Seminario de investigación sobre análisis estadístico de datos biográficos Análisis Armónico Cualitativo y datos biográficos, Maestria de Estadistica. Facultad de Matemáticas y Estadística, Universidad Nacional de Colombia, Bogota.

Olivier Barbary

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Matemáticas y
Estadística

Seminario de investigación sobre análisis estadístico de datos biográficos
Coordinadores: O. Barbary, C.E. Pardo, J. Ramos

ANÁLISIS ARMÓNICO CUALITATIVO Y DATOS BIOGRÁFICOS

Olivier Barbary : ORSTOM, profesor invitado U.N.C.

Santafé de Bogotá, febrero de 1996

INTRODUCCIÓN

El seminario de investigación sobre análisis estadísticos de datos biográficos que se enmarca dentro del programa de cooperación entre la Universidad Nacional de Colombia y el Instituto francés ORSTOM, tiene como objetivo principal el desarrollo de técnicas estadísticas de **análisis de datos adaptadas al análisis de las informaciones biográficas** recolectadas en las encuestas socio demográficas. Para tal propósito los integrantes del seminario han definido cinco lineas de investigación:

1. Análisis de correspondencia de datos longitudinales cualitativos
2. Análisis conjunto de tablas cualitativas
3. Estadística textual y datos biográficos
4. Modelos probabilísticos y datos biográficos
5. Análisis de datos biográficos y teoría de rachas

En estas sesiones, se presenta la primera linea metodológica definida : las aplicaciones del análisis de correspondencia y más precisamente el método llamado "**análisis armónico cualitativo**".

En la primera parte se quiere hacer una aproximación rápida a la **recolección y codificación de los datos longitudinales categóricos** y a los principales enfoques **metodológicos para el análisis** de estos, lo que nos permitirá llegar a la conclusión de la necesidad de una metodología que identifique grupos de individuos con trayectorias similares, es decir **un método de análisis tipológico de los datos biográficos**.

La segunda parte, como repuesta a esta necesidad, introduce el método **de análisis armónico cualitativo**, cuyos fundamentos teóricos se describen sintéticamente hasta llegar a una propiedad fundamental en la práctica : la equivalencia **con un análisis de correspondencia particular**. Luego se revisan las técnicas de aproximación numéricas empleadas para su ejecución informática.

Finalmente en la tercera y última parte, se presenta una **aplicación** del método a los datos de la encuesta realizada en 1993 por el programa desarrollado conjuntamente por el Cede (Universidad de los Andes) y Orstom, programa titulado “La movilidad de las poblaciones y su impacto sobre la dinámica del área metropolitana de Bogotá”.

PRIMERA PARTE : **Datos biográficos, conceptos, notaciones y metodologias de análisis.**

1.1 Conceptos generales, recolección y captura de la información.

En los estudios de ciencias sociales, los datos longitudinales también llamados **datos biográficos** aparecen cada vez con mayor frecuencia. En muchas encuestas sociodemográficas, se busca conocer, mediande cuestionarios de tipo "**calendarios**", la cantidad de tiempo que los individuos transcurren en una serie de estados para identificar la **trayectoria** residencial, educacional, profesional, matrimonial, etc. de los individuos.

Existe dos **técnicas principales de recolección** para datos biográficos : los **cuestionarios retrospectivos** y las encuestas por visitas repetidas.

En el primer método, el investigador registra aquellos estados que la persona especifica haber conocido durante determinado tiempo. Esta metodología tiene la ventaja que la observación puede ser **exhaustiva y continua sobre toda la vida del individuo** y por eso es la mas empleada. Pero también tiene la desventaja que la veracidad de las fechas y estados se basan en la buena memoria del entrevistado; por lo tanto la cualidad de la información depende en gran parte de un **diseño adecuado del cuestionario** y de la buena capacitación de los encuestadores.

El segundo método se puede llevar a cabo de dos maneras : el investigador hace visitas periódicas, y registra cada una de estas; este método es comúnmente llamado **panel**; la otra manera es mediante cuestionario **auto aplicado** en donde la persona va llevando un **diario** de las actividades que desarrolló.

La **captura** de los datos también se puede hacer de dos formas.

- La captura con **un registro por evento o etapa**. Así para una persona se tienen tantos registros como eventos o etapas haya presenciado esta. Esta forma de captura permite tener registros cortos y de tamaño fijo.

