



Buenas cifras,
mejores vidas

A solid yellow circle is positioned to the left of the main title, partially overlapping the red-tinted image area.

Encuesta Nacional Multipropósito de Hogares

(Seguimiento al Plan
Nacional de Desarrollo)
Marzo, 2021



01.

Documento metodológico de diseño muestral





Instituto Nacional de Estadística y Censos (INEC)

Documento Metodológico de Diseño Muestral de la Encuesta Multipropósito 2020

Dirección:

Dirección de Infraestructura Estadística y Muestreo

Elaborado por:

Francisco Céspedes
Ángel Gaibor
Javier Núñez

Revisado por:

Francisco Céspedes

Aprobado por:

David Sánchez

Quito -Ecuador, 2021





Contenido

Introducción.....	6
Diseño Muestral Encuesta Multipropósito 2020	7
Determinación del Marco de Muestreo para encuestas de hogares	7
Universo de investigación	7
Unidad de observación.....	8
Unidad de análisis	8
Cobertura geográfica	8
Tipo de muestreo.....	8
Dominios de estudio.....	8
Tamaño y selección de la muestra de la Encuesta Multipropósito	8
Tamaño de la muestra	8
Asignación de la muestra	11
Selección de la muestra.....	11
Cobertura de viviendas planificada y efectiva	13
Cobertura planificada.....	13
Cobertura efectiva	13
Cálculo de los factores de expansión	14
Esquema de ponderación de Encuesta Multipropósito	14
Factor de expansión teórico.....	14
Factor de expansión ajustado por cobertura	16
Calibración de los factores de expansión.....	17
Recorte de los factores de expansión extremos	20
Estimación de características de la población	21
Estimación de errores.....	21
Métodos de estimación de errores para diseños muestrales complejos	22
Referencias	24
Anexos	25
Anexo 1: Esquema de calibración propuesto	25
Anexo 2: Medidas para evaluar el esquema de calibración planteado (Esquema 1)	25
Anexo 3:.....	28





Lista de Tablas

Tabla 1. Conciliación entre variables para el Formulario 1.....	9
Tabla 2. Conciliación entre variables para el Formulario 2.....	9
Tabla 3. Tamaños de muestra Multipropósito diciembre 2020.	10
Tabla 4. Distribución de la cobertura planificada de viviendas por estrato socioeconómico	13
Tabla 5. Distribución de la cobertura efectiva de viviendas por estrato socioeconómico	13
Tabla 6. Variables requeridas para declaración del diseño muestral – ENEMDU.....	23

Lista de Gráficos

Gráfico 1. Comparación de los factores de expansión ajustados por cobertura y calibrados por dominio y grupo de edad para la base completa.....	19
Gráfico 2. Comparación de los factores de expansión ajustados por cobertura y calibrados por dominio y grupo de edad para la base del formulario 1	28
Gráfico 3. Comparación de los factores de expansión ajustados por cobertura y calibrados por dominio y grupo de edad para la base del formulario 2.....	29





Introducción

El Instituto Nacional de Estadística y Censos – INEC, como ente rector de la estadística nacional es el encargado de llevar a cabo una serie de encuestas que ayudan a la planificación nacional, además de la producción estadística del país. En el marco del monitoreo constante de los indicadores que forman parte de Plan Nacional de Desarrollo, Objetivos de Desarrollo Sostenible, se presenta la “Encuesta Multipropósito 2020”, en su tercera etapa para continuar con el monitoreo de los indicadores de avance en cada uno de los ejes de seguimiento.

Para cumplir con los propósitos de la planificación nacional e ir adecuando las diversas necesidades que se van presentando en el camino de la investigación, el Instituto Nacional de Estadística y Censos – INEC, ha tomado en cuenta un plan de mejoras adaptadas al Modelo de Producción Estadística, es por lo que es necesario actualizar la información de manera continua y dar cumplimiento a los diversos compromisos con la sociedad demandante de información de calidad y oportuna.

El Diseño Muestral es uno de los procesos medulares en el desarrollo de la investigación estadística, es por ello, la necesidad de elaborar un diseño muestral que cumpla con cada uno de los detalles demandado por los diversos usuarios, en este contexto, para el año 2020 dando cumplimiento a este compromiso de considerar un tamaño de muestra suficiente, mediante la información de la Encuesta Multipropósito 2019, con el indicador porcentaje de población que realiza actividad física más de 3,5 horas a la semana, se actualizó el tamaño de muestra para obtener los indicadores a nivel nacional urbano y rural.

El Modelo de Producción Estadística, sigue procesos estándares por cada una de las operaciones estadísticas, por esta razón el diseño muestral de la Encuesta Multipropósito 2019, sigue varias etapas desarrolladas en el presente documento.





Diseño Muestral Encuesta Multipropósito 2020

Determinación del Marco de Muestreo para encuestas de hogares

El Marco de Muestreo se basa en los resultados definitivos y la cartografía del VII Censo de Población y VI de Vivienda del 2010 (CPV-2010); este Marco tiene la característica de ser constituido por áreas geográficas, que tienen límites perfectamente definidos e identificables sobre el terreno.

El Marco ha tenido actualizaciones parciales en función a los cambios presentados en las unidades de observación (viviendas) en determinadas áreas geográficas, debido principalmente a los movimientos y dinámica demográfica propios de la población, así como los cambios en estructuras habitacionales; que en su conjunto hacen necesario un mantenimiento continuo del Marco de Muestreo.

Se han ejecutado, en el periodo 2014-2017, diferentes procesos de actualización cartográfica con diferente alcance, los cuales se detallan a continuación:

- Encuesta Condiciones de Vida 2013-2014: 2.425 sectores censales.
- Actualización ENEMDU 2014: 5.564 sectores censales.
- Proyecto 2015: 548 sectores censales.
- Actualización 2017: 1.779 sectores censales.

El Marco de Muestreo para encuestas de hogares se lo ha dividido por dominios de estudio, y dentro de ellos sus correspondientes UPM, a las cuales se asignó un estrato tomando principalmente sus características geográficas, socio-económicas y socio-demográficas, con la finalidad de mejorar la precisión y exactitud de los estimadores, minimizando su varianza¹.

Universo de investigación

El universo de estudio de la Encuesta Multipropósito 2020 son todas las personas residentes en las viviendas del Ecuador, exceptuando la población que reside en viviendas colectivas, viviendas flotantes y población indigente (sin techo).

