República del Ecuador

Diseño Muestral de la ENEMDU trimestral longitudinal

Julio-Septiembre 2022 y Julio-Septiembre 2023

INEC Buenas cifras, mejores vidas

Dirección

Dirección de Infraestructura Estadística y Muestreo (DINEM)

Elaborado por:

Pablo Peñafiel
Félix Encalada
Laura Tierra
Cristina Valladolid
William Constante

Revisado por:

Christian Garces

Aprobado por:

Christian Garces
Julio Muñoz

Quito -Ecuador, 2023

Contenido

Lista de tablas

Lista de gráficos

Introducción

El Instituto Nacional de Estadística y Censos (INEC) ejecuta la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) desde 1993. La ENEMDU es una encuesta de aplicación continua, la cual se realiza los doce meses del año. La información generada de la encuesta sirve de insumo al gobierno para la planificación del desarrollo nacional y su correspondiente monitoreo y evaluación, así como al sector privado y sociedad civil en general para su conocimiento y toma de decisiones.

Así, desde 2018, se plantean algunas mejoras puntuales al diseño muestral de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). La técnica de muestreo de la ENEMDU es similar a la utilizada en años anteriores y corresponde a un muestreo probabilístico en dos etapas, con estratificación geográfica por dominios de estudio y área urbana-rural. Los estimadores asociados al diseño se calibran por una proyección de población calculada según métodos demográficos. Sin embargo, el nuevo diseño contempla la afinación del marco de muestro que incluye la mejora en los siguientes aspectos:

- Equilibrio de las Unidades Primarias de Muestreo (UPM).
- Estratificación acorde a las UPM equilibradas.
- Optimización de la dispersión de la muestra.

Los cambios implementados tienen el objetivo de mejorar la precisión de los estimadores y oportunidad de la información.

Por otro lado, en el segundo semestre del año 2020, la Comisión Económica para América Latina y el Caribe (CEPAL), mediante un trabajo conjunto con el INEC, realizaron un rediseño de la ENEMDU en lo referente al cálculo del tamaño y selección de la muestra, esquema de rotación de paneles y cálculo de los factores de expansión. Estos aspectos se encuentran detallados en el informe de misión de asistencia técnica que fue entregado por CEPAL al INEC en febrero de 2021, el cual se titula "Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024)".
En este documento se expone de forma clara y detallada el proceso de diseño muestral de la ENEMDU trimestral longitudinal, referente a la formación de la base longitudinal, agregación de encuestas, esquema de rotación de paneles, cálculo de los factores de expansión longitudinales, estimaciones de características y errores muéstrales, etc.

1. Antecedentes

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 varianza.

El Instituto Nacional de Estadísticas y Censos (INEC) genera bases longitudinales puntuales de la Encuesta de Empleo, Desempleo y Subempleo (ENEMDU), mediante la Dirección de Infraestructura Estadísticas y Muestreo (DINEM) a través del proceso matching que es una herramienta que nos permite validar la información levantada en campo por periodos (1, 2, 3 y 4). Con esta validación obtenemos novedades que se reportan a la Dirección de cartografía y Operaciones de Campo (DICA) y Coordinaciones Zonales con la finalidad de que realicen los seguimientos y validaciones respectivas. Las bases longitudinales puntuales consisten en el emparejamiento de viviendas y personas que son miembros del hogar investigado en dos periodos de tiempo diferentes, de acuerdo al esquema de rotación de paneles de la ENEMDU. La base longitudinal puntual permite el cálculo de indicadores de permanencia y transición de la Población en Edad de Trabajar-PET(15 años y más), con la

finalidad de conocer el movimiento de los ocupados, desocupados e inactivos dentro del Mercado Laboral Ecuatoriano.

2. Características del diseño muestral de la ENEMDU trimestral longitudinal

Universo de investigación

El universo de estudio de la ENEMDU son personas de 5 y más años de edad, residentes en las viviendas del Ecuador, exceptuando la población que reside en viviendas colectivas, viviendas flotantes y población indigente (sin techo).

Unidad de observación

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

Unidad de análisis

Para el caso de los indicadores laborales, la población de referencia son todas las personas mayores o iguales a 15 años.

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 ENEMDU corresponde a un muestreo probabilístico estratificado bietápico de elementos.

Paneles de rotación

El esquema de rotación de paneles para la ENEMDU 2021-2024 presenta una rotación tanto de UPM como de viviendas dentro de las UPM, según CEPAL (2021) esto permitirá que se afiance el conocimiento de la construcción de pesos de muestreo longitudinales. El esquema de rotación para el trimestre III 2022-2023 se presenta a continuación:

Tabla 1: Rotación de paneles que se traslapan en el trimestre III 2022-2023

ROTACIÓN DE PANELES ENEMDU													
Año	Trimestre	Mes											
		M1				M2				M3			
2022	T1	a11	b11	c21	d21	e11	f11	g21	h21	i11	j11	k21	l21
	T2	a11	b12	c21	d22	e11	f12	g21	h22	i11	j12	k21	l22
	T3	a12	b12	c12	d22	e12	f12	g12	h22	i12	j12	k12	l22
	T4	a12	b21	c12	d21	e12	f21	g12	h21	i12	j21	k12	l21
2023	T1	a21	b21	c21	d21	e21	f21	g21	h21	i21	j21	k21	l21
	T2	a21	b22	c21	d22	e21	f22	g21	h22	i21	j22	k21	l22
	T3	a22	b22	c31	d22	e22	f22	g31	h22	i22	j22	k31	l22
	T4	a22	b21	c31	d31	e22	f21	g31	h31	i22	j21	k31	l31

Dominios de estudio y representatividad:

a) **ENEMDU mensual:** La ENEMDU mensual tiene como sus dominios de diseño y representatividad Nacional, Urbano-Rural.

b) **ENEMDU trimestral:** La ENEMDU trimestral tiene como dominios de diseño y representatividad Nacional, Urbano-Rural y 5 ciudades principales (Quito, Guayaquil, Cuenca, Machala y Ambato).

c) **ENEMDU anual:** La ENEMDU anual tiene como dominios de diseño y representatividad Nacional, Urbano-Rural, 5 ciudades principales (Quito, Guayaquil, Cuenca, Machala y Ambato) y 24 provincias del Ecuador.

d) **ENEMDU longitudinal trimestral:** La ENEMDU longitudinal trimestral tiene como dominio de diseño y representatividad Nacional, Urbano-Rural.

e) **ENEMDU longitudinal anual:** La ENEMDU longitudinal anual tiene como dominio de diseño y representatividad Nacional, Urbano-Rural y 5 ciudades principales (Quito, Guayaquil, Cuenca, Machala y Ambato).

Dado los tamaños de muestra de las ENEMDU longitudinales (trimestrales o anuales) en función al porcentaje de traslape, fue necesario realizar un análisis de calidad y confiabilidad de las estimaciones obtenidas a partir de estas encuestas. Para tal fin, se calcularon estimaciones correspondientes a la matriz de transición laboral, ya sea para totales o proporciones, las cuales se sometieron a diferentes criterios estadísticos para verificar qué tan confiable se considera dicha estimación, tales como tamaños de muestra, grados de libertad, tipo de indicador, máximo error estándar tolerable y coeficiente de variación.

A través de la aplicación y análisis de estos criterios mencionados, se obtuvo como resultado que a partir de la ENEMDU longitudinal trimestral se tendría estimaciones fiables a nivel nacional, urbano y rural; mientras que a partir de la ENEMDU longitudinal anual se conseguirían estimaciones fiables a nivel nacional, urbano, rural y para las 5 ciudades principales (Quito, Guayaquil, Cuenca, Machala y Ambato). Para una mayor información, se recomienda al usuario revisar el documento "Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares" (INE, 2020).