- La captura con **un registro por individuo**. Cada registro es extenso y de tamaño variable ya que contiene todos los tiempos y cambios de estado de cada individuo. Generalmente los Software estadísticos no manejan registros de tamaño variable. Por esta razón, se prefiere generalmente la otra forma de captura.

1.2 : Notaciones y estructuración de la información longitudinal.

Las notaciones que se utilizan para los datos longitudinales categóricos se desprenden de la noción clásica de **matriz indicadora** de una o varias variables cualitativas (Van Der Heidjen [1987]).

Considerando sobre un conjunto I de n individuos una variable categórica J con p modalidades, la matriz indicadora Z^{IJ} de tamaño n x p se conforma de la yuxtaposición de las p variables indicadoras de cada modalidad. Cuando se tiene un conjunto Q de variables categóricas (cada variables con J_q modalidades), dado que cada matriz tiene n filas, se pueden encadenar una tras otra en forma horizontal para obtener una super matriz Z^{IQ} de orden : n x ($\sum_{q=1}^{Card(Q)} Jq$).

Por ejemplo si en un conjunto de cinco individuos tenemos dos variables respectivamente con tres modalidades (a, b, c) y dos modalidades (r, s), y si las modalisades son exclusivas, la matriz indicadora seria de tipo :

	a	b	c	r	s
	1	0	0	1	0
objetos	0	1	0	1	0
	1	0	0	0	1
	0	0	1	0	1
	1	0	0	1	0

En el caso de los datos longitudinales lo que se tiene es un cubo de información con tres entradas : en las filas encontramos los n individuos de I, en las columnas una serie de bloques de J_q modalidades para cada variable de Q, y en la tercera dimensión cada periodo t de estabilidad del proceso (t = 1, ..τ) durante la duración de observación T. En este caso la matriz indicadora Z^{IQT} es de orden $n\times(\sum_q J_q)\times\tau$, y sus elementos se denotan z^q_{ijt}.

z^q_{ijt} = 1 cuando el individuo i se halla en el estado j de la variable q durante el periodo t, de lo contrario z^q_{ijt} = 0. En esta matriz puede variar sin restricciones el número de individuos, el número de variables, el número de modalidades de cada variable y el número de periodos de tiempo.

La matriz indicadora Z^{IQT}, llamada mas simplemente Z, no puede ser analizada directamente bajo análisis factoriales, luego esta debe ser transformada en una matriz de dos entradas, para este efecto existen diferentes posibilidades.

- *Cortes transversales sobre la matriz Z.*

Una forma de obtener una matriz de dos entradas es, si se consiente una perdida de información, tomar separadamente cortes transversales de la matriz Z. Los cortes transversales para el bloque de la variable q se denotan $Z^{i(q)}$, $Z^{t(q)}$, $Z^{j(q)}$, respectivamente para un corte de orden TxJ_q correspondiente al individuo i, un corte de orden IxJ_q correspondiente al periodo t, y un corte de orden IxT. correspondiente a la categoría j de la variable q. Evidentemente el análisis de estos cortes carece generalmente de interés porque la información se reduce demasiado, sin embargo algunos de ellos pueden aportar resultados de interés para ciertas problemáticas de análisis particulares.

- *Yuxtaposición de cortes transversales*

Una manera mas interesante de obtener una matriz de dos entradas sin perder información es yuxtaponiendo horizontal o verticalmente los cortes transversales, dos yuxtaposiciones son de uso frecuente.

La primera es cuando los $Z^{t(q)}$ son yuxtapuestos horizontalmente. La matriz se nota Z^{ijt} y tiene orden $n\times[(\sum_q J_q)\times\tau]$. Esta estructura corresponde a un registro por individuo y, como lo veremos, sirve para la ejecución práctica del análisis armónico cualitativo.

La segunda yuxtaposición también se hace sobre los $Z^{t(q)}$, ahora cada t es yuxtapuesto verticalmente, cada objeto i es representado por T filas. Esta yuxtaposición se nota Z^{itj} y es de orden $(n\times\tau)\times\sum_q J_q$ Esta estructura corresponde a un registro por combinación de individuos con periodos y es el punto de partida de los métodos llamados de “análisis conjunto”.

