regresión, para lo cual se hace necesario postular una estructura matemática para la relación entre las entradas y salidas (técnicas paramétricas); Métodos de estadística multivariada: Análisis de Correlaciones, Análisis Discriminante y Análisis de Componentes Principales; Modelos de programación matemática, como el Análisis Envolvente de Datos (AED).

La metodología de la programación matemática AED, es un método no paramétrico de optimización de funciones lineales, para medir la eficiencia relativa de las UTD. El AED proporciona una única medida para cada UTD en términos de la utilización de las entradas para generar las salidas deseadas.

El Análisis de Componentes Principales (ACP) es una técnica de reducción de dimensión, que se usa cuando se toman muchas variables, pues en la mayoría de los casos dichas variables están relacionadas y conviene reducir el número de éstas, identificando un conjunto pequeño de variables que explique una porción grande de la varianza total de las variables originales. El ACP es también un método de ordenamiento en el análisis multidimensional.

Teniendo en cuenta estas características del AED y del ACP, el objetivo de este trabajo es aplicar AED y ACP a un caso real, con la finalidad de verificar, la consistencia de los resultados obtenidos por ambas metodologías y proponer posteriormente un método integrado que tome lo mejor de cada una.

En la situación que estudiaremos se clasificarán las UTD para así refinar las políticas medioambientales relacionadas a la incidencia del melanoma, ya que en algunos países como Nueva Zelanda, Australia y los de Europa del Norte, donde las muertes debidas al cáncer de melanoma eran raras, tienen ahora mayor índice de mortalidad por este cáncer, siendo el cuarto tipo de cáncer más común [**11**]. La idea es encontrar una política medioambiental universal apropiada para proteger la capa de ozono y proponer medicinas para reducir la absorción de rayos ultravioletas A y B. Estas políticas se apuntan a reducir la tasa de mortalidad de melanoma y por lo tanto, aumentar la tasa de supervivencia del cáncer de piel o melanoma.

Este Trabajo está organizado de la siguiente forma: el Capítulo 1 presenta el Análisis Envolvente de Datos (AED) y la descripción del modelo BCC utilizado, propuesto por Banker, Charnes y Cooper; El Capítulo 2 desarrolla el método estadístico multivariado: Análisis de Componentes Principales (ACP) para medir el desempeño global de las UTD y buscar las relaciones entre las variables de entradas y salidas; En el Capítulo 3 se comparan los dos

procedimientos (AED y ACP) y se propone un método de integración AED-ACP; el Capítulo 4 presenta la aplicación del AED, el ACP y del método AED-ACP y su respectivo análisis de resultados; el último Capítulo presenta las conclusiones y recomendaciones del trabajo.

CAPíTULO 1

Análisis Envolvente de Datos(AED)

1. Introducción

En este primer capítulo se aborda lo relacionado al Análisis Envolvente de Datos (AED), sus orígenes, fundamentos, modelos, ventajas y desventajas.

El Análisis Envolvente de Datos (AED) es una técnica de medición de la eficiencia de unidades productivas (UTD) basada en la programación lineal o programación matemática.

En el AED la organización o ente a estudiar se denomina Unidad de Toma de Decisiones (UTD). Genéricamente una UTD es considerada como la entidad responsable de convertir las entradas en salidas y cuyas actuaciones han de ser evaluadas. En las aplicaciones de gestión, UTD pueden incluir bancos, departamentos, tiendas y supermercados, y se extienden a los fabricantes de automóviles, hospitales, escuelas, bibliotecas públicas, etc. En ingeniería, las UTD puede adoptar la forma de aviones o de sus componentes como los motores a reacción. A los efectos de asegurar la comparación relativa, un grupo de las UTD se utiliza para evaluar con cada UTD teniendo un cierto grado de libertad de gestión en la toma de decisiones.

2. Orígenes

Si bien no es extraño encontrar autores como Seiford [**15**] que afirman que el modelo del AED fue desarrollado por primera vez en el año 1978 por Charnes, Cooper y Rhodes [**7**]; existen otros como Charnes [**4**], que afirman que el origen de esta técnica es debido a Rhodes, quien, en 1978 la aplicó en su tesis doctoral dirigida por W.W. Cooper en el análisis de eficiencia del programa de educación Follow-Through de las escuelas públicas de los Estados Unidos.

Fundamentalmente, este método sigue los conceptos básicos de eficiencia y su medición propuestos por Farrell en 1957, quien define la eficiencia como producir lo máximo posible a partir de unos recursos (entradas) dados. No obstante, junto con este autor, hubo otros que proporcionaron los fundamentos necesarios para que el AED pudiera surgir y fuese utilizado, como Charnes y Cooper [**5**], Aigner y Chu [**3**] y Afriat [**2**].

En 1968, Aigner y Chu trataron de continuar el trabajo pionero de Farrell. Tras distinguir entre diversos conceptos de función de producción que hacían complicado su entendimiento, se dedicaron a completar el trabajo de Farrell en aquellos aspectos en los que este autor no logró ser lo suficientemente genérico (por ejemplo, en la estimación de la frontera eficiente, formada por el conjunto de unidades eficientes) utilizando métodos de programación matemática (Aigner y Chu [**3**]). Sin embargo, cabe mencionar que aunque estos autores perseguían una generalización del método de Farrell, introdujeron también un elemento restrictivo al considerar relaciones funcionales predeterminadas entre las entradas y las salidas, ya que una de las características que hacían amplia a la idea de Farrell era la no necesidad de considerar previamente una forma específica de función de producción (función que relaciona las entradas con las salidas). Además, un problema adicional con su procedimiento es que los test habituales de significatividad de las estimaciones se basaban en supuestos muy restrictivos acerca de las perturbaciones (Dunlop [**9**]).

Un enfoque diferente fué el que realizó Afriat en 1972 al desarrollar un método de análisis de la producción que, en la filosofía de Farrell, evitaba la consideración de especificaciones concretas de la función de producción y al mismo tiempo, convertía al método de Farrell en un caso particular (Afriat [**2**]). De todos modos, este autor basó su análisis en consideraciones específicas acerca de determinadas propiedades (no decrecimiento, concavidad, etc.) que deben tener las funciones de producción.

En definitiva todos los precedentes comentados generarían un método que compara entre sí UTD homogéneas respecto a entradas y salidas, dando así una medida de la eficiencia relativa. La eficiencia relativa de cada UTD es calculada computando la razón definida por el cociente entre la suma ponderada de las salidas y la suma ponderada de las entradas,

siendo los pesos calculados en función de criterios paretianos y considerando que la eficiencia de ninguna entidad puede superar la unidad (Charnes [**4**]). El Criterio o Eficiencia de Pareto considera que: "en todo sistema de asignación de recursos existe un punto óptimo en el cual ninguna entidad puede mejorar sin que otra empeore".

3. Fundamentos

Una vez expuestos sus orígenes, se continuará con sus fundamentos. Para ello, primero se procederá a exponer una serie de conceptos relacionados con la filosofía del modelo y después se comentarán los detalles del mismo.

El AED es una técnica de medición de la eficiencia basada en la obtención de una frontera de eficiencia a partir del conjunto de observaciones que se considere sin la estimación de ninguna función de producción, es decir, sin necesidad de conocer ninguna forma de relación funcional entre las entradas y las salidas. Es definitivamente una alternativa para extraer información de observaciones frente a los métodos paramétricos cuyo objetivo es la obtención de un hiperplano que se ajuste lo mejor posible al conjunto de observaciones. El AED, por el contrario, trata de optimizar la medida de eficiencia de cada unidad analizada para crear así una frontera eficiente basada en el criterio de Pareto (Charnes [**4**]). De este modo, primero se construye la frontera de producción empírica y después se evalúa la eficiencia de cada unidad observada que no pertenezca a la frontera de eficiencia. Así, además de no ser un método paramétrico (por no presuponer la existencia de una función que relacione entradas con salidas) tampoco es estadístico puesto que no asume que la eficiencia no captada siga algún tipo de distribución probabilística.

De cara al proceso de evaluación, se considera que una unidad productiva es eficiente y por tanto, que pertenece a la frontera de producción, cuando produce más de alguna salida sin generar menos del resto y sin consumir más recursos o entradas, o bien, cuando utilizando menos de alguna entrada, y no más del resto, genere los mismos productos (salidas).

Lo anterior explicaría el tipo de elementos que componen la frontera eficiente, pero deja sin aclarar cómo evaluar las UTD que no formen parte de ella. La idea es comparar cada

unidad no eficiente con aquélla que lo sea y a la vez, tenga una técnica de producción similar; es decir, que utilice entradas similares para producir salidas parecidas.

No necesariamente debe ocurrir que esa entidad, eficiente y homogénea técnicamente con la evaluada, deba tener su reflejo en la realidad. Puede ocurrir (de hecho es lo usual) que la unidad con la que se compare la entidad en evaluación no sea real sino una combinación lineal de otras existentes. El conjunto de unidades reales eficientes combinadas para generar otra unidad eficiente, pero ficticia, se denomina grupo o *Conjunto de Referencia* y su identificación permite planificar las mejoras de las UTD ineficientes sobre la base de niveles efectivamente alcanzados.

En cualquier caso, para medir la eficiencia de una unidad hay dos opciones. La primera, comprobar la cantidad de entradas utilizadas para obtener las mismas salidas (orientación a la entrada) y la segunda, lograr el máximo número de salidas manteniendo las entradas (orientación a la salida). Escoger una vía u otra debe depender de las características concretas del problema a analizar.

A continuación se expone el modelo de Análisis Envolvente de Datos con el cual se trabajará. Para detalles en relación a la Programación Lineal (PL) ver Apéndice 1.

4. Modelo Básico

La esencia del AED es la siguiente: Sea $\theta_i \in [0,1]$ la eficiencia operacional o medida de la eficiencia de la *i-ésima* UTD. La eficiencia operacional es el cociente de la suma ponderada de las salidas sobre la suma ponderada de las entradas. Los datos de entrada y salida se introducen en la estructura de Programación Lineal (PL) como en (1). Ambas variables de datos, entradas y salidas, se suponen deterministas en el AED. Con:

$$\underline{y} \;=\; \left(\; \underline{y}_1,\;\; \underline{y}_2,\;\; \ldots,\;\; \underline{y}_n \;\right) = \begin{pmatrix} y_{11} & \cdots & y_{1n} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mn} \end{pmatrix}$$

$$\underline{x} \overset{y}{=} \left(\underline{x}_1, \ \underline{x}_2, \ \ldots, \ \underline{x}_n \right) = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{k1} & \cdots & y_{kn} \end{pmatrix}$$

denotando las matrices de m salidas y k entradas, respectivamente, de las n UTD, sus eficiencias operacionales $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_n)$ con $\theta_i \in [0, 1]$, se obtienen encontrando la solución óptima en PL del Problema Primal (PP):

Maximizar: θ_i

$$\text{sujeto a:} \begin{cases} (11\ldots1)\underline{\lambda}' = 1, \\ \underline{x\lambda'} \leq \underline{x}_i, \\ \underline{y\lambda'} \geq \theta_i \underline{y}_i, \\ \underline{\lambda} \geq 0, \end{cases} \quad (1)$$

donde $\underline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ denota los pesos óptimos a ser escogidos de los datos de entrada-salida y λ' denota su traspuesto. El Modelo de AED utilizado en (1) es el propuesto por Banker, Charnes y Cooper (Modelo BCC). El proceso para encontrar la solución óptima de (1) se transforma empleando el Dual de (1), que es el siguiente:

Minimizar: $\underline{z} = \underline{vx}_i - v$

$$\text{sujeto a:} \begin{cases} \underline{uy}_i = 1, \\ \underline{uy} + \underline{v}_i(11\ldots1)' \leq \underline{vx}, \\ \underline{v} \geq 0, \underline{u} \geq 0, \ | \ \underline{v}_i \ | > 0, \end{cases} \quad (2)$$

donde $\underline{u}, \underline{v}, \underline{x}$ y $\underline{v}_i$ son llamados pesos de salida, pesos de entrada, un valor escalar y una constante arbitraria asociada con $(11\ldots1)\underline{\lambda}' = 1$, respectivamente. La solución óptima debe satisfacer simultáneamente a (1) y a (2). Los valores:

$$\underline{s}^- = \underline{x}_i - \underline{x\lambda}^* \geq \underline{0}$$

$$\overset{y}{=}$$

$$\underline{s}^+ = \underline{y\lambda}^* - \underline{y}_i \geq \underline{0}$$

son los vectores de *excesos de entradas* en R^m y *deficit de salidas* en R^k donde el $*$ indica el valor óptimo. Cuando el score θ_i es uno, la *i-ésima* UTD es considerada eficiente, por otra parte, cuando el score $\theta_i < 1$, la *i-ésima* UTD es ineficaz. El conjunto de referencia para cada unidad ineficaz puede identificarse con un valor.

