Indicadores ambientales sintéticos:

Una aproximación conceptual desde la estadística multivariante

Recibido para evaluación: 26 de Junio de 2007
Aceptación: 13 de Mayo de 2008
Recibido versión final: 06 de Mayo de 2008

Luis Alfonso Escobar Jaramillo[1]

RESUMEN

En este artículo se realiza una descripción general de la utilidad del análisis multivariante de datos, y formalmente se exponen dos metodologías (análisis de componentes principales y análisis de distancia P2) que usan técnicas de análisis multivariante para definir la dimensionalidad real de los datos para la estimación de indicadores sintéticos o índices de calidad ambiental.

PALABRAS CLAVE: Indicadores sintéticos, Índices ambientales, Análisis multivariante, Componentes principales, Análisis de distancia.

ABSTRACT

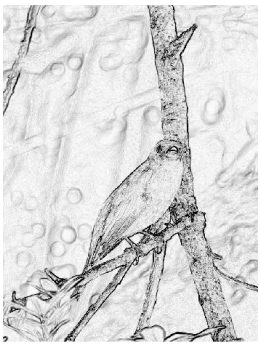
This paper presents a general description of multivariate statistical analysis and shows two methodologies: analysis of principal components and analysis of distance, DP2. Both methods use techniques of multivariate analysis to define the true dimension of data, which is useful to estimate indicators of environmental quality.

KEY WORDS: Recycler workers, Synthetic indicators, Environmental indices, Multivariate analysis, Principal components, Pistance analysis.

1. Economista Ph.D. Profesor Asociado de la Escuela de Ingeniería de Recursos Naturales y del Ambiente - EIDENAR, Facultad de Ingeniería, Universidad del Valle, Colombia.
lescobar@univalle.edu.co

1. INTRODUCCIÓN

La teoría estadística y los métodos aplicados a la elaboración de indicadores sintéticos por entidades del orden local, regional e internacional, recomiendan el uso de análisis multivariante para el tratamiento de datos que describen fenómenos que pueden ser explicados en común (Segnestan, 2002a, Castro, 2004). Se argumenta que pasar de la recopilación de indicadores simples a la generación de indicadores sintéticos es un imperativo para simplificar los datos y estructurarlos en información más elaborada, de acuerdo a las necesidades de los tomadores de decisión y el público en general [Polanco (2006); MMA (1996; 2000)].



En este artículo se realiza una descripción general de la utilidad del análisis multivariante de datos, empleando dos técnicas, en ocasiones complementarias: análisis de componentes principales y análisis de distancia P2 con el fin de definir la dimensionalidad real de los datos disponibles, para la estimación de un indicador sintético o índice de calidad ambiental a nivel urbano. El empleo del análisis multivariante para mejorar la interpretación de problemas no es nuevo en Colombia, trabajos realizados para la reducción de datos y la dimensionalidad de objetos de investigación pueden encontrarse en Ospina y Lema (2005) quien emplea Análisis de Componentes Principales (ACP) y Análisis de Factores para encontrar las relaciones entre variables e impactos y, calificar y clasificar los impactos por su grado de influencia. Escobar y Bermúdez (2004) presentan un índice de calidad ambiental para las localidades de Bogotá, empleando ACP, indicando con este estudio que la ciudad no es ambientalmente homogénea según el índice construido y que es posible priorizar las inversiones ambientales de acuerdo al valor resultante del índice. En este mismo sentido Escobar (2004; 2006a; 2006b) presenta los detalles de la aplicación del índice de calidad ambiental para la ciudad de Cali. A nivel internacional, se recomienda la revisión del estudio de Castro (2004) quien hace una excelente revisión bibliográfica sobre el uso de técnicas de análisis multivariante para la construcción de índices de desarrollo sostenible, y emplea las técnicas de ACP, Distancia P2 y lógica difusa para estimar índices de desarrollo sostenible para los municipios de la comunidad de Andalucía (España).

2. MÉTODOS DE ANÁLISIS MULTIVARIANTE

Es posible que cualquier investigador se enfrente a teorías científicas o criterios consensuados sobre algún aspecto a investigar, por ejemplo el desarrollo sostenible, de tal forma que en su explicación, se incluyan los componentes económicos, sociales y ambientales. En cada uno de estos componentes se pueden describir un conjunto grande, complicado y complejo de datos que representan las variables que explican el desarrollo sostenible en las diferentes unidades de observación (por ejemplo países, ciudades, comunas y barrios, etc). Es aquí donde tiene sentido el uso de métodos de análisis multivariante dado que ayudan al investigador a resumir grandes cantidades de variables, que pueden estar correlacionadas, por medio de relativamente pocos componentes que los simplifican.

Los métodos de análisis multivariante son un conjunto de técnicas de análisis de datos que permiten al investigador testar la utilidad conjunta de los datos que se emplean en la explicación de los fenómenos de interés analítico, como el bienestar social, la calidad ambiental urbana y el desarrollo sostenible, etc. (Dixon et al., 2002; Pardo et al., 2002; Visauta et al., 2003).

Para un empleo acertado de los métodos de análisis multivariante es pertinente primero definir la unidad de observación (barrios, comunas, municipios y países, etc) sobre la cual recae dicho análisis. Las unidades de observación son la base para el análisis del fenómeno que se quiere estudiar, dado que ello exige una medición y evaluación de los datos al nivel de agregación de la unidad seleccionada.

La aplicación de métodos de análisis multivariante en la medición de la calidad ambiental, el desarrollo sostenible, el bienestar social, la calidad de vida, etc, como una variable latente, conduce a la elaboración de un indicador sintético que resume la información contenida en múltiples datos, dado que este es determinado por una variedad de indicadores simples que inciden positiva y negativamente en su valoración.

Tipo de Problema	ACP	AF	ACC	AVC	ADC	DP_2	AD	AA	MANOVA
Exploración de las relaciones entre variables	*A veces*	*Indudablemente*	*A veces*	*Rara vez*	*Nunca*	*No lo hace*	*Nunca*	*Nunca*	*Nunca*
Cribado de datos	*Indudablemente*	*A veces*	*Nunca*	*Nunca*	*Nunca*	*No lo hace*	*Nunca*	*A veces*	*Nunca*
Creación de nuevas variables	*Lo hace*	*Lo hace*	*Lo hace*	*Lo hace*	*Lo hace*	*Lo hace*	*No lo hace*	*No lo hace*	*No lo hace*
Predicción de ser miembro de un grupo	*No lo hace*	*No lo hace*	*No lo hace*	*No lo hace*	*Lo hace*	*No lo hace*	*Lo hace*	*Lo hace*	*No lo hace*
Comparación de medidas grupales	*Posiblemente*	*Posiblemente*	*No lo hace*	*Lo hace*	*Rara vez*	*No lo hace*	*Rara vez*	*No lo hace*	*Lo hace*
Comparación de grupos de variables	*Posiblemente*	*Posiblemente*	*Indudablemente*	*Nunca*	*Nunca*	*No lo hace*	*Nunca*	*Nunca*	*Nunca*
Verificación de agrupamientos	*Indudablemente*	*Posiblemente*	*Nunca*	*Nunca*	*Nunca*	*Indudablemente*	*Nunca*	*Indudablemente*	*Nunca*
Reducción de dimensionalidad	*Indudablemente*	*Indudablemente*	*Indudablemente*	*Indudablemente*	*Indudablemente*	*Indudablemente*	*Nunca*	*Nunca*	*Nunca*
Creación de variables significativas	*No es probable*	*For lo comin*	*No es probable*	*No es probable*	*Posiblemente*	*Lo hace*	*Nunca*	*Nunca*	*Nunca*

Tabla 1.

Lista cruzada de métodos de análisis multivariante y tipos de problemas

Es importante aclarar que las técnicas de análisis multivariante se caracterizan por su tendencia exploratoria en lugar de confirmatoria. De esta forma, se debe entender que la estadística convencional exige al investigador comprobar hipótesis con el uso de los datos, mientras que con el análisis multivariante el investigador intenta derivar de los datos una explicación relevante y consistente, dando respuesta a si existe alguna información valiosa en la estructura de los datos (Johnson, 2000).

3. DESCRIPCIÓN GENERAL DE LAS TÉCNICAS DE ANÁLISIS MULTIVARIANTE

Se puede determinar dos grandes categorías de técnicas de análisis multivariante:

- Las técnicas dirigidas por variables que determinan las correlaciones que podrían existir entre las variables respuestas, a través del análisis de la matriz de correlación. Algunos ejemplos de estas técnicas son el Análisis de Componentes Principales (ACP)[1], el Análisis por Factores (AF), el Análisis de Regresión (AR), el Análisis de Correlación Canónica (ACC), Análisis de Variables Canónicas (AVC), Análisis Discriminante Canónico (ADC) y el Análisis de Distancia P_2 (DP_2)[2].
- Las técnicas dirigidas a las unidades de observación se dirigen a determinar la relación que podría existir entre éstas. Ejemplo de estas técnicas son el Análisis Discriminante (AD), el Análisis por Agrupación (AA) y el Análisis Multivariado de la Varianza (MANOVA).