3. Tamaño y selección de la muestra de la ENEMDU

Tamaño de la muestra

El tamaño de muestra por dominio de la ENEMDU fue calculado considerando los siguientes parámetros:

- **N:** número de personas en cada dominio en el Marco de Muestreo.
- **M:** número de UPM en cada dominio en el Marco de Muestreo.
- **R:** porcentaje de población económicamente activa (PEA) en cada dominio. Calculado a partir de la ENEMDU anual 2019.
- **B:** promedio de personas por hogar en cada dominio, calculado a partir de la ENEMDU anual 2019.
- **Rho (ρ):** coeficiente de correlación intraclase de la tasa de desempleo para cada uno de los dominios, valor que es calculado a partir de la información de la ENEMDU anual 2019.
- **P:** tasa de desempleo en cada dominio, estimado a partir de la ENEMDU anual 2019.
- **A:** amplitud del intervalo de confianza.
- **Delta (δ):** margen de error relativo asociado a la tasa de desempleo en cada dominio. Este valor se calcula como la mitad del ancho del intervalo (A) de confianza dividido para la tasa de desempleo -estimada a partir de la ENEMDU anual 2019- y este resultado elevado al cuadrado.
- **1-α:** nivel de confianza del 95%.

- **TNR:** tasa de no respuesta del 20%, valor calculado mediante la realización de un análisis histórico de la cobertura de la encuesta en los periodos 2018, 2019 y 2020.

Para cada uno de los cálculos y escenarios de tamaño de muestra, la confiabilidad estadística para la nueva ENEMDU 2021 – 2024 se fija en 95%, además, el número de viviendas seleccionadas dentro de cada una de las UPM se sigue manteniendo en 7.

Los parámetros utilizados para el cálculo de muestra en cada uno de los dominios para la encuesta se presentan en la Tabla 2:

Tabla 2. Parámetros utilizados para el cálculo del tamaño de muestra ENEMDU 2021-2024

Dominio	N	M	R	B	Rho	P	A
Azuay	450.652	2.643	54,90%	4,92	0,0440	2,25%	1,60%
Bolívar	224.270	1.279	51,30%	3,97	0,0170	1,31%	1,60%
Cañar	300.749	1.596	51,90%	4,33	0,0300	3,67%	1,60%
Carchi	192.573	1.288	48,60%	3,72	0,0440	4,60%	1,80%
Cotopaxi	514.752	2.807	58,10%	4,25	0,0190	1,83%	1,60%
Chimborazo	546.499	3.196	56,10%	4,52	0,0240	1,59%	1,60%
El Oro	428.880	2.862	48,60%	4,16	0,0200	4,35%	1,60%
Esmeraldas	627.896	3.392	40,00%	4,73	0,0700	10,17%	1,80%
Guayas	1.674.002	10.112	43,60%	5,51	0,0510	3,38%	1,60%
Imbabura	496.033	3.076	46,10%	4,26	0,0780	6,11%	1,60%
Loja	547.507	3.194	53,20%	4,54	0,0780	3,85%	1,60%
Los Ríos	898.647	5.459	44,30%	4,85	0,0410	2,89%	1,60%
Manabí	1.554.229	9.198	45,60%	5,40	0,0730	2,47%	1,60%
Morona Santiago	195.141	975	47,40%	4,54	0,0510	1,83%	1,60%
Napo	132.566	701	49,40%	4,36	0,0370	2,92%	1,80%
Pastaza	111.690	617	47,00%	4,05	0,0110	3,07%	1,80%
Pichincha	1.180.351	7.358	48,00%	4,78	0,0480	6,19%	1,60%
Tungurahua	399.561	2.525	60,10%	4,68	0,0130	1,15%	1,60%
Zamora Chinchipe	119.508	669	50,70%	4,13	0,0260	3,32%	1,80%
Galápagos	32.395	348	53,50%	3,22	0,0400	1,82%	1,80%
Sucumbíos	227.509	1.250	43,70%	4,21	0,0330	5,42%	1,80%
Orellana	161.472	982	46,90%	4,32	0,0420	2,69%	1,60%
Sto Domingo de los Tsáchilas	494.320	3.133	44,50%	4,21	0,0620	2,53%	1,60%
Santa Elena	386.346	2.025	40,50%	4,89	0,2000	3,20%	1,80%
Quito	1.950.476	13.811	46,40%	5,66	0,0260	9,12%	1,80%
Guayaquil	2.610.712	16.907	45,40%	6,71	0,0190	3,07%	1,00%
Cuenca	402.447	2.423	44,50%	6,08	0,0380	5,35%	1,80%
Machala	270.241	1.797	43,50%	6,38	0,0380	6,13%	1,80%
Ambato	192.509	1.376	49,50%	5,59	0,0420	4,79%	1,80%

Fuente: CEPAL. (2021). Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024).

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

La expresión matemática utilizada para el cálculo de tamaño de muestra de personas en cada dominio de diseño D es:

$$n_D \geq \frac{P(1-P)\ Deff}{\frac{\delta^2 P^2}{z_\alpha^2} + \frac{P(1-P)\ Deff}{N_D}} * \frac{1}{1-T_{NR}}$$

donde:

z = $percentil\ de\ la\ distribución\ normal\ estándar\ asociado\ al\ nivel\ de\ confianza\ 1-\alpha$
δ = $margen\ de\ error\ relativo\ máximo$
P = $estimación\ de\ la\ variable\ de\ diseño\ (tasa\ de\ desempleo)$
N_D = $tamaño\ de\ la\ población\ en\ cada\ dominio\ de\ diseño\ D$
T_{NR} = $tasa\ de\ no\ respuesta$

El efecto de diseño $Deff$, definido como una función de la correlación existente entre la variable de interés (desempleo) y la conformación de las UPM, está dado por la siguiente expresión:

$$Deff \approx 1 + (\bar{n} - 1) * \rho$$

Donde $\bar{n}$ es el número promedio de personas de la población económicamente activa (PEA) que serán encuestadas y ρ es la correlación intraclase entre el desempleo y la conformación de las UPM.

A su vez, $\bar{n}$ es calculada a través del siguiente algoritmo:

$$\bar{n} = 7 * r * b$$

Donde el número (7) se refiere al número de viviendas investigadas en cada UPM, r es el porcentaje de población económicamente activa (PEA) y b es el promedio de personas por hogar.
Por último, cabe señalar que la tasa de desempleo, la cual varía dependiendo del dominio de representatividad, el enfoque que se utilizó en cada dominio de diseño debería controlar el ancho del intervalo de confianza generado a partir de la encuesta. Es por esta razón que se fijó el error máximo relativo como función de la amplitud del intervalo de confianza A. Por lo tanto, se tiene que:

$$\delta = \left(\frac{A}{2}\right)^2$$

Tamaño de muestra de viviendas

El número de viviendas que deben ser seleccionadas estará determinado por la muestra de personas (n_D), número promedio de personas por vivienda (b) y el porcentaje de personas que presentan la característica de interés ($r = Proporción\ de\ la\ PEA$), de la siguiente forma:

$$n_{vD} = \frac{n_D}{r * b}$$

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:

$$n_{UPM_D} = \frac{n_{vD}}{Carga\ técnica\ operativa}$$

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 siete (7).

Considerando las restricciones presupuestarias presentes, y luego de aplicar los algoritmos de cálculo correspondientes, se obtiene un tamaño de muestra de 9.016 viviendas mensuales. Además, al fijar en 7 el número de viviendas levantadas por UPM, el tamaño de muestra es de 1.288 UPM a ser visitadas cada mes en la ENEMDU. Para el caso particular de la ENEMDU longitudinal cabe señalar que cuando el traslape de paneles es del 75%, 50% o 25% el tamaño de muestra por dominio de representatividad es el siguiente:

Tabla 3. Tamaños de muestra ENEMDU Longitudinal Trimestral 75% de traslape

Dominio	Viviendas del marco	UPM del marco	Muestra UPM	Muestra Viviendas
Urbano	2.715.812	79.647	2.142	14.994
Rural	1.002.506	27.352	756	5.292
Nacional	3.718.318	106.999	2.898	20.286

Fuente: CEPAL. (2021). Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024).

