UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS
DIVISION DE ESTUDIOS DE POSGRADO

CALIBRACION EN MUESTREO: UNA APLICACION A LA ENCUESTA NACIONAL DE INGRESOS Y GASTOS DE LOS HOGARES 1992 Y 1996

T E S I S

QUE PARA OBTENER
EL GRADO ACADEMICO DE

MAESTRA EN CIENCIAS
(MATEMATICAS)

P R E S E N T A :

MONICA TINAJERO BRAVO

DIRECTORA DE TESIS: GUILLERMINA ESLAVA GOMEZ

MEXICO, D. F. JUNIO 2000

Dedico este trabajo:

A mis papás,
por su inmenso amor.

A Brendis, Nando y Ricky,
por su cariño.

A mis amigos,
por serlo.

A Paco,
por su comprensión y ayuda.

¡Con todo mi amor!

Agradecimientos

A la profesora Guillermina Eslava por su valiosa ayuda y asesoría brindadas para la elaboración de la tesis. Gracias por todo el tiempo que me dedicó.

A los profesores Ignacio Méndez, Alberto Contreras, Carlos Díaz, Mogens Bladt, Patricia Romero y Salvador Zamora por el tiempo que invirtieron al revisar este trabajo y por sus importantes comentarios.

Al Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas por haberme proporcionado los medios necesarios para formarme en el área de la estadística y, especialmente, a los profesores del Instituto quienes nos compartieron sus conocimientos y experiencias.

A mis compañeros de la maestría Yazmín, Paty, Karim, Jaime, Oswaldo y José Luis por todos los momentos que vivimos.

¡Gracias UNAM!

INDICE

RESUMEN

En este trabajo se exponen los estimadores obtenidos mediante la calibración de los ponderadores de la encuesta, propuestos por Deville y Särndal (1992), y cuya principal finalidad es incrementar la precisión de los mismos. Los ponderadores calibrados son el resultado de ajustar los pesos dados por el diseño de muestreo con base en una medida de distancia, así como de información auxiliar proveniente de censos, registros administrativos e incluso de otras encuestas. Los nuevos ponderadores se obtienen minimizando la distancia entre ellos y los ponderadores originales, cumpliéndose al mismo tiempo las condiciones proporcionadas por la información externa. Lo anterior da lugar a un sistema de ecuaciones.

Pertenecen a este tipo de estimadores el de regresión, de razón y de posestratificación, que son casos especiales de la función de distancia de mínimos cuadrados. No obstante, en la literatura se han propuesto diversas funciones, cada una de ellas genera su propio factor de ajuste y posee tanto ventajas como desventajas, por ejemplo, la existencia de una solución al sistema de ecuaciones, pesos extremos, etc., que serán comentadas a lo largo del primer capítulo. Ya que es importante conocer la precisión de los estimadores, se presenta el método de Jackknife para calcular su varianza.

Con el objetivo de ilustrar y explorar la técnica empíricamente, se utiliza la Encuesta Nacional de Ingreso Gasto de los Hogares 1992 y 1996. Como fuentes de información auxiliar se usan el Censo de Población y Vivienda 1990, para calibrar los ponderadores de la ENIGH 1992; y el Conteo de Población 1995, para ajustar los pesos de la ENIGH 1996.

El ejercicio se hace en el contexto de tablas de contingencia, es decir, se aprovechan las frecuencias censales para algunas variables de infraestructura de las viviendas. En éste se comparan nueve medidas de distancia sugeridas en la literatura; la comparación se hace en términos de la media, rango y distribución de los factores de ajuste, así como también de las estimaciones y su varianza.

Adicionalmente, se comentan algunos problemas en cuanto a la implementación y se presenta el número de iteraciones requeridas por el algoritmo computacional, ya que como se mencionó al principio, para encontrar los factores de ajuste es necesario encontrar la solución a un sistema de ecuaciones, que en general es no lineal, y por ende se requiere de métodos iterativos. Los programas se elaboraron en MATLAB y se ejecutaron en una computadora personal.

INTRODUCCION

Las encuestas como medio para obtener información que permita explorar, analizar y comprender diversos aspectos de nuestro entorno son cada día más comunes y cobran mayor relevancia. En este sentido es deseable que las estimaciones que se deriven de una muestra sean muy cercanas a la realidad. Por ello, la teoría estadística, y en particular el muestreo, han desarrollado diversas técnicas de selección y estimación, que permiten que dichas estimaciones sean lo más confiable posibles y al menor costo.

En lo que corresponde a la estimación, una manera de mejorar los estimadores en términos de precisión es utilizar, adicionalmente a la información inherente a la muestra, información auxiliar; es decir, de otra fuente diferente a la encuesta de interés, por ejemplo, datos provenientes de censos, estadísticas vitales, registros administrativos, etc.

En esta tesis se presenta una clase de estimadores que utilizan información externa, a través de los denominados "ponderadores calibrados", los cuales son el resultado de ajustar los ponderadores originales de tal manera que se satisfagan las condiciones proporcionadas por la fuente externa, a la vez que la diferencia entre los nuevos factores de expansión y los dados por el diseño sea lo más pequeña posible, en términos de una medida de distancia. Casos particulares de esta familia de estimadores son el de razón, regresión y posestratificación.

El segundo propósito del presente trabajo es ilustrar la técnica anterior en el contexto de tablas de contingencia, utilizando la Encuesta Nacional de Ingresos y Gastos en los Hogares (ENIGH) 1992 y 1996, así como también algunas variables generadas por el Censo de Población y Vivienda 1990 y el Conteo de Población de 1995.

Para cumplir con los objetivos antes mencionados, el trabajo se dividió en tres capítulos. En el primero se presenta el planteamiento formal, se derivan los nuevos factores bajo diversas medidas de distancia, se exponen dos métodos para el cálculo de varianza y se muestra el algoritmo iterativo de Newton.

En el capítulo dos se proporciona una descripción de las fuentes de información. En el caso de la encuesta se señalan sus objetivos, metodología, diseño de muestreo y principales variables. En lo que se refiere al Censo y al Conteo de Población, brevemente se mencionan sus objetivos, alcances y principales temas que trata.

En el capítulo tres, se obtienen y analizan los resultados. En el primer apartado se presentan las variables que se utilizaron para calibrar y en el segundo se efectúa una comparación de los factores de ajuste, según diferentes funciones de distancia. Esta comparación se hizo en términos de sus características descriptivas, de la varianza de los estimadores y del trabajo de cómputo.

Finalmente se muestran las conclusiones y tres anexos. En el primero se plasman los algoritmos que se utilizaron para calcular los factores de ajuste, el segundo contiene una tabla de los principales rubros que integran el gasto y el ingreso y en el último se exhiben los histogramas de los nuevos factores.

Es importante señalar que aunque las conclusiones obtenidas se limitan al ejercicio que se llevó a cabo, se espera que sean de utilidad para trabajos futuros y para la puesta en marcha de nuevos sistemas de estimación.

CAPITULO 1
ESTIMADORES OBTENIDOS A PARTIR DE PESOS CALIBRADOS

1.1. INTRODUCCIÓN

Con el fin de mejorar las estimaciones obtenidas a partir de una encuesta es común que se utilice información auxiliar en la etapa de estimación, por ejemplo a través del uso de estimadores de razón, estimadores de regresión, etc. Dichos estimadores son el resultado de ajustar el ponderador original mediante un factor, y pertenecen a una familia más general conocida como estimadores obtenidos mediante la calibración de los pesos muestrales[1].

Los pesos muestrales, ponderadores o factores de expansión, se definen como el inverso de la probabilidad de que la unidad pertenezca a la muestra, y están ligados al esquema de muestreo utilizado. Por ejemplo, en el muestreo aleatorio simple, los factores de expansión están dados por la razón entre el número de elementos en la población y el tamaño de muestra, N/n , que en términos sencillos indica a cuántos elementos de la población representa cada unidad en la muestra. Análogamente, en el muestreo aleatorio simple estratificado los ponderadores para el estrato h, se obtienen mediante el cociente N_h/n_h. Por ende deben de ser considerados en la estimación de parámetros.

Los estimadores calibrados fueron desarrollados por Deville y Särndal (1992) quienes los definen como aquellos que usan "pesos calibrados", los cuales deben cumplir las condiciones siguientes:

i) Estar tan cerca como sea posible de los pesos muestrales originales, de acuerdo a una medida de distancia.

ii) Satisfacer un conjunto de restricciones. Estas se refieren a la información auxiliar (proveniente de censos, registros administrativos, estadísticas vitales, etc.), y en algunos casos, al intervalo en el que deben estar los nuevos ponderadores.

Resumiendo, emplear estos estimadores tiene como objetivo usar información auxiliar para ajustar los ponderadores, de tal manera que los pesos calibrados y los pesos originales estén lo más cerca posible, así como también que se cumplan las condiciones obtenidas de otra fuente.

Estos estimadores fueron ampliamente desarrollados por Deville y Särndal (1992). Sin embargo, uno de los primeros trabajos, en el caso de tablas de contingencia, corresponde a Deming y Stephan (1940), cuyo objetivo era estimar las frecuencias de las celdas de una tabla de contingencia de 2 ó 3 dimensiones. Deville (1988) hace una extensión de la idea de Lemel (1976) de modificar los pesos dados por el esquema de muestreo. Deville y Särndal (1992) propusieron una clase de medidas de distancia y posteriormente, Deville, Särndal y Sautory (1993), las aplican en el contexto de que la información auxiliar corresponda a los totales marginales de tablas de contingencia, utilizando la encuesta de condiciones de vida en Francia, 1990. Stukel, Hidiroglou y Särndal (1996) compararon varias medidas, simulando datos a partir de la Encuesta de Fuerza Laboral en Canadá, efectuada en 1990.

---

[1] En inglés son denominados *calibration estimators*.

Estevao, Hidiroglou y Särndal (1995) desarrollaron un sistema computacional en la oficina de Estadística de Canadá para la estimación de totales, razones de totales, promedios y proporciones, utilizando el estimador de regresión generalizado, que como se verá más adelante, es el originado por una de las medidas de distancia propuestas. Adicionalmente, Huang y Fuller (1978), y Singh y Mohl (1996) desarrollaron estimadores similares, que mantienen las propiedades mencionadas anteriormente, y los aplicaron a la Encuesta de Gasto Familiar en Canadá, 1990

Lundström y Särndal (1999) proponen este tipo de estimadores como una alternativa para reducir el sesgo por no respuesta. En general, la no respuesta no es el resultado de un mecanismo de selección aleatoria, motivo por el cual es necesario darle un tratamiento en la etapa de estimación. En la literatura se manejan dos clases de procedimientos: la imputación, es decir, reemplazar los valores faltantes por *proxys*; y la ponderación, es decir, modificar los pesos muestrales mediante un factor de ajuste. El método que proponen los autores pertenece a la segunda clase y tiene la ventaja que sólo requiere una etapa para calcular el factor de ajuste. Otros métodos se basan en dos fases: una relacionada con la información auxiliar y otra con el mecanismo de respuesta.

Antes de proseguir con la derivación general de los pesos calibrados se presentará un ejemplo tomado de Deming y Stephan (1940), en el que se persigue obtener una mejor estimación del número de casos que pertenecen a cada celda de una tabla de contingencia de dos dimensiones.

Para ello, supóngase que al tomar una muestra aleatoria simple sin reemplazo, se tienen las frecuencias muestrales que se exhiben en la tabla 1.1, de acuerdo a la edad del entrevistado y estado de residencia en los Estados Unidos. En realidad, las frecuencias fueron obtenidas artificialmente como una muestra del 5% de las personas blancas de ascendencia blanca, asistiendo a la escuela, *New England*, 1930.

Tabla 1.1. Frecuencias muestrales, n_{ij}, según estado de residencia y edad

		Grupo de edad				
Estado		*7 - 13*	*14 - 15*	*16 - 17*	*18 - 20*	*Total*
		j=1	*j=2*	*j=3*	*j=4*	
Maine	*i=1*	3,623	781	557	313	5,274
New Hampshire	*i=2*	1,570	395	251	155	2,371
Vermont	*i=3*	1,553	419	264	116	2,352
Massachusetts	*i=4*	10,538	2,455	1,706	1,160	15,859
Rhode Island	*i=5*	1,681	353	171	154	2,359
Connecticut	*i=6*	3,882	857	544	339	5,622
Total		22,847	5,260	3,493	2,237	33,837

Adicionalmente, para las dos variables anteriores, se dispone de la información censal que a continuación se muestra.

Tabla 1.2. Número de elementos en la población que pertenecen a cada estado, $N_{i.}$

Estado	*Número de elementos en la población*
Maine	105,040
New Hampshire	47,900
Vermont	48,640
Massachusetts	315,320
Rhode Island	46,600
Connecticut	113,240
Total	676,740

Tabla 1.3. Número de elementos en la población que pertenecen a cada grupo de edad, $N_{.j}$

Grupo de edad	*Número de elementos en la población*
7 - 13	457,540
14 - 15	105,700
16 - 17	69,240
18 - 20	44,260
Total	676,740

Dado el esquema de muestreo, el ponderador para todos los elementos en la muestra es N/n=20. Así, de no considerarse los datos censales, un estimador del número de individuos en la población que pertenece al estado i y al grupo de edad j está dado por la expresión (para las estimaciones ver tabla 1.4).

$$\hat{N}_{ij} = \frac{N}{n} n_{ij} .$$

Tabla 1.4. Número estimado de elementos de la población, $\hat{N}_{ij}$, según estado y grupo de edad

	Grupo de edad				
Estado	*7 - 13*	*14 - 15*	*16 - 17*	*18 – 20*	*Total*
Maine	72,460	15,620	11,140	6,260	105,480
New Hampshire	31,400	7,900	5,020	3,100	47,420
Vermont	31,060	8,380	5,280	2,320	47,040
Massachusetts	210,760	49,100	34,120	23,200	317,180
Rhode Island	33,620	7,060	3,420	3,080	47,180
Connecticut	77,640	17,140	10,880	6,780	112,440
Total	456,940	105,200	69,860	44,740	676,740

Entonces, el problema es encontrar factores de ajuste g_{ij}, tales que al multiplicarlos por las frecuencias estimadas y sumar por renglón o columna se tengan los totales censales, es decir,

$$\sum_j g_{ij}\hat{N}_{ij} = N_{i.}$$
$$\sum_i g_{ij}\hat{N}_{ij} = N_{.j}.$$

Si adicionalmente pedimos que las frecuencias ajustadas estén lo más cerca posible a las estimadas originalmente, y tomamos como distancia la de mínimos cuadrados, se obtienen las siguientes frecuencias ajustadas (Deming y Stephan, 1940).

$$\hat{M}_{ij} = g_{ij}\hat{N}_{ij} = \left(1 + r_i + c_j\right)\hat{N}_{ij},$$

donde r_i y c_j satisfacen el conjunto de ecuaciones:

$$\sum_j \left(r_i + c_j\right)\hat{N}_{ij} = N_{i.} - \hat{N}_{i.}$$
$$\sum_i \left(r_i + c_j\right)\hat{N}_{ij} = N_{.j} - \hat{N}_{.j}.$$

Finalmente, resolviendo el sistema de ecuaciones se obtienen las frecuencias calibradas que se presentan en la tabla 1.5.

Tabla 1.5. Número poblacional de elementos estimado por calibración, $\hat{M}_{ij}$, según estado y grupo de edad

Estado		*Grupo de edad*				
	c_j / r_i	*7 - 13*	*14 - 15*	*16 - 17*	*18 - 20*	*Total*
		0.0118	*0.0149*	*0.0012*	*0.0000*	
Maine	*-0.0146*	72,257	15,625	10,991	6,169	105,040
New Hampshire	*-0.0003*	31,761	8,015	5,025	3,099	47,900
Vermont	*0.0234*	32,153	8,701	5,410	2,374	48,640
Massachusetts	*-0.0162*	209,833	49,036	33,608	22,824	315,320
Rhode Island	*-0.0230*	33,243	7,003	3,345	3,009	46,600
Connecticut	*-0.0034*	78,292	17,337	10,856	6,757	113,240
Total		457,540	105,700	69,240	44,260	676,740

1.2. PLANTEAMIENTO DEL PROBLEMA

A continuación se expone el problema general de encontrar los nuevos ponderadores.

Notación:

$U=\{1,\ldots,k,\ldots,N\}$	Una población finita con N elementos.
$s\subseteq U$	Una muestra obtenida bajo un diseño de muestreo probabilístico dado.
$p(s)$	Probabilidad de que s sea seleccionada.
$\pi_k = P(k\in s)$	Probabilidad de que la unidad k pertenezca a la muestra.
$d_k = 1/\pi_k$	Peso de la unidad k asociado al diseño de muestreo.
y_k	Valor de la variable de interés, y, para el elemento k de la población.
$\mathbf{x}'_k=(x_{k1},\ldots,x_{kp})$	Vector auxiliar de valores para el elemento k.
$\mathbf{t}_x=\sum_U \mathbf{x}_k$	Total poblacional[2] para el vector $\mathbf{x}$, conocido de otra fuente.

El objetivo es estimar el total poblacional, $t_y=\sum_U y_k$, usando los totales de las variables auxiliares $\mathbf{t}_x$ para calibrar los pesos muestrales d_k.

Un estimador del total de y, que incorpora información de la etapa de diseño, es el propuesto por Horvitz y Thompson

$$\hat{t}_{y\pi}=\sum_s \frac{y_k}{\pi_k}=\sum_s d_k y_k \qquad (1)$$

Sin embargo, es deseable un estimador mejorado que también incorpore información en la etapa de estimación, para lo cual se construyen nuevos ponderadores o "pesos calibrados", denotados por w_k, bajo los cuales el nuevo estimador es de la forma

$$\hat{t}_{yw}=\sum_s w_k y_k \,. \qquad (2)$$

Los pesos serán construidos de tal forma que, en el sentido de una medida de distancia, los nuevos ponderadores $\{w_k : k\in s\}$ se encuentren lo más cerca posible de los ponderadores originales $\{d_k : k\in s\}$, en su caso, que pertenezcan a un intervalo preestablecido y que satisfagan las condiciones proporcionadas por la información auxiliar, es decir, que se cumplan las ecuaciones de calibración (3).

[2] Si $A\subseteq U$, para simplificar notación, $\Sigma_{k\in A}$ es equivalente a Σ_A.

$$\mathbf{t}_x = \sum_s w_k \mathbf{x}_k \,. \tag{3}$$

El problema de calibración corresponde a un problema de análisis numérico, se busca la minimización de una función multivariada sujeta a restricciones. La solución puede ser analítica pero la mayoría de los casos se encuentra numéricamente de forma iterativa.

1.3. MEDIDAS DE DISTANCIA

La medida de distancia es hasta cierto punto arbitraria, motivo por el cual se han propuesto varias alternativas, entre las que destacan las sugeridas por Deville y Särndal (1992), así como por Huang y Fuller (1978), y Singh y Mohl (1996). En este apartado se expondrá la solución general al enfoque manejado por Deville y Särndal, posteriormente se presentarán siete funciones estudiadas por ellos y finalmente dos medidas propuestas por los cuatro últimos.

Sea $F^*(w_k, d_k)$ una medida de distancia, tal que para d fijo cumple con las cuatro condiciones siguientes:

i) Es no negativa, diferenciable con respecto a w y estrictamente convexa.

ii) Está definida en un intervalo $D(d_k)$ que contiene a d.

iii) $F^*(d_k, d_k) = 0$.

iv) $f^*(w_k, d_k) = \dfrac{\partial F^*(w_k, d_k)}{\partial w_k}$ es una función continua y uno a uno.

De lo anterior se sigue que $f^*(w_k, d_k)$ es una función estrictamente creciente en w y $f^*(d_k, d_k) = 0$.

Deville y Särndal limitan su discusión a medidas del tipo $F^*(w_k, d_k) = {}^{d_k}\!/_{q_k}\, F(g_k)$, donde la razón entre el ponderador calibrado y el original, $g_k = w_k / d_k$, es denominada "factor-g" o factor de ajuste y $1/q_k$ es un peso positivo conocido no relacionado con d_k, en la práctica es común utilizar $q_k = 1$.

Así, la ecuación (2) puede reescribirse como

$$\hat{t}_{yw} = \sum_s d_k g_k y_k \,.$$

Se requiere minimizar la esperanza de la distancia, $E\left\{\sum_s F^*(w_k, d_k)\right\}$, sujeta a las restricciones especificadas por (3). Hay que observar que minimizar la esperanza anterior es equivalente a minimizar, para cualquier muestra, la cantidad (Deville y Särndal, 1992)

$$\sum_s F^*(w_k, d_k) = \sum_s \frac{d_k}{q_k} F(g_k), \qquad (4)$$

sujeta a $\mathbf{t}_x = \sum_s w_k \mathbf{x}_k$.

Utilizando multiplicadores de Lagrange para encontrar el mínimo de (4) se tiene

$$L(\mathbf{w}, \lambda) = \sum_s \frac{d_k}{q_k} F\left(\frac{w_k}{d_k}\right) - \left[\sum_s w_k \mathbf{x}'_k - \mathbf{t}'_x\right]\lambda$$

$$= \sum_s \frac{d_k}{q_k} F(g_k) - \left[\sum_s w_k \mathbf{x}'_k - \mathbf{t}'_x\right]\lambda$$

$$\Rightarrow \frac{\partial L(\mathbf{w}, \lambda)}{\partial w_k} = \frac{d_k}{q_k} f(g_k) \frac{1}{d_k} - \mathbf{x}'_k \lambda$$

$$= \frac{1}{q_k} f(g_k) - \mathbf{x}'_k \lambda$$

donde:

$$f(g_k) = \frac{\partial F(g_k)}{\partial g_k}$$

$$\lambda' = (\lambda_1, \ldots, \lambda_p).$$

Igualando a cero y despejando se tiene

$$f(g_k) = q_k \mathbf{x}'_k \lambda$$

$$\Rightarrow \; g_k = g(q_k \mathbf{x}'_k \lambda),$$

por lo tanto:

$$w_k = d_k g_k = d_k g(q_k \mathbf{x}'_k \lambda) \qquad (5)$$

donde $g(z)$ es la función inversa de $f(z)$, i.e., $g(z) = f^{-1}(z)$, que por iv en 1.3 está bien definida.

Finalmente, para calcular w_k se debe de obtener λ que cumpla con la ecuación

$$\sum_s w_k \mathbf{x}_k = \sum_s d_k g(q_k \mathbf{x}'_k \lambda) \mathbf{x}_k = \mathbf{t}_x . \qquad (6)$$

Tal sistema de ecuaciones con p variables y p incógnitas, dependiendo de la medida de distancia, posiblemente será no lineal. Su solución requiere de procedimientos iterativos, los cuales pueden clasificarse en dos familias:

i) Familia I. Pertenecen a ella los métodos que satisfacen las condiciones sobre los totales en cada iteración y estas continúan hasta que se cumplen las condiciones de tolerancia sobre el intervalo.

ii) Familia II. Está integrada por los procedimientos que satisfacen las condiciones del intervalo después de cada iteración y estas continúan hasta que se cumplen las condiciones de convergencia sobre los totales.

Es importante mencionar que si existe solución para el sistema de ecuaciones (6), las condiciones i)-iv) permiten que esta sea única. Así, para cada medida de distancia existe un conjunto de pesos calibrados y por lo tanto un estimador con base en ellos. En el caso de ciertas medidas de distancia, como las propuestas por Deville y Särndal (1992), en la literatura se proporcionan las expresiones para los ponderadores calibrados. En este trabajo, adicionalmente, se derivarán tales expresiones y se presentará el estimador correspondiente, suponiendo que el parámetro de interés es el total poblacional; así también se señalarán las ventajas y/o desventajas de cada método.

1.3.1. Distancia de mínimos cuadrados generalizados

Debido a la forma en que se modifican los ponderadores también se le denomina método lineal y es una de las medidas más conocida ya que da lugar al estimador de regresión generalizado, como se verá posteriormente. La función (Deville y Särndal, 1992) está definida por:

$$F^*(w_k, d_k) = \frac{(w_k - d_k)^2}{2 d_k q_k} = \frac{d_k}{2 q_k}\left(\frac{w_k}{d_k} - 1\right)^2 \qquad (7)$$

$$\Rightarrow \quad F(g_k) = \frac{1}{2}(g_k - 1)^2 .$$

Derivando con respecto a g_k

$$f(g_k) = (g_k - 1)$$
$$\Rightarrow g(z) = z + 1,$$

entonces los pesos calibrados tienen la forma

$$w_k = d_k g_k = d_k g(q_k \mathbf{x}'_k \lambda)$$
$$= d_k (q_k \mathbf{x}'_k \lambda + 1).$$

Sustituyendo en las ecuaciones de calibración se tiene

$$\mathbf{t}_x = \sum_s \mathbf{x}_k d_k (q_k \mathbf{x}'_k \lambda + 1)$$
$$= \sum_s d_k q_k \mathbf{x}_k \mathbf{x}'_k \lambda + \sum_s d_k \mathbf{x}_k$$

$$\Rightarrow \mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} = \sum_s d_k q_k \mathbf{x}_k \mathbf{x}'_k \lambda ,$$

donde:

$$\hat{\mathbf{t}}_{x\pi} = \sum_s \frac{1}{\pi_k} \mathbf{x}_k = \sum_s d_k \mathbf{x}_k .$$

Por lo tanto,

$$\lambda = \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right) \text{ donde } \mathbf{T}_s^{-1} = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1} .$$

Encontrando el estimador del total

$$\begin{aligned} \hat{t}_{yw} &= \sum_s w_k y_k = \sum_s d_k \left(q_k \mathbf{x}'_k \lambda + 1\right) y_k \\ &= \sum_s d_k y_k + \sum_s d_k q_k \mathbf{x}'_k \lambda y_k \\ &= \hat{t}_{y\pi} + \sum_s d_k q_k \mathbf{x}'_k \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right) y_k , \end{aligned}$$

ya que $\mathbf{T}_s = \mathbf{T}'_s$ y que el transpuesto de un escalar es igual al escalar

$$\begin{aligned} &= \hat{t}_{y\pi} + \sum_s d_k q_k \mathbf{t}'_x \mathbf{T}_s^{-1} \mathbf{x}_k y_k - \sum_s d_k q_k \hat{\mathbf{t}}'_{x\pi} \mathbf{T}_s^{-1} \mathbf{x}_k y_k \\ &= \hat{t}_{y\pi} + \mathbf{t}'_x \mathbf{T}_s^{-1} \sum_s d_k q_k \mathbf{x}_k y_k - \hat{\mathbf{t}}'_{x\pi} \mathbf{T}_s^{-1} \sum_s d_k q_k \mathbf{x}_k y_k . \end{aligned}$$

Finalmente, el estimador de t_y resultante es

$$\hat{t}_{yw} = \hat{t}_{yreg} = \sum_s w_k y_k = \hat{t}_{y\pi} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)' \hat{\mathbf{B}}_s$$

donde:

$\hat{t}_{y\pi} = \sum_s {}^{y_k}\!/_{\pi_k} = \sum_s d_k y_k$, y

$\hat{\mathbf{B}}_s = \mathbf{T}_s^{-1} \sum_s d_k q_k \mathbf{x}_k y_k$ es el estimador ponderado del coeficiente de regresión múltiple.

Como puede observarse, bajo esta distancia se genera el conocido estimador de regresión generalizado, el cual considera como casos especiales, al estimador de razón, al estimador de regresión simple y al estimador de posestratificación (ver siguiente sección).

Una ventaja de esta medida es que el sistema de ecuaciones para λ siempre tiene solución. Sin embargo, la función de distancia de mínimos cuadrados puede generar ponderadores negativos, los cuales en la práctica no son deseables porque es difícil darles una interpretación.

1.3.1.1. Derivación del estimador de razón

Sean:

$\mathbf{x}_k = x_k$ un escalar positivo

$q_k = 1/x_k$.

Entonces

$$\mathbf{t}_x = \sum_U \mathbf{x}_k = \sum_U x_k = t_x$$

$$\mathbf{t}_{x\pi} = \sum_s d_k \mathbf{x}_k = \sum_s d_k x_k = \hat{t}_{x\pi}$$

$$\mathbf{T}_s = \sum_s d_k q_k \mathbf{x}_k \mathbf{x}'_k$$

$$= \sum_s d_k \frac{1}{x_k} x_k x_k$$

$$= \sum_s d_k x_k = \hat{t}_{x\pi} ,$$

ya que

$$\lambda = \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)$$

$$= \frac{1}{\hat{t}_{x\pi}}\left(t_x - \hat{t}_{x\pi}\right),$$

encontrando el peso calibrado

$$w_k = d_k\left(q_k \mathbf{x}'_k \lambda + 1\right)$$

$$= d_k\left(\frac{t_x}{\hat{t}_{x\pi}} - 1 + 1\right)$$

$$= d_k \frac{t_x}{\hat{t}_{x\pi}} .$$

Por lo tanto, el estimador calibrado queda determinado por

$$\hat{t}_{yw} = \sum_s w_k y_k = \sum_s d_k \frac{t_x}{\hat{t}_{x\pi}} y_k$$

$$= \frac{t_x}{\hat{t}_{x\pi}} \sum_s d_k y_k$$

$$= \frac{t_x}{\hat{t}_{x\pi}} \hat{t}_{y\pi}$$ el cual corresponde al estimador de razón de y.

Arriba se obtuvo el estimador de razón para la distancia de mínimos cuadrados generalizados. Sin embargo, se obtiene el mismo estimador si se cumple que $q_k = 1/x_k$, $\mathbf{x}_k = x_k$ y la función es del tipo $f^*(w_k, d_k) = f\left(\frac{w_k}{d_k}\right)\frac{1}{q_k} = f(g_k)\frac{1}{q_k}$ (Deville y Särndal, 1992), como se verá enseguida.

Si $f^*(w_k, d_k) = f(g_k)\frac{1}{q_k}$

$\Rightarrow g_k = g(q_k \mathbf{x}'_k \lambda)$,

donde $g(z)$ es la función inversa de $f(z)$.