En la Tabla 1, se sintetizan las principales técnicas de análisis multivariante y su relevancia en la aplicación a problemas específicos[3].

En este artículo se presenta el marco conceptual para estimar indicadores sintéticos, aplicando las técnicas de ACP y DP_2. La selección de estas técnicas obedece a que el análisis que se realiza para estimar indicadores sintéticos consiste, a grandes rasgos, en la exploración de la correlación entre los indicadores simples, la reducción de la dimensionalidad de los datos, la agrupación de los indicadores a través de componentes que incorporen la mayor parte de la varianza contenida en los datos y por último la calificación y ordenación de las unidades de observación, que en el caso de este estudio se refieren a Comunas de la ciudad de Cali.

4. ASPECTOS TÉCNICOS PARA LA CONSTRUCCIÓN DE LOS INDICADORES SINTÉTICOS

4.1. Algunas precisiones acerca de los indicadores sintéticos

El indicador sintético o índice está formado por una serie de componentes que aportan información valiosa acerca del objetivo a medir, ya sea la calidad ambiental, el desarrollo sostenible, etc., en las distintas unidades de observación (Zarzosa, 1996:65).

Es deseable que la información recopilada para la medición de los componentes sea objetivamente medida, dado que es una expresión matemática (un dato) acerca del estado del mismo. Una medición subjetiva puede conducir erróneamente a múltiples estados y por lo tanto puede afectar de distintas formas y magnitudes el índice[4].

Esta medición de los componentes indica la magnitud del estado del mismo. Por ello técnicamente se denomina indicador simple o parcial. Este indicador puede ser un dato o la combinación de más de uno para conformar o calificar el estado de un componente de interés. Por ejemplo en el modelo presentado en la Tabla 2, Escobar (2006a) presenta 38 indicadores simples que determinan la calidad ambiental urbana, del cual se deduce que los datos originales con que se valorarían estos indicadores simples son ampliamente mayores. Imagine el indicador densidad de área verde por habitante, que demandaría la combinación matemática de dos tipos de datos: áreas verdes (en m^2) dividido la población total en una unidad de observación. En resumen, el indicador sintético o índice es una combinación matemática de los indicadores parciales.

1. Son extensas las referencias de trabajos, tanto conceptuales como empíricos, que describen las propiedades estadísticas de esta técnica de análisis multivariante. En el campo teórico y conceptual se pueden citar Anderson (1984) y Johnson (2000), entre otros. En campo del uso aplicado ver Pena Trapero (1977), Zarzosa (1996) y Castro (2004) para España. Ospina y Lema (2005), Escobar y Bermudez (2004) y Escobar (2004; 2006a; 2006b) en Colombia.

2. Los desarrollos de esta metodología son recientes y parten del trabajo seminal de Pena Trapero (1977).

3. Para más detalle sobre las técnicas de análisis multivariante, puede consultar Anderson (1984), Hair et al., (1999) y Johnson (2000). Aquí podrá encontrar los procedimientos generales para el análisis de datos en cada una de las técnicas descritas. Para detalles respecto a DP2, ver Pena Trapero (1977) y Zarzosa (1996).

Tabla 2.

Sistema de indicadores de calidad ambiental urbana

Indicadores simples	Área temática	Componentes	Índice
Nivel I	Nivel II	Nivel III	Nivel IV
RS1. Residuos sólidos generados (Ton)			
RS2. Cobertura de recolección (% de viviendas)	Residuos sólidos urbanos (*Irsu*)		
RS3. Basureros crónicos (No)			
RS4. Percepción social del servicio			
CE1. Cobertura de energía (% de viviendas)			
CE2. Conexiones ilegales (No.)			
CE3. Cobertura de gas (% de viviendas)	Consumo de energía (*Icen*)		
CE4. Uso de leña, carbón, etc. (No. viviendas)			
CE5. Fuentes de contaminación lumínica (No)		Índice de Flujo Urbano (IFLU)	
TR1. Densidad de coches (Coches/habitantes)			
TR2. Accidentes mortales de transito (No)	Trafico urbano (*Itru*)		
TR3. Kilómetros de vía pavimentada (No)			
TR4. Semáforos (No)			
CV1. Personas por vivienda (No)			
CV2. Mts^2 por vivienda (No)	Calidad de la vivienda (*Icav*)		
CV3. Densidad de viviendas (No/hectárea)			
AI1. Concentración de NOx (% del territorio)*			
AI2. Concentración de SOx (% del territorio)*			
AI3. Concentración de CO_2 (% del territorio)*	Aire (*Iair*)		Índice de Calidad Ambiental (ICA)
AI4. Concentración de material particulado*			
AI5. Denuncias por olores molestos (No)			
AG1. Cobertura de agua (% de viviendas)			
AG2. Cobertura de alcantarillado (% de viviendas)	Agua (*Iagu*)		
AG3. Riesgo de inundación (% del territorio)**			
RU1. Ruido diurno (% del territorio)*			
RU2. Ruido nocturno (% del territorio)*	Ruido (*Irui*)		
RU3. Denuncias ciudadana (No)			
SU1. Área urbanizable construida (% del total)		Índice de Medio Ambiente Urbano (IMAU)	
SU2. Área verde institucional (% del total)			
SU3. Densidad área verde (mt^2/habitante)	Suelo (*Isue*)		
SU4. Erosión (% del total)*			
ES1. Denuncias invasión del espacio público (No)			
ES2. Parques y plazas por localidad (No)	Espacio público (*Iepu*)		
ES3. Andenes peatonales (Mt^2)			
BI1. Densidad de árboles (No/habitante)			
BI2. Árboles sembrados (No/año)			
BI3. Fauna animal no nociva (No)	Biodiversidad (*Ibio*)		
BI4. Especies vegetales (No)			
BI5. Sitios de interés ecológico (No)			

** Estas variables pueden ser derivadas empleando un sistema de información geográfico, una vez sea diseñado un modelo cartográfico que permita estimar el porcentaje del territorio que supera los estándares. Para el caso del modelo aplicado solo se cuenta con información georefenciada para la concentración de material particulado. El proceso técnico consiste en el cruce de un mapa digital que represente el modelo de dispersión del contaminante en la zona urbana con el mapa de división político administrativa de la ciudad en comunas, derivando el porcentaje del territorio que supera los estándares.*

*** Al igual que con las variables que se derivan de modelos de dispersión, esta variable se puede estimar cruzando los mapas de riego de inundación y división político administrativas de la ciudad, previo a la construcción de un modelo cartográfico que identifique los procesos técnicos para derivar tal información.*

Fuente: Escobar (2006a)

De acuerdo a lo anotado hasta aquí, el éxito de un buen índice depende de unos buenos indicadores simples y a su vez éstos dependen de un buen conjunto de datos.

4.2. Supuestos relevantes para la formulación del índice

Medir la calidad ambiental, el desarrollo sostenible o cualquier otra variable latente, como la agregación de un conjunto de datos e indicadores simples en un índice, demanda la formulación de una serie de supuestos que aseguran la consistencia de los resultados, partiendo de la adecuada selección de los datos (Ebert et al, 2004; Segnestam, 2002a, 2002b; MMA, 1996, 2000). Estos supuestos consisten en:

- Completitud: Los indicadores seleccionados para construir el índice de calidad ambiental son todos los que explican el objetivo a medir[5].
- Bondad de los indicadores que determinan el índice: Los indicadores seleccionados deben medir adecuadamente los atributos que se describan en relación con el índice a estimar, tanto en el espacio como en el tiempo[6].
- Objetividad en la valoración de los indicadores: Los datos que se utilicen deben ser una representación objetiva de la valoración de su estado. Es decir, están ausentes de sus propiedades intrínsecas, los juicios de valor con respecto a su magnitud en el espacio y el tiempo.

Con los supuestos planteados se han asegurado las condiciones ideales de los datos utilizados. El siguiente paso es determinar las propiedades matemáticas que garantizan la consistencia de un indicador sintético. Al respecto Pena Trapero (1977) y Zarzosa (1996) presentan una serie de condiciones que a priori debe cumplir el indicador sintético para que sea consistente.

I. Existencia y Determinación. La función matemática que define el indicador sintético debe ser tal que exista y tenga solución para todos los valores del índice en cada unidad de observación.

II. Monotonía. El índice debe responder positivamente a una modificación positiva de los componentes y negativamente a una modificación negativa. En el modelo presentado en la tabla 3 Eq. (1 a 13)[7] se define el sentido que tienen los componentes de primer nivel cuando entran al indicador de segundo, tercer y cuarto nivel. En la práctica, esto exige que en el análisis multivariado se introduzca la información con el signo con que se espera estén relacionados los indicadores simples y el objetivo a medir.

