Revista Colombiana de Estadística
Volumen **23** (2000) No. 1 , páginas 1–13

ANÁLISIS DE LA MÉTRICA EN APLICACIONES DE LA ESTADÍSTICA TEXTUAL A LA TIPOLOGÍA DE TRAYECTORIAS

CAMPO ELIAS PARDO[1]

Resumen. Cuando se realizan las técnicas de estadística textual en el análisis de trayectorias se construyen tablas se contingencia, que cruzan los itinerarios con las palabras código, que indican las etapas del conjunto de individuos. A las tablas de contingencia se les aplica los análisis de correspondencias y de clasificación, utilizando la distancia ji–cuadrado. En este trabajo se analizan y comparan las distancias ji–cuadrado y los órdenes inducidas por éstas para trayectorias residenciales hipotéticas, considerando diferentes formas de codificación. Se observan los efectos de agregar palabras para tener en cuenta la noción de censura, agregar información y eliminar palabras de baja frecuencia. Se constata la alta conveniencia de introducir la censura en forma anualizada. Se muestra que la opción de adicionar información agreagada esta justificada desde el punto de vista de la distancia ji–cuadrado. Se observa que la eliminación de palabras por un umbral de frecuencia ocasiona un desajuste en los marginales de la tabla de contingencia, que queda fuera de control en el análisis.

Palabras claves: métrica, distancia ji–cuadrado, análisis de correspondencias, clasificación, análisis de datos biográficos, análisis de datos longitudinales, análisis de datos.

1. Introducción

Una de las estrategias posibles en el análisis de datos longitudinales cualitativos, en particular de datos biográficos, consiste en recodificar las trayectorias

(1) Profesor Departamento de Estadística, Universidad Nacional de Colombia; e-mail: cpardo@matematicas.unal.edu.co.

individuales en frases y aplicar luego las técnicas de la estadística textual presentadas por Lebart y Salem (1994). Una trayectoria se compone de estas etapas con información de la modalidad de una variable y la duración en ella del individuo que se analiza. La información de las trayectorias se suele almacenar en un archivo en el que cada registro corresponde a una etapa del individuo. La trayectoria de un individuo requiere de tantos registros como etapas haya tenido. Sea, por ejemplo, una persona que vivió en la localidad A entre los 20 y 25 años de edad, luego se pasó a B en donde permaneció hasta los 30 años y finalmente se movió a C (a los 31 años), en donde se encontraba que al momento de recoger la información y tenía una edad de 40 años. Esta información se puede disponer en tres registros así:

Individuo	Etapa	Modalidad	Edad Inicial
I1	1	A	20
I1	2	B	26
I1	3	C	31

En Montenegro y Pardo (1996) se presentan algunas recodificaciones y se seleccionan las más apropiadas para emplear la estadística textual en el análisis de las trayectorias residenciales de los datos de movilidad residencial en el área metropolitana de Bogotá de la encuesta de CEDE – ORSTOM (Dureau y colaboradores, 1994). Para el análisis textual, las trayectorias se comportan como respuestas a preguntas abiertas de encuestas y lo que se análiza son tablas léxicas construidas a partir de tales respuestas. Una tabla léxica es una tabla de contingencia que cruza las respuestas con las formas gráficas o palabras que cada individuo utiliza en su respuesta. En el caso de los itinerarios, una fila de la tabla léxica contiene la frecuencia de aparición de cada una de las palabras código en cada trayectoria. El significado preciso de la tabla léxica depende de la codificación adoptada. Sobre la tabla léxica se realiza un análisis de correspondencias simples, seguido de una clasificación y una descipción de las clases obtenidas utilizando palabras y respuestas características. En los métodos se utiliza la distancia ji–cuadrado que es una distancia entre perfiles. El objetivo de este trabajo es analizar el comportamiento de la distancia ji–cuadrado para las recodificaciones utilizadas en la referencia anterior.

2. Codificación

La codificación básica adoptada es la de una palabra por año de vida del individuo. La palabra contiene dos informaciones: la modalidad de la variable de estado y la edad del individuo. Para la trayectoria del ejemplo la frase es:

A20 A21 A22 A23 A24 A25 B26 B27 B28 B29 B30 C31 C32 C33 C34 C35 C36 C37 C38 C39 C40

Cada palabra de la frase indica, para un año, la modalidad de la variable en que se encuentra el individuo y la edad. Con las trayectorias de todos los individuos del estudio se construye una tabla léxica (individuos x palabras), la cual para este caso contiene ceros y unos, indicando si el individuo de la fila asume la palabra de la columna.