De la ecuación (6),

$$
\begin{aligned}
t_x &= \sum_s d_k g_k x_k \\
&= \sum_s d_k g(q_k x_k \lambda) x_k \\
&= \sum_s d_k g\left(\frac{1}{x_k} x_k \lambda\right) x_k \\
&= \sum_s d_k g(\lambda) x_k \\
\Rightarrow g(\lambda) &= \frac{t_x}{\hat{t}_{x\pi}}.
\end{aligned}
$$

Por lo tanto

$$
\begin{aligned}
\hat{t}_{yw} &= \sum_s w_k y_k \\
&= \sum_s d_k g(\lambda) y_k \\
&= g(\lambda) \sum_s d_k y_k \\
&= \frac{t_x}{\hat{t}_{x\pi}} \hat{t}_{y\pi}, \text{ que corresponde al estimador de razón.}
\end{aligned}
$$

1.3.1.2. Derivación del estimador de posestratificación

Dicho estimador corresponde al caso en que el vector de totales es el número de elementos que pertenecen a cada categoría de la variable de posestratificación, la derivada de la función de distancia es de la forma $f^*(w_k, d_k) = f(g_k)$ y $q_k = 1$, como se muestra a continuación.

Sean:

$\hat{t}_{y\pi h} = \sum_{k\in h} d_k y_k = \sum_h d_k y_k$, el estimador de y para el posestrato h

$\mathbf{t}'_x = (N_1,\ldots,N_h,\ldots,N_H)$

$\lambda' = (\lambda_1,\ldots,\lambda_h,\ldots,\lambda_H)$

$\mathbf{x}'_k = (0,\ldots,1,\ldots,0)$

$q_k = 1 \quad \forall k$

donde:

$$x_{kl} = \begin{cases} 1 & \text{si} \quad l = h \\ 0 & \text{en otro caso.} \end{cases}$$

Ya que $f^*(w_k, d_k) = f(g_k)$

$$\Rightarrow g_k = g(q_k \mathbf{x}'_k \lambda)$$
$$= g(\lambda_h).$$

De la ecuación (6),

$$\mathbf{t}_x = \sum_s d_k g(\lambda_h)\mathbf{x}_k$$
$$= \sum_{h=1}^{H} g(\lambda_h) \sum_{k\in h} d_k \mathbf{x}_k$$

$$\Rightarrow N_h = g(\lambda_h)\hat{N}_h$$

$$\Rightarrow g(\lambda_h) = \frac{N_h}{\hat{N}_h},$$

donde:

$$\hat{N}_h = \sum_{k\in h} d_k .$$

Finalmente, el estimador de posestratificación está dado por

$$\hat{t}_{yw} = \sum_s d_k g_k y_k$$
$$= \sum_s d_k \frac{N_h}{\hat{N}_h} y_k$$
$$= \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \sum_{k\in h} d_k y_k$$

$$= \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \hat{t}_{y\pi h} .$$

1.3.2. Distancia raking ratio

Debido a la forma que toman los pesos calibrados, a este método se le denomina multiplicativo. La función está definida por (Deville y Särndal, 1992):

$$F^*(w_k, d_k) = \frac{1}{q_k}\left\{ w_k \log\left(\frac{w_k}{d_k}\right) - w_k + d_k \right\} = \frac{d_k}{q_k}\left\{ \frac{w_k}{d_k} \log\left(\frac{w_k}{d_k}\right) - \frac{w_k}{d_k} + 1 \right\} \qquad (8)$$

$$\Rightarrow F(g_k) = g_k \log(g_k) - g_k + 1 .$$

Derivando con respecto a g_k

$$f(g_k) = g_k \frac{1}{g_k} + \log(g_k) - 1$$

$$= \log(g_k)$$

$$\Rightarrow g(z) = \exp(z)$$

$$\Rightarrow w_k = d_k g_k = d_k g(q_k \mathbf{x}'_k \lambda)$$

$$= d_k \exp(q_k \mathbf{x}'_k \lambda),$$

en donde λ debe de satisfacer el conjunto de ecuaciones

$$\sum_s w_k \mathbf{x}_k = \sum_s d_k \exp(q_k \mathbf{x}'_k \lambda)\mathbf{x}_k = \mathbf{t}_x .$$

Aunque con esta medida el sistema siempre tiene solución y los ponderadores modificados son positivos, éstos pueden tomar valores extremadamente grandes comparados con los originales.

1.3.3. Distancia de Hellinger

A continuación se muestra la función de Hellinger:

$$F^*(w_k, d_k) = \frac{2\left(\sqrt{w_k} - \sqrt{d_k}\right)^2}{q_k} = \frac{2d_k}{q_k}\left(\sqrt{\frac{w_k}{d_k}} - 1\right)^2 \qquad (9)$$

$$\Rightarrow F(g_k) = 2\left(\sqrt{g_k} - 1\right)^2 .$$

Derivando la expresión anterior con respecto a g_k se obtiene

$$f(g_k) = 4\left(\sqrt{g_k} - 1\right)\frac{1}{2} g_k^{-1/2}$$
$$= 2\left(1 - g_k^{-1/2}\right)$$

$$\Rightarrow g(z) = \left(1 - z/2\right)^{-2}$$

$$\Rightarrow w_k = d_k g_k = d_k g\left(q_k \mathbf{x}'_k \lambda\right)$$
$$= d_k \left(1 - \frac{q_k}{2} \mathbf{x}'_k \lambda\right)^{-2},$$

con λ que debe de cumplir

$$\sum_s w_k \mathbf{x}_k = \sum_s d_k \left(1 - \frac{q_k}{2} \mathbf{x}'_k \lambda\right)^{-2} \mathbf{x}_k = \mathbf{t}_x .$$

1.3.4. Distancia de entropía mínima

Esta se define por:

$$F^*\left(w_k, d_k\right) = \frac{1}{q_k}\left\{- d_k \log\left(\frac{w_k}{d_k}\right) + w_k - d_k\right\} = \frac{d_k}{q_k}\left\{- \log\left(\frac{w_k}{d_k}\right) + \frac{w_k}{d_k} - 1\right\} \qquad (10)$$
$$\Rightarrow F\left(g_k\right) = -\log\left(g_k\right) + g_k - 1 .$$

Derivando la función anterior

$$f\left(g_k\right) = -g_k^{-1} + 1$$
$$\Rightarrow g(z) = (1 - z)^{-1} .$$

Por ende,

$$w_k = d_k g_k = d_k g\left(q_k \mathbf{x}'_k \lambda\right)$$
$$= d_k \left(1 - q_k \mathbf{x}'_k \lambda\right)^{-1},$$

donde λ debe de satisfacer el conjunto de ecuaciones siguiente

$$\sum_s w_k \mathbf{x}_k = \sum_s d_k \left(1 - q_k \mathbf{x}'_k \lambda\right)^{-1} \mathbf{x}_k = \mathbf{t}_x .$$

1.3.5. Distancia de mínimos cuadrados modificada

Esta medida es muy similar a la de mínimos cuadrados generalizados (ver sección 1.3.1) y se define mediante:

$$F^*(w_k, d_k) = \frac{(w_k - d_k)^2}{2 w_k q_k} = \frac{d_k}{2 q_k}\left(\frac{w_k}{d_k}\right)^{-1}\left(\frac{w_k}{d_k} - 1\right)^2 \qquad (11)$$

$$\Rightarrow F(g_k) = \frac{1}{2}(g_k)^{-1}(g_k - 1)^2 .$$

Derivando con respecto a g_k

$$f(g_k) = \frac{1}{2}(g_k)^{-1} 2(g_k - 1) - \frac{1}{2}(g_k - 1)^2 (g_k)^{-2}$$

$$= \frac{1}{2}\left\{1 - (g_k)^{-2}\right\}$$

$$\Rightarrow g(z) = (1 - 2z)^{-1/2}$$

$$\Rightarrow w_k = d_k g_k = d_k g(q_k \mathbf{x}'_k \lambda)$$

$$= d_k (1 - 2 q_k \mathbf{x}'_k \lambda)^{-1/2},$$

donde λ es tal que

$$\sum_s w_k \mathbf{x}_k = \sum_s d_k (1 - 2 q_k \mathbf{x}'_k \lambda)^{-1/2} \mathbf{x}_k = \mathbf{t}_x .$$

Las distancias de Hellinger, entropía mínima y mínimos cuadrados modificada, no garantizan que el sistema de ecuaciones para λ tenga solución. Sin embargo, conforme el tamaño de muestra se incrementa, la probabilidad de que la tenga tiende a uno (Deville y Särndal, 1992). Además, aunque los pesos calibrados son positivos, pueden tomar valores extremos.

1.3.6. Distancia de mínimos cuadrados restringida

Como ya se mencionó, no es deseable la posibilidad de ponderadores negativos, como en el caso del método lineal, o bien de pesos muy grandes, como en el método multiplicativo. Por ello se han considerado funciones adicionales con la propiedad de generar ponderadores que se encuentren en un intervalo predeterminado. De esta manera se elimina el problema anterior y al mismo tiempo se mantienen las propiedades de los estimadores que utilizan pesos calibrados; no obstante, puede suceder que el sistema (6) no tenga solución.

Además de las restricciones sobre información auxiliar también es necesario imponer condiciones sobre el intervalo de los ponderadores, de tal manera que los factores g_k pertenezcan a un intervalo determinado, es decir, que se cumpla

$$L < g_k = \frac{w_k}{d_k} < U, \quad \text{con } L < 1 < U,$$

de esta manera, si L es positivo w_k también es positivo.

En el caso de mínimos cuadrados restringidos, la función de distancia es como sigue

$$F^*(w_k, d_k) = \begin{cases} \dfrac{(w_k - d_k)^2}{2 d_k q_k} & \text{si } L < \dfrac{w_k}{d_k} < U \\ \infty & \text{en otro caso.} \end{cases} \tag{12}$$

Así, al restringir los w_k obtenidos con la distancia de mínimos cuadrados generalizados, se tiene que

$$L < g_k = q_k \mathbf{x}'_k \lambda + 1 < U$$

$$\frac{L-1}{q_k} < \mathbf{x}'_k \lambda < \frac{U-1}{q_k}$$

$$\Rightarrow w_k = \begin{cases} d_k (q_k \mathbf{x}'_k \lambda + 1) & \text{si } \dfrac{L-1}{q_k} < \mathbf{x}'_k \lambda < \dfrac{U-1}{q_k} \\ d_k L & \text{si } \mathbf{x}'_k \lambda \le \dfrac{L-1}{q_k} \\ d_k U & \text{si } \mathbf{x}'_k \lambda \ge \dfrac{U-1}{q_k} \end{cases}$$

donde λ debe de satisfacer

$$\sum_s w_k \mathbf{x}_k = \sum_s d_k g(q_k \mathbf{x}'_k \lambda) \mathbf{x}_k = \mathbf{t}_x .$$

Si todos los ponderadores calibrados pertenecen al intervalo preestablecido $[L, U]$, se tiene el caso de la función de mínimos cuadrados generalizados (7).

1.3.7. Distancia logit

Otra función que evita generar pesos con valores no deseados es la denominada *logit*, la cual se define por:

$$F^*(w_k, d_k) = \begin{cases} \dfrac{d_k}{A q_k} \left[\left(\dfrac{w_k}{d_k} - L \right) \log \left(\dfrac{\dfrac{w_k}{d_k} - L}{1 - L} \right) + \left(U - \dfrac{w_k}{d_k} \right) \log \left(\dfrac{U - \dfrac{w_k}{d_k}}{U - 1} \right) \right] & \text{si } L < \dfrac{w_k}{d_k} < U \\ \infty & \text{en otro caso} \end{cases} \tag{13}$$

$$\Rightarrow F(g_k)=\frac{1}{A}\left[(g_k-L)\log\left(\frac{g_k-L}{1-L}\right)+(U-g_k)\log\left(\frac{U-g_k}{U-1}\right)\right] \text{ si } L<g_k<U,$$

donde:

$$A=\frac{U-L}{(1-L)(U-1)}.$$

Derivando con respecto a g_k

$$f(g_k)=\frac{1}{A}\left[(g_k-L)\left(\frac{1-L}{g_k-L}\right)\frac{1}{(1-L)}+\log\left(\frac{g_k-L}{1-L}\right)+\right.$$
$$\left.-(U-g_k)\left(\frac{U-1}{U-g_k}\right)\frac{1}{(U-1)}-\log\left(\frac{U-g_k}{U-1}\right)\right]$$

$$=\frac{1}{A}\left[\log\left(\frac{g_k-L}{1-L}\right)+\log\left(\frac{U-g_k}{U-1}\right)\right]$$
$$=\frac{1}{A}\left[\log\left(\frac{g_k-L}{U-g_k}\right)-\log\left(\frac{1-L}{U-1}\right)\right]$$

$$\Rightarrow g(z)=\frac{L+U\left(\frac{1-L}{U-1}\right)\exp(Az)}{1+\left(\frac{1-L}{U-1}\right)\exp(Az)}$$

$$=\frac{(U-1)L+(1-L)U\exp(Az)}{(U-1)+(1-L)\exp(Az)}$$

$$\Rightarrow w_k=d_k g_k=d_k g(q_k\mathbf{x}'_k\lambda)$$
$$=d_k\frac{(U-1)L+(1-L)U\exp(Aq_k\mathbf{x}'_k\lambda)}{(U-1)+(1-L)\exp(Aq_k\mathbf{x}'_k\lambda)}.$$

En donde λ debe de satisfacer:

$$\sum_s w_k\mathbf{x}_k=\sum_s d_k\frac{(U-1)L+(1-L)U\exp(Aq_k\mathbf{x}'_k\lambda)}{(U-1)+(1-L)\exp(Aq_k\mathbf{x}'_k\lambda)}\mathbf{x}_k=\mathbf{t}_x.$$

En el caso que L sea igual a cero y U muy grande se tiene la distancia *raking ratio* (8), como puede verse fácilmente.

Si $L=0,\quad U\to\infty$

$$\Rightarrow A \to 1$$

$$\Rightarrow lim_{U\to\infty} w_k = lim_{U\to\infty} d_k \frac{U \exp(q_k \mathbf{x}'_k \lambda)}{U + \exp(q_k \mathbf{x}'_k \lambda)}$$

$$= d_k \exp(q_k \mathbf{x}'_k \lambda).$$

1.3.8. Distancia de Huang-Fuller modificada

Corresponde a una pequeña modificación a la sugerida por Huang y Fuller en 1978, quienes propusieron un método para ajustar los pesos originados por la función de mínimos cuadrados generalizados tal que las ecuaciones de calibración (6) se satisfagan en cada iteración y los factores g_k sean cercanos a uno. Esta función pertenece a la familia I.

Singh y Mohl (1996) demostraron que el método puede escribirse en términos de minimizar una función de distancia, con cambios de iteración a iteración.

$$F^*\left(w_k^{(v)}, d_k\right) = \frac{\left(w_k^{(v)} - d_k\right)^2}{d_k a_k^{(v-1)^*}} \tag{14}$$

donde:

$a_k^{(v-1)^*} = a_k^{(v-1)} \cdots a_k^{(1)} a_k^{(0)}$ con $a_k^{(0)} = 1$

$v = 1,2,\ldots$ es el número de iteración.

$$a_k^{(v-1)} = \begin{cases} 1 & \text{si} \quad \xi_k^{(v-1)} < 0.5 \\ 1 - \beta\left(\xi_k^{(v-1)} - 0.5\right)^2 & \text{si} \quad 0.5 \le \xi_k^{(v-1)} < 1 \\ \dfrac{1 - \beta/4}{\xi_k^{(v-1)}} & \text{si} \quad \xi_k^{(v-1)} \ge 1 \end{cases}$$

$$\xi_k^{(v-1)} = \begin{cases} \dfrac{g_k^{(v-1)} - 1}{L' - 1} & \text{si} \quad g_k^{(v-1)} \le 1 \\ \dfrac{g_k^{(v-1)} - 1}{U' - 1} & \text{en otro caso} \end{cases}$$

con

$L' = \alpha L + 1 - \alpha$

$U' = \alpha U + 1 - \alpha$

$0 < \alpha, \beta < 1$ son elegidos arbitrariamente.

El modelo de ajuste para los ponderadores es lineal (como en el caso de regresión) y los factores de ajuste a_k $(0<a_k<1)$ se usan para facilitar que se cumplan las restricciones sobre el intervalo, de tal manera que los a_k hagan más pequeños los g_k que se localicen más alejados de los límites $[L,U]$. Por otra parte los parámetros α y β sirven para acelerar la convergencia del algoritmo. En el algoritmo el intervalo [*L, U*] es acortado a [*L', U'*] por medio de α. Esto implica que las unidades que están dentro de [*L, U*], pero cercanas a los límites del mismo, también son modificadas. Singh y Mohl (1996) probaron, empíricamente, diversos valores para esos parámetros, sugiriendo que $\alpha = 2/3$ y $\beta = 0.8$ funcionan bien en la práctica.

El factor g_k y el ponderador calibrado en cada iteración están dados, respectivamente, por:

$$g_k^{(v)} = 1 + a_k^{(v-1)^*}\mathbf{x}'_k\,\lambda^{(v)}$$
$$w_k^{(v)} = d_k\,g_k^{(v)}\,, \qquad v = 1,2,3,\ldots$$

donde:

$$\lambda^{(v)} = \left[\sum_s d_k a_k^{(v-1)^*}\mathbf{x}_k\mathbf{x}'_k\right]^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)$$
$$\hat{\mathbf{t}}_{x\pi} = \sum_{k\in s} d_k\mathbf{x}_k\,,$$

con valores iniciales $g_k^{(0)} = 1$ y $w_k^{(0)} = d_k$.

Las iteraciones continúan hasta que se cumple el criterio de convergencia para la restricción impuesta al intervalo de los g-factores.

1.3.9. Distancia de contracción[3]-minimización

Esta distancia, al igual que la distancia modificada de Huang-Fuller, fue propuesta por Singh y Mohl (1996) y pertenece a la familia I. El modelo para ajustar los ponderadores también es lineal, pero introduce un nuevo parámetro o factor de contracción, tal que al aplicárselo a los *g*-factores, éstos pertenezcan al intervalo $[L,U]$. Mediante un proceso iterativo se encuentran nuevos ponderadores que estén lo más cerca posible, en el sentido de la medida de distancia dada por (15), a los pesos de la iteración anterior.

$$F^*\left(w_k^{(v)}, w_k^{(v-1)^*}\right) = \frac{\left(w_k^{(v)} - w_k^{(v-1)^*}\right)^2}{w_k^{(v-1)^*}} \qquad (15)$$

donde:

[3] Del inglés *shrinkage*.

$$w_k^{(v-1)^*} = \begin{cases} L' d_k & \text{si} \quad w_k^{(v-1)} < L'' d_k \\ U' d_k & \text{si} \quad w_k^{(v-1)} > U'' d_k \\ w_k^{(v-1)} & \text{en otro caso} \end{cases}$$

$v = 1,2,3,\ldots$

$L' = \alpha L + 1 - \alpha$

$U' = \alpha U + 1 - \alpha$

$L'' = \eta L + 1 - \eta$

$U' = \eta U + 1 - \eta$

$0 < \alpha < \eta < 1$ son elegidos arbitrariamente.

Los parámetros α y η son elegidos arbitrariamente y sirven para acelerar la convergencia del algoritmo. Singh y Mohl sugieren $\alpha = 0.67$ y $\eta = 0.9$. Es importante notar que conforme α se acerca a 1, L' también se aproxima a L y U' a U. Lo mismo ocurre con η, L'' y U''. Además, de las condiciones impuestas, $L''<L'$ y $U'<U''$.

En cada iteración, el factor de ajuste y su correspondiente ponderador calibrado se obtienen mediante:

$$g_k^{(v)} = 1 + \mathbf{x}'_k \lambda^{(v)}$$

$$w_k^{(v)} = w_k^{(v-1)^*} g_k^{(v)}, \qquad v = 1,2,3,\ldots$$

donde:

$$\lambda^{(v)} = \left[\sum_s w_k^{(v-1)^*} \mathbf{x}_k \mathbf{x}'_k \right]^{-1} \left(\mathbf{t}_x - \hat{\mathbf{t}}_{xw}^{(v-1)^*} \right)$$

$$\hat{\mathbf{t}}_{xw}^{(v-1)^*} = \sum_s w_k^{(v-1)^*} \mathbf{x}_k$$

$$w_k^{(0)} = w_k^{(0)^*} = d_k .$$

El procedimiento continúa hasta que se cumple el criterio de convergencia para la restricción impuesta al intervalo de los *g*-factores.

El estimador resultante, al igual que el generado por la distancia Huang-Fuller, es asintóticamente equivalente al estimador de regresión.

Por otra parte, es importante notar que para las últimas dos funciones las ecuaciones o restricciones de calibración se cumplen en cada iteración y se continúa el algoritmo hasta que se logra la condición sobre el intervalo de los ponderadores modificados. Por otra parte, en las funciones 1.3.1 a 1.3.7 la restricción referente a los intervalos se satisface en cada iteración y el procedimiento se continúa hasta que se cumplen las ecuaciones de calibración.

Tabla 1.6. Funciones de distancia y pesos calibrados[a]

Distancia	$F^*(w_k, d_k)$	*Peso calibrado,* w_k
Mínimos cuadrados generalizados	$\frac{(w_k - d_k)^2}{2d_k q_k} = \frac{d_k}{2q_k}\left(\frac{w_k}{d_k} - 1\right)^2$	$d_k(q_k \mathbf{x}'_k \lambda + 1)$
Raking ratio	$\frac{1}{q_k}\left\{w_k \log\left(\frac{w_k}{d_k}\right) - w_k + d_k\right\}$	$d_k \exp(q_k \mathbf{x}'_k \lambda)$
Hellinger	$\frac{2\left(\sqrt{w_k} - \sqrt{d_k}\right)^2}{q_k}$	$d_k\left(1 - \frac{q_k}{2}\mathbf{x}'_k \lambda\right)^{-2}$
Entropía mínima	$\frac{1}{q_k}\left\{-d_k \log\left(\frac{w_k}{d_k}\right) + w_k - d_k\right\}$	$d_k(1 - q_k \mathbf{x}'_k \lambda)^{-1}$
Mínimos cuadrados modificada	$\frac{(w_k - d_k)^2}{2w_k q_k}$	$d_k(1 - 2q_k \mathbf{x}'_k \lambda)^{-1/2}$
Mínimos cuadrados restringida	$\begin{cases} \frac{(w_k - d_k)^2}{2d_k q_k} & \text{si } L < \frac{w_k}{d_k} < U \\ \infty & \text{en otro caso} \end{cases}$	$\begin{cases} d_k(q_k \mathbf{x}'_k \lambda + 1) & \text{si } \frac{L-1}{q_k} < \mathbf{x}'_k \lambda < \frac{L-1}{q_k} \\ d_k L & \text{si } \mathbf{x}'_k \lambda \le \frac{L-1}{q_k} \\ d_k U & \text{si } \mathbf{x}'_k \lambda \ge \frac{U-1}{q_k} \end{cases}$
Logit	$\frac{d_k}{Aq_k}\left[\left(\frac{w_k}{d_k} - L\right)\log\left(\frac{\frac{w_k}{d_k} - L}{1-L}\right) + \left(U - \frac{w_k}{d_k}\right)\log\left(\frac{U - \frac{w_k}{d_k}}{U-1}\right)\right]$	$d_k \frac{(U-1)L + (1-L)U\exp(Aq_k \mathbf{x}'_k \lambda)}{(U-1) + (1-L)\exp(Aq_k \mathbf{x}'_k \lambda)}$
Huang-Fuller modificada	$\frac{(w_k^{(v)} - d_k)^2}{d_k a_k^{(v-1)^*}}$	$d_k(1 + a_k^{(v-1)^*} \mathbf{x}'_k \lambda^{(v)})$
Contracción-minimización	$\frac{(w_k^{(v)} - w_k^{(v-1)^*})^2}{w_k^{(v-1)^*}}$	$w_k^{(v-1)^*}(1 + \mathbf{x}'_k \lambda^{(v)})$

[a] Las primeras siete distancias fueron propuestas por Deville y Särndal (1992) y las dos últimas por Singh y Mohl (1996). $A=(U-L)/[(1-L)(U-1)]$, $a_k^{(v-1)^*}$ es un factor de ajuste (ver sección 1.3.8), $w_k^{(v-1)^*}$ es el ponderador de la iteración anterior (ver sección 1.3.9).

A manera de resumen, en la tabla 1.6 se presentan las funciones de distancia así como sus respectivos ponderadores calibrados.

Es importante resaltar que desde un enfoque basado en el diseño[4], es decir, aquel que utiliza las probabilidades de selección para la inferencia a partir de la muestra y por lo tanto para determinar las propiedades del estimador[5], el estimador $\hat{t}_{yw}$ posee las siguientes propiedades, independientemente del método de calibración utilizado.

i) Es asintóticamente insesgado.

ii) Es consistente.

iii) Es asintóticamente equivalente al estimador de regresión $\hat{t}_{yreg}$ para cualquier función que cumpla las condiciones i) - iv) de la sección 1.3.

iv) Los ponderadores calibrados sólo dependen de la muestra a través de la(s) variable(s) x de los elementos que están en la misma, pero no dependen de la variable que se quiere estudiar y, lo cual implica que el conjunto de los w_k's puede ser usado para cualquier otra variable z que se desee estimar. Por supuesto que los beneficios serán mayores si la(s) variable(s) x están relacionadas con la variable z.

1.4. CÁLCULO DE LA VARIANZA DE LOS ESTIMADORES

Un requisito básico de casi todas las formas de análisis estadístico y por lo tanto del muestreo, es que se proporcione una medida de precisión para cada estimación derivada de la encuesta. La medida de precisión más común es la varianza del estimador. En general, ésta deberá ser estimada a partir de los datos.

La varianza de un estimador es una función de la forma del estimador y del diseño de muestreo. Así, para algunos diseños de muestreo básicos y algunos estimadores, existen fórmulas estándar para estimar de manera insesgada la varianza. No obstante, para las encuestas denominadas *complejas,* generalmente no es posible obtener un estimador insesgado de la varianza y por ende existen diferentes metodologías para estimarla de manera aproximada.

Aunque el término "encuestas de muestreo complejas" no se ha definido rigurosamente, y no hay un punto claro de separación entre una encuesta compleja y una simple, Wolter (1985) señala cinco características para distinguirlas:

[4] Del inglés *design-based approach*, otro enfoque sería el *model-based.*

[5] Las propiedades del estimador basado en el diseño son definidas en términos de su comportamiento sobre muestreo repetido, es decir, tomando la esperanza sobre el conjunto de todas las muestras posibles permitidas por el diseño.

i) El grado de complejidad del diseño de muestreo. Una encuesta compleja involucra características como estratificación, selección polietápica, diferentes probabilidades, múltiples marcos de muestreo, etc.

ii) El grado de complejidad del estimador(es). Una encuesta compleja genera estimadores no lineales, ajustes en la estimación por no respuesta o cobertura incompleta, estimadores de razón, de regresión, etc.

iii) Múltiples variables de interés. La mayoría de las encuestas complejas incluyen muchas características de interés.

iv) Usos descriptivos y analíticos de la encuesta. Varias de estas encuestas persiguen ambos objetivos.

v) El tamaño de la encuesta. Generalmente se involucra a cientos o miles de individuos que son entrevistados y un extenso trabajo de campo.

Así, entre los métodos de estimación de varianza utilizados en encuestas complejas se encuentran los siguientes:

a) Grupos aleatorios.

b) Muestreo por mitades balanceadas (*balanced half-sample*).

c) Jackknife.

d) Bootstrap.

e) Función generalizada de varianza (*generalized variance function*).

f) Linearización mediante series de Taylor.

En la elección del método de estimación de varianza se deberán considerar aspectos estadísticos como el sesgo, el error cuadrático medio, y aspectos administrativos como el tiempo, costo, simplicidad, etc.

En el caso que aquí nos ocupa, se requiere calcular la varianza del estimador para el total de y, $\hat{t}_{yw}$, obtenido utilizando los pesos calibrados w_k. Debido a que la varianza exacta de este estimador es difícil de obtener ya que éste es no lineal, además de que en la práctica es común el uso de diseños de muestreo complejos, se mostrará la varianza bajo dos métodos de los señalados anteriormente: el de linealización por series de Taylor y el Jackknife.

1.4.1. Método de linealización

En el caso del estimador de regresión generalizado, es decir, cuando los ponderadores calibrados están dados por $w_k = d_k\left(q_k \mathbf{x}'_k \lambda + 1\right)$, una aproximación para la varianza de $\hat{t}_{yreg}$ (ver sección 1.3.1), obtenida mediante el método de linealización por series de Taylor es (ver e.g. Särndal, Swensson y Wretman, 1989):

$$V\left(\hat{t}_{yreg}\right) = \sum_U \sum \left(\pi_{kl} - \pi_k \pi_l\right)\left(d_k E_k\right)\left(d_l E_l\right) \qquad (16)$$

donde:

$$E_k = y_k - \mathbf{x}'_k \mathbf{b}$$

$$\mathbf{b} = \left(\sum_U q_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \sum_U q_k \mathbf{x}_k y_k$$

E_k es el residuo para la unidad k

π_{kl} denota la probabilidad de inclusión de segundo orden, es decir, la probabilidad de que las unidades k y l pertenezcan a la muestra.

Un estimador de la varianza (16) está dado por (Särndal, Swensson y Wretman, 1989):

$$\hat{v}\left(\hat{t}_{yreg}\right) = \sum_s \sum \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \right) (w_k e_k)(w_l e_l) \qquad (17)$$

donde:

$$e_k = y_k - \mathbf{x}'_k \hat{\mathbf{b}}_s$$

$$\hat{\mathbf{b}}_s = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \sum_s d_k q_k \mathbf{x}_k y_k \ .$$

Deville y Särndal (1992) mostraron que cualquier función de distancia que obedezca el conjunto de condiciones generales i)-v) señaladas en la sección 1.3 producirá un estimador que es asintóticamente equivalente al de regresión generalizado. Así, para estas funciones, la varianza asintótica de $\hat{t}_{yw}$ es la misma que la del estimador de regresión.

$$\hat{v}\left(\hat{t}_{yw}\right) = \sum_s \sum \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \right) (w_k e_k)(w_l e_l) \qquad (18)$$

Este método es el único que permite estimar la contribución de las diversas etapas de muestreo en la varianza total (ver e. g. Shah, 1978). Desafortunadamente, este procedimiento para calcular la varianza enfrenta dos desventajas:

i) Se debe de disponer de las probabilidades de inclusión de segundo orden, π_{kl}, las cuales son difíciles de calcular en muchas ocasiones, o bien no son proporcionadas por la institución o persona encargada del diseño.

ii) Puede implicar gran esfuerzo computacional por la doble suma que involucra la ecuación 18, dependiendo del tamaño de muestra.