- *Tablas de contingencia marginales.*

Esta forma de transformar la super matriz de tres entradas Z en una matriz de dos entradas se hace mediante sumas, generalmente sobre la dimención del tiempo o sobre las filas (individuos), para obtener tablas de contigencia que tengan un sentido interpretativo :
- Tablas de contingencia individuos por estados : sumando sobre los tiempos se obtiene la matriz Z^{IQ} cuyos elementos z^{q}_{ij+} corresponden a la frequencia con la cual el objeto i se encuentra en el estado j de la variable q. El sentido que tiene esta frecuencia depende de la manera de discretizar el tiempo continuo de la observación.
- Tablas de contingencia estados por periodos : si, como en el caso precedente, suponemos una discretización del tiempo, también podemos sumar sobre los objetos, esta matriz se notara Z^{QT} cuyos elementos z^{q}_{+jt} representan el número de individuos que se encuentran en la categoría j de la variable q durante el periodo t.
Estos tipos de tablas de contingencia se pueden analizar con el análisis factorial de correspondencia, simple o múltiple según que hay una o varias variables categóricas (ver Barbary O. [1994], pgs 45-93).

- *Matrices de transición.*

En el caso más sencillo cuando T tiene dos "niveles" (que corresponden a dos intervalos en que se decida particionar el tiempo, ó dos fechas que se desea analizar), la variable q con J categorías y un número grande de objetos i, se genera la matriz llamada de transición F a partir de las matrices indicadoras Z^1 y Z^2 : $F=Z^{1'}Z^2$. La matriz F tiene orden JxJ y las frecuencias f_{ij} representan el número de objetos que pasan del estado i en el tiempo 1, al estado j en el tiempo 2, la diagonal f_{ii} indica el número de objetos que no cambian de estado.
Por ejemplo dada una variable con tres categorías a, b, c, dos fechas T_1, T_2 y cuatro individuos, una matriz de transición puede ser :

			T2			
		A	B	C		
	A	2			2	f_{at1}
T1	B		1		1	$f_{.bt1}$
	C	1			1	$f_{.ct1}$
		3	1	0		
		$f_{.at2}$	$f_{.bt2}$	$f_{.ct2}$		

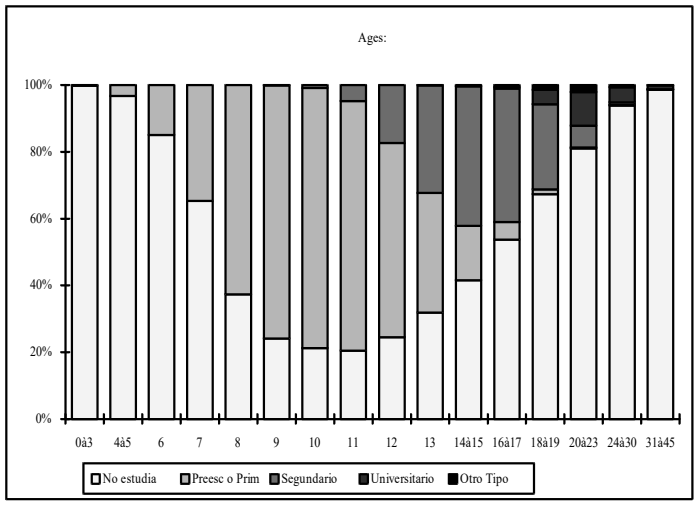
Las matrices de transición pueden ser analizada mediante análisis factoriales o modelos log lineales, pero tienen la desventaja de perder vínculo con los objetos en estudio y por esta razón no se puede generar tipologias de los objetos.

1.3 Metodologias de análisis

Representación gráfica.

Cuando el número de estados es pequeño la matriz $\underline{Z}$ es mostrada algunas veces gráficamente por medio de barras. Estas gráficas se presentan de acuerdo al objetivo del estudio y se trabaja con las marginales Z_{+jt}, Z_{i+t} o Z_{ij+}.

Representación del nivel de educación cursado



Estos gráficos son muy útiles cuando se manejan conglomerados de individuos que tienen comportamiento homogéneo y donde la agregación de la marginal puede ser interpretada como el promedio de una clase o tipo de individuos. En el gráfico se presenta por ejemplo el perfil promedio de la trayectoria educacional sobre la muestra biográfica de la encuesta CEDE/ORSTOM en Bogota.

