II Congreso de la Asociación Latinoamericana de Población
Guadalajara, México, 3 – 5 de Septiembre de 2006

La demografía latinoamericana del siglo XXI
Desafíos, oportunidades y prioridades

Estrategia metodológica para la reducción de datos aplicada en demografía.

Gerardo Correa
Instituto Nacional de Estadística. Venezuela
correaje@yahoo.com

Mesa 10. Métodos, Técnicas y Datos en la Investigación Demográfica en América Latina y el Caribe

ESTRATEGIA METODOLÓGICA PARA LA REDUCCIÓN DE DATOS APLICADA EN DEMOGRAFÍA

Autor: Gerardo Correa
correaje@yahoo.com
Instituto Nacional de Estadística. Venezuela

RESUMEN

En este trabajo se presenta una estrategia metodológica que permite reducir, en una importante proporción, el número de filas de una matriz de datos cualitativos, sobre la cual se requiera aplicar técnicas de análisis multivarinte de datos, tales como el Análisis de Correspondencias Múltiples o el Análisis de Cluster.
La estrategia consiste, en derivar a partir de la matriz original de datos, una matriz de prototipos, conformada con todas las combinaciones posibles de las modalidades correspondientes a las variables seleccionadas.
Se presentan dos aplicaciones de la estrategia, sobre la base de datos censales del Censo de 2001 realizado en Venezuela. La primera consiste en construir una tipología de viviendas para Venezuela, utilizando las técnicas de Análisis de Correspondencias Múltiples y Análisis de Clasificación, mientras que la segunda consiste analizar la Fecundidad Alta en Venezuela, desde una perspectiva multivarante.

1. INTRODUCCIÓN

En el último decenio se ha popularizado el término de Minería de Datos o Data Mining, en referencia al proceso a través del cual se extrae conocimientos de grandes volúmenes de datos.

Uno de los aspectos que debe enfrentar la minería de datos para lograr su objetivo principal, es reducir los grandes volúmenes de datos que se generan debido a las transacciones que ocurren en las grandes corporaciones.

No obstante, el problema de la reducción de datos es común en todas aquellas disciplinas que requieran extraer información de un conjunto de datos.

Desde hace bastante tiempo, la Demografía se ha ocupado de extraer conocimiento de bases de datos que en ocasiones pueden contener millones de registros, tal como es el caso de las bases de datos correspondientes a los registros vitales o a las operaciones censales.

En este sentido, la aplicación de técnicas de reducción de datos, pueden enriquecer de manera sustantiva el trabajo demográfico, porque permitiría aplicar herramientas estadísticas que resultan muy costosas de aplicar sobre grandes bases de datos, entre las cuales podemos destacar el Análisis de Correspondencias Múltiple.

El presente trabajo es parte de los resultados de un Trabajo Especial de Grado (Correa, 2005), en el cual se propone una estrategia metodológica de reducción de datos, para aplicar Análisis de Correspondencias Múltiple y Análisis de Conglomerados sobre la base de datos censales correspondiente al XIII Censo General de Población y Vivienda realizado en Venezuela en octubre de 2001, con el objeto de construir una tipología de viviendas. Así mismo, también se presenta una aplicación de la estrategia señalada anteriormente para caracterizar a los hogares en los cuales existan mujeres con fecundidad alta **(**Rodríguez, 2003).

El trabajo está organizado en cuatro secciones: en la segunda se presentan los aspectos metodológicos correspondientes a la estrategia propuesta, en la tercera se presentan dos aplicaciones de la misma, y en la cuarta las conclusiones respectivas.

