Anexo 4. Diseño Muestral. Efecto de Origen de Clase en Argentina (1955-2001).

Quartulli, Diego.

Anexo 4
Diseño Muestral

"Dados n casos al azar sobre alguna variable" son típicas
palabras iniciales para las fábulas estadísticas
y cada una de ellas es engañosa para los datos reales
Muchos datos no son "dados" –ellos tienen que tomarse, atraídos, capturados o extraídos.
El "n" generalmente no es fijo, pero varía por las variadas imperfecciones en la colección.
La selección no es simplemente "al azar" sino agrupada y estratificada o de otro modo complejo.
(Kish, 2004, p. vi)

La posibilidad de inferir dichas declaraciones depende
de la calidad de los datos, los cuales proveen las premisas de la inferencia
(Fisher, 1965, p. 35)

A4.1 Introducción[1]

Para que se puedan realizar inferencias estadísticas, esto es, aplicar la estadística inferencial, cuando se analizan datos estos deben haberse producido siguiendo un diseño muestral que contemple la aleatoriedad en todos sus pasos y una serie de operaciones empíricas en campo y en gabinete que intenten cumplir con aquel.[2]

En este sentido, los datos que se usaron en este trabajo se han aproximado a muchos de los requisitos anteriores, ofreciendo una muestra relativamente extensa (5706 hogares) con una amplia cobertura no sólo de la población urbana sino también de muchos aglomerados urbanos de la Argentina.

Como se destacó en la introducción, los datos de esta investigación se basan en una muestra realizada por el Observatorio de la Deuda Social (ODSA) en su edición 2010 de la Encuesta de la Deuda Social de la Argentina (EDSA).

En este marco, en el presente anexo se expone una serie de aspectos metodológicos vinculados con los dominios de estudio (A4.2), la estrategia, el plan

[1] Este anexo es una actualización y corrección de tres escritos anteriores (Quartulli et al., 2011)(Quartulli, 2012)(Adaszko, 2013).

[2] Esta proposición tiene su excepción en los casos límites que se suponga que la población misma a muestrar posee una aleatoriedad intrínseca en todos sus niveles. Cuanto más se suponga que aquella se aleja de estos supuestos, más importante se vuelven el diseño y las acciones del investigador para lograr la aleatoriedad en la muestra.

de muestreo y la selección de casos (A4.3), la elaboración de ponderadores (A4.4) y la estimación de los errores muestrales (AA.5).

A4.2 Dominios del Estudio

El objetivo del diseño muestral original fue poder contar con una muestra representativa, dentro de cierto intervalo de confianza y con determinados márgenes de error muestral, de 9 sub-dominios, que luego, agregándose de una manera específica, se puede expandir a dominios empíricos intermedios hasta llegar al total urbano de la Argentina.

En una etapa inicial, se tomaron en cuenta criterios geo-demográficos del siguiente modo. En primer lugar se seleccionaron el conjunto de aglomerados a incluir (por región y tamaño).[3]

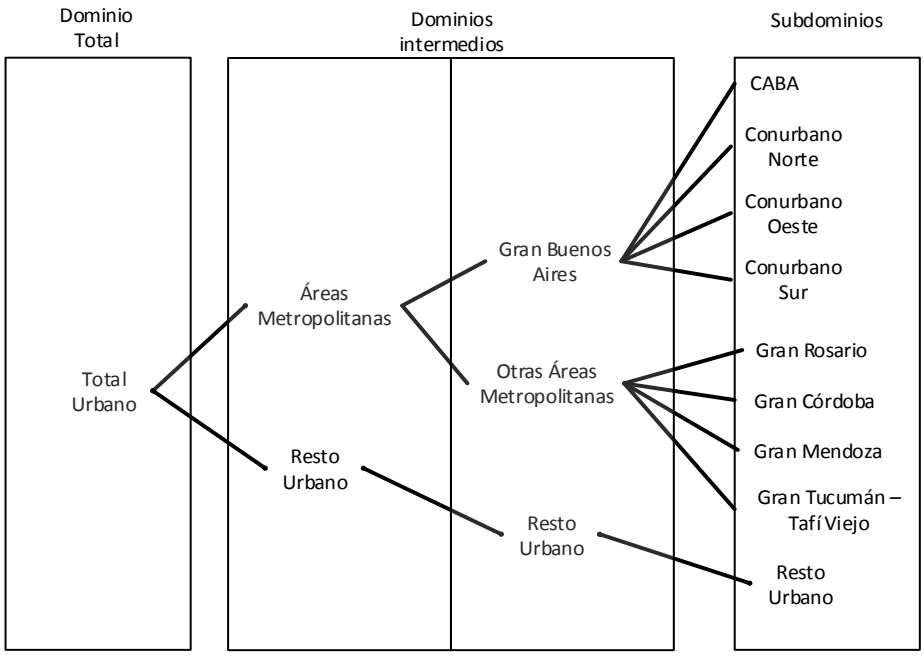
Los 9 sub-dominios de estudio para los que son representativos los resultados de la encuesta son: 1) CABA, 2) Conurbano Bonaerense Norte, 3) Conurbano Bonaerense Oeste, 4) Conurbano Bonaerense Sur, 5) Gran Rosario, 6) Gran Córdoba, 7) Gran Mendoza, 8) Gran Tucumán - Tafí Viejo y 9) Resto Urbano Interior (ciudades no metropolitanas de más de 80 mil habitantes).

