ANÁLISIS MULTIDIMENSIONAL DE POBLACIÓN CON DATOS INCOMPLETOS*

FRANS WILLEKENS**

1. INTRODUCCIÓN

EN LA DEMOGRAFÍA multidimensional se estudian simultáneamente diversas *dimensiones* de un sistema demográfico. En la multirregional, las dimensiones son la edad y la región de residencia. Los sistemas demográficos también se podrían definir a lo largo de las dimensiones de edad y status marital, edad y status ocupacional, etc. Para cada dimensión se pueden distinguir varios *estados* o categorías demográficas, por ejemplo, grupos quinquenales de edad; estados civiles, como: soltero, casado, viudo, divorciado, etc. También se puede considerar que los estados de vida y muerte constituyen una dimensión. Sin embargo, en este artículo no, porque las muertes están calculadas como residuos, es decir, como personas que no se mueven a ningún estado del sistema.

La aplicación de la demografía multidimensional puede ser limitada, debido a la disponibilidad de los datos. La demanda de muchos datos es una gran desventaja del análisis multidimensional. Las oportunidades de las técnicas avanzadas de la investigación demográfica sólo se pueden explorar completamente si se tienen suficientes datos disponibles o si se pueden usar procedimientos adecuados para estimar los datos faltantes. Recientemente, oficinas estadísticas en todo el mundo desarrollado han empezado la recolección y tabulación, en una forma regular, de los datos necesarios para el análisis multidimensional de población, o están considerando hacerlo. También en las oficinas estadísticas han surgido discusiones sobre si es suficiente recolectar información sobre tamaños y composiciones

* Artículo presentado en la Conferencia de Demografía Matemática Multidimensional, "College Park, Maryland, U.S.A.", 23 a 25 de marzo de 1981. Forma parte de un proyecto de investigación para el desarrollo de una metodología para obtener patrones de migración detallados, a partir de datos incompletos; el proyecto se lleva a cabo con la ayuda financiera de la Organización Holandesa para el Progreso de la Investigación Pura (Z.W.C.). El artículo también está incluido en las actas de la Conferencia que serán editadas por K. Land y A. Rogers y publicadas en "Estudios de Población" –serie de Academic Press de Nueva York.

** Desearía agradecer enormemente a Willem Schaafsma y Paul de Jong por sus extensos comentarios a mi borrador preliminar y a Jaine Koendering por su contribución para transformar el manuscrito en una versión diestramente mecanografiada.

de subpoblaciones mayores, o si es necesario cuantificar los flujos entre categorías relevantes de las subpoblaciones. En el primer capítulo de este artículo, revisamos algunos desarrollos positivos en la recolección de datos. Aunque los pasos que se han seguido para la solución del problema de datos son favorables, todavía hay un largo camino por recorrer. Incluso, los datos recolectados pueden no ser adecuados para el análisis multidimensional, debido a deficiencias en el tipo y tamaño de muestra. Por lo tanto, frecuentemente se tienen que complementar los datos disponibles con estimadores.

El objetivo del presente artículo es estimar los datos de flujos necesarios para el análisis multidimensional de población. La estimación de flujos o transiciones es un problema relativamente nuevo en demografía. Sin embargo, ha sido investigado por muchos años en la ciencia regional (migración), ciencia del transporte (tráfico entre nudos en una red de transporte) y economía (transacciones entre industrias en una matriz de insumo-producto de entradas y salidas). Los estadísticos han puesto su atención a este problema en la teoría de información, en las estadísticas bayesianas y en el análisis de tablas de contingencia. Continuaremos con esta investigación para proponer una estrategia para estimar datos faltantes en el análisis multidimensional de población. El aspecto principal de esta estrategia es la derivación del valor de cada elemento faltante en la matriz de flujo del conjunto total de datos. Esto significa que los elementos individuales no están considerados por separado, sino sólo en conexión con los otros elementos, es decir, como componentes de una gran estructura. El arreglo matricial proporciona un esquema para esta representación estructural.[1] La estrategia propuesta consta de cinco pasos:

a) Establecer la forma del arreglo. Determinar las dimensiones (clasificaciones) del sistema de población y los estados (categorías) de cada dimensión. Los arreglos estudiados en este artículo son clasificaciones cruzadas multidimensionales. Tan pronto como se ha establecido la forma del arreglo, la localización de las celdas dice algo acerca de las características de los individuos que caen en él. Por ejemplo, en un arreglo tridimensionado, los individuos de una celda específica tienen una característica en común con los individuos de todas las celdas en el mismo renglón, otra característica en común con todos los individuos de las celdas en la misma columna y aún otra con todos los individuos en las celdas del mismo estrato.

b) Desarrollar un modelo del conjunto de datos en el arreglo. Para poder considerar elementos individuales en conexión con otros elementos, las relaciones estructurales en los datos están representadas por un modelo paramétrico. Se pueden imaginar diversos modelos de estructura de datos, pero nosotros nos limitaremos al modelo log-lineal. Se puede demostrar que los modelos desarrollados en las disciplinas antes mencionadas son formulaciones particulares del modelo log-lineal, como se concibió en la investigación de patrones de asociación en tablas de contingencia.

c) Introducir al arreglo los datos disponibles, es decir, llenar el arreglo lo máximo posible. En general, no conocemos las celdas individuales, sino solamen-

1 Una forma diferente, pero relacionada se obtiene con la postulación de un modelo estocástico apropiado. Este enfoque es ilustrado en la sección 3.2 y está completamente desarrollado en De Jong (1981).

te los totales marginales. Si algunas de las celdas son observadas, incluirlas en el arreglo. Otra información previa es listada separadamente.

d) Determinar los valores de parámetros del modelo paramétrico en base a los diferentes tipos de información previa, complementados con hipótesis sobre ciertas relaciones estructurales en los datos por ser estimados. Una hipótesis usada frecuentemente, cuando sólo se tienen datos agregados, es el supuesto de independencia entre algunas de las variables. Los modelos de ajuste y prueba para independencia son equivalentes al modelo adoptado en este artículo. Esta equivalencia permite un enfoque notablemente simple y transparente para estimar los elementos faltantes en el arreglo.

e) Aplicar el modelo para inferir los valores de los elementos faltantes.

Al implementar este procedimiento de cinco pasos, pueden surgir problemas metodológicos y prácticos. Algunos de los potenciales problemas son tratados en este artículo. La sección 2 enfoca el requerimiento y disponibilidad de datos. Se da una revisión ilustrativa de los tipos de datos de flujo disponibles y de las fuentes de donde se obtienen datos relevantes para el análisis multidimensional. Esta revisión es muy incompleta, pero demuestra que es útil para buscar fuentes adicionales, y tal vez no comunes, de datos para poder llenar el arreglo lo máximo posible.

La sección 3 discute el modelo log-lineal. Están representados dos modelos de formulaciones: el modelo aditivo es popular en los análisis de interacción: sin embargo, el modelo multiplicativo puede ser más apropiado para tratar con problemas de estimación. Esta sección también revela qué información es suficiente para estimar los parámetros del modelo log-lineal. Estas condiciones de suficiencia llevan directamente al tema principal de este artículo; es decir, el cálculo de los valores esperados, a partir de cualesquier datos disponibles.

En la sección 4 se consideran diferentes situaciones hipotéticas de disponibilidad de datos:

i totales marginales solamente

ii totales marginales, complementados con estimadores preliminares de las celdas.

iii totales marginales y unas pocas celdas dadas (es decir, conocidas exactamente).

Éstas y otras condiciones relacionadas de conocimientos previos se pueden tratar de una manera unificada y se puede aplicar un procedimiento simple de ajuste multiproporcional para estimar los elementos faltantes.

Esta técnica general de estimación está basada en el análisis log-lineal de los conjuntos de datos e incluye diversos enfoques encontrados en la literatura, para inferir, de los datos disponibles, las entradas de una tabla n-dimensionada. Modelos gravitacionales, maximización de entropía, minimización de ganancia de información y otros métodos populares de estimación son equivalentes al ajuste multiproporcional. Esta equivalencia indica algunas directrices interesantes de la investigación futura para el mejoramiento de las técnicas de estimación.

Las técnicas presentadas en este artículo, así como algunos procedimientos equivalentes, son aplicadas a datos reales en la sección 5. Se escogen dos campos de aplicación: la estimación de tablas de movilidad social y de tablas de migración.

2. Requerimientos y disponibilidad de datos

El análisis demográfico multidimensional requiere datos brutos sobre flujos. El análisis demográfico convencional de poblaciones, descompuestas en subpoblaciones, enfoca el tamaño y cambios del tamaño de cada subpoblación. Esta perspectiva de stock hace uso de las tasas más importantes que indican en qué medida prevalece cada característica particular en la población (por ejemplo, tasa de participación en la actividad ocupacional, proporción de casados, etc.). En el análisis multidimensional el tamaño de la población es de importancia secundaria. El énfasis se pone en los flujos, es decir, en los pasajes a través de diversos estados; la magnitud de cada subpoblación es sólo el resultado de una condición inicial y un mecanismo de flujos. La razón para adoptar la perspectiva de flujos es que la dinámica que subyace al cambio poblacional puede ser representada más fácilmente. Se consideran explícitamente las entradas y salidas de cada subpoblación. Para adoptar la perspectiva de flujos, es necesario que existan datos sobre éstos, y que puedan acomodarse en un arreglo matricial. Este arreglo no es solamente un esquema de representación de datos convenientes, sino también una base útil para la integración de procedimientos de estimación de datos. En la primera sección, se presenta el arreglo para la representación y estimación multidimensional de datos. La segunda sección de esta parte nos da una revisión ilustrativa de fuentes de datos para el análisis multidimensional de población. También demuestra que el problema de datos no es tan grave. Aún no se han tocado algunas fuentes de datos con la información correcta.

2.1 Requerimiento de datos: un esquema del arreglo

El análisis multidimensional requiere flujos específicos por edad entre diversos estados considerados en el análisis. Los datos se pueden acomodar para constituir una tabla multidimensional de contingencia. Esta tabla representa el arreglo en el que se puede acomodar la información disponible y puede definirse el problema de estimación. En este artículo nos limitaremos a un sistema de población bidimensionado. Una de las dimensiones será la edad. Los descubrimientos se pueden generalizar a cualquier número de dimensiones o, alternativamente, se puede reducir un sistema de población de una dimensión mayor a dos dimensiones, incrementando el número de estados o categorías en una o más dimensiones (por ejemplo, cada uno de los cuatro estados civiles se pueden descomponer por sexo, dando así ocho estados a su dimensión).

Para cada grupo de edad, se tiene que conocer el número de pasajes de un estado a cualquiera de los otros. Esta información puede ser arreglada en estratos de una tabla de doble entrada (cuadro 1). Sea que k denote la edad (estrato); i, el estado de origen (renglón) y j, el estado de destino (columna). Hay L grupos de edad (estratos), K orígenes (renglones) y C destinos (columnas) (R = C). El número total de celdas en el arreglo es, por lo tanto, L x R x C. Sea que K, I y J representen los conjuntos de índices de k, i y j, respectivamente:

$$K = \{1, 2, \ldots, k, \ldots, L\}$$

$$I = \{1, 2, \ldots, i, \ldots, R\}$$

CUADRO 1

ARREGLO DE PASAJES PARA EL GRUPO K DE EDAD

Status de origen	Status de destino				
		1	2	3	Total
	1	m_{11k}	m_{12k}	m_{13k}	$m_{1.k}$
	2	m_{21k}	m_{22k}	m_{23k}	$m_{2.k}$
	3	m_{31k}	m_{32k}	m_{33k}	$m_{3.k}$
	Total	$m_{.1k}$	$m_{.2k}$	$m_{.3k}$	$m_{..k}$

$$J = \{1, 2, \ldots, j, \ldots, C\}$$

Algunos pasajes o agregados de pasajes pueden ser fijos, porque son realmente conocidos u observados, mientras que otros deben ser estimados. Sea S el conjunto de celdas que se deben estimar en un arreglo de entradas múltiples. Formalmente:
$S = \{(i, j, k)$ el pasaje de i a j por la categoría k es posible y no fijo$\}$ Si la celda (i, j, k) está en S, escribimos $(i, j, k) \in S$. Se hará referencia a cuatro tipos de arreglo: arreglo observado $\{x_{ijk}\}$, conteniendo los valores observados del número de personas en la categoría k, pasando de i a j; arreglo de valores esperados $\{m_{ijk}\}$; arreglo previo o de estimadores preliminares $\{m^{o}_{ijk}\}$ y arreglo de estimadores de máxima verosimilitud (EMV) de los valores esperados $\{\hat{m}_{ijk}\}$.

En aplicaciones prácticas el arreglo $\{x_{ijk}\}$ no es conocido, excepto para elementos $(i, j, k) \notin S$, si los hay. El uso principal de un arreglo observado es para analizar la validez de los métodos de estimación.

Los agregados a los conjuntos de índices I, J, K se denotan por puntos. Por ejemplo, el marginal bivariado $x_{.jk}$ es el total marginal de x_{ijk} sobre todos los $i \in I$. El marginal univariado $x_{..k}$ es la suma de x_{ijk} sobre todos los $i \in I$ y $j \in J$. El gran total es $x_{...} = N$ y es igual al total de pasajes en el sistema. Los totales marginales en el arreglo son de particular relevancia en el análisis multidimensional con datos incompletos, ya que la información disponible sobre los pasajes está generalmente limitada a los valores agregados de flujo.

En esta sección no se ha definido el término "pasaje". La definición depende mucho de la manera en que se miden los pasajes y, por lo tanto, del sistema para la recolección de datos. En los censos y encuestas retrospectivas, el pasaje se mide comparando el status en el momento de la enumeración con el status en una fecha previa. Sin embargo, en los sistemas de registro se contabiliza cada cambio en el status. Cualquiera que sea la definición de pasaje, el arreglo es una estrategia útil para integrar la disponibilidad y estimación de datos. Los arreglos tam-

bién proporcionan una conexión lógica entre datos y modelos demográficos. Rees y Wilson (1977) y Rees (1980) trabajan en las ventajas de los arreglos para la construcción del modelo demográfico y proponen algunas reglas para diseñarlo. La persona que se enfrenta con el problema de establecer un arreglo se refiere a la literatura, a Rees (1980) en particular. En este artículo suponemos que el arreglo está dado y enfoca los modelos de datos en el arreglo y la estimación de los elementos faltantes.

2.2. *Revisión ilustrativa de las fuentes de datos para el análisis multidimensional de población*

En esta sección listamos algunas fuentes de datos que proveen datos de suma relevancia para el análisis multidimensional. Esta revisión no es exhaustiva; su único propósito es ilustrar el tipo de datos disponibles y los tipos de problemas de estimación asociados con él. La disponibilidad de datos es un concepto complejo. Los datos pueden haber sido recolectados, pero no tabulados o pueden haber sido tabulados, pero no publicados. En el último caso, generalmente los datos están disponibles en microfichas; en el primero, pueden haber sido tabulados para su eventual consulta. Esta revisión considera datos para el análisis multirregional, análisis ocupacional y análisis de status marital. También se tocan algunas áreas nuevas para la aplicación de la demografía multidimensional.

a) Análisis multirregional

Los censos y sistemas de registro son las principales fuentes de datos de migración. En varios países europeos, se tiene que registrar cualquier cambio de residencia. Las oficinas centrales de estadísticas recolectan información de las oficinas de administración local para preparar las estadísticas de migración. La "tarjeta de movimiento" contiene información sobre origen y destino, así como sobre ciertas características, tales como la edad.