Tabla 4. Tamaños de muestra ENEMDU Longitudinal Trimestral 50% de traslape

Dominio	Viviendas del marco	UPM del marco	Muestra UPM	Muestra Viviendas
Urbano	2.715.812	79.647	1.428	9.996
Rural	1.002.506	27.352	504	3.528
Nacional	3.718.318	106.999	1.932	13.524

Fuente: CEPAL. (2021). Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024).

Tabla 5: Tamaños de muestra ENEMDU Longitudinal Trimestral 25% de traslape

Dominio	Viviendas del marco	UPM del marco/	Muestra UPM	Muestra Viviendas
Urbano	2.715.812	79.647	714	4.998
Rural	1.002.506	27.352	252	1.764
Nacional	3.718.318	106.999	966	6.762

Fuente: CEPAL. (2021). Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024).

Asignación de la muestra

La asignación del tamaño de muestra en cada uno de los 150 estratos de muestreo se llevó a cabo usando la asignación de Kish (Kish, 1987; Maligalig & Martinez, 2013) que permite distribuir en cada uno de los estratos un tamaño de muestra óptimo, a través del siguiente algoritmo:

$$n_h = n \cdot \frac{\sqrt{\frac{1}{H^2}\left(\frac{N_h}{N}\right)^2}}{\sum_{h=1}^{H}\sqrt{\frac{1}{H^2}\left(\frac{N_h}{N}\right)^2}}$$

donde:

n_h = $Tamaño\ de\ muestra\ para\ el\ estrato\ h$
N_h = $Número\ de\ UPM\ del\ estrato\ h$
N = $Número\ de\ UPM\ en\ el\ Marco\ de\ Muestreo.$
n = $Número\ de\ UPM\ seleccionadas\ en\ la\ muestra\ en\ la\ primera\ etapa\ de\ muestreo$
H = $Número\ de\ estratos\ en\ el\ dominio$

Tabla 6. Asignación longitudinal trimestral de la muestra de viviendas y UPM por territorio traslape del 75%

Subpoblación	Viviendas del marco	UPM del marco	Muestra UPM	Muestra Viviendas
Quito	473.957	13.811	306	2.142
Guayaquil	589.772	16.907	288	2.016
Cuenca	84.623	2.423	207	1.449
Machala	62.658	1.797	234	1.638
Ambato	47.817	1.376	189	1.323
Resto Sierra Urbano	592.918	17.565	432	3.024
Resto Costa Urbano	780.055	23.186	324	2.268
Amazonía Urbano	79.861	2.410	144	1.008
Sierra Rural	538.194	14.530	369	2.583
Costa Rural	360.130	9.862	198	1.386
Amazonía Rural	99.813	2.784	171	1.197
Región Insular	8.520	348	36	252
Total	**3.718.318**	**106.999**	**2.898**	**20.286**

Fuente: CEPAL. (2021). Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024).

Tabla 7. Asignación longitudinal trimestral de la muestra de viviendas y UPM por territorio traslape del 50%

Subpoblación	Viviendas del marco	UPM del marco	Muestra UPM	Muestra Viviendas
Quito	473.957	13.811	204	1.428
Guayaquil	589.772	16.907	192	1.344
Cuenca	84.623	2.423	138	966
Machala	62.658	1.797	156	1.092
Ambato	47.817	1.376	126	882
Resto Sierra Urbano	592.918	17.565	288	2.016
Resto Costa Urbano	780.055	23.186	216	1.512
Amazonía Urbano	79.861	2.410	96	672
Sierra Rural	538.194	14.530	246	1.722
Costa Rural	360.130	9.862	132	924
Amazonía Rural	99.813	2.784	114	798
Región Insular	8.520	348	24	168
Total	**3.718.318**	**106.999**	**1.932**	**13.524**

Fuente: CEPAL. (2021). Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024).

Tabla 8. Asignación longitudinal trimestral de la muestra de viviendas y UPM por territorio traslape del 25%

Subpoblación	Viviendas del marco	UPM del marco	Muestra UPM	Muestra Viviendas
Quito	473.957	13.811	102	714
Guayaquil	589.772	16.907	96	672
Cuenca	84.623	2.423	69	483
Machala	62.658	1.797	78	546
Ambato	47.817	1.376	63	441
Resto Sierra Urbano	592.918	17.565	144	1.008
Resto Costa Urbano	780.055	23.186	108	756
Amazonía Urbano	79.861	2.410	48	336
Sierra Rural	538.194	14.530	123	861
Costa Rural	360.130	9.862	66	462
Amazonía Rural	99.813	2.784	57	399
Región Insular	8.520	348	12	84
Total	**3.718.318**	**106.999**	**966**	**6.762**

Fuente: CEPAL. (2021). Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024).

En la Tabla 6, Tabla 7 y Tabla 8 se presenta la asignación de la muestra por subpoblación en la ENEMDU trimestral longitudinal, cabe recalcar que estos no son dominios de diseño de la encuesta trimestral - longitudinal, por cuanto la representatividad a nivel de las variables de mercado laboral agregado es **Nacional, Urbano-Rural.**

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.

4. Estructura Base Longitudinal

Cobertura Matching

El proceso matching o emparejamiento de la información consiste en cotejar los formularios de un mismo hogar, es decir, se juntan los datos de los mismos individuos encuestados en dos periodos de tiempo diferentes. La consideración principal para la construcción de la base longitudinal ENEMDU, consiste en el emparejamiento de los miembros del hogar que fueron investigados en la Encuesta Nacional de Empleo, Desempleo y Subempleo – ENEMDU, Trimestre III de 2022 al Trimestre III de 2023.

Tabla 9. Cobertura de viviendas por mes de levantamiento

Vivienda	Cobertura por mes TIII2022 - TIII2023		
	7	8	9
1	801	936	859
2	870	824	857
3	811	891	825
4	856	804	826
5	829	783	815
6	799	797	844
7	829	849	808

La clave única de emparejamiento matching que permite realizar este proceso por medios manuales o automáticos es el identificador de personas, se construye mediante la unión de las variables provincia, cantón, parroquia, conglomerado, vivienda, hogar y código de persona, con la finalidad de unir la información de los miembros del hogar que respondieron en los 2 periodos investigadas.

Etapas de emparejamiento

Etapa 1.- Se realiza el emparejamiento entre los dos periodos a través del identificador de personas, la mayoría de casos son coincidentes.

Etapa 2.- Para los casos no coincidentes entre los periodos se procede a realizar un emparejamiento manual caso por caso.

Una vez realizadas estas 2 etapas de emparejamiento, el resultado es la base matching, la misma que pasa a un proceso de validación.

Validación de la base matching

La validación de la base matching, se realiza comparando las variables nombre y apellidos completos, sexo, edad y cédula de identidad/ciudadanía. De este procedimiento es importante mencionar las respectivas validaciones de consistencia de la información que se realizan:

- Nombres y Apellidos: se presentan casos en que los nombres y/o apellidos cambian de posición entre un periodo a otro, la validación para confirmar si es la misma persona se realiza principalmente por medio del número de cédula de identidad o ciudadanía, edad, sexo y para algunos casos con la relación de parentesco.

- Escritura de Nombres y Apellidos: otro caso consiste en la escritura de los nombres y/o apellidos diferentes entre un periodo y otro, como por ejemplo: Jenny María en el periodo 1 y Maria Jeni en el periodo 2, sin embargo es la misma persona. Por tal razón esta persona forma parte de la base matching debido a que pertenece a los dos periodos.

- Sexo: se valida esta variable, considerando que en los 2 periodos, el sexo debe ser el mismo, es decir sexo masculino con código "1" y sexo femenino con código "2". Generalmente la inconsistencia de esta variable es por la digitación.