Si se agregan los sub-dominios 1,2,3,4 se obtiene el dominio del Gran Buenos Aires y si agregan los sub-dominios 5,6,7,8 se obtiene el complemento de las ciudades Metropolitanas, lo que permite a su turno, que la agregación de 1,2,3,4,5,6,7,8 conforme el dominio de las ciudades Metropolitanas. Luego, si a lo anterior, se agrega el sub-dominio 9, se obtiene el gran dominio del total urbano de la Argentina. Esto se expresa de modo más claro en la Figura A4.1.

[3] Este punto suele generar alguna controversia en algunos usuarios de muestras. Si la selección de los *aglomerados* no es aleatoria y luego no se realiza ninguna acción para su corrección, la muestra no es aleatoria.

Es cierto que luego, en base a información secundaria, se podría suponer que determinado aglomerado es similar a otro aglomerado, y por lo tanto, lo encontrado en determinado aglomerado se puede inferir a otro. Esto tipo de inferencias son válidas y legítimas, pero no es una inferencia estadística que se deriva del diseño muestral. Es una inferencia del investigador, como cuando asume que de la *población* efectivamente analizada se pueden inferir los datos observado al *universo* que hipoteticamente se quiere analizar. La inferencia muestral es desde la *muestra* de una *población* hacia esta, no de ella hacia el *universo*. Para una discusión similar (Kish, 2004, pp. 27-31)

Figura A4.1. Dominios y subdominios del estudio.



Todo diseño muestral tiene como componente necesario al menos un dominio sobre el cual se quiere realizar una inferencia. En este caso el dominio principal de la muestra fueron los hogares (urbanos) de los aglomerados de más de 80.000 habitantes.

Una manera de otorgarle racionalidad a esta decisión es suponer que se buscaba tener una muestra que fuera representativa de *cada uno* de los aglomerados de más de 500.000 habitantes de toda la Argentina y en forma adicional, del *conjunto* (ya no de cada uno) los aglomerados del país con más de 80.000 habitantes.[4]

A continuación se presentan en la tabla A4.1, para cada aglomerado urbano incluido en la muestra, los volúmenes poblacionales y de hogares según los datos del último Censo Nacional de Población, Hogares y Viviendas llevado adelante en el año 2010.

[4] Luego, aunque de forma estricta, se encuentre ajeno al diseño muestral, se asume que el dominio principal es representativo del universo conceptual sobre el cual se quiere realizar la inferencia, que en esta investigación fue el conjunto de los hogares urbanos de la Argentina para el período de estudio.

Tabla A4.1. Población y Hogares según aglomerado urbano de EDSA 2010. Datos CNPHyV 2010.