1.4.2. Método Jackknife

En este trabajo se utilizará un método alternativo para calcular la varianza de los estimadores, cuyo uso se ha incrementado en la práctica debido al desarrollo de los equipos de cómputo: el Jackknife. Este pertenece a los métodos de remuestreo, los cuales tienen la ventaja de que emplean una sola

fórmula para calcular el error estándar de cualquier estadística, $\hat{\theta}$. Además, el Jackknife tiene cierto apoyo teórico y ha revelado un buen comportamiento en diversos estudios empíricos.

Fue introducido por Quenouille en 1949 como un método para reducir el sesgo de un estimador del coeficiente de correlación serial, posteriormente, en 1956, el mismo autor generalizó la técnica en el contexto de poblaciones infinitas. En 1959, Durbin estudió su uso en estimadores de razón para poblaciones finitas. Actualmente, existe un gran número de investigaciones sobre las propiedades de la técnica, las cuales siguen dos líneas: su uso para reducir el sesgo y su uso para estimación de varianza.

El Jackknife deriva estimaciones de los parámetros de interés para cada una de las submuestras de la original y entonces calcula la varianza del estimador a partir de la variabilidad entre las estimaciones de las submuestras. Cada submuestra se obtiene al eliminar una de las unidades primarias de selección de la muestra total; así, se obtienen tantas submuestras como número de unidades primarias de muestreo.

En el caso de un diseño estratificado por conglomerados, en una o varias etapas, Wolter (1985) brinda un esquema para calcular la varianza por este método. Shao y Tu (1995) también dedican una parte de su libro al cálculo de la varianza. Sin embargo, Rao (1997) proporciona una manera más directa de efectuar el Jackknife, mediante un ajuste a los factores de expansión cada vez que se remueve una unidad, como se muestra a continuación.

i) Supóngase que se tienen H estratos y en cada estrato se seleccionaron n_h conglomerados. Se elimina el conglomerado l del estrato k.

ii) Defínanse los nuevos pesos

$$d^*_{hij} = \begin{cases} d_{hij} & (h \neq k) \\ d_{hij}\dfrac{n_h}{n_h - 1} & (h = k, i \neq l) \\ 0 & (h = k, i = l). \end{cases}$$

donde:

$h = 1,..,H$ es el subíndice para el estrato

$i = 1.,,.n_h$ corresponde al conglomerado

j se refiere a los elementos del conglomerado.

El primer caso corresponde a las unidades que no están en el estrato donde se está eliminando el conglomerado, el segundo a las unidades en el estrato donde está eliminando y el tercero al conglomerado que se está quitando.

iii) Calculénse los g-factores, g_{hij}, utilizando los factores d^*_{hij}, de acuerdo a la función de distancia elegida. En el caso de que no se ajusten los ponderadores con información auxiliar g_{hij}=1.

iv) Obténgase los nuevos factores de expansión $w^*_{hij} = d^*_{hij} g_{hij}$.

v) Obténgase $\hat{\theta}_{(kl)}$ de la misma forma que $\hat{\theta}$, quitando el conglomerado l del estrato k y utilizando los nuevos factores de expansión.

Finalmente, la varianza de $\hat{\theta}$ se estima mediante:

$$\hat{v}_J(\hat{\theta}) = \sum_h \frac{n_h - 1}{n_h} \sum_l \left(\hat{\theta}_{(hl)} - \hat{\theta}\right)^2 . \qquad (19)$$

En el caso de que el parámetro a estimar sea un total, como es de interés en este trabajo, la fórmula (19) es equivalente a:

$$\hat{v}_J(\hat{t}_{yw}) = \sum_h \frac{n_h - 1}{n_h} \sum_l \left(\hat{t}_{yw(hl)} - \hat{t}_{yw}\right)^2 \qquad (20)$$

donde:

$$\hat{t}_{yw(hl)} = \sum_s w^*_{hik} y_{hik} .$$

Para estadísticas suaves y sin posestratificación, $\hat{\theta}$, se ha establecido la consistencia asintótica de la varianza del estimador Jackknife, conforme el número de estratos crece. Por otra parte, es importante mencionar que el estimador dado por Wolter (1985) y éste dado por Rao (1997), no son iguales.

1.5. CALIBRACIÓN CUANDO SE CONOCEN LOS TOTALES DE UNA TABLA DE CONTINGENCIA.

Una aplicación importante de esta técnica es cuando se conocen las frecuencias de una tabla de contingencias (ya sea marginales o de las celdas). En este sentido se pueden señalar dos tipos de calibración:

i) Calibración cuando se conocen las frecuencias por celda, conocida como "posestratificación completa".

ii) Calibración cuando se conocen solamente las frecuencias marginales de la tabla o "posestratificación incompleta".

A continuación se describirán estos procesos en el caso de una tabla de contingencias de dos dimensiones, y suponiendo que el peso q_k, no relacionado con los factores de expansión originales, es igual a la unidad para todos los individuos en la muestra.

1.5.1. Posestratificación completa

Como se mencionó anteriormente, en este caso se conoce el número de elementos de cada celda en la población. El estimador bajo este método es igual al conocido estimador de posestratificación (ver e.g. Cochcran, 1977) como se observa a continuación, y de ahí se deriva su nombre.

Sean:

Una tabla de contingencia con r renglones y c columnas, por lo tanto se tienen rc celdas. Por ejemplo, cuando los elementos de la población son clasificados de acuerdo a grupos de edad y sexo.

Los elementos de la población se pueden clasificar en una y sólo una de las celdas denotadas por U_{ij}.

N_{ij} es el número de elementos que contiene cada celda U_{ij}, $i = 1,..,r$; $j = 1,..,c$;

$\Rightarrow \sum_i \sum_j N_{ij} = N$.

N_{ij} son conocidos.

s_{ij} denota el conjunto de elementos en la muestra que pertenecen a la celda U_{ij}.

n_{ij} es el número de elementos en la muestra que pertecenecen a la celda U_{ij}.

$q_k = 1 \quad \forall k$.

En este caso el vector $\mathbf{x}_k$ está compuesto de r x c entradas en donde una entrada es igual a 1 y las demás a cero, es decir,

$$x_{kij} = \begin{cases} 1 & \text{si } k \in U_{ij} \\ 0 & \text{en otro caso.} \end{cases}$$

Entonces, el vector de totales poblacionales está dado por

$$\mathbf{t}'_x = \sum_U \mathbf{x}'_k = \left(N_{11}, N_{12}, \ldots, N_{1c}, N_{21}, \ldots, N_{rc}\right),$$

y el vector de multiplicadores de Lagrange tiene la forma

$$\lambda' = \left(\lambda_{11}, \ldots, \lambda_{rc}\right).$$

Por lo tanto,

$$g_k = g\left(\mathbf{x}'_k \lambda\right) = g\left(\lambda_{ij}\right),$$

es decir, el factor es constante para todos los individuos que pertenecen a la celda U_{ij}.

Sustituyendo en las ecuaciones de calibración:

$$\mathbf{t}_x = \sum_s d_k g(\lambda_{ij}) \mathbf{x}_k,$$

$$\Rightarrow N_{ij} = \sum_{s_{ij}} d_k g(\lambda_{ij})$$

$$= g(\lambda_{ij}) \sum_{s_{ij}} d_k ,$$

ya que $\hat{N}_{ij} = \sum_{s_{ij}} d_k$

$$\Rightarrow N_{ij} = g(\lambda_{ij}) \hat{N}_{ij},$$

despejando

$$g(\lambda_{ij}) = \frac{N_{ij}}{\hat{N}_{ij}}.$$

Los nuevos factores de expansión están dados por:

$$\begin{aligned} w_k &= d_k g_k \\ &= d_k g\left(\lambda_{ij}\right) \\ &= d_k \frac{N_{ij}}{\hat{N}_{ij}}. \end{aligned}$$

Por lo tanto,

$$\begin{aligned} \hat{t}_{yw} &= \sum_s w_k y_k \\ &= \sum_s d_k \frac{N_{ij}}{\hat{N}_{ij}} y_k \\ &= \sum_i \sum_j N_{ij} \tilde{y}_{sij} \end{aligned}$$

donde:

$\tilde{y}_{sij} = \dfrac{\sum_{s_{ij}} d_k y_k}{\hat{N}_{ij}}$ es el estimador ponderado de la media muestral para la celda U_{ij}.

Como se puede observar el estimador anterior toma la forma del conocido posestratificado, con las celdas como posestratos.

1.5.2. Posestratificación incompleta

En este caso la información auxiliar es menos detallada en comparación con la posestratificación completa, ya que se tienen los totales marginales y no la información sobre cada una de las celdas.

Desde el punto de vista teórico, este método es interesante porque las funciones de distancia propuestas por Deville y Särndal (1992) llevan a una clase de estimadores conocidas como *generalized raking procedures*, Deville, Särndal y Sautory (1993).

Desde el punto de vista práctico, la calibración basada en totales marginales presenta las siguientes ventajas:

i) No es necesario conocer las frecuencias de cada celda.

ii) Los dos conjuntos de frecuencias marginales pueden provenir de fuentes diferentes.

iii) Cuando las celdas muestrales están vacías o tienen muy pocas observaciones, el estimador de posestratificación completa puede ser inestable o indefinido, siendo más estable el estimador basado en conteos marginales.

iv) La información auxiliar se puede obtener de otra encuesta independiente, que sea lo suficientemente grande como para generar estimaciones precisas de los totales marginales.

v) Utilizando un muestreo estratificado se pueden preservar las ventajas de éste y los beneficios de la posestratificación.

Sean:

Una tabla cruzada con r renglones y c columnas.

$N_{i.}$ el total marginal para el renglón i, i=1,...,r.

$N_{.j}$ el total marginal para la columna j, j=1,...,c.

$N_{i.}, N_{.j}$ son conocidos.

$$\mathbf{x}'_k = \left(\delta_{1.k}, \delta_{2.k}, \ldots, \delta_{r.k}, \delta_{.1k}, \ldots, \delta_{.ck}\right),$$

donde:

$$\delta_{i.k} = \begin{cases} 1 & \text{si el elemento } k \text{ esta en el renglon } i \\ 0 & \text{en otro caso} \end{cases}$$

$$\delta_{.jk} = \begin{cases} 1 & \text{si el elemento } k \text{ esta en la columna } j \\ 0 & \text{en otro caso} \end{cases}$$

$$\lambda' = \left(r_1, .., r_r, c_1, .., c_c\right).$$

Entonces, el total es un vector de $r + c$ entradas

$$\mathbf{t}'_x = \sum_u \mathbf{x}'_k = \left(N_{1.}, \ldots, N_{r.}, N_{.1}, \ldots, N_{.c}\right)$$

donde:

$$N_{i.} = \sum_{j=1}^{c} N_{ij}$$

$$N_{.j} = \sum_{i=1}^{r} N_{ij}$$

$$\Rightarrow \mathbf{x}'_k \lambda = r_i + c_j \quad \text{con } k \in U_{ij}.$$

Por lo tanto

$g_k = g(r_i + c_j)$ depende sólo de la celda, es decir, del renglón y la columna.

De las ecuaciones de calibración

$$\begin{aligned}\mathbf{t}_x &= \sum_s d_k g_k \mathbf{x}_k \\ &= \sum_s d_k g(r_i + c_j)\mathbf{x}_k\end{aligned}$$

$$\Rightarrow \sum_{j=1}^{c} \hat{N}_{ij} g(r_i + c_j) = N_{i.}, \qquad i=1,\ldots,r \tag{21}$$

$$\sum_{i=1}^{r} \hat{N}_{ij} g(r_i + c_j) = N_{.j}, \qquad j=1,\ldots,c \tag{22}$$

donde:

$$\hat{N}_{ij} = \sum_{s_{ij}} d_k .$$

Ya que una de las $r+c$ ecuaciones es redundante, para resolver el sistema, se fija una componente de λ, por ejemplo $\lambda_{r+c} = c_c = 0$, y se resuelve el sistema para $i = 1,..,r \;\; y \;\; j = 1,..,c-1$.

Una vez que se determinaron r_i y c_j, las frecuencias de la celda calibradas quedan como sigue

$$\begin{aligned}\hat{N}_{ij}^{w} &= \sum_{s_{ij}} d_k g_k \\ &= \sum_{sij} d_k g(r_i + c_j) \\ &= g(r_i + c_j)\hat{N}_{ij}\end{aligned}$$

$$\Rightarrow g(r_i + c_j) = \frac{\hat{N}_{ij}^{w}}{\hat{N}_{ij}},$$

por lo tanto, los pesos calibrados están dados por

$$\begin{aligned}w_k &= d_k g_k \\ &= d_k \frac{\hat{N}_{ij}^{w}}{\hat{N}_{ij}}.\end{aligned}$$

Por último, el estimador del total se obtiene mediante

$$\hat{t}_{yw} = \sum_s w_k y_k$$

$$= \sum_s d_k \frac{N_{ij}^w}{\hat{N}_{ij}} y_k$$
$$= \sum_i \sum_j \hat{N}_{ij}^w \tilde{y}_{sij} \qquad (23)$$

donde:

$$\tilde{y}_{sij} = \frac{\sum_{s_{ij}} d_k y_k}{\hat{N}_{ij}}.$$

El estimador (23) sólo difiere del estimador de posestratificación completa en que utiliza las frecuencias estimadas bajo calibración, $\hat{N}_{ij}^w$, en lugar de las frecuencias poblacionales N_{ij}.

Por otra parte, en el caso de un muestreo aleatorio simple o bien de un muestreo autoponderado[6], el ponderador es el mismo para todos los elementos de la muestra, y por lo tanto el ponderador calibrado será igual para todos los que pertenezcan a la misma categoría, como se ilustra a continuación

$$d_k = \frac{N}{n}$$
$$\Rightarrow \hat{N}_{ij} = \sum_{s_{ij}} d_k = \frac{N}{n} n_{ij},$$

por ende, el ponderador ajustado está dado por

$$w_k = d_k \frac{\hat{N}_{ij}^w}{\hat{N}_{ij}}$$
$$= \frac{\hat{N}_{ij}^w}{n_{ij}}.$$

Es importante notar que las ecuaciones de calibración (21) y (22) deben de resolverse para la función elegida por el estadístico. Así, los ponderadores calibrados serán de diferentes formas, de acuerdo con la medida de distancia que se utilice, como se muestra en la tabla 1.7.

[6] En un muestreo autoponderado la selección se hace de tal manera que la probabilidad de que un elemento pertenezca a la muestra sea igual para todos los elementos. Por ende la media de la población se puede estimar como la media simple de los casos en la muestra. Por ejemplo, en el muestreo aleatorio estratificado con asignación proporcional, el factor de expansión para los elementos que pertenecen al estrato h es $d_h=N_h/n_h$, pero como $n_h=nN_h/N$, entonces $d_h=N/n$.

Tabla 1.7. Ponderador calibrado (posestratificación incompleta), según función de distancia

Función de Distancia	*Ponderador calibrado*
Mínimos Cuadrados Generalizados	$w_k = d_k\left(1 + r_i + c_j\right)$
Raking Ratio	$w_k = d_k \exp\left(r_i + c_j\right)$
Hellinger	$w_k = d_k\left(1 - \dfrac{\left(r_i + c_j\right)}{2}\right)^{-2}$
Entropía Mínima	$w_k = d_k\left(1 - r_i - c_j\right)^{-1}$
Mínimos Cuadrados Modificada	$w_k = d_k\left(1 - 2r_i - 2c_j\right)^{-1/2}$
Mínimos Cuadrados Restringida	$w_k = \begin{cases} d_k L & \text{si } 1 + r_i + c_j < L \\ d_k U & \text{si } 1 + r_i + c_j > U \\ d_k\left(1 + r_i + c_j\right) & \text{otro caso} \end{cases}$
Logit	$w_k = d_k \dfrac{(U-1)L + (1-L)U\exp\left[A\left(r_i + c_j\right)\right]}{(U-1) + (1-L)\exp\left[A\left(r_i + c_j\right)\right]}$
Huang-Fuller Modificada	$w_k = d_k\left(1 + a_k^{(v-1)^*}\left(r_i^{(v)} + c_j^{(v)}\right)\right)$, v = iteración
Contracción-Minimización	$w_k = w_k^{(v-1)^*}\left(1 + r_i^{(v)} + c_j^{(v)}\right)$, v = iteración

$A=(U-L)/[(1-L)(U-1)]$, $a_k^{(v-1)^*}$ es un factor de ajuste (ver sección 1.3.8), $w_k^{(v-1)^*}$ es el ponderador de la iteración anterior (ver sección 1.3.9).

1.6. ALGORITMO DE CÓMPUTO

Para encontrar los nuevos factores w_k es indispensable hallar la solución al sistema de ecuaciones de calibración (6), en general este sistema no es lineal y por ello es necesario emplear un procedimiento iterativo. Así, uno de los métodos posibles es el de Newton, ver Deville, Särndal, Sautory (1993), el cual se expone a continuación[7].

[7] Los programas que se utilizaron en este trabajo se muestran en el Anexo A.

Sea:

$$\begin{aligned}\varnothing_s(\lambda) &= \sum_s d_k\, g(\mathbf{x}'_k\,\lambda)\mathbf{x}_k - \sum_s d_k \mathbf{x}_k \\ &= \mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} \\ \Rightarrow \varnothing'_s(\lambda) &= \frac{\partial \varnothing_s(\lambda)}{\partial \lambda}.\end{aligned}$$

El valor de λ en la iteracion $v+1$ está dado por:

$$\begin{aligned}\lambda^{(v+1)} &= \lambda^{(v)} + \left\{\varnothing'_s\left(\lambda^{(v)}\right)\right\}^{-1}\left\{\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} - \varnothing_s\left(\lambda^{(v)}\right)\right\} \\ &= \lambda^{(v)} + \left\{\varnothing'_s\left(\lambda^{(v)}\right)\right\}^{-1}\left\{\mathbf{t}_x - \hat{\mathbf{t}}^{(v)}_{xw}\right\}\end{aligned}$$

donde:

$v = 1,2,\ldots$

$\hat{\mathbf{t}}^{(v)}_{xw} = \sum_s d_k\, g(\mathbf{x}'_k\,\lambda^{(v)})\mathbf{x}_k$.

Los valores iniciales para el procedimiento son:

$\lambda^0 = \mathbf{0}$

$\varnothing_s(\mathbf{0}) = \mathbf{0}$

$\varnothing'_s(\mathbf{0}) = \mathbf{T}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}'_k$.

Como puede observarse, en la primera iteración se obtiene $\lambda^1 = \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)$, que es la solución al método lineal. Las iteraciones continúan hasta que se cumple el criterio de convergencia sobre la diferencia entre el total poblacional y el total estimado bajo calibración, $\mathbf{t}_x - \hat{\mathbf{t}}^{(v)}_{xw}$.

En el caso de que la información auxiliar corresponda a los totales marginales de una tabla de contingencia el vector de multiplicadores de Lagrange es del tipo $\lambda' = (r_1,..,r_r,..,c_1,..,c_c)$. Fijando $c_c = 0$, lo cual elimina el último renglón y la última columna de la matriz $\varnothing_s'(\lambda)$, se obtienen los siguientes elementos de la matriz:

$$\varnothing_s'(\lambda) = \begin{cases} m_{i,i} = \sum_{j=1}^{c} \hat{N}_{ij}\, g'(r_i + c_j) \\ m_{r+j,r+j} = \sum_{i=1}^{r} \hat{N}_{ij}\, g'(r_i + c_j) \\ m_{i,r+j} = m_{r+j,i} = \hat{N}_{ij}\, g'(r_i + c_j) \\ 0 \qquad\qquad \text{en otro caso} \end{cases}$$

donde:

$$g'\left(r_i + c_j\right) = \frac{\partial\ g(u)}{\partial u}.$$

Para las medidas propuestas por Deville y Särndal, la función $g'(u)$ tiene la siguiente forma.

Tabla 1.8. Función g'(u), según función de distancia

Función de Distancia	*Función* $g'(u)$
Mínimos Cuadrados Generalizados	$g'(u) = 1$
Raking Ratio	$g'(u) = \exp\left(r_i + c_j\right)$
Hellinger	$g'(u) = \left(1 - \frac{\left(r_i + c_j\right)}{2}\right)^{-3}$
Entropía Mínima	$g'(u) = \left(1 - r_i - c_j\right)^{-2}$
Mínimos Cuadrados Modificada	$g'(u) = \left(1 - 2r_i - 2c_j\right)^{-3/2}$
Mínimos Cuadrados Restringida	$g'(u) = \begin{cases} 1 & \text{si } L \le g_k \le U \\ 0 & \text{si } g_k < L \text{ o } g_k > U \end{cases}$
Logit	$g'(u) = \frac{(U-L)^2 \exp\left[A\left(r_i + c_j\right)\right]}{\left\{(U-1) + (1-L)\exp\left[A\left(r_i + c_j\right)\right]\right\}^2}$

La solución para la distancia *raking ratio* se puede obtener efectuando el algoritmo clásico *raking ratio* de Deming y Stephan (1940), algunas veces llamado ajuste proporcional iterativo (*iterative proportional fitting*). Deming y Stephan sugirieron tal algoritmo pensando que este convergía a la solución del método lineal.

CAPITULO 2
FUENTES DE INFORMACION

Como se mencionó en el capítulo anterior, para generar estimadores a partir de ponderadores calibrados es necesario disponer de información proveniente de dos fuentes:

i) De una encuesta (muestra).

ii) De datos censales, registros, estadísticas vitales, etc.

En lo que se refiere a la primera, en este trabajo se utilizó la Encuesta Nacional de Ingreso y Gasto de los Hogares para los años 1992 y 1996, se eligieron tales años porque corresponden a las fechas más cercanas a los datos censales. Con respecto al segundo tipo de información, se empleó tanto el XI Censo General de Población de Vivienda 1990, así como del Conteo de Población 1995. En el siguiente apartado se presenta una descripción de estas fuentes.

2.1. ENCUESTA NACIONAL DE INGRESO Y GASTO DE LOS HOGARES

El Instituto Nacional de Estadística, Geografía e Informática (INEGI) realiza en forma periódica diversas encuestas tanto en establecimientos económicos como en hogares. Es a este último tipo de unidades al que se aplica la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), la cual se ha levantado en los años 1984, 1989, 1992, 1994 y 1996.

2.1.1. Objetivos

De acuerdo con lo anterior, las encuestas tienen como objetivo general proporcionar información sobre la distribución, monto y estructura del ingreso y el gasto de los hogares mexicanos, y como objetivos específicos generar información acerca de:

i) La estructura del ingreso y gasto corrientes, percepciones y erogaciones financieras.

ii) Las características sociodemográficas de los miembros del hogar.

iii) La condición de actividad y características ocupacionales de los miembros de 12 años y más.

iv) Las características de infraestructura de la vivienda y de equipamiento del hogar.

2.1.2. Metodología

En las encuestas realizadas en 1992 y 1996 se aplicaron esencialmente los mismos aspectos metodológicos, lo cual permite la comparación de los resultados obtenidos en los diferentes levantamientos. A continuación se describen brevemente estos puntos.

i) Marco conceptual. Las encuestas de ingreso y gasto en los hogares están basadas en la consideración de que el monto del ingreso, su procedencia y su forma de distribución condiciona en gran medida, el nivel de bienestar de la población, puesto que es el ingreso el que determina la capacidad económica de los hogares para adquirir los bienes y servicios necesarios.

ii) Población objeto de estudio. La constituyen los hogares de nacionales y extranjeros que residen habitualmente en el país.

iii) Periodos de referencia. Los conceptos que componen el ingreso y el gasto son de muy diversa naturaleza, en cuanto a su ocurrencia y fluctuación en el tiempo, por lo que se planteó la necesidad de combinar periodos de diferente extensión. Por ejemplo, el gasto en alimentos se refiere al periodo de de una semana anterior a la fecha de la encuesta, en cambio, para el ingreso neto el periodo es de seis meses anteriores a la fecha de la entrevista. Para las variables sociodemográficas se consideró el momento de la entrevista como periodo de referencia.

iv) Unidades de análisis. Se seleccionó a la vivienda particular como unidad de muestreo y al hogar como unidad de observación, ubicando a partir de éstas a las unidades de análisis[1].

v) Instrumentos de captación. La recolección de información se llevó a cabo por medio de visitas a las viviendas seleccionadas, utilizando cuestionarios especializados que permitieron la operacionalización del marco conceptual.

vi) Periodo de levantamiento. Las encuestas fueron efectuadas durante el tercer trimestre del año, ya que en esta época se han observado menos variaciones en la percepción de los ingresos y los gastos. El periodo de levantamiento fue del 21 de agosto al 17 de noviembre del año respectivo, dándose a conocer los resultados en noviembre de 1993 la encuesta de 92 y en septiembre de 1998 la encuesta de 96.

vii) Diseño muestral. El esquema de muestreo fue estratificado, polietápico y con probabilidad proporcional al tamaño, éste se describirá con mayor detalle en el siguiente apartado.

viii) Procedimientos de operación de campo. Se contó con un equipo de entrevistadores, supervisores y jefes de área capacitados de manera especial sobre los procedimientos, lineamientos y criterios establecidos con base al marco de conceptos. Así también se desarrollaron mecanismos de control para asegurar la calidad de la información.

ix) Crítica-codificación. Tiene la finalidad de mantener la calidad de la información y asegurar la veracidad y confiabilidad de los resultados, se efectúa posteriormente a la recolección de los datos y antes de su procesamiento. Consiste en revisar los instrumentos de captación (cuestionarios) para eliminar errores, omisiones e inconsistencias que puedan ser recuperadas con retorno a campo o criterios de imputación.

x) Captura de la información. El sistema de captura incluye filtros de congruencia e integridad que sólo permite registrar a los cuestionarios correctamente requisitados. También se realizó una validación del archivo mediante frecuencias para algunas variables captadas.

2.1.3. Diseño Muestral

Aunque los diseños para las encuestas de 1992 y 1996 tienen varias características comunes, existen otras en las que difieren. A continuación se proporciona una descripción de cada uno de ellos.

[1] El INEGI define una vivienda particular como la vivienda destinada al alojamiento de familias o grupos de personas que forman hogares. Un hogar se define como el conjunto de personas unidas o no por lazos de parentesco que residen habitualmente en la misma vivienda particular y se sostienen de un gasto común, principalmente para comer.

i) Diseño 1992

El diseño muestral de la ENIGH 1992 considera los objetivos planteados por la encuesta, la población objeto de estudio, las principales variables sobre las que se desea generar información y niveles o dominios de estudio. Dominio de estudio es una parte de la población para la que se desea dar estimaciones con precisión y confianza estadísticas propias. En este caso, se deseaba generar resultados a nivel nacional así como para los estratos urbano (zonas de alta densidad) y rural (zonas de baja densidad). Así, el diseño de muestreo fue estratificado, polietápico y con probabilidad proporcional al tamaño.

Los estratos se conformaron de acuerdo a su densidad de población:

a) Estrato Urbano. Lo constituyen los municipios del país que cumplieron con alguna de las siguientes características: tener al menos una localidad de 15,000 o más habitantes, tener 100,000 habitantes o más, contener a la capital de la entidad, formar parte de las áreas metropolitanas consideradas en la Encuesta Nacional de Empleo Urbano (ENEU). Este estrato se dividió en cuatro subestratos:

a1) Areas metropolitanas y conurbadas.
a2) Localidades de 100,000 habitantes y más.
a3) Localidades de 15,000 a 99,999 habitantes.
a4) Localidades de 2,500 a 14,999 habitantes.

b) Estrato rural. Está formado por todos los municipios del país que no están considerados en ninguna de las categorías anteriores.

En cada estado se identificaron algunas ciudades que se incluyeron con certeza en la muestra y en el resto de la entidad se seleccionaron, de manera independiente en cada área (urbana, rural), unidades primarias de muestreo (UPM's) con probabilidad proporcional a su tamaño, la medida de tamaño fue el total de la población en la unidad. Las UPM's fueron grupos de municipios, localidades o grupos de manzanas que se agrupan conforme a características homogéneas.

En cada UPM seleccionada se eligieron viviendas con igual probabilidad. La selección fue en segmentos compactos de tamaño 5 y 20, para las áreas urbanas y rurales, respectivamente.

ii) Diseño 1996

En el caso de la ENIGH 1996, se deseaba generar resultados a nivel nacional, para los estratos urbano, rural y para cada uno de los estados de Campeche, Coahuila, Guanajuato, Hidalgo, Jalisco, México, Oaxaca y Tabasco.

Para cada estado (dominio o no de estudio) se tuvo un diseño independiente (diferentes criterios por estado). No obstante, el diseño de muestreo en cada estado fue estratificado, polietápico y con probabilidad proporcional al tamaño. Las unidades primarias de muestreo estuvieron constituidas por Areas Geoestadísticas Básicas[2] (AGEBs) y las unidades de segunda etapa por viviendas particulares.

[2] AGEB es la unidad mínima del Marco Geoestadístico, y posee tres atributos fundamentales: a) es reconocible en el terreno y está delimitada por rasgos identificables y perdurables, b) es homogénea, en cuanto a sus características sociales, económicas y geográficas c) su extensión es tal que se puede supervisar por una persona.

La estratificación se hizo de acuerdo a la población en las localidades, conformándose dos estratos (urbano/rural) en cada estado, dando lugar a 64 estratos a nivel nacional.

a) Estrato Urbano. Lo constituyen las localidades de 2,500 y más habitantes, y a su vez está formado por cuatro subestratos:

a1) Ciudades que entran con certeza: áreas metropolitanas y áreas conurbadas, así como ciudades de interés y/o capitales de estado.
a2) Localidades de 100,000 habitantes y más.
a3) Localidades de 15,000 a 99,999 habitantes.
a4) Localidades de 2,500 a 14,999 habitantes.

b) Estrato rural. Lo integran las localidades de menos de 2,500 habitantes.