- Edad: tomando en cuenta la subjetividad de esta variable, donde el verdadero valor depende de varios factores externos, como el tipo de Informante sea Directo o Calificado, es una de las razones para casos no coincidentes en la edad, por lo cual se consideran diferencias en su valor por rangos de edad, los cuales se han establecido de acuerdo a análisis realizados. La diferencia de años permitida por rangos de edad se describe en la siguiente tabla:

Tabla 10 . Diferencia de años por rangos de edad

Rangos de edad	Diferencia en años
0 a 10	2
11 a 20	3
21 a 50	5
51 y más	10

Cédula de Identidad/ciudadanía: para la validación de esta variable se compara el número de cédula de identidad o ciudadanía entre los 2 periodos, en el cual existen casos en que los números son diferentes entre una ronda y otra, para la validación de estos casos se verifica con las variables nombres y apellidos completos, sexo, edad, los mismos se presentan por falta de información por parte del Informante o por error en la digitación.

Construcción de bases longitudinales.

Los insumos necesarios para la construcción de las bases longitudinales trimestrales son:

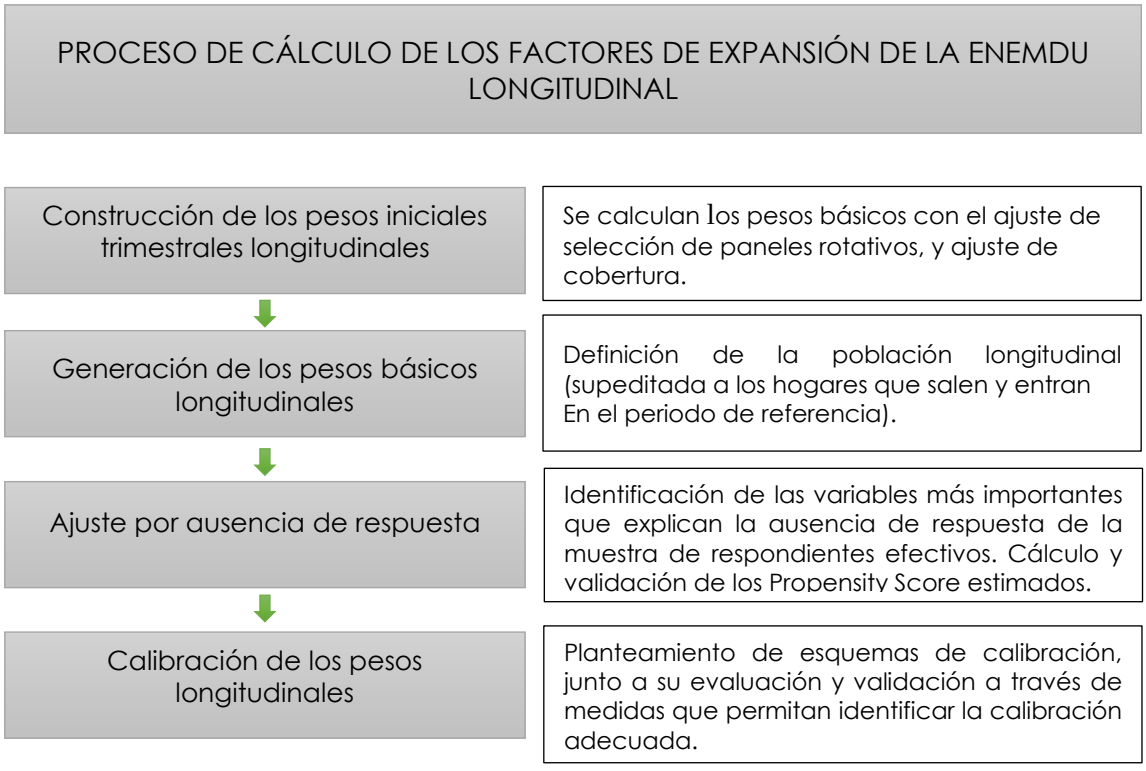
- Base matching julio2022_julio2023
- Base matching agosto2022_agosto2023
- Base matching septiembre2022_septiembre2023

Con las bases indicadas anteriormente se procede a la unión de las mismas para el posterior cálculo de los factores de expansión.

5. Calculo de los factores de expansión de la ENEMDU Longitudinal

El proceso de cálculo de los factores de expansión de la ENEMDU Longitudinal involucra varias actividades interrelacionadas, como la construcción de los pesos básicos de muestreo, generación de los pesos longitudinales, ajuste por probabilidad de respuesta y calibración de los factores de expansión de la ENEMDU Longitudinal. Cabe mencionar que en cada actividad de este proceso existe la validación correspondiente.

Gráfico 1. Diagrama de flujo del proceso de cálculo de los factores de expansión de la ENEMDU Longitudinal.



Construcción de los pesos básicos iniciales longitudinales:

Usando la Encuesta ENEMDU se calculó una estructura de ponderación longitudinal para los individuos pertenecientes al panel según la rotación 2-(2)-2). Para esto se inició con el cálculo de los pesos iniciales, dados por el primer periodo transversal (julio, agosto, septiembre 2022). El factor básico de expansión denotado como d_{1k} se define como el inverso multiplicativo de la probabilidad de inclusión, π_k.

Desde el principio todos los elementos seleccionados cuentan con un factor de expansión, se pueda o no recoger la información de la encuesta, así que posterior al proceso de recolección será necesario llevar a cabo un proceso adicional para ajustar los factores de expansión por la falta de cobertura o la ausencia de respuesta (Kalton & Flores-Cervantes, 2003).

Para la construcción de los pesos básicos iniciales es necesario mencionar que se usaran los factores iniciales ponderados calibradas de la ENEMDU acumulada trimestral julio, agosto, septiembre 2022.

Ajuste por probabilidad de inclusión del panel

Luego de haber realizado los ajustes necesarios en la base del primer periodo, es necesario filtrar cada una de las bases transversales con el identificador del panel de interés. La determinación de los pesos iniciales viene supeditada a los pesos básicos ajustados por cobertura d_{1k} del procesamiento transversal del primer trimestre que se quiere combinar. En general, dado que cada panel es representativo del país y tienen las mismas características al momento de la selección, De acuerdo con (LaRoche, 2003), el factor de expansión básico longitudinal se crea a partir del inverso de la probabilidad de inclusión de los paneles, de modo que:

$$d_{1k} = \frac{W_{k,pos}}{\Pr\ (selección\ de\ paneles)}$$

Dado que el esquema rotacional de la ENEMDU es 2-(2)-2 trimestral, entonces el traslape del primer periodo (trimestre III 2022) como del segundo periodo (trimestre III 2023), será del 25%. En particular para los dos periodos, el tamaño de la muestra de la ENEMDU fue de 966 unidades primarias de muestreo comunes (traslapadas en los periodos de interés), repartidas en cada uno de los 150 estratos de muestreo a lo lardo del territorio nacional. Por lo tanto.

$$d_{1k} = \frac{W_{k,pos}}{0{,}25}$$

donde:

d_{1k} = pesos básicos ajustados por cobertura.
$W_{k,pos} = factor\ de\ expansion\ inicial, calibrado\ ENEMDU\ trimestral.$

Generación de los pesos longitudinales

A partir del emparejamiento de estas bases y habiendo seguido un proceso riguroso de identificación secuencial de respondientes y no respondientes, se procede a realizar la combinación de las correspondientes bases de datos transversales. Este procedimiento debe tener en cuenta únicamente a las unidades muéstrales que respondieron sistemáticamente en cada uno de los periodos de interés. Por lo tanto, se sugiere que se sigan los siguientes pasos:

- Identificación de los respondientes en TIII 2022 y TIII 2023.
- Identificación de quienes respondieron en TIII 2022, pero no en TIII 2023.

En esta instancia se construye la base longitudinal que será usada para realizar los análisis de interés.