El proceso de formular la PL y encontrar una solución óptima debe ser repetido para cada una de las UTD. En realidad, no hay dificultades a nivel computacional para realizar este proceso, sólo se deben modificar las instrucciones adecuadamente.

Por último, cabría realizar una breve exposición de las ventajas e inconvenientes que posee la aplicación de la técnica AED.

5. Ventajas y Desventajas del AED

El AED presenta una serie de características en su metodología que la han convertido en una técnica muy utilizada. Charnes, Cooper, Lewin y Seiford [**6**] destacan las siguientes tres características como ventajas:

(1) Caracteriza cada una de las Unidades mediante una única puntuación de eficiencia (relativa).
(2) Al proyectar cada Unidad ineficiente sobre la frontera eficiente destaca áreas de mejora para cada una de las Unidades.
(3) La no consideración por el AED de la aproximación alternativa e indirecta de especificar modelos estadísticos y hacer inferencias basadas en el análisis de residuos y coeficientes de los parámetros.

Además de las tres características anteriores, Charnes, Cooper, Lewin y Seiford [**6**] aportan otras peculiaridades sobre el AED como por ejemplo, la posibilidad de ajustarse a variables exógenas, es decir, variables que están fuera del modelo utilizado; e incorporar variables categóricas o cualitativas, por ejemplo: nombres.

Otro aspecto resaltante del AED es su capacidad de manejar situaciones con múltiples entradas y salidas expresados en distintas unidades de medidas. Además el AED es una técnica no paramétrica, y por lo tanto, no supone ninguna forma funcional de relación entre las entradas y salidas, ni supone una distribución de la ineficiencia.

Continuando con las limitaciones o desventajas que presenta la técnica del AED, una de las mayores críticas recibidas es que se trata de una aproximación determinista y no tiene en cuenta influencias sobre el proceso productivo de carácter aleatorio, ni la incertidumbre (errores de medida o introducción incorrecta de datos). Así la precisión de los resultados alcanzados dependerá de la exactitud de las medidas de las Entradas y las Salidas consideras.

Además el AED es sensible a la existencia de observaciones extremas y toda desviación respecto de la frontera es tomada como ineficiente, lo que puede llevar a una sobreestimación de la misma.

CAPíTULO 2

Análisis de Componentes Principales (ACP)

1. Introducción

En este capítulo se presentan la metodología, ventajas y desventajas del método estadístico ACP.

El Análisis de Componentes Principales (ACP) es una técnica utilizada para la reducción de la dimensionalidad de un conjunto de datos. La idea del ACP es estudiar las relaciones entre un gran número de variables y explicar dichas variables en términos de un grupo menor de variables con una pérdida mínima de información. Estas "nuevas"variables se denominan Componentes Principales (CP), las mismas no están correlacionadas y en caso de que sigan una distribución normal multivariante, serán además independientes.

2. Metodología del ACP

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 la Primera 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 covarianza 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 los datos. Además las coordenadas en la nueva base dan la composición en factores subyacentes de los datos iniciales.

El ACP transforma p observaciones correlacionadas en $q \leq p$ scores de componentes principales. La mecánica del ACP puede resumirse de la siguiente manera:

Sea $\underline{D}_i = (D_{1i}, D_{2i}, \ldots, D_{pi})$ el vector que denota las p observaciones de variables de datos, tanto de entrada como de salida, para la *i-ésima* UTD, siendo $p \geq 1$ la dimensión con $i = 1, 2, \ldots, n$. Note que $\underline{D} = (\underline{D}'_1, \underline{D}'_2, \ldots, \underline{D}'_n)$ es una matriz de p filas y n columnas, donde $\underline{D}'_i$ es la traspuesta de $\underline{D}_i$ para cada $i = 1, 2, \ldots, n$. Cada vector $\underline{D}_i$ de p observaciones se supone con distribución Gaussiana $N_p(\underline{\mu}, \Sigma)$ de dimensión p, donde $\mu' = (\mu_1, \mu_2, \ldots, \mu_p)$ denota el vector de la media y

$$\Sigma = \begin{pmatrix} var(D_1) & cov(D_1, D_2) & \cdots & cov(D_1, D_p) \\ cov(D_2, D_1) & var(D_2) & \cdots & cov(D_2, D_p) \\ \vdots & \ddots & \vdots & \\ cov(D_p, D_1) & cov(D_p, D_2) & \cdots & var(D_p) \end{pmatrix}$$

denota una matriz de covarianza simétrica y positiva definida. Además la probabilidad máxima estimada de la media, varianza y parámetros de covarianza son: la media, la varianza y la covarianza muestrales, denotadas por:

$$\begin{aligned} \overline{\underline{D}} &= (\overline{D}_1, \overline{D}_2, \ldots, \overline{D}_p), \\ s_{jj} &= \sum_{i=1}^{n} (D_{ji} - \overline{D}_j)^2/(n-1), \\ s_{jl} &= \sum_{i=1}^{n} (D_{ji} - \overline{D}_j)(D_{li} - \overline{D}_l)/(n-1), j \neq l, \\ \overline{D}_j &= \sum_{i=1}^{n} D_{ji}/n, \\ j &= 1, 2, \ldots, p. \end{aligned}$$

Debido a que la matriz de covarianza es positiva definida, existe una matriz ortonormal $\underline{C} = (\underline{C}_1, \underline{C}_2, \ldots, \underline{C}_q)$ formada por q autovectores, cada uno con n filas, que cumple $\underline{C}'\underline{C} = \underline{I}$, que es una matriz unitaria, es decir, una matriz con 1 en la diagonal y 0 en las posiciones fuera de la diagonal. Debido a las bases ortonormales, el q autovector (vector propio) puede colocarse de manera que una versión ordenada de ellos pueda crearse. Las versiones ordenadas

se llaman **Componentes Principales (CP)**. Es decir,

$$\underline{CP}'_l = (c_{i1}, ci2, \ldots, c_{ip})\underline{D}'$$

es el vector de la *l-ésima* Componente Principal tal que:

$$\begin{aligned} var(\underline{CP}'_l) &= v_{(ll)}, \\ cov(CP_j, CP_l) &= \left\{ \begin{array}{lll} 0 & si & j \neq l \\ v_{(jj)} & si & j = l \end{array} \right\}, \\ l &= 1, 2, \ldots, q. \end{aligned}$$

donde $v_{(11)} \geq v_{(22)} \geq \ldots v_{(qq)} \geq 0$ son las dispersiones ordenadas de la CP q. Note que los scores de la CP son diferentes de los autovectores.

Las CP no sólo son no correlacionas sino que también son independientes si las variables de los datos siguen una distribución Gaussiana multivariante. En otras palabras,cuando las variables de los datos no tienen distribución Gaussiana, las q CP serían no correlacionadas, pero no necesariamente independiente. Note que la $\underline{CP_1}$ tiene la dispersión más grande, la segunda CP, $\underline{CP_2}$ tiene la segunda dispersión más grande, y así sucesivamente. Sólo un significante y óptimo $q*$ número de CP serán consideradas para la extensa discusión basada en el porcentaje deseado de variaciones explicado. En otras palabras, el primer número $q*$ significante de Componentes Principales captura un deseable $100 \sum_{j=1}^{q} v_{jj} / \sum_{j=1}^{p} v(jj)$ porcentaje de la dispersión total que existió en los datos originales.

Las q^* Componentes Principales son estocásticamente independientes unas a otras, es decir, independientes en el sentido aleatorio o probabilístico (CP_i y CP_j son independientes para $i \neq j$) con una suposición de que la distribución subyacente es una Gaussiana multivariante de dimensión q^*. Estos scores de Componentes Principales podrían usarse, en un sentido óptimo, en lugar de todas las observaciones originales de las variables de datos para clasificar las UTD. Esta práctica ofrecería algunas ventajas y unas desventajas. Una desventaja es que el ACP no debe realizarse con una mezcla de variables de datos de entrada y datos de salida. Una ventaja es que los valores extremos entre UTD se puede identificar.

Sin embargo, el método del ACP ofrece una oportunidad de identificar cualquier influencia extraordinaria en las UTD. Note que un score de la Componente Principal $CP_{l,j}$, sigue una distribución Gaussiana con alguna media desconocida μ y varianza σ^2. Por consiguiente, la suma de los scores $\sum_{j=1}^{n} CP_{l,j}$ sigue una distribución Gaussiana con media $n\mu$, y dispersión $n\sigma^2$. Con la condición dada sobre la suma $\sum_{j=1}^{n} CP_{l,j} = t$, el score $CP_{l,j}$ sigue una distribución Gaussiana con función de densidad de probabilidad:

$$\begin{aligned} f(x|\sum_{j=1}^{n} CP_{l,j}) &= \frac{f(x|\mu_l, \sigma_l^2) f(t-x|[n-1]\mu_l, [n-1]\sigma_l^2)}{f(t|n\mu_l, n\sigma_l^2)} \\ &\approx (\sigma_l[1-n^{-1}]\sqrt[2]{2\pi})^{-1} \exp^{-[x-\mu_l]^2/2[1-n^{-1}]\sigma_l^2} \end{aligned}$$

En consecuencia, la media condicional es μ y la varianza condicional es $[1-n^{-1}]\sigma_l^2$. Sus estimaciones de probabilidad máximas son:

$$\begin{aligned} \widehat{\mu_l} &= \sum_{j=1}^{n} CP_{l,j}/n \\ &\quad y \\ \widehat{\sigma_l}^2 &= (n-1)^{-1} \sum_{j=1}^{n} (CP_{l,j} - \widehat{\mu_l})^2. \end{aligned}$$

Se puede estandarizar los scores de las Componentes Principales como sigue:

$$z_{l,j} = \frac{CP_{l,j} - \widehat{\mu_l}}{(1-n^{-1})\widehat{\sigma_l}}$$

Si el score $z_{l,j}$ cae fuera del intervalo $(-z_{\alpha/2}, z_{\alpha/2})$, entonces la *j-ésima* UTD se considera inusualmente influenciada por la *l-ésima* Componente Principal.

3. Ventajas e Inconvenientes del ACP

Una de las principales ventajas del ACP es que permite capturar fácilmente las primeras variables importantes y luego el siguiente conjunto importante de variables de datos, y así sucesiva y jerárquicamente. Esto es porque la primera componente principal es la mejor combinación lineal que captura la máxima cantidad de dispersión de los datos. Los coeficientes de la primera CP revelan la primera colección de mejores variables de datos "influenciadas". Entonces, la segunda colección de próximos "mejores"variables de datos influenciadas se identifica en la segunda CP, y así sucesivamente. De esta manera, las variables son racimos basados en su importancia. Las variables en racimo son visibles en la gráfica de los

primeros y segundos scores de las Componentes Principales.

Dentro de cada una de las $q*$ Componentes Principales, pueden calcularse una media condicional esperada y varianza de un score para una suma dada de los scores en esa CP. Los scores se estandarizan a z valores usando la media y varianza condicionales. Cuando un z valor cae fuera de un rango normal $(-z_{\alpha/2}, z_{\alpha/2})$ para un $\alpha \in (0, 1)$ seleccionado, la pertinente UTD se afirma que tiene una alta influencia. Este rasgo de identificar la alta influencia de la unidad es un conocimiento extra.