III. Unicidad. El índice debe representar un único valor, para una situación dada.

IV. Invariancia. Como consecuencia de la propiedad III, el índice debe ser invariante respecto a un cambio de origen y/o de escala de medición de los indicadores simples.

V. Homogeneidad. La función matemática que defina el índice, debe ser homogénea de grado uno. Esto asegura que si los indicadores simples aumentan o disminuyen, en igual proporción lo hace el indicador sintético.

$$f(C\cdot I_1, C\cdot I_2, ..., C\cdot I_n) = C\cdot\ f(I_1, I_2, ..., I_n)$$

VI. Transitividad. Suponga que a, b y c son tres estados distintos medidos por el índice e I(a), I(b) e I(c) son valores del índice correspondiente a esos tres estados. Debe verificarse que:

$$\left.\begin{array}{l} I(a) > I(b) \\ I(b) > I(c) \end{array}\right\} \Rightarrow I(a) > I(c)$$

VII. Exhaustividad. El índice debe ser tal que aproveche al máximo y de forma útil la información suministrada por cada uno de los indicadores simples[8].

4. La objetividad en la medición de los componentes asegura la no violación de supuestos como el de invarianza del índice. Más adelante se detallan los postulados e hipótesis que se deben exigir a un índice para que sea consistente.

5. Este supuesto es bastante fuerte para el análisis multivariado aplicado a indicadores sintéticos sociales y ambientales, dado que su formulación teórica no está definida y depende en gran parte de criterios y acuerdos sociales, que pueden variar entre regiones o unidades de observación. Por ello, este supuesto debe ser relajado, para permitir que los criterios de selección de las variables para indicadores sociales tengan la flexibilidad que demanda el trabajo de formulación de indicadores simples. Para detalles sobre el concepto de flexibilidad, puede consultarse MMA (1996:15) y Milon et al, (1995:5).

6. Los indicadores seleccionados y los datos que lo valoran, deben ser una buena estimación de sus propiedades intrínsecas. Por lo tanto se asume que la base de datos que se construya debe cumplir con este supuesto.

7. Este modelo es tomado del trabajo de investigación tutelado y tesis doctoral del autor: Escobar (2004; 2006a; 2006b), en el que se construye un índice de calidad ambiental a un nivel de Comunas en la ciudad de Cali, para derivar el valor económico que los agentes sociales le asignan a contar con una determinada calidad ambiental relacionada con el entorno inmediato a su vivienda.

8. Zarzosa (1996:68) sostiene que esta propiedad garantiza que el índice haga una buena utilización de la información contenida en los indicadores parciales y elimine la duplicación de información existente en más de dos indicadores simples que explican el indicador sintético.

Tabla 3.
De indicadores de primer nivel a indicadores sintéticos

Ecuación	Hipótesis de relación lineal	
$Irsu_j = \sum_{j=1}^{n} W_i RS_{ij}$	$\frac{\partial Irsu_j}{\partial RS_{1j}} < 0 \; ; \; \frac{\partial Irsu_j}{\partial RS_{2j}} > 0 \; ; \; \frac{\partial Irsu_j}{\partial RS_{3j}} < 0 \; ; \; \frac{\partial Irsu_j}{\partial RS_{4j}} > 0$	*Eq.* (1)
$Icen_j = \sum_{j=1}^{n} W_i CE_{ij}$	$\frac{\partial Icen_j}{\partial CE_{1j}} > 0 \; ; \; \frac{\partial Icen_j}{\partial CE_{2j}} < 0 \; ; \; \frac{\partial Icen_j}{\partial CE_{3j}} > 0 \; ; \; \frac{\partial Icen_j}{\partial CE_{4j}} < 0 \; ; \; \frac{\partial Icen_j}{\partial CE_{5j}} < 0$	*Eq.* (2)
$Itra_j = \sum_{j=1}^{n} W_i TR_{ij}$	$\frac{\partial Itra_j}{\partial TR_{1j}} < 0 \; ; \; \frac{\partial Itra_j}{\partial TR_{2j}} < 0 \; ; \; \frac{\partial Itra_j}{\partial TR_{3j}} < 0 \; ; \; \frac{\partial Itra_j}{\partial TR_{4j}} < 0$	*Eq.* (3)
$Icav_j = \sum_{j=1}^{n} W_i CV_{ij}$	$\frac{\partial Icav_j}{\partial CV_{1j}} < 0 \; ; \; \frac{\partial Icav_j}{\partial CV_{2j}} > 0 \; ; \; \frac{\partial Icav_j}{\partial CV_{3j}} < 0$	*Eq.* (4)
$Iair_j = \sum_{j=1}^{n} W_i AI_{ij}$	$\frac{\partial Iair_j}{\partial AI_{1j}} < 0 \; ; \; \frac{\partial Iair_j}{\partial AI_{2j}} < 0 \; ; \; \frac{\partial Iair_j}{\partial AI_{3j}} < 0 \; ; \; \frac{\partial Iair_j}{\partial AI_{4j}} < 0 \; ; \; \frac{\partial Iair_j}{\partial AI_{5j}} < 0$	*Eq.* (5)
$Iagu_j = \sum_{j=1}^{n} W_i AG_{ij}$	$\frac{\partial Iagu_j}{\partial AG_{1j}} > 0 \; ; \; \frac{\partial Iagu_j}{\partial AG_{2j}} > 0 \quad \frac{\partial Iagu_j}{\partial AG_{3j}} > 0$	*Eq.* (6)
$Irui_j = \sum_{j=1}^{n} W_i RU_{ij}$	$\frac{\partial Irui_j}{\partial RU_{1j}} < 0 \; ; \; \frac{\partial Irui_j}{\partial RU_{2j}} < 0 \; ; \; \frac{\partial Irui_j}{\partial RU_{3j}} < 0$	*Eq.* (7)
$Isue_j = \sum_{j=1}^{n} W_i SU_{ij}$	$\frac{\partial Isue_j}{\partial SU_{1j}} > 0 \; ; \; \frac{\partial Isue_j}{\partial SU_{2j}} > 0 \; ; \; \frac{\partial Isue_j}{\partial SU_{3j}} > 0 \; ; \; \frac{\partial Isue_j}{\partial SU_{4j}} < 0$	*Eq.* (8)
$Iepu_j = \sum_{j=1}^{n} W_i EP_{ij}$	$\frac{\partial Icen_j}{\partial EP_{1j}} < 0 \; ; \; \frac{\partial Icen_j}{\partial EP_{2j}} > 0$	*Eq.* (9)
$Ibio_j = \sum_{j=1}^{n} W_i BI_{ij}$	$\frac{\partial Ibio_j}{\partial BI_{1j}} > 0 \; ; \; \frac{\partial Ibio_j}{\partial BI_{2j}} > 0 \; ; \; \frac{\partial Ibio_j}{\partial BI_{3j}} > 0 \; ; \; \frac{\partial Ibio_j}{\partial BI_{4j}} > 0 \; ; \; \frac{\partial Ibio_j}{\partial BI_{5j}} > 0$	*Eq.* (10)
$I_{mj} = \frac{\sum_{i=1}^{r} Z_{rj} \cdot \sqrt{\lambda_r}}{\sum_{i=1}^{r} \sqrt{\lambda_r}}$		Eq. (11)
$IU_{mj} = \frac{\sum_{i=1}^{r} Z_{rj} \cdot \sqrt{\lambda_r}}{\sum_{i=1}^{r} \sqrt{\lambda_r}}$		Eq. (12)
$ICA_j = \frac{\sum_{i=1}^{r} Z_{rj} \cdot \sqrt{\lambda_r}}{\sum_{i=1}^{r} \sqrt{\lambda_r}}$		*Eq.* (13)

Los j = 1, 2, 3,.., n representan las unidades experimentales. Las *i* representan el subíndice de cada una de las variables que componen la ecuación de cada indicador de segundo nivel y que se detallan en la tabla 2. W_i es el conjunto de pesos relativos de cada indicador de primer nivel. Como las variables que componen cada indicador de segundo nivel tienen diferentes unidades de medida y escala, se emplean técnicas de análisis multivariantes como el Análisis de Componentes Principales (ACP) (Johnson, 2000) y Análisis de Distancia P2 (DP2) (Pena Trapero, 1977; Zarzosa, 1996).

Fuente: Escobar (2006a)

Una vez definidas las propiedades matemáticas que deben cumplir los indicadores simples y el índice, se debe resolver el problema de agregación dado que lo más probable es que cada uno de los componentes del índice tenga distintas unidades de medida.

Se propone que técnicamente cualquier proceso de tipificación pueda transformar todos los componentes en las mismas unidades y escala, asegurando que esta transformación no modifique el orden del estado del indicador en las distintas unidades de observación.