¹ Para más información de la construcción y estratificación del Marco de Muestreo para encuestas de hogares referirse al siguiente link: https://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2018/Disenio_Muestral_2018/SIEH%20-MMM.pdf





Unidad de observación

La unidad de observación son todas las viviendas particulares ocupadas que se encuentran en territorio nacional, las mismas que tienen ligada su identificación geográfica mediante fuentes cartográficas.

Unidad de análisis

Debido a que se trata de una encuesta de propósitos múltiples existen dos unidades de análisis: el hogar particular y las personas miembros de hogar divididas en los diferentes grupos de interés de la operación estadística.

Cobertura geográfica

La cobertura geográfica está definida por las viviendas ocupadas que se encuentren ubicadas dentro del territorio ecuatoriano incluyendo la región insular.

Tipo de muestreo

El tipo de muestreo de la Encuesta Multipropósito 2020 corresponde a un muestreo probabilístico estratificado bietápico de elementos.

Dominios de estudio

Los dominios de estudio para la Encuesta Multipropósito 2020 son los siguientes:

- Nacional
- Nacional Urbano
- Nacional Rural

Tamaño y selección de la muestra de la Encuesta Multipropósito

A partir de las expresiones de tamaño de muestra para muestras complejas, se calcula el tamaño de muestra necesario para lograr una precisión adecuada en la inferencia.

Tamaño de la muestra

El tamaño de muestra de la Encuesta Multipropósito 2020 fue calculado considerando los siguientes parámetros:





- Margen de error relativo del 12%.
- Nivel de confianza del 95%.
- Tasa de no respuesta del 20%.

Los algoritmos que se utilizaron para el cálculo de los tamaños de muestra tanto de personas, como de viviendas y UPM, se detallan a continuación.

Tamaño de muestra de personas

Para la determinación del tamaño de la muestra se requiere establecer la característica o características a estimar, el nivel de confianza, la precisión requerida y la tasa de no respuesta.

En este sentido, la variable de investigación, en base a la cual se calcula el tamaño de la muestra es porcentaje de población que realiza actividad física más de 3,5 horas a la semana, cuyos tamaños son los que presentan mayor muestra una vez conciliada la muestra.

Tabla 1. Conciliación entre variables para el Formulario 1.

Desagregación	Estimación	Error Estándar	I. Inferior	I. Superior	CV	Efecto de Diseño
Urbano	0,12	0,003	0,11	0,13	2,82	2,41
Rural	0,07	0,004	0,06	0,08	5,00	1,88

Tabla 2. Conciliación entre variables para el Formulario 2.

Desagregación	Estimación	Error Estándar	I. Inferior	I. Superior	CV	Efecto de Diseño
Urbano	0,12	0,003	0,11	0,13	2,82	2,41
Rural	0,18	0,01	0,17	0,20	3,68	1,08

La expresión matemática utilizada para el cálculo de tamaño de muestra de personas es:

$$n_k \geq \frac{z_{1-\frac{\alpha}{2}}^2 N_k p(1-p)}{(e_r p)^2 (N_k - 1) + z_{1-\frac{\alpha}{2}}^2 p(1-p)} * \frac{deff}{1 - tnr}$$

Donde:

- n_k = Tama;o de la muestra de personas para el dominio de estudio k
- $z_{1-\frac{\alpha}{2}}$ = Nivel de confianza correspondiente al percentil α
- N_k = Total de personas de 0 a 5 a; os en el dominio de estudio k
- p = Probabilidad de incidencia de la variable a investigar
- e_r = Error relativo
- $deff$ = Efecto de dise;o
- tnr = tasa de no respuesta

A su vez, la varianza S^2 es calculada a través del siguiente algoritmo:





$$S^2 = P(1 - P) * deff$$

donde *deff* es el efecto de diseño calculado a partir del coeficiente de correlación intraclase entre las UPM y la variable de diseño (desempleo), y el número promedio de viviendas levantadas por UPM.

$$deff = 1 + (\bar{n} - 1) * \rho$$

Tamaño de muestra de viviendas

Es necesario calcular el número total de viviendas que deben ser seleccionados para lograr entrevistar a todas las personas que serán observadas en el punto anterior. El número de viviendas que deben ser seleccionados estará determinado por la muestra de personas (n_k), número promedio de personas por hogar (b_k) y el porcentaje de personas con la característica de interés, de la siguiente forma:

$$n_{vk} = \frac{n_k}{r * b_k}$$

Tamaño de muestra de UPM

Las viviendas y las personas que participan en la encuesta forman parte de UPM previamente seleccionadas. En este paso final, es necesario calcular el número de UPM que deben ser seleccionadas en la primera etapa de muestreo a partir de la relación:

$$n_{UPMk} = \frac{n_{vk}}{\text{Carga técnica operativa}}$$

La carga técnica operativa se refiere al número de viviendas asignadas a cada encuestador como carga de trabajo; que fue el resultado de un previo análisis de correlación intraclase donde se pudo verificar, mediante simulaciones matemáticas, el número de observaciones necesarias para minimizar la varianza dentro de cada UPM. Como resultado de este procedimiento se definió tanto operativa como técnicamente que el número de viviendas a investigarse por UPM será doce (12).

A través de la aplicación de los algoritmos de cálculo, y tomando en cuenta que se levantan 12 viviendas por UPM junto con el presupuesto disponible de la operación estadística, se obtuvo un tamaño de muestra de 12.060 viviendas, cuya distribución por dominio se muestra en la Tabla 1:

Tabla 3. Tamaños de muestra Multipropósito diciembre 2020.

Subpoblación	Viviendas del marco	UPM del marco	Muestra UPM	Muestra Viviendas
Urbano	2.715.812	79.647	390	4.680
Rural	1.002.506	27.352	615	7.380
Nacional	3.718.318	106.999	1.005	12.060





Asignación de la muestra

La muestra es distribuida proporcionalmente por cada estrato dentro de cada dominio de estudio.

$$n_{ki} = n_k \cdot \frac{N_{ki}}{N_k}$$

Donde:

- n_{ki} = *Tamaño de la muestra para el estrato i del dominio de estudio k.*
- n_k = *Tamaño de la muestra para el dominio de estudio k.*
- N_{ki} = *Total de viviendas ocupadas en el estrato i del dominio de estudio k.*
- N_k = *Total de viviendas ocupadas en el dominio de estudio k.*

Selección de la muestra

La selección de la muestra se realiza en forma aleatoria, en dos etapas:

- **Primera etapa:** selección de Unidades Primarias de Muestreo (UPM) por estrato.
- **Segunda etapa:** selección de viviendas ocupadas dentro de cada una de las UPM seleccionadas en la primera etapa.