Con datos complejos y grandes, la comprensión de la información se hace más difícil. He aquí la importancia de las gráficas, ya que explican más que mil palabras. El primer y segundo score de CP para cada UTD puede utilizarse para hacer una gráfica. Esta gráfica bidimensional muestra la configuración de los racimos de las UTD. Las cercanías entre los racimos de las UTD puede ser interpretadas fácilmente por el ACP.

CAPíTULO 3

Comparación e Integración del AED y el ACP

1. Introducción

Luego ver las características particulares, ventajas y desventajas del AED y ACP, en este tercer capítulo se compararán ambos métodos para observar sus semejanzas y diferencias, posteriormente integrarlos para obtener un método que supere las debilidades de cada uno de estos.

2. Similitudes y Diferencias

Al comparar el AED y ACP encontramos algunas similitudes y otras diferencias, entre las cuales se pueden destacar las presentadas por Shanmugam, R. y Johnson C.[**14**]:

- En el ACP las unidades (UTD) son seleccionadas aleatoriamente y se suponen estocásticas, de hecho, distribuidas como variables aleatorias Gaussianas. Además, en el ACP, las unidades representan a otras unidades no seleccionadas de la población. A diferencia de esto, en el AED las UTD no son seleccionadas aleatoriamente, no tienen que ser estocástica y constituyen la totalidad de las unidades.

- El AED requiere que se distingan las variables de Entrada de las variables de Salida, requisito este que no es necesario para el ACP.

- Con respecto a las Correlaciones entre las Variables (de Entrada-Salida), éstas son básicas para el ACP, sin ellas no tendría sentido realizarlo. Las CP, por supuesto, no están correlacionadas. Las correlaciones son irrelevantes para el AED.

- El ACP es una técnica empleada generalmente para disminuir la complejidad de los datos multivariantes, puesto que reduce la dimensionalidad. El AED, en cambio no reduce la dimensión de los datos, por lo que se pensaría que cuando los datos son

demasiado grandes, el AED no resultara suficientemente rápido debido a la enorme cantidad de cómputos. Pero, actualmente existen software de AED que resuelven un conjunto grande de datos en un tiempo razonable.

- El AED relaciona los datos de entrada y los de salida para clasificar óptimamente la eficiencia operacional de las UTD. El ACP no clasifica las UTD de la misma manera como lo hace el AED. Sin embargo, pueden extenderse los conceptos del ACP para clasificar las UTD como se verá más adelante.

- Una similitud y ventaja de ambos métodos es que, tanto en el AED como en el ACP, las variables de los datos pueden ser negativas, cero, o positivas.

Luego de estudiar las diferencias y similitudes ente el AED y el ACP, se pasará a describir un método de integración (AED-ACP) para superar las dificultades y mejorar las fortalezas de cada uno.

3. Método AED y ACP Integrados

Dependiendo de la orientación hacia la programación matemática o pensamiento estadístico, las personas emplean el AED o ACP, respectivamente. La mayoría de los libros y documentos de AED no menciona ACP, y viceversa. Incluso uno de los libros más importantes de AED Cooper-Seiford-Tone [**8**] sólo menciona el ACP una vez en una ilustración del caso.

AED y ACP tienen, como se vio en la parte anterior, fortalezas y debilidades en sí mismos. El ACP proporciona una conveniente reducción de los datos pero exige que los datos deben tener una distribución Gaussiana. Al contrario, el AED es más bien un procedimiento lento ya que requiere de un análisis separado para clasificar cada UTD, pues el proceso de formular la PL y encontrar su solución óptima debe ser repetido para cada una de las UTD. Varios software de AED (incluso uno en [**8**]) hacen los cómputos automáticamente y proporcionan extensas y excelentes salidas en menos de un fragmento de un minuto. Una pregunta natural para hacerse es: ¿por qué no combinar ambas técnicas para obtener lo mejor de ambas?.

Si la respuesta es afirmativa, entonces ¿cómo la integración debe funcionar para lograr una técnica superior?. En esta parte se intentará responder estas preguntas.

El método de integración que se verá es el propuesto en 2007 por Shanmugam y Johnson [**14**]. Otros autores han intentado realizar integraciones de ese tipo como Adler y Golany [**1**] quienes intentaron aplicar AED y ACP para evaluar las aerolíneas liberalizadas; Zhu [**16**] y Premachandra [**13**] aplicaron una integración para evaluar el rendimiento económico de las ciudades chinas. Desafortunadamente, algunas fallas serias existen en sus planteamientos, los cuales se señalan a continuación.

Como se mencionó anteriormente, el ACP exige que las variables de datos sigan una distribución Gaussiana multivariada. Si es así, entonces cada uno de los datos de entrada, así como los datos de salida deben ser variables Gaussianas. Por lo tanto, dividiendo una variable aleatoria Gaussiana por otra variable aleatoria Gaussiana, como se hace en Zhu y Premachandra, haría que la razón entre las variables aleatorias Gaussianas tuviese disribución Cauchy, mas no Gaussiana. El ACP ya no es apropiado para esta relación de datos pues el requisito necesario de distribución de Gauss es violado. Para empeorar las cosas, ni la media ni la varianza existe para una variable aleatoria con distribución Cauchy.

Veamos a continuación como al dividir dos variables aleatorias Gaussianas obtenemos una variable con distribución de Cauchy. Tenemos dos variables con distribución Gaussiana X e Y, calcularemos la función de distribución de X/Y y luego su función de densidad para observar que en efecto corresponde a la densidad de una variable Cauchy.

Sean X , Y $\sim$N(0,1), entonces:

$$\begin{aligned} F_{\frac{x}{y}}(t) &= P(\frac{X}{Y} \leq t) \\ &= P(X \leq tY, Y > 0) + P(X \geq tY, Y \leq 0) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [\mathbf{1}_{(-\infty,tY)}(x)\mathbf{1}_{(0,\infty)}(y) + \mathbf{1}_{(tY,\infty)}(x)\mathbf{1}_{(-\infty,0]}(y)]f(x)f(y)dxdy \end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty [\int_{-\infty}^{ty} f(x)dx]f(y)dy + \int_{-\infty}^0 [\int_{ty}^\infty f(x)dx]f(y)dy \\
&= \int_0^\infty \Phi(ty)f(y)dy + \int_\infty^0 [1-\Phi(ty)]f(y)dy \\
F_{\frac{x}{y}}(t) &= 2\int_0^\infty \Phi(ty)f(y)dy.
\end{aligned}
$$

Luego la densidad de X/Y es:

$$
\begin{aligned}
f_{\frac{x}{y}}(t) &= F'_{\frac{x}{y}}(t) \\
F'_{\frac{x}{y}}(t) &= 2\int_0^\infty yf(ty)f(y)dy \\
&= 2\frac{1}{2\pi}\int_0^\infty y\exp^{-\frac{y^2}{2}(1+t^2)}\,dy \\
&= \frac{1}{\pi\sqrt{1+t^2}}\int_0^\infty y\sqrt{1+t^2}\exp^{-\frac{(y\sqrt{1+t^2})^2}{2}}\,dy \\
&= \frac{1}{\pi\sqrt{1+t^2}}\int_0^\infty z\exp^{-\frac{z^2}{2}}\frac{dz}{\sqrt{1+t^2}} \\
&= \frac{1}{\pi(1+t^2)}
\end{aligned}
$$

Por lo tanto, la variable aleatoria X/Y, con X e $Y \sim N(0,1)$, tiene distribución de Cauchy.

Shanmugam y Johnson [**14**] proponen un método de integración que evita caer en el problema de la distribución Cauchy.

La idea principal de la integración del AED y ACP consiste en aplicar primero el ACP para la reducción de datos en $q \leq p$ scores de componentes principales de entrada y de salida separadamente. De acuerdo con el $\alpha \in (0,1)$ escogido, de manera óptima un $q*$ número de importantes componentes principales son elegidos y sus scores se consideran para el AED posterior. El ACP se realiza sobre los datos de entrada y se considera sólo un número importante de scores de CP de entradas para las calificaciones con AED. Asimismo, el ACP se realiza sobre las variables de datos de salida, y se considera sólo otro grupo significativo de score de CP de salidas para realizar el AED. En el método de regresión, se construirán

ecuaciones de regresión para las CP de las salidas y serán usadas para proyectar cada uno de los scores en términos de las CP de las entradas como variables de datos independientes. Los scores de CP de las salidas proyectadas entonces, se pueden convertir en porcentajes simples y la clasificación de las unidades puede hacerse con base en los porcentajes.

El AED se realiza usando los scores de CP de las entradas como datos de entradas y los scores de CP de las salidas como datos de salida, en lugar de los datos originales. Se utiliza el software DEA-solver contenido en Cooper [**8**] para hallar la solución óptima en la aplicación del método propuesto.

Con el objetivo de determinar la correlación entre los métodos AED, ACP y el método de integración AED-ACP, se realizará la prueba no paramétrica de correlación de Spearman, calculando el coeficiente ρ de la siguiente forma:

$$\rho \quad = \quad 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

donde $d_i = U_i - V_i$, U_i y V_i son los valores de eficiencia (eficiencias operacionales) de la i-ésima UTD bajo dos métodos de clasificación diferentes. El ρ de Spearman oscila entre -1 y 1, cuando es significativamente grande, es decir, si ρ es 1 o un valor muy cercano a 1, implica que los métodos no son estadísticamente diferentes, o en otras palabras que están altamente correlacionados.

Para la aplicación de los métodos, se usan datos médicos de 45 países, sobre la proporción de supervivencia de melanoma entre hombre y mujeres como variables de salida y los grados latitud, la capa de ozono, los rayos ultravioletas A y B, como variables de entrada. Los países serán clasificados usando AED, ACP, y luego el método integrado AED-ACP planteado anteriormente.

CAPíTULO 4

Aplicación de los Métodos: AED, ACP y el integrado AED-ACP

1. Introducción

Como una aplicación de los capítulos anteriores vamos a clasificar 45 países de acuerdo con la proporción de supervivencia al melanoma entre hombres y mujeres, teniendo en cuenta la acción de los rayos ultravioletas A y B, la Capa de Ozono y la Latitud.

2. Aplicación de los Métodos y Análisis de los Resultados

Cualquiera de las versiones de AED en el software [**8**] podría escogerse para trabajar, pero la versión BCC de Banker-Charnes-Cooper es seleccionada en vez de la versión Charnes-Cooper-Rhodes (CCR) u otra versión para los análisis en ésta sección pues, a pesar de que el modelo CCR tiene ambas orientaciones a las entradas y a las salidas, al igual que el BCC, el CCR se construye sobre la suposición de que constantes se vuelven escalares. Esto implica que para cada par entrada-salida factible (x, y), el CCR requiere que (tx, ty) sea factible para cada escalar positivo t. Esta suposición del CCR es demasiado fuerte, pero el BCC relaja esta suposición. Tampoco, el CCR toma en consideración los excesos de entradas y déficit de salidas, ni contempla la posibilidad de existencia de ineficiencias debidas a las diferencias entre las escalas operativas en cada UTD. La versión de BCC corrige estas deficiencias.

En la versión de BCC hay dos formas de encontrar la solución óptima, la primera forma es minimizar las entradas que mantienen los niveles de las salidas fijos llamada **solución orientada a las entradas** (BCC-I en [**8**]). La segunda forma, consiste en maximizar las salidas sin requerir entradas extras y se llama **solución orientada a las salidas** (BCC-O en [**8**]). La clasificación eficiente es la misma si se elige la primera o la segunda manera. En el siguiente ejemplo, se estudiarán los datos de Garland [**11**] sobre el melanoma en diversos

países, para la búsqueda de la solución óptima emplearemos ambos métodos a fin de ejemplificar la equivalencia de éstos.

Los datos del artículo de Garland [**11**] se muestran en la Tabla 1. En esta tabla se observa que el número de variables de salidas es $m = 2$, el número de variables de entradas es $k = 4$ y $n = 45$ es el número de países.