Uno de los problemas en el análisis de datos longitudinales es la censura. En el caso de trayectorias en donde se utiliza el tiempo biográfico, o sea, la edad del individuo se presenta censura a la derecha ya que para todos los individuos a información empieza una edad común pero termina a la edad que tenga al momento de la encuesta (Barbary, 1996). La censura implica que para los individuos más jovenes no hay información al final, puesto que no han vivido esas edades. Considérese un individuo con 30 años de edad que ha vivido entre los 20 y 25 en B y a los 26 años se pasó a A, en donde se encontraba al momento de la encuesta, teniendo 30 años de edad. Su trayectoria está censurada a la derecha pues pues aún no ha vivido entre los 31 y 40 años. Existe la opción de considerar o no la censura, una manera de registrarla es con una palabra común por cada año hasta llegar al final del tiempo considerado para el análisis. La frase con censura se puede escribir:

B20 B21 B22 B23 B24 B25 C26 C27 C28 C29 C30 Censura Censura Censura Censura Censura Censura Censura Censura Censura Censura

La otra opción es adicionar a cada palabra *Censura* la edad a que esta ocurre, en cuyo caso la trayectoria queda:

B20 B21 B22 B23 B24 B25 C26 C27 C28 C29 C30 Censura31 Censura32 Censura33 Censura34 Censura35 Censura36 Censura37 Censura38 Censura39 Censura40

Esta forma de introducir la censura se puede denominar: *censura anualizada*.

3. Tablas Léxicas y distancias ji-cuadrado

En el análisis de datos textuales primero se hace la codificación del "corpus" y se establece su vocabulario con las frecuencias de aparición de cada palabra, luego se suelen eliminar las palabras cuyas frecuencias sean iguales o inferiores a un umbral que se establece. Posteriormente se hace un análisis de correspondencias sobre la tabla léxica construida con el vocabulario retenido. Las filas de la tabla representan a los individuos y las columnas las palabras código retenidas.

El análisis de correspondencias simples utiliza la distancia ji-cuadrado entre perfiles para realizar una comparación entre los perfiles fila y los perfiles columna de una tabla de contingencia, en este caso la tabla léxica, construida a partir de las trayectorias residenciales de los individuos. Para tener un panorama inicial de los que sucede se analizan los cambios en la distancia ji-cuadrado entre los perfiles fila, a través de un ejemplo construido con siete trayectorias residenciales de individuos entre 20 y 30 años, las cuales se presentan en la tabla 1.

Tabla 1. Ejemplo de siete trayectorias residenciales entre 20 y 30 años

- *I1: Santafe20-Santafe30: ha vivido de los 20 a los 30 años en Santafe*
- *I2: Usaquén20-Usaquén30: ha vivido de los 20 a los 30 años en Usaquén*
- *I3: Usaquén20-Usaquén25: ha vivido de los 20 a los 30 años en Usaquén. (censurado 5 años, tiene 25 años de edad)*
- *I4: Santafe20-Santafe21: ha vivido de los 20 y 21 años en Santafe. (censurado 9 años, tiene 21 años)*
- *I5: Usaquén20-Usaquén25, Santafe26-Santafe30*
- *I6: Santafe20-Santafe25 (censurado 5 años)*
- *I7: Usaquén20-Usaquén25, Chapinero26-Chapinero30*

La tabla 2 presenta la tabla léxica, la cual, por comodidad, se ha escrito en forma traspuesta, a la derecha de ésta se presenta la suma de frecuencias de las palabras y abajo la suma de palabras para cada individuo. La suma inferior es 11 para todos los individuos no censurados. Los valores de cada celda sólo pueden tomar los valores 0 o 1; 1 en el caso que el individuo haya vivido en la modalidad y edad indicada por la palabra y cero en el caso contrario.

Con el objeto de observar la influencia de la censura y la eliminación de palabras por un umbral de frecuencias se calcula la distancia ji-cuadrado entre individuos de la tabla 2 y de otras tablas derivadas de ésta, como se describe a continuación:

1. (D1): Para la tabla 2.
2. (D2): Para una tabla formada por la tabla 2 y la tabla 2a, es decir adicionando la palabra Censura, con la frecuancia 0,0,5,9,0,5 y 0 respectivamente, para lograr la misma suma marginal en los individuos.
3. (D3): Para una tabla formada por la tabla 2 y la tabla 2b, o sea adicionando palabras Censura, con la edad en que ocurre (por ejemplo Censura25). Todos los individuos tienen 1 en 11 palabras.
4. (D4): Para una tabla con información agregada (a la tabla 2, se agrega la tabla 2b y la tabla 2c)
5. (D5): Para la tabla 2d, derivada de la utilizada en D3, eliminando las 3 series de palabras con frecuencia 1. *(Chapinero26-Chapinero30, Usaquén26-Usaquén30 y censura22 a Censura25)*

Tabla 2. Tabla léxica (traspuesta)

Localidad	I1	I2	I3	I4	I5	I6	I7	Marginal
Chapinero26							1	1
Chapinero27							1	1
Chapinero28							1	1
Chapinero29							1	1
Chapinero30							1	1
Santafe20	1			1		1		3
Santafe21	1			1		1		3
Santafe22	1					1		2
Santafe23	1					1		2
Santafe24	1					1		2
Santafe25	1					1		2

Tabla 2. Tabla léxica (traspuesta– Continuación)