2. METODOLOGÍA

Existen diversas estrategias para reducir el volumen de datos requeridos para una determinada investigación. Algunas de estas estrategias pueden ser tan sencillas, como por ejemplo, transformar datos cuantitativos en intervalos de clases para ser presentados en una tabla de frecuencias para datos agrupados, o tan complicadas como la utilización de una red neuronal del tipo backpropagation para comprimir información proveniente de señales eléctricas (Hilera, J.; Martinez V.; 2000). Los siguientes son algunos de los tipos de estrategias de reducción de datos (Han J., Lamber M. 2000), más utilizadas:

- **Reducción de la dimensionalidad o "features extraction"** Consiste en extraer algunas características o variables del conjunto de datos, conservando la mayor cantidad posible de información. Alguna de las técnicas más utilizadas son las siguientes: Análisis de Componentes Principales, Métodos Heurísticos, Compresión de Datos, algunos paradigmas de Redes Neuronales
- **Discretización de valores:** Consiste en convertir datos de naturaleza continua en valores discretos.
- **Reducción de la numerosidad o selección de instancias:** Consiste en reducir la cantidad de registros de la base de datos. Puede realizarse a través de métodos parametritos (estableciendo un modelo para la data) o por métodos no parametritos. (muestreo de elementos, construcción de prototipos a través del análisis de agrupamiento)

La estrategia que se propone en este trabajo se orienta a la reducción de la numerosidad de los datos, sobre los cuales se requiera aplicar técnicas de análisis multivariante de datos, tales como análisis de correspondencias múltiple o análisis de agrupamiento. La misma es aplicable en datos de tipo cualitativo.

Consiste en construir una matiz de prototipos, que denotaremos con la letra P, la cual contiene en sus filas (m filas) los prototipos p_{ij} (i=1,2,...m),(j=1,2...J); J representa el número de modalidades derivados de las distintas combinaciones de las modalidades que conforman las variables categóricas de la matriz de datos Z.

Se puede demostrar que aplicar Análisis de Correspondencias Múltiples sobre la matriz P es equivalente que aplicarlo sobre la matriz de datos original Z.

En efecto, un elemento cualquiera b^{qq}_{ij} de la tabla de Burt correspondiente a la matriz Z, se puede obtener a través de la siguiente expresión:

$$b^{qq'}_{jk} = \sum_{i}^{n} z^{q}_{ij} z^{q'}_{ik}$$

Como además para cada prototipo p_i en el que las modalidades z^{q}_{j} $z^{q'}_{k,}$ de las variables q y q' respectivamente sean distintas de cero, se repiten w_i veces, podemos expresar la fórmula anterior de la siguiente manera:

$$b^{qq'}_{jk} = \sum_{i}^{m} w_i p^{q}_{j} p^{q'}_{k}$$

Por otra parte la tabla de Burt de la matriz P ponderada por las frecuencias de los prototipos tiene la siguiente forma:

$$B = (W^{1/2}P)' (W^{1/2}P)$$

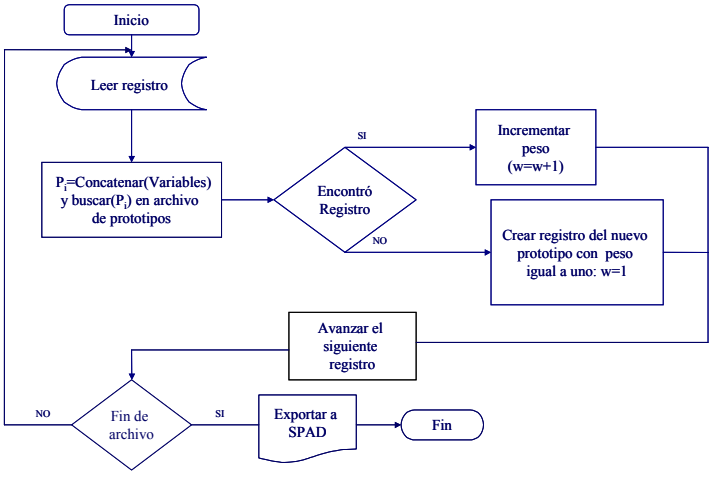
Y un elemento b^{qq}_{ij} cualquiera corresponde a :

$$b^{qq'}_{jk} = \sum_{i}^{m} w_i p^{q}_{j} p^{q'}_{k}$$

Con lo que se demuestra que la tabla de Burt de la matriz Z es la misma que la de la matriz P y por lo tanto el ACM aplicado sobre Z es equivalente al aplicado sobre la matriz P.