Modelización de procesos estocásticos.

En oposición al enfoque puramente descriptivo de las representaciones gráficas, existe una linea metodológica importante para la modelización de los datos longitudinales. Los modelos probabilísticos paramétricos o semi paramétricos desarrollados en un principio para el análisis de los datos de sobrevivencia (Cox [1972]) tienen como propósito formalizar matemáticamente las biografías individuales como procesos aleatorios. La biografía de los individuos esta representada por una serie de variables aleatorias que corresponden al tiempo de permanencia en los diferentes estados, tomados en forma cronológica. El objetivo consiste en estimar a partir de los datos de la encuesta, un modelo de distribución de esta serie de variables aleatorias (distribución conjunta). Este enfoque permite un análisis de las interacciones entre fenómenos y de los efectos de las características individuales o de los eventos exteriores (contexto) sobre el tiempo de permanencia en los diferentes estados pero no proporciona métodos de análisis tipológicos de los datos individuales sino modelos explicativos de la movilidad individual.

1.4. Conclusión

El análisis de las biografías, como datos longitudinales particulares, es un campo de investigación relativamente nuevo. En su extensión actual, se puede distinguir dos enfoques teóricos : el enfoque descriptivo (o de análisis de datos) y el enfoque inferencial (o probabilístico). Según muchos criterios, son más complementarios que concurrentes. No obstante, siendo dirigidos hacia objetivos y técnicas radicalmente distintos y practicados por "escuelas" académicas diferentes, estos dos enfoques provocan un debate científico. No es aquí el lugar para ello; solamente nos parece natural dar la **prioridad al enfoque descriptivo**. Como lo dice M. Volle [1981] :

> *"(...) el análisis de datos se ubica en una etapa del razonamiento lógicamente anterior a la inducción (inferencia) probabilística (...). Para el que practica el análisis de datos, es solo después de haber identificado las estructuras que sustenten los datos, después de haberlos clasificados y recortados, que es posible abordarlos con un enfoque probabilístico."* [1]

En la primera parte, hemos visto como el **análisis de correspondencia** se puede aplicar a varias estructuras matriciales que sintetizan la información biográfica proporcionada por las encuestas retrospectivas; el libro de Van Der Heidjen [1987] presenta un panorama bastante completo del tema. Sin embargo ninguna de las soluciones metodológicas abordas hasta ahora se pueden considerar como una herramienta completa para el análisis tipológico de las biografías. Un método recientemente desarrollado en Francia presenta precisamente estas características, se trata del **análisis armónico cualitativo**, una aplicación particular del análisis de las correspondencias múltiples (Deville J.C, Saporta G. [1980] y Deville J.C. [1982]). Es este método, lo cual parece de los más prometedores en el campo del análisis tipológico de los itinerarios biográficos, que vamos a presentar en la segunda parte.

[1] : M. VOLLE, Analyse des données, p 21, Economica, Paris, 1981.

SEGUNDA PARTE : El análisis armónico cualitativo

En esta segunda parte no pretendemos más que dejar planteados los grandes rasgos de la definición y justificación matemática del análisis factorial de un proceso cualitativo. La teoría completa con las demostraciones rigurosas se encuentra en Deville [1982].

2.1 Definiciones y principio

El marco probabilístico en el cual se puede formalizar la observación longitudinal de los eventos biográficos sobre una población es el siguiente.

-1: T=[0,T] es un intervalo de tiempo y (T, β_t, μ) el espacio de tiempo medible con :

β_t , conjunto de los borelianos de T

μ, medida de Lebesgue sobre (T, β_t)

-2: χ es un conjunto finito de estados con m elementos que permite definir una tribu de eventos A.

-3: (Ω, A, P) es el espacio de probabilidad asociado a la población Ω y los eventos A