Localidad	I1	I2	I3	I4	I5	I6	I7	Marginal
Santafe26	1				1			2
Santafe27	1				1			2
Santafe28	1				1			2
Santafe29	1				1			2
Santafe30	1				1			2
Usaquén20		1	1		1		1	4
Usaquén21		1	1		1		1	4
Usaquén22		1	1		1		1	4
Usaquén23		1	1		1		1	4
Usaquén24		1	1		1		1	4
Usaquén25		1	1		1		1	4
Usaquén26		1						1
Usaquén27		1						1
Usaquén28		1						1
Usaquén29		1						1
Usaquén20		1						1
Marginal	11	11	6	2	11	6	11	58

Tabla 2a. Con una palabra censura

Localidad	I1	I2	I3	I4	I5	I6	I7	Marginal
Censura	0	0	5	9	0	5	0	19
Marginal	11	11	11	11	11	11	11	77

Tabla 2b. Con una palabra censura por cada año

Localidad	I1	I2	I3	I4	I5	I6	I7	Marginal
Censura22				1				1
Censura23				1				1
Censura24				1				1
Censura25				1				1
Censura26			1	1		1		3
Censura27			1	1		1		3
Censura28			1	1		1		3
Censura29			1	1		1		3
Censura30			1	1		1		3
Marginal	11	11	11	11	11	11	11	77

Tabla 2c. Con información agregada

Localidad	I1	I2	I3	I4	I5	I6	I7	Marginal
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Chapinero							5	5
Santafe	I1			2	5	6		24
Usaquen		I1	6		6		6	29
Censura	0	0	5	9	0	5	0	19
Marginal	22	22	22	22	22	22	22	154

Tabla 2d. Tabla léxica eliminando palabras con frecuencia 1 (traspuesta)

Localidad	I1	I2	I3	I4	I5	I6	I7	Marginal
Santafe20	1			1		1		3
Santafe21	1			1		1		3
Santafe22	1					1		2
Santafe23	1					1		2
Santafe24	1					1		2
Santafe25	1					1		2
Santafe26	1				1			2
Santafe27	1				1			2
Santafe28	1				1			2
Santafe29	1				1			2
Santafe30	1				1			2
Usaquen20		1	1		1		1	4
Usaquen21		1	1		1		1	4
Usaquen22		1	1		1		1	4
Usaquen23		1	1		1		1	4
Usaquen24		1	1		1		1	4
Usaquen25		1	1		1		1	4
Censura26			1	1		1		3
Censura27			1	1		1		3
Censura28			1	1		1		3
Censura29			1	1		1		3
Censura30			1	1		1		3
Marginal	11	6	11	7	11	11	6	63

3.1 Distancias ji–cuadrado entre perfiles individuos de la tabla

Utilizando la nomenclatura de Lebart y colaboradores (1995) la distancia ji-cuadrado entre dos perfiles fila i y k se escribe:

$$d^2(i,k) = \sum_{j=1}^{p} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{kj}}{f_{k\cdot}} \right)^2 \tag{1}$$

donde $f_{ij}/f_{i\cdot}$ es un elemento del perfil para la fila i y $f_{kj}/f_{k\cdot}$ un elemento del perfil para la fila k, expresados apartir de la matriz de frecuencias relativas $\mathbf{F}$, siendo $f_{i\cdot}$ y $f_{k\cdot}$ las marginales respectivas y $f_{\cdot j}$ la marginal de la columna j. Los perfiles se pueden obtener directamente de la tabla léxica, denotada como $\mathbf{K}$ (traspuesta de la tabla 1) y los $f_{\cdot j}$ se encuentran dividiendo los marginales columna (filas de la tabla 1) por 58.

Si dos individuos tienen valores de sus marginales diferentes, entonces las coordenadas en el perfil para las modalidades en común que asuman van a ser también diferentes aportando valores a la distancia ji-cuadrado. La diferencia de las marginales

para los individuos se origina en la censura de uno de ellos. El valor del aporte a la distancia entre dos individuos depende también del inverso de la frecuencia marginal de la palabra ($f_{\cdot j}$), es decir 58/frecuencia marginal, o sea que a mayor frecuencia de las palabras la contribución a la distancia se hace menor.

Los anteriores hechos se constatan en la tabla 3 en donde se muestra la distancia D1 (columna 1) que es la raíz cuadrada de la expresión (1), entre los individuos indicados por la columna 2. Las distancias aparecen ordenadas de menor a mayor. La tercera columna indica las frecuencias marginales de los individuos de la tabla léxica. En las columnas 4 y 5 aparecen la trayectorias común y diferente, respectivamente, entre los individuos, se indica entre paréntesis la frecuencia de la serie de palabras. La última columna registra los sumandos a la distancia ji-cuadrado.

La menor distancia es *1.30* entre los individuos *I3* e *I5*, la trayectoria común aporta a la distancia debido a que *I3* esta censurado. Lo mismo sucede entre *I2* e *I3* (D1=1.70), ésta es mayor debido a que la parte diferente se debe a una serie de palabras de frecuencia 1 comparado con las del caso anterior que son de frecuencia 2.