La estrategia propuesta tiene la ventaja, en comparación con otras alternativas, pues resulta fácil de implementar, ya que solamente es necesario construir la matriz de prototipos, como se muestra en el Grafico 1, y aplicar sobre esta matriz el análisis requerido

Grafico1: Diagrama de flujo para generar la matriz de prototipos



3. RESULTADOS

Los resultados arrojados muestran una importante reducción en el volumen de los datos. Para el primer ejercicio, que consistió en la construcción de una tipología de viviendas, se logró reducir de 5.174.937 registros de viviendas a 357.140. En el Cuadro 1 se muestran las variables incluidas (16 en total) y el Gráfico 2 el primer plano factorial.

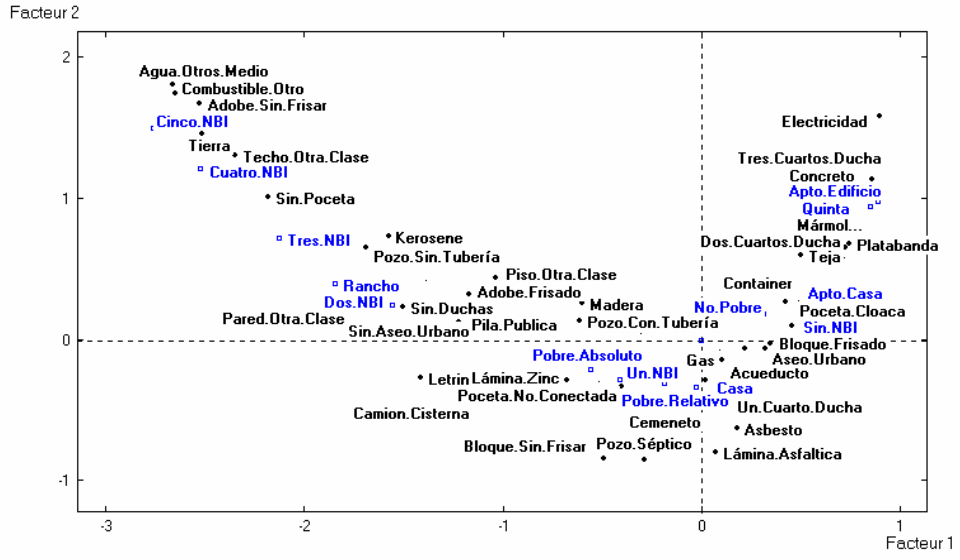
El segundo ejercicio fue caracterizar los hogares con fecundidad alta. La fecundidad alta se calcula de la siguiente manera: a) mujeres entre 15 y 19 años con un hijo o más; b) mujeres entre 20 y 24 años con dos o más hijos; c) mujeres entre 25 y 29 años con tres o más hijos; c) mujeres entre 30 y 34 años con cuatro o más hijos; d) mujeres entre 40 y 49 años con cinco o más hijos.

La reducción fue aún mayor: de 3.147.441 registros de hogares, a 46.762, lo cual obedece a que en este caso el número de variables incluidas fue casi la mitad que en el primer caso (9 variables en total), y por tanto fueron menos las combinaciones resultantes. En el Cuadro 3 se presentan las coordenadas de las variables incluidas y en el Grafico 3 el primer plano factorial.