-4: X_t el proceso cualitativo observado durante el tiempo T, es decir una variable aleatoria cualitativa sobre el producto cartesiano ΩxT con valores en χ

$$X_t : \Omega xT \rightarrow \chi$$
$$(\omega,t) \rightarrow X_t(\omega) = x$$

Si 1_t^x es la función indicadora del evento $X_t = x$ y notamos E(.) la esperanza matemática bajo P, tenemos :

$$P(X_t = x) = E(1_t^x),$$
$$P(X_t = x \cap X_s = y) = E(1_t^x \, 1_s^x)$$

La idea directriz del método de análisis armónico cualitativa es seguir los pasos de la construcción de cualquier análisis factorial en el marco del algebra de operadores : se busca primero una **codificación real** del proceso X_t con el fin de obtener un vector aleatorio con valores reales y luego se hace la **descomposición espectral del operador** de proyección asociado a dicho vector (descomposición llamada también análisis armónico). Una codificación real del proceso X_t, es una función f_t tal que :

$$f_t : \chi \times T \rightarrow \Re.$$
$$(x,t) \rightarrow y = f_t(x)$$

Obviamente dentro del espacio de todos las codificaciones reales de X_t posibles, unas son mas naturales y prácticas que otras.

2.2 Análisis espectral de un proceso cualitativo

Sea $L^2(\Omega,A,P)$ el conjunto de las variables aleatorias numéricas sobre Ω de cuadrado integrable. El subespacio $L^2(X_t)$ engendrado por las codificaciones reales $f_t(X_t)$ del proceso cualitativo X_t, se compone de variables numéricas de la forma :

$$\zeta_t = \sum_{x\in\chi} a^x 1_t^x, \quad (a^x \in \Re) \qquad (1)$$

Se nota E^t el operador de esperanza condicional a X_t, es decir la proyección ortogonal de $L^2(\Omega,A,P)$ sobre el subespacio $L^2(X_t)$. E^t transforma una variable positiva en otra positiva e inversamente, todo operador que cumpla esta propiedad es un operador de esperanza condicional. En otros términos cada variable X_t es estadísticamente equivalente al operador E^t de esperanza condicional. La consecuencia lógica de este principio consiste en afirmar que las dependencias estadísticas entre dos instante t y s del proceso son resumidas por el producto
$E^tE^s = K(t,s)$ de los operadores asociados a las v.a. X_t y X_S. El producto K es un operador sobre H, el espacio de Hilbert de los procesos de segundo orden de cuadrado integrable, y como tal admite una descomposición espectral en m vectores propios $.\zeta^i$ (i=1,m) llamados "procesos propios" :

$$K = \sum_{i=1}^{m} \lambda_i \zeta^i \otimes \zeta^i$$

donde el producto entre procesos notado $\zeta\otimes\eta$, transforma el proceso ζ_t en el proceso :

$$\zeta_t \int_T E(\zeta_s \eta_s)\, ds.$$

Los ζ^i, asociados a los valores propios positivos λ_i, forman un conjunto orto normal en H y K satisface entonces la ecuación llamada de valores propios :

$$\lambda_i \zeta_t^i = \int_T K(t,s)\zeta_s^i ds \qquad \forall i \in \{1, n\}$$

(2)

Siendo los procesos ζ_t medibles, es legitimo hacer intervenir en (2) las integrales y esperanzas condicionales y la ecuación puede ser escrita así :

$$\lambda_i \zeta_t^i = E^t\left(\int_T E^s\left(\zeta_s^i\right)ds\right)$$

luego, dado que $E^S\zeta_S = \zeta_S$, tenemos :

$$\lambda_i \zeta_t^i = E^t\left(\int_T \zeta_s^i ds\right)$$

Ahora se define la variable aleatoria z_i, llamada generatriz del proceso ζ_i, por :

$$z_i = \int_T \zeta_s^i ds$$

entonces utilizando la nueva variable y integrando la ecuación (2) se obtiene :

$$\lambda_i z_i = \int_T E^t(z_i)dt \qquad (3)$$

donde queda claro que las variables z_i no dependen del tiempo.

Una ultima forma de la ecuación de valores propios se obtiene si volvemos a la expresión (1) para ζ_t y recordamos que $E^s\zeta_s = \zeta_s$. Luego si olvidamos el índice i para simplificar la escritura, (2) se transforma en :

$$\sum_x \lambda a_t^x 1_t^x = \sum_y \int_T a_s^y E^t\left(1_s^y\right)ds$$

Ahora $E^t(1_s^y) = \sum_x 1_t^x E\left(1_t^x 1_s^y\right) / E\left(1_t^x\right) = \sum_x 1_t^x P_{x,y}(t,s) / P_x(t)$, donde $P_{x,y}(t,s) = P[X_t = x, X_s = y]$, y $P_x(t) = P[X_t = x]$, y (2) se convierte en :

$$\lambda a_t^x = \sum_y \int_T a_s^y \left[P_{x,y}(t,s) / P_x(t)\right]ds \qquad (4)$$

En resumen, las ecuaciones (3) y (4) muestran que los z_i y los a_t^x forman una descomposición espectral del proceso en una serie de variables aleatorias independientes del tiempo (z_i) y de códigos reales (a_t^x) que dependen del tiempo.