En la distancia entre I1 e I5 la parte común no suma pues ninguno de los dos individuos está censurado. Las distancias D1(2,7) es mayor que las distancias D1(2,5) y D1(5,7) debido a que en conjunto hay menor frecuencia entre las palabras que suman a la distancia.

Comparando las distancias D1(2,3) y D1(3,7) con D1(2,5) y D1(5,7) se observa la influencia de la censura de dos maneras: por un lado hace sumar a las distancias la parte común y por el otro hay un sumando menos, en comparación a donde no hay censura.

La mayor censura del individuo 4 hace que las distancias entre éste y los demás sean las mayores de todas, de hecho son las últimas filas de la tabla 3. Las dos mayores distancias son las de este individuo con las de trayectorias completas sin palabras en común.

Tabla 3. Elementos de la distancia ji–cuadrado de la tabla 2

1 D1	2 Entre	3 Margen	4 Trayectoria común	5 Trayectoria diferente	6 Sumandos de la distancia ji-cuadrado
1.30	I3,I5	6,I1	Usaquén20-25(4) (4)	C26-C30: Santafe26-30(2)	6x0.08+5x0.24
1.41	I1,I5	I1,I1	Santafe26-30 (2)	Santafe20-21(3)-Santafe22-25(2): Usaquén20-25(4)	2x0.16+4x0.24+6x0.12
1.44	I1,I6	I1,6	Santafe20-25 (3,2)	Santafe26-30(2): C26-C30	2x0.11+4x0.17+5x0.24
1.70	I2,I3	I1,6	Usaquén20-25 (4)	Usaquén26-30(1): C26-C30	6x0.08+5x0.48
1.70	I3,I7	6,I1	Usaquén20-25 (4)	C26-C30: Chapinero26-30(1):	6x0.08+5x0.48
1.90	I2,I5	I1,I1	Usaquén20-25 (4)	Usaquén26-30(1): Santafe26-30(2)	5x0.48+5x0.24
1.90	I5,I7	I1,I1	Usaquén20-25 (4)	Santafe26-30(2): Chapinero26-30(1)	5x0.48+5x0.24
2.19	I2,I7	I1,I1	Usaquén20-25 (4)	Usaquén26-30(1): Chapinero26-30(1)	10x0.48
2.21	I1,I3	I1,6		Santafe20-21(3)-Santafe22-30(2): Usaquén20-Usaquén25(4)	2x0.16+9x0.24+6x0.40

Tabla 3. Elementos de la distancia ji–cuadrado de la tabla 2. (Continuación)

1 D1	2 Entre	3 Margen	4 Trayectoria común	5 Trayectoria diferente	6 Sumandos de la distancia ji-cuadrado
2.36	I1,I2	I1,I1	No censurados	Santafe20-21(3)-Santafe22-30(2): Usaquén20-Usaquén25(4)	2x0.16+9x0.24+6x0.12+5x0.48
2.36	I1,I7	I1,I1	No censurados	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Chapinero26-30(1)	2x0.16+9x0.24+6x0.12+5x0.48
2.49	I5,I6	I1,6		Usaquén20-25(4)-Santafe26-30(2): Santafe20-21(3)-Chapinero22-25(2)	2x0.54+4x0.81+5x0.24+6x0.12
2.59	I3,I6	6,6	Censurado 5 años	Usaquén20-25(4): Santafe20-21(3)-Santafe22-25(2)	2x0.54+4x0.81+5x0.40
2.72	I2,I6	I1,6		Usaquén20-25(4)-Usaquén26-30(1): Santafe20-21(3)-Santafe22-25(2)	2x0.54+4x0.81+6x0.12+5x0.48
2.72	I6,I7	6,I1		Santafe20-21(3)-Santafe22-25(2): Usaquén20-25(4)-Chapinero26-30(1)	2x0.54+4x0.81+6x0.12+5x0.48
2.74	I4,I6	2,6	Censura26-30	Santafe20-21(3): Santafe20-21(3)-Santafe22-25(2)	2x2.15+4x0.81
2.94	I1,I4	I1,2	Santafe20-21 (3)	Santafe20-21(3)-Santafe22-30(2): Santafe20-21(3)	2x3.24+9x0.24
3.40	I4,I5	2,I1		Santafe20-21(3): Usaquén20-25(4)-Santafe26-30(2)	2x4.83+5x0.24+6x0.12
3.48	I3,I4	6,2	Censura26-30	Usaquén20-25(4): Santafe20-21(3)	6x0.402x4.83
3.58	I2,I4	I1,2		Usaquén20-25(4)-Usaquén26-30(1): Santafe20-21(3)	2x4.83+6x0.12+5x0.48
3.58	I4,I7	2,I1		Santafe20-21(3): Usaquén20-25(4)-Chapinero26-30(1)	5x0.48+2x4.83+5x0.12

3.2 Modificación de la distancia al tener en cuenta la censura

La introducción de la censura a la tabla léxica se puede hacer de dos maneras: introduciendo una sola palabra *censura* (tabla 2a); o introduciendo palabras compuestas por *censura* y la edad en que ocurre (tabla 2b). En los dos casos la frecuencia total de las nuevas tablas léxicas aumenta en 19 = 0+0+5+9+0+5+0, es decir a 77. Hay dos causas para aumentar las distancias: más palabras y menores valores de las frecuencias marginales. Pero también hay una razón para que disminuyan entre individuos con palabras comunes, pues ahora el aporte de la parte común es nulo. De este modo se produce también un acercamiento relativo de los individuos simultáneamente censurados.