La selección de las UPM que forman parte de la muestra se realiza de manera independiente en cada uno de los dominios de forma aleatoria, asignando a cada UPM igual probabilidad de ser seleccionada. De la misma forma, la selección de viviendas es aleatoria dentro de cada UPM seleccionada.

Nota: La información levantada en campo de la encuesta Multipropósito diciembre 2020 se agrupa en 3 bases de datos, ya que se utilizaron tres tipos de formularios:

Formulario	Variables consideradas por formulario
Formulario 1	Porcentaje de la población que realiza actividad física más de 3,5 horas a la semana
	Porcentaje de personas que utiliza internet
	Confianza en el Consejo de la Judicatura / rural
	Confianza en la Fiscalía General del Estado
	Confianza en las Fuerzas Armadas
	Confianza en el Consejo de la Judicatura
	Confianza en la Defensoría Pública
	Confianza en la Policía Nacional





Formulario 2

Porcentaje de hogares con percepción positiva en relación a los servicios de salud pública

Porcentaje de hogares con al menos un computador

Porcentaje de hogares con acceso a internet

Por tal hecho la muestra se dividió en dos grupos iguales dentro de cada UPM seleccionada, de las doce viviendas por UPM seis tenían las preguntas comunes y del formulario 1 y las otras seis tenían las preguntas comunes y del formulario 2, es necesario mencionar una vez más que la variable que representa la muestra y es común en los dos formularios es el porcentaje de la población que realiza actividad física más de 3,5 horas a la semana.





Cobertura de viviendas planificada y efectiva

Cobertura planificada

A continuación, se presenta la cobertura planificada de viviendas de la Encuesta Multipropósito diciembre 2020:

Tabla 4. Distribución de la cobertura planificada de viviendas por estrato socioeconómico

Formulario	Dominio	Estrato Socioeconómico			Total
		Alto (1)	Medio (2)	Bajo (3)	
Formulario 1	Urbano	750	972	618	2.340
	Rural	1.362	1.482	846	3.690
Formulario 2	Urbano	750	972	618	2.340
	Rural	1.362	1.482	846	3.690
Formulario común	Urbano	1.500	1.944	1.236	4.680
	Rural	2.724	2.964	1.692	7.380
Total viviendas	Nacional	4.224	4.908	2.928	12.060

Cobertura efectiva

Por otro lado, la cobertura efectiva de viviendas levantadas en la Multipropósito 2020, se muestra a continuación:

Tabla 5. Distribución de la cobertura efectiva de viviendas por estrato socioeconómico

Formulario	Dominio	Estrato Socioeconómico			Total
		Alto (1)	Medio (2)	Bajo (3)	
Formulario 1	Urbano	717	939	587	2.243
	Rural	1.314	1.444	835	3.593
Formulario 2	Urbano	723	930	587	2.240
	Rural	1.296	1.439	832	3.567
Formulario común	Urbano	1.440	1.869	1.174	4.483
	Rural	2.610	2.883	1.667	7.160
Total viviendas	Nacional	4.050	4.752	2.841	11.643





Cálculo de los factores de expansión

El procedimiento de ponderación general para la ENEMDU requiere de dos etapas.

- La **primera etapa** consiste en calcular una ponderación de UPM. De ser necesario, se aplica un ajuste por no respuesta a nivel de UPM.
- La **segunda etapa** consiste en calcular una ponderación de viviendas dentro de cada UPM. De ser necesario se aplica un ajuste por no respuesta a nivel de vivienda.

La “falta de respuesta” se produce cuando no se llega a obtener respuesta de algunas unidades de la muestra. Resulta útil pensar en la población de la muestra dividida en dos conjuntos, el primero, formado por todas las unidades de la muestra de las que se ha obtenido respuesta; y el segundo, por todas las unidades de la muestra de las que no ha podido obtenerse respuesta.

La tasa de falta de respuesta puede estimarse con exactitud si se llevan a cabo recuentos de todos los elementos que cumplen los requisitos que se incluyen en la muestra. La tasa de respuesta en una encuesta se define como el coeficiente entre el número de cuestionarios completados por las unidades y el número total de unidades de la muestra que cumplen los requisitos. La falta de respuesta puede deberse a la ausencia del domicilio de las personas seleccionadas, a los cambios en la condición de ocupación de la vivienda, a la negación de dichas personas a participar o a su incapacidad para responder a las preguntas, entre otras. Otra posible causa de la falta de respuesta puede ser la imposibilidad de llevar a cabo la encuesta en determinadas zonas por razones climatológicas, dificultades del terreno o cuestiones de seguridad.

El factor de expansión final, es el resultado de un procedimiento que involucra el cálculo del factor de expansión de diseño, el ajuste por cobertura, la calibración y el recorte de los pesos de muestreo extremos, los cuales se describen a continuación.

Esquema de ponderación de Encuesta Multipropósito

Como se explica en la Sección anterior, la encuesta Multipropósito cuenta de tres bases de publicación razón por la cual se generan tres factores de expansión cuyo procedimiento es el mismo y se detalla a continuación.

Factor de expansión teórico

Probabilidad de inclusión de primera etapa (UPM)

Las UPM fueron seleccionadas con igual probabilidad en los estratos de muestreo h , por tanto, la probabilidad de inclusión de la primera etapa (UPM) es calculada a partir del siguiente algoritmo:





$$\pi_i = \frac{n_{i_h}}{N_{i_h}}$$

Donde:

n_{i_h} = es el número de UPM investigadas en el estrato h en la muestra.

N_{i_h} = es el número de UPM del estrato h en la población.

Por tanto, la ponderación básica de primera etapa para la i -ésima UPM muestreada en el estrato h es la inversa de la probabilidad de inclusión de la primera etapa y queda definida por:

$$w_i = \frac{1}{\pi_i}$$

Probabilidad de inclusión de segunda etapa (viviendas)

Las viviendas fueron escogidas con igual probabilidad de selección en cada UPM levantada en la primera etapa, por tanto, la probabilidad de inclusión de la segunda etapa (vivienda) es calculada a partir del siguiente algoritmo:

$$\pi_{j|i} = \frac{n_{j_i}}{N_{j_i}}$$

Donde:

n_{j_i} = es el número de viviendas investigadas en la UPM i en la muestra.

N_{j_i} = es el número de viviendas de la UPM i en la población.