2.3 Equivalencia con el análisis de correspondencia.

Consideramos ahora los p+1 instantes $t_o=0 < t_1 ... < t_{p-1} < t_p = T$ tales que todas las trayectorias del proceso son constantes sobre los p intervalos $[t_{j-1}, t_j[$.

Notemos ζ_j el valor de ζ_t sobre $[t_{j-1}, t_j[$, E^j la esperanza condicional de ζ_j y L_j la longitud del intervalo j, $L_j = t_j - t_{j-1}$.

Así, las anteriores ecuaciones se convierten en :

$$\lambda \zeta_k = \sum_{j=1}^{p} L_j E^k E^j \zeta_j \qquad (2')$$

$$z = \sum_{j=1}^{p} L_j \zeta_j \qquad (3')$$

Y los códigos asociados con las variables ζ_k son funciones constantes sobre cada uno de los intervalos de la partición, luego se puede escribir la ecuación (4) así :

$$\lambda a_k^x = \sum_y \sum_{j=1}^{p} L_j a_j^y \, [P_{x,y}(k,j) / P_x(k)] \qquad (4')$$

En el caso en que Ω es finito, compuesto por n individuos, la ecuación (4') equivale a la ecuación de base del análisis canónico o del análisis de correspondencia (ver Deville J.C. [1982], pgs 68-74). El código real a_k^x es el vector proprio asociado al valor propio λ en el análisis de correspondencia de una tabla disyuntiva particular a n filas y mp columnas (ver sección siguiente). Así el análisis armónico cualitativo viene a ser una generalización de métodos conocidos de análisis de datos (análisis factorial canónico en general o más específicamente el análisis de correspondencia) y desde el punto de vista numérico son equivalentes a estos. Todas las técnicas de interpretación que habitualmente se dan dentro de estos métodos son válidos para este caso.

2.4 Aproximación del análisis Armónico Cualitativo por métodos numéricos

Hasta el momento, hemos considerado el intervalo de tiempo [0,T] como continuo o, en el último párrafo, discretizado según los intervalos de estabilidad del proceso; es decir que conservamos toda la información conocida. Y por otro lado tenemos una solución teórica para hacer el análisis factorial del proceso. Veamos ahora cual es la aproximación práctica de esta solución.

2.4.1 El A.A.C. como análisis de correspondencia de un cuadro disyuntivo particular.

Como en el párrafo 2.3, tenemos los p+1 instantes $t_o=0 <t_1...<t_{p-1}<t_p=T$; los t_i limitan los intervalos durante los cuales ningún individuo cambia de estado.
Retomando la ecuación (4) : $\lambda z = \int_T \zeta_s ds$, la idea es aproximar la integral por una suma de Rieman. Con las mismas notaciones que en 2.3, se puede ver la suma :

$Q = \sum_{j=1}^{p} L_j E^{\theta j}$, $\theta_j \in [\, t_j - 1, t_j \,]$ como un nuevo operador mucho más simple que K.

La descomposición espectral de Q proporcionará una aproximación de los valores propios del análisis armónico cualitativo del proceso X_t. Esta descomposición no es otra que el análisis de correspondencia múltiple de las variables cualitativas $X_{\theta j}$ ponderadas por el peso L_j.

Dicho de otra forma, el análisis armónico en tiempo continuo se convierte, en tiempo discreto, en la descomposición en factores resultante **del análisis de correspondencias múltiples de un cuadro disyuntivo particular** con n líneas y mp columnas (recordamos que n es el número de individuos, m el número de estados posibles y p el número de períodos en los que el proceso se mantiene estable). La casilla elemental del cuadro vale uno si el individuo está en el estado considerado durante el período y cero si no lo está. Esta tabla disyuntiva completa puede ser sometida al análisis de correspondencia, pero inmediatamente surge un problema : en los casos concretos de aplicación**, los números n, m y p generan un cuadro de tamaño asombroso, lleno de ceros**, y su análisis no proporciona resultados interesantes. La solución práctica consiste en dividir el intervalo de observación en un número razonable de períodos (de duración constante o no**) sin tener en cuenta los cambios de estado individuales**. Se construye luego el cuadro calculando la proporción de tiempo que ha permanecido cada individuo en cada uno de los estados lo largo de cada período (**densidad individual de presencia en los estados**). En seguida se aplica el análisis de correspondencias y las técnicas de interpretación habituales. Aparecen entonces dos posibilidades para el cálculo de las frecuencias.