Dominios	Aglomerados	Población			Hogares 2001
		Total poblacional	0-17 años	18 años y +	
CABA	Ciudad Autónoma de Buenos Aires	2.890.151	565.032	2.325.119	1.150.134
Conurbano Bonaerense[5]	Conurbano Zona Norte	2.679.738	794.726	1.885.012	806.001
	Conurbano Zona Oeste	3.960.318	1.228.645	2.731.673	1.137.591
	Conurbano Zona Sur	3.909.613	1.196.099	2.713.514	1.160.884
	Total	10.549.669	3.219.470	7.330.199	3.104.476
Otras grandes áreas metropolitanas	Gran Rosario	1.270.103	328.476	941.627	417.690
	Gran Córdoba	1.505.131	426.980	1.078.151	468.922
	Gran Mendoza	933.526	205.952	727.574	317.578
	Gran Tucumán – Tafí Viejo	797.557	241.773	555.785	216.182
	Total	4.506.317	1.203.180	3.303.137	1.420.373
Resto Urbano	Mar del Plata	614.350	160.242	454.108	208.222
	Gran Salta	536.113	181.099	355.014	137.000
	Gran Paraná	339.930	99.223	240.707	105.030
	Gran Resistencia	390.874	127.843	263.031	110.100
	Gran San Juan	441.477	142.504	298.973	119.049
	Neuquén-Plottier-Cipoletti	345.097	107.185	237.912	108.346
	Zárate	114.269	35.940	78.329	34.013
	La Rioja	180.995	60.373	120.622	48.916
	Goya	89.959	32.247	57.712	24.344
	San Rafael	188.018	57.415	130.603	56.391
	Comodoro Rivadavia	161.326	48.398	112.928	48.398
	Ushuaia y Río Grande	126.998	42.188	84.810	38.948
	Total	3.529.406	1.094.657	2.434.749	1.038.756
Total general		21.475.543	6.082.339	15.393.204	6.713.739

Fuente: Elaboración propia en base a Censo de Población, hogares y Vivienda de 2010.

[5] Zona Norte: Vicente López, San Isidro, San Fernando, Tigre, General San Martín, San Miguel, Malvinas Argentinas, José C. Paz y Pilar. Zona Oeste: La Matanza, Merlo, Moreno, Morón, Hurlingham, Ituzaingó, Tres de Febrero, Cañuelas, General Rodríguez y Marcos Paz. Zona Sur: Avellaneda, Quilmes, Berazategui, Florencio Varela, Lanús, Lomas De Zamora, Almirante Brown, Esteban Echeverría, Ezeiza, Presidente Perón y San Vicente.

A4.3 Estrategia y plan de muestreo[6]

La explicitación de las acciones realizadas en las distintas etapas del muestreo es importante porque permite, *a posteriori*, la realización de estimaciones del error muestral para muestras complejas. En efecto, sin un detalle de aquellas acciones es inviable llevar a cabo un análisis de muestras complejas.

A4.3.1 Primera Etapa

En la primera etapa del muestreo, el dominio total del estudio (ciudades de 80 mil habitantes o más) es dividido en dos sub- dominios:

a) grandes áreas metropolitanas y

b) resto urbano interior conformado por ciudades no metropolitanas de más de 80 mil habitantes.

Con cada uno de esos dos conjuntos se adoptan estrategias diferentes. Por un lado, en el primero de los dos dominios se relevan siete aglomerados urbanos, que por un lado constituyen una buena representación de la realidad urbana de las grandes áreas metropolitanas, cuyos habitantes y hogares representan el universo de estudio de la EDSA–Bicentenario, a la vez que en 2010 abarcaban a casi el 50% de la población del país.

Los siete aglomerados urbanos de este primer conjunto son: 1) Gran Buenos Aires, y dentro de este, 2) la Ciudad Autónoma de Buenos Aires y 3) el Conurbano Bonaerense; 4) Gran Rosario; 5) Gran Córdoba; 6) Gran Mendoza; y 7) Gran Tucumán–Tafí viejo.

Para el segundo gran dominio (el Resto Urbano Interior conformado por ciudades no metropolitanas de más de 80 mil habitantes) se adopta la siguiente estrategia. En base a información censal se estratifica a los aglomerados en dos grupos de acuerdo con el volumen poblacional.

Por un lado, ciudades de 80 mil a 200 mil habitantes, y por otro ciudades o aglomerados intermedios en los que la población supera los 200 mil residentes, pero que no incluyen a las áreas metropolitanas del primer dominio. Mientras que, como se indicó, en el primer dominio se seleccionan aglomerados fijos, en cada uno de los dos estratos del segundo dominio se aplica una selección de ciudades mediante una estrategia de conglomeración muestral, aplicando como método de selección un muestreo aleatorio con probabilidades proporcionales al tamaño de cada ciudad o aglomerado.

[6] La explicitación de las etapas es importante porque permite, a posteriori, la realización de estimaciones del error muestral para muestra complejas. En efecto, para la realización de las correcciones se debe saber qué acciones se realizaron en cada paso.