Los efectos mencionados se pueden apreciar en la tabla 4, en donde las columnas 4, 6 y 9 representan el orden obtenido según las distancias D1, D2 y D3, respectivamente. D2 es la distancia para el caso de introducir una sola palabra y D3 para una por cada año *(censura+edad)*. En las distancias D2 y D3 se observa un aumento a 115.2%, igual a $\sqrt{\frac{77}{58}}$, para todas las parejas de individuos sin censura. Este incremento se debe al aumento del tamaño de la tabla por la presencia de nuevas palabras, lo que ocasiona un aumento del peso en el cálculo de la distancia. Este valor de 115.2% se constituye en referencia para ver las variaciones en la distancias debidas a otras causas. La tabla 4 está ordenada de menor a mayor por la columna 7 (D2/D1). Las reducciones más drásticas se dan para las dos distancias que involucran las

mayores censuras, debido a la disminución de las coordenadas de sus perfiles y a un parecido en las palabras de censura. La introducción de la censura ocasiona una reducción de la distancia en la mayoría de los casos (comparado con el 115%), sólo se excluyen las tres últimas filas de la tabla, en donde hay un aumento de las distancias. En estos tres últimos casos las parejas de individuos considerados tienen los primeros 6 años comunes con las palabras de alta frecuencia, entonces el aumento por la inclusión de las palabras de censura es mayor que la disminución por la trayectoria común.

Las columnas 4, 6 y 9 se han incluido para comparar los órdenes con las tres distancias, puesto que las disminuciones y aumentos relativos obviamente cambian tales ordenes. Comparando la distancias D2 y D3 se observa una disminución mayor cuando la censura se introduce como una sola palabra.

Tabla 4. Comparaciones de distancias D1, D2 y D3

1 D1	2 Entre	3 Margen	4 O1	5 D2	6 O2	7 D2/D1	8 D3	9 O3	10 D3/D1	11 Trayectoria completa
3.48	I3,I4	6,2	19	1.38	2	39.8%	1.98	6	57.0%	Usaquén20-25(4)-C26-30: Santafe20-21(3):C22-30
2.74	I4,I6	2,6	16	1.34	1	49.0%	1.95	5	71.3%	Santafe20-21(3):C22-30: Santafe20-25(3,2):C26-30
2.59	I3,I6	6,6	13	1.63	5	62.8%	1.63	1	62.8%	Usaquén20-25(4)-C26-30: Santafe20-25(3,2):C26-30
3.40	I4,I5	2,11	18	2.38	14	70.0%	2.56	15	75.3%	Santafe20-21(3)-C26-30: Usaquén20-25(4)-Santafe26-30(2):
3.58	I2,I4	11,2	20	2.70	18	75.4%	2.86	20	79.9%	Usaquén20-25(4)-Usaquén26-30(1): Santafe20-21(3)-C22-30
3.58	I4,I7	2,11	20	2.70	18	75.4%	2.86	20	79.9%	Santafe20-21(3)-C22-30: Usaquén20-25(4)-Chapinero26-30(1)
2.94	I1,I4	11,2	17	2.36	13	80.4%	2.54	14	86.6%	Santafe20-21(3)-Santafe22-30(2): Santafe20-21(3)-C22-30
2.49	I5,I6	11,6	12	2.25	11	90.4%	2.30	11	92.4%	Usaquén20-25(4)-Santafe26-30(2): Santafe20-25(3,2)-C26-30
2.72	I2,I6	11,6	14	2.58	16	94.9%	2.63	16	96.4%	Usaquén20-25(4)-Usaquén26-30(1): Santafe20-25(3,2)-C26-30
2.72	I6,I7	6,11	14	2.58	16	94.9%	2.63	16	96.4%	Santafe20-21(3)-Santafe22-25(2): Usaquén20-25(4)-Chapinero26-30(1)
2.21	I1,I3	11,6	9	2.25	11	101.9%	2.30	11	104.1%	Santafe20-21(3)-Santafe22-30(2): Usaquén20-Usaquén25(4)-C26-30
1.44	I1,I6	11,6	3	1.56	3	107.9%	1.63	1	112.7%	Santafe20-21(3)-Santafe22-30(2): Santafe20-25(2,3)-C26-30
1.41	I1,I5	11,11	2	1.63	5	115.2%	1.63	1	115.2%	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Santafe26-30(2)
1.90	I2,I5	11,11	6	2.18	9	115.2%	2.18	9	115.2%	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-Santafe26-30(2)
1.90	I5,I7	11,11	6	2.18	9	115.2%	2.18	9	115.2%	Usaquén20-25(4)-Santafe26-30(2): Usaquén20-25(4)-Chapinero26-30(1)
2.19	I2,I7	11,11	8	2.52	15	115.2%	2.52	13	115.2%	Usaquén20-25(4)-Santafe26-30(1): Usaquén20-25(4)-Chapinero26-30(1)
2.36	I1,I2	11,11	10	2.72	20	115.2%	2.72	18	115.2%	Santafe20-21(4)-Santafe22-30(2): Usaquén20-25(4)-Usaquén26-30(1)
2.36	I1,I7	11,11	10	2.72	20	115.2%	2.72	18	115.2%	Santafe20-21(4)-Santafe22-30(2): Usaquén20-25(4)-Chapinero26-30(1)
1.70	I2,I3	11,6	4	2.00	7	117.8%	2.06	7	121.0%	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-C26-30
1.70	I3,I7	6,11	4	2.00	7	117.8%	2.06	7	121.0%	Usaquén20-25(4)-C26-C30: Usaquén20-25(4)-Chapinero26-30(1)
1.30	I3,I5	6,11	1	1.56	3	119.6%	1.63	1	125.0%	Usaquén20-25(4)-C26-C30: Usaquén20-25(4)-Santafe26-30(2)