Cuadro 1: Variables incluidas en el estudio

Dimensión materialidad	Dimensión servicios y confort	Dimensión socioeconómica
Tipo de vivienda. Material en paredes exteriores. Material en el techo. Material en el piso.	Eliminación de excretas. Abastecimiento de agua. Eliminación de la basura Acceso a servicio eléctrico. Acceso al servicio telefónico Número de duchas. Número de cuartos para dormir de la vivienda. Ubicación de la cocina.	Tenencia de la vivienda Niveles de ingreso Condición de ocupación

Grafico 2: Primer plano factorial para la tipología de viviendas



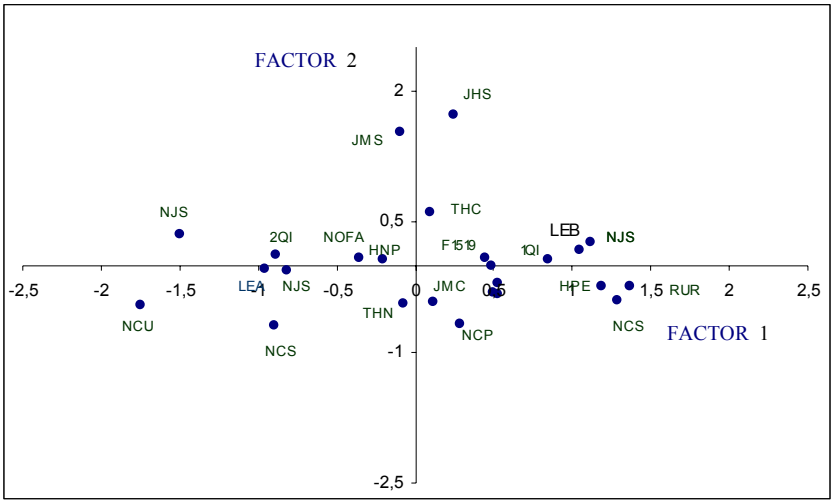
Cuadro 2:
Principales características de los grupos obtenidos a través del análisis de cluster automático

Modalidad	% respecto al total de la modalidad	% respecto al total de la clase	Modalidad	% respecto al total de la modalidad	% respecto al total de la clase
Clase 1 / 10			**Clase 2 / 10**		
Mármol	26,8	81,0	Platabanda	31,8	75,7
Tres cuartos con duchas	33,4	41,5	Mármol...	25,8	65,8
Bloque frisado	10,1	83,8	Concreto	46,2	26,4
Platabanda	17,8	50,0	Dos cuartos con duchas	15,7	27,6
Clase 3 / 10			**Clase 4 / 10**		
Bloque sin frisar	26,8	51,2	Bloque frisado	20,2	88,0
Container	32,1	30,8	Mármol	24,2	38,5
Poceta.Cloaca	19,7	58,1	Platabanda	20,0	29,6
Clase 5 / 10			**Clase 6 / 10**		
Pozo séptico	29,8	68,0	Adobe.Frisado	79,5	77,2
Bloque frisado	21,6	78,2	Teja	36,8	32,1
Asbesto	41,9	31,5	Casa	11,2	98,1
Lmina.Asfaltica	32,3	39,9	Cemeneto	10,1	80,8
Clase 7 / 10			**Clase 8 / 10**		
Sin aseo urbano	31,7	77,3	Camion.Cisterna	32,6	34,4
Combustible.Otro	39,7	36,0	Pared.Otra.Clase	32,0	32,7
Adobe sin frisar	45,9	24,9	Letrina	31,2	29,7
Sin poceta	35,4	41,1	Un.Cuarto	25,0	41,7
Clase 9 / 10			**Clase 10 / 10**		
Pared.Otra.Clase	47,1	79,5	Combustible.Otro	33,6	77,8
Rancho	36,1	86,6	Sin.Electricidad	35,5	69,6
Tierra	30,6	69,3	Sin.Poceta	26,7	79,3
Un.Cuarto	25,4	70,1	Tierra	28,5	67,9

Cuadro 3:
Coordenadas correspondientes a las variables incluidas para la caracterización de hogares con fecundidad alta.