A partir de lo anterior, en el primer estrato quedaron seleccionadas Mar del Plata, Gran Salta, Gran Paraná, Gran Residencia, Gran San Juán y Neuquén-Plottier. En el segundo estrato de ciudades de 80 mil a 200 mil habitantes quedaron incluidas Zárate, la Rioja Capital, Goya, San Rafaél, Comodoro Rivadavia y el eje Ushuaia-Río Grande. Nuevamente, ambos grupos conforman un segundo gran dominio de estudio –el Resto Urbano Interior- y las estimaciones que realiza la EDSA–Bicentenario son representativas de ese total, mas no así de cada una de las ciudades o aglomerados que lo componen.

A4.3.2 Segunda Etapa

En una segunda etapa, al interior de cada aglomerado urbano se eligen radios censales (unidades secundarias de muestreo), siguiendo una estrategia de estratificación a los efectos de minimizar los coeficientes de variación de las principales estimaciones a realizar.

La variable utilizada como criterio de estratificación es el promedio de años de educación del jefe del hogar por radio censal. Las razones para utilizar esta variable son cuatro:

a) muestra un importante grado de correlación con las principales variables de interés;

b) es 'proxi' del nivel socioeconómico de los hogares y, a la vez, también da cuenta del capital cultural de los mismos, con lo que constituye una buena aproximación a la estratificación social;

c) permite una pos estratificación de la muestra; y

d) ha dado buenos resultados en todos los relevamientos anteriores de la propia EDSA.

El total de radios censales o puntos de muestreo seleccionados son 950, los que, a partir de la variable de estratificación, son divididos en 5 estratos en el caso de los aglomerados con al menos 200 mil habitantes y en 3 estratos en los aglomerados que cuentan con una población menor. En esta línea, con el propósito de mejorar la captación de los casos extremos (de mayor y menor nivel socioeducativo), los 5 estratos de los aglomerados con al menos 200 mil habitantes se dividen en 3 grupos centrales con el 1/4 de los casos cada uno y 2 grupos en los extremos con 1/8 cada uno.

Si bien el procedimiento descripto respeta la proporcionalidad del tamaño de los radios y estratos a nivel de cada aglomerado, esto no es así para el total urbano, lo que lleva a la necesidad de la construcción posterior de ponderadores correctores y expansores que permitan replicar la estructura censal. En cada uno de los dos tipos

de aglomerados (clasificados por su tamaño) los estratos quedan conformados como se detalla en la tabla A4.2.

Tabla A4.2. Porcentajes de radios censales y de hogares según estratificación muestral, tamaño muestral y tamaño del aglomerado de la EDSA.

	Estratos muestrales	Proporción asignada en la estratificación
Ciudades o aglomerados de 200 mil habitantes o más	Muy bajo	12,5% (1/8)
	Bajo	25% (1/4)
	Medio	25% (1/4)
	Medio alto	25% (1/4)
	Alto	12,5% (1/8)
Ciudades o aglomerados de 80 mil habitantes a 200 mil habitantes	Bajo	33,3 (1/3)
	Medio	33,3 (1/3)
	Alto	33,3 (1/3)

Tabla A4.3. Cantidad de hogares y puntos muestrales según aglomerado urbano para EDSA.

Grupo	Aglomerado	Hogares	Puntos muestra
CABA	Ciudad Autónoma de Buenos Aires	438	72
Conurbano Bonaerense	Conurbano Zona Norte	432	72
	Conurbano Zona Oeste	432	72
	Conurbano Zona Sur	432	72
	Total	1296	216
Otras grandes áreas metropolitanas	Gran Rosario	624	104
	Gran Córdoba	624	104
	Gran San Miguel de Tucumán y Tafí Viejo	624	104
	Gran Mendoza	624	104
	Total	2496	416
Resto urbano	Mar del Plata	192	32
	Gran Salta	192	32
	Gran Paraná	192	32
	Gran Resistencia	192	32
	Gran San Juan	192	32
	Neuquén-Plottier-Cipolletti	192	32
	Zárate	54	9
	La Rioja	54	9
	Goya	54	9
	San Rafael	54	9
	Comodoro Rivadavia	54	9
	Ushuaia y Río Grande	54	9
	Total	1476	246
Total General		5706	951

A4.3.3 Tercera Etapa

En la tercera etapa, una vez elegidas las unidades secundarias, se aplica una selección sistemática de viviendas y hogares (unidades terciarias). Dado que se estableció un tamaño de muestra alrededor de 5700 hogares, con una asignación esperada de 6 hogares por punto muestra, el total de estos últimos son distribuidos entre los aglomerados siguiendo un criterio de no proporcionalidad, a los efectos de encontrar un óptimo en materia de reducción de los márgenes de error muestral.