En la tabla 5 se comparan los órdenes incluidos por las distancias, los cuales aparecen en las columnas O1, O2 y O3 respectivamente. La tabla 5 yuxtapone dos tablas: la primera, ordenada por O2 y la segunda por O3. El ordenamiento O1 está en la tabla 3. Este ejemplo parece indicar que la codificación con censura anualizada es la que más se ajusta a la similitud entre las trayectorias.

Tabla 5. Comparación de órdenes D2 y D3

O1	O2	O3	Trayectoria completa	O1	O2	O3	Trayectoria completa
16	1	5	Santafe20-21(3)-C22-30: Santafe20-25(3,2)-C26-30	3	3	1	Santafe20-21(3)-Santafe22-30(2): Santafe20-25(2,3)-C26-30
19	2	6	Usaquén20-25(4)-C26-30: Santafe20-21(3):C22-30	1	3	1	Usaquén20-25(4)-C26-30: Usaquén20-25(4)-Santafe26-30(2):
3	3	1	Santafe20-21(3)-Santafe22-30(2): Santafe20-25(2,3)-C26-30	13	5	1	Usaquén20-25(4)-C26-C30: Santafe20-25(3,2)-C26-30
1	3	1	Usaquén20-25(4)-C26-30: Usaquén20-25(4)-Santafe26-30(2):	2	5	1	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Santafe26-30(2)
13	5	1	Usaquén20-25(4)-C26-C30: Santafe20-25(3,2)-C26-30	16	1	5	Santafe20-21(3)-C22-30: Santafe20-25(3,2)-C26-30
2	5	1	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Santafe26-30(2)	19	2	6	Usaquén20-25(4)-C26-30: Santafe20-21(3):C22-30
4	7	7	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-C26-30	4	7	7	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-C26-30
4	7	7	Usaquén20-25(4)-C26-30: Usaquén20-25(4)-Chapinero26-30(1)	4	7	7	Usaquén20-25(4)-C26-30: Usaquén20-25(4)-Chapinero26-30(1)
6	9	9	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-Santafe26-30(2)	6	9	9	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-Santafe26-30(2)
6	9	9	Usaquén20-25(4)-Santafe26-30(2): Usaquén20-25(4)-Chapinero26-30(1)	6	9	9	Usaquén20-25(4)-Santafe26-30(2): Usaquén20-25(4)-Chapinero26-30(1)
12	11	11	Usaquén20-25(4)-Santafe26-30(2): Santafe20-25(3,2)-C26-30	12	11	11	Usaquén20-25(4)-Santafe26-30(2): Santafe20-25(3,2)-C26-30
9	11	11	Santafe20-21(3)-Santafe22-30(2): Usaquén20-Usaquén25(4)-C26-30	9	11	11	Santafe20-21(3)-Santafe22-30(2): Usaquén20-Usaquén25(4)-C26-30
17	13	14	Santafe20-21(3)-Santafe22-30(2): Santafe20-21(3)-C22-30	8	15	13	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-Chapinero26-30(1)
18	14	15	Santafe20-21(3)-C26-30: Usaquén20-25(4)-Santafe26-30(2)	17	13	14	Santafe20-21(3)-Santafe22-30(2): Santafe20-21(3)-C22-30
8	15	13	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-Chapinero26-30(1)	18	14	15	Santafe20-21(3)-C26-30: Usaquén20-25(4)-Santafe26-30(2)
14	16	16	Usaquén20-25(4)-Usaquén26-30(1): Santafe20-25(3,2)-C26-30	14	16	16	Usaquén20-25(4)-Usaquén26-30(1): Santafe20-25(3,2)-C26-30
14	16	16	Santafe20-21(3)-Santafe22-25(2): Usaquén20-25(4)-Chapinero26-30(1)	14	16	16	Santafe20-21(3)-Santafe22-25(2): Usaquén20-25(4)-Chapinero26-30(1)
20	18	20	Usaquén20-25(4)-Usaquén26-30(1): Santafe20-21(3)-C22-30	10	20	18	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Usaquén26-30(1)
20	18	20	Santafe20-21(3)-C22-30: Usaquén20-25(4)-Chapinero26-30(1)	10	20	18	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Chapinero26-30(1)
10	20	18	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Usaquén26-30(1)	20	18	20	Usaquén20-25(4)-Usaquén26-30(1): Santafe20-21(3)-C22-30
10	20	18	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Chapinero26-30(1)	20	18	20	Santafe20-21(3)-C22-30: Usaquén20-25(4)-Chapinero26-30(1)