2.4.2. Métodos de recodificación de la información

Consideramos ahora el intervalo [0,T] dividido en p intervalos :
$[t_o,t_1[\ [t_k,t_{k+1}[\ [t_{p-1},t_p]$, de longitud igual o no. Los t_K son escogido sin tomar en cuenta los cambios de estados individuales, según el conocimiento previo del proceso (argumentos de orden demográfico, sociológico, etc.) o con base en la distribución de frecuencia de los cambios de estado en el tiempo (argumentos estadísticos). Además, abandonando el marco probabilístico definido en la parte teórica, podemos notar I el conjunto finito de los individuos (card(I) = n) y J el conjunto finito de los estados del proceso (card(J) = m).

La codificación real del proceso cualitativo es una función real sobre el producto cartesiano IxJx[0,T] :

$$Y : IxJx[0,T] \rightarrow R$$

Sea τ_{ijk} (i=1,n; j=1,m; k=1,p) el tiempo pasado por el individuo i en el estado j durante el intervalo de tiempo $[t_{k-1},t_k[$, podemos definir el valor de Y de dos manera :

- asignándole la proporción del tiempo total de observación del Proceso (T) que el individuo i a pasado en el estado j durante el intervalo $[t_{k-1},t_k[$, es decir :

$$\mathbf{Y(i,j,k) = \tau_{ijk} / T};$$

en este caso, cualquier sea los cortes del intervalo [0,T], la métrica sobre el tiempo es uniforme y la suma de cada fila de la tabla de frecuencia (dimensión individuos) vale 1.

- asignándole la proporción de la duración del intervalo $[t_{k-1},t_k[$ que el individuo i a pasado en el estado j durante el intervalo $[t_{k-1},t_k[$, es decir :

$$\mathbf{Y(i,j,k) = \tau_{ijk} / (t_k - t_{k-1})};$$

en este caso, si los intervalos $[t_{k-1},t_k[$ no son de longitud constante, la métrica sobre el tiempo no es uniforme y la suma de cada fila de la tabla de frecuencia vale p, el número de periodos definidos en [0,T].

Cabe primero resaltar que bajo esta recodificación de los datos originales, se conserva la integridad de la información relativa a la duraciones individuales en los estados, pero se pierde el orden cronológico de los cambios de un estado al otro cuando ocurren en un mismo intervalo de codificación.

Para el cálculo algebraico efectuado en el análisis de correspondencia, de nada es obligatorio que los distintos periodos de codificación sean de duración constante o la métrica sobre el tiempo uniforme. Al contrario son varios los argumentos a favor de la segunda alternativa. Por una parte, del punto de vista estadístico, tenemos interés en detallar la codificación durante los periodos en donde ocurren muchos cambios de estados, fijándose en el histógrama de la distribución temporal de los eventos (ver J.C. Deville [1982] y A. Florette [1988]). Por otra parte estos argumentos estadísticos coinciden con frecuencia con preocupaciones de orden temáticas : parece natural por ejemplo, en una problemática de análisis de las trayectorias residenciales de los individuos, hacer más énfasis en los cambios durante la edad adulta, cuando las decisiones pertenecen más generalmente al individuo que durante edades anteriores o posteriores. Así en la práctica se adoptará con frecuencia la segunda solución que vamos a ilustrar ahora con un ejemplo artificial.

Consideremos datos longitudinales observados sobre una duración de 10 horas y una partición de T en cuatro periodos : la duración de p_1 es de 3 horas, p_2 de 2 horas, p_3 y p_4 de 4 horas. Sea I el conjunto de los individuos (card(I) = 4) y J un conjunto de estados con tres modalidades a, b, c. Notaremos Π, de orden n x (m.p) es decir 4x12, la matriz de los datos codificados. De acuerdo con la solución escogida Π_{ijk} contiene la proporción de la duración del intervalo k que el individuo i a pasado en el estado j.