El censo es la principal fuente de datos de migración en la mayoría de los países. Datos relevantes se pueden derivar de respuestas a preguntas como edad, lugar de enumeración, lugar de nacimiento, lugar de residencia en una fecha previa fija (uno o cinco años antes) y duración de residencia. Una ventaja de los censos es que proporcionan información detallada sobre características migratorias. Sin embargo, una desventaja es que la información proporcionada puede no estar actualizada, debido al intervalo en que se llevan a cabo los censos. Por lo tanto, los datos censales pueden ser complementados con información de migración obtenida a través de otras fuentes, tales como encuestas de hogares y encuestas de ocupación, que también se llevan a cabo a intervalos regulares. Sin embargo, surge el problema de combinar los datos de fuentes separadas; un problema que será tratado a lo largo de este artículo.

b) Análisis ocupacional

Se construyeron tablas multidimensionales de vida activa para Dinamarca (Hoem y Fong, 1976; Willekens, 1980b) y para Estados Unidos (Schoen

y Woodrow, 1980; Smith, 1980). Hoem y Fong utilizan datos de flujo generados por una encuesta ocupacional especial que se llevó a cabo en 1973-1974. Los estudios de Schoen y Woodrow y los de Smith están basados en los datos de la "Current Population Survey" (CPS), proporcionados por la Oficina de Censos. Schoen y Woodrow dan una descripción detallada de los datos. Desde enero de 1973, la CPS contiene una pregunta retrospectiva sobre el status ocupacional de exactamente un año antes.

Sin embargo, esta pregunta se hizo solamente a personas de 16 años o más que estuvieran empleadas en el momento de la enumeración (cerca de 60% del total de los respondentes). Como consecuencia, sólo se puede llenar parte del arreglo con datos observados. En el futuro, puede ser que se encuentre disponible en los Estados Unidos un conjunto más completo de datos de flujos ocupacionales. The National Comission on Employment and Unemployment Statistics recomendó recientemente que la CPS resuma la publicación de datos de flujos en movilidad ocupacional y prepare mensualmente cintas con series de tiempo de datos brutos de flujos para uso público. También recomendó que el próximo cuestionario censal incluya una pregunta sobre ocupación, industria y lugar de residencia de un año antes (Stein, 1980).

Una fuente de datos potencialmente útil para el análisis multidimensional de la ocupación es la Encuesta de Ocupación semestral que se llevó a cabo en 1973 en cada país miembro de la Comunidad Europea. Esta encuesta distingue ocho status ocupacionales (de los cuales uno es empleado) y contiene para cada status una pregunta retrospectiva sobre el status ocupacional de un año antes (desde 1977, también se pregunta sobre el status de dos años antes). La encuesta, que está diseñada con base comparable en cada país, proporciona rica información demográfica y socioeconómica sobre los respondentes. Debe tenerse cuidado de extraer solamente la información que esté sustentada por el tamaño de muestra.

El "Dutch National Program for Demographic Research" inició un proyecto enfocado en el desarrollo de tablas multidimensionales de vida activa para Holanda en base a los datos de transición proporcionados por la Encuesta Ocupacional. En una etapa futura se investigará si estos datos pueden ser útiles para proyectar la ocupación. En el apéndice A, el Sr. A. Struyk, que está llevando a cabo el proyecto de investigación, presenta una revisión general de la organización y del contenido de datos de las Encuestas Ocupacionales de la Comunidad Europea.

c) *Análisis de nupcialidad*

Hay disponibles más datos de flujo sobre la formación y disolución matrimonial que para otras aplicaciones del análisis multidimensional. El registro de cambios en el status marital es común en la mayoría de los países. Los datos de situaciones belgas y holandesas están discutidos en Willekens, et al. (1979) y en Koesoebjono (1981), respectivamente. Schoen y Nelson (1974), proporcionan profundizaciones del tipo de datos disponibles en los Estados Unidos sobre cambios maritales.

d) Análisis de la educación y otras posibles aplicaciones de la demografía multidimensional

Los análisis multirregional, ocupacional y marital son solamente algunas de las muchas situaciones en que las técnicas de la demografía multidimensional pueden ser aplicadas en forma fructífera. De hecho, cualquier investigación demográfica de eventos renovables se puede beneficiar con la aplicación del modelo de tabla de vida de incrementos-decrementos y, por lo tanto, de la demografía multidimensional. El creciente número de encuestas retrospectivas puede suministrar los datos de entrada requeridos. El análisis de la educación es un nuevo campo posible para su aplicación. En Holanda, el "Central Bureau of Statistics" recolectó datos sobre flujos hacia, de y dentro del sistema educacional, a través de una encuesta retrospectiva de alumnos durante 1978. En la encuesta se preguntó sobre tipo y nivel de educación un año antes.

La matriz de flujos publicada cubre un amplio conjunto de categorías educacionales; la dimensión de edad, sin embargo, ha sido abandonada, pero puede ser recuperada fácilmente. Entre otros tópicos de análisis, en los que la aplicación de las técnicas de la demografía multidimensional podría llevar a una mejor comprensión, se incluyen el estudio de movilidad social, la planificación familiar, la participación de seguridad social, etc. Para algunos de estos estudios, los datos se pueden obtener de las encuestas retrospectivas. Sin embargo, aunque la disponibilidad de datos no sea adecuada, puede tenerse en consideración la demografía multidimensional. Con métodos de estimación apropiados, podrán inferirse los datos requeridos de los conocimientos que se tengan. Dichas técnicas son el tema principal de este artículo.

3. Modelos de datos en el arreglo

La investigación de grandes conjuntos de datos se vuelve relativamente sencilla ajustando modelos a los datos. En la década pasada, el análisis estructural de tablas de contingencia atrajo un interés considerable y los resultados de esta investigación están bien documentados (Bishop, Fienberg y Holland, 1975; Goodman, 1978; Gokhale y Kullback, 1978; Haberman, 1979). Técnicas analíticas, desarrolladas originalmente para identificar patrones de asociación entre diversas variables categóricas, pueden ser aplicadas fructíferamente en estimaciones.[2] El modelo log-lineal es una de ellas. Es parte de una clase de modelos lineales generalizados que describen los valores de las celdas en términos de totales marginales y de interacciones entre clasificaciones cruzadas de variables.

[2] En las estadísticas matemáticas se hace la distinción entre "estimación" y "predicción". La estimación se refiere a los parámetros del modelo, mientras que la predicción está relacionada con los resultados (de variables aleatorias) obtenidos con la aplicación o postulación de un modelo particular. En este artículo no se hace distinción, debido a que en la literatura demográfica "predicción de migración" tiene un significado bien definido, pero diferente. Para la teoría estadística de predicción, aplicada a migración, remitirse a De Jong (1981).

En la primera sección se presenta el modelo log-lineal y se demuestra que es una herramienta eficiente para el análisis estructural de los datos categóricos. La segunda sección da una regla simple para determinar la información necesaria para estimar los parámetros del modelo.

3.1 *El modelo log-lineal*

El modelo log-lineal no es desconocido en demografía. En forma reciente, un creciente número de autores ha adoptado esta perspectiva de modelo para estudiar dependencias entre tabulaciones cruzadas de variables demográficas (Little, 1978, 1980; Little y Pullum, 1979; Hobcraft, 1978; Clogg, 1978, 1980; Fienberg y Mason, 1978; y otros).

La formulación actual del modelo log-lineal se debe a Birch (1963) y se parece al modelo del análisis de variación. Hay una revisión clara hecha por Payne (1977). El cuadro 1 muestra el modelo para los valores esperados de las celdas en dos formas equivalentes. La formulación aditiva es más popular, debido a que se asemeja bastante al paradigma del análisis de variación. La formulación multiplicativa sin embargo, es conveniente para resolver problemas de estimación, debido a que está directamente relacionada con los modelos convencionales de estimación para datos sobre flujos. Esta relación con los modelos convencionales simplifica la interpretación de los parámetros del modelo en términos de datos disponibles. Además, algoritmos desarrollados en la ciencia regional y la ciencia del transporte pueden ser aplicados fructíferamente para resolver problemas de estimación en el análisis multidimensional de población.

En el modelo de ambas formulaciones, la aditiva y la multiplicativa, hay ocho términos. El número de términos depende de la dimensión del arreglo y no está relacionado con el número de celdas o, en forma equivalente, con el número de estados o categorías a lo largo de las dimensiones. Sin embargo, el número de valores de los parámetros depende del número de celdas en el arreglo. En los modelos 1 y 11, hay tantos valores de los parámetros independientes como celdas en el arreglo. El modelo es, por consiguiente, conocido como *modelo log-lineal saturado*. Cada parámetro en el modelo representa un efecto estructural particular en m_{ijk}. De acuerdo con el modelo log-lineal, el valor esperado es la suma de diversos efectos. El efecto total es el efecto del tamaño; es la media geométrica de todos los valores de las celdas. Los principales efectos denotan los efectos en m_{ijk} de diferencias relativas en tamaño entre los diversos marginales univariados. Por ejemplo, w_k^C y u_k^C son los efectos de clasificación de la edad en el número de pasajes m_{ijk}. Siendo todo lo demás igual, grandes grupos de edad llevan a grandes valores de m_{ijk} El efecto de la edad es la razón entre la media geométrica del k-ésimo estrato y la media geométrica del total. Comparando diversas medias geométricas, se puede determinar el rango de efectos ejercidos por m_{ijk}. Por ejemplo, para determinar si el efecto de edad difiere del estado de origen i, basta con calcular w_{ik}^{AC} o u_{ik}^{AC} . Un valor diferente de cero de u_{ik}^{AC} significa la existencia de una interacción entre edad y origen. Nótese que el patrón de interacción determinado de esta manera representa el promedio de interacción de todas las tablas AC (para todas las regiones de destino posibles; es decir, los valores j de la variable B). El patrón puede diferir para cada nivel de B, resultando en

un u_{ijk}^{ABC} diferente de cero. Si $u_{ijk}^{ABC} \neq o$, entonces la interacción entre pares de A, B y C debe ser también diferente de cero. El principio de que para cada término u diferente de cero, sus relativos de menor orden también deben ser diferentes de cero es conocido como el principio jerárquico (Bishop, Fienberg y Holland, 1975, p. 34). Recíprocamente, si cualquier término u se considera igual a cero, sus relativos de orden superior también deben ser cero. En este artículo sólo se consideran los modelos jerárquicos log-lineales.

Con la introducción del modelo log-lineal, se transforma el problema de estimación de las celdas en un problema de estimación de parámetros; por ejemplo, la cuantificación de diversos efectos. Los efectos son determinados basándose en los datos disponibles, aumentados por supuestos.

La siguiente conclusión es la base para la derivación del procedimiento de estimación. *Para estimar los valores de las celdas en un arreglo multidimensional, se deben cuantificar los efectos de interacción. Por lo tanto, la estimación de los datos está estrechamente relacionada con la prueba de hipótesis.*

Si todos los datos están disponibles, es decir, si se desconoce el arreglo $\{x_{ijk}\}$, entonces todos los valores de los parámetros se pueden derivar de dichos datos. En el cuadro 4 de la sección 5 se muestran los parámetros del modelo log-lineal de un conjunto de datos de movilidad social.

Se puede ver fácilmente que el modelo log-lineal saturado es una réplica exacta de los datos observados, es decir,

$$m_{ijk} = x_{ijk}.$$

En aplicaciones prácticas, no se conoce $\{x_{ijk}\}$ y se tienen que calcular los parámetros, de cualquier información previa existente. Si algunos parámetros se pueden estimar de datos disponibles, se considera su valor igual a cero y se supone que su efecto de interacción asociado está ausente. El modelo log-lineal con algunos términos ausentes es el *modelo no saturado.* La siguiente sección describe un enfoque integrado para la estimación de parámetros. Se debe tener en mente que con la estimación de un parámetro del modelo log-lineal, se está imponiendo un patrón particular de interacción en el arreglo $\{m_{ijk}\}$. Recíprocamente, si se quiere que los estimadores expresen que ciertas variables dependen de alguna manera particular unas de otras, se debe introducir esto a través de valores apropiados de los parámetros modelo relevantes.

3.2 *Estimador suficiente para obtener celdas*

Primero se determina qué información se requiere para estimar los parámetros del modelo log-lineal. Para hacer esto, se considera que el arreglo es el resultado de un esquema multinomio simple de muestreo (patrón en el que el tamaño total de muestra es fijo, sea N, y cada celda tiene una distribución independiente de Poisson). Entonces, el arreglo denota una distribución multinomial $M(N; m_{ijk}/m_{...})$ con una función de densidad de probabilidad (Fisher, 1922; Bishop, Fienberg y Holland, 1975, p. 63):

$$P\,(X_{ijk} = x_{ijk} \text{ para todas } i,j,k \mid X_{...} = N) = \frac{N!}{\prod_{ijk} x_{ijk}!} \prod_{ijk} \left(\frac{m_{ijk}}{m}\right)^{x_{ijk}} \quad (21)$$

Cuadro 2

El modelo Log-Lineal

Formulación multiplicativa		Formulación aditiva	
Modelo $m_{ijk} = w_i^A w_j^B w_k^C w_{ij}^{AB} w_{ik}^{AC} w_{jk}^{BC} w_{ijk}^{ABC}$	(1)	$\ln m_{ijk} = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}$	(11)
Media de efecto total $w = \left[\prod_{i,j,k} m_{ijk}\right]^{\frac{1}{RCL}}$	(2)	$u = \frac{1}{RCL} \sum_{i,j,k} \ln m_{ijk}$	(12)
Principales efectos $w_i^A = \frac{1}{w}\left[\prod_{j,k} m_{ijk}\right]^{\frac{1}{CL}}$	(3)	$u_i^A = \frac{1}{CL} \sum_{j,k} \ln m_{ijk} - u$	(13)
w_j^B y w_k^C: análogas		u_i^B y u_k^C: análogas	
Efectos de interacción de primer orden (interacción de dos entradas o en forma de par)			
$w_{ij}^{AB} = \frac{1}{z}\left[\prod_k m_{ijk}\right]^{\frac{1}{L}}$	(4)	$u_{ij}^{AB} = \frac{1}{L} \sum_k \ln m_{ijk} - c$	(14)
con $z = w w_i^A w_j^B w_k^C$		con $c = u + u_i^A + u_j^B + u_k^C$	
w_{ik}^{AC} y w_{jk}^{BC}: análogas		u_{ik}^{AC} y u_{jk}^{BC}: análogas	
Efectos de interacción de segundo orden (interacción de tres entradas)			
$w_{ijk}^{ABC} = \frac{1}{z'} m_{ijk}$,	(5)	$u_{ijk}^{ABC} = \ln m_{ijk} - c'$,	(15)
con $z' = w w_i^A w_j^B w_k^C w_{ij}^{AB} w_{ik}^{AC} w_{jk}^{BC}$		con $c' = u + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC}$	
Restricciones $\prod_i w_i^A = \prod_j w_j^B = \prod_k w_j^C = 1$	(6)	$\sum_i u_i^A = \sum_j u_j^B = \sum_k u_k^C = 0$	(16)
$\prod_i w_{ij}^{AB} = \prod_j w_{ij}^{AB} = \prod_i w_{ik}^{AC} = \prod_k w_{ik}^{AC} = \prod_j w_{jk}^{BC} = \prod_k w_{jk}^{BC} = 1$	(7)	$\sum_i u_{ij}^{AB} = \sum_j u_{ij}^{AB} = \sum_i u_{ik}^{AC} = \sum_k u_{ik}^{AC} = \sum_j u_{jk}^{BC} = \sum_k u_{jk}^{BC} = 0$	(17)
$\prod_i w_{ijk}^{ABC} = \prod_j w_{ijk}^{ABC} = \prod_k w_{ijk}^{ABC} = 1$	(8)	$\sum_i u_{ijk}^{ABC} = \sum_j u_{ijk}^{ABC} = \sum_k u_{ijk}^{ABC} = 0$	(18)
$w = \exp u$	(9)	$u = \ln w$	(19)
$w_i^A = \exp u_i^A$		$u_i^A = \ln w_i^A$	
$w_{ij}^{AB} = \exp u_{ij}^{AB}$	(10)	$u_{ij}^{AB} = \ln w_{ij}^{AB}$	(20)
$w_{jik}^{ABC} = \exp u_{ijk}^{ABC}$		$u_{ijk}^{ABC} = \ln w_{ijk}^{ABC}$	