Otro problema que quedaría por resolver es cómo ponderar los indicadores simples en un indicador agregado o sintético. Ello demandaría un sistema de ponderación que otorgue peso o importancia a cada indicador simple. Algunos criterios podrían ser:

- Mediante la técnica delphi es posible determinar ponderaciones de expertos.
- Desde el punto de vista de los objetivos sociales, se puede asignar ponderaciones de acuerdo a la importancia de cada uno de los indicadores simples.
- Asignar igual ponderación a cada uno de los indicadores simples.
- El peso de cada indicador simple vendría dado por la información útil que tuviera cada uno de acuerdo a la varianza explicada del índice[9].

Este último criterio es el que emplea la mayoría de estudios que estiman índices, a través de los métodos de análisis multivariante que se emplean en este artículo: ACP y DP_2.

Se debe entender que el empleo de estas técnicas multivariante tiene como objeto estimar el valor de un indicador sintético que resuma o determine las dimensiones reales de todas las variables o indicadores simples, de acuerdo a un modelo conceptual formulado (desarrollo sostenible, calidad ambiental, etc,.). Este valor, para el caso de las metodologías que se trabajan en este estudio, presenta una combinación lineal del conjunto de variables respuestas, tal como se describen en las trece ecuaciones presentadas en la Tabla 3 (Escobar, 2006a).

5. ANÁLISIS DE COMPONENTES PRINCIPALES[10]

En muchas situaciones, los científicos de todas las disciplinas se enfrentan al manejo de un conjunto amplio de datos que representan y explican el comportamiento de un vasto número de unidades de observación. En estos casos, el uso de técnicas multivariante para el análisis de los datos debe partir inicialmente de un examen de los mismos, asegurando el cumplimiento de las propiedades estadísticas básicas para realizar análisis entre una variable latente (el índice) y las variables respuestas (indicadores simples).

Es probable que las variables respuestas puedan presentar propiedades no deseables para los análisis multivariados. Por ello se requiere muchas veces tratamientos previos del conjunto de datos, que prueben si la distribución de este conjunto de variables es una distribución normal multivariada. Si esto no es así, es altamente deseable una transformación del conjunto de variables. Este es el espíritu que hay en el trasfondo del ACP.

5.1. Definición

El ACP es un procedimiento matemático que transforma un conjunto de variables respuestas correlacionadas en un conjunto menor de variables ortogonales (no relacionadas) llamadas componentes principales que tienen como fundamento explicar la mayor parte de la varianza contenida en los datos originales[11]. En este sentido, los componentes principales son combinaciones lineales de las variables originales[12].

5.2. Utilidades del análisis de componentes principales

La utilidad del ACP se ha resumido en la Tabla 1 donde se define el tipo de problemas que generalmente resuelve esta técnica. En este orden de ideas, Johnson (2000) se refiere a la

9. Muchos autores afirman que asignar ponderaciones con base en correlaciones empíricas, puede llevar a resultados que distan de la realidad. Por ello es importante valorar los otros criterios, según sea el índice que se esté calculando.

10. Los desarrollos teóricos de esta técnica se pueden remontar a los trabajos Hotelling (1933) y Rao (1965). Manuales y trabajos aplicados se pueden encontrar en Jolliffe (1986), Johnson (2000) y Hair et al. (1999).

11. Muchos autores advierten que si las variables ya están casi no correlacionadas, el analista no gana nada con aplicar ACP porque la dimensionalidad real del conjunto de datos es igual al número de variables respuestas. Ello demanda la necesidad de aplicar una prueba estadística previa que defina si los datos siguen una distribución normal multivariada. Para detalles de estas pruebas se puede consultar a Johnson (2000).

12. Esta propiedad de relaciones lineales entre las variables respuestas es altamente aplicado al marco conceptual del índice, en la medida en que el modelo utilizado parta de relaciones lineales causa - efecto para explicar fenómenos sociales y ambientales.

utilidad del ACP por las siguientes funciones:

- Cribado de datos multivariados. El ACP es útil para el análisis previo de cualquier tipo de análisis multivariado, y puede ayudar a revelar anormalidades en los datos y a descubrir datos atípicos (outliers).
- Agrupación de unidades de observación. El ACP ayuda el investigador a detectar subgrupos de acuerdo al comportamiento común de la varianza de las variables respuesta.
- Análisis discriminantes. En este tipo de análisis, la matriz de variaza- covarianza requiere ser invertida para realizar una regla de discriminación. Cuando el número de variables respuestas es superior al número de unidades de observación, la matriz no se puede invertir. Allí juega un papel importante el ACP para reducir todas las variables respuestas en un número menor de componentes principales que expliquen el mayor porcentaje de varianza de los datos originales.
- Regresión. El ACP ayuda a determinar y corregir los problemas derivados de regresiones que presenten problemas de multicolinealidad.

5.3. Objetivos del análisis de componentes principales

Son dos los objetivos más relevantes asociados al análisis de componentes principales:

1. Reducir la dimensionalidad de un conjunto de datos, descubriendo la verdadera dimensión contenida en ellos.

   Cuando la dimensionalidad real de los datos es inferior al número original de los datos, entonces el análisis sugiere remplazar estos últimos por los componentes principales estimados, ayudando a mejorar la interpretación de los datos[13].

2. Identificar nuevas variables significativas contenidas en la estructura de los datos.

   La creación de nuevas variables componentes principales es una combinación lineal de las variables originales que debe seguir un orden de importancia, teniendo en cuenta que éstas no están correlacionadas. La primera componente explica la mayor varianza posible en los datos y cada componente adicional adquiere la mayor variabilidad posible restante.

La mayoría de manuales y estudios aplicados de análisis multivariado presentan dos procedimientos generales para definir los componentes principales. Mediante el análisis de la matriz de varianza covarianza y a través de la matriz de correlación (P)[14]. Por ejemplo, para la definición formal de la estructura analítica que define los componentes principales, en el estudio de Escobar (2006b) se desarrolla el análisis de la matriz de correlaciones, aclarando que el uso de la matriz de varianza-covarianza es pertinente cuando el investigador está seguro de que todas las variables del análisis guardan las siguientes condiciones:

- Todas las variables deben estar medidas en las mismas unidades o comparables.
- Las unidades deben tener varianzas similares.

El argumento de estas exigencias para el uso de la matriz de varianza covarianza es que si alguna de las variables respuesta tiene una unidad de medida o escala distinta, ésta puede tener un efecto significativo sobre los componentes principales, presentándola como importante, cuando en realidad no lo sea[15].

Cuando las variables no presentan fundamentos de escala similares, es necesario la aplicación del ACP a la matriz de correlación de las **P** variables respuestas tipificadas[16] y no a los datos originales. Por lo tanto, exige al investigador una transformación previa de los datos de tal forma que garanticen una escala y unidad de medida común[17].

5.4. ACP a partir de la matriz de correlaciones

Estimar los componentes principales a partir de la matriz de correlaciones implica un análisis matricial de los datos tipificados, reduciendo la estructura de ellos a una misma unidad de escala. A continuación se presenta formalmente el procedimiento matemático y los algoritmos

13. En concreto, el procedimiento realizado define geométricamente un subespacio creado con las "m" primeras componentes, mejorando el ajuste de toda la información de las variables utilizadas mediante la estimación de la suma de los cuadrados de las distancias perpendiculares desde cada punto al subespacio determinado. Es importante reiterar que el ACP agrega valor a los datos, cuando la dimensionalidad de p variables es mayor que los m componentes determinados.

14. Puede consultarse algunos manuales en Anderson (1984); Hair, et al (1995); Jolliffe (1986) y Johnson (2000).

15. Esto es lo que autores como Jolliffe (1986) llaman sesgo de componentes iníciales. Johnson (2000) afirma que "..si una de las variables tiene una varianza mucho mayor que las demás, dominará la componente principal, sin importar la estructura de las covarianzas de las variables, y, en este caso tiene poco objeto la realización de un PCA".

16. Muchos autores recomiendan tener cuidado cuando se elige algún criterio de tipificación para el conjunto de variables, dado que con ello se está diciendo a priori que todas las variables respuestas tienen igual importancia. Es conveniente por ello hacer un examen de los datos, de tal forma que permita al investigador observar la relevancia de utilizar la matriz de varianza-covarianza o la matriz de correlación (P).

17. Existen distintos criterios de tipificación. Para una revisión al respecto puede ver Johnson (2000).

asociados a la estimación de las m componentes principales.

Sea $X = (X_{n1}, X_{n2}, .., X_{np})$ un conjunto de **p** variables respuestas que representan **n** observaciones seleccionadas de una población con media, y matriz de correlación **P**.