Por tanto, la ponderación básica de segunda etapa para la j -ésima vivienda en la i -ésima UPM viene dada por:

$$w_j = \frac{1}{\pi_{j|i}}$$

Factor de expansión teórico o de diseño

El factor de expansión teórico o de diseño es el inverso de la multiplicación de las probabilidades de inclusión de la primera y de la segunda etapa, tal como se expresa a continuación:

$$w_k = \frac{1}{\pi_i * \pi_{j|i}}$$





Factor de expansión ajustado por cobertura

Ajuste por cambio de ocupación (no elegibilidad)

Si las viviendas seleccionadas presentan cambio en su estado de ocupación (viviendas temporales, desocupadas, en construcción, destruidas, convertidas en negocio), se realiza el siguiente ajuste:

$$a_1 = 1 - \frac{\# \text{ de viviendas no elegibles}}{\# \text{ de viviendas seleccionadas}}$$

Por lo tanto, el factor de expansión ajustado por cambio de ocupación es:

$$w_{k1}^{ne} = w_k * a_1,$$

donde:

- w_{k1}^{ne} = es el factor de expansión ajustado por no elegibilidad.
- w_k = es el factor de expansión teórico.
- a_1 = ajuste por cambio de ocupación (no elegibles).

Ajuste por nadie en casa

El siguiente ajuste consiste en redistribuir los pesos de las unidades secundarias de muestreo con elegibilidad desconocida (nadie en casa) utilizando la siguiente expresión:

$$a_2 = \frac{\# \text{ de viviendas investigadas en la UMPs}}{\# \text{ de viviendas completas en la UMPs}}$$

Por lo tanto, el factor de expansión ajustado por cambio de ocupación es:

$$w_{k2}^{ed} = w_{k1}^{ne} * a_2,$$

donde:

- w_{k2}^{ed} = es el factor de expansión ajustado por elegibilidad desconocida.
- w_{k1}^{ne} = es el factor de expansión ajustado por no elegibilidad.
- a_2 = ajuste por nadie en casa (elegibilidad desconocida).

Cabe recalcar que este ajuste se realiza únicamente en las UPM que cuentan con al menos una vivienda con elegibilidad desconocida.

Ajuste por rechazo (no respondientes)

El último ajuste en los pesos de muestreo se realiza si el resultado de la entrevista de las viviendas, al momento de realizar la investigación, fue rechazo. Si este es el caso, el ajuste a realizar es:





$$a_3 = \frac{\# \text{ de viviendas investigadas en la UMPs}}{\# \text{ de viviendas complatas en la UMPs}}$$

Cabe recalcar, que si una vivienda fue rechazado esta sigue siendo parte del conjunto de viviendas elegibles. Además, este ajuste no se realiza si $a_2 \neq 1$.

Por tanto, el factor de expansión ajustado por cobertura viene dado por la siguiente multiplicación:

$$w_k^{(*)} = w_{k2}^{ed} * a_3$$

donde:

- $w_k^{(*)}$ = es el factor de expansion ajustado por elegibilidad conocida.
- w_{k2}^{ed} = es el factor de expansión ajustado por elegibilidad desconocida.
- a_3 = ajuste por rechazo (no respondientes).

Calibración de los factores de expansión

La calibración de los factores de expansión (Deville J.C., Särndal C.E. y Sautory O., 1993) es un ajuste que se realiza a los ponderadores con el propósito de que las estimaciones de algunas variables de control reproduzcan con exactitud los totales poblacionales de dichas variables.

Cuando los estudios por muestreo están afectados por la ausencia de respuesta, es deseable tener las siguientes propiedades en la estructura inferencial que sustenta el muestreo:

1. Sesgo pequeño o nulo.
2. Errores estándares pequeños.
3. Un sistema de ponderación que reproduzca la información auxiliar disponible.
4. Un sistema de ponderación que sea eficiente al momento de estimar cualquier característica de interés en un estudio multipropósito.

Heredia (2010) manifiesta que para la calibración de los factores de expansión es necesario tomar en cuenta la siguiente información:

- Considerar una fuente de información auxiliar que se utiliza como "población", que puede ser censos, registros administrativos u otras encuestas; en las cuales se conocen los totales por variables o características que se desee estudiar.
- Otra fuente de información es la "muestra", de la cual procederán los estadísticos que infieren a los parámetros poblacionales.
- Identificar las variables de interés.
- Estimación de los totales de las variables de interés de la muestra.

El objetivo de la calibración es obtener un nuevo sistema de factores de expansión w_k que se encuentren cerca de los ponderadores de diseño d_k , de tal forma que cuando los ponderadores sean usados para estimar los totales de las variables





auxiliares, dichos totales sean reproducidos con exactitud, de manera que los nuevos factores conserven cualquier propiedad buena de estimación de los pesos básicos.

Estimador de calibración

El estimador de calibración se define de la siguiente manera:

Considere una función de distancia G con argumentos $x = w_k/d_k$ con las siguientes propiedades:

- G es positiva y estrictamente convexa,
- $G(1) = G'(1) = 0$, y
- $G''(1) = 1$.

Bajo esta definición, $G(w_k/d_k)$ mide la distancia de los factores de expansión originales d_k a los nuevos factores de expansión w_k , siendo $\sum_s d_k G(w_k/d_k)$ la medida de distancia para toda la muestra s . Por lo tanto, el problema de optimización es:

Minimizar $\sum_s d_k G(w_k/d_k) - \lambda'(\sum_s w_k x_k - \sum_U x_k)$, donde U hace referencia a la población, $x_k = (x_{k1}, \dots, x_{kj}, \dots, x_{kj})'$ es un vector de valores auxiliares y $\lambda = (\lambda_1, \dots, \lambda_j, \dots, \lambda_j)'$ es un J - vector de multiplicadores de Lagrange. Para calcular los nuevos factores de expansión, primero se debe determinar el valor de λ , el cual se obtiene resolviendo las ecuaciones de calibración:

$$\sum_s d_k F(x_k' \lambda) x_k = \sum_U x_k.$$

Luego, el estimador de calibración queda definido por:

$$\hat{t}_{yc} = \sum_s w_k y_k = \sum_s d_k F(x_k' \lambda) y_k$$

para los (y_k, x_k) datos observados en la muestra ($k \in s$) y una función de distancia G dada. Nótese que $w_k = d_k F(x_k' \lambda)$ es el nuevo factor de expansión calibrado. En este contexto, notaremos $g = F(x_k' \lambda)$.

Cabe mencionar que, en este proceso de construcción de factores de expansión para la ENEMDU, se calibra los pesos de muestreo ajustados por cobertura, por tanto, los ponderadores calibrados son calculados con la siguiente expresión:

$$w_k^{(**)} = w_k^{(*)} * g_k$$

Donde $w_k^{(**)}$ son los factores de expansión calibrados, mientras que $w_k^{(*)}$ son los pesos de muestreo ajustados por cobertura y g_k son los pesos de calibración.