La verosimilitud logarítmica del multinomio es

$$\ln\left\{\frac{N!}{\prod_{i,j,k} x_{ijk}\,!}\right\} + \sum_{i,j,k} x_{ijk} \ln m_{ijk} - N \ln nm \quad \ldots \tag{22}$$

Maximizando la verosimilitud logarítmica bajo el supuesto de que $\sum_{ijk} (m_{ijk}/m_{...}) = 1$, podemos hacer caso omiso del primer y tercer términos y considerar solamente el restante, el núcleo de la función de verosimilitud logarítmica. La forma de producto de este núcleo.

$$W = \prod_{ijk} (m_{ijk})^{x_{ijk}} \tag{23}$$

representa el número esperado de individuos que caen en la celda (i, j, k) si se escogen aleatoriamente los miembros x_{ijk} de la población de tamaño N. La cantidad $\frac{W}{N}$ es, por lo tanto, la probabilidad de que un individuo escogido aleatoriamente caiga en la celda (i, j, k). Sustituyendo $\ln m_{ijk}$ en el núcleo del modelo log-lineal (11), da que:

$$\begin{aligned}\sum_{i,j,k} x_{ijk} \ln m_{ijk} = Nu &+ \sum_i x_{i..}\, u_i^A + \sum_j x_{.j.}\, u_j^B \\ &+ \sum_k x_{..k}\, u_k^C + \sum_{i,j} x_{ij.}\, u_{ij}^{AB} \\ &+ \sum_{i,k} x_{i.k}\, u_{ik}^{AB} + \sum_{j,k} x_{.jk}\, u_{jk}^{BC} \\ &+ \sum_{i,j,k} x_{ijk}\, u_{ijk}^{ABC}\end{aligned} \tag{24}$$

El estimador suficiente para obtener parámetros del modelo log-lineal consiste en términos x adyacentes a los parámetros desconocidos. Por ejemplo, para determinar el efecto columna u_j^B, se necesita saber el total $x_{.j.}$; y para determinar la interacción entre A y B, se debe saber el total marginal $x_{ij.}$. Nótese que la interacción derivada de $x_{ij.}$ es un promedio de los patrones de interacción en varios niveles k de la tercer variable C.

El conocimiento del marginal bivariado $x_{ij.}$ implica el conocimiento de los marginales univariados $x_{i..}$ y $x_{.j.}$, necesarios para determinar u_i^A y u_j^B, respectivamente. Por lo tanto, se puede definir un estimador mínimo suficiente (en este caso $x_{ij.}$). En la práctica, el estimador mínimo suficiente se puede obtener inspeccionando el modelo log-lineal.

La expresión (24) relaciona el modelo log-lineal con la función de verosimilitud. Birch (1963) demostró que existe un conjunto único de estimadores de celdas elementales que

(i) satisface las condiciones impuestas por la forma del modelo paramétrico (log-lineal),

(ii) satisface las restricciones de que los totales marginales de los estimadores m_{ijk} correspondan a los totales marginales dados ($x_{ij.}$, $x_{.j.}$, etc.) y

(iii) maximiza la función de verosimilitud.

Por lo tanto, con la aplicación del modelo log-lineal, se pueden derivar, del estimador suficiente solo, estimadores de máxima verosimilitud $\hat{m}_{ijk}$ de los valores esperados m_{ijk}.

Formulada de una manera diferente y más de acuerdo con el tema principal del presente artículo, la aplicación del modelo log-lineal a los datos disponibles da estimadores que no sólo son consistentes con lo que ya se sabe, sino también tienen una verosimilitud máxima de ocurrencia. Si se está bien informado, se notará la equivalencia entre estos aspectos y las características de los estimadores obtenidos con la aplicación de técnicas de maximización de entropía, que fueron desarrolladas por Wilson (1980) en el campo de la ciencia regional. La equivalencia entre maximización de entropía y estimación de máxima verosimilitud fue demostrada formalmente por Batty y Mackie (1972) y Willekens (1980).

4. Estimación de parámetros, a partir de datos disponibles

Para estimar elementos faltantes en el arreglo multidimensional, se sugirió un modelo de enfoque y se escogió el modelo log-lineal como una representación adecuada de los datos. La forma funcional del modelo paramétrico, subyacente al procedimiento de estimación, es, por lo tanto, fija. El estimador suficiente para estimar los parámetros del modelo consiste de términos x adyacentes a los parámetros desconocidos. En este capítulo se revisa el procedimiento para derivar, de datos incompletos, los valores de parámetros para este modelo. Primero, se supone que la información previa está limitada únicamente a totales marginales. Después, se demostrará cómo se pueden introducir otros tipos de información. Cualesquiera que sean los datos previos que se utilicen, la estrategia de estimación es la misma: el conocimiento previo da estimadores de los parámetros del modelo log-lineal y, por lo tanto, impone una estructura a los valores esperados m_{ijk}. Mientras mejor se pueda describir la relación estructural entre las clasificaciones cruzadas de variables, mejores serán los estimadores. Si no se pueden obtener algunos valores de los parámetros por falta de información, observada o aproximada, sobre términos x adyacentes, los valores se consideran igual a cero, lo cual implica la ausencia del patrón de interacción al que representan.

4.1 *Métodos para estimar las celdas a partir de totales marginales*

Los EMV que estamos buscando satisfacen el modelo log-lineal y las restricciones marginales. El modelo y las restricciones forman un sistema de ecuacio-

nes, las ecuaciones de máxima verosimilitud, cuya solución da los EMV. En algunos casos, las ecuaciones de los EMV tienen una forma cerrada de solución.

a) Expresiones en forma cerrada

Los estimadores pueden ser expresados en forma cerrada si los totales marginales conocidos satisfacen condiciones particulares. Por ejemplo, si la información disponible está limitada a $x_{i..}$, $x_{.j.}$ y $x_{..k}$, entonces los EMV $\hat{m}_{ijk}$ son la solución del siguiente sistema de ecuaciones:
–ecuaciones modelo:

$$m_{ijk} = w\, w_i^A\, w_j^B\, w_k^C$$
$$\text{o} \quad \ln m_{ijk} = u + u_i^A + u_j^B + u_k^C \tag{25}$$

con los parámetros w y u que satisfacen a la (6) y (16), respectivamente.
– ecuaciones del estimador mínimo suficiente

$$\sum_{j,k} \hat{m}_{ijk} = x_{i..} \tag{26}$$

$$\sum_{i,k} \hat{m}_{ijk} = x_{.j.} \tag{27}$$

$$\sum_{i,j} \hat{m}_{ijk} = x_{..k} \tag{28}$$

$$\sum_{i,j,k} m_{ijk} = x_{...} = N \tag{29}$$

Los estimadores de celdas para el modelo (25) y el estimador suficiente que consta de la (26) a la (29) son

$$\hat{m}_{ijk} = \frac{x_{i..}\, x_{.j.}\, x_{..k}}{N^2}$$

Esta fórmula proporciona una expresión en forma cerrada de EMV en términos de estimadores suficientes. Es la ilustración más simple del problema de estimación multidimensional. Willekens, Por y Raquillet (1979) nombran este caso el problema 3E (orillas) ya que la información conocida se puede arreglar en las orillas de una caja, cuyo contenido se tiene que estimar. Existen otros varios estimadores en forma cerrada en tres dimensiones. Se pueden derivar resolviendo el conjunto apropiado de ecuaciones de máxima verosimilitud. El cuadro 2 resume

Cuadro 2

EMV en arreglos tridimensionados

Caso	Datos Disponibles	Modelo log-lineal	Estimaciones en forma cerrada de celdas	Interpretación del modelo log-lineal
3E	{A, B, C}	$m_{ijk} = w\, w_i^A w_j^B w_k^C$	$\hat{m}_{ijk} = \frac{1}{N^2} \times_{i..} \times_{.j.} \times_{..k}$	-Independencia mutua (variables A, B y C son independientes).
1FE	{AB, C}	$m_{ijk} = w\, w_i^A w_j^B w_k^C w_{ij}^{AB}$	$\hat{m}_{ijk} = \frac{1}{N} \times_{ij.} \times_{..k}$	-Independencia múltiple (variable unida AB es independiente de C)
2F	{AB, BC}	$m_{ijk} = w\, w_i^A w_j^B w_k^C w_{ij}^{AB} w_{jk}^{BC}$	$\hat{m}_{ijk} = \times_{ij.} \times_{.jk} / \times_{.j.}$	-Independencia condicionada (A independiente de C, B dada)
3F	{AB, AC, BC}	$m_{ijk} = w\, w_i^A w_j^B w_k^C w_{ij}^{AB} w_{jk}^{BC} w_{ik}^{AC}$	solución en forma no cerrada.	-Asociación en par (cada interacción de dos entradas es independiente del nivel de la tercer variable).
CI	{ABC}	$m_{ijk} = w\, w_i^A w_j^B w_k^C w_{ij}^{AB} w_{jk}^{BC} w_{ik}^{AC} w_{ijk}^{ABC}$	$\hat{m}_{ijk} = \times_{ijk}$	-Interacción de tres entradas (la asociación entre cada par de variables varía con el nivel de la tercer variable).

estos resultados. Los datos disponibles están representados como un conjunto de marginales univariados y/o bivariados. Por ejemplo, en un sistema de población multirregional, podría surgir un caso en el que el patrón de migración es conocido para la población total y sólo es dada una sola estructura de migrantes por edad. Esta información previa podría ser arreglada en una cara o una orilla de una caja y, por lo tanto, se puede usar el llamado problema 1FE (cara, orilla). El procedimiento de los EMV se reduce a aplicar la composición de edad $\frac{1}{N} x_{..k}$ a cada elemento de la matriz de migración $x_{ij.}$. La aplicación de una composición de edad individual implica que en el arreglo m_{ijk} la edad es independiente de la región de origen y la región de destino. Aunque sea un supuesto no realista, es válido en el modelo, debido a la información previa tan limitada.

Como no tenemos estimadores suficientes para estimar todos los parámetros del modelo log-lineal, postulamos que algunos parámetros son cero, con la consecuencia de que los efectos de interacción a los que representan están ausentes. En la siguiente sección, se verá que aún con la ausencia de datos "duros", se puede imponer una estructura al arreglo $\{m_{ijk}\}$ con el uso de información "suavizada" como los datos recolectados en una fecha previa en la opinión de un experto, medidas relativas obtenidas en encuestas por muestreo, etc.

La existencia de expresiones en forma cerrada para los EMV es muy conveniente para el análisis multidimensional con datos incompletos. Bishop, Fienberg y Halland (1975, pp. 76-82) proporcionan algunas reglas para detectar la existencia de estimadores directos. Una característica interesante de estas reglas es que se aplican a arreglos de cualquier dimensión. La idea principal es suprimir configuraciones redundantes de datos disponibles o subconfiguraciones traslapadas. Si no quedan más de dos configuraciones, existen estimadores en forma cerrada. En general (para cualquier dimensión), se puede demostrar que esta declaración implica que al menos un efecto de dos factores debe estar ausente para que existan estimadores directos. La forma general de los estimadores directos es predecible: el numerador tiene entradas de cada configuración suficiente; el denominador tiene entradas de configuraciones redundantes causadas por el traslape; términos en potencias de N aseguran el orden de magnitud correcto. El traslape está ilustrado en el caso 2F: el subíndice j aparece en ambas configuraciones, la AB y la BC.

b) Arreglo iterativo por ajuste multiproporcional

Para derivar los EMV, el problema 3F requiere un estimador mínimo suficiente, que consiste de tres marginales bivariados. No existe solución en forma cerrada y el ajuste iterativo de las configuraciones suficientes o datos previos es la única salida. Para comenzar el procedimiento, se puede escoger cualquier conjunto de estimadores preliminares que no exhiba un efecto de tres factores ($u_{ijk}^{ABC} = 0$). Por ejemplo, la distribución uniforme satisface esta condición y, por lo tanto, un valor inicial conveniente es $m_{ijk}^{(0)} = 1$ para todos los i, j y k.

El algoritmo iterativo va como sigue:

Paso 0: $s = 0$

Paso 1: Ajuste proporcional a lo largo de la dimensión C

$$\hat{m}_{ijk}^{(3s+1)} = \hat{m}_{ijk}^{(3s)} \frac{x_{ij.}}{\sum_k \hat{m}_{ijk}^{(3s)}}$$

Paso 2: Ajuste proporcional a lo largo de la dimensión A

$$\hat{m}_{ijk}^{(3s+2)} = \hat{m}_{ijk}^{(3s+1)} \frac{x_{.jk}}{\sum_i \hat{m}_{ijk}^{(3s+1)}}$$

Paso 3: Ajuste proporcional a lo largo de la dimensión B

$$\hat{m}_{ijk}^{(3s+3)} = \hat{m}_{ijk}^{(3s+2)} \frac{x_{i.k}}{\sum_j \hat{m}_{ijk}^{(3s+2)}}$$

Si el criterio para detenerse $\left| \frac{\hat{m}_{ijk}^{(3s+3)}}{\hat{m}_{ijk}^{(3s+2)}} - 1 \right| \leqslant \epsilon$

se satisface para cada i, j y k, entonces detenga la iteración; de otra manera, s = s + 1 y vaya al paso 1.

El algoritmo es una variante especial de un algoritmo más general discutido en la siguiente sección. Este método de ajuste proporcional sucesivo se conoce bajo varios nombres. Fue desarrollado originalmente por Barlett en 1935, quien lo llamó el "modelo de interacción no de segundo orden". En el análisis de la tabla de contingencia, se hizo conocido como el método de ajuste proporcional iterativo (API) (ver e.g. Bishop, Fienberg y Holland, 1975, pp. 83-97). En este artículo, el algoritmo es referido como algoritmo de ajuste multiproporcional (AJM), por razones que se explicarán más adelante. Aplicando el principio de descomposición de Rockafellar; Willekens, Por y Raquillet (1979) muestran que este algoritmo puede ser derivado de una maximización de la función de entropía

$$\sum_{i,j,k} m_{ijk} \ln m_{ijk}, \tag{30}$$

sujeta a las restricciones bivariadas $\{AB, AC, BC\}$ y que el método converge a una solución única (existen otras pruebas de convergencia API; ver referencias en Bishop. Fienberg y Holland, 1975, p. 85). Ellos demuestran que el algoritmo API es equivalente al algoritmo primordial directo del problema no lineal matemático de programación. Además, los autores derivan un algoritmo basado en la formulación dual del problema de programación. La ventaja de esta formulación dual es que se relaciona más directamente con los parámetros del modelo log-lineal. La dualidad será discutida en la siguiente sección.