Antes de aplicar el ACP para clasificar los países, algunos análisis estadísticos básicos podrían ofrecer características interesantes sobre los datos. La media y desviación estándar de las variables de datos Latitud, UVA, UVB, Capa de Ozono, proporción de supervivencia de hombres y de mujeres al melanoma se muestran en la Tabla 1. En esta tabla se muestra que, prácticamente la totalidad de las observaciones de latitud, UVA, UVB, capa de ozono, proporción de supervivencia de hombres y de mujeres se encuentran a tres desviaciones típicas de la media. Éstos junto con sus gráficas de probabilidad (Figuras 1-6), inducen a pensar que:

(1) El conjunto de variables de datos de entrada: Latitud, UVA, UVB, Capa de Ozono están distribuidos como una Gaussiana multivariante con dimensión $p = 4$.
(2) El conjunto de variables de datos de salida, proporción de supervivencia de hombres y proporción de supervivencia de mujeres siguen una distribución Gaussiana multivariante con dimensión $p = 2$.
(3) Las variables de entrada son los factores causales de las variables de salida, según sus correlaciones, que se muestran en la Tabla 2.

Tabla 1. Datos Originales

PAÍS	ENTRADAS				SALIDAS	
	LATITUD	UVA	UVB	CAPA DE	Coeficiente de supervivencia	
	(Grados)	(Watts /m2)	(Watts /m2)	OZONO	Hombres	Mujeres
Alemania	52	34800	694	275	0.82	0.88
Argentina	34	51300	1409	335	0.93	0.96
Australia	38	48100	1253	370	0.5	0.74
Austria	48	38800	850	275	0.76	0.84
Bélgica	51	35800	732	275	0.86	0.9
Bulgaria	43	43600	1052	310	0.93	0.96
Canadá	50	36900	770	320	0.8	0.91
Chile	34	51300	1409	300	0.93	0.92
Costa Rica	10	63600	2064	275	0.89	0.99
Cuba	23	58500	1783	275	0.96	0.98
Dinamarca	56	30600	546	275	0.66	0.74
Escocia	56	30600	546	275	0.88	0.9
España	41	45500	1133	275	0.9	0.94
Estados Unidos	37	48900	1293	300	0.74	0.87
Finlandia	60	26200	409	240	0.79	0.89
Francia	49	37900	810	275	0.87	0.9
Grecia	37	48900	1293	325	0.95	0.96
Hong Kong	22	59000	1812	260	0.98	0.99
Hungría	48	38800	850	320	0.76	0.86
Inglaterra	52	34800	694	275	0.84	0.86
Irlanda	53	33800	656	275	0.9	0.89
Irlanda del Norte	54	32700	618	275	0.8	0.86
Islandia	68	17800	197	270	0.91	0.97
Israel	32	52800	1484	275	0.77	0.82
Italia	43	43600	1052	300	0.85	0.89
Japón	36	49700	1332	275	0.98	0.99
Luxemburgo	50	36900	770	275	0.96	0.8
Malta	36	49700	1332	310	0.97	0.98
México	19	60500	1893	275	0.96	0.98
Noruega	60	26200	409	245	0.82	0.78
Nueva Zelanda	42	44500	1092	385	0.41	0.7
Países Bajos	53	33800	656	275	0.82	0.84
Polonia	53	33800	656	275	0.85	0.87
Portugal	40	46200	1175	285	0.94	0.94
Puerto Rico	18	60900	1916	275	0.93	0.97
República Checa	50	36900	770	275	0.76	0.84
República de Corea	37	48900	1293	275	0.99	0.99
Rumania	44	42700	1011	310	0.9	0.92
Singapur	1	64700	2127	225	0.97	0.99
Suecia	58	28400	478	275	0.74	0.83
Suiza	47	39800	890	290	0.75	0.82
Trinidad	10	63600	2064	275	0.97	0.98
Uruguay	35	50600	1372	240	0.87	0.95
Venezuela	10	63600	2064	260	0.97	0.97
Yugoslavia	45	41800	970	300	0.85	0.89
Promedio	40.78	43728.89	1103.98	285	0.85	0.90

TABLA 2. Coorrelaciones

		Latitud	UVA	UVB	Ozono	Hombres	Mujeres
Latitud	Correlación de Pearson	1	-0.978**	-0.991**	0.078	-0.384**	-0.553**
	Sig. (bilateral)		0.00	0.00	0.608	0.009	0.000
UVA	Correlación de Pearson	-0.978**	1	0.994**	0.035	0.360*	0.534**
	Sig. (bilateral)	0		0	0.817	0.015	0.00
UVB	Correlación de Pearson	-0.991**	0.994**	1	-0.02	0.384**	0.562**
	Sig. (bilateral)	0.000	0		0.895	0.009	0.000
Ozono	Correlación de Pearson	0.078	0.035	-0.02	1	-0.504**	-0.326*
	Sig. (bilateral)	0.608	0.817	0.895		0.000	0.029
Hombres	Correlación de Pearson	-0.384**	0.360*	0.384**	-0.504**	1	0.859
	Sig. (bilateral)	0.009	0.015	0.009	0		0.000
Mujeres	Correlación de Pearson	-0.553**	0.534**	0.562**	-0.326*	0.858**	1
	Sig. (bilateral)	0.000	0.000	0.000	0.029	0.000	
**	La correlación es significativa al nivel 0,01 (bilateral).						
*	La correlación es significante al nivel 0,05 (bilateral).						

A continuación se muestran los gráficos de probabilidad (Gráficos Normal P-P) de la Latitud, los Rayos Ultravioletas A y B, la Capa de Ozono (Variables de Entrada) y la Supervivencia de Hombres y Mujeres al cancer de piel (Variables de Salida), los cuales se concluye que tanto las Entradas como las Salidas siguen una distribución Gaussiana.

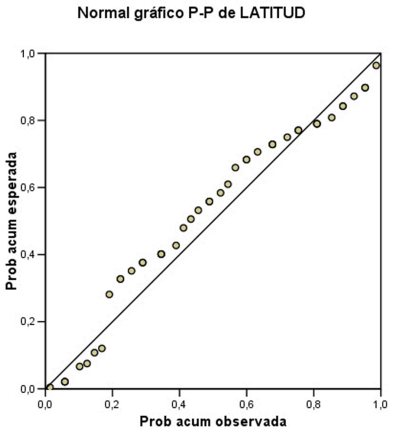


FIGURA 4.1. Gráfico Normal PP de Latitud

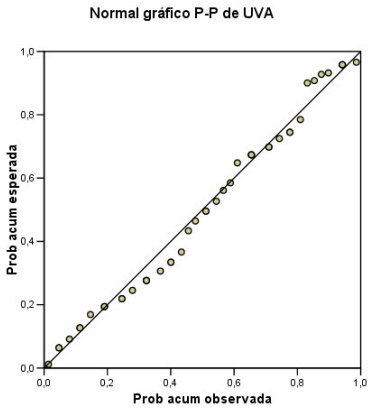


FIGURA 4.2. Gráfico Normal PP de UVA

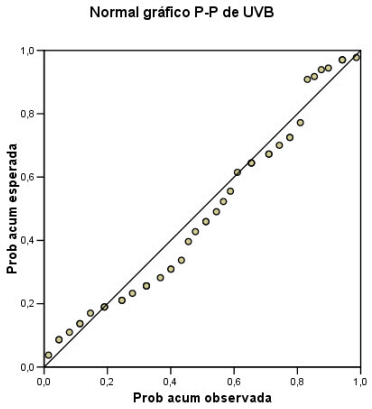


FIGURA 4.3. Gráfico Normal PP de UVB

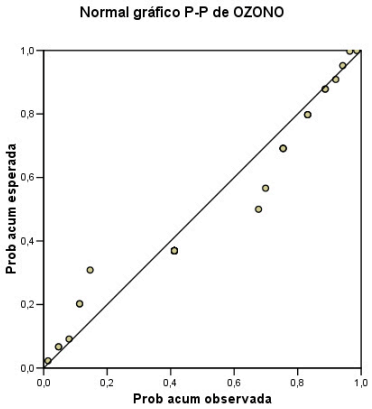


FIGURA 4.4. Gráfico Normal PP de Capa de Ozono

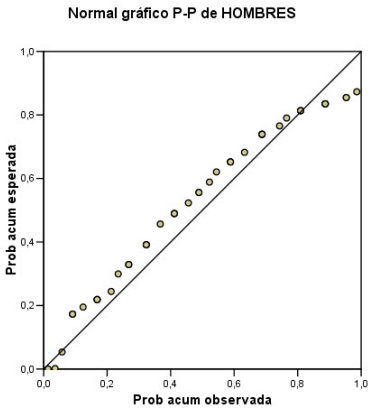


FIGURA 4.5. Gráfico Normal PP de Supervivencia de Hombres

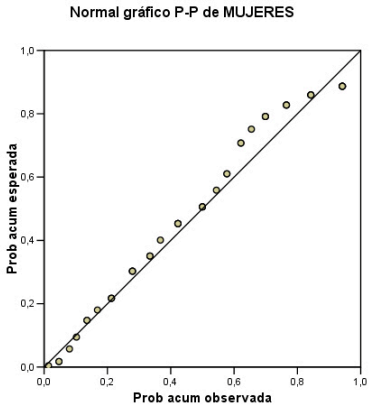


FIGURA 4.6. Gráfico Normal PP de Supervivencia de Mujeres

La correlación entre los rayos UVA, UVB y la latitud son un hecho científico. Las variables latitud, UVA, o UVB son apoderadas unas de otras. Por lo tanto, se seleccionan las correlaciones de latitud en la Tabla 2 para comentarlas. La variable latitud está considerable pero negativamente correlacionada con la proporción de supervivencia de hombres (correlación $r = -0,38$; $p - valor = 0,009$) y con la proporción de supervivencia de mujeres (correlación $r = -0{,}55$; $p - valor = 0{,}0001$), pero insignificantemente correlacionada con la capa de ozono (correlación $r = 0{,}07$; $p - valor = 0{,}60$). Las proporciones de supervivencia de hombres y de mujeres están positiva y significativamente correlacionada entre sí ($r = 0{,}859$; $p - valor = 0{,}0001$). Lo que es muy sorprendente es que mientras la capa de ozono no está significativamente correlacionada con la latitud, UVA, o UVB, si está fuerte, significativa, y positivamente correlacionada con la proporción de supervivencia de cáncer de

melanoma de hombres (correlación $r = 0{,}54$; $p - valor = 0{,}0004$), pero ligera, significativa, y negativamente correlacionada con la proporción de supervivencia de mujeres (correlación $r = -0{,}32$; $p - valor = 0{,}02$).

Luego de estos comentarios sobre las correlaciones podrían surgir algunas preguntas, como por ejemplo, si no se correlacionan significativamente la capa de ozono con la cantidad de radiación UVA y UVB que se recibe, ¿cómo ellos influyen significativa y positivamente en la proporción de supervivencia de hombres pero negativamente en la proporción de supervivencia de mujeres?. Este rompecabezas plantea una interesante pregunta sobre ¿cómo influencia la capa de ozono?, tal vez, las relaciones de regresión entre la capa de ozono, intensidades de UVA, y UVB podrían revelar algún dato que ayude a responder ésta interrogante. Estas preguntas no se responderán en este trabajo pero podrán servir para realizar otros trabajos.

Los Detalles sobre las correlaciones se muestran en la Tabla 2. Estos resultados implican relaciones causales de las variables de entrada y las variables de salida.

Los datos de la Tabla 1 son introducidos en la versión BCC-O (BCC orientada a las salidas) y en la versión BCC-I (BCC orientada a las entradas) del DEA-solver dado en [**8**]. Los scores de clasificación para estos 45 países se obtienen con el software, y se muestran en Tabla 3. En esta tabla se observa que el primer grupo de países con una tasa de supervivencia de melanoma excelente, según el BCC-O, está formado por Costa Rica, Finlandia, Hong Kong, Islandia, Japón, República de Corea, Luxemburgo, Noruega, Singapur, países que forman el Conjunto de Referencia. En cambio, se observa que para el BCC-I, sólo los países: Finlandia, Islandia, República de Corea, Luxemburgo, Noruega, Singapur del Conjunto de Referencia anterior son eficientes, incluyendo además a Suecia; esto se debe a que en el modelo orientado a las salidas (BCC-O) toma sólo los países que maximizan las salidas con las entradas disponibles y el modelo orientado a las entradas (BCC-I), toma como eficientes a los países que con mínimas entradas producen más salidas.