La matriz de correlaciones **P**, derivada de la correlación lineal de cada par de variables originales, está dada por:

$$P = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

En su forma matricial ampliada las componentes principales están definidas por y son una combinación lineal de las p variables originales, donde a_{1j} representa las ponderaciones, de tal forma que se puede representar como:

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_p \end{bmatrix} = \begin{bmatrix} a_{11}X_1 & a_{12}X_2 & \dots & a_{1p}X_p \\ a_{21}X_1 & a_{22}X_2 & \dots & a_{2p}X_p \\ \dots & \dots & \dots & \dots \\ a_{p1}X_1 & a_{p2}X_2 & \dots & a_{pp}X_p \end{bmatrix} \} \Rightarrow Z = AX$$

En forma resumida, el primer componente estaría determinado por $Z_1 = a_1'(X - \mu)$, en donde a_1 se elige teniendo en cuenta que la varianza de $a_1'(X - \mu)$ se maximice sobre todos los vectores a_1 que satisfagan $a_1'a_1 = 1$

Como la condición más importante que se le ha impuesto al primer componente es que maximice la varianza contenida en todos los datos, esta ocurriría cuando a_1 es un autovector de P correspondiente al autovalor λ_1 y que satisface $a_1'a_1 = 1$ Esto implica maximizar la función objetivo, que es la varianza de Z_1, de tal forma que $Var(Z_1) = \frac{\sum_{j=1}^{n} Z_{1j}^2}{n} = \frac{1}{n} u_1' X' X u_1 = u_1' \left[\frac{1}{n} X' X\right] u_1$ donde $\left[\frac{1}{n} X' X\right]$ es equivalente a la matriz **P** de las variables estandarizadas. En este caso la expresión de la varianza del primer componente se describe como $Var(Z_1) = a_1' P a_1$. Esta función se maximiza, sujeta a la restricción $a_1'a_1 = 1$ y resolviendo el siguiente lagrangiano: $L = a_1' P a_1 - \lambda (a_1' a_1 - 1)$. Derivando con respecto a a_1 e igualando a cero se obtiene: $\frac{\partial L}{\partial u_1} = 2 P a_1 - 2 \lambda a_1 = 0$; $(P - \lambda I) a_1 = 0$. La solución de esta expresión permite derivar los autovectores y autovalores del primer componente a_1 Y λ_1 respectivamente (Johnson, 2000).

La segunda componente principal se define por $Z_2 = a_2^{'}(X-\mu)$, en donde a_2 se elige teniendo en cuenta que la varianza de $a_2^{'}(X-\mu)$ se maximice entre todas las combinaciones lineales de X que no estén correlacionadas con la primera componente principal ($a_2^{'}a_1 = 0$) y que satisfagan $a_2^{'}a_2 = 1$. Dando como resultado el autovector a_2 y el autovalor λ_2. Esto garantiza, como se ha indicado anteriormente, que el segundo componente capturará la mayor parte de la varianza restante en los datos, luego de obtener el primer componente.

En forma general, cada uno de los j-ésimos $(j = 3,4,...,p)$ componentes principales restantes se puede representar por $Z_j = a_1^{'}(X-\mu)$ en donde a_j se elige teniendo en cuenta que la varianza de $a_j^{'}(X-\mu)$, se maximice entre todas las combinaciones lineales de X que no estén correlacionadas con la j-1 componentes principales y que satisfagan $a_j^{'}a_j = 1$. Obteniendo los j-ésimos autovalor λ_j y autovector a_j mas grande de **P**.

De lo anterior se generaliza que los autovalores y autovectores de P se expresan así $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p$ Y $a_1, a_2, ..., a_p$, respectivamente.

Para analizar la importancia de los componentes estimados, se parte de la varianza de la j-ésima componente (Z_j) que es λ_j.Si definimos la sumatoria de las λ_j, ello es equivalente a la Traza de la matriz de correlación [Tr(P)], en el sentido que mide la variación total explicada por todas las variables componentes principales[18] (Johnson, 2000:98). Así, la importancia de la j-ésima componente principal está definida por $P_j = \frac{\lambda_j}{Tr(P)}$ para $j = 1,2,...., p$.

De la formulación presentada hasta aquí se definen tres condiciones básicas:

1. Los vectores son normalizados de tal forma que: $|a_1| = |a_2| = ... = |a_p| = 1$

   $\Leftrightarrow a_i^{'}a_i = 1$

   $\Leftrightarrow a_i^{'}a_j = 0$

2. Los componentes $Z_1, Z_2, .., Z_p$ no están correlacionados entre sí.

3. $Var(Z_1) \geq Var(Z_2) \geq ... \geq Var(Z_p)$, siendo $\sum_{i=1}^{p} Var(X_i) = \sum_{h=1}^{p} Var(Z_h)$

18. Esto es equivalente a la cantidad total de la variación medida por las variables originales

$Tr(\Sigma) = \sigma_{11} + \sigma_{22} + ... + \sigma_{pp}$

5.5. Calificación de componentes principales

Una vez estimados los componentes principales, su utilidad en el análisis estadístico está asociada a que su calificación pueda dar contenido a la variable latente que se intenta explicar. Esto nos lleva a preguntarnos ¿Cuál es el procedimiento para calificar cada una de las unidades de observación?

Siguiendo el análisis matricial resumido que se ha presentado en la sección 5.4, si X_r es el vector de variables observadas para la r-ésima unidad de observación, entonces, la calificación de la j-ésima unidad de observación viene dada por: $Z_{rj} = a'_j(X_r - \mu)$, para $j = 1,2,....,p$ Y $r = 1,2,....,n$

Esta calificación proporciona un valor ordinal de las unidades de observación, en la medida en que indican su ubicación relativa en un conjunto de datos o variables con respecto a los ejes componentes principales que se han definido.

5.6. Determinación del número de componentes principales

La mayoría de los manuales consultados argumentan que son tres los criterios que emplean los investigadores para determinar la verdadera dimensionalidad de los datos a través de los componentes principales (Jolliffe, 1986; Johnson, 2000).

- El porcentaje total de varianzia explicada por los componentes principales. Para el caso de estudios que toman muestras de laboratorio, se espera reunir del 90 al 95% de la variabilidad total con pocos componentes. Sin embargo para datos de investigación social, es posible que se requieran más de tres componentes principales para explicar entre el 70 y 75% de la variación total.
- El criterio gráfico que parte de la estimación de una gráfica de sedimentación (Scree) que se define a partir de los autovalores estimados contra el reciproco. Lo que se representa son las combinaciones de $(1,\lambda_1),(2,\lambda_2),...,(p,\lambda_p)$. Se supone que cuando la gráfica tiende a nivelarse, sus autovalores tienden a cero y su explicación como un componente adicional probablemente pueda estar midiendo ruido aleatorio. Ello demanda el criterio del especialista.
- Cuando los componentes principales son estimados a través de la matriz de correlación, el criterio a elegir, además de los dos anotados antes, es el que la dimensión real de los datos se encuentra cuando los autovalores son mayor o igual a 1. La lógica de este criterio es que cuando se usan variables tipificadas para estimar los componentes principales, la varianza explicada de ellos no puede ser inferior a la que resulta de las variables tipificadas, que en este caso es igual a la unidad.

5.7. Estimación del índice derivado del ACP

Una vez definido los componentes principales, la pregunta obligada es: ¿cómo incorporarlos al análisis del indicador sintético?. Si sólo es seleccionado un componente principal, éste actuaría como un indicador sintético del conjunto de variables respuesta, pero cuando los componentes son más de uno, ello demanda un tratamiento para utilizarlos en el análisis de los resultados. Aquí se propone, siguiendo a Peters et al., (1970), que el índice sea calculado empleando un promedio de las puntuaciones de cada componente principal, ponderados por la raíz cuadrada de la varianza de cada componente. En este sentido, el índice para cada unidad de observación se debe calcular como:

$$Indice_{\ j} = \frac{\sum_{i=1}^{r} Z_{rj} \cdot \sqrt{\lambda_r}}{\sum_{i=1}^{r} \sqrt{\lambda_r}} \quad ; j = 1,2,\ldots,21$$ y i = 1, 2, 3...r componentes.

Siendo Z_{rj} la puntuación del componente r-ésimo para la unidad de observación j-ésima, y $\sqrt{\lambda_r}$ la raíz cuadrada del autovalor para dicho componente, garantizando así que los componentes con una mayor varianza explicada tenga una mayor ponderación en la calificación del índice.

5.8. Resumen del procedimiento a emplear para el ACP

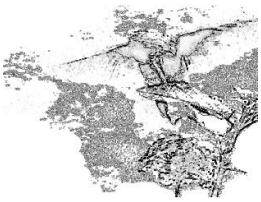
A continuación se resumen los pasos a seguir para la aplicación del análisis de componentes principales al estudio de indicadores sintéticos.