Para demostrar que AJM no introduce un efecto de tercer orden, podemos re-escribir los estimadores como un producto de funciones entre dos variables únicamente:

$$\hat{m}_{ijk} = f_1(i,j) \cdot f_2(j,k) \cdot f_3(i,k), \tag{31}$$

$$\text{con} \quad f_i(i,j) = \prod_s \frac{\hat{m}_{.ij}}{\sum_k \hat{m}^{(3s)}_{.ijk}}$$

$$f_2(j,k) = \prod_s \frac{\hat{m}_{.jk}}{\sum_i \hat{m}^{(3s+1)}_{ijk}}$$

$$f_3(i,k) = \prod_s \frac{\hat{m}_{i.k}}{\sum_j \hat{m}^{(3s+2)}_{ijk}}$$

donde s denota la interacción.

Chilton y Poet (1973) y Caussinus y Thelot (1976) presentan algoritmos equivalentes para calcular las funciones bivariadas de (31).

4.2 *Métodos para estimar celdas a partir de totales marginales complementados con otra información previa*

Los términos x adyacentes a los parámetros desconocidos (ver expresión 24) son suficientes para estimar los parámetros del modelo log-lineal. En la sección anterior, se supuso que los parámetros eran cero y que el patrón de interacción al que representan estaba ausente, si el término x requerido no estaba disponible. En particular, no se pudo asignar ningún valor al término u^{ABC}, ya que se desconocían las celdas individuales x_{ijk}. Si no están disponibles algunos términos x, los estimadores de parámetros se pueden derivar de otras fuentes de información previa. Una combinación de diferentes fuentes de datos puede permitir la derivación de estimadores adecuados para los parámetros del modelo log-lineal y para los valores de las celdas. Llamaremos *fuente principal de datos* al conjunto de términos x conocidos; *fuente(s) auxiliar(es) de datos* a la información previa de la que se derivan los parámetros del modelo que no se pueden obtener directamente de los términos x. La información previa puede venir de varias maneras. Por ejemplo, pueden faltar datos detallados de movilidad para un país, pero pueden existir para otro. Si los dos países son similares, entonces los parámetros del modelo log-lineal que se obtuvieron para un país se pueden aplicar para derivar los estimadores de movilidad para el otro país. Análogamente, se pueden combinar fuentes similares de datos de diferentes períodos. Los censos pueden tener la información detallada requerida para el análisis multidimensional, pero puede ser que las estadísticas ya sean anticuadas. Una combinación de información censal

y datos más recientes puede dar una base adecuada para el análisis multidimensional. Algunas veces se puede introducir la opinión de un experto, la intuición y el sentido común para aumentar la calidad de esta base de datos. Podemos saber anticipadamente que algunas transiciones son imposibles o que tienen que tomar ciertos valores. Por ejemplo, es imposible un reingreso al estado de soltero. En algunos casos, se hace caso omiso de algunas transiciones: en el análisis multirregional, generalmente no se consideran las migraciones intra-regionales, debido a que no afectan la redistribución de la población. Valores fijos de las celdas surgen cuando algunos elementos x_{ijk} son observados. Esto es ilustrado, por ejemplo, por la base de datos utilizada por Schoen y Woodrow (1980) para construir las tablas de vida activa. Se observaron transiciones ocupacionales solamente para un subgrupo de la población; es decir, para aquellas personas empleadas en el momento de la enumeración. Se tuvieron que estimar las transiciones hechas por personas en una categoría ocupacional diferente.

El propósito de esta sección es demostrar cómo se puede generar una fuente auxiliar de datos apropiada, si no existe, y cómo se puede combinar con la fuente principal de datos para dar estimadores exactos para los parámetros y celdas del modelo log-lineal. La idea principal es que los parámetros que no se pueden estimar de la fuente principal de datos se derivan ("toman prestados") de la fuente auxiliar.

4.2.1 *Combinación de las fuentes de datos*

La fuente auxiliar de datos se denota por $\{x^{o}_{ijk}\}$ y da origen al arreglo de estimadores preliminares $\{m^{o}_{ijk}\}$.

Este arreglo es de la misma dimensión y magnitud que el arreglo $\{x_{ijk}\}$. Si todos los x^{o}_{ijk} son observados, entonces $m^{o}_{ijk} = x^{o}_{ijk}$. En este artículo suponemos que todos los x^{o}_{ijk} son observados.

Ambos conjuntos de datos, el principal y el auxiliar, se pueden incorporar en un modelo por las expresiones log-lineales:

$$\ln m_{ijk} = u + u^{A}_{i} + u^{B}_{j} + u^{C}_{k} + u^{AB}_{ij} + u^{AC}_{ik} + u^{BC}_{jk} + u^{ABC}_{ijk} \qquad (32)$$

$$\ln m^{o}_{ijk} = {}^{o}u + {}^{o}u^{A}_{i} + {}^{o}u^{B}_{j} + {}^{o}u^{C}_{k} + {}^{o}u^{AB}_{ij} + {}^{o}u^{AC}_{ik} + {}^{o}u^{BC}_{jk} + {}^{o}u^{ABC}_{ijk}, \qquad (33)$$

y

$$\ln \frac{m_{ijk}}{m^{o}_{ijk}} = {}^{r}u + {}^{r}u^{A}_{i} + {}^{r}u^{B}_{j} + {}^{r}u^{C}_{k} + {}^{r}u^{AB}_{ij} + {}^{r}u^{AC}_{ik} + {}^{r}u^{BC}_{jk} + {}^{r}u^{ABC}_{ijk}, \qquad (34)$$

con

$$u = u - {}^{o}u$$

$$ {}^{r}u_i^A = u_i^A - {}^{o}u_i^A \qquad {}^{r}u_j^B = u_j^B - {}^{o}u_j^B \qquad {}^{r}u_k^C = u_k^C - {}^{o}u_k^C $$

$$ {}^{r}u_{ij}^{AB} = u_{ij}^{AB} - {}^{o}u_{ij}^{AB} \qquad {}^{r}u_{ik}^{AC} = u_{ik}^{AC} - {}^{o}u_{ik}^{AC} \qquad {}^{r}u_{jk}^{BC} = u_{jk}^{BC} - {}^{o}u_{jk}^{BC} $$

$$ {}^{r}u_{ijk}^{ABC} = u_{ijk}^{ABC} - {}^{o}u_{ijk}^{ABC} \, . $$

La formulación multiplicativa del modelo log-lineal (34) es

$$ m_{ijk} = m_{ijk}^{o} \; {}^{r}w \, {}^{r}w_i^A \, {}^{r}w_j^B \, {}^{r}w_k^C \, {}^{r}w_{ij}^{AB} \, {}^{r}w_{ik}^{AC} \, {}^{r}w_{jk}^{BC} \, {}^{r}w_{ijk}^{ABC} \, , \qquad (35) $$

donde el sobreíndice r denota la razón del término w en el modelo log-lineal de los estimadores finales y el término ^{o}w, de los estimadores preliminares.

Suponiendo que el arreglo auxiliar es completamente observado ($m_{ijk}^{o} = x_{ijk}^{o}$), el modelo log-lineal asociado está saturado y todos los parámetros ^{o}u, incluyendo el término de interacción de segundo orden, se pueden calcular de los estimadores preliminares. Por otro lado, los valores de los parámetros u se determinan de la siguiente manera:

i – para términos x disponibles, los parámetros u asociados se calculan de la fuente principal de datos;

ii – los parámetros u que no se pueden determinar de la fuente principal de datos se consideran iguales a los términos ^{o}u equivalentes.

De acuerdo con este procedimiento, los patrones de interacción entre variables en el arreglo $\{m_{ijk}\}$ se derivan, hasta donde es posible, de la fuente principal de datos. Los patrones de interacción que no se pueden determinar de esta manera se consideran iguales a los patrones de interacción observados en el arreglo auxiliar. Ahora podemos contestar una importante pregunta: ¿Cómo contribuye la fuente auxiliar de datos a la calidad de los estimadores m_{ijk}? El conjunto auxiliar de datos introduce patrones de interacción a los estimadores finales que no se pueden derivar de la fuente principal de datos. Como corolario, se puede concluir que a través de la selección de un arreglo apropiado $\{m_{ijk}^{o}\}$, combinado con el conjunto correcto de términos x, se puede imponer cualquier patrón de interacción de orden superior en el arreglo $\{m_{ijk}\}$. Este enfoque puede ser de gran utilidad en el análisis multidimensional de población con datos incompletos. Los estimadores obtenidos son EMV o aproximaciones de EMV, dependiendo de los datos disponibles (Haberman, 1979, pp. 519-540).

El arreglo $\{m_{ijk}\}$ que satisface las condiciones descritas arriba, se puede obtener sin que el cálculo de los parámetros del modelo log-lineal sea un paso intermedio. Ajustes proporcionales de las celdas del arreglo $\{m_{ijk}^{o}\}$ para formar un conjunto de marginales predefinidos (términos x) dan estimadores apropiados. Si se conocen tres totales marginales bivariados ($x_{ij.}$, $x_{i.k}$, $x_{.jk}$), el arreglo se asemeja estrechamente al procedimiento de Barlett, mencionado en la sección anterior y conocido como ajuste multiproporcional. En lugar de empezar con la distribución uniforme que da un modelo de "interacción no de segundo orden", se usa el arreglo $\{m_{ijk}^{o}\}$:

Paso 0: $m_{ijk}^{(0)} = m_{ijk}^{o}$ para todos los i, j, k ϵ S

El algoritmo converge a los EMV (Haberman, 1979, p. 540). En este caso 3F, los efectos principales y los efectos de interacción de primer orden se derivan de los marginales dados; sólo los efectos de interacción de segundo orden se "toman prestados" del arreglo $\{m_{ijk}^{o}\}$, es decir, $u_{ijk}^{ABC} = {}^{0}u_{ijk}^{ABC}$ y, por lo tanto ${}^{r}u_{ijk}^{ABC} = 0$.

El modelo log-lineal de este caso 3F se deriva fácilmente de la (34) con la supresión del término de interacción de segundo orden. El modelo log-lineal es equivalente a la expresión obtenida por Willekens, Por y Raquillet (1979, p. 23) a través de la maximización de la dualidad de la función de entropía

$$\sum_{i,j,k} m_{ijk} \ln \frac{m_{ijk}}{m_{ijk}^{o}}$$

sujeta a los tres conjuntos de restricciones bivariadas que constituyen el estimador mínimo suficiente.

La solución para el problema dual de entropía es

$$m_{ijk} = m_{ijk}^{o} \exp - \left[1 + \lambda_{ij} + \nu_{ik} + \xi_{jk} \right] , \qquad (36)$$

con las variables duales λ_{ij}, ν_{ik} y ξ_{jk} asociadas con las restricciones $x_{ij.}$, $x_{i.k}$ y $x_{.jk}$, respectivamente. En la sección anterior (ver también Willekens, 1980) se demostraron algunas analogías entre el modelo log-lineal y la maximización de entropía. Las variables duales de la maximización de entropía se pueden expresar en términos de los parámetros u del modelo log-lineal y viceversa.

El algoritmo del ajuste multiproporcional es ligeramente diferente si está dada otra combinación de totales marginales: por ejemplo, considerando el caso 1FE con $x_{ij.}$ y $x_{..k}$ dados, el algoritmo entonces va de la siguiente manera:

Paso 0: $s = 0$

$$\hat{m}_{ijk}^{(o)} = m_{ijk}^{o} \text{ para todos los i,j,k } \epsilon \text{ S}$$

Paso 1: para cada valor k, ajústese la matriz (i,j) al total $x_{..k}$

$$\hat{m}_{ijk}^{(2s+1)} = \hat{m}_{ijk}^{(2s)} \frac{x_{..k}}{\sum_{i,j} \hat{m}_{ijk}^{(2s)}} \quad \text{para todos los k}$$

Paso 2: Ajústese a lo largo de la dimensión de edades k

$$\hat{m}_{ijk}^{(2s+2)} = \hat{m}_{ijk}^{(2s+1)} \frac{x_{ij.}}{\sum_k \hat{m}_{ijk}^{(2s+1)}} \qquad \text{para todos los i y j}$$

Si no se alcanza el criterio para detenerse, entonces s = s + 1 y vaya al paso 1.

Los estimadores son de máxima verosimilitud con efectos principales y el efecto de interacción (AB), derivado de los totales marginales y con los efectos (AC), (BC) y (ABC) "se toman prestados" de los estimadores preliminares, es decir,

$u_{ik}^{AC} = {}^{o}u_{ik}^{AC}$; $u_{jk}^{BC} = {}^{o}u_{jk}^{BC}$ y $u_{ijk}^{ABC} = {}^{o}u_{ijk}^{ABC}$. El modelo log-lineal (34) se reduce a

$$\ln \frac{m_{ijk}}{m^{o}_{ijk}} = {}^{r}u + {}^{r}u_{i}^{A} + {}^{r}u_{j}^{B} + {}^{r}u_{k}^{C} + {}^{r}u_{ij}^{AB}$$

o

$$m_{ijk} = m^{o}_{ijk} \; {}^{r}w \; {}^{r}w_{i}^{A} \; {}^{r}w_{j}^{B} \; {}^{r}w_{k}^{C} \; {}^{r}w_{ij}^{AB}$$

Análogamente, se puede derivar un algoritmo para el problema 3E. Este problema lo estudiaron Evans y Kirby (1974) durante un intento de generalizar modelos (gravitacionales) de interacción espacial desarrollados en la ciencia del transporte para inferir flujos de tráfico de un área a otra por tipos de productos. El modelo Evans-Kirby es como sigue: $m_{ijk} = r_i \, s_j \, p_k \, m^{o}_{ijk}$, con r_i, s_j y p_k como factores de balanceo. Este modelo no es más que el modelo log-lineal del caso 3E, con sólo los efectos principales presentes en los términos x.

Nótese que cualquier conjunto de datos preliminares que exhiba los mismos efectos de interacción no encontrados en los términos x (marginales), da los mismos estimadores finales (para una prueba formal, ver Bishop, Fienberg y Holland, 1975, p. 93). Por lo tanto, en el caso 3F, cualquier arreglo con valores u_{ijk}^{ABC} dados da los mismos estimadores. Esto no es sorprendente, ya que la única contribución del arreglo inicial a los estimadores finales es la interacción de segundo orden.

El algoritmo de ajuste multiproporcional tiene una larga historia. En la literatura estadística, frecuentemente está asociado con Deming y Stephan (1940) y es conocido como el procedimiento "clásico" de ajuste proporcional iterativo (Bishop, Fienberg y Holland, 1975, p. 84). Sin embargo, para tablas (matrices) bidimensionadas, este método fue desarrollado también en otras áreas de investigación científica para resolver problemas de estimación. De acuerdo con Murchland (1978), la primera aplicación de esta técnica de ajuste biproporcional fue durante la proyección del tráfico telefónico para evaluar las necesidades de expansión de la red y data de 1937 (Kruithof, 1937).

El algoritmo es sumamente utilizado en la ciencia del transporte, donde Fratar lo introdujo en 1954 en un intento por pronosticar la demanda de transporte personal de un área a otra. Desde entonces, la ciencia del transporte, en donde la técnica también se conoce como procedimiento Furness, elaboró este método, para poder estimar flujos de interacción espacial. La directriz de esta elaboración fue una reformulación para un modelo gravitacional (ver abajo). Los científicos del transporte también aplicaron el modelo gravitacional para inferir flujos de migración y más tarde, los analistas de migración adoptaron la técnica. Leontief (1940) y Stone (1962) hicieron un desarrollo adicional independiente del procedimiento de ajuste biproporcional, ambos en el campo de la economía. En el modelo de insumo-producto, la técnica se conoce como el método RAS, y está asociado con Stone, y como una técnica no de encuesta para estimar matrices de insumo-producto. Ésta ha sido investigada extensamente (Bacharach, 1970, MacGill, 1977). Sin embargo, no se ha ligado al modelo log-lineal de datos categóricos. Como consecuencia, la contribución de cada fuente de información previa a los estimadores no podría haber sido descifrada.