Para la ENEMDU, se aplicó un esquema de calibración en el cual se presentan 8 celdas o post estratos de calibración (Ver Anexo 1), y se evaluó cada uno a través de criterios para validar la calibración propuestas por Silva (2004) (Ver Anexo 2).

Validación de la calibración de los factores de expansión





Silva (2004) propone 7 medidas para evaluar la calidad de la calibración de los factores de expansión, las cuales se detallan a continuación:

- Error relativo promedio sobre las variables auxiliares

$$M1 = \frac{1}{p} \sum_{j=1}^p \frac{|\hat{t}_{xc} - t_x|}{t_x}.$$

- Coeficiente de variación HT relativo promedio

$$M2 = \frac{1}{p} \sum_{j=1}^p \frac{(Var(\hat{t}_{x\pi}))^{1/2}}{t_x}.$$

- Proporción de pesos extremos (límite inferior)

$$M3 = \frac{1}{n} \sum_{k \in S} I(g_k < L).$$

- Proporción de pesos extremos (límite superior)

$$M4 = \frac{1}{n} \sum_{k \in S} I(g_k > U).$$

- Coeficiente de variación de los g_k

$$M5 = \frac{\sigma(g)}{\bar{g}}.$$

- Distancia entre los pesos de calibración y los pesos originales

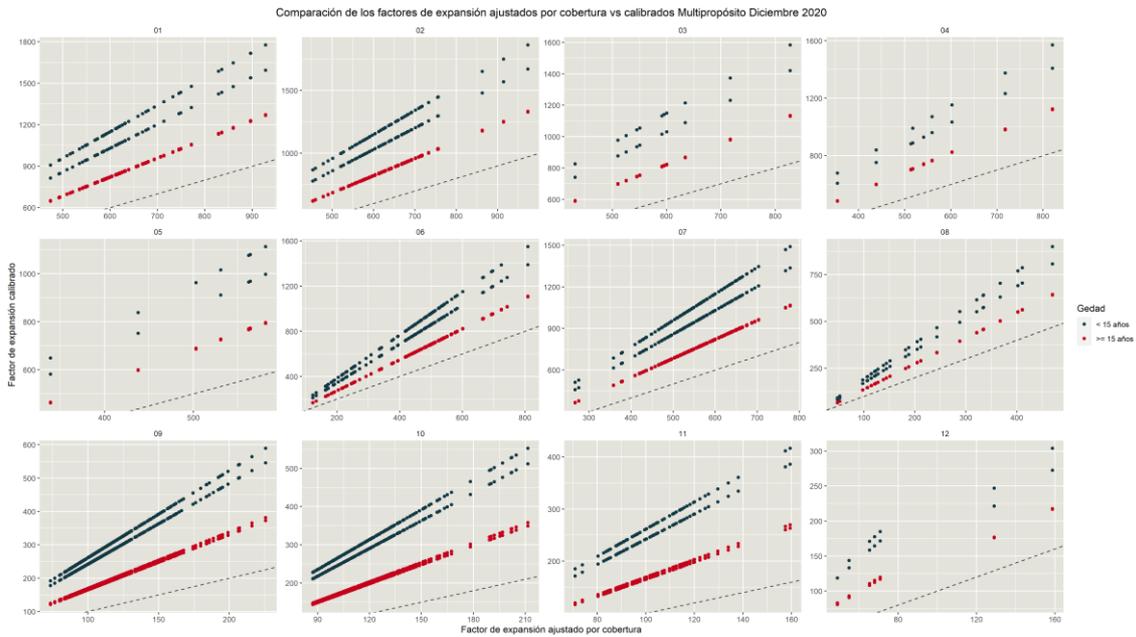
$$M6 = \frac{1}{n} \sum_{k \in S} \frac{(w_k - d_k)^2}{d_k} = \frac{1}{n} \sum_{k \in S} d_k (g_k - 1)^2.$$

- Eficiencia de los estimadores de calibración sobre la estimación del diseño de muestreo

$$M7 = \frac{1}{J} \sum_{j=1}^J \frac{\hat{V}_g(\hat{T}_{yjc})}{\hat{V}(\hat{T}_{y_j})}.$$

Gráfico 1. Comparación de los factores de expansión ajustados por cobertura y calibrados por dominio y grupo de edad para la base completa





Según el Gráfico 1, en la mayoría de dominios, los factores de expansión calibrados son más grandes que los ponderadores ajustados por cobertura, debido a la condición de reproducir con exactitud los totales poblacionales por sexo y grupo de edad (menores a 15 años, y mayores o iguales a 15 años) en los diferentes posts estratos de calibración. Los gráficos para el formulario 1 y 2 se encuentran en el Anexo 3.

Recorte de los factores de expansión extremos

Debido a los ajustes por cobertura y calibración, la variabilidad de los factores de expansión se ha visto incrementada. Potter (1990) señala que la variación extrema en los factores de expansión puede resultar en varianzas muestrales excesivamente grandes.

Para solventar este inconveniente se han desarrollado algunos procedimientos para limitar o reducir el tamaño de los factores de expansión extremos, algunos de los cuales están descritos en Potter (1990, 1993). Para la ENEMDU se utiliza el método de recorte y redistribución de factores de expansión más común, descrito en Valliant (2013), el cual propone el siguiente método iterativo:

1. Fijar una cota inferior y superior para los factores de expansión.
2. Todo factor de expansión mayor a la cota superior (o menor a la cota inferior) se fija en el valor de la cota, definiendo:

$$3. \quad fexp_{trim} = \begin{cases} U & \text{si } w_k^{(**)} \geq U, \\ w_k^{(**)} & \text{si } L < w_k^{(**)} < U, \\ L & \text{si } w_k^{(**)} \leq L. \end{cases}$$

4. Donde $fexp_{trim}$ es el conjunto de los factores de expansión recortados.
5. Calcular $K = \sum_k |fexp - fexp_{trim}|$.
6. Distribuir K equitativamente entre todos los pesos no recortados.





7. Repetir los pasos (1)-(4) hasta que $K = 0$.
Para la encuesta Multipropósito se evaluó el procedimiento de recorte y redistribución a diferentes umbrales (3, 3.5, 4, 4.5 y 5 veces la mediana), sin presentar necesidad de recorte, razón por la cual este procedimiento no se aplicó para el cálculo de los factores de expansión.