El número de radios asignados a cada aglomerado depende de la manera en que se determinaron los dominios de representatividad estadística y de la necesidad de poder predicar sobre cada dominio dependiendo del número de hogares esperados en cada caso.

A4.3.4 Cuarta Etapa

Mientras que en la tercera etapa se aplica un muestreo sistemático de hogares, en la cuarta y última etapa se busca llegar al segundo universo a describir, las personas de 18 años y más, las que responden por sí mismas y por el hogar del que forman parte. En este caso, se utiliza un criterio de cuotas por sexo y grupo de edad de acuerdo a la estructura demográfica según datos censales. Las cantidades de radios / puntos muestra y de hogares asignados a cada aglomerado urbano en 2010 son los que se detallan en la tabla A4.3.

A4.4 Ponderadores para los dominios originales de la muestra

Como se indicó previamente, dada la estrategia de muestreo y el diseño adoptado, la muestra final de alrededor de 5700 hogares y sus respectivas personas es autoponderada dentro de cada aglomerado, pero no así en lo que respecta al total urbano.

A partir de esto, una vez finalizado el trabajo de campo se procede a la construcción de diferentes tipos de ponderadores que, por un lado corrigen el peso que cada caso tiene en la población real y, por otro, permiten expandir la muestra a los diferentes dominios de estudio.

Como es habitual en el muestreo, los ponderadores se construyen a partir de la inversa del producto de las probabilidades de inclusión de primer orden en cada etapa, las que, a su vez, se encuentran sujetas al diseño y a las estrategias utilizadas en cada una de aquellas.

A continuación se presenta una formalización de la construcción de los ponderadores de cada hogar y de cada persona de la muestra de la EDSA-Bicentenario. Para simplificar, se toma en principio un único aglomerado urbano, pero mediante una serie de sumatorias finales puede generalizarse al conjunto de los 20 aglomerados del estudio.

De este modo, divídase al conjunto R compuesto de radios censales ($R = r_1, r_2, r_3, \dots r_n$) del aglomerado urbano a_1 del conjunto de aglomerados A ($A = a_1, a_2, a_3, \dots a_{20}$) en $E = 5$ estratos muestrales de tamaño M_e. De este modo:

A4.1

$$R = \sum_{e=1}^{E} M_e = EM_e$$

Como se indicó, la variable de estratificación es el promedio de años de educación del jefe del hogar por radio censal. Dentro de cada estrato e se extrae una muestra S_e del conjunto de radios R_e de tamaño M_e. El diseño de esta primera etapa bajo un muestreo aleatorio simple de radios dentro de un estrato dado puede expresarse así:

A4.2

$$p(S_e) = 1/(\frac{M_e}{m_e})$$ si y sólo sí s es de tamaño m_e y $p(s) = 0$.

Con esto, dentro de cada estrato e, la probabilidad de inclusión de un radio r va a estar dada por:

A4.3

$$\pi_{e,r} = m_e/M_e.$$

Sin embargo, como se indicó, cada estrato tiene un peso w_e diferente en el conjunto de la muestra. Entonces, la probabilidad de inclusión de ese radio, dado que pertenece al estrato e, se corrige como:

A4.4

$$\pi_{e,r} = w_e E(\frac{m_e}{M_e})$$

y su ponderador va a ser:

A4.5

$$w_{e,r} = M_e w_e E / m_e$$

Mientras que en la etapa anterior se utiliza un muestreo aleatorio simple de radios (con iguales probabilidades) dentro de $E = 5$ estratos muestrales, a los que posteriormente se los pondera de manera diferente a partir de un vector de pesos W, en la etapa de selección de hogares dentro de cada radio censal se aplica un muestreo sistemático.