Siguiendo con el análisis de los resultados de la Tabla 3, para el BCC-O, el tercer grupo de países con peor proporción de supervivencia de muertes por melanoma es el conformado

Tabla 3. Análisis Envolvente de Datos BCC

PAÍS	Score AED BCC-O	Conj. Ref. según AED BCC-O	Score AED BCC-I	Conj. Ref. según AED BCC-I
Alemania	0.90	Islandia	0.98	Finlandia
Argentina	0.96	Japón	0.90	Islandia
Australia	0.74	Islandia	0.91	Finlandia
Austria	0.85	Islandia	0.96	Finlandia
Bélgica	0.92	Islandia	0.97	Finlandia
Bulgaria	0.97	Islandia	0.94	Islandia
Canadá	0.93	Islandia	0.97	Finlandia
Chile	0.94	Rep. de Corea	0.90	Singapur
Costa Rica	1.00	Rep. de Corea	0.96	Singapur
Cuba	0.98	Japón	0.89	Singapur
Dinamarca	0.76	Islandia	0.99	Finlandia
Escocia	0.95	Islandia	0.99	Islandia
España	0.95	Islandia	0.93	Singapur
Estados Unidos	0.87	Rep. de Corea	0.91	Noruega
Finlandia	1.00	Finlandia	1.00	Finlandia
Francia	0.92	Islandia	0.97	Noruega
Grecia	0.96	Rep. de Corea	0.93	Singapur
Hong Kong	1.00	Rep. de Corea	0.96	Rep. de Corea
Hungría	0.87	Islandia	0.96	Finlandia
Inglaterra	0.89	Islandia	0.98	Noruega
Irlanda	0.96	Islandia	0.99	Islandia
Irlanda del Norte	0.88	Islandia	0.98	Finlandia
Islandia	1.00	Islandia	1.00	Islandia
Israel	0.82	Rep. de Corea	0.89	Singapur
Italia	0.90	Islandia	0.94	Finlandia
Japón	1.00	Rep. de Corea	0.99	Rep. de Corea
Luxemburgo	1.00	Luxemburgo	1.00	Luxemburgo
Malta	0.98	Rep. de Corea	0.96	Rep. de Corea
México	0.99	Rep. de Corea	0.90	Singapur
Noruega	1.00	Noruega	1.00	Noruega
Nueva Zelanda	0.71	Islandia	0.93	Noruega
Países Bajos	0.87	Islandia	0.98	Finlandia
Polonia	0.91	Islandia	0.98	Noruega
Portugal	0.95	Islandia	0.93	Islandia
Puerto Rico	0.97	Rep. de Corea	0.89	Singapur
República Checa	0.86	Islandia	0.97	Noruega
República de Corea	1.00	Rep. de Corea	1.00	Rep. de Corea
Rumania	0.93	Islandia	0.94	Singapur
Singapur	1.00	Singapore	1.00	Singapur
Suecia	0.88	Finlandia	1.00	Finlandia
Suiza	0.83	Islandia	0.95	Finlandia
Trinidad	1.00	Rep. de Corea	0.95	Singapur
Uruguay	0.98	Finlandia	0.97	Singapur
Venezuela	0.99	Rep. de Corea	0.95	Singapur
Yugoslavia	0.90	Islandia	0.95	Finlandia

por Inglaterra, Irlanda del Norte, Suecia, Hungría, Países Bajos, Estados Unidos, República Checa, Austria, Suiza, Israel, Dinamarca, Australia, y Nueva Zelanda. Todos los otros 23 países que están en el segundo grupo, caen entre la proporción de supervivencia excelente y peor de muertes por cáncer de melanoma y estos son Venezuela, Trinidad, México, Malta, Cuba, Puerto Rico, Uruguay, Bulgaria, Grecia, Argentina, Irlanda, Portugal, Escocia, España, Chile, Rumania, Canadá, Francia, Bélgica, Polonia, Yugoslavia, Italia, y Alemania.

Según el BCC-I el grupo de países con peor índice de supervivencia al melanoma estaría formado por: México, Cuba, Israel y Puerto Rico; mientras que los demás países que están entre los de eficiencia 1 y los peores, forman el segundo grupo, a saber: Escocia, Japón, Dinamarca, Irlanda, Irlanda del Norte, Polonia, Países Bajos, Inglaterra, Alemania, Bélgica, Uruguay, Canadá, República Checa, Francia, Hong Kong, Hungría, Austria, Malta, Costa Rica, Suiza, Venezuela, Trinidad, Yugoslavia, Rumania, Bulgaria, Italia, Nueva Zelanda, España, Portugal, Grecia, Australia, Estados Unidos, Argentina y Chile.

El AED identifica un conjunto de referencia formado por una lista de países a ser imitados. Con respecto a la proporción de supervivencia de hombres y de mujeres, los países Australia, Bulgaria, Chile, España, Estados Unidos, Grecia, Hong Kong, Israel, Italia, Japón, Malta, Nueva Zelanda, Portugal, Rumania, Trinidad y Yugoslavia podrían imitar a la República de Corea en el primer grupo; los países Alemania, Austria, Bélgica, Canadá, Dinamarca, Escocia, Francia, Hungría, Inglaterra, Irlanda, Irlanda del Norte, Países Bajos, Polonia, República Checa, Suecia y Suiza deben imitar, en el primer grupo a Islandia; a Costa Rica, Cuba, México, Puerto Rico, Uruguay y Venezuela le corresponde simular a Singapur en el primer grupo.

El ACP ofrece los siguientes resultados. Los primeros dos autovalores son mayores que uno, capturando 99,53 % de la variación en las variables de entrada. Los scores de las CP se calculan usando:

$$CP1_i = -0{,}994 * latitud_i + 0{,}994 * UVA_i + 0{,}999 * UVB_i - 0{,}03 * Ozono_i$$

$$CP2_i = 0{,}048 * latitud_i + 0{,}068 * UVA_i + 0{,}012 * UVB_i + 0{,}999 * Ozono_i$$

Estos resultados del ACP revelan lo siguiente: (i) Las variables de entrada latitud, UVA, y UVB son un racimo importante de variables y ellas están identificadas en la primera CP; (ii) La variable de entrada ozono también es importante y está sola, según la segunda CP. Esto se muestra en la siguiente Figura:

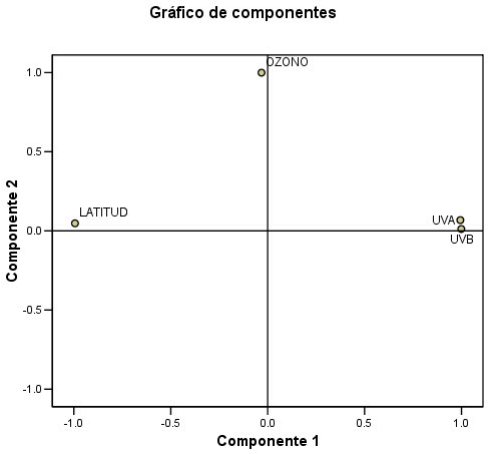


FIGURA 4.7. Gráfico de Componentes

Sustituyendo las observaciones: $\underline{D_i'} = (latitud_i, UVA_i, UVB_i, Ozono_i)$ en las ecuaciones anteriores, los scores de CP1 y de CP2 se calculan para los países $i = 1, 2, ..., 45$. Igualmente, el score de la CP3 puede calcularse como una combinación lineal óptima de las variables de salida, proporción de supervivencia de hombres y proporción de supervivencia de mujeres, usando: $CP3_i = 0{,}964 * (Hombres_i) + 0 - 964 * (Mujeres_i)$ para los países $i = 1, 2, ..., n = 45$. Con el primer autovalor igual a 1,859, la CP3 captura 92,94 % de la variación total en las proporciones de supervivencia de hombres y de mujeres.

En la Tabla 4 se presentan los valores de las Componentes Principales: CP1, CP2, y CP3. Las correlaciones de las variables de entrada CP1 y CP2 con la variable de salida CP3 son 0,44 y 0,46 con 0,001 y 0,001 como sus p-valores. Estas correlaciones son significantes.

Como se mencionó anteriormente, se ajusta una regresión y se proyectan los valores de la CP3 por los valores dados de las Componentes Principales CP1 y CP2 para cada UTD. Estos valores proyectados $\widehat{CP3}$ pueden convertirse no sólo en scores de porcentajes sino también

en scores de tasas de eficacia, dividiéndolos por el máximo porcentaje de los scores. Estos scores de tasa de eficacia se muestran en la Tabla 4. Estos números sugieren que el país Singapur está en la línea fronteriza y todos los países deben imitar a Singapur con respecto a la supervivencia del cáncer del melanoma.

La mayoría de las implicaciones en los resultados del AED se repite aquí en los resultados del ACP. Hay algunas sorpresas. Como en los resultados del AED (BCC-O), los países Venezuela, Uruguay, Hong Kong, Trinidad, Costa Rica, Puerto Rico, México, y Cuba tienen la clasificación eficaz en la proporción de sobrevivencia de melanoma. Los países Nueva Zelanda, Australia, Islandia, Canadá, Hungría, Suecia, y Argentina están teniendo peor supervivencia. Esta aproximación para clasificar, basada completamente en la técnica del ACP no es tan convincente ni satisfactoria como el AED porque no hay un conjunto de referencia de países. Se evidencia acá una de las ventajas de integrar ACP con AED.

Siguiendo la regla se aplica para identificar el outliers en la técnica del ACP. Si el score

$$z_{l,j} = \frac{CP_{lj} - n^{-1}\sum_{j=1}^{n} CP_{lj}}{n^{-1}\sum_{j=1}^{n}(CP_{lj} - n^{-1}\sum_{j=1}^{n} CP_{lj})^2}$$

cae fuera del intervalo $(-z_{\alpha/2}, z_{\alpha/2})$ entonces la j-ésima UTD es considerada extraordinariamente influenciada por la l-ésima CP. En los datos de melanoma, con $\alpha = 0{,}05$ es interesante notar que los países de Islandia y Singapur tienen valores z_{1j} y z_{2j} fuera de lo común en las CP1 y CP2, respectivamente. Los países Australia y Nueva Zelanda tienen números z_{3j} inusual en la CP3. Estos resultados confirman la sospecha de Garland [**11**] de que Australia y Nueva Zelanda ahora tienen una rara incidencia de mortalidad por cáncer de melanoma a pesar de que la mortalidad por melanoma era antes inexistente en ambos países. Tal hallazgo no fue posible con el AED.

El ACP se ha realizado en los datos de entrada: latitud, capa de ozono, UVA, y UVB, dando como resultado que las dos primeras CP capturan 99,53 % de la variación total en los datos de entrada, y por lo tanto los dos scores de CP, CP1 y CP2 son suficientes y se generan. Sus autovalores son 2,97 y 1,01, respectivamente. Igualmente, el ACP se ha realizado en las variables de datos de salida proporción de supervivencia de hombres y proporción

de supervivencia de mujeres. Sólo una CP es necesaria según el autovalor 1,86, y es la CP3. Entonces se calculan los scores de CP3.

La Tabla 4 contiene los scores de CP1, CP2, y CP3. El AED se realiza usando CP1 y CP2 como variables de entrada y CP3 como una variable de salida. Los 45 países se clasifican usando el DEA-solver [**8**], y las clasificaciones se muestran en la Tabla 4.

El primer grupo de países con el score de clasificación perfecto es el formado por Islandia y la República de Corea según el valor de CP3, una combinación lineal óptima de la proporción de supervivencia de hombres y de mujeres a partir de la proporción de supervivencia de melanoma.

El tercer grupo de países de los pésimos, es decir, por debajo del score de la media de clasificación 0.57 es: Alemania, Inglaterra, Bélgica, Rep. Checa, Luxemburgo, Canadá, Francia, Bulgaria, Austria, Hungría, Suiza, Yugoslavia, Rumania, Italia, Argentina, España, Nueva Zelanda, Portugal, Puerto Rico, Estados Unidos, Australia, Uruguay, Chile, Israel, y Costa Rica en un orden de peor a mucho peor.