Estimación de características de la población

Una vez construidos los factores de expansión, se calcula los estimadores provenientes de la ENEMDU, se utiliza el estimador de Horvitz-Thompson, el cual sirve para estimar el valor total de una característica determinada (tasa de empleo adecuado, desempleo, subempleo, etc.). Está dado por (Carl-Erik Särndal, 1992).

$$\hat{Y}_{mHT} = \sum_{i=1}^{C_m} \sum_{j=1}^{v_i} P_T^j y_{ij}$$

Donde:

- \hat{Y}_{mHT} = estimador insesgado para el total Y_k de una variable y en el dominio de estudio m
- C_m = número de conglomerados investigadas en el dominio de estudio m .
- v_i = número de viviendas investigadas en el conglomerado i del dominio de estudio m
- y_{ij} = valor de la variable y para la vivienda j del conglomerado i del dominio de estudio m

Estimación de errores

Una vez realizada la estimación respectiva para la variable de interés a nivel de dominio de estudio, el error de muestreo es calculado a partir de la estimación de la varianza del estimador del total \hat{Y}_{mHT} . Para calcular adecuadamente los errores de muestreo de cada estimador, se debe tomar en cuenta los diferentes aspectos del diseño muestral, es decir, las dos etapas de muestreo, la estratificación presente en los dominios de estudio y los procesos de selección en cada una de las etapas.

Con todos estos elementos, el coeficiente de variación para el estimador \hat{Y}_{kHT} viene dado por la siguiente expresión:

$$CV(\hat{Y}_{kHT}) = \frac{\sqrt{\hat{V}_{2st}(\hat{Y}_{kHT})}}{\hat{Y}_{kHT}}$$

Donde:

- $\hat{V}_{2st}(\hat{Y}_{kHT})$ estimación de la varianza de segunda etapa del estimador del total del individuo k
- \hat{Y}_{kHT} estimador del total de la característica de interés del individuo k

Un estimador insesgado para la varianza está dado por:





$$\hat{V}_{2st}(\hat{t}_\pi) = \sum \sum_{s_i} \check{\Delta}_{Iij} \frac{\hat{t}_{i\pi}}{\pi_{Ii}} \frac{\hat{t}_{j\pi}}{\pi_{Ij}} + \sum_{s_i} \frac{\hat{V}_i}{\pi_{Ii}}$$

En el cual el \hat{V}_i apropiado es:

$$\hat{V}_i = \sum \sum_{s_{Iii}} \check{\Delta}_{IIqr|i} \frac{\hat{t}_{iq\pi}}{\pi_{IIq|i}} \frac{\hat{t}_{ir\pi}}{\pi_{IIr|i}}$$

Donde:

- i, j = Índice que recorre las UPM i, j en el dominio de estudio m.
- q, r = Índices que recorren las viviendas de la UPM i en el dominio de estudio m.
- π_{Ii} = Probabilidad de selección de la Etapa I para la i – ésima UPM en el dominio de estudio m.
- $\pi_{IIq|i}$ = Probabilidad de selección de la Etapa II para la q – ésima vivienda, dada la i – ésima UPM.
- $\check{\Delta}_{Iij}$ = Cantidad Δ expandida asociada a las UPM i, j.
- $\check{\Delta}_{IIqr|i}$ = Cantidad Δ expandida asociada a las viviendas q, r dada la selección de la i – ésima UPM.

Métodos de estimación de errores para diseños muestrales complejos

Aunque la selección del diseño de muestreo y el estimador sean de libre elección para los investigadores, no lo es el cálculo de las medidas de confiabilidad y precisión. Dado que la base científica sobre la cual descansa el muestreo es la inferencia estadística, se deben respetar las normas básicas para la asignación y posterior cálculo del margen de error, que constituye una medida unificada del error total de muestreo el cual cuantifica la incertidumbre acerca de las estimaciones en una encuesta.

Los métodos de estimación de los errores muestrales pueden clasificarse en cuatro categorías:

- a) Métodos exactos
- b) Métodos del último conglomerado
- c) Aproximaciones por linealización
- d) Técnicas de replicación

Para la descripción de los métodos se ha tomado como referencia los textos de Kish y Frankel (1974), Wolter (1985) y Lehtonen y Pahkinen (1995), que se encuentran descritos en el documento “**ENEMDU: Cálculo de errores estándar y declaración de muestras complejas**”, donde se realiza una breve descripción de los métodos convencionales para estimar varianzas o errores muestrales para estimaciones basados en muestreo complejo, que es una característica de la ENEMDU.

A continuación, se describirá las principales características de cada uno de los métodos de estimación de errores para el muestreo complejo:

- Los métodos exactos pueden ser utilizados para estimar totales, medias, tamaños y proporciones.





- La linealización de Taylor debe ser utilizada para estimar parámetros no lineales como razones, medias dentro de dominios, cuartiles o funciones de distribución.
- La técnica del último conglomerado junto con la linealización de Taylor puede ser utilizada para estimar la varianza de los indicadores de interés de las encuestas dirigidas a hogares que tengan diseños muestrales complejos. Esta es la técnica que por defecto utiliza el software SPSS.
- Las técnicas de replicación pueden ser usadas para estimar eficientemente todos los parámetros de interés, sin importar su forma funcional.
- La comparación general entre los métodos de linealización y replicación, es que no generan resultados idénticos del error de muestreo, pero hay que señalar que existen estudios (Kish y Frankel, 1974) que concluyen que las diferencias presentadas no son significativas cuando se trata de grandes muestras.

El INEC utiliza para la estimación de los parámetros de interés y sus correspondientes errores de muestreo diversos programas estadísticos tales como SPSS, Stata y R. En virtud de las características de cada uno de los métodos, es la técnica del último conglomerado en combinación con la linealización de Taylor, la cual induce a una muy buena aproximación del error muestral sobre los indicadores más importantes de las encuestas dirigidas a hogares, además de su facilidad de cálculo y replica. En este sentido, será esta la técnica la utilizada para la estimación de los errores muestrales en la Encuesta Multipropósito.

Las variables requeridas para declarar el diseño muestral en los programas estadísticos (SPSS, Stata y R) y ejecutar el cálculo de los errores de muestreo son presentadas en la Tabla 6, donde se describe las etiquetas de las variables identificadoras de las UPM, estratos y factores de expansión.