- Se parte de un conjunto de variables que conceptualmente responden a un modelo de explicación de la variable latente.
- Se analizan los datos para determinar si se emplea el análisis de varianzas-convarianzas o el de correlaciones. Ello depende del grado de homogeneidad de la escala en que se encuentran los datos de las variables respuesta.
- Se prueba la independencia de las variables respuestas. Es decir, se debe probar si estas variables son independientes o no correlacionadas. En cuyo caso el ACP no operaría[19].
- Determinar si existen datos ausentes, outliers, etc, y definir un procedimiento a emplear para su tratamiento en el conjunto del ACP.
- Modelar la base datos (puede ser en SPSS o cualquier otro paquete estadístico), obteniendo los componentes principales y el conjunto de estimaciones que permitirán probar la consistencia de los resultados. Estos paquetes estadísticos estiman automáticamente los valores de los componentes seleccionados y a juicio del investigador se deben definir los componentes relevantes para la construcción del índice. Un criterio útil es que se deben seleccionar los componentes que tengan autovalores mayor o igual a 1.
- Estimar el orden de las unidades de observación, de acuerdo al procedimiento de agregación de componentes descrito en la sección 5.7.
- Análisis espacial de los resultados, de tal forma que se puedan identificar patrones espaciales del índice por cada unidad de observación.
- Presentación de los resultados, mediante informe escrito, cuadros, gráficos o mapas según a quien se dirija la investigación. Aquí juega un papel importante la vinculación de las distintas bases de datos a un Sistema de Información Geográfica, que permita representar los resultados espacialmente.

6. ANÁLISIS DE LA DISTANCIA P_2 (DP_2)

La elección de indicadores sintéticos para valorar una variable latente como el desarrollo, el bienestar, la calidad ambiental, etc., puede hacerse desde criterios de reducción de la información de los indicadores simples (análisis factorial) o mediante los indicadores de distancia

19. Las variables respuestas son independientes (no correlacionadas) si matriz de correlaciones P = I (identidad) o, la matriz de varianzas-covarianzas es una matriz diagonal. El estadístico de prueba de Ho: P=I se compara con respecto los valores de $|R|$, para el caso de la matriz de correlación. Para valores grandes de N, se rechaza Ho si $-a\log|R| > \chi^2_{\alpha,p(p-1)/2}$, en donde $a = N - 1 - (2p+5)/6$. El valor de $|R|$ es el determinante de la matriz de correlación, que es estimado por el paquete estadístico SPSS 10.0.



como la DP_2. Dependiendo del objetivo que se persiga con el indicador sintético, es conveniente uno u otro método, sin embargo Pena Trapero (1977:77) indica que en algunas situaciones el análisis factorial y el análisis de distancias pueden ser complementarios porque "frente a la misma matriz de observación X, los indicadores de distancia pretenden comparar de alguna forma la posición relativa de las filas, mientras que en el análisis factorial, pretendería obtener los factores comunes, contenidos en las columnas". En este sentido, Zarzosa (1996:82) indica que esta complementariedad puede llevar a que se utilice el análisis factorial para reducir los indicadores simples de la matriz inicial de observaciones y luego utilizar indicadores de distancia para hacer comparaciones entre filas o unidades de observación.

El análisis de distancia que se realiza en estudios como los de Castro (2004) y Escobar (2004; 2006b), presentan, una metodología alternativa que permite comparar los resultados de la medición de índices con la resultante del ACP. La DP_2 es un indicador de distancia que presentamos en este artículo, como alternativa metodológica para estimar indicadores sintéticos[20].

A continuación se define la DP_2, las propiedades matemáticas que la hacen una buena aproximación para estimar un "buen" indicador sintético, y el procedimiento para abordar su aplicación empírica.

6.1. Condiciones básicas de una medida de distancia

Las medidas de distancia (cuadrática o euclidiana) satisfacen las condiciones exigidas en un espacio métrico:

1) No negatividad: La distancia es un número real único, no negativo, que vale cero únicamente cuando dos vectores, Xj y X*, son iguales.

2) Conmutatividad: La distancia entre Xj y X*, es igual a la distancia entre X* y Xj.

3) Desigualdad triangular: Dados tres vectores Xj, Xk y X*, definidos en el mismo espacio vectorial, se debe verificar que la suma de la distancia de dos de ellas al tercero ha de ser mayor o igual a la distancia existente entre ambos:

Otras condiciones adicionales que debe cumplir cualquier indicador de distancia, es la verificación de la propiedad de Exhaustividad (Propiedad VII) basados en:

- Aprovechar al máximo la información proporcionada por los indicadores simples.
- El indicador sintético debe solo tener en cuenta la información útil y eliminar la doble información o la que se encuentre repetida entre dos o más indicadores simples.

Esto plantea la necesidad de cumplir con cuatro condiciones adicionales que aseguran la no duplicidad de información. Para el caso de la distancia que interesa a este estudio (D1)[21], Siguiendo a Zarzosa (1996), éstas se pueden describir como: $d_i^{(1)} = \left| X_{ji} - X_{*j} \right|^2$ donde X_{ji} es el valor observado del indicador simple i en la unidad de observación j y es el valor de referencia del indicador simple *i*.[22]

De forma general, entonces: $D_1 = \sum_i d_i^{(1)}$ [23]

En el marco de esta medida de distancia, las condiciones para probar la propiedad de Exhaustividad son:

1) De independencia. Si todos los indicadores parciales son mutuamente independientes, el indicador sintético de distancia es la suma de todos ellos.

2) Dependencia funcional. Si la información que recogen uno o varios indicadores parciales está incluida en la que recoge otro, aquellos deben ser eliminados, al no aportar información adicional.

20. Pena Trapero (1977:65) y Zarzosa (1996:70) presentan un análisis en extenso de los métodos basados en el concepto de distancia para medir diferencias entre unidades de experimentales (países, regiones, etc.). Estos métodos son la Distancia CRL de Pearson, Distancia de Frechet, Distancia Generalizada de Mahalanobis, Distancia de Stone, Distancia - I de Ivanovic, Distancia - P1 y la que se trabaja en esta investigación, DP2, basada en la Distancia de Frechet.

21. Se le conoce como distancia euclidiana y la DP2 se basa en este tipo de distancia.

22. Este valor refleja el objetivo deseable, la meta de política o un indicador objetivo definido por la norma para cada indicador simple. En este sentido, el resultado numérico de un indicador sintético de distancia indica lo que falta para alcanzar el nivel de calidad mínimo objetivo o deseable mediante la política ambiental o social. También puede ser un indicador de un país, región, ciudad o comuna de referencia.

$$d_k^{(1)} = f(d_h^{(1)})$$

$$D_1 = \sum_{i \neq k} d_i^{(1)}$$

En el caso de que un solo indicador parcial recoja la información de los demás, el índice se debe expresar solo en función del mismo: $D_1 = d_i^{(1)}$.

3) De dependencia parcial. Si alguno de los indicadores simples contiene información de otros indicadores, el índice de distancia debería ser modificado para eliminar la duplicación de información.

4) De partición. Si los indicadores simples pueden partirse en dos o más grupos independientes por la información contenida, el índice de distancia igualmente se puede partir en los grupos que se deriven, siendo igual a la suma de ellos.

6.2. Definición de la DP2

La DP2 es un procedimiento matemático empleado para la estimación de indicadores sintéticos de distancia, que ha sido utilizado para medir variables latentes como el bienestar, desarrollo sostenible, etc., entre distintas regiones de un país o unidades de observación. Para el caso de España, son buenas las referencias de algunos estudios realizados por Pena Trapero (1977), Zarzosa (1996) y Castro (2004). En Colombia se pueden revisar los trabajos de Escobar y Bermúdez (2004) y Escobar (2004; 2006b).

Este indicador sintético se basa en el concepto de distancia elaborado por Pena Trapero a partir de una modificación de la distancia de Ivanovic (dI) (1974)[24].

$$d_{I(i,i^*)} = \sum_{i=1}^{p} \frac{\left| X_{ij} - X_{i^*j} \right|}{\sigma_i} \prod_{i=1}^{j-1} (1 - r_{ij.1,2,\ldots,i-1}), i < j$$

Donde $r_{ij.1,2,\ldots,i-1}$ es el coeficiente de correlación parcial entre el componente i-ésimo y el j-ésimo, dI es una medida que refleja el valor absoluto de la diferencia entre el conjunto de indicadores ideales (X_{i^*j}) con relación a un conjunto de p indicadores simples (X_{ij}), tipificados por la inversa de la desviación estándar del indicador simple y corrigiendo la información redundante mediante la inclusión del coeficiente de correlación parcial[25].

La DP_2 es desarrollada por Pena Trapero, basado en la distancia de Frechet y utiliza como factor de ponderaciones de las distancias estimadas, el coeficiente de determinación (R2) como se indica a continuación: $DP_2 = \sum_{i=1}^{p} \frac{\left| X_{ij} - X_{i^*j} \right|}{\sigma_i} \left(1 - R^2_{i.i-1,i-2,\ldots 1}\right)$

donde, $d_i = \left| X_{ji} - X_{*i} \right|$[26] para el caso donde se mide distancia del indicador simple de un país, región, ciudad, comuna, etc., con respecto a un parámetro o base de referencia del indicador simple X_* , σ_i es la desviación estándar de los valores que toma el indicador simple i-ésimo.