En las ciudades con certeza, la selección de las unidades primarias de muestreo (UPM) fue con probabilidad proporcional al tamaño de la misma; la medida de tamaño para las UPMs fue el número de viviendas de acuerdo al Censo de Población y Vivienda de 1990. En el resto de localidades urbanas se efectuó una agrupación previa de las AGEBs de acuerdo con variables socioeconómicas censales y a la UPM se le asignó el estrato de la AGEB de la que provenía, efectuándose también una selección con probabilidad proporcional al tamaño. En cada UPM seleccionada se eligieron 20 viviendas en segmentos compactos de tamaño cinco.

En las localidades con menos de 2,500 habitantes también se efectuó previamente una estratificación de las AGEBs con base en variables socioeconómicas, se eligieron unidades primarias de muestreo con probabilidad proporcional al tamaño y en la segunda etapa se seleccionaron 20 viviendas en segmentos compactos de tamaño 10.

En cada estado dominio de estudio se eligieron 50 unidades primarias de muestreo y un promedio de 20 viviendas por UPM mediante muestreo sistemático de las áreas de listado de viviendas. Un listado es el registro de las viviendas particulares que se encuentran en una manzana o localidad. La asignación del número de UPMs dentro de los estratos fue proporcional al número de viviendas de cada unidad seleccionada para la Encuesta Nacional de la Dinámica Demográfica (ENADID).

iii) Factores de expansión

Para generar las estimaciones correspondientes es necesario el uso de ponderadores, de acuerdo con el esquema de selección. El factor de expansión está dado por el producto del inverso de la probabilidad de selección en cada etapa.

De acuerdo con la nota metodológica proporcionada por el INEGI, la probabilidad de selección de una vivienda del área de listado L, en la *i-ésima* UPM del estrato h está dado por:

$$\pi_{hiL} = \frac{n_h V_{hi}}{V_h} \frac{M_L m_L}{V_L},$$

En las localidades de 2,500 y más habitantes el AGEB es el agrupamiento convencional de las manzanas, su tamaño puede variar de 20 a 80 manzanas aproximadamente. En las localidades de menos de 2,500 habitantes es el espacio que incluye una o más localidades.

donde:

n_h Número de UPM's seleccionadas por estrato

V_{hi} Total de viviendas en la UPM i del estrato h al momento de construirse el marco

V_h Total de viviendas en estrato h

V_L Total de viviendas del listado L

M_L Número de segmentos seleccionados por listado

m_L En el estrato urbano es igual a 5 y a 10 en el rural.

Entonces, el factor de expansión o ponderador es el inverso de la probabilidad de selección:

$$d_{hiL} = 1/\pi_{hiL}.$$

iv) Tamaño de muestra

El cálculo del tamaño de muestra involucra el parámetro que se desea estimar (proporción, media, total, etc.), el nivel de confianza deseado y el error que se está dispuesto a aceptar. Así, el tamaño de muestra para estimar la media de una variable, suponiendo un muestreo aleatorio simple sin reemplazo, está dado:

$$n_o = \frac{t_{\alpha/2}^2 S^2}{d^2}, \qquad n = \frac{n_o}{1 + n_o/N}$$

donde:

n Tamaño de muestra

S^2 Varianza poblacional

$t_{\alpha/2}$ Corresponde al porcentil de una distribución normal (cuando la varianza es conocida), o bien al porcentil de una distribución t-student con $n-1$ grados de libertad, asociado al nivel de confianza deseado α

d Es el error absoluto que se está dispuesto a aceptar

N Número de elementos en la población.

Adicionalmente, cuando el diseño de muestreo no es aleatorio simple, se hace necesario considerar el efecto del diseño, y por otra parte una tasa de no respuesta. Lo anterior da lugar al siguiente tamaño de muestra final:

$$n_f = n\frac{deff}{(1 - TNR)}$$

donde:

n_f Es el número de viviendas a seleccionar

TNR Es la tasa de no respuesta esperada

deff Es el efecto de diseño, con $deff = \dfrac{Var\left(\hat{\theta}_{muestreo\ complejo}\right)}{Var\left(\hat{\theta}_{m.a.s.}\right)}$.

El efecto de diseño es la razón entre la varianza de un estimador bajo un esquema de muestreo dado y la varianza de un muestreo aleatorio simple del mismo número de elementos, es decir, es la eficiencia relativa de un m.a.s. respecto a un muestreo dado (Kish, 1989).

De acuerdo con el material acumulado por la Dirección General de Estadística del INEGI, en particular de las encuestas en los hogares, se han calculado varianzas, coeficientes de variación y efectos de diseño para variables correlacionadas con el ingreso. De esta experiencia, el valor máximo obtenido para un efecto de diseño fue de 2.3, cuando el segmento de viviendas seleccionadas fue de 5. A partir de este valor se estimó el efecto de diseño cuando el segmento era de 10 viviendas, obteniéndose un efecto de 3.9. En cuanto a la tasa de no respuesta, se esperaba que fuera del 15%.

Así, el tamaño de muestra (planeado) para 1992 fue aproximadamente de 10,000 viviendas distribuidas de manera independiente en todos los estados del país, el tamaño de muestra por estado se asignó de manera proporcional al número de viviendas en el mismo. En lo correspondiente a 1996, en cada estado dominio de estudio se eligieron aproximadamente 1,000 viviendas y a nivel nacional un total de 14,000 viviendas.

Una vez efectuados los procesos de crítica-codificación y validación de la información, el número efectivo de viviendas así como el número de personas entrevistadas, por estado y estrato quedó como se muestra en las tablas 2.1 y 2.2. Como se puede observar el número de viviendas del Distrito Federal, Nuevo León y Tlaxcala disminuye en 1996 con respecto a 1992; desafortunadamente el documento metodológico de la encuesta no explica la razón.

Tabla 2.1. Número de viviendas en muestra por estado y estrato

Estado	1992			1996		
	Urbano	*Rural*	*Total*	*Urbano*	*Rural*	*Total*
1 Aguascalientes	184	75	259	152	94	246
2 B.C.	132	58	190	226	47	273
3 B.C.S.	90	133	223	165	71	236
4 Campeche	95	146	241	546	281	**827**
5 Coahuila	193	70	263	650	209	**859**
6 Colima	100	150	250	142	84	226
7 Chiapas	126	159	285	129	153	282
8 Chihuahua	159	105	264	196	73	269
9 D.F.	979	45	1,024	805	35	840
10 Durango	108	150	258	159	103	262
11 Guanajuato	174	141	315	574	299	**873**
12 Guerrero	118	130	248	140	135	275
13 Hidalgo	70	186	256	432	423	**855**
14 Jalisco	432	71	503	624	217	**841**
15 México	787	95	882	1,026	72	**1,098**
16 Michoacán	122	69	191	168	117	285
17 Morelos	132	104	236	191	55	246
18 Nayarit	63	187	250	109	149	258
19 Nuevo León	392	106	498	197	57	254
20 Oaxaca	85	150	235	370	438	**808**
21 Puebla	127	116	243	152	118	270
22 Querétaro	66	195	261	138	120	258
23 Q. Roo	156	102	258	178	72	250
24 S.L.P.	83	153	236	153	122	275
25 Sinaloa	110	152	262	190	79	269
26 Sonora	136	108	244	199	80	279
27 Tabasco	100	158	258	501	451	**952**
28 Tamaulipas	150	108	258	205	70	275
29 Tlaxcala	502	360	862	175	107	282
30 Veracruz	157	116	273	177	113	290
31 Yucatán	106	145	251	174	89	263
32 Zacatecas	101	152	253	115	151	266
Nacional	6,335	4,195	10,530	9,358	4,684	14,042

Fuente: ENIGH 92 y 96.
Negritas: Estados dominio de estudio en 1996.

Tabla 2.2. Número de personas seleccionadas por estado y estrato

	1992			*1996*		
Estado	*Urbano*	*Rural*	*Total*	*Urbano*	*Rural*	*Total*
1 Aguascalientes	925	413	1,338	681	484	1,165
2 B.C.	525	254	779	890	195	1,085
3 B.C.S.	406	610	1,016	625	266	891
4 Campeche	380	745	1,125	2,402	1,418	**3,820**
5 Coahuila	866	324	1,190	2,949	867	**3,816**
6 Colima	386	680	1,066	575	382	957
7 Chiapas	624	895	1,519	557	931	1,488
8 Chihuahua	633	503	1,136	784	292	1,076
9 D.F.	4,279	208	4,487	3,174	166	3,340
10 Durango	517	762	1,279	692	468	1,160
11 Guanajuato	850	742	1,592	2,764	1,680	**4,444**
12 Guerrero	536	725	1,261	633	725	1,358
13 Hidalgo	313	918	1,231	1,854	2,002	**3,856**
14 Jalisco	2,069	383	2,452	2,933	1,060	**3,993**
15 México	3,718	536	4,254	4,775	389	**5,164**
16 Michoacán	581	368	949	781	588	1,369
17 Morelos	625	516	1,141	832	262	1,094
18 Nayarit	279	849	1,128	442	661	1,103
19 Nuevo León	1,804	500	2,304	823	284	1,107
20 Oaxaca	415	811	1,226	1,613	2,170	**3,783**
21 Puebla	610	736	1,346	693	658	1,351
22 Querétaro	312	1,055	1,367	612	601	1,213
23 Q. Roo	640	540	1,180	669	315	984
24 S.L.P.	409	769	1,178	696	626	1,322
25 Sinaloa	503	789	1,292	851	410	1,261
26 Sonora	597	502	1,099	859	375	1,234
27 Tabasco	426	906	1,332	2,227	2,402	**4,629**
28 Tamaulipas	615	466	1,081	815	303	1,118
29 Tlaxcala	2,549	2,068	4,617	826	528	1,354
30 Veracruz	697	624	1,321	742	579	1,321
31 Yucatán	498	773	1,271	759	473	1,232
32 Zacatecas	504	801	1,305	495	776	1,271
Nacional	29,091	21,771	50,862	41,023	23,336	64,359

Fuente: ENIGH 92 y 96
Negritas: Estados dominio de estudio en 1996.

2.1.4. Información generada

Como se mencionó anteriormente, estas encuestas proporcionan información que se puede clasificar en los siguientes rubros:

Tabla 2.3. Tipo de información ENIGH

Rubro	*Nivel*	*Ejemplos*
Características sociodemográficas	Personas	Edad, sexo, perfil educativo, características ocupacionales.
Infraestructura y servicios	Vivienda	Disponibilidad de agua, drenaje, material en muros, techos, pisos, tenencia de la propiedad.
Ingreso Gasto	Hogar	Salarios, rentas, intereses. Consumo comida, educación.

Adicionalmente, el ingreso y el gasto, según la fuente de donde provienen, puede dividirse en monetario, no monetario y percepciones financieras (monetarias y no monetarias) o erogaciones financieras (monetarias y no monetarias). En el Anexo B se presentan los principales conceptos que integran cada uno de dichos puntos.

2.1.4.1. Características sociodemográficas

A continuación se presenta la distribución porcentual de las principales variables que proporciona la encuesta referente a las características sociales y demográficas.

Tabla 2.4. Variables demográficas

	1992		*1996*	
Variable	*%*	*Tamaño de muestra*	*%*	*Tamaño de muestra*
Tipo de área				
Urbana	72.6	28830	72.7	41023
Rural	27.4	21548	27.3	23336
Total	100.0	50378	100.0	64359
Edad				
0 - 11 años	30.3	15846	29.1	18958
12 - 24 años	28.6	14309	27.3	17852
25 - 39 años	20.5	10091	21.9	13543
40 - 64 años	16.1	7872	17.2	10904
65 y más años	4.4	2260	4.7	3102
Total	100.0	50378	100.0	64359
Sexo				
Hombre	48.9	24904	48.7	31485
Mujer	51.1	25474	51.3	32874
Total	100.0	50378	100.0	64359

Tabla 2.4. Variables demográficas (continuación)

Variable	1992 %	1992 Tamaño de muestra	1996 %	1996 Tamaño de muestra
Estado civil (personas de 12 años y más)				
Unión libre			7.1	3214
Casado			45.8	20992
Separado			2.9	1284
Divorciado			0.8	352
Viudo			4.5	1952
Soltero			38.8	17607
Total			100.0	45401

Fuente: Cálculos propios basados en la ENIGH 92 y 96.
No se preguntó el estado civil en la encuesta de 1992.

Como puede observarse en la tabla 2.4, en lo referente al tipo de área, edad, sexo y estado civil, entre 1992 y 1996 no hubo cambios importantes en los porcentajes para las categorías que las integran. Respecto a las variables nivel de escolaridad y tipo de escuela a la que se asiste se observan diferencias entre ambos años.

Tabla 2.5. Variables de escolaridad

Variable	1992 %	1992 Tamaño de muestra	1996 %	1996 Tamaño de muestra
Sabe leer y escribir (personas de 6 años y más)				
Si	87.7	36842	88.2	48025
No	12.3	5736	11.8	7081
Total	100.0	42578	100.0	55106
Asistencia a centro educativo (personas de 5 años y más)				
Si	34.0	14665	33.5	19061
No	66.0	29208	66.5	37584
Total	100.0	43873	100.0	56645
Tipo de escuela a la que asiste				
Pública	87.0	13213	89.9	17665
Privada	12.6	1390	10.0	1379
Otro	0.4	62	0.1	17
Total	100.0	14665	100.0	19061
Nivel de escolaridad (personas 6 años y más)				
Sin primaria	11.1	5130	8.8	5320
Primaria incompleta	36.4	16202	34.1	20076
Primaria completa, secundaria incompleta	25.1	10749	25.0	14144
Secundaria completa, preparatoria incompleta	17.6	6992	20.1	10043
Preparatoria completa, universidad incompleta	6.9	2517	8.5	3932
Universidad completa y más	3.0	988	3.5	1588
Total	100.0	42578	100.0	55103

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

Tabla 2.6. Variables de condición de actividad

Variable	1992		1996	
	%	Tamaño de muestra	%	Tamaño de muestra
Percepción de ingresos				
Sí	35.8	17404	39.3	24966
No	64.2	32974	60.7	39393
Total	100.0	50378	100.0	64359
Condición de actividad (personas de 12 años y más)				
Ocupado	49.7	17270	53.8	24217
Desocupado	1.9	692	2.5	1076
Población económicamente inactiva	48.4	16570	43.7	20108
Total	100.0	34532	100.0	45401
Población económicamente inactiva (12 años y más)				
Quehaceres domésticos	54.8	9515	53.4	11027
Estudiante	34.4	5421	35.4	6861
Pensionado	3.3	462	3.9	680
Otros	7.4	1172	7.3	1540
Total	100.0	16570	100.0	20108
Posición en el trabajo (ocupados)				
Empleado u obrero	59.1	9350	55.7	12523
Jornalero o peón	6.8	1543	7.5	2320
Patrón o empresario	5.4	909	4.7	1156
Cuenta propia	20.5	3828	22.2	5664
Sin retribución	8.1	1625	9.9	2554
No especificado	0.1	15	0.0	0
Total	100.0	17270	100.0	24217
Número de horas trabajadas a la semana (ocupados)				
1 - 20	9.9	1813	13.4	3261
21 - 39	16.7	3096	18.1	4800
40 - 48	43.6	7266	39.3	9179
48 y más	27.6	4714	29.2	6977
No especificado	2.2	381	0.0	0
Total	100.0	17270	100.0	24217
Sector económico de actividad (ocupados)				
Agricultura, ganadería, etc.	21.6	4923	21.2	6318
Industria manufacturera	18.7	2746	18.0	3975
Industria no manufacturera	9.0	1561	7.2	1823
Comercio	17.2	2728	16.8	3780
Transportes, servicios financieros, etc.	4.9	815	5.3	1069
Servicios comunitarios y sociales	28.5	4482	31.3	7252
No especificado	0.1	15		
Total	100.0	17270	100.0	24217
Afiliación a algún sindicato (ocupados)				
Sí pertenece	14.1	2118	10.0	2368
No pertenece	51.9	8846	53.2	12478
No especificado	34.0	6306	36.8	9371
Total	100.0	17270	100.0	24217

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

En lo que se refiere a las características relacionadas con la condición de actividad existen más diferencias entre las distribuciones estimadas para 1992 y 1996. Aunque en sentido estricto sería necesario efectuar (en todos los casos) las pruebas de hipótesis pertinentes con la finalidad de ver si tales diferencias son estadísticamente significativas. Por otra parte, es importante aclarar que la ENIGH no es una encuesta especializada en empleo ya que el Instituto también lleva a cabo la Encuesta Nacional de Empleo (ENE) y la de Empleo Urbano (ENEU).

2.1.4.2. Características de las viviendas

Las principales características de las viviendas captadas por las encuestas se pueden dividir: en tipo de vivienda, número de habitantes por hogar y de cuartos para diferentes usos, materiales de los que está construida, disponibilidad de diversos servicios y disponibilidad de vehículos. En las siguientes tablas se presentan las proporciones estimadas para cada categoría, según el año[3].

Tabla 2.7. Tipo de vivienda

	1992			*1996*	
Variable	*%*	*Tamaño de muestra*	*Variable*	*%*	*Tamaño de muestra*
Tipo de vivienda					
Casa sola	84.0	9198	Casa sola	85.1	12475
Departamento o casa en vecindad	14.3	1121	Casa con áreas comúnes	2.6	283
Cuarto de azotea, vivienda móvil	0.2	15	Departamento	11.0	1082
			Otros	0.1	11
No especificado	1.5	196	No especificado	1.3	191
Total	100.0	10530	Total	100.0	14042
Tenencia de la vivienda					
Propia y totalmente pagada en terreno propio	53.1	5313	Propia y totalmente pagada en terreno propio	52.5	7290
Propia y totalmente pagada en terreno ejidal	16.8	2283	Propia en terreno ejidal	15.1	2715
Propia y totalmente pagada en terreno de asentamiento irregular	1.7	179	Propia en terreno de asentamiento irregular	1.3	145
Propia y la están pagando	5.1	481	Propia y la están pagando	7.2	801
Rentada	11.4	947	Rentada	11.9	1458
Recibida como prestación	0.9	107	Recibida como prestación	0.8	99
Prestada	9.3	991	Prestada	9.6	1321
Otros	0.2	33	Otros	0.2	22
No especificado	1.5	196	No especificado	1.3	191
Total	100.0	10530	Total	100.0	14042

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

[3] En algunos casos las categorías cambiaron de un año a otro, en caso de que no se pudieran formar las mismas para ambos años, en las tablas se presenta la descripción para cada periodo.

Tabla 2.8. Número de ocupantes y de cuartos para usos diversos

Variable	*1992*		*1996*	
	%	*Tamaño de muestra*	*%*	*Tamaño de muestra*
Tamaño del hogar				
1-2 personas	16.5	1709	17.8	2445
3-4 personas	34.1	3551	36.8	5053
5-6 personas	30.9	3162	29.7	4199
7 y más personas	18.5	2108	15.6	2345
Total	100.0	10530	100.0	14042
Número de cuartos				
1	25.2	2855	21.3	3196
2	26.3	2969	22.6	3398
3	22.1	2218	25.7	3482
4	15.2	1441	17.6	2385
5 y más	11.2	1047	12.8	1581
Total	100.0	10530	100.0	14042
Número de cuartos p/dormir				
1	40.0	4460	39.3	5645
2	34.7	3697	37.5	5285
3	19.2	1792	18.0	2455
4 y más	6.0	581	5.2	656
Total	100.0	10530	100.0	14042
Existencia de cuarto para cocina				
Sí	84.4	8767	84.6	11755
No	14.1	1567	14.1	2096
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042
Duermen en la cocina				
Sí	2.8	321	2.2	369
No	81.5	8437	82.5	11386
No especificado	15.7	1772	15.4	2287
Total	100.0	10530	100.0	14042
Existencia de cuarto para baño				
Sí	79.6	7937	86.1	11556
No	18.8	2397	12.6	2295
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042
Número de recamaras				
0			19.0	2770
1			20.0	2954
2			34.3	4805
3			20.0	2682
4 y más			6.6	831
Total			100.0	14042

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

Tabla 2.9. Material predominante en la vivienda

Variable	*1992*		*1996*	
	%	*Tamaño de muestra*	*%*	*Tamaño de muestra*
Material predominante en muros				
Lámina de cartón	0.7	120	0.7	93
Carrizo, bambú o palma	1.0	165	1.0	197
Embarro o bajareque	3.0	387	1.5	329
Madera	7.1	800	6.4	992
Lámina de asbesto o metálica	1.0	82	0.6	119
Adobe	12.2	1527	12.1	1850
Tabique, ladrillo, block, piedra o cemento	72.7	7183	74.6	10048
Otros materiales	0.7	70	1.9	223
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042
Material predominante en techos				
Lámina de cartón	8.5	968	5.6	770
Palma, tejamanil o madera	7.3	925	5.8	944
Lámina de asbesto o metálica	19.7	2158	18.8	3299
Teja	6.7	639	6.0	694
Losa de concreto, tabique o ladrillo	54.4	5344	53.9	6496
Otros materiales	1.8	300	8.6	1648
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042
Material predominante en pisos				
Tierra	16.4	1941	12.6	2068
Cemento o firme	53.4	6000	52.3	7691
Madera, mosaico u otros	28.7	2393	33.9	4092
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

Tabla 2.10. Disponibilidad de servicios

Variable	1992		1996	
	%	Tamaño de muestra	%	Tamaño de muestra
Disponibilidad de drenaje				
Conectado al de la calle	54.2	4768	60.2	7073
Fosa séptica	13.2	1682	4.2	878
Desagüe al suelo, río o lago	5.3	658	2.4	310
No dispone	25.8	3226	32.0	5590
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042
Disponibilidad de luz eléctrica				
Sí	91.6	9325	94.7	13102
No	6.9	1009	4.0	749
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042
Disponibilidad de teléfono				
Sí	23.3	1954	28.6	3282
No	75.2	8380	70.1	10569
No especificado	1.5	196	1.3	191
Total	100.0	10530	100.0	14042
Disponibilidad de servicio público de recolección de basura				
Sí			75.2	9531
No			24.8	4511
Total			100.0	14042
Frecuencia del servicio				
Diario			29.2	2746
Cada tercer día			38.7	3699
Cada tres días			11.1	1023
Entre cuatro y siete días			20.0	1948
Ocho días o más			1.0	115
Total			100.0	9531

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

Tabla 2.11. Disponibilidad de servicios

	1992			*1996*	
Variable	*%*	*Tamaño de muestra*	*Variable*	*%*	*Tamaño de muestra*
Disponibilidad de agua			*Disponibilidad de agua*		
Entubada dentro de la vivienda	57.4	5572	Entubada dentro de la vivienda	56.5	7072
Entubada fuera de la vivienda pero en el edificio, vecindad o terreno	21.8	2619	Entubada fuera de la vivienda pero en el edificio, vecindad o terreno	28.0	4502
Pozo dentro del terreno	5.3	547	Llave pública	1.2	149
Por acarreo	11.8	1395	No dispone	12.9	2128
Pipa	2.2	201			
No especificado	1.5	196	No especificado	1.3	191
Total	100.0	10530	Total	100.0	14042

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

Tabla 2.12. Disponibilidad de vehículos

Variable	*1992*		*1996*	
	%	*Tamaño de muestra*	*%*	*Tamaño de muestra*
Número de automóviles propios para uso del hogar				
0	80.0	8755	78.9	11492
1	16.4	1504	17.5	2174
2	2.9	229	3.0	314
3 y más	0.7	42	0.6	62
Total	100.0	10530	100.0	14042
Número de camionetas propias para uso del hogar				
0	92.0	9578	88.9	12374
1	7.4	887	10.2	1529
2	0.5	59	0.8	127
3 y más	0.1	6	0.1	12
Total	100.0	10530	100.0	14042
Número de motocicletas propias para uso del hogar				
0	98.7	10368	98.9	13855
1	1.3	154	1.0	176
2	0.1	8	0.0	11
Total	100.0	10530	100.0	14042
Número de bicicletas propias para uso del hogar				
0	71.7	7604	84.4	11318
1	18.3	1989	11.7	2070
2	7.5	705	2.9	497
3 y más	2.5	232	1.0	157
Total	100.0	10530	100.0	14042
Número de vehículos de tracción animal propios para uso del hogar				
0	98.8	10340	99.1	13890
1	1.1	180	0.9	146
2	0.0	9	0.0	5
3 y más	0.0	1	0.0	1
Total	100.0	10530	100.0	14042

Fuente: Cálculos propios basados en la ENIGH 92 y 96.

2.2. XI CENSO GENERAL DE POBLACIÓN Y VIVIENDA 1990, CONTEO DE POBLACIÓN 1995

La información auxiliar que se necesita para mejorar los estimadores mediante el método comentado en el primer capítulo, se obtuvo del Censo de población 1990 y del Conteo de 1995. Al respecto es importante mencionar que dada la naturaleza de dichas fuentes, la información que proporciona el censo es más rica en comparación con la del conteo. Lo anterior constituye una limitante al análisis realizado, puesto que se tienen que encontrar variables disponibles y categorías compatibles tanto de la encuesta como de la información auxiliar.

i) XI Censo General de Población y Vivienda 1990

El Censo de Población y Vivienda 1990 tiene como objetivo proporcionar información confiable y oportuna sobre la población y las viviendas de nuestro país, manteniendo la comparabilidad histórica con censos anteriores. La semana de referencia para el levantamiento fue del 5 al 11 de marzo de 1990 y se publicó en febrero de 1992, sólo dos años después de su levantamiento, en comparación con el censo de 1980 que fue publicado en 1986.

Los temas incluidos en este levantamiento fueron:

a) Características de la vivienda. Comprende preguntas sobre material predominante en paredes, techos y pisos, número de cuartos y dormitorios, tenencia, disponibilidad de agua entubada, drenaje y electricidad, etc.

b) Ocupantes de la vivienda. Con preguntas sobre el número total personas que residían habitualmente en la vivienda y parentesco con respecto al jefe de familia.

c) Características demográficas. Contiene variables como edad, sexo, lugar de nacimiento, estado civil, número de hijos nacidos vivos y sobrevivientes, etc.

d) Características educativas y culturales. Comprende rubros referentes a condición de alfabetismo, asistencia escolar, nivel de instrucción de la población, lengua indígena y religión.

e) Características económicas. Incluye puntos como condición de actividad, ocupación, situación en el trabajo, rama de actividad económica e ingresos.

Así, una de las ventajas del Censo, además de la gama de aspectos que toca, es que provee resultados a varios niveles de estudio: nacional, estatal, municipal, localidad, e inclusive de AGEB.

ii) Conteo de Población y Vivienda 1995

El Conteo de Población y Vivienda 1995, cuya fecha de referencia es el 5 de noviembre de 1995, constituye un proyecto novedoso en el país debido a sus características metodológicas y a la estrategia operativa establecida. Los objetivos generales que persigue este levantamiento son:

a) Proporcionar información básica sobre la población y las viviendas.

b) Mantener actualizadas las estadísticas demográficas y socioeconómicas del país.

c) Incrementar la serie histórica de información sociodemográfica, conservando en la medida de lo posible, la comparabilidad de la información con otros censos y encuestas de población y vivienda.

d) También se plantearon como metas lograr la máxima cobertura, óptima calidad y publicación oportuna de los resultados. La primera edición de los resultados definitivos del Conteo de Población fue en diciembre de 1996, poco más de un año después.

Para cumplir los objetivos propuestos se plantearon dos estrategias:

a) Enumeración exhaustiva de toda la población y viviendas del país. En ésta se captó información referente a servicios básicos de la vivienda (agua, drenaje y electricidad), características de los residentes habituales (sexo, edad, condición de alfabetismo, de habla indígena y española).

b) Encuesta, basada en una muestra de la población. El diseño de muestreo fue estratificado y bietápico; estratificado porque se formaron grupos de viviendas con características similares y bietápico porque en la primera etapa se eligieron grupos de viviendas y posteriormente viviendas. En ella se profundizó sobre temas como características de la vivienda (materiales de construcción, número de cuartos y servicios básicos), estructura por edad y sexo, migración, características educativas, nupcialidad, características económicas (condición de actividad, situación en el trabajo, ocupación, rama de actividad, ingresos, etc.), subsidios sociales, servicios de salud y discapacidad.

La enumeración, cuyos resultados se utilizaron en uno de los ejercicios que se llevaron a cabo en este estudio, proporciona indicadores a nivel nacional, estatal, municipal, localidad y AGEB´s, mientras que la encuesta proporciona resultados a nivel nacional, por entidad federativa y regiones.

CAPITULO 3
ANALISIS DE RESULTADOS

El presente capítulo tiene como objetivo aplicar la técnica expuesta en el capítulo uno, es decir, generar estimaciones mediante pesos calibrados. Así también, se quiere analizar y comparar los resultados generados, de acuerdo a las principales medidas de distancia mencionadas en la literatura. Tal aplicación se hará en el contexto de tablas de contingencia de dos dimensiones[1], llevándose a cabo tres ejercicios en los que se utilizan diferentes variables para calibrar. Asimismo, en cada uno de ellos, se obtuvieron los ponderadores ajustados según las funciones de distancia citadas en la sección 1.3.

3.1. VARIABLES DE CALIBRACIÓN

Como se mencionó en el capítulo anterior, por una parte se utilizó la encuesta denominada ENIGH 1992 y 1996, y por otra parte, la fuente auxiliar la constituyeron algunas estadísticas generadas por el XI Censo General de Población y Vivienda 1990 (para calibrar la de 1992), así como por el Conteo de Población 1995 (para calibrar la de 1996).

En particular, para los tres ejercicios que se llevaron a cabo en esta tesis, se eligieron los siguientes pares de variables con la finalidad de calibrar los ponderadores originales o calculados según el diseño de muestreo:

i) Disponibilidad de drenaje y disponibilidad de luz eléctrica. La información reportada por el censo de 1990 se utilizará para calibrar los factores de la encuesta de 1992.

ii) Disponibilidad de drenaje y disponibilidad de luz eléctrica. La información reportada por el conteo de población 1995 se usará como información auxiliar para calibrar los factores de expansión de la ENIGH 1996.