En primer lugar, se define la muestra longitudinal $s_r(1-2)$como aquella constituida por las unidades seleccionadas en ambos periodos de interés para los paneles coincidentes:

$$s_r(1-2) = s^1 \cap s^2$$

La muestra $s_r(1-2)$ es representativa de la población longitudinal en los dos periodos combinados. En esta etapa, el factor de expansión longitudinal se define como idéntico al peso resultante de la sección anterior; es decir $d_{12.k}^{básico} = d_{1,k}^{básico}$. Es necesario identificar las unidades que no respondieron en alguna ocasión para asignarles un peso longitudinal nulo; es decir:

$$d_{12.k}^{básico} = \begin{cases} d_{1,k}^{básico}, & \forall k \in s_r(1-2) \\ 0, & \forall k \notin s_r(1-2) \end{cases} .$$

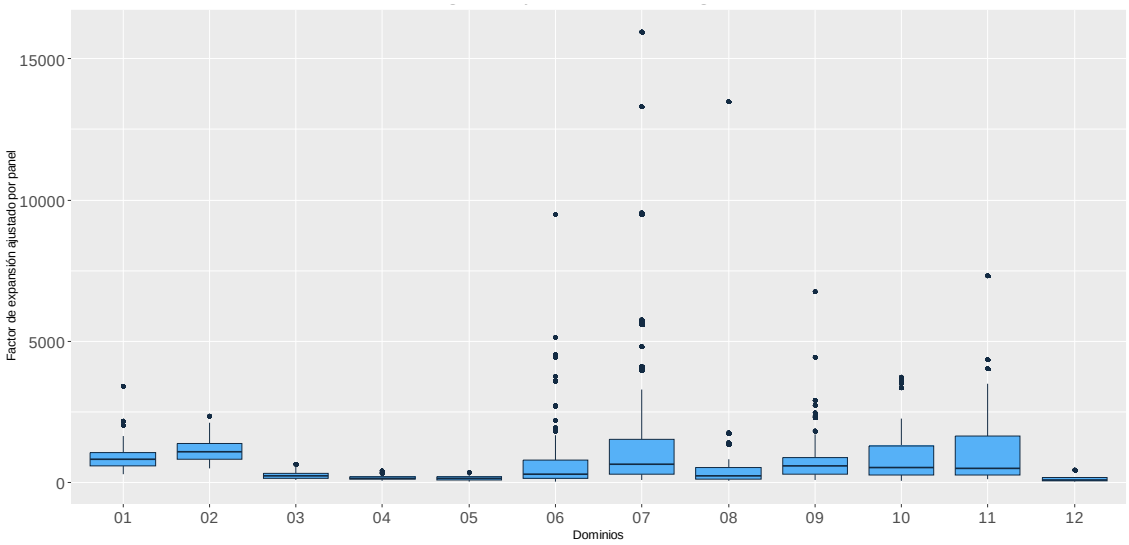
En donde el conjunto $s_r(1-2)$ representa a las unidades que respondieron la encuesta en los dos periodos de la combinación. Las unidades que no

respondieron en alguna ocasión deberán ser excluidas de la base de datos puesto que su peso de muestreo es nulo.

Para la construcción del factor básico longitudinal correspondiente al panel conformado por los dos periodos (Trimestre III 2022 y Trimestre III 2023), se tienen en cuenta todas las unidades que estuvieron en ambos periodos de interés. Sea $s^{(2)}$ el conjunto de la muestra longitudinal constituida por las unidades seleccionadas en los dos periodos para los paneles coincidentes, es decir, la intersección de las muestras transversales del TIII-2022 y el TIII-2023,

$$s^{(2)} = s^{TIII-2022} \cap s^{TIII-2023}.$$

Gráfico 2. Distribución por dominio de los factores básicos longitudinales para el TIII 2022 – TIII 2023



Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). INEC - Ecuador.

En el Gráfico 2 se presenta la distribución del factor básico longitudinal. Usando las 966 UPM del panel de interés, se cuenta con una muestra de 22.459 personas comunes (7.600 entre julio 2022 y julio 2023, 7.497 entre agosto 2022 y agosto 2023 7.362 entre septiembre 2022 y septiembre 2023).

Ajuste por ausencia de respuesta

Según Rosenbaum y Rubin (1983), propusieron una técnica útil para dilucidar la estructura de la ausencia de respuesta y, por consiguiente, corregir el sesgo de cobertura y el sesgo por ausencia de respuesta (Lensvelt-Mulders, Lugtig y Hubregtse, 2009).

Para el manejo efectivo de la ausencia de respuesta se consideran las variables dicotómicas I_k y D_k, que indican si el individuo pertenece a la muestra original y

si ha respondido a la ENEMDU Longitudinal, respectivamente. Suponiendo que la distribución de las respuestas efectivas puede ser estimada, la probabilidad de respuesta (Propensity Score) de una persona en la muestra está dada por:

$$\emptyset_k = \Pr(D_k = 1 \mid I_k = 1).$$

Esta probabilidad es distinta para cada persona y puede ser estimada usando los datos del panel (muestra original). Contar con la muestra original, para la cual se obtuvo toda la información del cuestionario en un período anterior, constituye un excelente punto de partida para tratar de eliminar el sesgo, puesto que se cuenta con un conjunto robusto de covariables para determinar el mejor modelo a fin de estimar el patrón de ausencia de respuesta en la muestra de respondientes efectivos (CEPAL, 2020).

Es importante mencionar que de las 22459 personas que deberían estar traslapadas en los dos periodos de interés, únicamente 17513 respondieron a los 2 periodos (5.795 entre julio 2022 y julio 2023, 5.884 entre agosto 2022 y agosto 2023, 5.834 entre septiembre 2022 y septiembre 2023). Por ende, 4.946 personas son no respondientes en este panel en particular. Dado que las tres distribuciones son similares. Bajo un modelo de regresión logística, la estimación de las probabilidades de respuesta tendrá la siguiente forma:

$$\widehat{\emptyset}_{123,k} = \frac{\exp\left(x^{'}\hat{\beta}\right)}{1 + \exp\left(x^{'}\hat{\beta}\right)}$$

Los pesos básicos son ajustados utilizando el inverso de la probabilidad de respuesta sobre los respondientes efectivos en el primer periodo de interés, así se conforma el primer conjunto de pesos iniciales de las bases de datos longitudinales, mediante la siguiente formula:

$$d_{12,k}^{propensity} = \frac{d_{12,k}^{inicial}}{\widehat{\emptyset}_{12,k}}$$

Es recomendable corroborar que la suma de los pesos ajustados por la ausencia de respuesta esté cercana al tamaño de la población que se quiere representar.

Identificación de las variables más relevantes que explique la respuesta o ausencia de respuesta

Según CEPAL (2020) manifiesta que "se debe prestar especial atención a la elección de predictores en el modelo de regresión logística, que debería

funcionar bien si las variables de información auxiliar disponibles son relevantes y explicativas de la respuesta ENEMDU longitudinal; de otra forma, esta metodología no tendrá ningún beneficio para la reducción del sesgo (y posiblemente lo exacerbará) y dará como resultado errores estándares más grandes".

Se debe tener en cuenta que la variable dependiente es una variable dicotómica: 1 si la persona respondió a la encuesta en los dos periodos (17.513 individuos) y 0 si la persona solo respondió a la encuesta en el primer periodo del 2021 (4.946 individuos), dando el total de 22459 registros de la muestra original. A continuación, se detallarán las variables explicativas usadas en cada uno de los modelos analizados:

Tabla 11. Variables explicativas utilizadas en el modelo

Variable	Nombre en base de datos	Tipo	Niveles
Sexo	sexo	Factor	Hombre Mujer
Edad	gedad	Factor	Seis grupos de edad
Ingreso per-cápita	ingreso	Factor	6 Categorías formadas por los quintiles del ingreso
Condicion de ocupacion	condactn	Factor	9 categorías derivadas de la condición de ocupación
Mes del primer levantamiento	mes	Factor	1 = enero 2 = febrero 3 = marzo
Estrato	estrato	Factor	150 estratos de muestreo
Pobreza	Recategorización de la pobreza monetaria	Factor	1 = pobre extremo 2 = pobre relativo 3 = No pobre 4 = Sin ingresos reportados
Formación laboral	Recategorización de la formalidad laboral	Factor	1 = formal 2 = informal 99 = no aplica
Condición de parentesco	p04	Factor	9 categorías de parentesco en el hogar; desde jefe de hogar hasta otra clase de parientes.