Sea H el total de hogares que residen en los R radios censales del aglomerado urbano a. Sea, a su vez, $H_{e,r}$ el total de hogares en el radio r seleccionado en la primera etapa dentro del estrato e, y $h_{e,r}$ el tamaño muestral (de hogares) en ese mismo radio y estrato; bajo un muestreo sistemático de hogares, el intervalo de selección k va a estar dado por:

A4.6

$k = H_{e,r}/h_{e,r}$ y comienza a partir de un número aleatorio x, donde $1 \leq x \leq k$

Los hogares seleccionados son $x, x + k, x + 2k, \ldots$, hasta llegar a completar $h_{e,r}$. Pero dado que en un muestreo sistemático los hogares solo pueden ser incluidos si y solo si caen dentro de las sub muestras S en las que es particionado el radio censal, el diseño muestral para esta etapa es:

A4.7

$p(s) = 1/k$ si $s \in S_{e,r}$ o $p(s) = 0$ si $\notin S_{e,r}$

Debido a lo anterior, la probabilidad de inclusión del hogar i en un radio muestral determinado (prescindiendo momentáneamente en la notación del estrato) va a estar dada por:

A4.8

$$\pi_{r,i} = \frac{1}{k} = h_r/H_r \; \forall i \in r$$

Donde r es el conjunto de radios censales que ya habían sido seleccionados en la etapa anterior.

Integrando la etapa previa y la estratificación, la probabilidad de inclusión $\pi_{e,r,i}$ del hogar i del radio censal r perteneciente al estrato e en el aglomerado urbano a, va a estar dada por:

A4.9

$$\pi_{e,r,i} = w_e E(\frac{m_e}{M_e})(\frac{h_{e,r}}{H_{e,r}})$$

y su ponderador va a ser:

A4.10

$$w_{e,r,i} = \frac{M_e H_{e,r} w_e E}{m_e h_{e,r}}$$

Por último, los ponderadores de la EDSA–para los encuestados adultos del hogar, que corresponde a la cuarta etapa del muestreo, se calculan del siguiente modo: sea N el total de personas de 18 años o más que residen en los H hogares de los R radios organizados en E estratos del aglomerado urbano a, y sea n la cantidad de personas que son seleccionadas para la muestra del aglomerado, estrato y radio específico; siendo que se utilizan cuotas de edad y sexo, en esta última etapa no entra el azar, pero dado que dichas cuotas respetan la proporcionalidad censal de esas categorías en $a, e,$ y r, la probabilidad de inclusión $\pi_{a,e,r,i,j}$ del sujeto j, va a estar dada por:

A4.11

$$\pi_{a,e,r,i,j} = w_e E \left(\frac{m_e}{M_e}\right)\left(\frac{h_{e,r}}{H_{e,r}}\right)\left(\frac{n_{a,e,r,i}}{N_{a,e,r,i}}\right)$$

y su ponderador va a ser:

A4.12

$$w_{a,e,r,i,j} = \frac{M_e H_{e,r} N_{a,e,r,i} w_e E}{m_e h_{e,r} n_{a,e,r,i}}$$

Finalmente, si se toma en cuenta al conjunto A de aglomerados, de la muestra y siendo que cada uno de ellos tiene un peso w_a en el conjunto de los aglomerados, el total de hogares que representa la EDSA–Bicentenario se reconstruye como:

A4.13

$$\sum_{a=1}^{A} w_a \sum_{e=1}^{E} w_e E \sum_{r=1}^{me} w_r \sum_{i=1}^{he,r} w_i = H$$

y el total de personas de 18 años o más, como

A4.14

$$\sum_{a=1}^{A} w_a \sum_{e=1}^{E} w_e E \sum_{r=1}^{me} w_r \sum_{j=1}^{he,r} w_j = N$$

Cabe señalar que, con posterioridad a la construcción de los ponderadores, también se efectúan arreglos, habida cuenta que en la práctica no es posible contar con una distribución libre de sesgos producidos por la 'no-respuesta'.

En este sentido, si bien los diseños muestrales y los trabajos de campo prevén estrategias para disminuir este tipo de problema, los sesgos logran atenuarse, pero no corregirse en su totalidad.

Como consecuencia de lo anterior existe una ‘post-calibración’ en donde se calibran los pesos de los hogares w_i y de las personas w_j con la ayuda de información auxiliar conocida o preestablecida a partir de registros o fuentes externas validadas, como los datos censales e información ad-hoc proveniente de la propia encuesta.

Esta segunda corrección atiende a considerar las diferencias entre la muestra observada y la esperada de acuerdo con los atributos de los hogares y/o las personas que componen los hogares seleccionados. Para ello se utiliza el procedimiento de ‘calibración por marginales fijos’ que estima las frecuencias ‘condicionales’ de una tabla de contingencia según los parámetros poblacionales conocidos (Neville y Sarndall, 1992).