3.3 Distancias con información agregada (D4)

En el análisis todas las formas gráficas (palabras) son diferentes sin importar los caracteres que contienen, por ejemplo *santafe20 y Santafe21* son palabras tan diferentes como *Chapinero20 y Santafe30*. Sin embargo *Santafe20 y Santafe 21* se parecen en el sentido que corresponden a la misma localidad, mientras que *Chapinero 20* corresponde a otra localidad. Una de las codificaciones propuestas en Montenegro y Pardo (1996) agrega información para introducir este

parecido en la codificación textual de la treyectorias. Al obtener la tabla léxica aparecen nuevas columnas que serían como la yustaposición de un nueva tabla a la inicial. Para verlo en el ejemplo se introducen las palabras *Santafe, Usaquén y Chapinero* con frecuencias que indican el número de años que el individuo estuvo en cada localidad. Para no desbalancear las frecuencias marginales de los individuos se incluye también la palabra *Censura*. En estas condiciones la tabla es la tabla 2 adicionada de la 2b y de la 2c.

En la tabla 6 se encuentran las distancias D4 y el orden inducido por ésta (O4), para comparación se incluyen los ordenamientos según la distancia D3. Las distancias D4 son siempre menores que las D3 (ver columna D4/D3 de la tabla 6), mostrando el acercamiento logrado por la información agregada. Esto se debe a una disminución global de $\frac{1}{\sqrt{2}}$ equivalente al 70.7%, debida al aumento del tamaño de la tabla al doble y a un aumento relativo a las 4 palabras nuevas, que es inferior debido a que las frecuencias son mayores en comparación con las palabras originales. De esa forma se logra un acercamiento por el hecho de haber estado en la localidad aunque sea en diferentes tiempos.

3.4 Efecto de la eliminación de palabras

En el análisis de tablas léxicas se suelen eliminar las palabras que tienen frecuencia igual o menor a un umbral que se establece, ya que estas palabras crearían los primeros ejes, lo cual no es interesante cuando se buscan individuos que se expresen con vocabulario similar. En el análisis de trayectorias esa eliminación implica perder etapas que son menos comunes y que no interesan para establecer grupos con trayectorias semejantes. Al reducir el número de palabras en la tabla léxica se desbalancea de nuevo la marginal de los individuos, lo que acarrea los problemas ya anotados en el cálculo de D1. Para ilustrar se retiran las series de palabras de frecuencia 1 de la tabla léxica con censura anualizada (tabla 2d) y se calcula la distancia D5. Los valores de D5 y su orden inducido O5 se encuentra también en la tabla 6.

Se observa en la columna D5/D3 de la tabla 6 una reducción de 90.4%, para las cuatro primeras distancias, ellas involucran individuos que no han perdido información al eliminar las palabras, con lo cual conservan su marginal en 11, entonces el 90.4% es simplemente $\sqrt{\frac{63}{77}}$ o sea se debe al cambio del tamaño de la tabla. Los individuos I2 e I7 quedan a distancia cero ya que al eliminar las palabras de frecuencia uno quedó para ellos solo la información de la parte común. Por el mismo motivo aparecen otras reducciones importantes en parejas de individuos.

