GOBIERNO DEL
ESTADO DE MÉXICO

GOBIERNO QUE TRABAJA Y LOGRA
enGRANDE

IGECEM

TALLER DE ANÁLISIS MULTIVARIABLE

TÉCNICA DE COMPONENTES PRINCIPALES

SEPTIEMBRE DE 2012

INSTITUTO DE INFORMACIÓN E INVESTIGACIÓN GEOGRÁFICA,
ESTADÍSTICA Y CATASTRAL DEL ESTADO DE MÉXICO

ANÁLISIS MULTIVARIABLE

El análisis multivariable es el conjunto de técnicas estadísticas que de forma simultánea **miden, explican y predicen todas las relaciones existentes entre los elementos que conforman una tabla de datos**, proporcionando un resultado que debe ser interpretado minuciosamente por el analista.

En esencia, la mayoría de las técnicas de análisis multivariable tienen como fundamento la búsqueda de una **combinación óptima** de las variables implicadas en el análisis; más en concreto, dicha combinación **debe representar las interrelaciones** que existen entre esas variables **o explicar el comportamiento** de alguna otra variable –ya sea con fines **predictivos o clasificatorios** – de la mejor manera posible.

Tal combinación representa el modelo ajustado, es decir, la mezcla de variables que se adapta a los hechos observados.

GOBIERNO QUE TRABAJA Y LOGRA
enGRANDE

TÉCNICAS DE ANÁLISIS MULTIVARIABLE

Técnicas de análisis de la dependencia

Técnica	Variable dependiente	Variables independientes
Análisis de la varianza y la covarianza	Métrica	No métricas
Análisis discriminante	No métrica	Métricas
Regresión lineal múltiple	Métrica	Métricas
Regresión lineal múltiple con variables ficticias	Métrica	No métricas
Modelos de elección discreta	No métrica	Métricas
Modelos de elección discreta con variables ficticias	No métrica	No métricas
Análisis conjunto	Métrica o no métrica	No métricas
Segmentación jerárquica	No métrica o métrica	No métricas
Análisis de ecuaciones estructurales	Métrica	Métricas o no métricas
Análisis con clases latentes	No métrica latente	No métricas observables

Técnicas de análisis de la interdependencia

Técnica	Variables	Forma grupos de:
Análsis factorial y por componentes principales	Métricas	Variables
Análisis de correspondencia	No métricas	Categorías de variables
Análsis de conglomerados	Métricas y no métricas	Objetos
Escalamiento multidimensional	Métricas y no métricas	Objetos
Análisis con clases latentes	No métricas	Objetos y categorías de variables

SUPUESTOS SUBYACENTES EN LOS MÉTODOS MULTIVARIABLES

Normalidad

La hipótesis de partida que debe cumplir cualquier análisis multivariable es la normalidad de todas y cada una de las variables que formen parte del estudio, si este supuesto no se cumple, el resto de tests estadísticos no serán válidos, puesto que se requiere de la normalidad para el uso de los estadísticos t y F.

Linealidad

Conviene examinar si las relaciones entre las variables que intervienen en el estudio son lineales. La linealidad indica que el modelo predice los valores de las variables dependientes siempre que se produzca una modificación en las variables independientes.

Homocedasticidad

La homocedasticidad es el último supuesto que deben cumplir los datos antes de iniciar su tratamiento multidimensional. Concretamente se verifica esta hipótesis cuando la varianza de los errores es constante. Es decir, la variación de la variable dependiente que se intenta explicar a través de las variables independientes no se concentra en un pequeño grupo de valores.

ANÁLISIS FACTORIAL Y POR COMPONENTES PRINCIPALES

Tanto el análisis factorial como el análisis de componentes principales son técnicas de análisis de la interdependencia o covariación presentada por un cierto número de variables métricas, susceptible de ser sintetizada en un conjunto de factores comunes que subyacen tras ella.

El número de variables que se extraen es inferior al número de variables analizadas, sin embargo, dichos factores serán suficientes para resumir la mayor parte de la información contenida en las variables originales. Como consecuencia, los factores podrán ser utilizados en sustitución de éstas, lo que explica que a menudo se haga referencia a estos análisis como técnicas de reducción de datos.