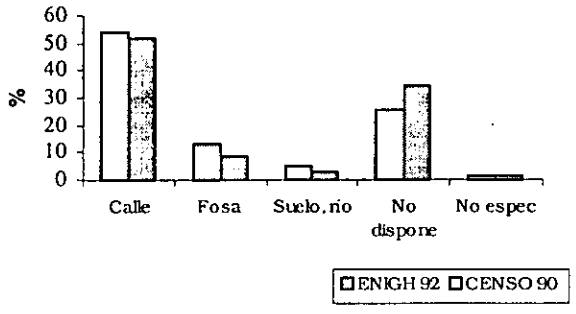
iii) Disponibilidad de agua y material predominante en pisos. Al igual que en el caso anterior, la información reportada por el censo de 1990 se aplicará en la encuesta de 1992.

En las gráficas 3.1 y 3.2 se muestran las diferencias entre los porcentajes estimados y los poblacionales para el primer conjunto de variables. Como se puede observar, la ENIGH 92 subestima la proporción de viviendas que no disponen de drenaje, así como la proporción de las que no cuentan con luz eléctrica.

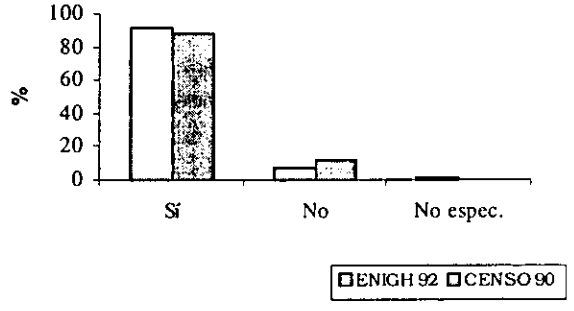
Para este mismo año, la encuesta también subestima la proporción de viviendas cuyo piso es de tierra, así como las que disponen de agua conectada fuera de la vivienda.

[1] Los resultados dados en las secciones 1.5 y 1.6 son fácilmente generalizables a tres o más dimensiones.

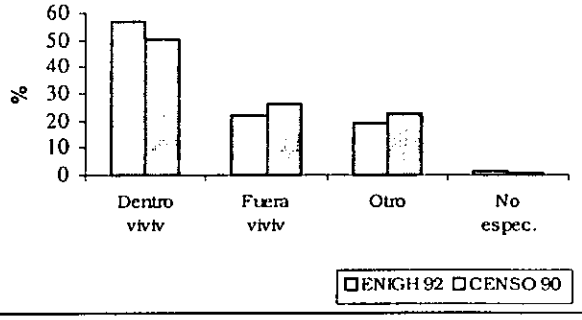
Gráfica 3.1. Disponibilidad de drenaje, según fuente de información



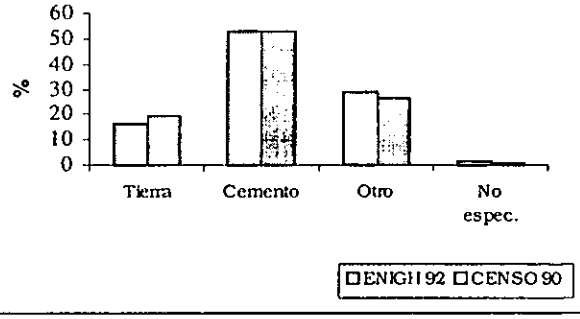
Gráfica 3.2. Disponibilidad de energía eléctrica, según fuente de información



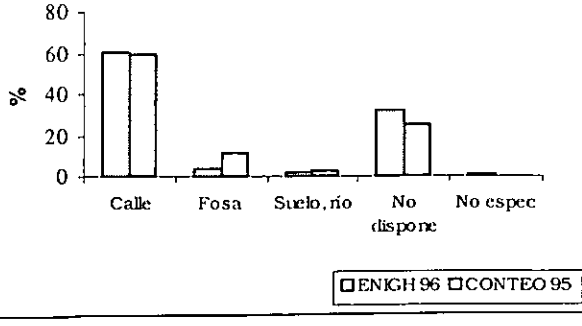
Gráfica 3.3 Disponibilidad de agua entubada, según fuente de información



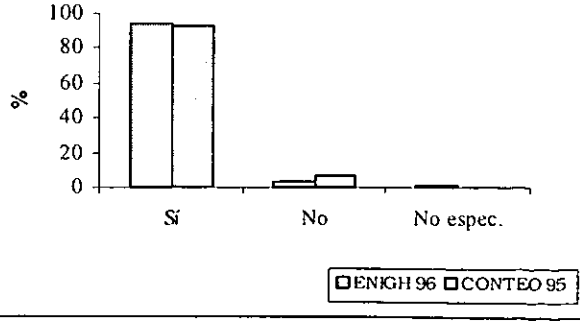
Gráfica 3.4 Material predominante en pisos, según fuente de información



Gráfica 3.5 Disponibilidad de drenaje, según fuente de información



Gráfica 3.6 Disponibilidad de energía eléctrica, según fuente de información



En lo que se refiere a la ENIGH 1996, la encuesta sobreestima el porcentaje de viviendas que disponen de energía eléctrica, así como el porcentaje de las que no disponen de drenaje, en comparación con el Conteo de Población y Vivienda 1995.

3.2. COMPARACIÓN DE LOS FACTORES

En el análisis se utilizaron las nueve medidas de distancia expuestas en el primer capítulo:

i) Mínimos cuadrados generalizados.

ii) *Raking ratio.*

iii) Hellinger.

iv) Entropia mínima.

v) Mínimos cuadrados modificada.

vi) Mínimos cuadrados restringida.

vii) *Logit.*

viii) Huang-Fuller modificada.

ix) Contracción-minimización.

Como se mencionó en ese apartado, el procedimiento requiere de encontrar la solución a un sistema de ecuaciones que, con excepción de la primer función de distancia, es no lineal. Para resolverlo se elaboró un programa en MATLAB[2] en el que se utiliza el método iterativo de Newton (ver sección 1.6).

Aunque no todas las medidas de distancia garantizan la existencia de una solución al sistema de ecuaciones (6), la probabilidad que de que exista tiende a uno (Deville y Särndal, 1992). Sin embargo, en el primer ejercicio no se encontró solución cuando se utilizó la distancia de mínimos cuadrados modificada; en el segundo, cuando se usó la distancia de entropía, así como la de mínimos cuadrados modificada; y en el tercero, para la de contracción-minimización.

Los límites L y U para los factores de ajuste de las distancias mínimos cuadrados restringida, *logit*, Huang-Fuller y contracción-minimización se determinan por ensayo-error, en cada uno de los tres ejercicios se trató que fueran iguales, sin embargo, para algunas funciones el sistema de ecuaciones (6) no tenía solución por lo que se amplió un poco el intervalo. Por ejemplo en la distancia *logit* del segundo ejemplo, el intervalo fue (0.55, 5.9), en comparación con (0.66, 3.6) para las demás.

Adicionalmente a la existencia de una solución, cada una de las medidas presenta ventajas y desventajas, las cuales serán comentadas en el contexto de esta aplicación. La comparación se hará en términos de las características descriptivas de los factores de ajuste, de estimadores y sus varianzas, y por último, de aspectos computacionales.

[2] En el Anexo A se muestran los programas utilizados.

3.2.1. Características descriptivas de los factores de ajuste

Cuando el vector de totales poblacionales corresponde a los totales marginales de una tabla de contingencia, los factores de ajuste o g-factores dependen de la celda a la que pertenece el elemento en cuestión (en este caso la vivienda), por lo tanto tienen la forma:

$$g_k = g\left(\mathbf{x}'_k \lambda\right) = g\left(r_i + c_j\right),$$

donde:

i corresponde a la categoría (renglón) *i* de una de las dos variables
j corresponde a la categoría (columna) *j* de la otra variable.

Los factores de ajuste resultantes se muestran en los cuadros 3.1 a 3.3. Para una categoría dada, existen diferencias importantes sólo en las celdas con tamaño de muestra muy pequeño. En general, para tamaños de muestra más grandes, los factores originados por las diversas medidas son similares.

Como se mencionó anteriormente, es deseable que los ponderadores calibrados estén lo más cerca posible de los originales, es decir, que los g-factores sean cercanos a uno. Bajo esta óptica, no existe un patrón común en los tres diferentes ejercicios. Mientras que cuando las variables de calibración fueron disponibilidad de drenaje y luz (1992), la distancia con valores más alejados de uno fue la de Huang-Fuller; en el ejercicio para 1996, los factores más extremos fueron los de Hellinger, siguiéndole mínimos cuadrados generalizados. La situación es diferente para disponibilidad de agua y pisos (1992), ya que la amplitud de los g-factores por categoría es menor que en los otros dos ejemplos, siendo las distancias de cuadrados mínimos modificada y *logit* las que originaron factores un poco más distantes. Lo anterior se puede explicar porque el menor número de casos fue de 87, mientras que para los otros dos pares de variables fue de 14 casos y tres casos. De hecho, las estimaciones basadas en un tamaño de muestra muy pequeño son de poca utilidad ya que el error de muestreo asociado a ellas es muy grande, recuérdese que la precisión de un estimador y el tamaño de muestra guardan una relación directa.

Tabla 3.1. Factores de ajuste[a] utilizando disponibilidad de drenaje y luz eléctrica, 1992

Categoría	*Mín. Cuad. Gener.*	*Raking Ratio*	*Hellinger*	*Entropía*	*Mín. Cuad. Modif.*	*Logit*	*Mín. Cuad. Restr.*	*Huang - Fuller*	*Contr. - Min.*	*Amplitud*	*Tamaño de muestra*
Fosa, no luz	1.35	1.08	0.96	0.86	---	1.81	1.72	**1.90**	1.89	1.04	111
Suelo, no luz	1.16	0.81	0.69	0.62	---	0.64	0.65	**0.52**	0.62	0.64	71
Calle, no luz	1.68	1.61	1.56	1.49	---	1.88	**1.90**	1.89	1.89	0.41	14
No drenaje, no luz	1.90	1.96	1.98	**2.00**	---	1.89	1.90	1.89	1.89	0.11	813
Suelo, luz	**0.45**	0.48	0.50	0.50	---	0.50	0.50	0.51	0.50	0.07	587
Fosa, luz	0.63	0.65	0.65	0.66	---	**0.61**	0.62	**0.61**	**0.61**	0.05	1571
No drenaje, luz	1.19	1.17	1.16	1.16	---	1.19	1.19	1.19	1.19	0.03	2413
Calle, luz	0.96	0.96	0.96	0.96	---	0.96	0.96	0.96	0.96	0.00	4754

[a] Los límites del intervalo de los g-factores para las funciones *logit*, mínimos cuadrados restringida, Huang-Fuller y contracción-minimización fueron *L=0.5, U=1.9*.
Negritas: valores más alejados de la unidad.

Tabla 3.2. Factores de ajuste[a] utilizando disponibilidad de drenaje y luz eléctrica, 1996

Categoría	*Mín. Cuad. Gener.*	*Raking Ratio*	*Hellinger*	*Entropía*	*Mín. Cuad. Modif.*	*Logit*	*Mín. Cuad. Restr.*	*Huang - Fuller*	*Contr. - Min.*	*Amplitud*	*Tamaño de muestra*
Fosa, no luz	3.78	6.82	**27.04**	---	---	5.32	3.60	2.95	3.12	24.09	3
Suelo,no luz	2.19	2.94	**3.96**	---	---	3.80	2.97	3.11	3.25	1.77	6
Calle, no luz	1.96	2.40	2.81	---	---	**3.21**	2.75	3.20	2.93	1.25	15
No drenaje, no luz	**1.63**	1.61	1.53	---	---	1.59	1.61	1.61	1.61	0.10	725
Fosa, luz	**2.81**	2.80	2.74	---	---	2.80	**2.81**	**2.81**	**2.81**	0.07	875
Suelo, luz	1.21	1.20	1.20	---	---	1.20	1.20	1.20	1.20	0.02	304
No drenaje, luz	0.66	0.66	0.67	---	---	0.66	0.66	0.66	0.66	0.01	4865
Calle, luz	0.98	0.98	0.98	---	---	0.98	0.98	0.98	0.98	0.00	7058

[a] Los límites del intervalo de los g-factores para la función *logit* fueron *L=0.55, U=5.9*; para mínimos cuadrados restringida, Huang-Fuller y contracción-minimización fueron *L=0.66, U=3.6*.
Negritas: valores más alejados de la unidad.

Tabla 3.3. Factores de ajuste utilizando[a] disponibilidad de agua y material en pisos, 1992

Categoría	*Mín. Cuad. Gener.*	*Raking Ratio*	*Hellinger*	*Entropía*	*Mín. Cuad. Modif.*	*Logit*	*Mín. Cuad. Restr.*	*Huang - Fuller*	*Contr. - Min.*	*Amplitud*	*Tamaño de muestra*
Otro agua, otro piso	1.17	1.18	1.18	1.19	1.20	**1.23**	1.21	**1.23**	---	0.06	87
Dentro vivienda, tierra	0.92	0.91	0.90	0.90	0.89	**0.87**	0.89	0.89	---	0.05	257
Fuera vivienda, otro piso	1.21	1.22	1.23	1.23	**1.25**	1.23	1.24	1.23	---	0.04	137
Fuera vivienda, tierra	1.24	1.25	1.25	1.25	**1.26**	1.23	1.23	1.22	---	0.04	603
Otro agua, cemento	1.13	1.13	1.13	1.12	1.12	1.10	1.11	1.10	---	0.03	975
Otro agua, tierra	1.20	1.20	1.20	1.20	1.20	1.22	1.22	1.22	---	0.02	1081
Fuera vivienda, cemento	1.17	1.17	1.17	1.17	1.16	1.17	1.17	1.18	---	0.01	1879
Dentro vivienda, cemento	0.85	0.85	0.85	0.85	0.86	0.86	0.86	0.85	---	0.01	3146
Dentro vivienda, otro piso	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89	---	0.00	2169

[a] Los límites del intervalo para los g-factores de las funciones *logit* y mínimos cuadrados restringida fueron *L=0.855, U=1.235*, para la función Huang-Fuller *L=0.854, U=1.235*.
Negritas: valores más alejados de la unidad.

También es importante explorar las estadísticas descriptivas de los g-factores. Por construcción el valor esperado (media) es igual a uno, como se puede demostrar fácilmente. En lo que se refiere a la mediana, todas las funciones de distancia arrojan factores con medianas muy parecidas. Para los primeros dos ejercicios son muy cercanas a uno, mientras que para el tercero son menores a uno (ver tablas 3.4 a 3.6).

Se puede observar que para cada ejercicio, la varianza de los factores de ajuste entre las diversas medidas es muy similar. Solamente la varianza inducida por la distancia de Hellinger, cuando las variables de calibración fueron luz y drenaje 1996, resultó mayor en comparación con las demás medidas.

En lo que se refiere a la amplitud de los g-factores, como era de esperarse, en general resultó ser menor para las medidas que imponen restricciones a los límites entre los que deben de estar tales factores. La mayoría son muy parecidos, y nuevamente, la distancia de Hellinger es la que representó un intervalo muy grande. Al respecto es importante comentar que aunque tanto la varianza como la amplitud de esta función son altos, sólo tres casos tienen g-factores tan grandes ($g_k = 27.04$), por lo que no se pueden hacer inferencias válidas con un tamaño de muestra tan pequeño.

Al comparar entre los tres ejercicios, la varianza y la amplitud de los factores de ajuste resultaron menores cuando se calibró de acuerdo a disponibilidad de agua y material predominante en pisos, en comparación con los otros dos pares de variables, esto se puede deber a que las diferencias entre los porcentajes marginales poblacionales y los estimados (ponderador original) para esta tabla de contingencia son menores.

Tabla 3.4. Estadísticas descriptivas de los factores de ajuste utilizando disponibilidad de drenaje y luz eléctrica, 1992

Función de distancia	*Media*	*Mediana*	*Varianza*	*Amplitud*	*Mínimo*	*Máximo*	*% < L=0.5*	*% > U=1.9*
Mínimos cuadrados generalizados	1.00	0.96	0.088	1.45	0.45	1.90	4.8	0.0
Raking ratio	1.00	0.96	0.089	1.47	0.48	1.96	4.8	5.7
Hellinger	1.00	0.96	0.090	1.49	0.50	1.98	4.8	5.7
Entropía mínima	1.00	0.96	0.092	1.50	0.50	2.00	0.0	5.7
Mínimos cuadrados modificada	---	---	---	---	---	---	---	---
Logit	1.00	0.96	0.091	1.39	0.50	1.89	0.0	0.0
Mínimos cuadrados restringida	1.00	0.96	0.090	1.40	0.50	1.90	0.0	0.0
Huang - Fuller modificada	1.00	0.96	0.092	1.39	0.51	1.90	0.0	0.0
Contracción - minimización	1.00	0.96	0.091	1.38	0.50	1.89	0.0	0.0

Tabla 3.5. Estadísticas descriptivas de los factores de ajuste utilizando disponibilidad de drenaje y luz eléctrica, 1996

Función de distancia	*Media*	*Mediana*	*Varianza*	*Amplitud*	*Mínimo*	*Máximo*	*%* < *L=.66*	*%* > *U=3.6*
Mínimos cuadrados generalizados	1.00	0.98	0.187	3.12	0.66	3.78	28.1	0.012
Raking ratio	1.00	0.98	0.188	6.16	0.66	6.82	0.0	0.012
Hellinger	1.00	0.98	0.256	26.37	0.67	27.04	0.0	0.035
Entropía mínima	---	---	---	---	---	---	---	---
Mínimos cuadrados modificada	---	---	---	---	---	---	---	---
Logit	1.00	0.98	0.189	4.65	0.66	5.32	0.0	0.035
Mínimos cuadrados restringida	1.00	0.98	0.187	2.94	0.66	3.60	0.0	0.0
Huang - Fuller modificada	1.00	0.98	0.188	2.54	0.66	3.20	0.0	0.0
Contracción - minimización	1.00	0.98	0.188	2.59	0.66	3.25	0.0	0.0

Tabla 3.6. Estadísticas descriptivas de los factores de ajuste utilizando disponibilidad de agua y material predominante en pisos, 1992

Función de Distancia	*Media*	*Mediana*	*Varianza*	*Amplitud*	*Mínimo*	*Máximo*	*%* < *L=.855*	*%* >*U=1.235*
Mínimos cuadrados generalizados	1.00	0.89	0.024	0.40	0.85	1.24	29.2	4.7
Raking ratio	1.00	0.89	0.024	0.40	0.85	1.25	29.2	4.7
Hellinger	1.00	0.89	0.024	0.40	0.85	1.25	29.2	4.7
Entropía mínima	1.00	0.89	0.024	0.40	0.85	1.25	29.2	4.7
Mínimos cuadrados modificada	1.00	0.89	0.024	0.40	0.86	1.26	0.0	6.4
Logit	1.00	0.89	0.024	0.38	0.86	1.23	0.0	0.0
Mínimos cuadrados restringida	1.00	0.89	0.024	0.38	0.86	1.24	0.0	0.0
Huang - Fuller modificada	1.00	0.89	0.024	0.38	0.85	1.23	29.2	0.0
Contracción - minimización	---	---	---	---	---	---	---	---

Tabla 3.7. Porcentaje de los factores de ajuste según el intervalo al que pertenecen, utilizando disponibilidad de drenaje y luz eléctrica, 1992

	Intervalo al que pertenecen los g-factores					
Función de distancia	*[0 - .5)*	*[.5 - 1)*	*[1 - 1.5)*	*[1.5 - 2)*	*[2 - +)*	*Total*
Mínimos cuadrados generalizados	4.8	66.6	22.8	5.8	0.0	100.0
Raking ratio	4.8	67.1	22.3	5.8	0.0	100.0
Hellinger	4.8	67.7	21.7	5.8	0.0	100.0
Entropía mínima	0.0	72.6	21.8	0.0	5.7	100.0
Mínimos cuadrados modificada	---	---	---	---	---	---
Logit	0.0	72.1	21.6	6.4	0.0	100.0
Mínimos cuadrados restringida	0.0	72.0	21.7	6.4	0.0	100.0
Huang - Fuller modificada	0.0	72.0	21.7	6.4	0.0	100.0
Contracción - minimización	0.0	72.0	21.7	6.4	0.0	100.0

Adicionalmente, los factores de ajuste se categorizaron por intervalos (ver tablas 3.7 a 3.9), encontrándose en cada uno de los ejercicios que la mayoría de las distribuciones originadas por las diversas funciones de distancia son muy parecidas. Únicamente en el primero de ellos la distancia de entropía mínima dio lugar a factores más grandes que las demás.

Tabla 3.8. Porcentaje de los factores de ajuste según el intervalo al que pertenecen, utilizando disponibilidad de drenaje y luz eléctrica, 1996

	Intervalo al que pertenecen los g-factores						
Función de distancia	*[0.5 - 1)*	*[1 - 1.5)*	*[1.5 - 2)*	*[2 - 2.5)*	*[2.5 - 3)*	*[3 - +)*	*Total*
Mínimos cuadrados generalizados	88.2	3.7	4.0	0.0	4.1	0.0	100.0
Raking ratio	88.2	3.7	3.9	0.1	4.2	0.0	100.0
Hellinger	88.2	3.7	3.9	0.0	4.2	0.0	100.0
Entropía mínima	---	---	---	---	---	---	---
Mínimos cuadrados modificada	---	---	---	---	---	---	---
Logit	88.2	3.7	3.9	0.0	4.1	0.1	100.0
Mínimos cuadrados restringida	88.2	3.7	3.9	0.0	4.2	0.0	100.0
Huang - Fuller modificada	88.2	3.7	3.9	0.0	4.2	0.1	100.0
Contracción – minimización	88.2	3.7	3.9	0.0	4.2	0.0	100.0

Tabla 3.9. Porcentaje de los factores de ajuste según el intervalo al que pertenecen, utilizando disponibilidad de agua y material de pisos, 1992

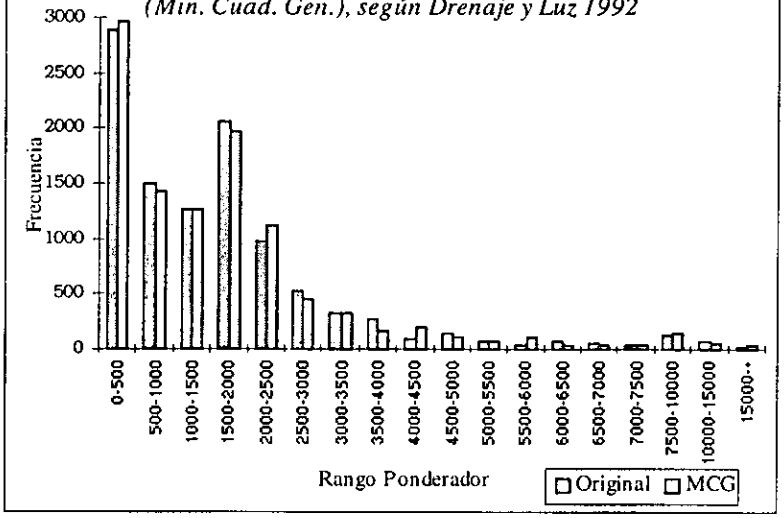
	Intervalo al que pertenecen los g-factores		
Función de distancia	*[0.5 - 1)*	*[1 - 1.5)*	*Total*
Mínimos cuadrados generalizados	57.4	42.6	100.0
Raking ratio	57.4	42.6	100.0
Hellinger	57.4	42.6	100.0
Entropía mínima	57.4	42.6	100.0
Mínimos cuadrados modificada	57.4	42.6	100.0
Logit	57.4	42.6	100.0
Mínimos cuadrados restringida	57.4	42.6	100.0
Huang – Fuller modificada	57.4	42.6	100.0
Contracción - minimización	---	---	---

Otro punto a considerar es la distribución de los pesos calibrados, que en general resultó parecida a la de los originales. Además, la distribución para las diferentes medidas es muy similar (ver gráficas 3.7 a 3.9) [3]. Cuando los ponderadores se corrigieron por agua y pisos (1992), se aprecia

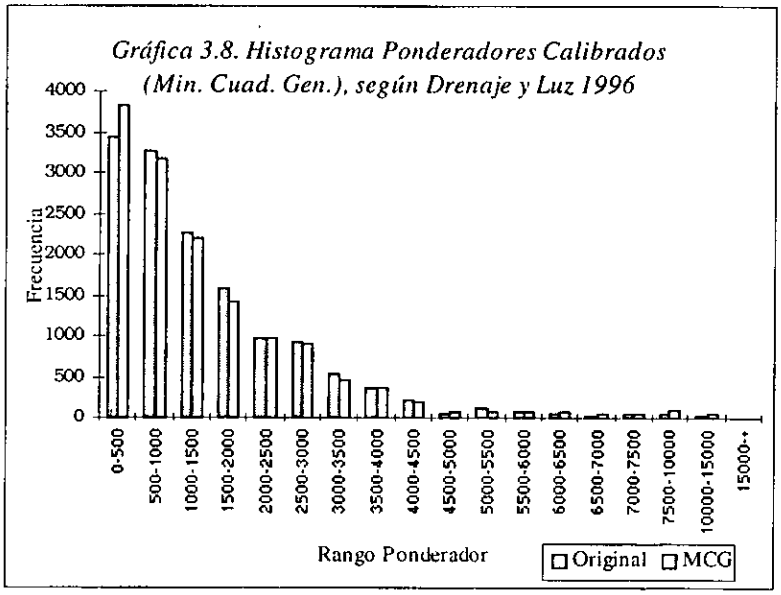
[3] Debido a la similitud entre las gráficas para las demás funciones de distancia se muestran en el Anexo C.

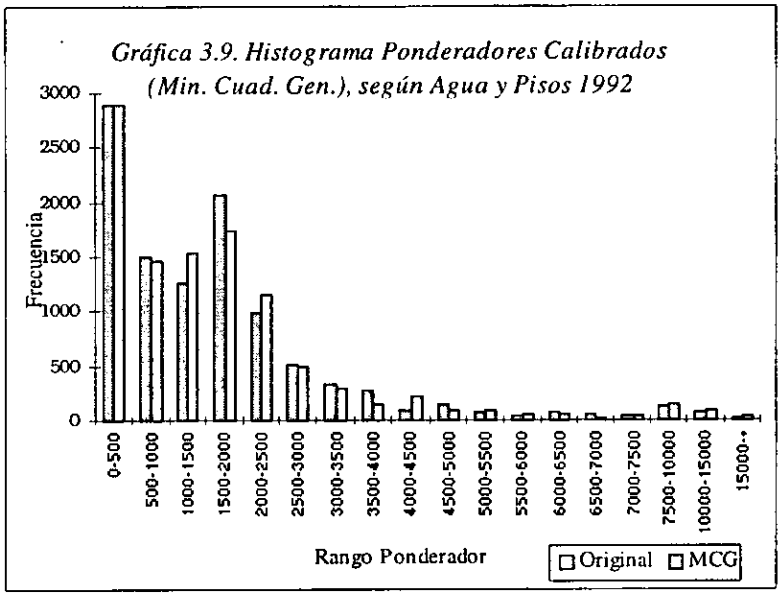
que la nueva distribución es un poco más suave, esto se debe a que algunos de los ponderadores cambian de intervalo.

Gráfica 3.7. Histograma Ponderadores Calibrados (Min. Cuad. Gen.), según Drenaje y Luz 1992



Gráfica 3.8. Histograma Ponderadores Calibrados (Min. Cuad. Gen.), según Drenaje y Luz 1996





3.2.2. Estimadores y varianzas de los estimadores

Uno de los posibles objetivos al ajustar los marginales de una tabla de contingencia a la información censal, es mejorar la estimación de las celdas de la misma. En las tablas 3.10 a 3.12 se presenta el porcentaje estimado del total de cada celda. Como puede observarse, existen diferencias entre los porcentajes estimados usando el ponderador original y los estimados utilizando el ponderador calibrado. Las estimaciones bajo las diversas medidas son parecidas, existiendo mayores diferencias, al menos en términos relativos, en las celdas pequeñas. Esto es consistente con las diferencias entre los g-factores correspondientes a las categorías con tamaños de muestra pequeños (tablas 3.1-3.3).