Estado civil	p06	Factor	1 = casado 2 = separado 3 = divorciado 4 = viudo 5 = unión libre 6 = soltero 99 = no aplica
Escolaridad	p10a	Factor	10 categorías de nivel de instrucción

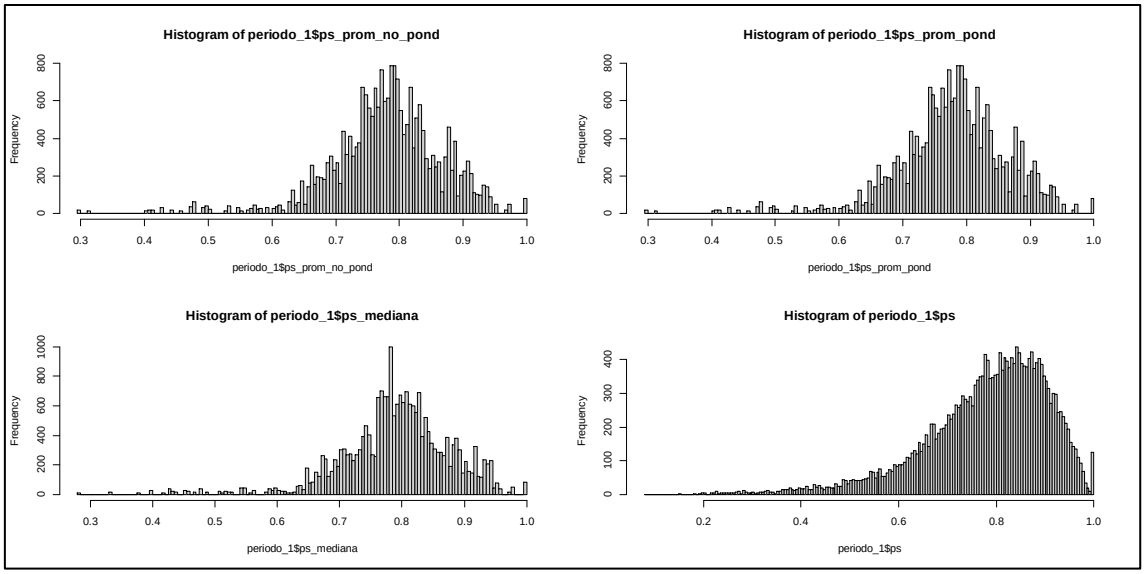
Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). INEC - Ecuador.

Validación de la probabilidad de respuesta (ps)

Dado que las covariables del modelo están disponibles a nivel de persona, entonces es natural que las probabilidades de respuesta estimadas por el modelo sean desiguales entre los individuos. En este paso, es recomendable utilizar una de las diferentes opciones disponibles en la literatura para controlar la variabilidad de los pesos inducida por esta heterogeneidad. Además, es posible aumentar la eficiencia del estimador si se crean categorías homogéneas de individuos que tengan la misma probabilidad de responder. En este escenario complejo, en el cual las probabilidades de respuesta fueron estimadas con un modelo de propensity score y teniendo en cuenta que las estimaciones de estas probabilidades varían entre cero y uno, es posible crear clases de individuos (respondientes y no respondientes) con probabilidades similares. En este caso, se asumiría que las unidades dentro de una misma clase tendrían la misma configuración de covariables, o al menos, una probabilidad de respuesta estimada similar. Así, dentro de cada clase, las unidades serían tratadas como si fuesen sido aleatorizadas al tratamiento (responder) o al control (no responder). Por lo tanto, el objetivo de este proceso es asegurar que cualquier diferencia en las covariables pueda ser ajustada dentro de la clase. Teniendo en cuenta que, si el modelo es adecuado, la estimación resumiría los efectos de las covariables en la respuesta del individuo, entonces una vez hayan sido creadas las clases es posible realizar el ajuste mediante alguna medida de localización en cada clase y, de esta forma, todos los individuos de una misma clase se ajustarían de la misma manera. Valliant and Dever (2017) muestran algunas medidas comúnmente adoptadas sobre las probabilidades de respuesta estimadas por el modelo ajustado:

- Promedio no ponderado.
- Promedio ponderado.
- Mediana no ponderada.

Gráfico 3. Distribuciones del promedio y mediana



Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). INEC - Ecuador.

En el Gráfico 3 se visualiza el promedio no ponderado (izquierda-arriba), promedio ponderado (derecha-arriba), mediana (izquierda-abajo) de las probabilidades de respuesta en clases definidas por las UPM, comparadas con la distribución original (derecha-abajo). Nótese que, si todas las unidades dentro de una clase tienen la misma probabilidad de responder, entonces la tasa de repuesta no ponderada es la mejor opción. Además, si dentro de las clases las unidades tienen una probabilidad de responder muy disímil, entonces el promedio no ponderado (o ponderado) del ps puede usarse. De la misma manera, la tasa estimada de repuesta puede ser ineficiente si los pesos de muestreo varían demasiado, pero la probabilidad de respuesta es similar en cada clase. Por último, la mediana se considera si la distribución de la probabilidad de respuesta es sesgada. Con estos resultados a nivel de UPM (definida como las clases) son satisfactorios para el promedio ponderado y no ponderado y para la mediana.

Formulación de los modelos de regresión logísticos

Una investigación realizada por CEPAL (2020) señala que si se asume que la probabilidad de respuesta depende de alguna combinación lineal de las covariables disponibles en la muestra original, es posible ajustar un modelo en que la variable dependiente es D_k y un vector de covariables independientes.

Kim y Riddles (2012) muestran que es posible utilizar un modelo basado en el ajuste de la probabilidad de respuesta mediante la siguiente expresión:

$$logit\left(\widehat{\emptyset}_k\right) = X_k\hat{\beta},$$

Donde $\hat{\beta}$ es el vector de coeficientes estimado de la regresión logística y X_k representa la matriz de covariables más importantes.

Con la finalidad de obtener la probabilidad estimada de respuesta o no respuesta, se efectuaron 3 modelos regresión logísticos, en los cuales la variable dependiente es dicotómica e identifica si el individuo respondió a la ENEMDU (1) o no (0).

A continuación, se detallan las variables independientes utilizadas en los 3 modelos efectuados:

- Modelo 1: Sexo, grupo de edad, relación de parentesco, estado civil y nivel de instrucción.
- Modelo 2: Las variables descritas en el Anexo 1 de este documento más las interacciones.
- Modelo 3: Las variables descritas en el Anexo 2.

Los resultados de los 3 modelos planteados sugieren que el mejor modelo logístico es el tercero, por lo que presenta el menor criterio de decisión por Akaike (Ver Anexo 3); llegando a la conclusión que el mejor modelo para explicar la respuesta o ausencia de la misma en la muestra de respondientes efectivos es el Modelo 3.
Con este modelo, se obtienen las probabilidades estimadas $\widehat{\emptyset}_k$ (Propensity Score) para respondientes y no respondientes de la muestra de respondientes efectivos, y a su vez estas probabilidades son transferidas a la muestra longitudinal, entonces el factor de expansión ajustado por probabilidad de respuesta toma la siguiente forma:

$$w_k = \frac{d_k}{\widehat{\emptyset}_k}$$

donde:

$\widehat{\emptyset}_k = \ ps\ estimados, de\ igual\ manera, transferidos\ a\ la\ ENEMDU\ Longitudinal$
$d_k = ajustados\ por\ la\ probabilidad\ de\ inclusion\ de\ los\ paneles$

Utilizar el factor de expansión ajustado por probabilidad de respuesta en el cálculo de los estimadores deseados minimizaría el sesgo de selección (CEPAL, 2020).