Por ejemplo, el primer individuo en el periodo p_1 ha pasado 2 horas en el estado a, luego una hora en el estado b y no ha conocido el estado c; para este individuo :
$\Pi_{1a1}=2/3$, $\Pi_{1b1}=1/3$ y $\Pi_{1c1}=0$; de donde $\sum_{j\in J}\Pi_{1j1}=1$ y $\sum_{j\in J,p\in\{1,4\}}\Pi_{1jp}=4$.
La matriz completa tendrá la siguiente forma :

Periodo	1			2			3			4			
Estado	a	b	c	a	b	c	a	b	c	a	b	c	
Individuos													
ind 1	2/3	1/3	0	0	1	0	1/2	1/2	0	1	0	0	4
ind 2	0	1	0	1/2	1/2	0	1/2	0	1/2	0	0	1	4
ind 3	1/2	0	1/2	1	0	0	0	1	0	1/2	1/2	0	4
ind 4	1	0	0	0	0	1	0	0	1	0	0	1	4
	13/6	4/3	1/2	3/2	3/2	1	1	3/2	3/2	3/2	1/2	2	16

BIBLIOGRAFÍA:

BARBARY O. [1994] : "El análisis estadístico de datos biográficos en ciencias sociales, curso introductivo al uso de nuevas metodologias de análisis multivariado en el estudio de las trayectorias individuales", dictado en el Simposio de Estadística de la Universidad Nacional de Colombia, 6-10 de junio 1994, Santafé de Bogotá 116 p.

CAILLIEZ F., PAGES J.P., [1976] Introduction à l'analyse des données, Smash, Paris.

COURGEAU D., LELIEVRE E. [1989] Analyse démographique des biographies, éditions de l'INED, Paris, 268 p.

COX D.B. [1972] : Regression models and life tables (with discussion), "Journal of Royal Statistical Society", B 34, pp 187-220.

DEVILLE J.C. & SAPORTA G. [1980] : L'Analyse harmonique qualitative, in E. DIDAY (ed.) : "Data analysis and informatics", North Holand, Amsterdam, pg 375-389.

DEVILLE J.C. [1982] : Analyse des données chronologiques qualitatives, comment analyser les calendriers ? "Annales de l'INSEE", n° 45, pp 45-104.

DUREAU F., FLOREZ C.E., HOYOS M.C. [1993] : La movilidad de las poblaciones y su impacto sobre la dinámica del área metropolitana de Bogotá, análisis de los datos existentes. Documento de trabajo No 1, CEDE/ORSTOM, Bogotá, 286 p.

DUREAU F., FLOREZ C.E., BARBARY O., GARCIA L., HOYOS M.C. [1994 -a] : La movilidad de las poblaciones y su impacto sobre la dinámica del área metropolitana de Bogotá, metodología de la encuesta cuantitativa. Documento de trabajo No 2, CEDE/ORSTOM, Bogotá, vol. 1 : 98 p. multigr., vol. 2 : annexes, 295 p.

DUREAU F., BARBARY O., FLOREZ C.E. [1994 -b] : La movilidad de las poblaciones y su impacto sobre la dinámica del área metropolitana de Bogotá, resultados preliminares de la encuesta cuantitativa. Documento de trabajo No 3, CEDE/ORSTOM, Bogotá, 309 p.

FINE J. [1994] Producción y tratamientos de datos de investigación en ciencias humanas, programa PRESTA, Fascículo 4. pg 7-12.

FLORETTE A. [1980] : Approximation et choix du découpage dans le cadre de l'analyse harmonique qualitative, Mémoire de DEA, ENSAE, Paris.

FLOREZ C.E. [1990] La transición demográfica en Colombia, efectos en la formación de la familia, ediciones Uniandes - Universidad de las Naciones Unidas, Bogotá, 242 p.

VAN DER HEIJDEN P. G. M. [1987] : Correspondence analysis of longitudinal categorical data, DSWO PRESS, Leiden.

VERNET E. [1983] L'analyse harmonique qualitative appliquée aux données de déplacements, rapport de stage ENSIMAG à l'INRETS, 120 p.

VOLLE M. [1981] Analyse des données, Economica, Paris.