ANÁLISIS FACTORIAL Y POR COMPONENTES PRINCIPALES

La principal diferencia entre el análisis factorial común y el análisis por componentes principales radica en el método de extracción de factores. Los factores extraídos mediante análisis factorial captan la variabilidad común o compartida por todas las variables; mientras que la variabilidad específica, propia de cada variable y sin relación con las demás se recogerá en factores únicos o específicos.

Sin embargo, en el análisis por componentes principales se busca explicar la mayor proporción posible de variabilidad total –común y específica– con el menor número posible de factores o componentes.

En este sentido se le puede considerar un caso particular del análisis factorial en el que los factores comunes explican el cien por ciento de la varianza total y, por consiguiente no existen factores específicos.

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

El análisis de componentes principales es un método de geométrico de carácter descriptivo.

El objetivo es descubrir la estructura subyacente en un conjunto de n individuos estudiados bajo una serie de p variables cuantitativas.

Imaginemos que tenemos n individuos medidos bajo una sola variable, es fácil describir a estos individuos representándolos en una recta. Si fuesen dos variables se representan en un plano, incluso si fuesen tres el recurso es representarlos en una gráfica de nube de puntos . ¿Pero qué ocurre si el número de variables es igual o superior a cuatro?

El análisis de componentes principales es un método que permite transformar un conjunto de variables originales en otro conjunto de variables llamado conjunto de componentes principales. Estas componentes principales son combinación lineal de las variables originales y se caracterizan por estar incorrelacinadas entre sí.

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Si se toman las variables originales y se calcula su matriz de correlaciones se puede observar que, habitualmente, existe un alto grado de correlación entre algunas de las variables. Esto nos lleva a pensar que, quizá podríamos trabajar sobre un conjunto de variables incorrelacionadas de menor dimensión.

Así, si las variables originales están muy correlacionadas entre sí, es esperable que su información se pueda expresar a través de unas pocas componentes principales.

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Ejemplo:

Para la construcción del índice de Desarrollo Social y Humano, se emplea el siguiente modelo:

IDSH= a1(IND1)+a2(IND2)+….+a9(IND9)
Donde:

Porcentaje de población mayor de 15 años analfabeta	IND1
Porcentaje de población de 3 a 14 años que no asiste a la escuela	IND2
Porcentaje de población de 15 años o más sin primaria completa	IND3
Promedio de hijos nacidos vivos	IND4
Porcentaje de la población sin derechohabiencia a servicios de salud	IND5
Porcentaje de ocupantes en viviendas particulares habitadas con piso de tierra	IND6
Porcentaje de ocupantes en viviendas particulares habitadas que no disponen de drenaje no sanitario	IND7
Porcentaje de ocupantes en viviendas particulares habitadas que no disponen de agua entubada	IND8
Porcentaje de población ocupada con ingresos de hasta dos salarios mínimos	IND9
Índice de desarrollo social y humano	IDSH

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

MATRIZ DE CORRELACIONES

	IND1	IND2	IND3	IND4	IND5	IND6	IND7	IND8	IND9
IND1	1.000	.622	.960	.954	-.568	.580	.856	.674	.790
IND2	.622	1.000	.682	.575	-.141	.562	.527	.321	.482
IND3	.960	.682	1.000	.958	-.546	.628	.853	.603	.833
IND4	.954	.575	.958	1.000	-.547	.582	.825	.632	.838
IND5	-.568	-.141	-.546	-.547	1.000	-.182	-.601	-.376	-.447
IND6	.580	.562	.628	.582	-.182	1.000	.424	.301	.654
IND7	.856	.527	.853	.825	-.601	.424	1.000	.587	.715
IND8	.674	.321	.603	.632	-.376	.301	.587	1.000	.448
IND9	.790	.482	.833	.838	-.447	.654	.715	.448	1.000

a. Determinante = 1.968E-005

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Partimos de que se pretende obtener un nuevo conjunto de variables que son combinación lineal de las variables originales.
La varianza será una medida de información que contiene cada variable:
Con estas hipótesis el problema de componentes principales puede ser descrito de la siguiente manera:

Toda combinación lineal c de las variables originales puede expresarse de la siguiente manera.

$$c = Yv$$

Donde ***v*** es el vector que nos permite obtener la combinación lineal y ***Y*** es el vector de variables originales.

La primera componente principal es la combinación lineal de variables originales de varianza máxima.

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Por lo tanto, buscamos **v1** de norma tal que la varianza de la primera componente principal **c1** sea máxima.