Calculo y validación de los Propensity Score estimados

En la Tabla 10 se aprecia que la muestra original está conformada por 22.459 individuos, de los cuales 17.513 personas respondieron, mientras que 4.946 individuos no respondieron a la ENEMDU Longitudinal.

Tabla 12. Total de individuos respondientes y no respondientes de la ENEMDU Longitudinal

Muestra original	Individuos	Porcentaje
Respondientes	17.513	22,02%
No respondientes	4.946	77,98%
Total	22.459	100,00%

Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). INEC - Ecuador.

Los puntajes de propensión estimados a partir de un modelo logístico, deben estar en un rango entre 0 y 1, además, la sumatoria de estos puntajes debe ser igual al número de respondientes efectivos de la ENEMDU Longitudinal, que en este caso son 17.513 individuos.

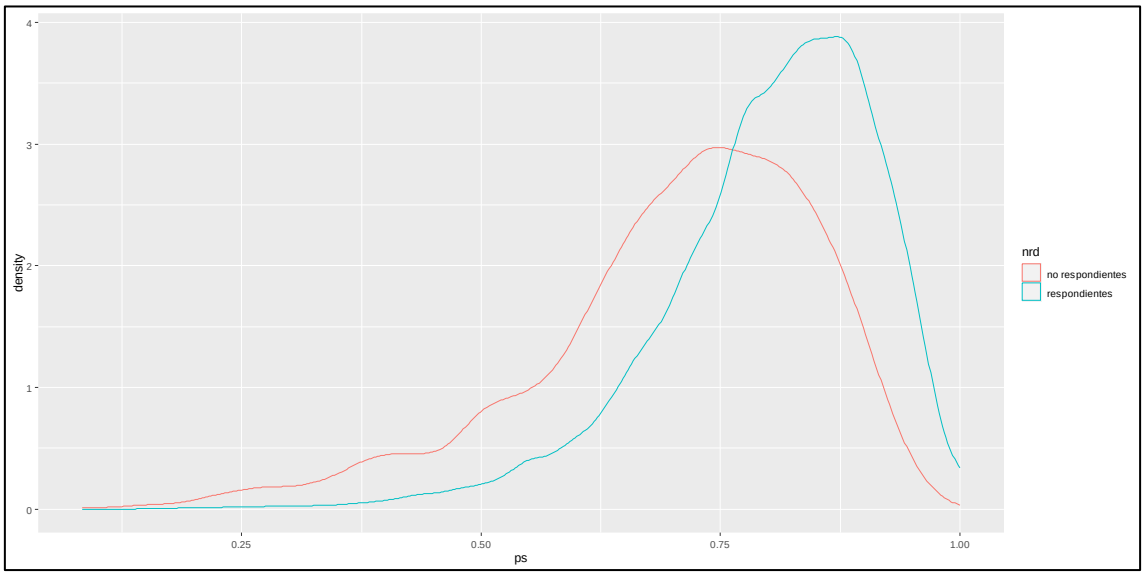
Tabla 13. Estadísticos descriptivos de los Propensity Score estimados

Estadístico	Valor
Mínimo	0,0776
Primer cuartil	0,7128
Mediana	0,7991
Media	0,7797
Tercer cuartil	0,8729
Máximo	1,0000

Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). INEC - Ecuador.

Por otra parte, otra validación que se debe realizar corresponde a la verificación de soporte común, el cual compara las funciones de densidad de los individuos de la muestra original que respondieron o no respondieron la ENEMDU Longitudinal.

Gráfico 4. Soporte común de los individuos respondientes y no respondientes



Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). INEC - Ecuador.

Como se ilustra en el Gráfico 4, tanto las personas de la muestra original que respondieron como los que no respondieron comparten las mismas cotas o extremos inferiores o superiores de los Propensity Score estimados.

Calibración de los factores de expansión Longitudinales

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 estándar 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 muestras. 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 $g_k = F(x_k'\lambda)$.

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

$$w_{k_c} = w_k \ * g_k$$

Donde w_{k_c} son los factores de expansión calibrados, mientras que w_k son los pesos de muestreo ajustados por Propensity score y g_k los pesos de calibración.

Cabe señalar que la calibración de los factores de expansión se realizó a nivel de UPM, es decir, todos los individuos de una UPM presentaban un mismo ponderador, independientemente de sus características demográficas como edad y sexo.

Para la ENEMDU Longitudinal, se aplicó un esquema de calibración en el cual se presentan 8 celdas o post estratos de calibración con la información auxiliar correspondiente a las proyecciones de población del mes intermedio del periodo 1 (en este caso es el mes de febrero de 2021) (Ver Anexo 4), y se evaluó cada uno a través de criterios para validar la calibración propuestas por Silva (2004) (Ver Anexo 5).

Es necesario mencionar que las proyecciones poblacionales son elaboradas en un área diferente a la Gestión de Diseño Muestral. Los criterios demográficos aplicados en la generación de las proyecciones poblacionales son explícitamente responsabilidad de los funcionarios que las elaboraron y quienes en la actualidad forman parte del equipo del Censo de Población y Vivienda 2021-2022 (Equipo de Proyecciones de Población del CPV).

Validación de la calibración de los factores de expansión

Silva (2004) propone 6 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

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

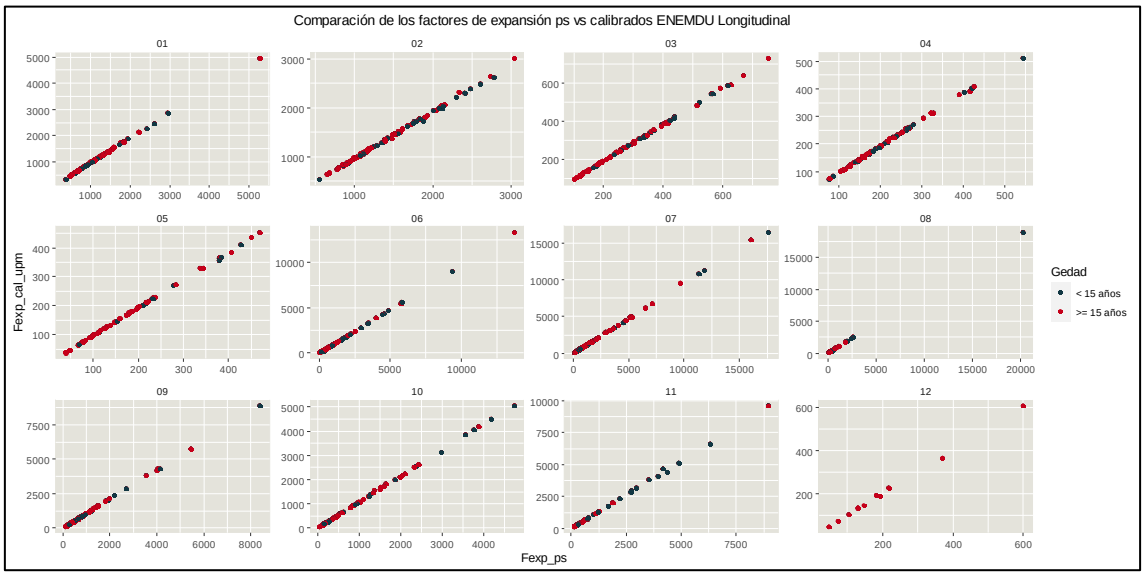
- Distancia entre los pesos de calibración y los pesos originales

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

Según el Gráfico 5, en la mayoría de dominios, los factores de expansión calibrados son cercanos a los factores ajustados por Propensity score, debido a la condición de reproducir con exactitud los totales poblacionales por sexo y grupo de edad[1], en las diferentes celdas o post estratos de calibración.

[1] Menores a 15 años y mayores o iguales a 15 años.