Tabla 6: Comparaciones de distancias D3, D4 y D5

Entre	D3	O3	D4	D/D3	O4	D5	D5/D3	O5	Trayectoria completa
11.16	1.63	1	1.44	88.4%	1	1.47	90.4%	8	Santafe20-21(3):C22-30: Usaquén25(4)-Santafe26-30(2)
I3.I5	1.63	1	1.44	88.4%	1	1.47	90.4%	8	Santafe20-21(3):C22-30: Santafe20-25(2.3)-C26-C30
I1.I5	1.63	1	1.48	91.0%	3	1.47	90.4%	8	Usaquén20-25(4)-C26-30: Santafe20-25(3,2):C26-30
I3.I6	1.63	1	1.48	91.0%	3	1.47	90.4%	8	Usaquén20-25(4)-C26-C30: Usaquén20-25(4)-Santafe26-30(2):
I4.I6	1.95	5	1.55	79.3%	5	1.20	61.5%	4	Santafe20-21(3)-C22-30 Santafe20-25(3,2)-C26-30
I3.I4	1.98	6	1.64	82.7%	6	1.39	70.0%	7	Usaquén20-25(4)-C26-C30: Santafe20-21(3)-C22-30:
I2.I3	2.06	7	1.68	81.4%	7	1.19	57.6%	2	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-C26-30
I2.I5	2.18	9	1.73	79.3%	8	1.36	62.3%	5	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-Santafe26-30(2):
I5.I6	2.3	11	1.87	81.1%	9	2.08	90.6%	12	Usaquén20-25(4)-Santafe26-30(2): Santafe20-25(3,2)-C26-30
I3.I7	2.06	7	2.03	98.7%	10	1.19	57.6%	2	Usaquén20-25(4)-C26-30: Usaquén20-25(4)-Chapinero26-30(1)
I5.I7	2.18	9	2.08	95.2%	11	1.36	62.3%	5	Usaquén20-25(4)-Santafe26-30(2): Usaquén20-25(4)-Chapinero26-30(1)
I3.I5	2.30	11	2.25	97.9%	12	2.18	90.6%	12	Santafe20-21(3)-Santafe22-30(2): Usaquén20-Usaquén25(4)-C26-30
I2.I7	2.52	13	2.25	89.1%	12	0.00	0.0%	1	Usaquén20-25(4)-Usaquén26-30(1): Usaquén20-25(4)-Chapinero26-30(1)
I4.I5	2.56	15	2.27	88.7%	14	2.25	88.1%	17	Santafe20-21(3)-C26-30: Usaquén20-25(4)-Santafe26-30(2)
I1.I4	2.54	14	2.38	92.7%	15	2.14	84.4%	14	Santafe20-21(3)-Santafe22-30(2): Santafe20-21(3)-C22-30
I2.I6	2.63	16	2.38	90.5%	15	2.21	84.0%	15	Santafe20-21(3)-Santafe22-25(2): Usaquén20-25(4)-Chapinero26-30(1)
I6.I7	2.63	16	2.52	95.7%	17	2.21	84.0%	15	Usaquén20-25(4)-Usaquén26-30(1) Santafe20-25(3,2)-C26-30
I1.I2	2.72	18	2.58	94.8%	18	2.31	84.8%	18	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Chapinero26-30(1)
I2.I4	2.86	20	2.61	91.3%	19	2.37	82.9%	20	Santafe20-21(3)-C22-30: Usaquén20-25(4)-Chapinero26-30(1)
I1.I7	2.72	18	2.70	99.3%	20	2.31	84.8%	18	Santafe20-21(3)-Santafe22-30(2): Usaquén20-25(4)-Usaquén26-30(1)
I4.I7	2.86	20	2.73	95.6%	21	2.37	82.9%	20	Usaquén20-25(4)-Santafe26-30(2): Santafe20-21(3)-C22-30

4. Conclusiones

Se constata la alta conveniencia de introducir la censura en la codificación de las trayectorias como frases. La opción de introducir la censura en forma anualizada (censura + edad) se comporta mejor que introduciendo la sola palabra censura repetida en cada año que ocurre. La segunda opción tiende a acercar más a los individuos por el solo hecho de estar censurados.

La opción de adicionar información agregada esta justificada desde el punto de vista de la distancia ji-cuadrado. La distancia se calcula sobre dos tablas yuxtapuestas y entonces conviene agregar información que mantenga el marginal de los individuos para que la distancia tenga interpretación.

la eliminación de palabras por un umbral además de la pérdida de información altera los marginales de los individuos y entonces vuelven a aparecer

los problemas de la tabla original, cuando no se incluía la censura (tabla 2): sin embargo, en el análisis de correspondencias los individuos que han perdido información tendrán menos peso (menos frecuencia marginal) lo cual podría contrarrestar un poco ese efecto.

Se destaca la necesidad de considerar lo que sucede con la distancia ji–cuadrado cuando se construye y transforman tablas que se van a someter al análisis de correspondencias.

Referencias

[1] Barbary O. (1996). *A*nálisis tipológico de datos biográficos en Bogotá, Departamento de Matemáticas y Estadística. Universidad Nacional de Colombia. Santafe de Bogotá.

[2] Durea F., Florez C. E., Barbary O., Garcia L., Hoyos M. C. (1994 *L*a movilidad de las poblaciones y su impacto sobre la dinámica del área metropolitana de Bogotá, metodología de la encuesta cuantitativa, Docuemnto de Trabajo no. 2, CEDE/ORSTOM, Bogotá.

[3] Lebart L, Salem A., (1994) *S*tatistique Textuelle. DUNOD, París.

[4] Lebart L, Salem A. (1995). *S*tatistique Exploratoire Multidimensionnelle. DUNOD, París.

[5] Montenegro A., Pardo C.E., (1996) *L*os itinerarios individuales interpretados como frases, una aplicación de la estadística textual a la tipología de trayectorias, en: Memorias del seminario de Capacitación e Investigación: Recolección y análisis de datos longitudinales. Bogotá 9–13 de diciembre de 1996. Universidad Nacional de Colombia, PRESTA, ORSTOM, Bogotá.