23. Esta medida de distancia aún no garantiza la propiedad de agregación porque no resuelve el problema de las escalas ni de las unidades de medida diferentes. Cuando se plantee la medida de distancia que utiliza la DP2, se volverá sobre este asunto.

24. Su cálculo de distancia está basado en la medida de distancia de Frechet $\sum_{i=1}^{p} \frac{\left| X_{ij} - X_{i^*j} \right|}{\sigma_i}$.

25. Pena Trapero (1977) indica que como el resultado del índice varía al cambiar el orden en que se introducen los indicadores simples, ello hace conveniente definir un orden para incorporar los indicadores simples al índice, dependiendo de la varianza explicada de cada uno de ellos. La propuesta de Ivanovic a este problema, consistió en describir un método iterativo de aproximación que concluyera en una solución convergente o estable, basado en el coeficiente de correlación entre el resultado del indicador sintético estimado y los valores de cada indicador simple.

26. Esto confirma que el indicador DP2 parte de una distancia euclidiana tipo D1.

El factor $\frac{d_i}{\sigma_i}$ soluciona el problema de la heterogeneidad de las unidades de medida de cada indicador simple, al tipificar la diferencia estimada por σ_j, garantizando la propiedad de aditividad de los componentes del índice. Además este divisor actúa como ponderador que da mayor importancia a las distancias con valores de mayor dispersión con respecto a la media.

$R^2_{i.i-1,i-2,\ldots 1}$, es el coeficiente de determinación en la regresión de X_i sobre $X_{i-1}.X_{i-2}\ldots.X_1$; Este coeficiente es un número abstracto, es decir que no importa la unidad de medida en la que se encuentren los indicadores simples. Además este factor permite eliminar la información ya contenida en los indicadores simples precedentes (Zarzosa, 1996:84).

$R^2_1 = 0$; Porque la primera componente aporta toda la información al no existir un componente previo. Por ello su ponderación es la unidad.

El orden de introducción de los componentes también hace variar el resultado final, lo que requiere el procedimiento iterativo de Ivanovic, tal como se describió antes en el pie de página 24. La diferencia básica es que en la DP_2 el procedimiento interactivo está asociado al coeficiente de determinación y no al de correlación. Por ello, para estimar la DP_2 es necesario realizar una jerarquización previa de los componentes, aproximándose mediante un método interactivo que parte de la solución inicial hasta encontrar la solución de convergencia[27].

El resultado generalizado de la ordenación de los indicadores simples en un índice mediante la DP_2 es:

$$DP_2 = \frac{d_1}{\sigma_1} + \frac{d_2}{\sigma_2}\left(1 - R^2_{2.1}\right) + \frac{d_3}{\sigma_3}\left(1 - R^2_{3.2,1}\right) + \ldots. + \frac{d_p}{\sigma_p}\left(1 - R^2_{p.p-1,p-2,\ldots 1}\right)$$

7. SÍNTESIS DEL ANÁLISIS DE LAS TÉCNICAS DE ACP Y DP_2

Como síntesis de los dos métodos que se utilizan para estimar indicadores sintéticos, es pertinente indicar de manera general cuales son las propiedades matemáticas que cumplen uno y otro método.

7.1. Cumplimiento de las propiedades matemáticas

Los indicadores sintéticos elaborados con la DP_2 cumplen con las VII condiciones presentadas en la sección 5.2 y dos propiedades adicionales: Aditividad[28] e Invarianza respecto a la base de referencia. Al respecto Zarzosa (1996:88-98) hace una demostración rigurosa del cumplimiento de estas propiedades, en la cual se puede profundizar sobre el tema.

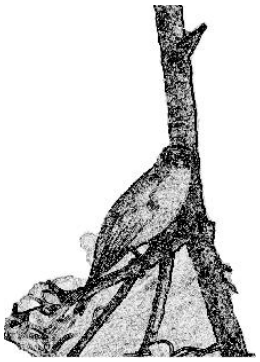
De otro lado, los indicadores basados en el ACP no cumplen satisfactoriamente las condiciones de:

- Invarianza. Sólo cumple con esta propiedad si los indicadores simples son tipificados. Es decir si se trabaja con la matriz de correlación. Esta propiedad no se cumple para "datos en bruto" (matriz de varianza-covarianza) sin resolver el problema de escala entre los indicadores simples.
- Homogeneidad. Las medidas derivadas del ACP son ordinales, por lo tanto las estimaciones que se realicen cumplen esta propiedad.

27. Este procedimiento es fundamental para garantizar la propiedad de unicidad del índice. El criterio utilizado por Pena Trapero consiste en ordenar los componentes por la cantidad de información que aporta al índice. Entraría el componente que tiene la mayor correlación con el indicador sintético y así sucesivamente con el segundo componente. Es conveniente aclarar que este procedimiento interactivo fue desarrollado en el programa FELIX, elaborado por Zarzoza y Zarzoza (1994).

28. De acuerdo a Zarzosa (1996), la DP2 no verifica estrictamente la propiedad de aditividad. Sin embargo ella demuestra como se cumple de manera restringida.

- Transitividad. El valor del índice puede variar si se incluye una unidad de observación adicional al conjunto de datos.



Lo anterior indica que con base en las propiedades matemáticas que deben cumplir los indicadores sintéticos, el DP_2 es más consistente que el ACP. Sin embargo debe recordar que antes, citando a Pena Trapero, se argumentaba la complementariedad de fases en los dos métodos, lo cual hace que el investigador vea la utilidad que reportan estos métodos para la simplificación de variables que explican un mismo fenómeno.

7.2. Algunos resultados empíricos de referencia

A continuación presentamos a manera de ejemplos, los resultados obtenidos por el autor en Escobar (2004; 2006a; 2006b), en el que se estimó el índice de calidad ambiental urbano, a nivel de comunas, empleando ACP y DP_2 para sintetizar un conjunto de variables o indicadores simples.

En este estudio Escobar, (2004; 2006a; 2006b) estimó el índice de calidad ambiental y presentó sus resultados en un SIG, para derivar de éste, un análisis espacial que le permitiera definir zonas ambientalmente homogéneas de la ciudad de Cali. Los resultados indican que la correlación de los dos índices es superior al 70% y que espacialmente la distribución del ICA en las distintas unidades de observación (21 comunas) es muy similar. Sin embargo las diferencias que se observan (ver Figuras 1 y 2), están asociadas a que en la aplicación del ICA mediante ACP, se emplearon todas las variables o indicadores simples que explicaban el ICA, y en el resultado mediante DP_2, el criterio de selección (correlación superior al 40% entre el conjunto de indicadores simples y los índices de componente) sólo dejó como relevantes siete indicadores simples, excluyendo la variable calidad del aire, siendo una de las más importantes en el ACP. Por ello los resultados del índice estimado por DP_2 en zonas como las comunas 5 y 6 no reflejan la contaminación atmosférica importada del parque industrial de Acopi-Yumbo, que si se registra con el índice construido con ACP.

Para un mayor detalle sobre la importancia y utilidad de los índices estimados, mediante técnicas de análisis multivariado, en la identificación y gestión de problemas ambientales urbanos, remitimos al lector a consultar en detalle el estudio de Escobar (2004; 2006a; 2006b) y Castro (2004), dado que éstos son aspectos que rebasan los objetivos que nos habíamos trazado en este artículo, como es el de presentar la relevancia de dos técnicas de análisis matemático (ACP y DP_2) para la estimar indicadores sintéticos.

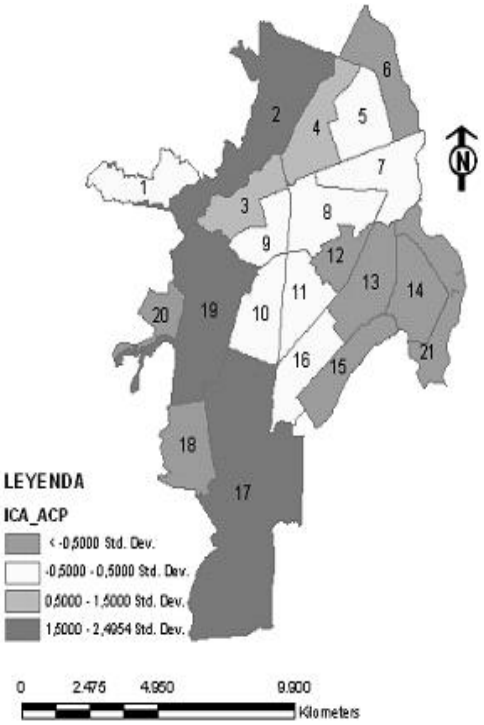
8. CONCLUSIONES

En este artículo se han presentado dos procedimientos matemáticos que pueden ser empleados para estimar indicadores sintéticos. En su exposición formal, fueron presentadas las propiedades matemáticas que idealmente deben cumplir los indicadores sintéticos y, de manera general, se definió que de las dos técnicas de análisis multivariante, la DP_2 es más consistente que el ACP para la estimación de un índice o indicador sintético objetivo. Sin embargo, como lo anota Pena Trapero (1977), la complementariedad en sus fases de desarrollo hace que estos dos métodos puedan ser usados de manera conjunta.