Tabla 3.10. Porcentajes estimados para disponibilidad de drenaje y luz eléctrica, 1992

	Ponderador original				*Ponderador calibrado usando mínimos cuadrados gen.*			
	Disponibilidad de luz eléctrica				Disponibilidad de luz eléctrica			
Disponibilidad de drenaje	Sí	No	No especif.	Total	Sí	No	No especif.	Total
Conectado al de la calle	54.1	0.1		54.2	52.2	0.2		52.3
Fosa séptica	12.6	0.6		13.2	8.0	0.8		8.8
Desagüe al suelo, río o lago	4.8	0.5		5.3	2.2	0.6		2.7
No dispone	20.2	5.7		25.8	23.9	10.8		34.7
No especificado			1.5	1.5			1.5	1.5
Total	91.6	6.9	1.5	100.0	86.2	12.3	1.5	100.0

Tabla 3.10. Porcentajes estimados para disponibilidad de drenaje y luz eléctrica, 1992 (continuación)

Ponderador calibrado usando raking ratio

Disponibilidad de drenaje	Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total
Conectado al de la calle	52.2	0.1		52.3
Fosa séptica	8.1	0.6		8.8
Desagüe al suelo, río o lago	2.3	0.4		2.7
No dispone	23.6	11.1		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

Ponderador calibrado usando Hellinger

Disponibilidad de drenaje	Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total
Conectado al de la calle	52.2	0.1		52.3
Fosa séptica	8.2	0.6		8.8
Desagüe al suelo, río o lago	2.4	0.3		2.7
No dispone	23.4	11.2		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

Ponderador calibrado usando entropia mínima

Disponibilidad de drenaje	Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total
Conectado al de la calle	52.2	0.1		52.3
Fosa séptica	8.3	0.5		8.8
Desagüe al suelo, río o lago	2.4	0.3		2.7
No dispone	23.3	11.3		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

Ponderador calibrado usando logit, L=.5 U=1.9

Disponibilidad de drenaje	Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total
Conectado al de la calle	52.1	0.2		52.3
Fosa séptica	7.7	1.1		8.8
Desagüe al suelo, río o lago	2.4	0.3		2.7
No dispone	24.0	10.7		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

Ponderador calibrado usando mín. cuad. restr., L=.5 U=1.9

Disponibilidad de drenaje	Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total
Conectado al de la calle	52.1	0.2		52.3
Fosa séptica	7.7	1.0		8.8
Desagüe al suelo, río o lago	2.4	0.3		2.7
No dispone	23.9	10.8		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

Ponderador calibrado usando Huang-Fuller, L=.5 U=1.9

Disponibilidad de drenaje	Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total
Conectado al de la calle	52.1	0.2		52.3
Fosa séptica	7.6	1.1		8.8
Desagüe al suelo, río o lago	2.5	0.3		2.7
No dispone	24.0	10.7		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

Ponderador calibrado usando contracción-minimización, L=.5 U=1.9

Disponibilidad de drenaje	Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total
Conectado al de la calle	52.1	0.2		52.3
Fosa séptica	7.6	1.1		8.8
Desagüe al suelo, río o lago	2.4	0.3		2.7
No dispone	24.0	10.7		34.7
No especificado			1.5	1.5
Total	86.2	12.3	1.5	100.0

Tabla 3.10a. Porcentajes estimados para disponibilidad de drenaje y luz eléctrica, 1992

Categoría	*Original*	*Mín. Cuad. Gener.*	*Raking Ratio*	*Hellinger*	*Entropía*	*Mín. Cuad. Modif.*	*Logit*	*Mín. Cuad. Restr.*	*Huang - Fuller*	*Contr. – Mín.*
Calle, luz	54.1	52.2	52.2	52.2	52.2	---	52.1	52.1	52.1	52.1
Fosa, luz	12.6	8.0	8.1	8.2	8.3	---	7.7	7.7	7.6	7.6
Suelo, luz	4.8	2.2	2.3	2.4	2.4	---	2.4	2.4	2.5	2.4
No drenaje, luz	20.2	23.9	23.6	23.4	23.3	---	24.0	23.9	24.0	24.0
Calle, no luz	0.1	0.2	0.1	0.1	0.1	---	0.2	0.2	0.2	0.2
Fosa, no luz	0.6	0.8	0.6	0.6	0.5	---	1.1	1.0	1.1	1.1
Suelo, no luz	0.5	0.6	0.4	0.3	0.3	---	0.3	0.3	0.3	0.3
No drenaje, no luz	5.7	10.8	11.1	11.2	11.3	---	10.7	10.8	10.7	10.7

Tabla 3.11. Porcentajes estimados para disponibilidad de drenaje y luz eléctrica, 1996

	Ponderador original				*Ponderador calibrado usando mínimos cuadrados gen.*			
	Disponibilidad de luz eléctrica				Disponibilidad de luz eléctrica			
Disponibilidad de drenaje	Sí	No	No especif.	Total	Sí	No	No especif.	Total
Conectado al de la calle	60.1	0.1		60.2	59.1	0.1		59.3
Fosa séptica	4.1	0.0		4.2	11.6	0.0		11.7
Desagüe al suelo, río o lago	2.4	0.0		2.4	2.9	0.0		2.9
No dispone	28.1	3.9		32.0	18.5	6.3		24.8
No especificado			1.3	1.3			1.3	1.3
Total	94.7	4.0	1.3	100.0	92.1	6.6	1.3	100.0

	Ponderador calibrado usando raking ratio				*Ponderador calibrado usando Hellinger*			
	Disponibilidad de luz eléctrica				Disponibilidad de luz eléctrica			
Disponibilidad de drenaje	Sí	No	No especif.	Total	Sí	No	No especif.	Total
Conectado al de la calle	59.1	0.2		59.3	59.1	0.2		59.3
Fosa séptica	11.6	0.1		11.7	11.3	0.3		11.7
Desagüe al suelo, río o lago	2.9	0.1		2.9	2.8	0.1		2.9
No dispone	18.6	6.2		24.8	18.9	5.9		24.8
No especificado			1.3	1.3			1.3	1.3
Total	92.1	6.6	1.3	100.0	92.1	6.6	1.3	100.0

Tabla 3.11. Porcentajes estimados para disponibilidad de drenaje y luz eléctrica, 1996 (continuación)

Disponibilidad de drenaje	*Ponderador calibrado usando logit, L=0.55 U=5.9*				*Ponderador calibrado usando mín. cuad. restr., L=0.66 U=3.6*			
	Disponibilidad de luz eléctrica				Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total	Sí	No	No especif.	Total
Conectado al de la calle	59.1	0.2		59.3	59.1	0.2		59.3
Fosa séptica	11.6	0.1		11.7	11.6	0.0		11.7
Desagüe al suelo, río o lago	2.9	0.1		2.9	2.9	0.1		2.9
No dispone	18.6	6.2		24.8	18.5	6.3		24.8
No especificado			1.3	1.3			1.3	1.3
Total	92.1	6.6	1.3	100.0	92.1	6.6	1.3	100.0

Disponibilidad de drenaje	*Ponderador calibrado usando Huang-Fuller, L=0.66 U=3.6*				*Ponderador calibrado usando contracción-minimización, L=0.66 U=3.6*			
	Disponibilidad de luz eléctrica				Disponibilidad de luz eléctrica			
	Sí	No	No especif.	Total	Sí	No	No especif.	Total
Conectado al de la calle	59.1	0.2		59.3	59.1	0.2		59.3
Fosa séptica	11.6	0.0		11.7	11.6	0.0		11.7
Desagüe al suelo, río o lago	2.9	0.1		2.9	2.9	0.1		2.9
No dispone	18.6	6.2		24.8	18.6	6.2		24.8
No especificado			1.3	1.3			1.3	1.3
Total	92.1	6.6	1.3	100.0	92.1	6.6	1.3	100.0

Tabla 3.11a. Porcentajes estimados para disponibilidad de drenaje y luz eléctrica, 1996

Categoría	*Original*	*Mín. Cuad. Gener.*	*Raking Ratio*	*Hellinger*	*Entropía*	*Mín. Cuad. modif..*	*Logit*	*Mín. Cuad. Restr.*	*Huang - Fuller*	*Contr. - Min.*
Calle, luz	60.1	59.1	59.1	59.1	---	---	59.1	59.1	59.1	59.1
Fosa, luz	4.1	11.6	11.6	11.3	---	---	11.6	11.6	11.6	11.6
Suelo, luz	2.4	2.9	2.9	2.8	---	---	2.9	2.9	2.9	2.9
No drenaje, luz	28.1	18.5	18.6	18.9	---	---	18.6	18.5	18.6	18.6
Calle, no luz	0.1	0.1	0.2	0.2	---	---	0.2	0.2	0.2	0.2
Fosa, no luz	0.0	0.0	0.1	0.3	---	---	0.1	0.0	0.0	0.0
Suelo, no luz	0.0	0.0	0.1	0.1	---	---	0.1	0.1	0.1	0.1
No drenaje, no luz	3.9	6.3	6.2	5.9	---	---	6.2	6.3	6.2	6.2

Tabla 3.12. Porcentajes estimados para disponibilidad de agua y material en pisos, 1992

	Ponderador original					*Ponderador calibrado usando mínimos cuadrados gen.*				
	Material predominante en pisos					Material predominante en pisos				
Disponibilidad de agua	Tierra	Cemento	Otro	No especif.	Total	Tierra	Cemento	Otro	No especif.	Total
Dentro de la vivienda	2.3	29.2	26.0		57.4	2.1	24.8	23.1		50.0
Fuera de la vivienda	4.7	15.3	1.8		21.8	5.8	18.0	2.2		25.9
Otro	9.5	8.9	1.0		19.3	11.4	10.0	1.1		22.6
No especificado				1.5	1.5				1.5	1.5
Total	16.4	53.4	28.7	1.5	100.0	19.3	52.8	26.5	1.5	100.0

	Ponderador calibrado usando raking ratio					*Ponderador calibrado usando Hellinger*				
	Material predominante en pisos					Material predominante en pisos				
Disponibilidad de agua	Tierra	Cemento	Otro	No especif.	Total	Tierra	Cemento	Otro	No especif.	Total
Dentro de la vivienda	2.1	24.8	23.1		50.0	2.0	24.9	23.1		50.0
Fuera de la vivienda	5.8	17.9	2.2		25.9	5.8	17.9	2.2		25.9
Otro	11.4	10.0	1.2		22.6	11.4	10.0	1.2		22.6
No especificado				1.5	1.5				1.5	1.5
Total	19.3	52.8	26.5	1.5	100.0	19.3	52.8	26.5	1.5	100.0

	Ponderador calibrado usando entropía mínima					*Ponderador calibrado usando mínimos cuadrados modificada*				
	Material predominante en pisos					Material predominante en pisos				
Disponibilidad de agua	Tierra	Cemento	Otro	No especif.	Total	Tierra	Cemento	Otro	No especif.	Total
Dentro de la vivienda	2.0	24.9	23.1		50.0	2.0	24.9	23.0		50.0
Fuera de la vivienda	5.8	17.9	2.2		25.9	5.9	17.8	2.2		25.9
Otro	11.4	10.0	1.2		22.6	11.4	10.0	1.2		22.6
No especificado				1.5	1.5				1.5	1.5
Total	19.3	52.8	26.5	1.5	100.0	19.3	52.8	26.5	1.5	100.0

Tabla 3.12. Porcentajes estimados para disponibilidad de agua y material en pisos, 1992 (continuación)

Disponibilidad de agua	Ponderador calibrado usando logit, L=.855 U=1.235					Ponderador calibrado usando mín. cuad. restr., L=.855 U=1.235				
	Material predominante en pisos					Material predominante en pisos				
	Tierra	Cemento	Otro	No especif.	Total	Tierra	Cemento	Otro	No especif.	Total
Dentro de la vivienda	2.0	25.0	23.1		50.0	2.0	24.9	23.1		50.0
Fuera de la vivienda	5.7	18.0	2.2		25.9	5.8	18.0	2.2		25.9
Otro	11.6	9.8	1.2		22.6	11.5	9.9	1.2		22.6
No especificado				1.5	1.5				1.5	1.5
Total	19.3	52.8	26.5	1.5	100.0	19.3	52.8	26.5	1.5	100.0

Disponibilidad de agua	Ponderador calibrado usando Huang-Fuller, L=.855 U=1.235				
	Material predominante en pisos				
	Tierra	Cemento	Otro	No especif.	Total
Dentro de la vivienda	2.0	24.9	23.1		50.0
Fuera de la vivienda	5.7	18.1	2.2		25.9
Otro	11.6	9.8	1.2		22.6
No especificado				1.5	1.5
Total	19.3	52.8	26.5	1.5	100.0

Tabla 3.12a. Porcentajes estimados para disponibilidad de agua y material en pisos, 1992

Categoría	*Original*	*Mín. Cuad. Gener.*	*Raking Ratio*	*Hellinger*	*Entropía*	*Mín. Cuad. Modif.*	*Logit*	*Mín. Cuad. Restr.*	*Huang - Fuller*
Dentro vivienda, tierra	2.3	2.1	2.1	2.0	2.0	2.0	2.0	2.0	2.0
Fuera vivienda, tierra	4.7	5.8	5.8	5.8	5.8	5.9	5.7	5.8	5.7
Otro agua, tierra	9.5	11.4	11.4	11.4	11.4	11.4	11.6	11.5	11.6
Dentro vivienda, cemento	29.2	24.8	24.8	24.9	24.9	24.9	25.0	24.9	24.9
Fuera vivienda, cemento	15.3	18.0	17.9	17.9	17.9	17.8	18.0	18.0	18.1
Otro agua, cemento	8.9	10.0	10.0	10.0	10.0	10.0	9.8	9.9	9.8
Dentro vivienda, otro piso	26.0	23.1	23.1	23.1	23.1	23.0	23.1	23.1	23.1
Fuera vivienda, otro piso	1.8	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2
Otro agua, otro piso	1.0	1.1	1.2	1.2	1.2	1.2	1.2	1.2	1.2

Otra de las metas que persigue la calibración es mejorar la estimación de una característica que se supone está relacionada con las variables usadas para calibrar. En este sentido se eligió como variable a estimar la existencia de teléfono en la vivienda, la cual se cree está relacionada con los servicios con los que cuenta la misma.

Como era de esperarse, dado el tamaño de muestra, los porcentajes estimados son muy parecidos[4], independientemente del tipo de ajuste a los ponderadores. En el caso de la encuesta efectuada en 1992 y para los dos ejercicios realizados, la proporción estimada de viviendas con teléfono resultó menor a la estimada originalmente. En 1996, el porcentaje estimado bajo la técnica de calibración fue mayor que el estimado con los ponderadores originales.

Tabla 3.13. Total y porcentaje estimado para disponibilidad de teléfono, utilizando drenaje y luz eléctrica, 1992

Función de distancia	*Total estimado*		*Porcentaje estimado*	
	Sí	*No*	*Sí*	*No*
Original	4,147,383	13,403,447	23.63%	76.37%
Mínimos cuadrados generalizados	3,914,837	13,635,992	22.31%	77.69%
Raking ratio	3,918,672	13,632,157	22.33%	77.67%
Hellinger	3,920,580	13,630,249	22.34%	77.66%
Entropía mínima	3,922,232	13,628,597	22.35%	77.65%
Mínimos cuadrados modificada	---	---	---	---
Logit	3,909,038	13,641,791	22.27%	77.73%
Mínimos cuadrados restringida	3,909,925	13,640,904	22.28%	77.72%
Huang - Fuller modificada	3,908,125	13,642,704	22.27%	77.73%
Contracción - minimización	3,908,201	13,642,628	22.27%	77.73%

Tabla 3.14. Total y porcentaje estimado para disponibilidad de teléfono, utilizando drenaje y luz eléctrica, 1996

Función de distancia	*Total estimado*		*Porcentaje estimado*	
	Sí	*No*	*Sí*	*No*
Original	5,860,726	14,338,673	29.01%	70.99%
Mínimos cuadrados generalizados	6,312,240	13,887,158	31.25%	68.75%
Raking ratio	6,306,419	13,892,979	31.22%	68.78%
Hellinger	6,285,427	13,913,971	31.12%	68.88%
Entropía mínima	---	---	---	---
Mínimos cuadrados modificada	---	---	---	---
Logit	6,302,102	13,897,296	31.20%	68.80%
Mínimos cuadrados restringida	6,306,944	13,892,454	31.22%	68.78%
Huang - Fuller modificada	6,304,540	13,894,858	31.21%	68.79%
Contracción - minimización	6,306,082	13,893,316	31.22%	68.78%

[4] Deville, Särndal (1992) demuestran que bajo ciertas condiciones de regularidad los estimadores bajo las diversas funciones de distancia son asintóticamente equivalentes al de mínimos cuadrados generalizados o estimador de regresión generalizado.

Tabla 3.15. Total y porcentaje estimado para disponibilidad de teléfono, utilizando agua y pisos, 1992

Función de distancia	*Total estimado*		*Porcentaje estimado*	
	Sí	*No*	*Sí*	*No*
Original	4,147,383	13,403,447	23.63%	76.37%
Mínimos cuadrados generalizados	3,707,558	13,843,272	21.12%	78.88%
Raking ratio	3,707,068	13,843,762	21.12%	78.88%
Hellinger	3,706,616	13,844,214	21.12%	78.88%
Entropía mínima	3,706,016	13,844,814	21.12%	78.88%
Mínimos cuadrados modificada	3,704,352	13,846,478	21.11%	78.89%
Logit	3,706,488	13,844,342	21.12%	78.88%
Mínimos cuadrados restringida	3,705,664	13,845,166	21.11%	78.89%
Huang - Fuller modificada	3,705,797	13,845,033	21.11%	78.89%
Contracción - minimización	---	---	---	---

Tabla 3.16. Error estándar estimado para el total y el porcentaje de viviendas con teléfono, utilizando drenaje y luz eléctrica, 1992

Función de distancia	*Total*		*Porcentaje*	
	Sí	*No*	*Sí*	*No*
Original	230,993	230,993	1.3161%	1.3161%
Mínimos cuadrados generalizados	195,696	195,696	1.1150%	1.1150%
Raking ratio	195,760	195,760	1.1154%	1.1154%
Hellinger	195,787	195,787	1.1155%	1.1155%
Entropía mínima	195,814	195,814	1.1157%	1.1157%
Mínimos cuadrados modificada	---	---	---	---
Logit	195,809	195,809	1.1157%	1.1157%
Mínimos cuadrados restringida	196,710	196,710	1.1208%	1.1208%
Huang - Fuller modificada	196,589	196,589	1.1201%	1.1201%
Contracción - minimización	195,448	195,448	1.1136%	1.1136%

Otro de los atractivos de la técnica propuesta es mejorar la precisión de las estimaciones. Por ello se estimaron las varianzas del total y de la proporción utilizando el método de Jackknife (ver sección 1.4). La ganancia en términos de precisión se corrobora al observar que el error estándar estimado tanto para el total, como para la proporción de viviendas con teléfono[5], es menor cuando se utilizan los pesos calibrados en comparación a cuando se usan los pesos originales. Asimismo,

[5] No se estimaron las varianzas utilizando la encuesta de 1996 porque no se contó con un identificador de la unidad primaria de muestreo (ver sección 1.4).

estos errores son muy parecidos, independientemente del tipo de medida utilizada. No obstante, para los dos ejercicios evaluados, la distancia de mínimos cuadrados restringa resultó ser ligeramente mayor que las demás; en contraste, la de mínimos cuadros generalizados (disponibilidad de agua y pisos) y la de contracción-minimización (disponibilidad de drenaje y luz) fueron las menores.

Tabla 3.17. Error estándar estimado para el total y el porcentaje de viviendas con teléfono , utilizando agua y pisos, 1992

Función de distancia	*Total*		*Porcentaje*	
	Sí	*No*	*Sí*	*No*
Original	230,993	230,993	1.3161%	1.3161%
Mínimos cuadrados generalizados	151,332	151,332	0.8623%	0.8623%
Raking ratio	151,630	151,630	0.8639%	0.8639%
Hellinger	151,843	151,843	0.8652%	0.8652%
Entropía mínima	152,113	152,113	0.8667%	0.8667%
Mínimos cuadrados modificada	152,874	152,874	0.8710%	0.8710%
Logit	153,764	153,764	0.8761%	0.8761%
Mínimos cuadrados restringida	154,590	154,590	0.8808%	0.8808%
Huang - Fuller modificada	153,537	153,537	0.8748%	0.8748%
Contracción – minimización	---	---	---	---

3.2.3. Aspectos computacionales

No se observó un patrón sistemático entre la función de distancia y el número de iteraciones requeridas por el algoritmo para encontrar una solución. Sin embargo, el método más sencillo de desarrollar y que sólo necesita una iteración es el de mínimos cuadrados generalizados, los demás métodos requieren mayor número de iteraciones. No obstante, el tiempo que se requiere para efectuar una iteración adicional no es significativo.

Por otra parte, las distancias que involucran restricciones al intervalo de los g-factores presentaron algunos problemas. Primero, los límites se determinan por ensayo y error, es decir, se prueba para diferentes valores hasta que el sistema tiene solución. Además, los límites no pueden ser muy diferentes a los originados por las distancias sin restricciones, ya que si son muy estrechos, no existe solución. Por ejemplo, al calibrar por luz y drenaje (1996), para la distancia *logit* se tuvo que usar una amplitud mayor que para las demás distancias (0.55, 5.9 vs 0.66, 3.6), de lo contrario el sistema de ecuaciones no tenía solución.

Lo anterior se complica cuando se calcula la varianza del estimador, ya que aunque el sistema tenga solución para la totalidad de la muestra, puede ocurrir que al quitar una unidad primaria de muestreo (UPM), como lo indica el método de Jackknife, el sistema de ecuaciones de calibración (6) ya no tenga solución. En la práctica esto se puede remediar aumentando la amplitud para los factores de ajuste. En el ejercicio que se llevó a cabo en este trabajo, la encuesta comprendió alrededor de 741 unidades primarias de muestreo, de las cuales al quitar aproximadamente 25 de ellas (una por una) no se encontró solución bajo los límites L y U preestablecidos inicialmente, se

procedió entonces a incrementar (lo menos posible) los límites para tales UPM's hasta que existiese una solución.

Tabla 3.18. Número de iteraciones requeridas por el algoritmo para encontrar la solución

Función de distancia	*Disponibilidad de drenaje y luz eléctrica, 1992*	*Disponibilidad de drenaje y luz eléctrica, 1996*	*Disponibilidad de agua y material predominante en pisos, 1992*
Mínimos cuadrados generalizados	1	1	1
Raking ratio	6	6	4
Hellinger	7	11	4
Entropía mínima	9	---	4
Mínimos cuadrados modificada	---	---	5
Logit	9	9	8
Mínimos cuadrados restringida	5	3	4
Huang - Fuller modificada	14	4	4
Contracción - minimización	7	4	---

4. CONCLUSIONES

Los estimadores que utilizan ponderadores calibrados se han desarrollado con el objetivo de obtener mejores estimaciones de los parámetros de interés. Sin embargo, la ganancia en precisión tiene un costo, por una parte, la técnica implica trabajo adicional para obtener la solución al sistema de ecuaciones, y por otra, se debe contar con la información censal o auxiliar. Afortunadamente, el primero se ve aminorado con la disponibilidad de paquetes computacionales adecuados, y el segundo con la existencia de centros de información estadística como el INEGI. Al respecto, hay que considerar la oportunidad con la que se obtienen los resultados censales, afortunadamente con el desarrollo en el campo informático cada vez se reducen más los tiempos de publicación, por ejemplo los resultados del X Censo de Población (1980) fueron publicados seis años más tarde, mientras que los del XI Censo tan sólo dos años después a su levantamiento.

Una vez que se cuenta con las herramientas de cómputo y de información necesarias, surge una interrogante ¿cuál medida de distancia utilizar? Como en muchas ocasiones, la respuesta no es única, ya que cada una de ellas presenta ventajas y desventajas. No obstante, la decisión deberá de tomarse considerando diversos puntos como las características de los nuevos ponderadores, la existencia de una solución y aspectos computacionales.

En el contexto de la aplicación que se hizo en este trabajo, de manera general se observó que en términos de las características de los factores de ajuste (mediana, rango, varianza, etc.), no existen diferencias sustanciales entre las diversas medidas, y donde existen las diferencias, el tamaño de muestra asociado a la celda es muy pequeño. Así, no se identificó un patrón para alentar o descartar alguna de las funciones.

En cambio, la existencia de una solución es de vital importancia. Desde el punto de vista teórico, la distancia de mínimos cuadrados generalizados y la distancia *raking ratio* garantizan la existencia de una solución. Aunque la probabilidad de que las otras medidas tengan solución tiende a uno conforme se incrementa el tamaño de muestra, en este trabajo no se encontró solución para tres funciones: mínimos cuadrados (modificada), entropía y contracción-minimización.

A pesar de que la distancia de mínimos cuadrados generalizados puede dar lugar a pesos negativos y la de *raking ratio* a pesos muy grandes, en ninguno de los ejercicios la primera generó ponderadores negativos, ni la segunda pesos excesivamente grandes.

Si bien es cierto que las distancias con límites para los g-factores evitan pesos negativos y con valores extremos, computacionalmente dichas distancias presentaron algunos problemas. Primero, porque mediante ensayo-error se deben encontrar los límites para los cuales existe una solución; segundo, al calcular la varianza para algunas unidades primarias de muestreo se deben de cambiar los límites para que el sistema tenga solución; finalmente, al menos en esta aplicación, los límites no resultaron muy diferentes a los de las distancias sin restricciones a los mismos.

Hay que resaltar, que independientemente de la función de distancia que se haya utilizado, los estimadores con ponderadores calibrados tuvieron errores estándar más pequeños que el estimador que empleó el ponderador original, es decir, se ganó en términos de precisión. En todos los casos se utilizó el método de Jackknife para estimar la varianza, el cual proporciona un estimador global de la misma. Sin embargo, en ocasiones es necesario estimar la contribución de las diversas etapas de

muestreo en la varianza, en cuyo caso se deberá de usar el método de linealización de Taylor que es el único que permite estimar cada uno de estos componentes.

Si bien anteriormente se esbozan algunos límites y alcances de las diversas funciones de distancia, queda mucho trabajo por realizar con el propósito de obtener conclusiones más certeras y generales, como por ejemplo, probar otras variables para calibrar, considerar tablas de mayores dimensiones y utilizar variables continuas.

No obstante, dado que las características de los factores producidos bajo las diferentes funciones de distancia no son muy diferentes, en mi opinión, la existencia de una solución y los aspectos computacionales deberían de tener mayor importancia. Bajo esta óptica, la distancia de mínimos cuadrados generalizados o bien la de *raking ratio*, que asegura pesos positivos, podrían ser las mejores opciones. Sin embargo, el estadístico o el técnico encargado de tomar la decisión deberá de poner en la balanza los puntos anteriores, así como los recursos con los que cuenta a fin de elegir la mejor distancia para el problema de su interés.

Finalmente, es importante señalar que en la medida en que se apliquen herramientas y métodos estadísticos sustentados en la teoría, se aumentará la calidad de las estimaciones y por ende la credibilidad en las encuestas como un medio para analizar, conocer y entender en mayor grado diferentes problemas de nuestro entorno.

ANEXO A
ALGORITMOS COMPUTACIONALES

A1. Programa para leer los datos: VIVI92VA.M

```
%Totales de la fuente auxiliar (cuatro primeras para drenaje y dos siguientes para luz 92)
T=[9320579; 1563147; 487683; 6179420; 15360174; 2190655 ];

%Totales de la fuente auxiliar (variables agua  y pisos)
%T=[8908785; 4619985; 4022060; 3434213; 9402723; 4713894];

%Totales de la fuente auxiliar (drenaje y luz 96)
%T=[12136241; 2386375; 601507; 5075275; 18853058; 1346340];


%Lectura de datos
%Datos de la encuesta en el archivo vivi92va.m.
%La matriz de datos se denomina MAT, con variables en el siguiente orden :
%dren1, dren2, dren3, dren4, luz1, luz2 (variables dummy),
%factor de expansion, folio, num. de upm, tel1, tel2 (ultimas 2 dummy para la variable y)
vivi92va


% Condiciones iniciales
% tol   Margen de error para los totales marginales calibrados
% r     Número de renglones
% c     Número de columnas

tol=.03;
r=4;
c=2;
tam2=size(MAT);
r2=tam2(1);
c2=tam2(2);
X=MAT(:,1:c2-5);
WEI=MAT(:,c2-4);
WEI2=MAT(:,c2-4);
UPM=MAT(:,c2-2);
YCAL1=zeros(750,1);
YCAL2=zeros(750,1);
NOITER=zeros(750,1);
```

A2. Programa que calcula los g-factores utilizando el método de Newton Distancia mínimos cuadrados generalizados: CUAD2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% iter        Número de iteración
% NES         Vector con los totales por celda (ponderador original)
% NCAL        Vector con los totales por celda calibrados
% TCAL        Vector con los totales por categoría calibrados
% L           Vector de multiplicadores de Lagrange

iter=0;
NES=zeros(r*c,1);
L=zeros(r+c,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);


%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
  for j=1:c
    for z=1:r2
      TEMPO(z)=X(z,i)*X(z,j+r);
    end
    tempo2=dot(TEMPO,WEI);
    NES(p)=tempo2;
    p=p+1;
  end
end


%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];
```

```
% Iteraciones para encontrar L
while  any(abs(TCAL-T)>tol) & iter<1

%Estimación de los totales calibrados (antes de la iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p);
end
TCAL=(NCAL'*MATDIS)'

%Construcción de la matriz derivada de phi
PHIDE=zeros(r+c,r+c);

%Totales por celda ajustados por la derivada de la función
L2=MATDIS*L;
for p=1:r*c
 NCALDE(p)=NES(p);
end

%Totales por categoría
TCALDE=(NCALDE'*MATDIS)';

for i=1:r+c
 PHIDE(i,i)=TCALDE(i);
end

for i=1:r
 for j=1:c
  for  p=1:r*c
   TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
  end
  tempo4=dot (NCALDE,TEMPO3);
  PHIDE (i,j+r)=tempo4;
  PHIDE (j+r,i)=tempo4;
 end
end


%Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
SPHIDE=zeros(r+c-1,r+c-1);
for i=1:r+c-1
 for j=1:r+c-1
  SPHIDE(i,j)=PHIDE(i,j);
 end
end

SPHIDEIN=inv(SPHIDE);
```

```
for i=1:r+c-1
 ST(i)=T(i);
 STCAL(i)=TCAL(i);
end

%Se obtiene el vector de multiplicadores de Lagrange
SL=SL+SPHIDEIN*(ST-STCAL);
for i=1:r+c-1
   L(i)=SL(i);
end

%Estimación de los totales para verificar criterio de convergencia (después de iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p);
end
TCAL=(NCAL'*MATDIS)';

iter=iter+1;

end

% Despliegue de resultados
format long
L
TCAL
iter

%Estimación de los totales para la variable y
TEMPO5=X*L;

for i=1:r2
 WCAL(i)=WEI(i)*(1+TEMPO5(i));
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A3. Programa para calcular la varianza por el método Jackknife: VAR2CUAD.M

```
% Número de la UPM
nupm=0;

for l=1:749
 nupm=nupm+1
 for k=1:r2
  if UPM(k)==nupm
   WEI(k)=0;
  end
  if UPM(k)>=1 & UPM(k)<=181 &  nupm>=1 & nupm<=181
   WEI(k)=181/180*WEI(k);
  end
  if UPM(k)>=182 & UPM(k)<=328 & nupm>=182 & nupm<=328
   WEI(k)=147/146*WEI(k);
  end
  if UPM(k)>=329 & UPM(k)<=443 & nupm>=329 & nupm<=443
   WEI(k)=115/114*WEI(k);
  end
  if UPM(k)>=444 & UPM(k)<=529 & nupm>=444 & nupm<=529
   WEI(k)=86/85*WEI(k);
  end
  if UPM(k)>=530 & UPM(k)<=749 & nupm>=530 & nupm<=749
   WEI(k)=220/219*WEI(k);
  end
 end

%Calcular los ponderadores calibrados
  cuad2
 YCAL1(nupm)=ycal1;
 YCAL2(nupm)=ycal2;
 NOITER(nupm)=iter;
 WEI=WEI2;

end

nupm=nupm+1
cuad2

YCAL1(nupm)=ycal1;
YCAL2(nupm)=ycal2;
WEI=WEI2;

%Se graban los resultados
save cuad1.txt YCAL1 /ascii /double
save cuad2.txt YCAL2 /ascii /double
save cuadite.txt NOITER /ascii /double
```