Variable	Modalidad	Etiqueta	Coordenadas			
			1	2	3	4
Logro educativo colectivo en el hogar	Menos del 43 %	LEB	1,05	0,18	-0,58	0,09
	60% o mas	LEA	-0,96	-0,04	-0,35	0,42
Nivel educativo del conyuge	Sin nivel	NCS	1,29	-0,41	-0,91	0,11
	Primaria	NCP	0,29	-0,67	0,66	-0,68
	Secundaria	NCM	-0,9	-0,69	-0,23	1,07
	Universitaria	NCU	-1,75	-0,46	-2,14	-1,81
Nivel educativo del jefe	Sin nivel	NJS	1,12	0,27	-0,64	0,15
	Secundaria	NJM	-0,82	-0,06	-0,15	1,03
	Universitaria	NJU	-1,5	0,36	-1,53	-1,38
Situacion del jefe	Jefe mujer con conyuge	JMC	0,12	-0,43	0,19	-0,38
	Jefe hombre sin conyuge	JHS	0,25	1,72	0,13	-0,18
	Jefe mujer sin conyuge	JMS	-0,1	1,52	0,22	0,21
Urbano Rural	Hogar rural	RUR	1,37	-0,24	-0,81	0,14
Quintil de ingreso	1er quintil de ingresos	1QI	0,846	0,06	-0,26	0,1
	5to quintil de ingresos	2QI	-0,89	0,11	-0,78	-0,79
Pobreza	Hogar no pobre	HNP	-0,36	0,08	0,02	-0,06
	Hogar en pobreza extrema	HPE	1,19	-0,24	-0,44	0,12
Tipo de hogar	Hogar nuclear	THN	-0,08	-0,45	-0,06	0,14
	Hogar compuesto	THC	0,1	0,61	-0,16	-0,02
Fecundidad Alta	Fecundidad alta 15-19 años	F15-19	0,45	0,08	0,12	-0,07
	Fecundidad alta 40-19 años	F40-49	0,49	-0,01	-0,02	-0,08
	Fecundidad alta 20-24	F20-24	0,53	-0,2	0,11	0,02
	Fecundidad alta 25-26	F25-26	0,5	-0,31	0,08	0,06
	Fecundidad alta 30-34	F30-34	0,53	-0,34	0,05	0,05
	Hogar sin fecundidad alta	NOFA	-0,21	0,06	-0,02	0,01

Grafico 2: Primer plano factorial para el análisis de la Fecundidad Alta



4. CONCLUSIONES

La conclusión mas importante que se puede derivar de este trabajo, es que existen diversas estrategias de reducción de datos que permiten aplicar, sobre grandes bases de datos, diversas técnicas estadísticas, en especial de análisis multivariante de datos.

La estrategia de reducción de datos que se propuso solamente es aplicable en variables de tipo cualitativo, pero tiene la ventaja que es fácil de implementar, puede lograr un importante porcentaje de reducción y es equivalente a trabajar con la data original

Con relación a las aplicaciones que se presentaron, estas sirvieron para mostrar las potencialidades de la estrategia propuesta; pero un análisis más profundo de estos resultados sobrepasa los objetivos de este trabajo y las restricciones del mismo.

5. REFERENCIAS BIBLIOGRAFICAS

Correa E., Gerardo: Aplicación de técnicas de análisis multivariante para construir una tipología de viviendas en Venezuela basada en el censo 2001. Trabajo especial de grado para optar al titulo de Especialista en Estadística (No publicado). Universidad Central de Venezuela, Caracas 2005

Han J., Lamber M.: Data Mining: Concepts and Techniques. San Francisco, Estados Unidos, 2001

Hilera, J; Martinez V: Redes Neuronales Artificiales. Bogota, Colombia 2000.

Rodríguez V, Jorge: La fecundidad alta en América Latina y el Caribe: un riesgo en transición. (CELADE) – División de Población. Santiago de Chile, octubre de 2003