En la técnica RAS, el formato del modelo es como sigue (para una tabla bidimensionada):

$$m_{ij} = r_i \, s_j \, m^o_{ij}. \tag{37}$$

Los factores de balanceo r_i y s_j se deben determinar de los datos. Como la suma de totales renglón debe ser igual a la suma de totales columna, los factores de balanceo son únicos, hasta un valor escalar. Se pueden imaginar diversos procedimientos escalares. En las ilustraciones de la sección 5, postulamos que $s_1 = 1$. Stone (1962) sugirió el siguiente procedimiento iterativo:

Paso 0: $s\,(\text{paso}) = 0$

$$r_i^{(0)} = 1$$

Paso 1: $$s_j^{(2s+1)} = \frac{x_{.j}}{\sum_i r_i^{(2s)} \, m^o_{ij}} \qquad \text{para } j = 2, 3, \ldots C$$

Paso 2: $$r_i^{(2s+2)} = \frac{x_{i.}}{\sum_j s_j^{(2s+1)} \, m^o_{ij}} \qquad \text{para } i = 1, 2, \ldots R$$

Si no se alcanza el criterio para detenerse, entonces $s = s + 1$ y vaya al paso 1.

La técnica Fratar o Furness sigue el mismo algoritmo. Como ya se dijo, el problema de ajuste bi(multi)proporcional también se puede resolver con la maximización de una función de entropía. En el caso bidimensional, la solución entrópica es

$$m_{ij} = m^o_{ij} \exp - \left[1 + \lambda_1 + u_j\right], \tag{38}$$

donde λ_i y u_j son variables duales asociadas con las restricciones de renglón y de columna, respectivamente[3].

Los factores de balanceo de la técnica RAS y las variables duales o multiplicadores Lagrange del método de entropía están relacionados entre sí. Podemos escribir que:

$$\lambda_i = -(1 + \ln r_i)$$

$$u_j = -\ln s_j$$

Relacionando los multiplicadores y los factores de balanceo a los parámetros del modelo log-lineal, se les puede dar una interpretación estadística. Como

$$-\left[1 + \lambda_i + u_j\right] = {}^{r}u + {}^{r}u_i^A + {}^{r}u_j^B \quad ,$$

y como $\sum_j {}^{r}u_j^B = 0$, tenemos que

$$\sum_j -\left[1 + \lambda_i + u_j\right] = C\,{}^{r}u + C\,{}^{r}u_i^A \quad ,$$

y, consecuentemente, que

$$\begin{aligned} \lambda_i &= -\left[1 + \frac{1}{C}\sum_j u_j + {}^{r}u + {}^{r}u_i^A\right] \\ u_j &= -\left[1 + \frac{1}{R}\sum_i \lambda_i + {}^{r}u + {}^{r}u_j^B\right] \quad , \end{aligned} \tag{39}$$

siendo C y R, respectivamente, el número de columnas y renglones en la matriz. Expresando los parámetros del modelo log-lineal en términos de multiplicadores Lagrange, nos da que:

[3] A veces, la (38) se escribe como

$$m_{ij} = m_{ij}^o \exp\left[\hat{\lambda}_i + \hat{u}_j\right] \quad .$$

Ambas expresiones son equivalentes, dado que podemos definir

$$\hat{\lambda}_i = 1 - \lambda_i \text{ y } \hat{u}_j = -u_j$$

Mientras la (38) se sigue de la maximización de entropía

$$W = \sum_{i,j} m_{ij} \ln \frac{m_{ij}}{m_{ij}^o} \quad ,$$

la expresión de arriba se sigue de la minimización de la entropía negativa (−W), también conocida como ganancia de información o divergencia de información. Por lo tanto, también se demuestra la semejanza estructural entre el ajuste biproporcional y la minimización de información.

$$r_u = -\left[1 + \frac{1}{R}\sum_i \lambda_i + \frac{1}{C}\sum_j u_j\right]$$

$$r u_i^A = -\left[1 + \lambda_i + \frac{1}{C}\sum_j u_j\right] - r_u$$

$$r u_j^B = -\left[1 + \frac{1}{R}\sum_i \lambda_i + u_j\right] - r_u \qquad (40)$$

Las igualdades se mantienen para cualquier escalar aplicado a los factores de balanceo y a los multiplicadores Lagrange.

4.2.2 *Estimación de las celdas a partir de totales marginales, cuando algunos valores de las celdas son fijos*

En la sección anterior se supuso que no se conocía exactamente ninguno de los valores individuales x_{ijk}. En la práctica, sucede frecuentemente que algunas celdas son dadas. Un caso como éste es el conjunto de datos utilizados por Schoen y Woodrow (1980).

Recuérdese que S es el conjunto de celdas (i,j,k) que no son fijas. El tratamiento de (i,j,k) $\notin$ S es muy simple. Se construye un arreglo $\{m^o_{ijk}\}$ que contenga un cero en todas las celdas (i,j,k) que no están en S. Para construir $\{m^o_{ijk}\}$, primero se meten todos los ceros estructurales ($x_{ijk} = 0$) al arreglo de estimadores preliminares. Después, se sustraen los x_{ijk} diferentes de cero de los totales marginales asociados y se sacan de la tabla. Se mete un cero en las celdas apropiadas del arreglo m^o_{ijk}. A las celdas restantes se les da un valor de uno (o cualquier valor que exprese un efecto de interacción deseado). De este procedimiento, resultan los siguientes arreglos:

a) $m^o_{ijk} = 0$ para celdas (i,j, k) $\notin$ S

$m^o_{ijk} = 1$ para celdas (i,j,k) ϵ S

b) el arreglo $\{x_{ijk}\}$ contiene los totales marginales revisados.

Los valores esperados m_{ijk} se calculan por el algoritmo de ajuste multiproporcional. Sin embargo, sólo se toman en cuenta las celdas en S y, por lo tanto, los ceros estructurales se conservan en el procedimiento de estimación. La estructura de interacción impuesta en los estimadores es *casi independiente* (Bishop, Fienberg y Holland, 1975, p. 179). Casi independencia implica independencia entre clasificaciones cruzadas de variables, siempre que no se consideren renglones, columnas o estratos que tengan entradas de ceros estructurales en cuando menos una de las celdas. El modelo log-lineal de una tabla que contiene ceros estructurales está definido para las subtablas con celdas (i,j,k) ϵS; los parámetros se calculan en base a celdas diferentes de cero únicamente.

Este simple procedimiento puede abrir el camino para una nueva combinación de datos observados y estimados. Los valores de las celdas para categorías críticas se pueden obtener por una encuesta especial; mientras que las celdas restan-

tes se pueden estimar por otros medios. Hewings y Janson (1980) proponen tal procedimiento para el análisis de entradas y salidas.

4.2.3 *Construcción de arreglos de estimadores preliminares*

A través de un arreglo de estimadores preliminares, se pueden imponer en los estimadores finales, patrones de interacción entre clasificaciones cruzadas de variables, que no están contenidas en términos x. Cualquier arreglo $\{m_{ijk}\}$ que exhiba los patrones de interacción deseados da el mismo conjunto de estimadores finales.

En problemas de estimación, los patrones de interacción verdaderos, es decir, los del arreglo $\{x_{ijk}\}$, no son todos conocidos. En general, se desconocen los patrones de interacción de orden superior y sólo se pueden aproximar escogiendo un arreglo $\{m^{o}_{ijk}\}$ apropiado. Se pueden adoptar varias estrategias, de las que aquí sólo se mencionan algunas.

a) *Tabla anticuada*

Si el cambio estructural es menor, entonces el arreglo $\{m^{o}_{ijk}\}$ puede consistir en una tabla recolectada en una fecha previa. Dicha tabla puede ser ya anticuada, pero el patrón de interacción puede seguir siendo válido. Ajustando la tabla vieja a marginales nuevos, se puede mantener el patrón de interacción de orden superior. Generalmente, se sigue este procedimiento en el análisis de entradas y salidas, donde los estimadores preliminares se generan de una encuesta hecha en una fecha previa y los marginales nuevos se toman de los arreglos nacionales. Shulman y Chaddha (1978) y Shulman (1979) proponen el mismo procedimiento para inferir características detalladas de la población en un período intercensal, por la combinación de datos agregados actualizados y tabulaciones de censos anticuadas, pero detalladas. En el análisis multidimensional de población, los datos vienen frecuentemente de censos o encuestas especiales, hechas periódicamente, pero en intervalos largos. Se pueden generar conjuntos de datos adecuados con la actualización de la información de los censos, utilizando datos agregados recientes.

b) *Tabla de variables intermedias o explicativas*

Otro enfoque es la derivación de un patrón de interacción obtenido de una clasificación cruzada de variables intermedias o explicativas. En este caso, el arreglo de estimadores preliminares consiste en variables intermedias o explicativas y se supone que los efectos de interacción de orden superior en este arreglo se aplican al arreglo de estimadores finales. Este enfoque se usa mucho, aunque sólo implícitamente, en el análisis de interacción espacial. Los modelos de interacción espacial, que se desarrollaron en la ciencia del transporte y la ciencia regional, jamás se han considerado como técnicas discretas de análisis multivariado. Sólo hasta hace muy poco se descubrieron (Willekens, 1980) analogías con modelos de datos categóricos. Los modelos desagregados seleccionados para el análisis de interacción espacial, que cada vez son más populares, están también más relacionados con el análisis de datos cate-

góricos, en particular con los análisis del logito y del probito (McFadden, 1978).

En modelos de interacción espacial, el reto es obtener estimadores exactos de flujos de bienes (transporte, tráfico) o personas (migración) de un área a otra, en base a algunos datos agregados e información sobre medidas de "fricción", "impedancia" o "disuasión" espacial. Para derivar estos estimadores, los análisis han utilizado generalmente el modelo gravitacional. El origen del modelo gravitacional no fue estadístico, sino mecánico.

A fines del siglo XIX, científicos como Carey y Ravenstein declararon la analogía entre fenómenos físicos gobernados por la ley de gravedad de Newton e interacciones sociales. En 1948, Stewart formalizó la idea proponiendo que la interacción entre el lugar i y el j está directamente relacionada con el número de personas en ambos lugares y está inversamente relacionada con la distancia al cuadrado:

$$m_{ij} = g P_i P_j / d_{ij}^2, \tag{41}$$

donde g es una constante por ser determinada de los datos. Se ha intentado modificar el modelo gravitacional social sin lastimar su estructura fundamental. El modelo gravitacional es un modelo log-lineal y se puede escribir en los términos del modelo discutido anteriormente en este artículo:

$$m_{ij} = w\, w_i^A\, w_j^B\, m_{ij}^o, \tag{42}$$

con $w = g$

$$w_i^A = P_i \qquad w_j^B = P_j$$

$$m_{ij}^o = d_{ij}^{-2}$$

Esta transformación explica el papel que juega cada término en el modelo gravitacional. Las cifras de población generan los efectos principales. El término de distancia contribuye al efecto de interacción de primer orden en la matriz $\{ m_{ij} \}$. La distancia es un factor de fricción espacial, que inhibe la interacción. Las modificaciones del modelo gravitacional, hechas durante un intento por aumentar la exactitud de los estimadores, afectan la manera en que los efectos principales son medidos y la interacción de primer orden queda descrita. Se puede encontrar una revisión de este enfoque en Wilson (1970) y más recientemente en Hua y Porell (1979). Particularmente interesante a seguirse son los diseños de las funciones de distribución espacial $\{ m_{ij}^o \}$, normalmente representados como F_{ij}. March (1971) resume los principales tipos de funciones propuestos en la literatura. Algunas de estas funciones más comúnmente usadas son:

i función de potencia inversa

$$F_{ij} = d_{ij}^{-j} \tag{43}$$

ii función exponencial negativa

$$F_{ij} = \exp\left[-\delta d_{ij}\right] \tag{44}$$

iii función de Tanner

$$F_{ij} = \exp\left[-\delta_1 d_{ij}\right] d_{ij}^{-\delta 2}, \tag{45}$$

y, δ, δ_1 y δ_2 son parámetros por ser estimados (Batty y Mackie, 1972 proporcionan una extensa revisión de procedimientos de estimación). En lugar de distancia d_{ij}, se puede usar el costo de transporte de i a j, c_{ij}. Cualquiera que sea la función de distribución utilizada, su propósito principal es el mismo, es decir, obtener la mejor aproximación cuantitativa de los patrones de interacción de orden superior, suponiendo que están presentes en el arreglo que será estimado. La función de distribución afecta a los estimadores de una manera completamente análoga al arreglo de estimadores preliminares. Como resultado, la recopilación de investigación en el análisis de interacción espacial puede aplicarse fructíferamente al análisis de datos categóricos y viceversa.

5. Aplicaciones numéricas

Para ilustrar las técnicas presentadas en este artículo, se aplican a dos conjuntos de datos de movilidad. El primero consiste en datos de movilidad social agregados por edad, de Inglaterra y Dinamarca (cuadro 3). Esta simple aplicación permite poner la atención en los parámetros del modelo log-lineal y su relación con los factores de balanceo del método de ajuste biproporcional y con las variables duales del problema de maximización de entropía, sin pérdida en el volumen de números. El conjunto de datos ya ha sido investigado a fondo por Bishop, Fienberg y Holland (1975), con la perspectiva de probar la presencia de un patrón de interacción particular en los datos[4]. El objetivo del presente artículo es estimar los elementos faltantes. La comparación de los ejemplos trabajados en ambas perspectivas ilustra uno de los principales puntos de este artículo, es decir, la estimación de elementos faltantes en clasificaciones cruzadas de datos es equivalente a probar las hipótesis de independencia estadística. Una consideración adicional para seleccionar esta ilustración es que el estudio de patrones de movilidad social es un área en la que se pueden aplicar fructíferamente las técnicas de demografía multidimensional, si se puede aumentar una dimensión de edad a las tablas de movilidad social (las cifras de movilidad en el cuadro 3 son medidas de

4 Un análisis similar fue llevado a cabo por Hauser (1979) para datos de movilidad de Estados Unidos.

CUADRO 3

DATOS DE MOVILIDAD SOCIAL OBSERVADOS, DE DINAMARCA E INGLATERRA

A. Datos daneses

Status de los Padres	Status de los Hijos					Total
	1	2	3	4	5	
1	18	17	16	4	2	57
2	24	105	109	59	21	318
3	23	84	289	217	95	708
4	8	49	175	348	198	778
5	6	8	69	201	246	530
Total	79	263	658	829	562	2391

B. Datos británicos

Status de los Padres	Status de los Hijos					Total
	1	2	3	4	5	
1	50	45	8	18	8	129
2	28	174	84	154	55	495
3	11	78	110	223	96	518
4	14	150	185	714	447	1510
5	3	42	72	320	411	848
Total	106	489	459	1429	1017	3500

Fuente: Bishop, Fienberg y Holland, 1975, p. 100.

tiempo de vida). El segundo conjunto de datos consiste en flujos de migración específicos por edad, de Austria y Suecia. Ambos países publican flujos específicos por edad, por región de origen y de destino. Por lo tanto, el arreglo $\{x_{ijk}\}$ es completamente conocido. Willekens, Por y Raquillet (1979) utilizaron estos conjuntos de datos para probar la validez de los métodos de estimación que desarrollaron para inferir los datos de migración necesarios para el análisis multirregional de población. En este artículo se resumen los principales resultados.