Como la varianza de las componentes principales **c1** se escribe como:

$$S^2\mathrm{c} = \mathrm{v}^{\mathrm{t}} * \mathrm{V}_{\mathrm{y}} * \mathrm{v}$$

Se puede resumir el planteamiento del problema de la siguiente manera:

$$Max\ (\mathrm{v}^{\mathrm{t}} * \mathrm{Vy} * \mathrm{v})$$

Sujeto a $\mathrm{v}^{\mathrm{t}}\mathrm{v}=1$

Para lo cual tenemos el siguiente lagrangiano:

$$L = v^t V_y v - \lambda(v^t v - 1)$$

Ahora derivados con respecto a v e igualamos a cero:

$$\frac{\partial L}{\partial v} = 2V_y v - 2\lambda v = 0$$

$$Vyv = \lambda v$$

$$(Vy - \lambda I)v = 0$$

GOBIERNO QUE TRABAJA Y LOGRA enGRANDE

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

De esto se concluye que **v** es el vector propio de la matriz de varianzas –covarianzas de los datos originales.

Como se ha puesto la condición de que la varianza sea máxima, se escoge el vector propio con mayor valor propio asociado.

El resto de los componentes principales se obtienen de forma sucesiva de manera que **no** estén correlacionados con la componente anterior y que la varianza sea máxima.

Los vectores propios de la matriz **Vy** asociados a los valores propios escritos en forma decreciente son por lo tanto los vectores buscados (los factores principales). Estos vectores nos permiten calcular las componentes principales a través de la expresión **c=Yv.** La varianza de cada componente principal viene dada por los valores propios.

La interpretación de los componentes principales se hace a través del estudio de las correlaciones entre las componentes principales y las variables originales. Aquellas variables originales cuya correlación con una componente principal dada esté muy próxima a 1 en valor absoluto, serán las que más contribuyan a la explicación de dicha componente principal.

EMPLEO DE LA TÉCNICA DE COMPONENTES PRINCIPALES PARA MEDIR EL AVANCE DE LOS OBJETIVOS DEL MILENIO

En el año 2000, un grupo de 189 países, entre ellos México, suscribieron la Declaración del Milenio. En ella se plantearon los Objetivos del Desarrollo del Milenio (ODM), cuya fecha límite de consecución es el año 2015, que sintetizan la aspiración de un mundo mejor para todos y representan además el mayor consenso activo de la humanidad para lograr el desarrollo humano sustentable, entendido como la ampliación permanente de las capacidades y las libertades de cada persona para poder alcanzar una vida digna, sin comprometer el patrimonio de las generaciones futuras.
Los objetivos del milenio son los siguientes:

Objetivo 1. Erradicar la pobreza extrema y el hambre
Objetivo 2. Lograr la enseñanza primaria universal
Objetivo 3. Promover la igualdad de género y el empoderamiento de la mujer
Objetivo 4. Reducir la mortalidad de los niños menores de 5 años
Objetivo 5. Mejorar la salud materna
Objetivo 6. Combatir el VIH/SIDA, el paludismo y otras enfermedades
Objetivo 7. Garantizar la sostenibilidad del medio ambiente
Objetivo 8. Fomentar una alianza mundial para el desarrollo