A4.5 Estimación de los errores muestrales

En la EDSA-Bicentenario (2010-2016), al ser una muestra multipropósito, no hay una única variable a ser estudiada y, por lo tanto, no existe un único margen de error muestral. Cada estimación cuenta con su propio margen de error, el cual depende fundamentalmente de tres aspectos centrales:

a) la varianza o dispersión del indicador a estimar,
b) el intervalo de confianza en el que se pretenda realizar las estimaciones y
c) el tamaño de la muestra y de las sub-muestras (para cuando no se examina el conjunto de la población).

En efecto, cuando se trabajan con análisis multivariados de datos categóricos, la cantidad de casos de cada celda son muy sensibles no sólo a la cantidad de variables, sino también a la cantidad de categorías de cada una de ellas. De este modo, es usual que el investigador llega a tener problemas con las frecuencias de cada celda y con la propia estimación de las frecuencias esperadas de algunos modelos.

El problema se agudiza cuando se tratan fenómenos como los tratados en esta tesis. Por ejemplo, podría parecer que utilizar un esquema de clases de 4 categorías, en donde una sola de ellas consuma el 50% de los casos es un claro indicio de ineficiencia estadística. Sin embargo, suponer que los datos de una muestra se aprovechan mejor, desde el punto de vista de la estadística inferencial, si se construyen sistemas de categorías que posean frecuencias similares es propio de los análisis univariados.

En los análisis multivariados también cuenta si existe una fuerte asociación entre las diferentes categorías de una variable y las categorías de otra. La razón es que a medida que agregan variables (y sus respectivas categorías) al análisis, se aumentan la cantidad de celdas a comparar al tiempo que se reducen la cantidad de casos en cada una de ellas. En otras palabras, importa la cantidad de casos que efectivamente se ubican en cada celda.[7]

A modo de resumen se reproducen los márgenes de error de las principales variables utilizadas en el presente trabajo. Estos se calcularon en base a una heterogeneidad proporción poblacional de un 50 % y un nivel de confianza del 95%.[8]

[7] Lo expuesto en el cuerpo del texto, en el caso del trabajo con muestras, debe complementarse con que la eficiencia de una estimación se reduce fuertemente y de forma no lineal cuando uno se acerca a una escasa cantidad de casos, por lo que la decisión final implica un *trade-off* difícil de explicitar en donde existe una fuerza, empujada por las leyes del muestreo a alejarse de categorías con pocos casos y otra, en sentido opuesto, empujada por una racionalidad categorial que teme que bajo el género de la categorías colapsadas se escondan especies con comportamientos diferenciales. Ver al respecto (King et al., 1994, pp. 66-74).

[8] Para mayores ejemplos puede consultarse (Adaszko, 2013).

Teniendo en cuenta las condiciones antes mencionadas, para el total de la muestra original (5682 casos post consistencia) el margen de error es de +-1,3pp..

Para los análisis aquí propuestos el *n* sobre el cual se hicieron muchos de los análisis del capítulo empírico es 3317. La deserción de casos con respecto al párrafo anterior, se debe principalmente, al filtro etario utilizado, que restringía los casos útiles a los comprendidos entre aquellos entre 27 y 70 años. De forma secundaria, la baja de casos también se tiene su explicación en la falta de respuesta de alguna de preguntas necesarias para hacer los análisis trivariados de origen de clase, bien posicional y período histórico.

A continuación se presentan, en la tabla A4.4, los errores muestrales del esquema de clase construido, diferenciado para cada período histórico analizado.

Tabla A4.4. Estimación de los errores muestrales del esquema de clase de origen según período. Supuestos de un intervalo de confianza del 95% y una proporción del 50%.

Clase de Origen	Tamaño de muestra					Margen de error				
	1955 1965	1966 1976	1977 1990	1991 2001	Total	1955 1965	1966 1976	1977 1990	1991 2001	Total
Clase de Servicio	40	67	103	93	303	15,5	11,9	9,6	10,1	5,6
Clase Intermedia	70	142	193	163	568	11,7	8,2	7,6	7,6	4,0
Pequeños Autónomos	99	157	209	218	617	9,8	7,8	6,6	6,6	3,9
Clase Trabajadora	310	434	570	515	1829	5,5	4,6	4,3	4,3	2,2
Total	519	800	1076	989	3317	4,2	3.4	2,9	3,0	1,5