El segundo grupo de países que caen entre el primer grupo y tercer grupo de países es Japón, Malta, Hong Kong, Finlandia, Noruega, Singapur, Suecia, Cuba, Trinidad, Dinamarca, Escocia, México, Venezuela, Irlanda del Norte, Grecia, los Países Bajos e Irlanda. Los países Cuba, Hong Kong, Japón, Malta, México, Puerto Rico, Singapur, Trinidad y Venezuela podrían copiar a la República de Corea para mejorar la proporción de supervivencia del cáncer del melanoma. Todos los otros países en el segundo o tercer grupo podrían considerar a Islandia para mejora la proporción de supervivencia del cáncer del melanoma.

Las correlaciones de Spearman (vea la Tabla5) para comparar el AED, el ACP, y el método integrado AED-ACP, son grandes y éstas implican que los tres métodos no son estadísticamente diferentes.

Tabla 4. AED para Componentes Principales

	ENTRADAS		SALIDA	Proyec.	AED	Conj. Ref.	AED	Conj. Ref.
PAÍS	CP 1	CP 2	CP 3	de CP 3	BCC-I	AED BCC-I	BCC-O	AED BCC-O
Alemania	35224.57	2651.95	1.64	0.82	0.56	Islandia	0.88	Rep. de Corea
Argentina	52355.95	3841.61	1.82	0.80	0.44	Islandia	0.95	Rep. de Corea
Australia	49014.28	3657.29	1.20	0.73	0.41	Islandia	0.63	Rep. de Corea
Austria	39360.39	2925.63	1.54	0.84	0.51	Islandia	0.82	Rep. de Corea
Bélgica	36257.52	2720.36	1.70	0.83	0.55	Islandia	0.91	Rep. de Corea
Bulgaria	44337.31	3289.18	1.82	0.81	0.52	Islandia	0.96	Rep. de Corea
Canadá	37388.53	2840.52	1.65	0.77	0.52	Islandia	0.88	Rep. de Corea
Chile	52357.00	3806.64	1.78	0.85	0.39	Islandia	0.93	Rep. de Corea
Costa Rica	65262.15	4624.77	1.81	0.92	0.32	Islandia	0.95	Rep. de Corea
Cuba	59899.11	4275.23	1.87	0.91	0.65	Rep. de Corea	0.98	Rep. de Corea
Dinamarca	30897.94	2364.77	1.35	0.80	0.63	Islandia	0.73	Islandia
Escocia	30897.94	2364.77	1.72	0.80	0.63	Islandia	0.93	Islandia
España	46309.86	3384.29	1.77	0.87	0.44	Islandia	0.93	Rep. de Corea
Estados Unidos	49852.53	3642.19	1.55	0.84	0.41	Islandia	0.81	Rep. de Corea
Finlandia	26384.55	2029.15	1.62	0.84	0.73	Islandia	0.88	Islandia
Francia	38424.83	2864.00	1.71	0.84	0.52	Islandia	0.91	Rep. de Corea
Grecia	49851.78	3667.17	1.84	0.80	0.58	Islandia	0.96	Rep. de Corea
Hong Kong	60426.52	4294.54	1.90	0.93	0.79	Rep. de Corea	0.99	Rep. de Corea
Hungría	39359.04	2970.58	1.56	0.77	0.50	Islandia	0.83	Rep. de Corea
Inglaterra	35224.57	2651.95	1.64	0.82	0.56	Islandia	0.88	Rep. de Corea
Irlanda	34191.61	2583.54	1.73	0.82	0.58	Islandia	0.93	Rep. de Corea
Irlanda del Norte	33059.26	2508.33	1.60	0.81	0.59	Islandia	0.86	Islandia
Islandia	17814.31	1485.76	1.81	0.75	1.00	Islandia	1.00	Islandia
Israel	53925.66	3884.47	1.53	0.89	0.38	Islandia	0.80	Rep. de Corea
Italia	44337.61	3279.19	1.68	0.82	0.45	Islandia	0.89	Rep. de Corea
Japón	50688.43	3672.04	1.90	0.88	0.93	Rep. de Corea	0.99	Rep. de Corea
Luxemburgo	37389.88	2795.57	1.70	0.83	0.53	Islandia	0.91	Rep. de Corea
Malta	50687.38	3707.00	1.88	0.83	0.80	Rep. de Corea	0.98	Rep. de Corea
México	62000.97	4412.35	1.87	0.91	0.63	Rep. de Corea	0.98	Rep. de Corea
Noruega	26384.40	2034.14	1.54	0.83	0.73	Islandia	0.84	Islandia
Nueva Zelanda	45270.61	3425.74	1.07	0.70	0.43	Islandia	0.56	Rep. de Corea
Países Bajos	34191.61	2583.54	1.60	0.82	0.58	Islandia	0.86	Rep. de Corea
Polonia	34191.61	2583.54	1.66	0.82	0.58	Islandia	0.89	Rep. de Corea
Portugal	47048.32	3442.34	1.81	0.85	0.43	Islandia	0.95	Rep. de Corea
Puerto Rico	62422.54	4439.78	1.83	0.91	0.43	Islandia	0.96	Rep. de Corea
República Checa	37389.88	2795.57	1.54	0.83	0.53	Islandia	0.82	Rep. de Corea
República de Corea	49853.28	3617.22	1.91	0.88	1.00	Rep. de Corea	1.00	Rep. de Corea
Rumania	43400.75	3227.53	1.75	0.80	0.46	Islandia	0.93	Rep. de Corea
Singapur	66428.93	4649.95	1.89	1.00	0.69	Rep. de Corea	0.99	Rep. de Corea
Suecia	28641.22	2214.45	1.51	0.79	0.67	Islandia	0.82	Islandia
Suiza	40394.89	3009.05	1.51	0.82	0.49	Islandia	0.80	Rep. de Corea
Trinidad	65262.15	4624.77	1.88	0.92	0.64	Rep. de Corea	0.98	Rep. de Corea
Uruguay	51625.04	3698.70	1.75	0.93	0.40	Islandia	0.92	Rep. de Corea
Venezuela	65262.60	4609.79	1.87	0.94	0.60	Rep. de Corea	0.98	Rep. de Corea
Yugoslavia	42464.50	3155.90	1.68	0.81	0.47	Islandia	0.89	Rep. de Corea

Tabla 5. Coeficiente de Correlación de Spearman

	AED vs ACP	AED vs método integrado	ACP vs método integrado
ρ de Spearman	0,9994	0,9999	0,9986

Algunos comentarios valen la pena sobre el ejemplo. ¿Los resultados de este ejemplo son útiles? Los resultados son de tipo exploratorio pero no de tipo conclusivo. Los países que se clasificaban muy bajo en este artículo no podrían, por supuesto, cambiar la latitud que controla los UVA o UVB. Conocer la posición de un país específico comparando con otros países en las clasificaciones de la proporción de supervivencia de melanoma es, en sí mismo, un beneficio de los resultados de este ejemplo. Debe haber otras variables que también se conectan al cáncer del melanoma entre los hombres y mujeres. Vale la pena reunir e involucrar datos con esas otras variables omitidas, y esto es conocimiento aislado. Este caso ejemplifica satisfactoriamente que el AED y el ACP se complementan mutuamente.

A continuación se realizaron algunas modificaciones a los datos originales para ver la importancia de que los datos sean Gaussianos en el ACP y para ver el comportamiento del AED con datos grandes.

En la Tabla 6 se muestran los resultados del ACP aplicado a los datos originales de la Tabla 1 pero alterando la variable UVA para que no sea Gaussiana. En este caso sólo se obtiene una Componente Principal para las Entradas que captura el $74,8\,\%$ de la variación total. Los scores de esta componente principal se calculan de la siguiente forma:

$$CP1_i = -0{,}997 * latitud_i + 0{,}999 * UVA_i + 0{,}998 * UVB_i - 0{,}069 * Ozono_i$$

Se puede observar que en esta componente principal queda prácticamente fuera la variable del espesor de la Capa de Ozono, que en el análisis previo se vio que la misma era de gran importancia. Esto no muestra que se pierde una parte de información importante.

Para las salidas obtenemos de nuevo la misma Componente Principal puesto que no se alteraron los datos de las variables de proporción de supervivencia de hombres y de mujeres.

En cambio, al aplicar el AED a estos datos modificados se obtiene resultados similares a los anteriores aún al aplicar el AED a las dos Componentes Principales resultantes de estos "nuevos"datos.

En este ejemplo se evidencia la importancia de revisar la distribución de las variables antes de realizar el Análisis de Componentes Principales (ACP).

Observamos en la Tabla 6 que el primer grupo de países con una tasa de supervivencia de melanoma excelente, está formado por Costa Rica, Finlandia, Hong Kong, Islandia, Japón, Rep. de Corea, Luxemburgo, Noruega, Singapur y Yugoslavia, que es similar al primer grupo que obtuvimos en la Tabla 3, sólo que en este caso resultó eficiente también Yugoslavia. Esto muestra que el Análisis Envolvente de Datos no se ve afectado significativamente, por el cambio en la Gaussianidad de una de las variables de Entrada. Se evidencia nuevamente con este ejemplo la manera como un método soluciona o corrige las deficiencias del otro.

Tabla 6. AED y ACP a Datos Modificados

PAÍS	ENTRADA CP 1	SALIDA CP 2	AED Datos "nuevos"	Conj. Ref. AED BCC-O	AED para CP	Conj. Ref. AED BCC-O
Alemania	1,209,829,582	1.64	0.90	Islandia	0.88	Islandia
Argentina	2,629,059,659	1.82	0.97	Japón	0.95	Rep. de Corea
Australia	2,311,297,577	1.20	0.75	Rep. de Corea	0.63	Rep. de Corea
Austria	1,503,935,341	1.54	0.86	Islandia	0.83	Rep. de Corea
Bélgica	1,280,359,021	1.70	0.92	Islandia	0.91	Islandia
Bulgaria	1,899,060,026	1.82	0.98	Rep. de Corea	0.97	Rep. de Corea
Canada	1,360,249,087	1.65	0.93	Islandia	0.89	Rep. de Corea
Chile	2,629,059,662	1.78	0.94	Rep. de Corea	0.93	Rep. de Corea
Costa Rica	4,040,917,071	1.81	1.00	Singapur	0.95	Rep. de Corea
Cuba	3,418,829,488	1.87	0.99	Hong Kong	0.98	Rep. de Corea
Dinamarca	935,424,110	1.35	0.77	Islandia	0.73	Islandia
Escocia	935,424,110	1.72	0.97	Islandia	0.93	Islandia
España	2,068,180,821	1.77	0.95	Rep. de Corea	0.94	Rep. de Corea
Estados Unidos	2,388,820,023	1.55	0.88	Rep. de Corea	0.81	Rep. de Corea
Finlandia	685,753,892	1.62	1.00	Finlandia	0.89	Islandia
Francia	1,434,974,331	1.71	0.93	Islandia	0.92	Rep. de Corea
Grecia	2,388,820,021	1.84	0.97	Rep. de Corea	0.96	Rep. de Corea
Hong Kong	3,477,520,769	1.90	1.00	Rep. de Corea	0.99	Rep. de Corea
Hungría	1,503,935,338	1.56	0.88	Islandia	0.84	Rep. de Corea
Inglaterra	1,209,829,582	1.64	0.90	Islandia	0.88	Islandia
Irlanda	1,141,298,143	1.73	0.97	Islandia	0.93	Islandia
Irlanda del Norte	1,068,221,254	1.60	0.89	Islandia	0.87	Islandia
Islandia	316,523,270	1.81	1.00	Islandia	1.00	Islandia
Israel	2,785,053,590	1.53	0.83	Japón	0.80	Rep. de Corea
Italia	1,899,060,026	1.68	0.91	Rep. de Corea	0.89	Rep. de Corea
Japón	2,467,621,184	1.90	1.00	Rep. de Corea	0.99	Rep. de Corea
Luxemburgo	1,360,249,090	1.70	1.00	Luxemburgo	0.91	Rep. de Corea
Malta	2,467,621,182	1.88	0.99	Rep. de Corea	0.98	Rep. de Corea
México	3,656,591,601	1.87	0.99	Singapur	0.98	Rep. de Corea
Noruega	685,753,891	1.54	1.00	Noruega	0.84	Islandia
Nueva Zelanda	1,978,270,771	1.07	0.71	Rep. de Corea	0.57	Rep. de Corea
Países Bajos	1,141,298,143	1.60	0.88	Islandia	0.86	Islandia
Polonia	1,141,298,143	1.66	0.92	Islandia	0.90	Islandia
Portugal	2,132,306,673	1.81	0.96	Rep. de Corea	0.96	Rep. de Corea
Puerto Rico	3,705,103,065	1.83	0.98	Singapur	0.96	Rep. de Corea
República Checa	1,360,249,090	1.54	0.86	Islandia	0.83	Rep. de Corea
República de Corea	2,388,820,025	1.91	1.00	Rep. de Corea	1.00	Rep. de Corea
Rumania	1,821,467,654	1.75	0.94	Rep. de Corea	0.93	Rep. de Corea
Singapur	4,181,906,016	1.89	1.00	Singapur	0.99	Rep. de Corea
Suecia	805,753,840	1.51	0.89	Finlandia	0.82	Islandia
Suiza	1,582,456,781	1.51	0.84	Islandia	0.81	Rep. de Corea
Trinidad	4,040,917,071	1.88	0.99	Singapur	0.98	Rep. de Corea
Uruguay	2,557,800,958	1.75	0.98	Singapur	0.92	Rep. de Corea
Venezuela	4,040,917,072	1.87	0.99	Singapur	0.98	Rep. de Corea
Yugoslavia	1,745,493,662	1.68	1.00	Yugoslavia	0.89	Islandia