5.1 *Cuadros bidimensionados de movilidad social*

El cuadro 4 muestra los valores de los parámetros del modelo log-lineal saturado

$$\ln m_{ij} = u + u_i^A + u_j^B + u_{ij}^{AB}$$

La variable A es el status de los padres; la B, es el status de los hijos. Nótese que los parámetros del modelo multiplicativo se pueden derivar fácilmente de los parámetros u, como se demuestra en el cuadro 1.

Los estimadores de parámetros se calculan por la fórmula demostrada en el cuadro 1, con el conjunto índice $K = \{1\}$, es decir, sólo se considera un estrato. Aunque los estimadores se podrían haber obtenido con la aplicación del paquete ECTA o GLM, se desarrolló un programa especial de cálculo. Este programa también calcula indicadores no dados por los programas estándar y permite la entrada flexible de información previa. El efecto total de u = 3.7831 se muestra en la esquina inferior derecha. La última columna contiene parámetros u_i^A, que representan el efecto de diferencias de tamaño en las categorías de status social de los padres (efectos renglón). El último renglón muestra los términos u_j^B, que miden los efectos de diferencias de tamaños en la categoría del status de los hijos. Los términos de interacción a_{ij}^{AB} constituyen los elementos de la matriz. Un término negativo u_{ij}^{AB} indica que hay menos hijos en el status j, suponiendo que el status de padres era i, que los que se podría esperar si el status de hijos y el de padres fueran independientes. Los parámetros demuestran que la mayoría de los hijos se queda en el status de sus padres y que, si cambian de status, generalmente el cambio es a un status adyacente. Nótese que $\sum_i u_i^A = \sum_j u_j^B = \sum_i u_{ij}^{AB} = \sum_j u_{ij}^{AB} = 0$, como lo requieren las especificaciones de restricción en el cuadro 1.

Los cuadros 5 y 6 presentan estimaciones de celdas bajo condiciones variantes de disponibilidad de información.

Para demostrar el impacto de datos incompletos en los patrones de interacción exhibidos por las estimaciones (arreglo $\{m_{ij}\}$), los cuadros también dan los parámetros del modelo log-lineal asociados con las diversas condiciones de disponibilidad de datos. Además, se demuestran los factores de balanceo del modelo de ajuste (técnica RAS) adoptados en el análisis de entradas y salidas y las variables duales o multiplicadores Lagrange del método de entropía.

El cuadro 5 demuestra que si la información previa está limitada a totales marginales únicamente, el efecto de interacción estará ausente en las estimaciones. Los valores esperados se pueden expresar de la siguiente manera:

CUADRO 4

PARÁMETROS DEL MODELO LONG-LINEAL SATURADO

A. Cuadro de movilidad social danesa

Status de los Padres	Status de los Hijos					Efecto Renglón
	1	2	3	4	5	
1	1.94	.92	-.07	-1.35	-1.46	-1.67
2	.42	.93	.04	- .47	- .92	.15
3	- .33	.00	.31	.13	- .11	.85
4	1.21	- .36	-.01	.78	.80	.67
5	- .82	-1.50	-.27	.90	1.69	.00
Efecto Columna	-1.17	- .21	.72	.62	.03	3.78

B. Cuadro de movilidad social británica

Status de los Padres	Status de los Hijos					Efecto Renglón
	1	2	3	4	5	
1	2.47	.62	-.83	-1.01	-1.24	-1.23
2	.45	.54	.08	- .31	- .76	.21
3	- .38	- .16	.46	.17	- .09	.10
4	- .96	- .33	.15	.51	.62	.92
5	-1.57	- .67	.14	.64	1.47	.00
Efecto Columna	-1.51	.23	-.04	.95	.37	4.18

CUADRO 5

CUADRO DE MOVILIDAD SOCIAL DANESA, ESTIMADO DE TOTALES MARGINALES SOLAMENTE ($m^{o}_{ij} = 1$ para todos i, j)

A. Estimadores

Status de los Padres	Status de los Hijos					Total
	1	2	3	4	5	
1	1.9	6.3	15.7	19.8	13.4	57
2	10.5	35.0	87.5	110.3	74.7	318
3	23.4	77.9	194.8	245.5	166.4	708
4	25.7	85.6	214.1	269.7	182.9	778
5	17.5	58.3	145.9	183.8	124.6	530
Total	79	263	658	829	562	2391

B. Parámetros del modelo log-lineal

Status de los Padres	Status de los Hijos					Efecto Renglón
	1	2	3	4	5	
1	.00	.00	.00	.00	.00	1.82
2	.00	.00	.00	.00	.00	- .10
3	.00	.00	.00	.00	.00	.70
4	.00	.00	.00	.00	.00	.80
5	.00	.00	.00	.00	.00	.41
Efecto Columna	-1.53	-.32	.59	.82	.44	3.98

C. Factores de balanceo (r_i, s_j) y multiplicadores Lagrange (λ_i, μ_j)

	r_i	λ_i	s_j	μ_j
1	1.88	-1.63	1.00	0.00
2	10.51	-3.35	3.33	-1.20
3	23.39	-4.15	8.33	-2.12
4	25.71	-4.25	10.49	-2.35
5	17.51	-3.86	7.11	-1.96

D. Bondad de ajuste

Chi-cuadrados Pearson: 754 para 16 grados de libertad

CUADRO 6

CUADRO DE MOVILIDAD SOCIAL DANESA, ESTIMADO DE TOTALES MARGINALES AUMENTADOS POR UN CUADRO DE MOVILIDAD SOCIAL BRITÁNICA

A. Estimadores

Status de los Padres	Status de los Hijos					Total
	A	B	C	D	E	
1	26.7	14.8	6.4	6.3	2.8	57
2	22.3	85.3	100.6	81.1	28.6	318
3	17.9	78.2	269.6	240.2	102.2	708
4	9.5	62.5	188.5	319.7	197.8	778
5	2.6	22.2	93.0	181.7	230.6	530
Total	79	263	658	829	562	2391

B. Parámetros del modelo log-lineal

Status de los Padres	Status de los Hijos					Efecto Renglón
	A	B	C	D	E	
1	2.47	.62	-.83	-1.01	-1.24	-1.68
2	.45	.54	.08	- .31	- .76	.16
3	- .38	-.16	.46	.17	- .09	.77
4	- .96	-.33	.15	.51	.62	.72
5	-1.57	-.67	.14	.64	1.47	.03
Efecto Columna	-1.32	-.07	.55	.72	.13	3.82

C. Factores de balanceo (r_i, s_j) y multiplicadores Lagrange (λ_i, μ_j)

	r_i	λ_i	s_j	μ_j
1	0.53	-0.37	1.00	0.00
2	0.80	-0.77	0.62	0.49
3	1.63	-1.49	1.50	-0.41
4	0.68	-0.61	0.66	-0.41
5	0.86	-0.85	0.65	0.43

D. Bondad de ajuste

Chi-cuadrado Pearson: 68

$$m_{ij} = \exp\left[u + u_i^A + u_j^B\right].$$

Por ejemplo,

$$m_{32} = \exp\left[3.9765 + 0.7030 - 0.3244\right] = 77.9$$

Los estimadores también se pueden expresar en términos de factores de balanceo: $m_{ij} = r_i s_j$. Por ejemplo, $m_{32} = 23.39 \times 3.33 = 77.9$, y en términos de multiplicadores Lagrange $m_{ij} = \exp -\left[1 + \lambda_i + u_j\right]$, lo que da

$$m_{32} = \exp -\left[1 - 4.15 - 1.20\right] = 77.9.$$

El cuadro 5 ilustra cómo aumenta la calidad de los estimadores cuando se añade una tabla de movilidad social de otro país el paquete de información previa. El valor de chi-cuadrado baja de 754 a 68. Comparando los cuadros 3 y 5, se observa que los datos británicos contribuyen a los parámetros de interacción de primer orden. El impacto de los datos británicos en el total de efectos renglón y columna se evalúa por comparación de los valores de los parámetros de los cuadros 4 y 5. Los valores esperados se pueden expresar en los siguientes términos:
– parámetros modelo log-lineales:

$$m_{ij} = m_{ij}^o \exp\left[{}^r u + {}^r u_i^A + {}^r u_j^B\right]$$

$$\text{ex.}: m_{32} = 78 \exp\left[(3.8189 - 4.1842) + (0.7712 - 0.1012) + (-0.0722 - 0.2300)\right]$$

$$= 78 \exp\left[-0.3653 + 0.6700 - 0.3022\right]$$

$$= 78 \exp\left(0.0025\right) = 78.2$$

– factores de balanceo:

$$m_{ij} = m_{ij}^o \; r_i \; s_j$$

$$\text{ex.}: m_{32} = 78 \cdot 1.6302 \cdot 0.6150 = 78.2$$

– multiplicadores Lagrange:

$$m_{ij} = m_{ij}^o \exp -\left[1 + \lambda_i + u_j\right]$$

$$\text{ex.}: m_{32} = 78 \exp -\left[1 - 1.4887 + 0.4861\right] = 78.2$$

5.2 *Cuadros tridimensionados de migración*

Los datos de migración requeridos para el análisis multirregional de población consisten en datos de flujos específicos por edad, por región de origen y de destino. Pocos países tabulan regularmente estos datos. En el contexto del proyecto del Estudio Comparativo de Migración y Establecimiento en la IIASA, cuyo objetivo era un estudio comparativo de patrones de distribución de población en los 17 países del "National Member Organization Countries" de la IIASA que aplican las técnicas de demografía multirregional, Willekens, Por y Raquillet (1979) desarrollaron, con la información disponible, una metodología para inferir los datos de migración necesarios. La metodología se basaba en el principio de maximización de entropía, pero se obtienen los mismos resultados que con la perspectiva adoptada en este artículo. Suponiendo diversas combinaciones de totales marginales (términos x), se obtuvieron, a través del algoritmo de ajuste multiproporcional y de procedimientos relativos, estimadores de flujos de migración detallados. Como en el análisis multirregional no se consideran flujos de migración intra-regional, se dejaron fuera del problema de estimación. El arreglo de estimadores preliminares era como sigue:

$$m^{o}_{ijk} = 1 \text{ para } i \neq j, \text{ para todo } k$$

$$m^{o}_{ijk} = 0 \text{ para } i = j, \text{ para todo } k.$$

El arreglo log-lineal implícito en los análisis es, por lo tanto, casi independiente.

Para probar la validez de las técnicas, se llevó a cabo un análisis de errores (bondad de ajuste) para los datos austriacos y suecos[5]. El número de grupos de edad en ambos conjuntos de datos era 18; el número de regiones en Austria era 4 y en Suecia, 8. El cuadro 7 resume los resultados principales. El caso 3F da estimaciones notablemente exactas. En el cuadro 8, se dan las estimaciones austriacas. Estas estimaciones no exhiben efecto de interacción de segundo orden (ver el caso 3F en el cuadro 2). La calidad de las estimaciones se puede explicar por la casi ausencia del efecto de interacción de segundo orden en los datos observados. Esto significaría que la interacción en forma de par entre el origen (A) y el destino (B) es la misma en cada grupo de edad (C); en otras palabras, el patrón de movilidad es relativamente independiente de la edad. Esto es completamente realista.

El análisis de errores del caso austriaco 3F revela que cerca de la mitad de las celdas fueron estimadas con menos de 4 por ciento de errores, casi dos tercios del volumen de migración tiene menos de 4 por ciento de errores de estimación. El cuadro 9 da resultados completos del análisis de errores. Una observación importante es que cerca del 60 por ciento del total absoluto del porcentaje de errores se debe a flujos de migración menores (menos de 200 migrantes) que representan solamente 11 por ciento del volumen de flujo (cuadro 9a). Se obtiene un

[5] El método (caso 3F) se aplicó realmente para estimar los datos de migración (faltantes para Bulgaria (Philipov, 1978), Holanda (Drewe y Willekens, 1980) y Bélgica (Willekens, 1977; Tan, 1980).

CUADRO 7

TABLAS DE MIGRACIÓN AUSTRIACA Y SUECA, ESTIMADAS DE DIVERSAS COMBINACIONES DE TOTALES MARGINALES: ANÁLISIS DE ERROR

Caso	Datos disponibles	Grados de libertad *	Chi-cuadrada Pearson		Promedio absoluto de error porcentual	
			Austria	Suecia	Austria	Suecia
3E	$\{x_{i..}, x_{.j.}, x_{..k}\}$	IJK - I - J - K + 2	18590	-	31.09	34.58
1FE	$\{x_{ij.}, x_{..k}\}$	(IJ - 1) (K - 1) - KI	3662	-	16.24	15.26
2F	$\{x_{ij.}, x_{.jk}\}$	(I - 1) (J - 1) K - KI	2006	-	12.08	11.90
3F	$\{x_{ij.}, x_{j.k}, x_{.jk}\}$	(I - 1) (J - 1) (K - 1) - KI	371	1262	4.27	6.32

* El número de grados de libertad asociado con el modelo ajustado se encuentra sustrayendo el número de parámetros independientes usado en el modelo del total de celdas a las que está ajustando el modelo. Para el modelo de casi-independencia con entradas diferentes de cero en las configuraciones marginales, el número de grados de libertad es igual al número de grados de libertad asociado con la tabla completa, menos el número de ceros estructurales. Para ambas, Austria y Suecia, K = 18; para Austria, I = J = 4 y para Suecia, I = J = 8.

Nota: el caso 3E no conserva los ceros estructurales.

CUADRO 8

FLUJOS DE MIGRACIÓN OBSERVADOS Y ESTIMADOS (3F) POR EDAD, AUSTRIA, CUATRO REGIONES, 1966-1971

		migración de este a					migración de sur a			
	Total	este	sur	norte	oeste	Total	este	sur	norte	oeste
0	1783.	0.	674.	874.	234.	1909.	882.	0.	556.	471.
		0-	-670-	-877-	-236-		-853-	0-	-575-	-481-
5	930.	0.	328.	482.	120.	1115	493.	0.	348.	274.
		0-	-328-	-468-	-134-		-530-	0-	-329-	-256-
10	1597.	0.	483.	821.	293.	3662	1371.	0.	1075.	1216.
		0-	-537-	-828-	-232-		-1342-	0-	-1044-	-1276-
15	4172.	0.	1351.	2029.	793.	8323	3800.	0.	2022.	2501.
		0-	-1280-	-2192-	-700-		-3760-	0-	-1950-	-2613-
20	4227.	0.	1336.	2246.	645.	4625	2097.	0.	1324.	1203.
		0-	-1289-	-2231-	-707-		-2081-	0-	-1381-	-1163-
25	2807.	0.	939.	1477.	392.	2625	1187.	0.	781.	656.
		0-	-910-	-1464-	-433-		-1225-	0-	-800-	-600-
30	1123.	0.	378.	596.	149.	1062	465.	0.	333.	263.
		0-	-368-	-588-	-167-		-464-	0-	-346-	-252-
35	915.	0.	295.	493.	127.	903	392.	0.	281.	230.
		0-	-293-	-480-	-142-		-408-	0-	-276-	-219-
40	871.	0.	280.	474.	117.	807	409.	0.	223.	174.
		0-	-312-	-438-	-121-		-411-	0-	-240-	-156-
45	579.	0.	204.	306.	69.	517	277.	0.	140.	100.
		0-	-222-	-289-	-68-		-269-	0-	-149-	-99-
50	618.	0.	229.	314.	75.	514	256.	0.	147.	111.
		0-	-241-	-312-	-65-		-263-	0-	-134-	-117-
55	689.	0.	261.	346.	81.	521	289.	0.	133.	99.
		0-	-280-	-331-	-78		-296-	0-	-132-	-93-
60	700.	0.	259.	373.	67.	455	246.	0.	133.	76.
		0-	-269-	-357-	-74-		-253-	0-	-137-	-65-
65	543.	0.	203.	290.	51.	340	185.	0.	100.	55.
		0-	-210-	-280-	-53-		-190-	0-	-102-	-48-
70	353.	0.	131.	189.	33.	217	118.	0.	64.	35.
		0-	-138-	-182-	-33-		-121-	0-	-65-	-31-
75	194.	0.	71.	105.	18.	118	63.	0.	35.	19.
		0-	-74-	-101-	-19-		-65-	0-	-36-	-17-
80	68.	0.	24.	37.	6.	39	22.	0.	11.	6.
		0-	-26-	-35-	-7-		-22-	0-	-12-	-5-
85	34.	0.	13.	18.	3.	21	11.	0.	7.	3.
		0-	-13-	-18-	-3-		-11-	0-	-7-	-3-
Total	22203.	0.	7460.	11471.	3272.	Total	27773	12564.	0.	7715.	7494.