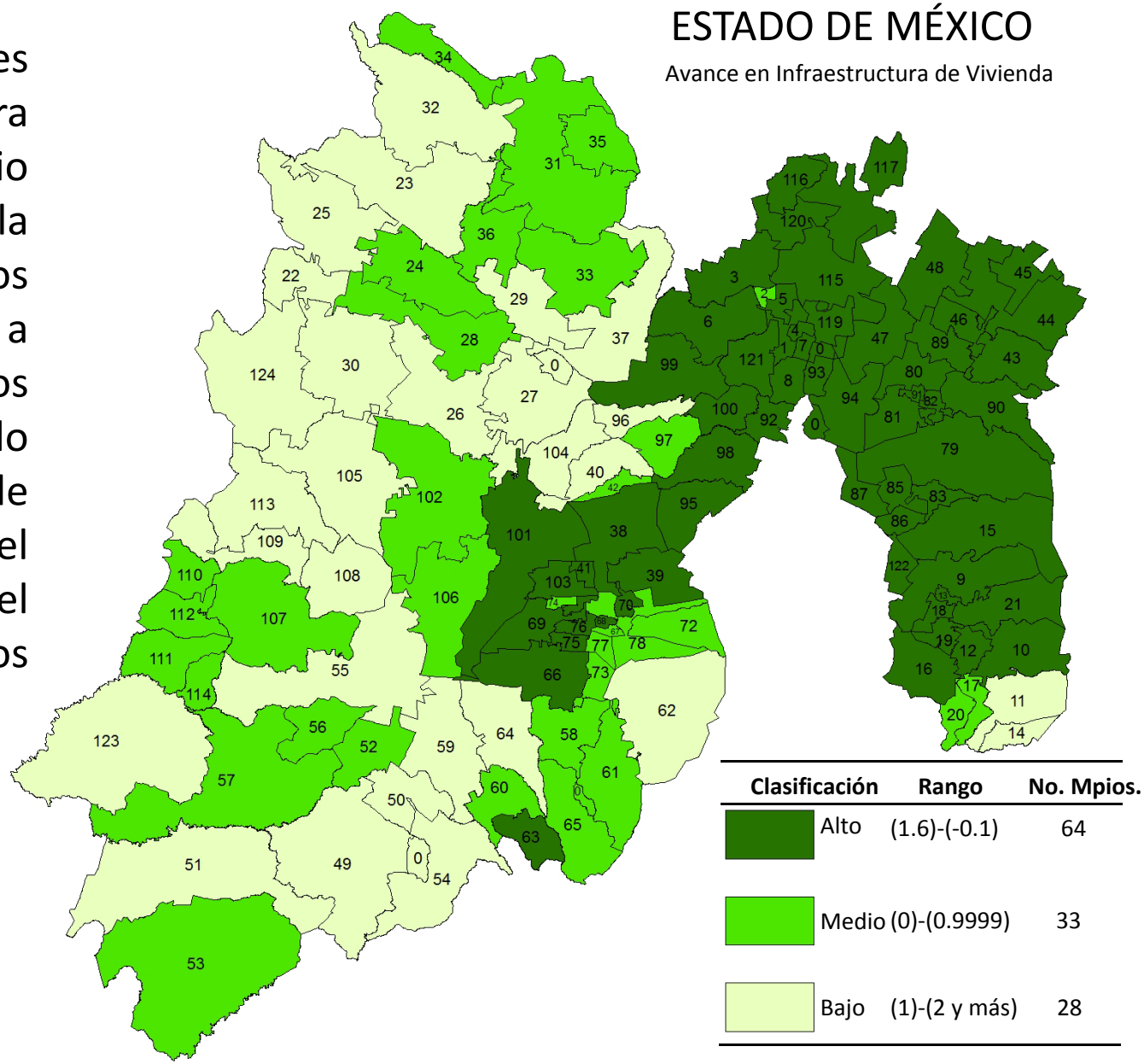
PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Debido a que los indicadores determinados por la SEDESOL pueden agruparse de acuerdo al cumplimiento del objetivo del milenio que pretenden medir, se aplicará la técnica de componentes principales, que es una técnica de reducción de la dimensionalidad. Su meta es explicar la mayor parte de la variabilidad total de un conjunto de variables cuantitativas con el menor número de componentes o factores comunes posibles.

Con este método es posible construir para cada objetivo del milenio un índice único que sea la combinación lineal de los indicadores asociados a ese objetivo, y que nos permita observar el grado de avance de cada uno de los 125 municipios del Estado de México en el cumplimiento de los objetivos del milenio.

PLANTEAMIENTO DEL PROBLEMA DEL ANÁLISIS DE COMPONENTES PRINCIPALES

Con este método es posible construir para cada objetivo del milenio un índice único que sea la combinación lineal de los indicadores asociados a ese objetivo, y que nos permita observar el grado de avance de cada uno de los 125 municipios del Estado de México en el cumplimiento de los objetivos del milenio.



GOBIERNO QUE TRABAJA Y LOGRA enGRANDE

Bibliografía recomendada:

Lévy Mangin Jean Pierre. “Análisis Multivariable para las Ciencias Sociales” Ed. Pearson Prentice Hall.

Pérez López César. Técnicas de Análisis Multivariable de Datos”. Ed. Pearson Prentice Hall.

Kleiman Ariel “ Matrices, aplicaciones matemáticas en economía y administración”. Ed Noriega Limusa.