Una observación final desde el punto de vista operativo lleva a recomendar que la bondad del uso de uno u otro método está, en últimas, altamente relacionada con una adecuada selección de los indicadores simples, siempre entendiendo que es una aproximación a la medición de una variable latente objetivo. Sin embargo cualquier técnica de análisis multivariado, ya sea ACP o DP_2, basada en información dudosa debe conllevar a igual resultado del indicador sintético objetivo.

En este artículo, se han presentado, a manera de ejemplo, los resultados de Escobar (2004; 2006a; 2006b) en el que se aplican las dos técnicas descritas, haciendo una síntesis y definición de los principales indicadores ambientales simples, consecuente con el nivel de

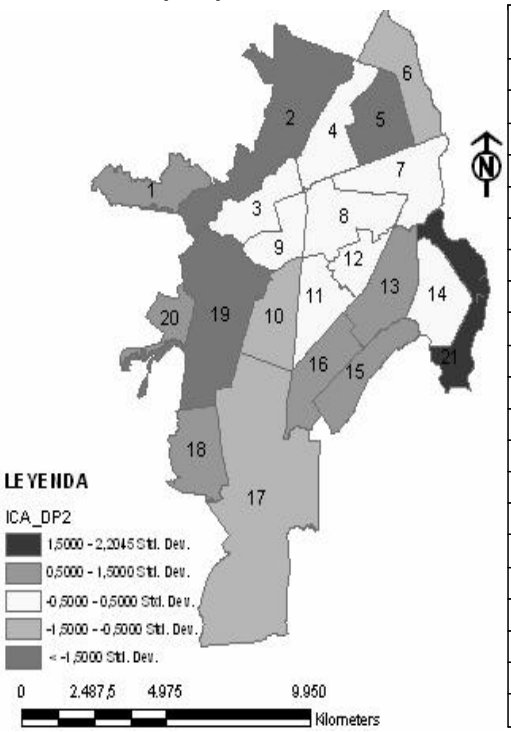
agregación adoptado (comunas), para derivar el Índice de Calidad Ambiental de la ciudad de Cali, y su análisis espacial mediante su representación cartográfica.



COMUNA	ICA
17	1,0535
2	0,7469
19	0,7186
3	0,3830
4	0,2686
5	0,0906
8	0,0999
9	0,0660
10	0,0555
11	-0,0210
7	-0,1155
16	-0,1740
1	-0,1945
18	-0,2214
21	-0,2562
15	-0,2895
12	-0,3156
6	-0,3376
20	-0,4718
13	-0,4781
14	-0,6076

Fuente: Escobar (2006a)

Figura 1.
Índice de Calidad Ambiental ACP



Comuna	ICA_DP2
5	3,1011
2	3,4661
19	3,7651
17	5,3363
10	5,9255
6	6,0123
8	6,4931
4	7,1656
11	7,3953
7	7,5588
12	7,6265
3	7,9050
9	8,2081
14	8,4766
16	9,2709
15	9,2950
18	9,6103
13	9,9103
1	10,2196
20	11,3077
21	13,2111

Fuente: Escobar (2006a)

Figura 2.
Índice de Calidad Ambiental DP_2

AGRADECIMIENTOS

El autor agradece la colaboración recibida por Don Diego Azqueta Ayarzun como director de su trabajo de tesis doctoral y por facilitar mi estancia en el Grupo de Economía Ambiental de la Universidad de Alcalá (España). En este periodo se desarrollaron muchas de las ideas expuestas en este artículo. También agradece la colaboración de Don Juan Marcos Castro Bonaño y Doña Pilar Zarzoza por facilitar la información y el software para correr los modelos que se han presentado en este estudio.

BIBLIOGRAFÍA

Anderson, T.W., 1984. An introduction to multivariate statistical analysis. Wiley and Sons Ltd. New York.

Castro B., J.M., 2004. Indicadores de desarrollo sostenible urbano. Una aplicación para Andalucía. Institututo de Estadística de Andalucía. Consejería de Economía y Hacienda.

Dixon, J. y Segnestamet, L., 2002. Environmental Indicators An Overview of Selected Initiatives at the World Bank. World Bank. Environment Department.

Ebert U. y Welsch, H., 2004. Meaningful environmental indices: a social choice approach. Journal Environmental Economics and Management. 47, pp. 270-283.

Escobar, L. A., 2004. Construcción de Índices de Calidad Ambiental Urbana: Un Modelo general y aplicación para Cali-Colombia. Trabajo de investigación tutelado para obtener el título de Diploma de Estudios Avanzados. Universidad de Alcalá. Comunidad de Madrid-España.

Escobar L. A., 2006a. Indicadores sintéticos de calidad ambiental: un modelo general para grandes zonas urbanas. Revista EURE. Revista Latinoamericana de Estudios Urbanos Regionales. Vol. XXXII. No. 96. Agosto. Santiago de Chile. Versión electrónica: www.scielo.cl/eure.htm. pp.73-98.

Escobar, L. A., 2006b. Valoración económica de la calidad ambiental desde una perspectiva geográfica. Tesis doctoral. Universidad de Alcalá, Diciembre. Comunidad de Madrid-España.

Escobar L. y Bermúdez T., 2004. Evaluación de la calidad ambiental por localidades en Bogotá: Una aproximación a la construcción de índices de calidad ambiental. Revista Gestión y Ambiente. Volumen 7- No. 2.

Hair, J.F., Anderson R., Tatham R. y Black, W.C, 1999. Análisis multivariante. 5ta edición. Prentice Hall Iberia. Madrid.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, pp. 417-520.

Jolliffe, I.T., 1986. Principal Components Analysis. Springer-Verlag. New York.

Johnson, D.E., 2000. Métodos multivariados aplicados al análisis de datos. México. Internacional Thomson editores. Edición No. 1. 566 P.

Ospina J. y Lema, A., 2005. Tratamiento estadístico para indicadores cuantitativos de impactos, aplicados a líneas de transmisión eléctrica. Revista Facultad Nacional Agricultura. Medellín.Vol.58, No.2946 2, pp.2945-2961.

Milon, J.W y Shogren, J. F., 1995. Economics, Ecology and the Art of integration. Tomado de Integreting Economic and Ecological Indicators: Practical methods for environmental policy analysis. Edited by Milon and Shogren. Library of Congress. US.

Ministerio de Medio Ambiente, MMA, 1996. Indicadores ambientales. Una propuesta para España. Dirección General de Calidad y Evaluación Ambiental. Ministerio de Medio Ambiente. Madrid.

Ministerio de Medio Ambiente, 2000. Sistema español de indicadores ambientales: Área de medio urbano. Centro de Publicaciones. Ministerio de Medio Ambiente. Madrid.

Pardo, A . y Ruiz, M. A., 2002. Guía para el análisis de datos. McGraw-Hill Interamericana de España.

Pena Trapero, J. B., 1977. Problemas de la medición del bienestar y conceptos afines. Una aplicación al caso español. INE. Madrid.

Peters, W.S. y Butler, J.Q., 1970. The construction of Regional Economic Indicators by principal components. Annals of Regional Science, IV, pp. 1-14.

Polanco C., 2006. Indicadores ambientales y modelos internacionales para toma de decisiones. Revista Gestión y Ambiente. Volumen 9, No. 2.

Rao, C.R., 1965. The use and interpretation of principal components analysis in applied research. Sankhya (A), 26, pp. 329-358.

Segnestam, L., 2002a. Indicators of Environment and Sustainable Development: Theories and Practical Experience. Environmental Eonomics Series. Paper no. 89. The World Bank Environment Department.

Ministerio de Medio Ambiente, 2002b. Indicators of environment and sustainable development. Stockholm Environment Institute. Policy & Institutions. www.Sei.se/policy.html.

Visauta, B. y Martori, J. C., 2003. Análisis estadístico con SPSS para Windows. Vol. II. Estadistica Multivariante. McGraw Hill.

Zarzosa, F. y Zarzosa, P., 1994. Programa de cálculo del Indicador Sintético de Distancia DP2 para medir el bienestar social. Número de Registro 655. Registro provincial de la propiedad intelectual de Valladolid.

Zarzosa, P., 1996. Aproximación a la medición del Bienestar Social. Universidad de Valladolid. Valladolid.