Conclusión

En este Trabajo Especial de Grado se presentaron dos métodos alternativos para ordenar UTD que tengan múltiples entradas y salidas, desde el punto de vista de la eficiencia relativa y compara la ordenación proporcionada por cada uno de ellos. Para posteriormente realizar una integración de ambos y obtener mejores resultados de clasificación de la UTD.

Uno de estos métodos: El AED, es un método no paramétrico, que usa programación lineal para medir eficiencia, utilizando cocientes de la suma de las salidas ponderadas sobre la suma de las entradas ponderadas. El otro es un método estadístico multivariado: ACP, que usa información de autovalores para combinar componentes principales obtenidas a partir de cocientes individuales de cada salida sobre cada entrada.

Cabe destacar, como parte final, algunas características sobresalientes de ambos métodos. El AED no requiera que los datos sean estocástico y por lo tanto no es necesario realizar ninguna comprobación de distribución de probabilidad para las observaciones de las variables de datos (de entrada o de salida). Antes de aplicar ACP, debe hacerse una revisión de la distribución subyacente, ya que el ACP requiere que las observaciones de las variables de datos (sean de tipo de entrada o de salida) sigan una distribución Gaussiana multivariada. El AED separa las variables de los datos en dos tipos de entrada o de salida.

En el procedimiento planteado de integración, primero debe realizarse el ACP separadamente en las observaciones de las variables de datos de entrada y en las de salida para obtener un número significante de Componentes Principales y generar los scores de entradas y de salidas. El AED es entonces aplicado en los scores de las entradas y salidas generando la tasa de la eficacia operacional de la UTD.

Es importante resaltar, que los modelos del AED presentan la ventaja de haber sido construidos específicamente para medir eficiencia técnica. Además, proporcionan un conjunto de información complementaria, que puede servir para fijar políticas tendientes a mejorar la eficiencia de determinadas UTD o del conjunto de referencia.

El método de integración AED-ACP ofrece varias ventajas tales como, el evitar caer en una distribución de Cauchy al realizar el cociente de las observaciones y el logro de lo mejor de las metodologías del AED y del ACP.

En el capítulo 4, se hace la aprobación del método de integración propuesto utilizando el Test de correlación de Spearman. La correlación es significativamente grande y confirma que el método propuesto no es estadísticamente diferente del AED y del ACP. El método integrado AED-ACP ofrece, como ya se vio varias ventajas, y por lo tanto se recomienda su uso.

Asimismo, resulta conveniente aclarar que con esto no se intenta proponer la sustitución del método AED por ACP, sino sólo disponer de un instrumento complementario de análisis, que sirva para advertir la posible necesidad de reanalizar el modelo utilizado (variables seleccionadas, modelo particular AED utilizado, etc.).

En relación con el resultado obtenido al aplicar los métodos, se sugiere a los países menos eficientes seguir las políticas medioambientales del país de referencia correspondiente para así aumentar la proporción de hombres y mujeres que sobreviven al cáncer de piel.

Debido a que los países ineficientes no pueden cambiar su latitud, cantidad de rayos UVA y UVB o el espesor de la capa de ozono, se propone realizar campañas para proteger la capa de ozono y crear medicinas que reduzcan la absorción de los rayos ultravioletas A y B (UVA y UVB).

Finalmente, este resultado, como otros similares realizado por otros autores, constituye un aliciente para un estudio teórico general sobre las relaciones de los dos métodos utilizados.

Bibliografía

[1] Adler N, Golany B.*Evaluation of deregulated airline networks using data envelopment analysis combined with principal component analysis with an application to Western Europe.* European Journal of Operational Research 2001;132(2):260-73.

[2] Afriat, S.N.*Efficiency estimation of production functions.*International Economic Review. 1972, 13, 3, 568-98.

[3] Aigner, D. J. y Chu, S.F.*On Estimating the Industry Production Function* American Economic Review,1968, vol. 58, nº 4, pp. 826-39.

[4] Charnes, A. *Data Envelopment Analysis: Theory,Methodology and Applications* New York,1997, Kluwer Academic Publishers, Second Edition.

[5] Charnes, A. y Cooper, W.W.*Programming with Linear Fractional Functionals* Naval Research Logistics Quaterly,1962, Vol. 9, pp. 181-6.

[6] Charnes, A.; Cooper, W.W.; Lewin, A.Y. y Seiford, L.M. *Data Envelopment Analysis: Theory, Methodology and Applications*, Kluwer Academic Publishers,1994. Boston.

[7] Charnes, A.; Cooper, W.W. y Rhodes, E. *Measuring efficiency of decision making unit.*European Journal of Operational Research. 1978;2:429-444.

[8] Cooper W, Seiford, L.M. y Tone, K *Data Envelopment Analysis* Second Edition, Springer 2007. NY. isbn 978-0387-45281-4.

[9] Dunlop, W. *The elusive concept of efficiency: a survey of the conceptual and measurement issues.* Occasional Paper, 109, Department of Economics.1985. University of Newcastle, Australia.

[10] Farrel, M.J.*The Measurement of Efficiency Productive*, Journal of the Royal Statistical Society ,1957, serie A, vol. 120.

[11] Garland, C.F.; Garland, F.C. y Gorham, E.D. *Epidemiologic evidence for different roles of ultraviolet A and B radiation in melanoma mortality rates.*Annals of Epidemiology. 2003;13(6):395-404.

[12] Jolliffe, I.T.*Principal components analysis.*New York: Springer. 2002.

[13] Premachandra IM. *A note on DEA vs. principal component analysis: an improvement to Joe Zhu's approach.* European Journal of Operational Research 2001;132(3):553-60.

[14] Shanmugam, R. y Johnson, C. *At a crossroad od data envelopment and principal component analyses.* Omega, 35 (2007) 351-364.

[15] Seiford, L.M.*DEA: The Evolution of the State of the Art(1978-1995)* The Journal of Productivity Analysis,1996, 7, 99-137.

[16] Zhu J. *Data Envelopment analysis vs. principal component analysis: an illustrative study of economic performance of Chinese cities.* European Journal of Operational Research 1998;111(1):50-61.

Apéndices

Apéndice 1

1. Programación Lineal (PL)

La Programación Lineal (PL) es una técnica matemática que consiste en una serie de métodos y procedimientos que permiten resolver problemas que están formulados a través de ecuaciones lineales, optimizando (minimizando o maximizando) la función objetivo, también lineal. La Programación Lineal se emplea sobre todo, en el ámbito de Ingeniería y Ciencias Sociales, permitiéndole a empresas y organizaciones, importantes beneficios y ahorro asociado a su utilización.

Un modelo de Programación Lineal proporciona un método eficiente para determinar una decisión óptima, escogida de un gran número de decisiones posibles. Las funciones que lo componen, es decir, función objetivo y restricciones, son funciones lineales en las variables de decisión.

En todos los problemas de Programación Lineal, la idea es maximizar o minimizar un objetivo que está sujeto a una lista de restricciones.

En un problema de Programación Lineal de dos variables x e y, se trata de optimizar (hacer máxima o mínima, según los casos) una función (llamada función objetivo) de la forma:

$$F(x, y) = A \cdot x + B \cdot y$$

y sujeta a una serie de restricciones dadas mediante un sistema de inecuaciones lineales del tipo:

$$\begin{cases} a_1x + b_1y \leq c1, \\ a_2x + b_2y \leq c2, \\ \vdots \\ a_mx + b_my \leq cm. \end{cases}$$

Los puntos del plano que cumplen el sistema de desigualdades forman un recinto convexo acotado (poligonal) o no acotado, llamado *región factible* del problema.

Todos los puntos de dicha región cumplen el sistema de desigualdades. Se trata de buscar, entre todos esos puntos, aquel o aquellos que hagan el valor de $F(x, y)$ máximo o mínimo, según sea el problema.

Los puntos de la región factible se denominan *soluciones factibles*.

De todas esas soluciones factibles, aquellas que hacen óptima (máxima o mínima) la función objetivo se llaman *soluciones óptimas*.

En general, un problema de programación lineal puede tener una, infinitas o ninguna solución.

Si hay una única solución óptima, ésta se encuentra en un vértice de la región factible y si hay infinitas soluciones óptimas, se encontrarán en un lado de la región factible.

Es posible que no haya solución óptima, pues cuando la región es no acotado, la función objetivo puede crecer o decrecer indefinidamente.

Para resolver el problema de PL con dos variables, se tienen dos formas la geométrica y la algebraica, pero antes a aplicar cualquiera de ellas siempre hay que tratar de dibujar la región factible, resolviendo el sistema de inecuaciones lineales correspondiente (la región factible puede estar acotada o no), y luego se calculan los vértices de dicha región.

Forma Geométrica

En este caso se representa el vector director de la recta que viene dada por la ecuación de la función objetivo, $F(x, y) = A \cdot x + B \cdot y$, que hay que maximizar o minimizar.

El vector director de la recta $A \cdot x + B \cdot y$, viene dado por $\overrightarrow{v} = (-B, A)$. Además, como lo único que nos importa es la dirección del vector y no su módulo, podemos dividir a las coordenadas del vector si los números son muy grandes, puesto que vectores con coordenadas

proporcionales tienen la misma dirección.

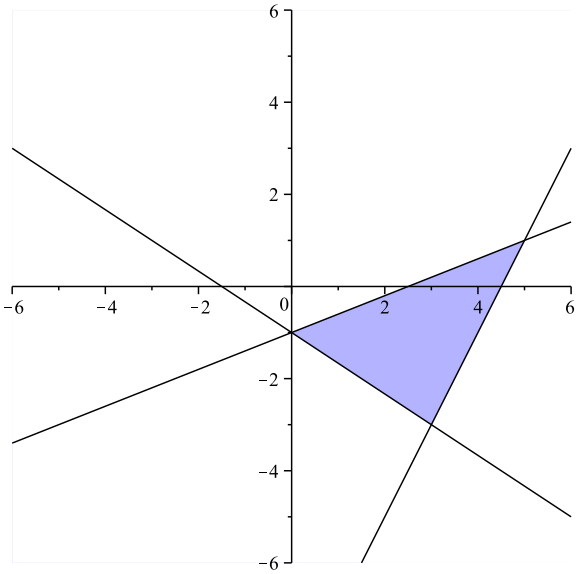
Posteriormente, se trazan rectas paralelas a este vector que pasen por los vértices de la región factible (si es acotada), o por todo el borde de la región factible (cuándo no es acotada) y se observa en qué vértice la función F se hace máxima (o mínima) sin más que tener en cuenta cuál de las rectas tiene mayor (o menor) ordenada en el origen, es decir, cuál recta corta en un punto mayor o menor al eje y.