	Total	migración de norte a este	sur	norte	oeste		Total	migración de oeste a este	sur	norte	oeste
0	1445.	764.	402.	0.	280.	0	905.	250.	352.	303.	0.
		-814-	-363-	0-	-268-			-229-	-395-	-281-	0-
5	917.	482.	251.	0.	184.	5	416.	111.	155.	150.	0.
		-448-	-282-	0-	-187-			-108-	-124-	-184-	0-
10	1568.	745.	371.	0.	452.	10	568.	148.	197.	222.	0.
		-771-	-344-	0-	-453-			-151-	-171-	-246-	0-
15	5297.	2888.	1107.	0.	1302.	15	2240.	736.	754.	750.	0.
		-2892-	-1123-	0-	-1282-			-772-	-809-	-659-	0-
20	3250.	1806.	734.	0.	710.	20	2235.	678.	736.	821.	0.
		-1861-	-701-	0-	-688-			-640-	-816-	-779-	0-
25	1885.	1029.	466.	0.	390.	25	1373.	395.	479.	499.	0.
		-998-	-482-	0-	-405-			-389-	-491-	-493-	0-
30	924.	492.	241.	0.	191.	30	606.	165.	216.	226.	0.
		-485-	-255-	0-	-184-			-173-	-212-	-221-	0-
35	750.	402.	187.	0.	162.	35	408.	113.	140.	155.	0.
		-379-	-213-	0-	-158-			-119-	-116-	-173-	0-
40	688.	419.	146.	0.	122.	40	361.	121.	113.	127.	0.
		-411-	-141-	0-	-136-			-128-	-87-	-146-	0-
45	447.	277.	101.	0.	68.	45	208.	71.	69.	69.	0.
		-275-	-102-	0-	-70-			-81-	-50-	-77-	0-
50	478.	273.	124.	0.	81.	50	241.	73.	88.	80.	0.
		-273-	-119-	0-	-86-			-65-	-81-	-95-	0-
55	489.	304.	114.	0.	71.	55	237.	82.	82.	72.	0.
		-295-	-114-	0-	-80-			-84-	-64-	-89-	0-
60	439.	271.	110.	0.	58.	60	185.	59.	64.	61.	0.
		-263-	-114-	0-	-62-			-60-	-51-	-74-	0-
65	328.	204.	83.	0.	41.	65	131.	42.	46.	43.	0.
		-198-	-84-	0-	-46-			-43-	-37-	-51-	0-
70	209.	130.	53.	0.	26.	70	82.	26.	29.	27.	0.
		-125-	-54-	0-	-30-			-28-	-21-	-33-	0-
75	109.	67.	27.	0.	14.	75	46.	15.	16.	16.	0.
		-66-	-27-	0-	-16-			-14-	-13-	-19-	0-
80	35.	22.	8.	0.	4.	80	14.	5.	5.	5.	0.
		-22-	-8-	0-	-5-			-5-	-3-	-6-	0.
85	19.	11.	6.	0.	2.	85	7.	2.	3.	2.	0.
		-11-	-6-	0-	-2-			-2-	-2-	-3-	0.
Total	19277.	10587.	4532.	0.	4158.	Total	10263.	3091.	3543.	3629.	0.

Fuente: Datos de migración observados: Sanberer (1981)
Estimaciones: Willekens, Por y Raquillet (1979, pp. 34-35)

Cuadro 9
Análisis (3f) de errores de estimaciones de migración, Austria

a. Análisis por tamaño de clase (volumen de flujo) y categoría migratoria

tamaño de clase	número de flujos* total	%	volumen de flujos total	%	cum. abs. valor	% error %	chi-cuadrada valor	%
0-200	112	51.85	8452.	10.63	1043.	68.56	0.9121e 02	33.70
200-400	45	20.83	12742.	16.02	241.	15.81	0.5711e 02	21.10
400-600	20	9.26	9481.	11.92	74.	4.87	0.2255e 02	8.33
600-800	11	5.09	7687.	9.67	73.	4.81	0.4191e 02	15.49
800-1000	9	4.17	7705.	9.69	36.	2.36	0.1924e 02	7.11
1000-1200	3	1.39	3330.	4.19	8.	0.52	0.2466e 01	0.91
1200-1400	7	3.24	9075.	11.41	25.	1.63	01295e 02	4.78
1400-1600	1	0.46	1464.	1.84	1.	0.06	01074e 00	0.04
1600-1800	0	0.00	0.	0.00	0.	0.00	0.0000e 00	0.00
1800-2000	2	0.93	3811.	4.79	7.	0.43	0.4201e 01	1.55
2000+	6	2.78	15769.	19.83	14.	0.95	0.1887e 02	6.97
total	216.	100.00	79516.	100.00	1522.	100.00	0.2706e 03	100.00

b. Análisis por categoría de error

categoría de error	porcentaje de error	número de flujos total	%	volumen de flujos flujos	%	flujo promedio
1	0 - 2	46.	21.30	24037.	30.23	522.543
2	2 - 4	57.	26.39	24756.	31.13	434.316
3	4 - 6	31.	14.35	13604.	17.11	438.839
4	6 - 8	18.	8.33	6463.	8.13	359.056
5	8 - 10	12.	5.56	4026.	5.06	335.500
6	10 - 15	28.	12.96	5021.	6.31	179.321

7	15 - 20	10.	4.63	798.	1.00	79.800
8	20 - 30	10.	4.63	650.	0.82	65.000
9	30 - 40	3.	1.39	158.	0.20	52.667
10	40 - 60	1.	0.46	3.	0.00	3.000
11	60 - 100	0.	0.00	0.	0.00	0.000
12	100 +	0.	0.00	0.	0.00	0.000
total		216.	100.00	79516.	100.00	368.130

promedio absoluto de porcentaje de errores = 4.27
(media relativa de desviación)

c. Análisis por tamaño de clase y categoría de error

Tamaño de clase	categoría de error												
	0-2	2-4	4-6	6-8	8-10	10-15	15-20	20-30	30-40	40-60	60-100	100+	total
0-200	21	23	13	8	4	20	10	9	3	1	0	0	112
200-400	9	12	10	5	3	5	0	1	0	0	0	0	45
400-600	6	9	0	2	2	1	0	0	0	0	0	0	20
600-800	1	2	4	0	2	2	0	0	0	0	0	0	11
800-1000	2	4	0	2	1	0	0	0	0	0	0	0	9
1000-1200	1	2	0	0	0	0	0	0	0	0	0	0	3
1200-1400	1	3	3	0	0	0	0	0	0	0	0	0	7
1400-1600	1	0	0	0	0	0	0	0	0	0	0	0	1
1600-1800	0	0	0	0	0	0	0	0	0	0	0	0	0
1800-2000	0	2	0	0	0	0	0	0	0	0	0	0	2
2000-	4	0	1	1	0	0	0	0	0	0	0	0	6
total	46	57	31	18	12	28	10	10	3	1	0	0	216

* Los flujos de ceros estructurales están excluidos.

patrón semejante si se usa la estadística de chi-cuadrado. Sin embargo, la distribución de errores es más explícita: los flujos menores son 34 por ciento del valor total de chi-cuadrado. La contribución de flujos menores al total de errores está mejor ilustrada por las clasificaciones cruzadas de las categorías de errores y los tamaños de clase de flujo (cuadro 9c). Los flujos menores están concentrados en las categorías de error más grandes. Bacharach (1970), Hewings (1977), Hinojosa (1978) y otros hicieron una observación semejante en un análisis de errores de coeficientes de entradas y salidas, estimado por el método RAS.

Del análisis de errores de flujos menores surgen dos problemas adicionales: la validez de las medidas de error utilizadas y el redondeo de estimaciones de flujo al valor entero más cercano.

La literatura estadística demuestra que se ha prestado atención a ambos problemas. Existen diversas sugerencias para sustituir la medida de chi-cuadrado en el caso de valores de celdas menores, así como sugerencias para ajustar las celdas menores (menos de 5, digamos).

Una solución pragmática a los problemas encontrados puede ser, forzar los flujos menores para que sean iguales a los estimadores preliminares. Su efecto en el resultado final sería insignificante y simplificaría la evaluación comparativa de los métodos de estimación. Hewings y Janson (1980, p. 847) proponen esta estrategia para la predicción de matrices de insumo-producto.

El análisis de errores de los datos de movilidad social y de los datos de migración muestra que un aumento de la información previa de mejores estimaciones. Sin embargo, la contribución de cada parte de la informacion previa no es igual. En el análisis de migración, por ejemplo, se podría observar que las estimaciones no mejoraron sustancialmente con el aumento del conocimiento de la estructura de edad de los inmigrantes (compárese el caso 2F con el 1F), mientras que la información del patrón total de migración era esencial (1FE contra 3E). En el análisis de movilidad social danés, el conocimiento de la tabla de movilidad social británica tuvo un impacto muy significativo en la calidad de las estimaciones.

Otros autores han experimentado observaciones semejantes. Snickars y Weibull (1977) compararon la capacidad descriptiva de cuatro modelos alternativos de distribución de viajes entre 12 regiones, en el condado de Estocolmo. En cada modelo se usaron volúmenes diferentes de información previa. Una observación interesante fue que el ajuste biproporcional de una matriz histórica de viaje, a marginales nuevos (método de Fratar) realizó el modelo gravitacional clásico con la función de distribución $\exp\left[-\beta c_{ij}\right]$ y el costo de transporte c_{ij}. La aplicación de una matriz histórica en lugar de una función de costo redujo la desviación de la media absoluta de un porcentaje de 20% a 7%, mientras que chi-cuadrado bajó de 731 a 107. El resultado indica que los patrones de viaje no se forman originalmente por diferenciales de costo de viaje. Por lo tanto, los efectos de interacción exhibidos por la función de distribución no son apropiados para describir el patrón de viaje. La conclusión de los autores respecto a que el modelo gravitacional tiene una capacidad descriptiva menor que el modelo de Fratar es, estrictamente hablando, incorrecta. No es la diferencia en la estructura del modelo lo que determina el resultado, sino la diferencia en los estimadores preliminares. La aplicación del modelo de Fratar con $m^{o}_{ij} = \exp\left[-\beta\, c_{ij}\right]$ daría resultados idénticos a los del modelo gravitacional. Como demuestran estos ejemplos, el valor de cada par-

te de información previa para propósitos de estimación se determina por la relevancia del efecto de iteración que exhibe y que impone a las estimaciones. El conocimiento previo no tiene ningún valor en sí mismo; sólo contribuye a través de los efectos de interacción que acarrea. Mientras más se asemejen los patrones de asociación en las estimaciones previas a aquéllos en los datos que serán estimados, mejores serán las estimaciones.

6. CONCLUSIÓN

La demografía multidimensional proporciona nuevas oportunidades para obtener una mejor comprensión demográfica. Sin embargo, sólo se pueden explorar completamente estas oportunidades si se tienen disponibles abundantes datos estadísticos o si se pueden aplicar métodos apropiados de estimación. Aunque los pasos seguidos por las oficinas estadísticas en todo el mundo para resolver el problema de datos son favorables, la falta de datos adecuados sigue siendo una gran desventaja para el análisis multidimensional de población. Por lo tanto, se necesitan métodos de estimación que se adapten a cualquier situación particular de datos. Este artículo sugiere una perspectiva unificada sobre técnicas de estimación para el análisis multidimensional con datos incompletos.

Un factor clave del enfoque unificado es su énfasis en las estructuras de datos y no en los valores de los elementos individuales de los datos. El conjunto de datos está enfocado como un sistema jerárquico interdependiente que puede ser introducido en un modelo. El modelo relaciona los valores que toman los elementos individuales de los datos con las características estructurales del sistema de datos; lo que es una ayuda para la exploración de estructuras de datos. En esta perspectiva, los elementos faltantes no son más que la expresión de nuestro conocimiento incompleto de la estructura de los sistemas y el problema para estimar exactamente los elementos faltantes es el de hipotetizar la estructura apropiada. Para ello, debe hacerse un uso óptimo de toda la información disponible en el sistema de datos o en el fenómeno o proceso al que representa. Las técnicas presentadas en este artículo tienen la intención de facilitar la formulación de hipótesis estructurales basadas en el conocimiento previo incompleto. Una importante ventaja de la perspectiva y las técnicas es que no son válidas solamente para conjuntos de datos convencionales (clasificaciones cruzadas de dos o tres variables), sino que se pueden aplicar igualmente a conjuntos de datos multidimensionales.

Para implementar la perspectiva unificada en la estimación de elementos faltantes, se sugiere una forma de arreglo. Las ventajas de los arreglos para el modelo y análisis multidimensional fueron discutidas por Rees (1980). Se distinguen dos tipos de arreglos. Uno contiene lo que se conoce de los datos de flujo reales. En general, el conocimiento previo está limitado a totales marginales, ceros estructurales y tal vez unos pocos elementos. El segundo arreglo contiene estimadores preliminares de los flujos reales. Los datos en este arreglo contribuyen a la calidad de los estimadores, imponiéndoles patrones de asociación, que son realistas, pero que no se pueden derivar de la información limitada en los flujos reales considerados. A través de los modelos log-lineales que describen a los dos arre-

glos (conjuntos de datos), se puede ver fácilmente cómo trabaja el mecanismo. La introducción de un arreglo de estimadores preliminares m^{o}_{ijk} es un compromiso entre la estimación de máxima verosimilitud bajo el modelo de independencia y la estimación de máxima verosimilitud que supone un modelo saturado.

A través del arreglo $\{m^{o}_{ijk}\}$ se pueden formular hipótesis estructurales apropiadas con respecto a los estimadores. En el artículo se muestra que estas hipótesis se pueden derivar de datos análogos de un país diferente pero similar, de datos estructurales o de tabulaciones cruzadas de variables intermedias o explicativas. El modelo gravitacional clásico para estimar la migración es una ilustración de este último caso.

El algoritmo de estimación que se presentó en el presente artículo es muy simple y no requiere calcular los parámetros del modelo log-lineal para poder determinar los valores esperados de las celdas.