A4. Programa que calcula los g-factores utilizando el método de Newton Distancia raking ratio: RAKI2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol      Margen de error para los totales marginales calibrados
% NES      Vector con los totales por celda (ponderador original)
% NCAL     Vector con los totales por celda calibrados
% TCAL     Vector con los totales por categoría calibrados
% L        Vector de multiplicadores de Lagrange

iter=0;
NES=zeros(r*c,1);
L=zeros(r+c,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);

%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
   for j=1:c
    for z=1:r2
       TEMPO(z)=X(z,i)*X(z,j+r);
    end
    tempo2=dot(TEMPO,WEI);
    NES(p)=tempo2;
    p=p+1;
   end
end

%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];
```

```
% Iteraciones para encontrar L
while  any(abs(TCAL-T)>tol) & iter<50

%Estimación de los totales calibrados (antes de la iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*exp(L2(p));
end
TCAL=(NCAL'*MATDIS)'

%Construcción de la matriz derivada de phi
PHIDE=zeros(r+c,r+c);

%Totales por celda ajustados por la derivada de la función
L2=MATDIS*L;
for p=1:r*c
 NCALDE(p)=NES(p)*exp(L2(p));
end

%Totales por categoría
TCALDE=(NCALDE'*MATDIS)';

for i=1:r+c
 PHIDE(i,i)=TCALDE(i);
end

for i=1:r
 for j=1:c
  for  p=1:r*c
   TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
  end
  tempo4=dot (NCALDE,TEMPO3);
  PHIDE (i,j+r)=tempo4;
  PHIDE (j+r,i)=tempo4;
 end
end


%Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
SPHIDE=zeros(r+c-1,r+c-1);
for i=1:r+c-1
 for j=1:r+c-1
  SPHIDE(i,j)=PHIDE(i,j);
 end
end

SPHIDEIN=inv(SPHIDE);
```

```
for i=1:r+c-1
 ST(i)=T(i);
 STCAL(i)=TCAL(i);
end

%Se obtiene el vector de multiplicadores de Lagrange
SL=SL+SPHIDEIN*(ST-STCAL);
for i=1:r+c-1
   L(i)=SL(i);
end

%Estimación de los totales para verificar criterio de convergencia (después de iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*exp(L2(p));
end
TCAL=(NCAL'*MATDIS)';
iter=iter+1;

end

% Despliegue de resultados
format long
L
TCAL
iter

%Estimación de los totales para la variable y
TEMPO5=X*L;
for i=1:r2
 WCAL(i)=WEI(i)*exp(TEMPO5(i));
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A5. Programa que calcula los g-factores utilizando el método de Newton Distancia Hellinger: HELL2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol      Margen de error para los totales marginales calibrados
% NES      Vector con los totales por celda (ponderador original)
% NCAL     Vector con los totales por celda calibrados
% TCAL     Vector con los totales por categoría calibrados
% L        Vector de multiplicadores de Lagrange

iter=0;
NES=zeros(r*c,1);
L=zeros(r+c,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);

%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
   for j=1:c
     for z=1:r2
        TEMPO(z)=X(z,i)*X(z,j+r);
     end
     tempo2=dot(TEMPO,WEI);
     NES(p)=tempo2;
     p=p+1;
   end
end

%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];
```

```
% Iteraciones para encontrar L
while  any(abs(TCAL-T)>tol) & iter<50

%Estimación de los totales calibrados (antes de la iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*(1-L2(p)/2)^(-2);
end
TCAL=(NCAL'*MATDIS)'

%Construcción de la matriz derivada de phi
PHIDE=zeros(r+c,r+c);

%Totales por celda ajustados por la derivada de la función
L2=MATDIS*L;
for p=1:r*c
 NCALDE(p)=NES(p)*(1-L2(p)/2)^(-3);
end

%Totales por categoría
TCALDE=(NCALDE'*MATDIS)';
for i=1:r+c
 PHIDE(i,i)=TCALDE(i);
end
for i=1:r
 for j=1:c
  for  p=1:r*c
   TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
  end
  tempo4=dot (NCALDE,TEMPO3);
  PHIDE (i,j+r)=tempo4;
  PHIDE (j+r,i)=tempo4;
 end
end

%Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
SPHIDE=zeros(r+c-1,r+c-1);
for i=1:r+c-1
 for j=1:r+c-1
   SPHIDE(i,j)=PHIDE(i,j);
 end
end

SPHIDEIN=inv(SPHIDE);

for i=1:r+c-1
 ST(i)=T(i);
 STCAL(i)=TCAL(i);
end
```

```
%Se obtiene el vector de multiplicadores de Lagrange
SL=SL+SPHIDEIN*(ST-STCAL);
for i=1:r+c-1
   L(i)=SL(i);
end

%Estimación de los totales para verificar criterio de convergencia (después de iteración)
L2=MATDIS*L;
for p=1:r*c
  NCAL(p)=NES(p)*(1-L2(p)/2)^(-2);
end
TCAL=(NCAL'*MATDIS)';
iter=iter+1;
end

% Despliegue de resultados
format long
L
TCAL
iter


%Estimación de los totales para la variable y
TEMPO5=X*L;
for i=1:r2
  WCAL(i)=WEI(i)*(1-TEMPO5(i)/2)^(-2);
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A6. Programa que calcula los g-factores utilizando el método de Newton Distancia entropía mínima: ENTR2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol     Margen de error para los totales marginales calibrados
% NES     Vector con los totales por celda (ponderador original)
% NCAL    Vector con los totales por celda calibrados
% TCAL    Vector con los totales por categoría calibrados
% L       Vector de multiplicadores de Lagrange

iter=0;
NES=zeros(r*c,1);
L=zeros(r+c,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);

%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
  for j=1:c
    for z=1:r2
      TEMPO(z)=X(z,i)*X(z,j+r);
    end
    tempo2=dot(TEMPO,WEI);
    NES(p)=tempo2;
    p=p+1;
  end
end

%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];
```

```
% Iteraciones para encontrar L
while  any(abs(TCAL-T)>tol) & iter<50

%Estimación de los totales calibrados (antes de la iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*(1-L2(p))^(-1);
end
TCAL=(NCAL'*MATDIS)'

%Construcción de la matriz derivada de phi
PHIDE=zeros(r+c,r+c);

%Totales por celda ajustados por la derivada de la función
L2=MATDIS*L;
for p=1:r*c
 NCALDE(p)=NES(p)*(1-L2(p))^(-2);
end

%Totales por categoría
TCALDE=(NCALDE'*MATDIS)';

for i=1:r+c
 PHIDE(i,i)=TCALDE(i);
end

for i=1:r
 for j=1:c
  for  p=1:r*c
   TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
  end
  tempo4=dot (NCALDE,TEMPO3);
  PHIDE (i,j+r)=tempo4;
  PHIDE (j+r,i)=tempo4;
 end
end


%Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
SPHIDE=zeros(r+c-1,r+c-1);
for i=1:r+c-1
 for j=1:r+c-1
   SPHIDE(i,j)=PHIDE(i,j);
 end
end

SPHIDEIN=inv(SPHIDE);
```

```
for i=1:r+c-1
 ST(i)=T(i);
 STCAL(i)=TCAL(i);
end

%Se obtiene el vector de multiplicadores de Lagrange
SL=SL+SPHIDEIN*(ST-STCAL);
for i=1:r+c-1
   L(i)=SL(i);
end

%Estimación de los totales para verificar criterio de convergencia (después de iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*(1-L2(p))^(-1);
end
TCAL=(NCAL'*MATDIS)';
iter=iter+1;
end

% Despliegue de resultados
format long
L
TCAL
iter


%Estimación de los totales para la variable y
TEMPO5=X*L;
for i=1:r2
 WCAL(i)=WEI(i)*(1-TEMPO5(i))^(-1);
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A7. Programa que calcula los g-factores utilizando el método de Newton Distancia mínimos cuadrados restringidos: MCRE2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol      Margen de error para los stotales marginales calibrados
% NCAL  Vector con los totales por celda calibrados
% TCAL   Vector con los totales por categoría calibrados
% L       Vector de multiplicadores de Lagrange


% Condiciones iniciales
linf=0.66;
lsup=3.6;
iter=0;
G=ones(r*c,1);
A=ones(r*c,1);
UNO=ones(r*c,1);
L=zeros(r+c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);

%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
   for j=1:c
    for z=1:r2
       TEMPO(z)=X(z,i)*X(z,j+r);
    end
    tempo2=dot(TEMPO,WEI);
    NES(p)=tempo2;
    p=p+1;
   end
end
```

```
%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
       0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];


% Iteraciones para encontrar L
while  any(abs(TCAL-T)>tol) & iter<150

%Construcción de la matriz derivada de phi
PHIDE=zeros(r+c,r+c);

%Totales por celda ajustados por el factor A(i)
L2=MATDIS*L;
for p=1:r*c
 NCALDE(p)=NES(p)*A(p);
end

%Totales por categoría
TCALDE=(NCALDE'*MATDIS)';

for i=1:r+c
 PHIDE(i,i)=TCALDE(i);
end

for i=1:r
 for j=1:c
  for  p=1:r*c
   TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
  end
  tempo4=dot (NCALDE,TEMPO3);
  PHIDE (i,j+r)=tempo4;
  PHIDE (j+r,i)=tempo4;
 end
end

for  p=1:r*c
   NCAL(p)=NES(p)*G(p);
 end
 TCAL= (NCAL'*MATDIS)';
```

```
%Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
SPHIDE=zeros(r+c-1,r+c-1);
for i=1:r+c-1
 for j=1:r+c-1
   SPHIDE(i,j)=PHIDE(i,j);
 end
end

SPHIDEIN=inv(SPHIDE);

for i=1:r+c-1
 ST(i)=T(i);
 STCAL(i)=TCAL(i);
end

SL=L(1:r+c-1,:);
SL=SL+SPHIDEIN*(ST-STCAL);

 for i=1:r+c-1
  L(i)=SL(i);
 end

G=UNO+ MATDIS*L;

A=ones(r*c,1);
 for i=1:r*c
  if G(i)<linf
   G(i)=linf;
   A(i)=0;
  end
  if G(i)>lsup
   G(i)=lsup;
   A(i)=0;
  end
 end

iter=iter+1;

end

% Despliega resultados
L
format long
TCAL
iter
```

```
%Estimación de los totales para la variable y
Gi=zeros(r2,1);
for i=1:r2
 Gi(i)=1+X(i,:)*L;
end

Ai=ones(r2,1);
 for i=1:r2
  if Gi(i)<linf
   Gi(i)=linf;
   Ai(i)=0;
  end
  if Gi(i)>lsup
   Gi(i)=lsup;
   Ai(i)=0;
  end
 end


for i=1:r2
 WCAL(i)=WEI(i)*Gi(i);
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A8. Programa que calcula los g-factores utilizando el método de Newton Distancia logit: LOGI2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol      Margen de error para los totales marginales calibrados
% NES      Vector con los totales por celda (ponderador original)
% NCAL     Vector con los totales por celda calibrados
% TCAL     Vector con los totales por categoría calibrados
% L        Vector de multiplicadores de Lagrange

iter=0;
NES=zeros(r*c,1);
L=zeros(r+c,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);
inf=0.55;
sup=5.9;
a=(sup-inf)/((1-inf)*(sup-1));

%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
   for j=1:c
    for z=1:r2
       TEMPO(z)=X(z,i)*X(z,j+r);
    end
    tempo2=dot(TEMPO,WEI);
    NES(p)=tempo2;
    p=p+1;
   end
end

%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];
```

```
% Iteraciones para encontrar L
while  any(abs(TCAL-T)>tol) & iter<50

%Estimación de los totales calibrados (antes de la iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*((sup-1)*inf+(1-inf)*sup*exp(a*L2(p)))/ ((sup-1)+(1-inf)*exp(a*L2(p)));
end
TCAL=(NCAL'*MATDIS)'

%Construcción de la matriz derivada de phi
PHIDE=zeros(r+c,r+c);

%Totales por celda ajustados por la derivada de la función
L2=MATDIS*L;
for p=1:r*c
 NCALDE(p)=NES(p)*(sup-inf)^2*exp(a*L2(p))/((sup-1)+(1-inf)*exp(a*L2(p)))^2;
end

%Totales por categoría
TCALDE=(NCALDE'*MATDIS)';

for i=1:r+c
 PHIDE(i,i)=TCALDE(i);
end

for i=1:r
 for j=1:c
  for  p=1:r*c
   TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
  end
  tempo4=dot (NCALDE,TEMPO3);
  PHIDE (i,j+r)=tempo4;
  PHIDE (j+r,i)=tempo4;
 end
end


%Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
SPHIDE=zeros(r+c-1,r+c-1);
for i=1:r+c-1
 for j=1:r+c-1
   SPHIDE(i,j)=PHIDE(i,j);
 end
end

SPHIDEIN=inv(SPHIDE);
```

```
for i=1:r+c-1
 ST(i)=T(i);
 STCAL(i)=TCAL(i);
end

%Se obtiene el vector de multiplicadores de Lagrange
SL=SL+SPHIDEIN*(ST-STCAL);
for i=1:r+c-1
   L(i)=SL(i);
end

%Estimación de los totales para verificar criterio de convergencia (después de iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)* ((sup-1)*inf+(1-inf)*sup*exp(a*L2(p)))/ ((sup-1)+(1-inf)*exp(a*L2(p)));
end
TCAL=(NCAL'*MATDIS)';
iter=iter+1;
end

% Despliegue de resultados
format long
L
TCAL
iter

%Estimación de los totales para la variable y
TEMPO5=X*L;

for i=1:r2
 WCAL(i)=WEI(i)*((sup-1)*inf+(1-inf)*sup*exp(a*TEMPO5(i)))/ ((sup-1)+(1-
inf)*exp(a*TEMPO5(i)));
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A9. Programa que calcula los g-factores utilizando el método de Newton Distancia mínimos cuadrados modificada: MODI2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol      Margen de error para los totales marginales calibrados
% NES      Vector con los totales por celda (ponderador original)
% NCAL     Vector con los totales por celda calibrados
% TCAL     Vector con los totales por categoría calibrados
% L        Vector de multiplicadores de Lagrange

iter=0;
NES=zeros(r*c,1);
L=zeros(r+c,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);

%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
  for j=1:c
    for z=1:r2
      TEMPO(z)=X(z,i)*X(z,j+r);
    end
    tempo2=dot(TEMPO,WEI);
    NES(p)=tempo2;
    p=p+1;
  end
end

%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];
```

```
% Iteraciones para encontrar L
while  any(abs(TCAL-T)>tol) & iter<50

%Estimación de los totales calibrados (antes de la iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*(1-2*L2(p))^(-1/2);
end
TCAL=(NCAL'*MATDIS)'

%Construcción de la matriz derivada de phi
PHIDE=zeros(r+c,r+c);

%Totales por celda ajustados por la derivada de la función
L2=MATDIS*L;
for p=1:r*c
 NCALDE(p)=NES(p)*(1-2*L2(p))^(-3/2);
end

%Totales por categoría
TCALDE=(NCALDE'*MATDIS)';

for i=1:r+c
 PHIDE(i,i)=TCALDE(i);
end

for i=1:r
 for j=1:c
  for  p=1:r*c
   TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
  end
  tempo4=dot (NCALDE,TEMPO3);
  PHIDE (i,j+r)=tempo4;
  PHIDE (j+r,i)=tempo4;
 end
end


%Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
SPHIDE=zeros(r+c-1,r+c-1);
for i=1:r+c-1
 for j=1:r+c-1
   SPHIDE(i,j)=PHIDE(i,j);
 end
end

SPHIDEIN=inv(SPHIDE);

for i=1:r+c-1
```

```
 ST(i)=T(i);
 STCAL(i)=TCAL(i);
end

%Se obtiene el vector de multiplicadores de Lagrange
SL=SL+SPHIDEIN*(ST-STCAL);
for i=1:r+c-1
   L(i)=SL(i);
end

%Estimación de los totales para verificar criterio de convergencia (después de iteración)
L2=MATDIS*L;
for p=1:r*c
 NCAL(p)=NES(p)*(1-2*L2(p))^(-1/2);
end
TCAL=(NCAL'*MATDIS)';

iter=iter+1;

end

% Despliegue de resultados
format long
L
TCAL
iter


%Estimación de los totales para la variable y
TEMPO5=X*L;

for i=1:r2
 WCAL(i)=WEI(i)*(1-2*TEMPO5(i))^(-1/2);
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A10. Programa que calcula los g-factores utilizando el método de Newton Distancia Huang - Fuller: HUAN2.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol     Margen de error para los factores de ajuste
% NES     Vector con los totales por celda (ponderador original)
% TES     Vector con los totales por categoría
% NCAL    Vector con los totales por celda calibrados
% TCAL    Vector con los totales por categoría calibrados
% L       Vector de multiplicadores de Lagrange

alfa=2/3;
beta=0.8;
linf1=0.662;
lsup1=3.598;
linf2=alfa*linf1+1-alfa;
lsup2=alfa*lsup1+1-alfa;
tol=.002;
iter=0;
G=zeros(r*c,1);
A=ones(r*c,1);
EPSI=zeros(r*c,1);
L=zeros(r+c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
TES=zeros(r+c,1);
STES=zeros(r+c-1,1);
TCAL=zeros(r+c,1);
NES=zeros(r*c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
STCAL=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);


%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
   for j=1:c
    for z=1:r2
      TEMPO(z)=X(z,i)*X(z,j+r);
```

```
    end
    tempo2=dot(TEMPO,WEI);
    NES(p)=tempo2;
    p=p+1;
   end
end

%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];

% Iteraciones para encontrar los g factores
while any(G>lsup1+tol | G<linf1-tol) & iter<100

 %Construcción de la matriz derivada de phi
 PHIDE=zeros(r+c,r+c);

 %Totales por celda ajustados por el factor A(i)
 L2=MATDIS*L;
 for p=1:r*c
  NCALDE(p)=NES(p)*A(p);
 end

 %Totales por categoría
 TCALDE=(NCALDE'*MATDIS)';
 for i=1:r+c
  PHIDE(i,i)=TCALDE(i);
 end

 for i=1:r
  for j=1:c
   for p=1:r*c
    TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
   end
   tempo4=dot (NCALDE,TEMPO3);
   PHIDE (i,j+r)=tempo4;
   PHIDE (j+r,i)=tempo4;
  end
 end

 TES= (NES'*MATDIS)';
```

```
 %Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
 SPHIDE=zeros(r+c-1,r+c-1);
 for i=1:r+c-1
  for j=1:r+c-1
   SPHIDE(i,j)=PHIDE(i,j);
  end
 end

 SPHIDEIN=inv(SPHIDE);

 for i=1:r+c-1
  ST(i)=T(i);
  STES(i)=TES(i);
 end

 SL=L(1:r+c-1,:);
 SL=SPHIDEIN*(ST-STES);

 for i=1:r+c-1
  L(i)=SL(i);
 end

 for p=1:r*c
  G(p)=1+A(p)*MATDIS(p,:)*L;
 end

 for p=1:r*c
  if G(p)<=1
   EPSI(p)=(G(p)-1)/(linf2-1);
  else
   EPSI(p)=(G(p)-1)/(lsup2-1);
  end
 end

 for p=1:r*c
  if EPSI(p)<.5
   A(p)=1*A(p);
  end
  if EPSI(p)>=0.5 & EPSI(p)<1
   A(p)= (1-beta*(EPSI(p)-0.5)^2)*A(p);
  end
  if EPSI(p)>=1
   A(p)=((1-beta/4)/EPSI(p))*A(p);
  end
 end
 iter=iter+1;
end
```

```
% Despliega resultados
L
format long
TCAL
iter

%Estimación de los totales para la variable y
Gi=zeros(r2,1);
for i=1:r2
 for p=1:r*c
  if X(i,:)==MATDIS(p,:)
   Gi(i)=G(p);
  end
 end
end

for i=1:r2
 WCAL(i)=WEI(i)*Gi(i);
end

ycal1=dot(WCAL, TEL1);
ycal2=dot(WCAL, TEL2);
```

A11. Programa que calcula los g-factores utilizando el método de Newton Distancia contracción-minimización: CONT.M

```
%Lectura de datos
Vivi92va

% Condiciones iniciales
% tol     Margen de error para los factores de ajuste
% NES     Vector con los totales por celda (ponderador original)
% TES     Vector con los totales por categoría
% NCAL    Vector con los totales por celda calibrados
% TCAL    Vector con los totales por categoría calibrados
% L       Vector de multiplicadores de Lagrange

alfa=2/3;
eta=0.9;
linf=0.662;
lsup=3.598;
linf1=alfa*linf+1-alfa;
lsup1=alfa*lsup+1-alfa;
linf2=eta*linf+1-eta;
lsup2=eta*lsup+1-eta;
iter=0;
G=zeros(r2,1);
L=zeros(r+c,1);
SL=zeros(r+c-1,1);
ST=zeros(r+c-1,1);
TCAL=zeros(r+c,1);
NCAL=zeros(r*c,1);
NCALDE=zeros(r*c,1);
STCALDE=zeros(r+c-1,1);
TEL1=MAT(:,c2-1);
TEL2=MAT(:,c2);
WEIO=WEI;
tol=.002;

%Matriz diseño, es decir, de las variables y sus categorías
MATDIS=[1, 0, 0, 0, 1, 0
        1, 0, 0, 0, 0, 1
        0, 1, 0, 0, 1, 0
        0, 1, 0, 0, 0, 1
        0, 0, 1, 0, 1, 0
        0, 0, 1, 0, 0, 1
        0, 0, 0, 1, 1, 0
        0, 0, 0, 1, 0, 1];
```

```
% Iteraciones para encontrar los g factores
while any(G>lsup+tol | G<linf-tol) & iter<51

WEI3=WEI;
 for i=1:r2
  if WEI(i)<linf2*WEIO(i)
   WEI3(i)=linf1*WEIO(i);
  end
  if WEI(i)>lsup2*WEIO(i)
   WEI3(i)=lsup1*WEIO(i);
  end
 end

%Se calculan los totales por cada celda y se almacenan en un vector (11,12,...,rc)
p=1;
for i=1:r
  for j=1:c
   for z=1:r2
     TEMPO(z)=X(z,i)*X(z,j+r);
   end
   tempo2=dot(TEMPO,WEI3);
   NCALDE(p)=tempo2;
   p=p+1;
  end
end

 %Construcción de la matriz derivada de phi
 PHIDE=zeros(r+c,r+c);

 %Totales por categoría
 TCALDE=(NCALDE'*MATDIS)';

 for i=1:r+c
  PHIDE(i,i)=TCALDE(i);
 end

 for i=1:r
  for j=1:c
   for p=1:r*c
    TEMPO3(p)=MATDIS(p,i)*MATDIS(p,j+r);
   end
   tempo4=dot (NCALDE,TEMPO3);
   PHIDE (i,j+r)=tempo4;
   PHIDE (j+r,i)=tempo4;
  end
 end
```

```
 %Se eliminan una columna y un renglón para que la matriz tenga inversa (información
redundante)
 SPHIDE=zeros(r+c-1,r+c-1);
 for i=1:r+c-1
  for j=1:r+c-1
   SPHIDE(i,j)=PHIDE(i,j);
  end
 end

 SPHIDEIN=inv(SPHIDE);

for i=1:r+c-1
  ST(i)=T(i);
  STCALDE(i)=TCALDE(i);
 end

 SL=L(1:r+c-1,:);
 SL=SPHIDEIN*(ST-STCALDE);

 for i=1:r+c-1
  L(i)=SL(i);
 end

 for i=1:r2
  WEI(i)=WEI3(i)*(1+ X(i,:)*L);
 end

 for i=1:r2
  if WEIO(i)>0
   G(i)=WEI(i)/WEIO(i);
  else
   G(i)=1;
  end
 end

iter=iter+1;