Ejemplo: Maximizar la función $F(x, y) = 2000x + 5000y$ sujeta a las restricciones:

$$\begin{cases} 2x + 3y \geq -3, \\ 2x - y - 9 \leq 0, \\ 2x - 5y - 5 \geq 0. \end{cases}$$

La región factible en este caso se obtiene representando las rectas:

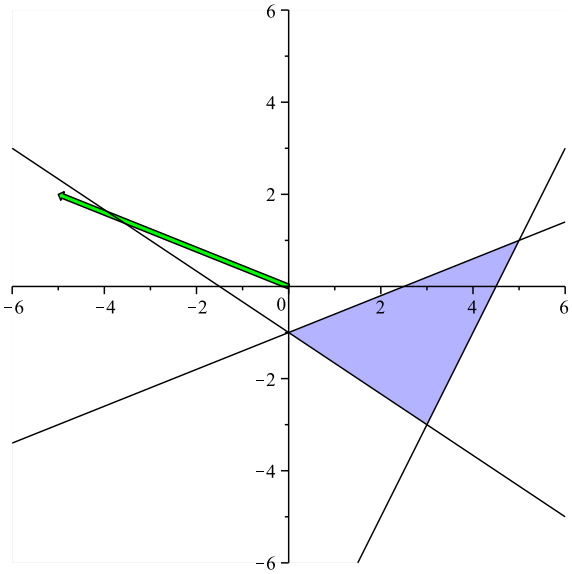
$$\begin{cases} 2x + 3y = -3, \\ 2x - y - 9 = 0, \\ 2x - 5y - 5 = 0. \end{cases}$$



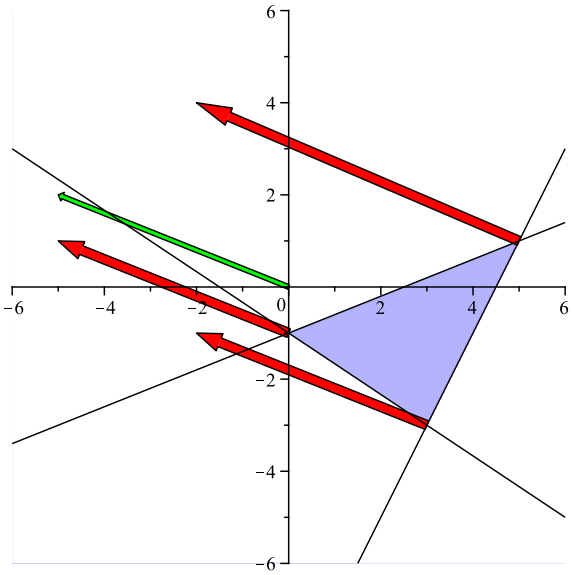
El triángulo sombreado es la solución del sistema.

Los vértices son los puntos $(0, -1)$, $(5, 1)$ y $(3, -3)$, puntos de cortes de las rectas.

Como la función es $F(x, y) = 2000x + 5000y$, el vector director es $\overrightarrow{v} = (-5000, 2000)$, que tiene la misma dirección que el $\overrightarrow{v} = (-5, 2)$ representado en la Figura anterior.



Se trazan las paralelas al vector que pasen por los vértices de la región factible, obteniéndose:



Se observa que de las tres paralelas, la que corta al eje y en un punto mayor es la que pasa por el vértice $(5, 1)$, que por tanto será la solución óptima al problema de máximos planteado.

Para saber cuál es este valor máximo sustituimos en la función:

$$F(5, 1) = 2000 \cdot 5 + 5000 \cdot 1 = 10000 + 5000 = 15000$$

.

Luego la función tiene su solución óptima en $(5, 1)$ donde toma el valor 15000.

Forma Algebraica

Consiste, simplemente, en sustituir cada uno de los vértices de la región en la función objetivo. La solución óptima vendrá dada por aquel que tome el mayor (o menor) valor.

En el ejemplo anterior, luego de hallar la región factible y los vértices $(0, -1)$, $(5, 1)$ y $(3, -3)$, éstos se sustituyen en la función F, obteniéndose:
$F(5, 1) = 2000 \cdot 5 + 5000 \cdot 1 = 10000 + 5000 = 15000$,
$F(0, -1) = 2000 \cdot 0 + 5000 \cdot (-1) = 0 - 5000 = -5000$,
$F(3, -3) = 2000 \cdot 3 + 5000 \cdot (-3) = 60000 - 15000 = -9000$.

Se puede notar que el valor máximo de F se alcanza para el vértice $(5, 1)$ y que dicho valor es 15000. La misma solución que se obtenía antes con la forma geométrica.

Un ejemplo típico de un problema que dio origen a la programación lineal es el problema del transporte:

Una empresa tiene 2 plantas de producción (P1 y P2) de cierto artículo que vende en 3 ciudades (C1,C2 y C3). En P1 se producen 5000 unidades, y en P2 7000 unidades. Estas 12000 unidades las vende así: 3500 es C1, 4000 en C2 y 4500 en C3. Los costos de transporte, en euros por unidad de producto, desde las plantas de producción a las ciudades son:

Envíos	Hasta C1	Hasta C2	Hasta C3
Desde P1	3	2.5	3.5
Desde P2	2.25	3.75	4

Determinar el número de artículos que debe enviar la empresa desde cada planta a cada ciudad para que los costos de transporte sean mínimos.

Para problemas de este tipo necesitamos tres variables.

Sea x=unidades de P1 a C1, y=unidades de P1 a C2 y z=unidades de P1 a C3. Tiene que verificarse entonces que $x + y + z = 5000$.

Si desde P1 a C1 se envían x unidades, como en C1 necesitan 3500, desde P2 se mandarán a C1 $3500 - x$. Razonando del mismo modo con y y z, se obtiene la tabla:

Envíos	Hasta C1	Hasta C2	Hasta C3
Desde P1	x	y	$z = 5000 - x - y$
Desde P2	$3500 - x$	$4000 - y$	$4500 - z = 4500 - (5000 - x - y)$

Se sustituyó $z = 5000 - x - y$ para transformar las 3 incógnitas en sólo 2.

Para obtener las restricciones imponemos que cada cantidad ha de ser mayor o igual que cero, es decir:

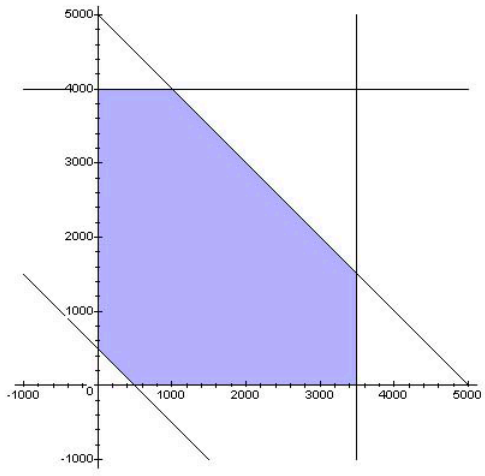
$$\begin{aligned} x &\geq 0, \\ 3500 - x &\geq 0, \\ y &\geq 0, \\ 4000 - y &\geq 0, \\ 5000 - x - y &\geq 0, \\ -500 + x + y &\geq 0. \end{aligned}$$

Por tanto el sistema de inecuaciones es:

$$\begin{cases} x \geq 0, \\ x \leq 3500, \\ y \geq 0, \\ y \leq 4000, \\ x + y \geq 5000, \\ x + y \geq 500. \end{cases}$$

Como se trata de minimizar costos, la función objetivo es:
$C(x, y) = 3 \cdot x + 2{,}5 \cdot y + 3{,}5 \cdot (5000 - x - y) + 2{,}25 \cdot (3500 - x) + 3{,}75 \cdot (4000 - y) + 4 \cdot (-500 + x + y)$
$C(x, y) = 1{,}25 \cdot x - 0{,}75 \cdot y + 22625.$

Siendo la región factible:



donde los vértices son A=$(0, 500)$, B=$(0, 4000)$, C=$(1000, 4000)$,D=$(3500, 1500)$,E=$(3500, 0)$y F=$(500, 0)$.
Sustituyendo, se obtiene:

$$\begin{aligned} C(0,500) &= 22250, \\ C(0,4000) &= 19625, \\ C(1000,4000) &= 20875, \\ C(3500,1500) &= 25873, \\ C(3500,0) &= 27000, \\ C(500,0) &= 23250. \end{aligned}$$

El mínimo se da en B, cuando $x = 0$ e $y = 4000$.Es decir, las unidades a distribuir son:

Envíos	Hasta C1	Hasta C2	Hasta C3
Desde P1	0	4000	1000
Desde P2	3500	0	3500

A todo problema de PL, llamado Problema Primal (PP), le corresponde otro denominado Problema Dual (PD). Entre ambos existen una serie de relaciones y propiedades importantes,a saber:

- El PD tiene tantas variables como restricciones tiene el PP.
- El PD tiene tantas restricciones como variables tiene el PP.
- Los coeficientes de la función objetivo del PP son los términos independientes (con su signo) de las restricciones del PD.
- Los términos independientes de las restricciones del PP(con su signo), son los coeficientes de la función objetivo del PD.
- La matriz de coeficientes de las restricciones del PD es igual a la traspuesta de la matriz de coeficientes de las restricciones del PP.
- A un PP de maximización, le corresponde un PD de minimización y viceversa.
- El Dual del PD es su Primal.
- El sentido de las desigualdades de las restricciones del PD y el signo de las variables del mismo problema, dependen de la forma de que tenga el signo de las variables del PP y del sentido de las restricciones del mismo problema.

A continuación, algunos ejemplos para entender mejor éstas relaciones y propiedades.

Ejemplos: En la Función objetivo, maximizar Z en PP equivale a minimizar $-Z$ en PD y minimizar Z (PP) a maximizar $-Z$ (PD).

(1) Max $Z = 3x_1 + 5x_2 + 6x_3$, equivale a: Min $-Z = -3x_1 - 5x_2 - 6x_3$.
(2) Min $Z = 2x_1 + 3x_2 + x_3$ equivale a: Max $-Z = -2x_1 - 3x_2 - x_3$.

Por otra parte, en relación con las restricciones:

$$AX \geq b \text{ equivale a } -AX \leq -b,$$
$$AX \leq b \text{ equivale a } -AX \geq -b,$$
$$AX = b \text{ equivale a } AX < b,\ AX > b.$$

Siendo la siguiente estructura canónica en la PL, el problema primal (PP):

$$\begin{aligned} MaxZ &= CX, \\ S.a: AX &\leq b, \\ X &\geq 0. \end{aligned}$$

Definimos el Problema Dual, asociado al problema anterior, de la siguiente forma:

$$\begin{aligned} MinG &= b^T Y, \\ S.a: A^T Y &\geq C^T, \\ Y &\geq 0. \end{aligned}$$

Se denomina a las anteriores formas *Problemas Duales Simétricos*, en contraposición con los *Asimétricos* o aquellos en que las restricciones aparecen con el signo igual, es decir: $AX = b$.

Apéndice 2

2. Distribución Gaussiana

La distribución Gaussiana, recibe también el nombre de distribución Normal, ya que una gran mayoría de las Variables Aleatorias continuas siguen esta distribución.

Una variable aleatoria X se dice que tiene distribución Normal o que está normalmente distribuida, con parámetros μ y σ^2 ($\sigma > 0$), si su función de densidad de probabilidad está dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}; \text{ para todo } x \in \mathbb{R}.$$

La función de distribución de una variable aleatoria normal es:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}} dx; \text{ para todo } t \in \mathbb{R},$$

y utilizaremos la notación $X \sim N(\mu, \sigma^2)$ para decir que X está normalmente distribuida con parámetros μ y σ^2.

En el caso en que $\mu = 0$ y $\sigma^2 = 1$ se dice que X tiene distribución normal estándar o unitaria y se utiliza la notación:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{(x)^2}{2}}; y$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp^{-\frac{(x)^2}{2}} dx;$$

para la función de densidad y la función de distribución de X respectivamente.