Tabla 6. Variables requeridas para declaración del diseño muestral – ENEMDU

Característica	Variable	Descripción
UPM	Upm	Agrupación de viviendas ocupadas en un número entre 30 a 60, próximas entre sí y con límites definidos.
Estratos	estrato	Identificación de estrato muestral (aproximación clasificación socio-económica)
Ponderación	Fexp	Factor de expansión calculado y ajustado (no cobertura)

Es importante indicar que los estratos de muestreo están definidos por el cruce entre Provincia (25 grupos) + Área (2 grupos) + estrato socioeconómico de la UPM (3 grupos). Además, las UPM deben tener identificadores únicos dentro de cada estrato y a través del tiempo. Por último, los hogares deben estar unívocamente identificados, así como su pertenencia a las UPM, a los estratos de muestreo y a las rondas del panel correspondiente.





Referencias

Cox, B.G. and McGrath, D.S. (1981), "An Examination of the Effect of Sample Weight Truncation on the Mean Square Error of Survey Estimates," presented at Biometrics Society ENAR meeting, Richmond, VA. March 1981.

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

Gutiérrez, A. (2018). *Revisión del diseño de muestreo y esquema de análisis de la Encuesta Nacional de Empleo, Desempleo y Subempleo-ENEMDU*. Misión de Asistencia Técnica. Quito.

Heredia, O. (2010). *El problema de la Calibración*. Aguascalientes: Dirección de diseño y marcos estadísticos - Instituto Nacional de Estadística y Geografía – INEGI.

Hidiroglu, M., Särndal, C., & Binder, D. (1995). Weighting and Estimation in Business Surveys. En Cox, Binder, Chinnappa, Christianson, Colledge, & Kott, *Business Survey Methods* (págs. 477-502). Jhon Wiley & Sons.

Kish, L. (1972). *Muestreo de Encuestas*. México: Trillas.

Potter F.J. (1990). A study of procedures to identify and trim extreme sampling weights. *American Statistical Association*, pp 225–230.

Särndal, C.E. y Swensson, B. y. W. J. (2003). *Model Assisted Survey Sampling*. Springer Series in Statistics. Springer: New York.

Silva A. (2004). *Calibration Estimation: When and Why, How much and How*. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística - IBGE.

Valliant R., Dever J.A. y Kreuter F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer International Publishing.





Anexos

Anexo 1: Esquema de calibración propuesto

El post estrato al que pertenece cada observación se puede identificar mediante un id de 5 dígitos, por ejemplo "00_1_1_1", que hacen referencia a los cruces utilizados para definir cada post estrato de acuerdo al siguiente detalle:

Dígito	Descripción y valores posibles
00	Dominio geográfico: nacional
1	Área: urbana (1) o rural (2)
1	Sexo: hombre (1) mujer (2)
1	Grupo de edad: menor a 15 años (1) 15 o más años (2)

En el ejemplo citado anteriormente, el post estrato "00_1_1_1" identifica a los individuos hombres menores a 15 años de la parte urbana a nivel nacional.

id_calib	Población
00_1_1_1	1.588.848
00_1_1_2	913.097
00_1_2_1	1.699.436
00_1_2_2	901.595
00_2_1_1	4.222.432
00_2_1_2	1.905.058
00_2_2_1	4.508.222
00_2_2_2	1.892.002

Anexo 2: Medidas para evaluar el esquema de calibración planteado (Esquema 1)

Medida 1:

Dominio	Formulario 1	Formulario 2	Formulario común
00_1_1_1	0	1,47E-16	4,40E-16
00_1_1_2	2,21E-16	0	6,62E-16
00_1_2_1	2,33E-15	9,73E-15	7,12E-15
00_1_2_2	0	0	1,03E-15
00_2_1_1	9,76E-11	4,11E-10	3,05E-10
00_2_1_2	9,78E-16	9,78E-16	4,89E-16
00_2_2_1	1,35E-09	1,06E-08	4,93E-09
00_2_2_2	3,69E-16	3,70E-16	1,11E-15





Medida 2:

	Formulario 1	Formulario 2	Formulario común
Nacional	0,0268	0,0271	0,0204

Medida 3:

Formulario	Dominio	L	N	Esquema 1
Formulario 1	00_1_1_1	1,73	902	115
	00_1_1_2	1,37	2.959	0
	00_1_2_1	1,88	888	0
	00_1_2_2	1,34	3.213	3.149
	00_2_1_1	2,35	1.642	0
	00_2_1_2	1,69	4.754	0
	00_2_2_1	2,52	1.508	0
	00_2_2_2	1,63	4.907	0
Formulario 2	00_1_1_1	1,67	934	0
	00_1_1_2	1,35	2.946	2.920
	00_1_2_1	1,92	864	0
	00_1_2_2	1,36	3.156	75
	00_2_1_1	2,44	1.573	0
	00_2_1_2	1,67	4.807	0
	00_2_2_1	2,67	1.425	0
	00_2_2_2	1,64	4.835	0
Formulario común	00_1_1_1	1,72	1.836	0
	00_1_1_2	1,37	5.905	0
	00_1_2_1	1,91	1.752	0
	00_1_2_2	1,36	6.369	0
	00_2_1_1	2,42	3.215	0
	00_2_1_2	1,70	9.561	0
	00_2_2_1	2,61	2.933	0
	00_2_2_2	1,65	9.742	0





Medida 4:

Formulario	Dominio	U	N	Esquema 1
Formulario 1	00_1_1_1	1,73	902	80
	00_1_1_2	1,37	2.959	2.877
	00_1_2_1	1,88	888	888
	00_1_2_2	1,34	3.213	0
	00_2_1_1	2,35	1.642	1.642
	00_2_1_2	1,69	4.754	4.754
	00_2_2_1	2,52	1.508	1.508
	00_2_2_2	1,63	4.907	4.907
Formulario 2	00_1_1_1	1,67	934	873
	00_1_1_2	1,35	2.946	0
	00_1_2_1	1,92	864	864
	00_1_2_2	1,36	3.156	81
	00_2_1_1	2,44	1.573	1.573
	00_2_1_2	1,67	4.807	4.807
	00_2_2_1	2,67	1.425	1.425
	00_2_2_2	1,64	4.835	4.835
Formulario común	00_1_1_1	1,72	1.836	1.836
	00_1_1_2	1,37	5.905	5.905
	00_1_2_1	1,91	1.752	1.752
	00_1_2_2	1,36	6.369	6.369
	00_2_1_1	2,42	3.215	3.215
	00_2_1_2	1,69	9.561	9.561
	00_2_2_1	2,61	2.933	2.933
	00_2_2_2	1,65	9.742	9.742

Medida 5:

Dominio	Formulario 1		Formulario 2		Formulario común	
	Media peso g	CV peso g	Media peso g	CV peso g	Media peso g	CV peso g
00_1_1_1	1,73	1,50E-16	1,67	1,31E-17	1,72	4,83E-17
00_1_1_2	1,37	1,07E-16	1,35	1,21E-16	1,37	2,57E-16
00_1_2_1	1,88	1,31E-16	1,92	1,97E-17	1,91	9,70E-17
00_1_2_2	1,34	9,96E-17	1,36	1,74E-17	1,36	6,53E-17
00_2_1_1	2,35	6,53E-17	2,44	6,89E-17	2,42	1,26E-16
00_2_1_2	1,69	5,54E-17	1,67	5,57E-17	1,69	6,58E-17
00_2_2_1	2,52	3,63E-17	2,67	9,69E-17	2,61	5,97E-17
00_2_2_2	1,63	5,46E-17	1,64	9,72E-18	1,65	1,32E-16





Medida 6:

Dominio	Formulario 1	Formulario 2	Formulario común
00_1_1_1	542,32	454,33	258,53
00_1_1_2	142,13	132,49	71,73
00_1_2_1	783,08	866,61	423,53
00_1_2_2	119,69	138,31	68,91
00_2_1_1	430,97	492,55	236,51
00_2_1_2	111,64	105,80	56,33
00_2_2_1	547,27	661,29	305,80
00_2_2_2	92,94	98,36	49,74

Anexo 3:

Gráfico 2. Comparación de los factores de expansión ajustados por cobertura y calibrados por dominio y grupo de edad para la base del formulario 1

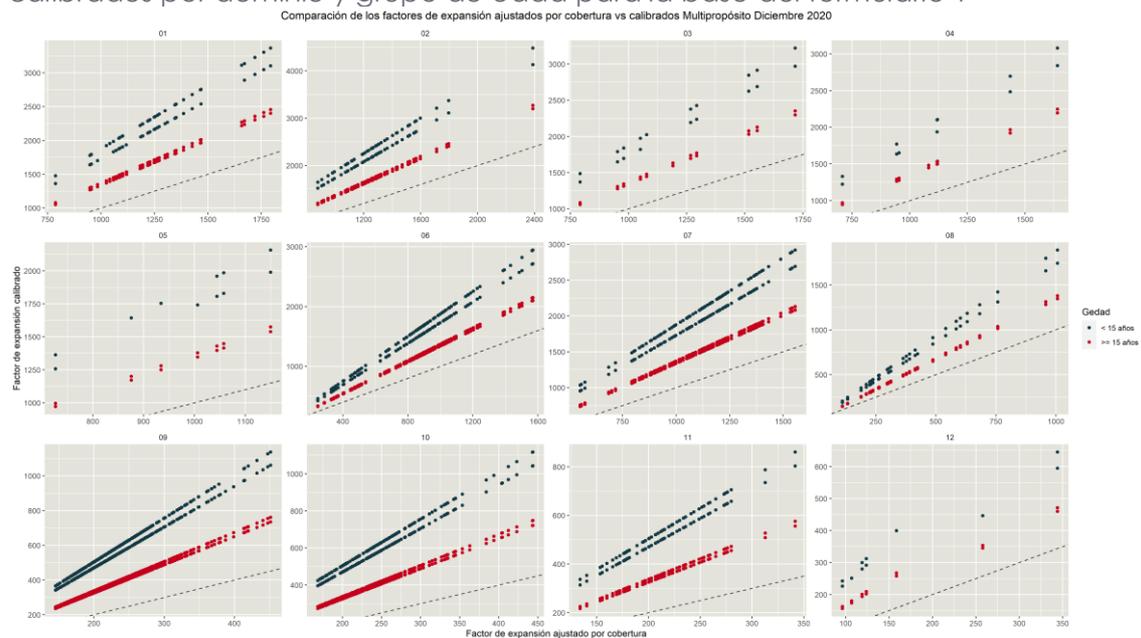
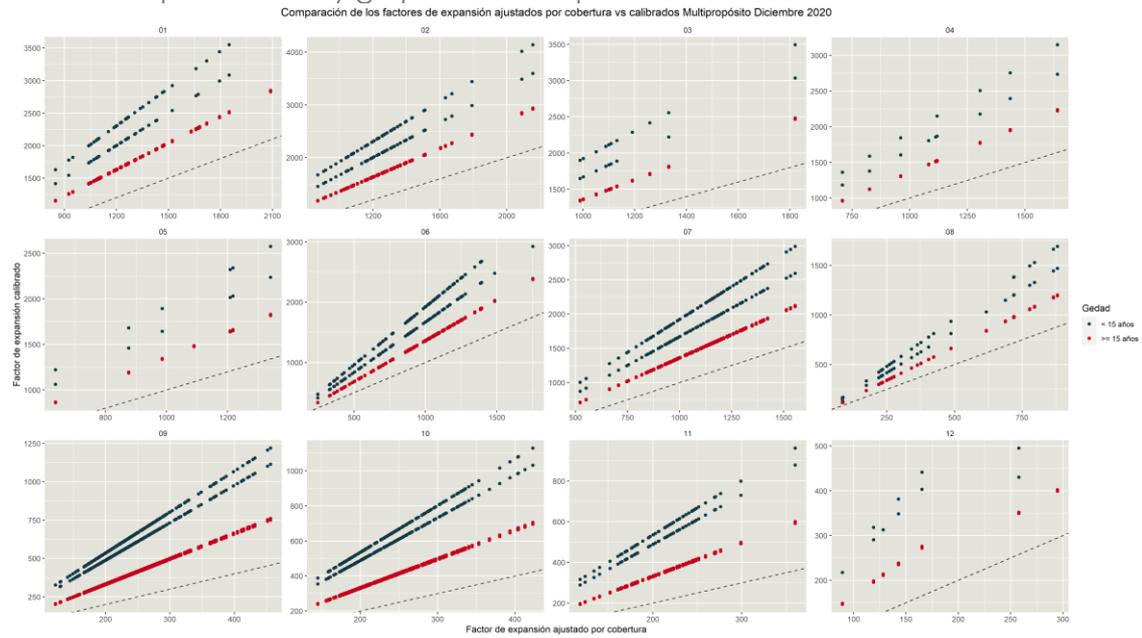




Gráfico 3. Comparación de los factores de expansión ajustados por cobertura y calibrados por dominio y grupo de edad para la base del formulario 2





Buenas cifras,
mejores vidas



@ecuadorencifras



@ecuadorencifras



@InecEcuador



t.me/euadorencifras



INEC/Ecuador



INECEcuador



INEC Ecuador