end

% Despliega resultados
L
format long
iter

%Estimación de los totales para la variable y
ycal1=dot(WEI, TEL1);
ycal2=dot(WEI, TEL2);
```

ANEXO B
PRINCIPALES RUBROS CAPTADOS POR LA ENIGH QUE CONFORMAN EL INGRESO - GASTO

Conceptos de Ingreso - Gasto

INGRESO

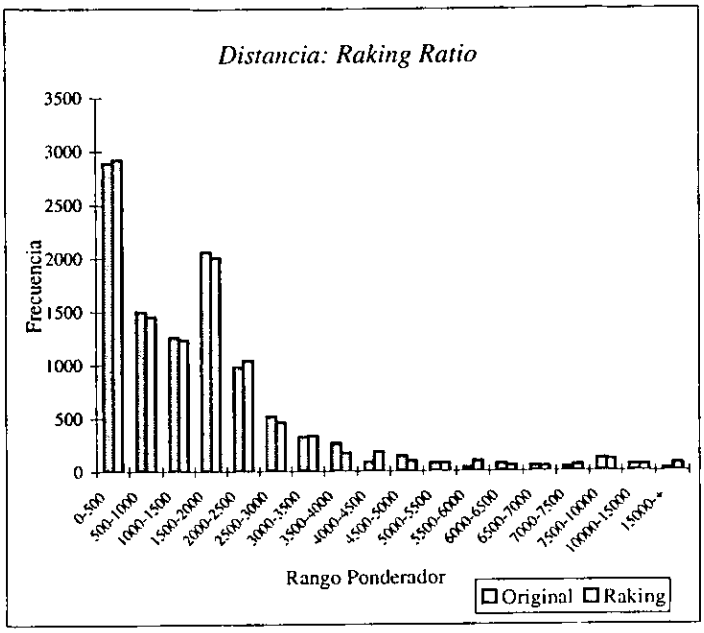
Monetario	**No monetario**	**Percepciones financieras**
• Remuneraciones al trabajo • Negocios propios • Renta de la propiedad (alquiler de casas, intereses provenientes de inversiones, cuentas de ahorros, etc.) • Cooperativas • Transferencias (pensiones, indemnizaciones, becas, regalos en dinero, donativos, ingresos de otros países, etc.) • Otros ingresos (venta de automóviles, aparatos eléctricos de segunda mano, viáticos, etc.)	A excepción de las percepciones financieras no monetarias se genera por: • Autoconsumo • Pago en especie • Regalos • Valor estimado del alquiler de la vivienda	*Monetarias:* • Retiro de inversiones, ahorros, tandas, etc. • Ingresos por préstamos a terceros • Préstamos de personas no miembros del hogar o instituciones • Venta de casas, monedas, maquinaria, etc. • Préstamos hipotecarios • Seguros de vida *No monetarias* (autoconsumo, pago en especie, regalos): • Materiales, reparación y ampliación de la vivienda • Depósitos en cuentas de ahorros • Préstamos a terceros • Pagos a tarjeta de crédito • Compra de monedas, alhajas, etc. • Seguro de vida • Herencias • Compra de casas, condominios, locales, terrenos, etc. • Pago de hipotecas • Compra de maquinaria, equipo, etc.

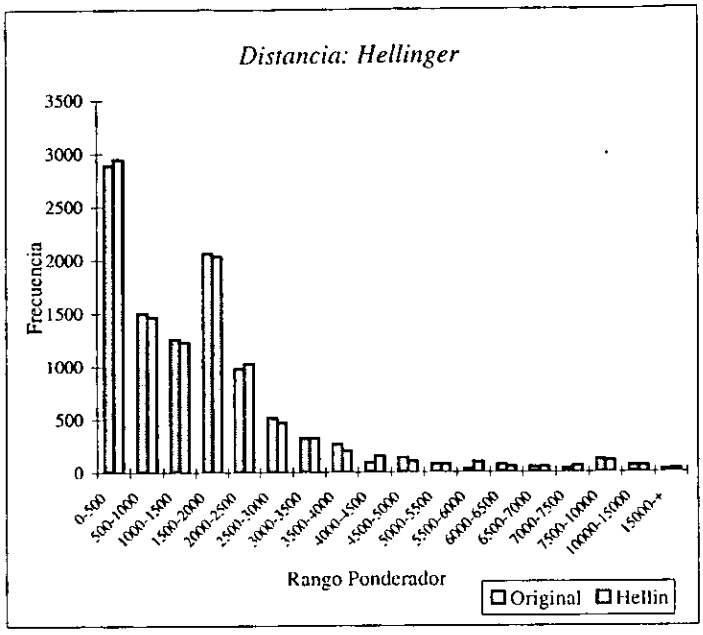
GASTO

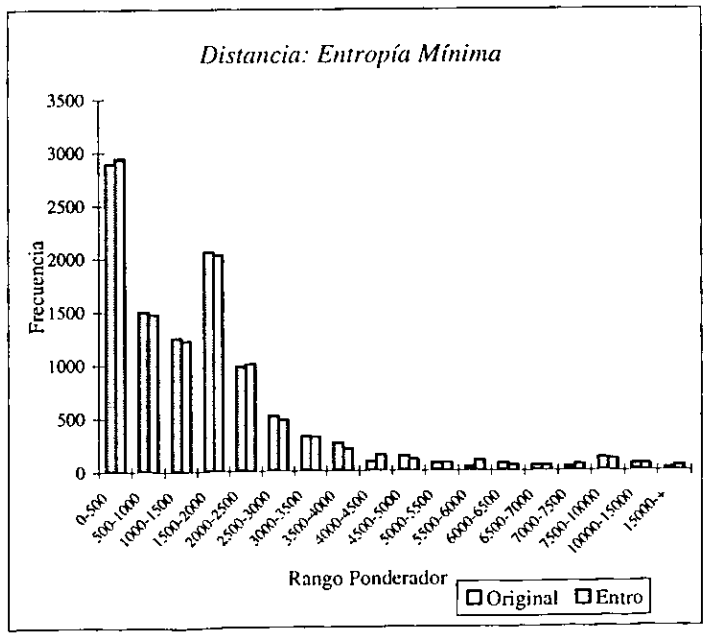
Monetario	**No monetario**	**Erogaciones financieras**
• Alimentos, bebidas y tabaco • Ropa, calzado y accesorios • Vivienda (renta, impuestos, servicios, etc.) • Mobiliario, enseres domésticos y mantenimiento • Cuidados de la salud • Transportes • Comunicaciones • Educación, esparcimiento y cultura • Otros bienes y servicios	A excepción de las erogaciones financieras, está conformado por: • Autoconsumo • Pago en especie • Regalos recibidos • Renta estimada de la vivienda propia o prestada	Monetarias: • Materiales, reparación y ampliación de la vivienda • Depósitos en cuentas de ahorros, • Préstamos a terceros • Pagos a tarjeta de crédito • Compra de monedas, alhajas, etc. • Seguro de vida • Herencias • Compra de casas, condominios, locales, terrenos, etc. • Pago de hipotecas • Compra de maquinaria, equipo, etc. No monetarias (autoconsumo, pago en especie, regalos): • Mismos que en monetarias

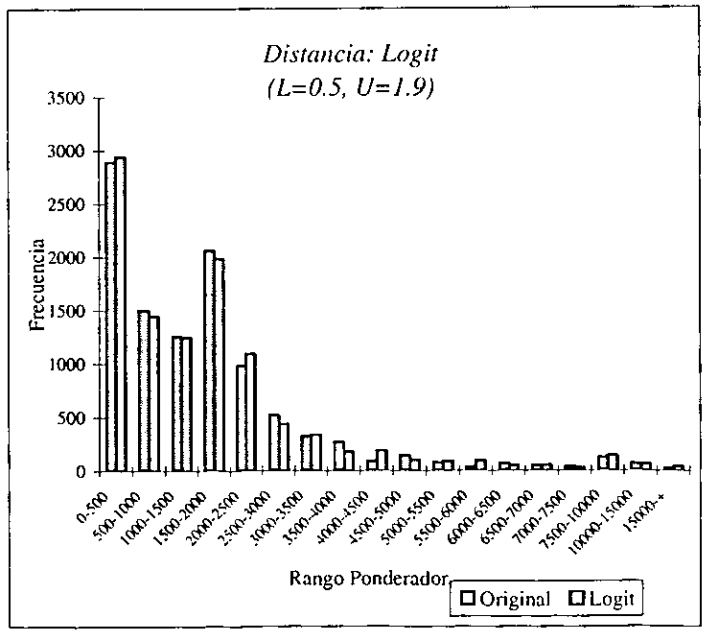
ANEXO C
HISTOGRAMAS DE LOS PONDERADORES

C1. Disponibilidad de drenaje y energía eléctrica, 1992

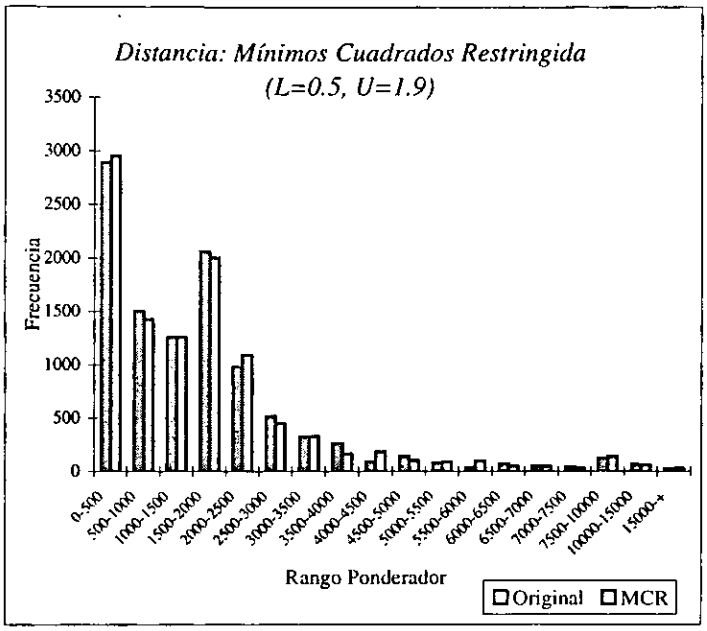


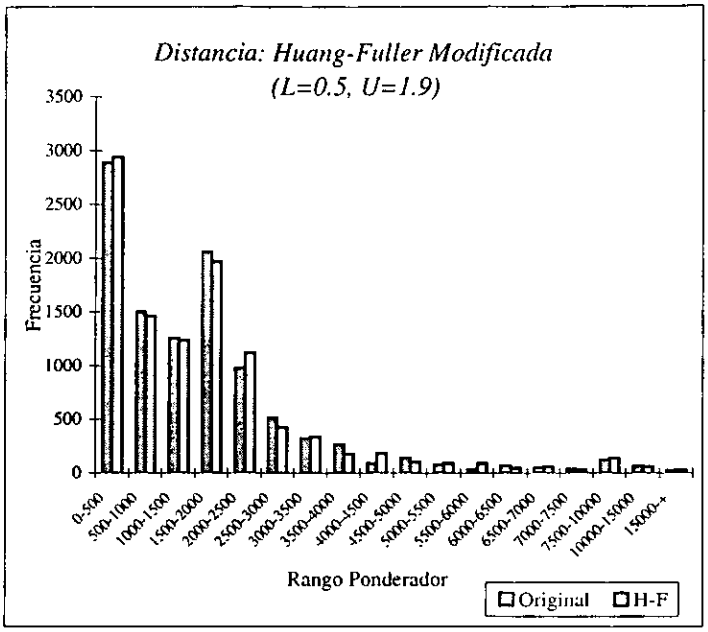


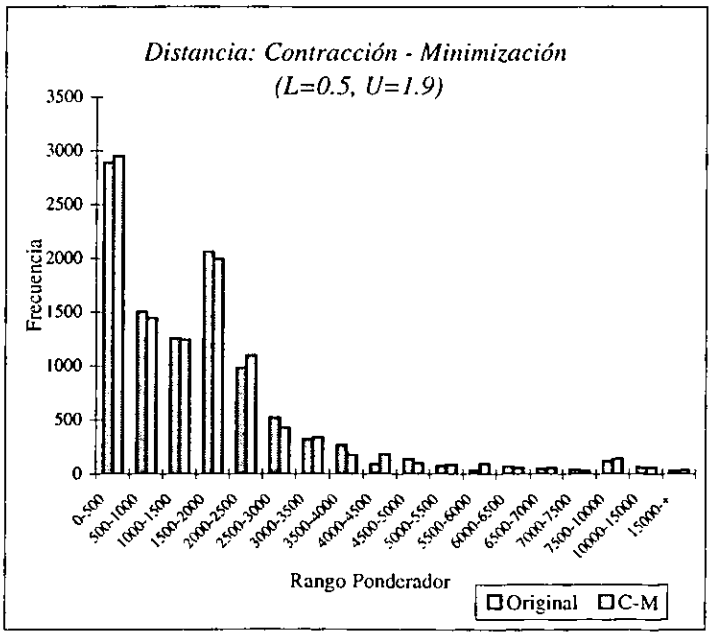




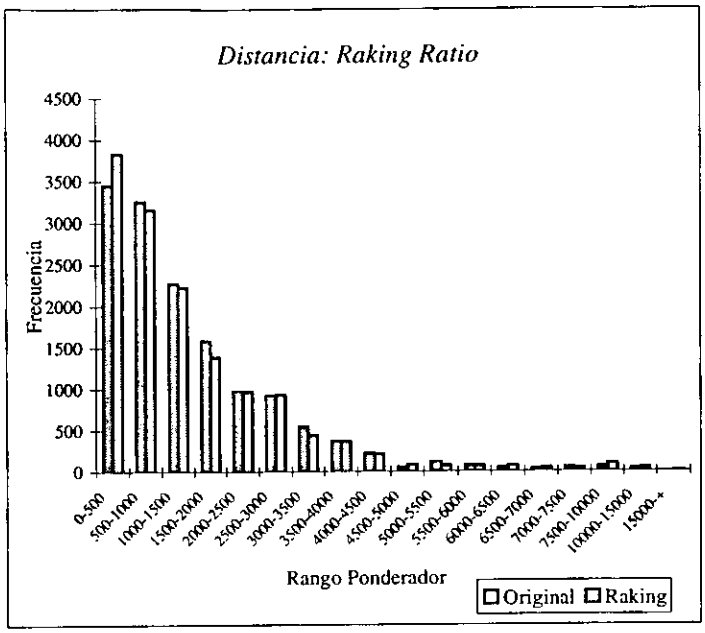
Disponibilidad de drenaje y energía eléctrica, 1992 (continuación)

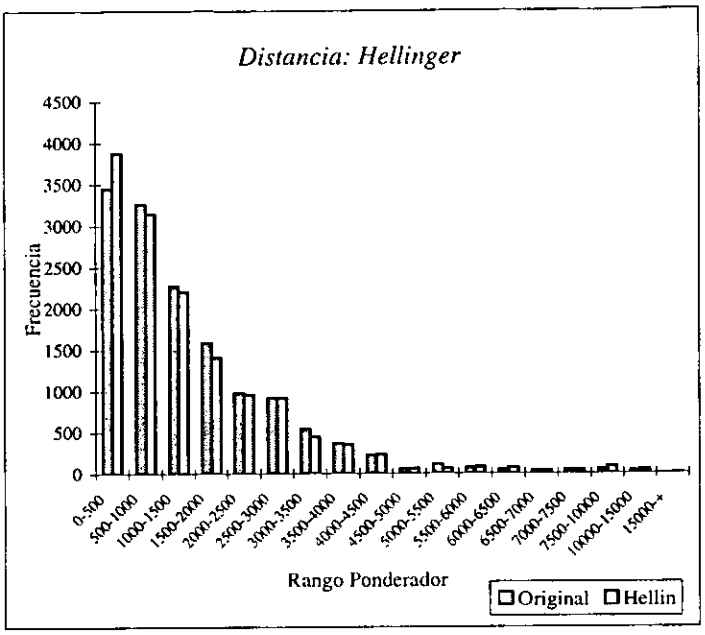


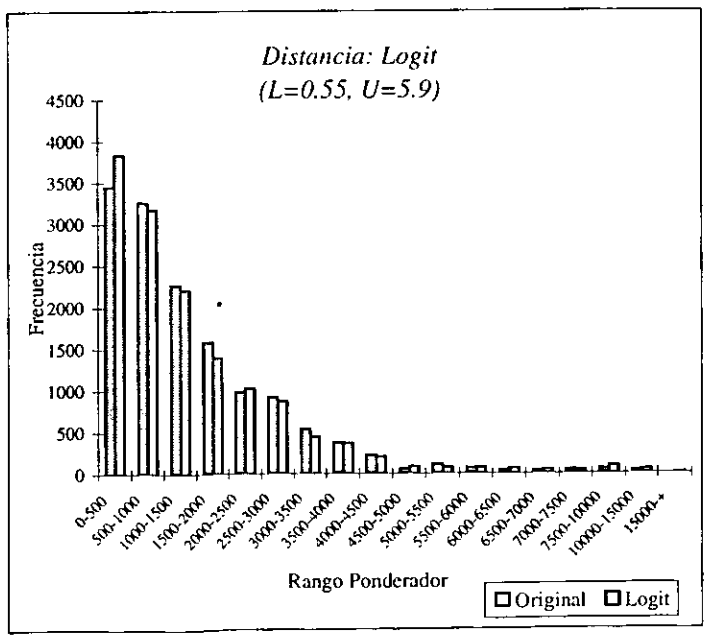


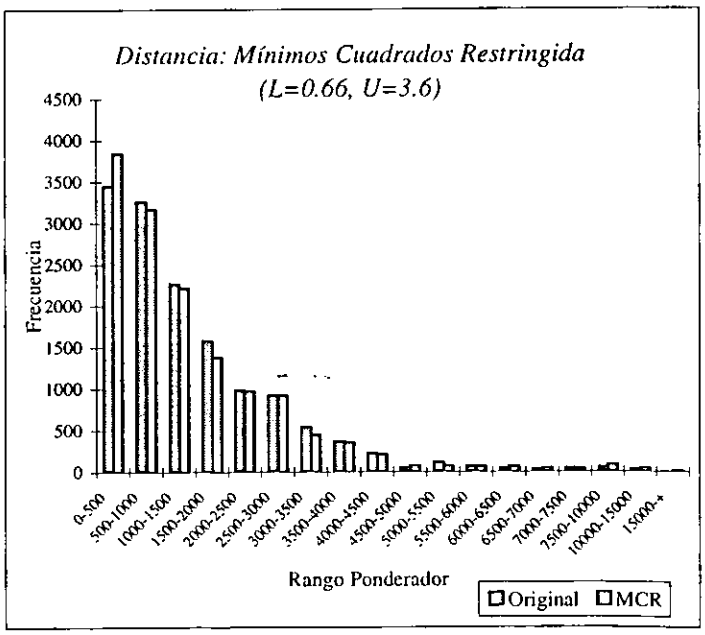


C2. Disponibilidad de drenaje y energía eléctrica, 1996

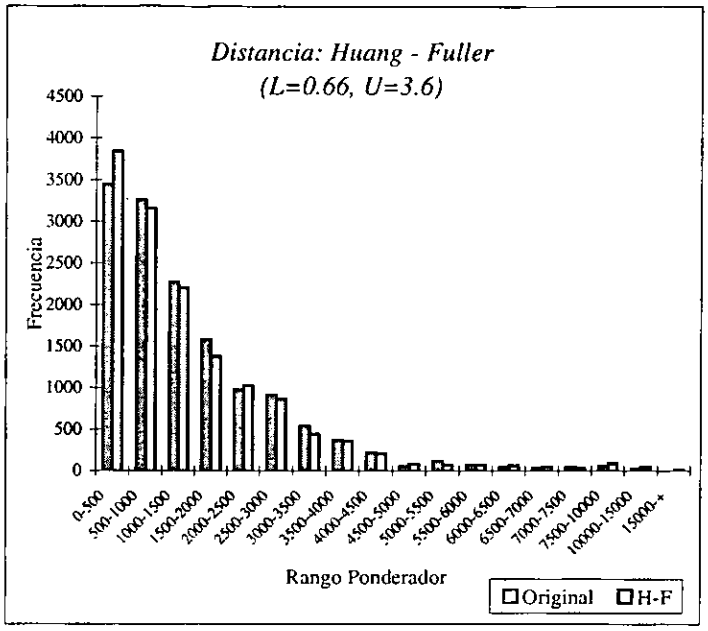


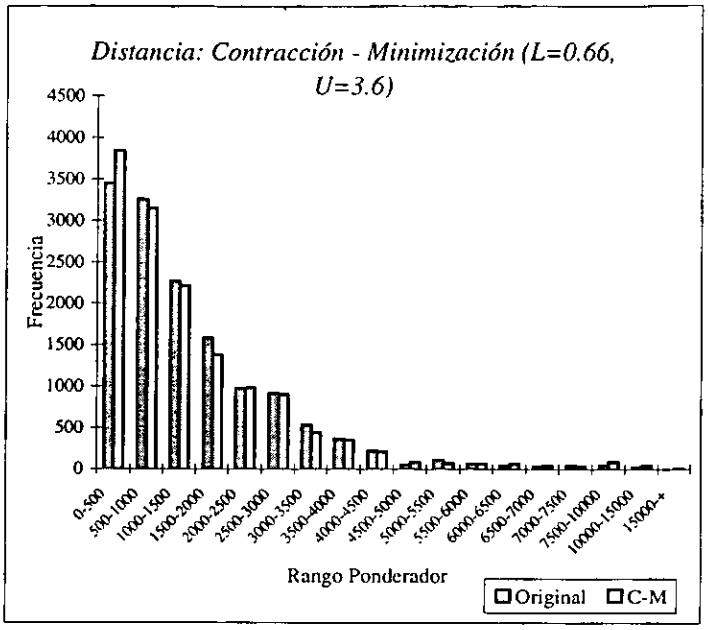




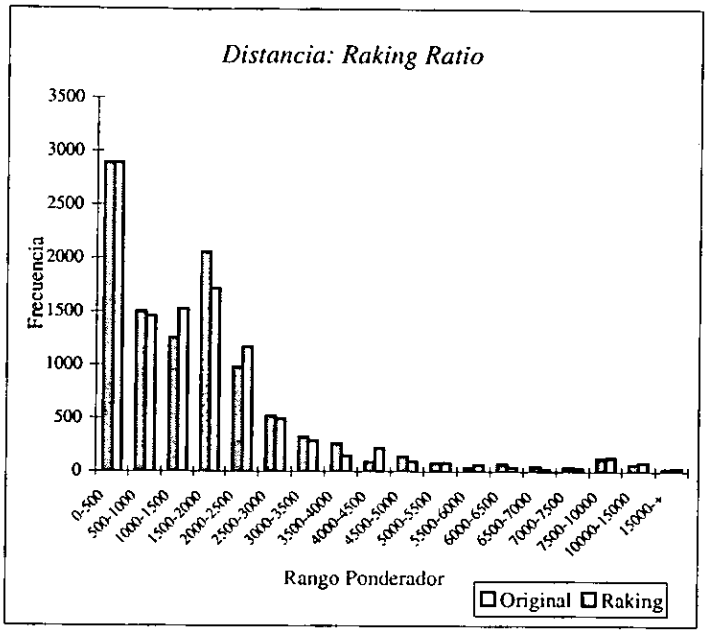


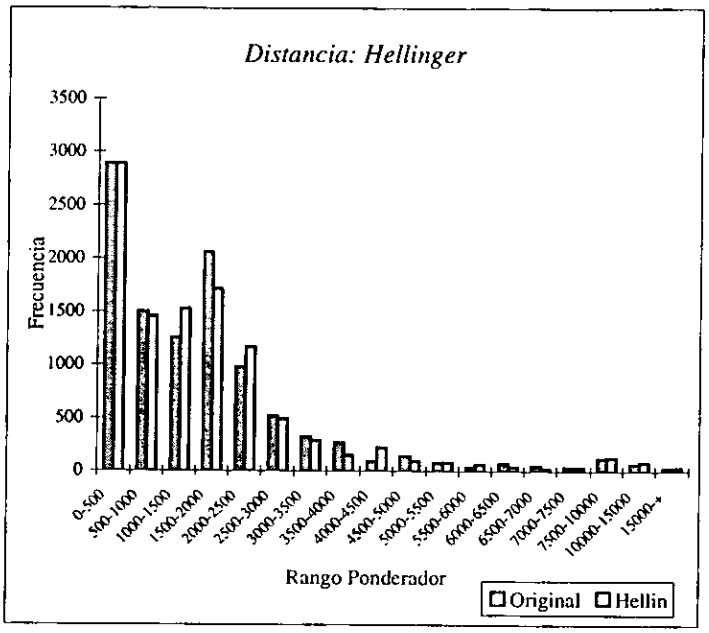
Disponibilidad de drenaje y energía eléctrica, 1996 (continuación)

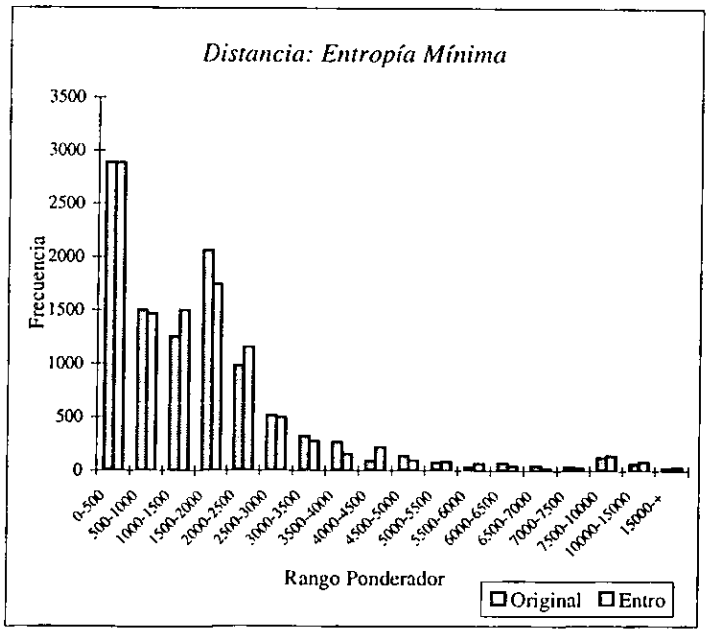


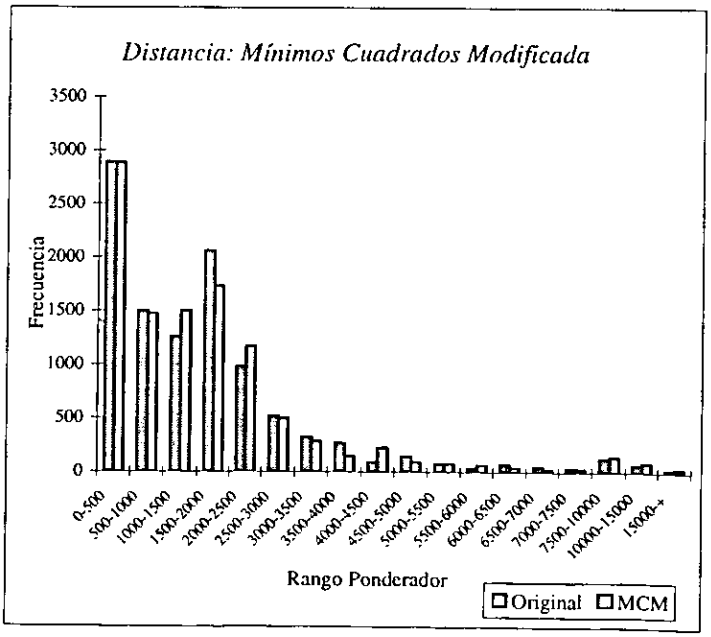


C3. Disponibilidad de agua y material en pisos, 1992

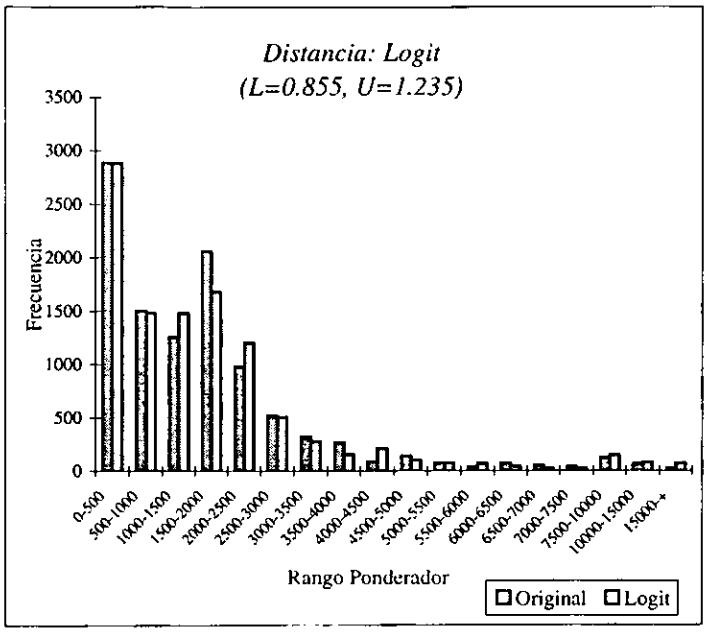


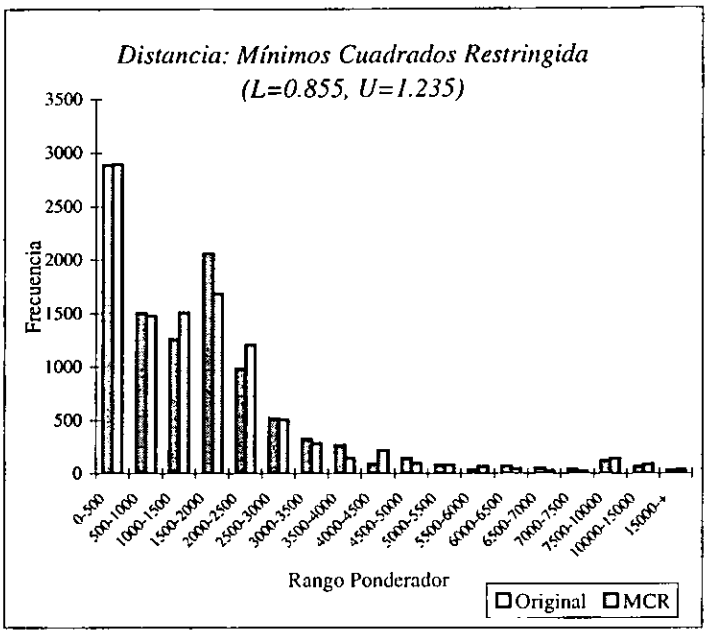


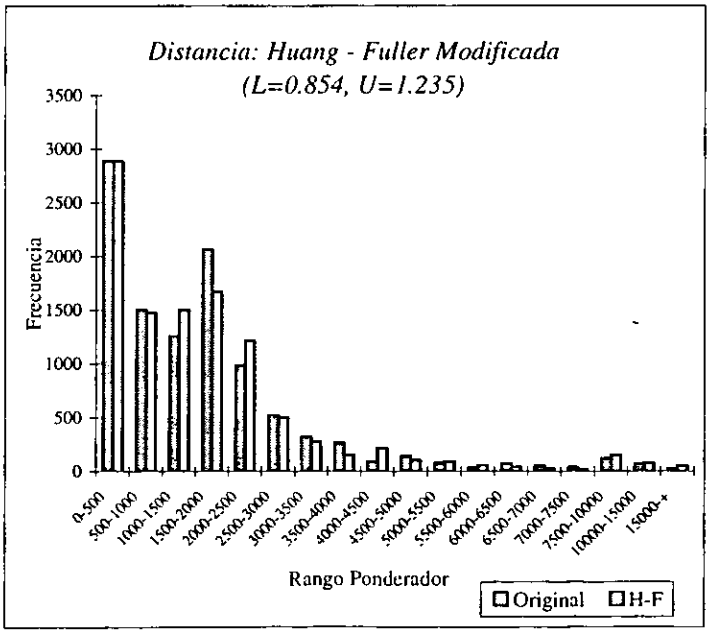




Disponibilidad de agua y material en pisos, 1992 (continuación)







BIBLIOGRAFIA

Cochran W. G., (1977) *"Sampling Techniques"*, 3th ed., John Wiley & Sons.

Deming W. E., Stephan F. F. (1940), *"On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known"*, The Annals of Mathematical Statistics, 11, 427-444.

Deville J. C. (1988), *"Estimation Linéaire et Redressement sur Informations Auxiliaires d'Enquêtes par Sondage"*, Essais en l'Honneur d'Edmond Malinvaud, Paris: Economica, pp. 915-927.

Deville J. C., Särndal C. E., Sautory O. (1993), *"Generalized Raking Procedures in Survey Sampling"*, Journal of the American Statistical Association, 88, 1013-1020.

Deville J. C., Särndal C. E. (1992), *"Calibration Estimators in Survey Sampling"*, Journal of the American Statistical Association, 87, 376-382.

Estevao V., Hidiroglou M. A. y Särndal C. E. (1995) *"Methodological Principles for a Generalized Estimation System at Statistics Canada"*. Journal of Official Statistics, 11, 181-204.

Huang E. T. y Fuller W. A. (1978), *"Nonnegative Regression Estimation for Sample Survey Data"*, Proceedings of the Social Statistics Section, American Statistical Association, 300-305.

Kish, L. (1989), *"Deffs: why, when and how? a review"*. ASA Proceedings of Survey Research Methods Section. 209-211 p.

INEGI, *"Conteo de Población y Vivienda 1995. Resultados Definitivos. Tabulados Básicos"*, 1997. 2ª. Reimpresión, INEGI.

INEGI, *"XI Censo General de Población y Vivienda, 1990. Tabulados Básicos"*, 1992.

INEGI, *"Encuesta Nacional de Ingresos y Gastos de los Hogares 1992"*, noviembre 1993.

INEGI, *"Encuesta Nacional de Ingresos y Gastos de los Hogares 1996"*, septiembre 1998.

Lemel, Y. (1976), *"Une Généralisation de la Méthode du Quotient por le Redressement des Enquêtes par Sondage"*, Annales de l'INSEE, 22-23, 272-282.

Lundström S., Särndal C. E. (1999), *"Calibration as a Standard Method por Treatment of Nonresponse"*, Journal of Official Statistics, 15, 305-327.

MATLAB for Windows, versión 4.2c.1 (1994). The Mathworks, Inc.

Palmer Arrache Catalina *"El cálculo de Varianza en Muestreos Complejos. Aplicación en la Encuesta de Nutrición 1996.". Tesis de Maestría*, IIMAS-UNAM, 1999.

Rao J. N. K. (1997), *"Developments in Sample Survey Theory: an Appraisal"*, The Canadian Journal of Statistics, 25, 1-21.

Särndal C. E., Swensson B. y Wretman J. H. (1989), *"The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total"*. Biometrika, 76, 3, 527-537.

Shah B. V. (1978), *"Variance Estimates for Complex Statistics from Multistage Sample Surveys"*. Survey Sampling and Measurement, Academic Press, pp. 25-33.

Shao J., Tu D. (1995), *"The Jackknife and Bootstrap"*, Springer Series in Statistics. Springer-Verlag.

Singh A. C., Mohl C. A. (1996), "*Understanding Calibration Estimators in Survey Sampling*", Survey Methodology, 22, 107-115.

Stukel D. M., Hidiroglou M. A., Särndal C. E. (1996), "*Variance Estimation for Calibration Estimators: a Comparison of Jackknifing versus Taylor Linearization*", Survey Methodology, 22, 117-125.

Wolter K. M. (1985), *"Introduction to Variance Estimation"*, Springer-Verlag, 427 pp.