Los valores de los flujos más probables dados, en el conocimiento previo limitado, se obtienen por el ajuste multiproporcional de los estimadores preliminares hasta que se ajustan exactamente a la información dada en los flujos reales (primer arreglo o fuente principal de datos). El ajuste multiproporcional no es el único método para inferir los estimadores requeridos. Otros algoritmos han sido discutidos en la literatura, pero son menos transparentes y, por lo tanto, menos apropiados si se quiere explorar las consecuencias para la estructura de datos y a través de ella, para estimadores individuales, si cambia el conocimiento disponible estadístico y real en el fenómeno o sistema. Este tipo de análisis exploratorio de datos, sin embargo, es crucial para determinar la contribución de cada parte de la información previa a los estimadores finales y para evaluar cuánta información se necesita realmente para asegurar que un conjunto de datos, requerido para la aplicación de las técnicas de la demografía multidimensional, puede ser estimado con un nivel de exactitud aceptable.

APÉNDICE A

Encuestas ocupacionales en la Comunidad Europea (EC)

por Albert Struyk.

En la Universidad de Tilburg, se inició un proyecto, financiado por el NPDR (Netherlands Programme for Demographic Research), para construir tablas de vida activa basadas en una encuesta por muestreo ocupacional, que desde 1973 fue organizada por la Oficina Estadística de la Comunidad Europea, en intervalos semestrales regulares y llevada a cabo –para Holanda– por la CBS, Oficina Estadística Nacional. El propósito de este trabajo es proveer un conjunto de datos, armonizados y comparables, sobre las principales características de empleo y desempleo en la comunidad.

Para poder facilitar un mejor uso e interpretación de los resultados de este tipo de encuestas, parece útil delinear las principales características metodológicas en esta contribución concisa.

La encuesta está organizada en base a propuestas de la Oficina Estadística de las Comunidades Europeas (OECE); dentro de la Encuesta Ocupacional por Muestreo, la parte de ocupación determina el contexto, la lista de preguntas y la codificación común de las respuestas individuales.

Los institutos estadísticos nacionales son responsables de seleccionar la muestra, preparar los cuestionarios, conducir las entrevistas de hogares y enviar los resultados a la OECE, de acuerdo con un esquema de codificación estándar.

La fecha de la encuesta está sincronizada de tal manera que siempre se lleva a cabo en primavera en todos los países. La fecha exacta en que se llevan a cabo, obviamente varía de país a país y es determinada por los institutos estadísticos nacionales, basados en la situación particular de cada país. La encuesta ha intentado cubrir al total de la población residente. Por razones técnicas y metodológicas, sin embargo, no es posible incluir a la población que vive en hogares colectivos. Consecuentemente, con el propósito de armonizar el campo de encuesta, los resultados de la comunidad se compilan en base a la población de hogares privados únicamente. La unidad estadística de la encuesta es el hogar.

La metodología de muestreo (tamaño de muestra, selección y muestreo de hogares, nivel de confiabilidad de los resultados, etc.) es determinada por los institutos estadísticos nacionales en base a las facilidades técnicas y administrativas de cada país. Para Holanda, la base de muestreo comprende los registros (listas de direcciones) para el censo de población y hogares de 1971, actualizado con las direcciones de viviendas construidas posteriormente. La unidad de la encuesta no es, por lo tanto, el hogar, sino la dirección. Cuando varias familias viven en una misma dirección, la encuesta las cubre a todas.

La muestra es subdividida en cinco estratos (regiones). En cada estrato se hace una muestra con una base proporcional, es decir, el número de personas incluidas en la muestra en cada comuna es proporcional al total de direcciones de la comuna en cuestión. La Encuesta Ocupacional de 1977 usó una muestra de 3% y cubrió aproximadamente 138 000 direcciones.

Los resultados son tratados en dos etapas: estimación de la población de referencia (universo) y cálculo de factores brutos. En Holanda la población total se estima en abril, excluyendo a las personas que viven en instituciones, marineros en el mar, y personas que viven en viviendas móviles (barcos, caravanas, etc.), basados en las estadísticas demográficas de 1970.

El esquema de codificación de la comunidad para la encuesta ocupacional comprende básicamente cinco partes:

1. Características principales de las personas entrevistadas (sexo, año de nacimiento, status marital, nacionalidad, región de residencia, relación con el (la) jefe del hogar);
2. Posición usual con respecto a la actividad económica;
3. Características ocupacionales;
4. Personas en busca de empleo;
5. Cambios en la situación comparados con el año anterior a la encuesta.

El propósito del punto 5 es determinar los principales cambios geográficos y ocupacionales en la población. En este caso, se usan preguntas retrospectivas, que incluyen preguntas similares, relativas al tiempo de la encuesta y a un punto previo en el tiempo (un año antes) a todas las personas entrevistadas. Sin embargo, las dificultades del registro y menor confiabilidad requieren cuidado, pero no impiden el propósito de armonizar la encuesta. A las personas que toman parte en la encuesta se les pide que declaren su situación un año antes a la encuesta, es decir:

a) Si regularmente están empleadas, desempleadas o inactivas;
b) Su status ocupacional previo y el sector y la rama de actividad de su ocupación regular a ese tiempo;
c) Si están fuera de su país; y si es así, en qué país;
d) Si, por el contrario, estaban en el país de la encuesta; y si era así, en qué región vivían.

La comparación de las respuestas a las preguntas sobre la situación habitual en el momento de la encuesta y la situación habitual un año antes, entonces, hace posible determinar:

—cambios en la situación, es decir, movimientos de desocupación a ocupación y viceversa;

—cambios en la actividad, es decir, posibles cambios del status ocupacional, sector y rama de actividad en el caso de personas empleadas regularmente en ambos tiempos, el de la encuesta y un año antes.

Con respecto a la movilidad geográfica, la encuesta puede determinar:

—personas que en el período considerado cambiaron su país de residencia y que vivían fuera del país un año antes;

—personas que cambiaron su región de residencia.

Es necesario enfatizar los límites de confiabilidad de estos datos, que obviamente se refieren a movilidad a niveles regionales dados y que, por lo tanto, no incluyen todos los movimientos de la población. En general, se puede decir que los resultados de la encuesta ocupacional están sujetos a errores que se pueden medir en términos de probabilidades, para poder determinar el grado de confiabilidad de los resultados. Sin embargo, proporciona estimadores suficientemente exactos para los niveles y estructuras de los diversos agregados en los que se divi-

de la ocupación, siempre que los análisis de este tipo se confinen a niveles de un cierto tamaño.

Se debería recomendar a la comunidad internacional que intentara llevar a cabo tales encuestas, enfatizando la importancia y necesidad de incluir, además de preguntas retrospectivas que consideren la movilidad geográfica y ocupacional, tópicos de nupcialidad y educación para poder usar completamente las herramientas que ofrece la demografía multidimensional.

BIBLIOGRAFÍA

Bacharach, M. (1970), Biproportional matrices and input-output analysis. London: Cambridge University Press.

Batty, M. and S. Mackie (1972), The calibration of gravity, entropy, and related models of spatial interaction. In: Environment and Planning, 4, pp. 205-233.

Birch, M. (1963), Maximum-likelihood in three-way contingency tables. In: Journal of the Royal Statistical Society, B 25, pp. 220-233.

Bishop, Y.M., S.E. Fienberg and P.W. Holland (1975), Discrete multivariate analysis: theory and practice. Cambridge, Mass.: M.I.T. Press.

Caussinus, H. and C. Thelot (1976), Note complémentaire sur l'analyse statistiques des migrations. (Further note on the statistical analysis of migrations). In: Annales de l'INSEE, 22-23, pp. 135-146.

Chilton, R. and R. Poet (1973), An entropy maximizing approach to the recovery of detailed migration patterns from aggregate census data. In: Environment and Planning A, 5, pp. 135-146.

Clogg, C.C. (1978), Adjustment of rates using multiplicative models. In: Demography, 15, pp. 523-539.

Clogg, C.C. (1980), Measuring underemployment. Demographic indications for the United States. New York: Academic Press.

Deming, W. and F. Stephan (1940), On a least square adjustment of a sampled frequency table when the expected marginal totals are known. In: Annals of Mathematical Statistics, 11, pp. 427-444.

Drewe, P. and F. Willekens (1980), Maximum likelihood estimation of age-specific migration flows in the Netherlands. In: Delft Progress Report, 5, pp. 92-111.

Evans, S.P. and H.R. Kirby (1974), A three-dimentional furness procedure for calibrating gravity models. In: Transportation Research, 8, pp. 105-122.

Fienberg, S.E. and W.M. Mason (1978), Identification and estimation of age-period-cohort models in the analysis of discrete archival data. In: K.F. Schuessler ed. Sociological Methodology 1979. San Francisco: Jossey-Bass Publishers, pp. 1-67.

Fisher, R.A. (1922), On the interpretation of Chi-square from contingency tables, and the calculation of P. In: Journal of the Royal Statistical Society, 85, pp. 87-94.

Fratar, T.J. (1954), Forecasting distribution of interzonal vehicular trips by successive approximation. In: Highway Research Board Proceedings. 33. pp. 376-385.

Gokhale, D. and S. Kullback (1978), The information in contingency tables. New York: Dekker.

Goodman, L. (1978), Analyzing qualitative/categorical data. Cambridge, Mass.: Abt Books, Abt Associates.

Haberman, S.J. (1979), Analysis of qualitative data (2 vols). New York: Academic Press.

Hauser, R.M. (1979), Some exploratory methods for modeling mobility tables and other cross-classified data. In: K.F. Schuessler ed. Sociological methodology 1980. San Francisco: Jossey-Bass Publishers, pp. 413-458.

Hewings, G.J.D. (1977), Evaluating the possibilities for exchanging regional input-output coefficients. In: Environment and Planning, A, 9, pp. 927-944.

Hewings, G.J.D. and B.N. Janson (1980), Excnanging regional input-output coefficients: a reply and further comments. In: Environment and Planning, A, 12, p. 843-854.

Hinojosa, R.C. (1978), A performance test of the biproportional adjustment of input-output coefficients. In: Environment and Planning, A, 10 pp. 1047-1052.

Hoem, J. and M. Fong (1976), A Markov chain model of working life tables. Copenhagen University, Laboratory of Actuarial Mathematics, Working Paper no. 2.

Hua, C. and F. Porell (1979), A critical review of the development of the gravity model. In: International Regional Science Review, 4, pp. 97-126.

Jong, P.M. de (1981), The reliability of methods for predicting missing figures in migration tables. Paper prepared for presentation at the Conference on the "Analysis of Multidimensional Contingency Tables", Rome, June 25-26, 1981.

Koesoebjono, S. (1981), Marital status life tables of female population in The Netherlands (1978); an application of the multidimensional demography. Working Paper no. 20, NIDI, Voorburg, The Netherlands.

Kruithof, J. (1937), Calculation of telephone traffic. In: De Ingenieur, 52, pp. E 15-E 25.
English translation by UK Post Office Research Department Library (no. 2663). London, Dollis Hill.

Leontief, W. (1941), The structure of the American economy, 1919-1939. New York, Oxford University Press.

Little, R.J.A. (1978), Generalized linear models for cross-classified data from the WFS. Technical Bulletin no. 5/Tech. 834, World Fertility Survey, London.

Little, R.J.A. (1980), Linear models from WFS data. Technical Bulletin no. 9/Tech. 1282P. World Fertility Survey, London.

Little, R.J.A. and T.W. Pullum (1979), The generalized linear model and direct standardization: a comparison. In: Sociological Methods and Research, 7, pp. 475-501.

Mac Gill, S.M. (1977), Theoretical properties of biproportional matrix adjustments. In: Environment and Planning, A 9, p. 687-701.

Mac Fadden, D. (1978), Modelling the choice of residential location. In: A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull eds. Spatial interaction theory and planning models. Amsterdam: North-Holland Publ. Co., pp. 75-96.

March, L. (1971), Urban systems: a generalised distribution function. In: A.G. Wilson ed. Urban and Regional Planning. (London Papers in Regional Science, vol. 2), London: Pion Ltd.

Murchland, J.D. (1978), Application, history and properties of bi- and multipro-

portional models. London: University College, Traffic Studies Group, JDM-292.

Nijkamp, P. (1979), Gravity and entropy models: the state of the art. In: G.R.M. Jansen, P.H.L. Bovy, J.P.J.M. Van Est and F. le Clerq. New developments in modelling travel demand and urban systems. Some results of recent Dutch research. Westmead, Farnborough, England: Saxon House, pp. 281-319.

Payne, C. (1977), The log-linear model of contingency tables. In: C.O. Muircheartaigh and C. Payne eds. The analysis of survey data. Vol. 2: Model fitting. New York: Wiley, pp. 105-144.

Philipov, D. (1978), Migration and settlement in Bulgaria. In: Environment and Planning, 10, p. 593-617.

Rees, P.H. and A.G. Wilson (1977), Spatial Population Analysis. London: Edward Arnold.

Rees, P.H. (1980), Multistate demographic accounts: measurement and estimation procedures. In: Environment and Planning A, 12, pp. 449-531.

Sauberer, M. (1981), Migration and settlement: Austria. Research Report, RR-81-00, IIASA, Laxenburg, Austria.

Schoen, R. and V. Nelson (1974), Marriage, divorce and mortality: a life table analysis. In: Demography, 12, pp. 313-324.

Schoen, R. and K. Woodrow (1980), Labor force status life tables for the United States, 1972. In: Demography, 17, pp. 297-322.

Shulman, S.A. (1979), Raking of state CPS labor force data. In: Proceedings of the 1979 Social Statistics Section, American Statistical Association, pp. 256-260.

Shulman, S.A. and R.L. Chaddha (1978), Updating 1970 census data on the race-sex-occupation distribution of a state. In: Proceedings of the 1978 Social Statistics Section, American Statistical Association, pp. 607-611.

Smith, S.J. (1980), Tables of working life for the United States, 1977: substantive and methodological implications. Paper presented at the annual meeting of the Population Association of America, Denver, Colorado, April 1980.

Snickars, F. and J.W. Weibull (1977), A minimum information principle. Theory and practice. In: Regional Science and Urban Economics, 7, pp. 137-168.

Stein, R.L. (1980), National Commission recommends changes in labor force statistics. In: Monthly Labor Review, April 1980, pp. 11-21.

Stewart, J.Q. (1948), Demographic gravitation: evidence and application. In: Sociometry, 1, pp. 31-58.

Stone, R. (1962), Multiple classifications in social accounting. In: Bulletin of the International Statistical Institute, 39, pp. 215-233.

Tan, E. (1980), On the estimation of migration flows by migrant categories. MA Thesis, Interuniversity Programme in Demography, Brussels.

Willekens, F. (1977), The recovery of detailed migration patterns from aggregate data: an entropy-maximizing approach. Research memorandum RM-77-58, IIASA, Laxenburg, Austria.

Willekens, F. (1980), Entropy, multiproportional adjustment and analysis of contingency tables. In: Systemi Urbani, 2 (nr. 2-3).

Willekens, F. (1980 b), Multistate analysis: tables of working life. In: Environment and Planning, 12, pp. 563-588.

Willekens, F., A. Por and R. Raquillet (1979), Entropy, multiproportional and quadratic techniques for inferring detailed migration patterns from

aggregate data. Mathematical theories, algorithms, applications and computer programs. Working Paper WP-79-88, IIASA, Laxenburg, Austria.

Willekens, F., I. Shah, J.M. Shah and P. Ramachandran (1980), Multistate analysis of marital status life tables. Theory and application. Working Paper no. 17, NIDI, Voorburg, Netherlands. (Revised version forthcoming in Population Studies).

Wilson, A.G. (1970), Entropy in urban and regional modelling. London: Pion Ltd.