Gráfico 5. Comparación de los factores de expansión por probabilidad de respuesta y calibrados por dominio y grupo de edad



Fuente: Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). INEC – Ecuador.

6. Estimaciones de características

Estimación de características de la población:

Una vez construidos los factores de expansión se calculan los estimadores provenientes de la ENEMDU, para ello se utiliza el estimador de Horvitz-Thompson, el cual sirve para estimar el valor total de una característica determinada. Está dado por (Carl-Erik Särndal, 1992):

$$\hat{Y}_{HT} = \sum_{k \in D} \sum_{l \in k} w_{k_c} y_{k_l}$$

donde:

$\hat{Y}_{HT} = estimador\ HT\ para\ el\ total\ de\ la\ característica\ de\ interes\ Y\ de\ la\ variable\ y$
$w_{k_c} = factor\ de\ expansión\ calibrado\ de\ la\ vivienda\ k$
$y_{k_l} = valor\ de\ la\ variable\ y\ para\ la\ persona\ l\ de\ la\ vivienda\ k.$

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 $\hat{Y}_{HT}$. 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 $\hat{Y}_{HT}$ viene dado por la siguiente expresión:

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

donde:

$\hat{V}_{2st}(\hat{Y}_{HT})$ = *estimación de la varianza de dos etapas del estimador HT del total de la variable* y.
$\hat{Y}_{HT}$ = *estimador HT del total de la característica de interés* Y.

Un estimador insesgado para la varianza está dado por:

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

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

$$\hat{V}_{\mathrm{i}} = \sum\sum_{s_{IIi}} \breve{\Delta}_{IIqr|i} \frac{\hat{t}_{iq\pi}}{\pi_{IIq|i}} \frac{\hat{t}_{ir\pi}}{\pi_{IIr|i}},$$

donde:

i, j = *Índice que recorre las UPM* i, j *en el dominio de estudio* d.
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* d.
$\pi_{IIq|i}$ = *Probabilidad de selección de la Etapa II para la* $q-$*ésima vivienda, dada la* $i-$*ésima UPM*.
$\breve{\Delta}_{Iij}$ = *Cantidad* Δ *expandida asociada a las UPM* i, j.
$\breve{\Delta}_{IIqr|i}$ = *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[2]" 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, y R. En virtud de las características de cada uno de los métodos es

[2] El documento se encuentra disponible en: https://www.ecuadorencifras.gob.ec/documentos/web-inec/EMPLEO/2021/Enero-2021/202101_ENEMDU_Calculo%20de%20errores%20estandar%20y%20declaracion%20de%20muestras%20complejas.pdf

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 ENEMDU.

Las variables requeridas para declarar el diseño muestral en los programas estadísticos (SPSS, y R) y ejecutar el cálculo de los errores de muestreo son presentadas en la Tabla 11, donde se describe las etiquetas de las variables identificadoras de las UPM, estratos y factores de expansión.

Tabla 14. 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
Ponderación	fexp	Factor de expansión calibrado

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

CEPAL. (2021). *Recomendaciones Metodológicas para el Rediseño de la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU 2021 – 2024)*. Informe de misión al Instituto Nacional de Estadística y Censos (INEC) de Ecuador.

CEPAL. (2019). *Revisión del esquema de agregación y análisis de la Encuesta Nacional de Empleo, Desempleo y Subempleo – ENEMDU. Misión de Asistencia Técnica.*

Kim, J. K. y M. K. Riddles (2012), *"Some theory for propensity-score-adjustment estimators in survey sampling", Survey Methodology, vol. 38, N° 2*

Kalton, G., & Flores-Cervantes, I. (2003). *Weighting Methods. Journal of official statistics,19(2), 81--97.*

Gutiérrez, Andrés. 2016. *Estrategias de muestreo: diseño de encuestas y estimación de parámetros. Segunda edición. Ediciones de la U.*

LaRoche, S. (2003). *Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics. Statistics Canada.*

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*.

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

Nicolas Privault, Understanding Markov Chains, Examples and Applications, Nanyang Technological University, Springer Singapore Heidelberg New York Dordrecht London, Library of Congress Control Number: 2013942497, 2013.

Rosenbaum, P. R. y D. B. Rubin (1983), "The central role of the propensity score in observational studies for causal effects", Biometrika, vol. 70, N° 1.

Lensvelt-Mulders, G., P. Lugtig y M. Hubregtse (2009), "Separating selection bias and non-coverage in Internet panels using propensity matching", Survey Practice, 2, N° 6.

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

INE (2020). Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares. Departamento de Metodología e Innovación Estadística. Instituto Nacional de Estadísticas. Chile.

Anexos

Anexo 1: Variables utilizadas en modelo 2

Variable	Pr(> |z|)	Hipótesis	Decisión
Grupo de Edad	0,026736	Rechazo h0	Significativo
Relación de parentesco	0,017954	Rechazo h0	Significativo
Estado Civil	0,000021	Rechazo h0	Significativo
Nivel de instrucción	0,015413	Rechazo h0	Significativo
Ingreso	0,000903	Rechazo h0	Significativo
Condición de actividad	0,002482	Rechazo h0	Significativo
Mes	0,000176	Rechazo h0	Significativo
Estrato	0,000792	Rechazo h0	Significativo
Gedadxcondactn	0,000344	Rechazo h0	Significativo
Gedadxingreso	0,002399	Rechazo h0	Significativo

Anexo 2: Variables utilizadas en modelo 3

Variable	Pr(> |z|)	Hipótesis	Decisión
Grupo de Edad	0,000113	Rechazo h0	Significativo
Relación de parentesco	0,000002	Rechazo h0	Significativo
Estado Civil	0,000816	Rechazo h0	Significativo
Nivel de instrucción	0,004093	Rechazo h0	Significativo
Ingreso	0,000259	Rechazo h0	Significativo
Condición de actividad	0,000132	Rechazo h0	Significativo
Mes	0,001890	Rechazo h0	Significativo
Estrato	0,000826	Rechazo h0	Significativo

Anexo 3: Criterios de información de Akaike de los modelos logísticos realizados.

Modelo	Akaike
Modelo 1	23.986,11
Modelo 2	22.043,97
Modelo 3	22.012,28

Anexo 4: Esquema de calibración propuesto.

El post estrato al que pertenece cada observación se puede identificar mediante un identificador de 4 dígitos (id_calib), 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	Grupo de edad: menor a 15 años (1), y mayor o igual a 15 años (2)

En el ejemplo citado anteriormente, el post estrato "00_1_1" identifica a los individuos menores a 15 años de áreas urbanas a nivel nacional.

id_calib	Población(t)	Calibradas(d)
00_1__1	3.366.715	3.534.433
00_1__2	8.973.527	9.372.117
00_2__1	1.853.554	1.729.883
00_2__2	3.907.406	3.631.216

Anexo 5: Medidas para evaluar el esquema de calibración planteado

Medida 1:

Id_calib	er_upm
00_1__1	5,60999E-13
00_1__2	3,47889E-13
00_2__1	8,99013E-12
00_2__2	1,32809E-11

Medida 2:

n	cv_upm
4	0,087653053

Medida 3:

Id_calib	L	n	M3_upm
00_1__1	1	2.668	100%
00_1__2	1	9.938	100%
00_2__1	1	1.167	0
00_2__2	1	3.740	0

Medida 4:

Id_calib	n	U	U3	M4_upm_U	M4_upm_U3
00_1__1	2.668	0,95	3	70,95%	0,00%
00_1__2	9.938	0,96	3	69,76%	0,00%
00_2__1	1.167	1,07	3	54,93%	0,00%
00_2__2	3.740	1,08	3	53,45%	0,00%

Medida 5:

Id_calib	n	cv_g_upm
00_1__1	2.668	0,01
00_1__2	9.938	0,01
00_2__1	1.167	0,02
00_2__2	3.740	0,02

Medida 6:

Id_calib	n	cv_g_upm
00_1_1_1	2.668	3,27
00_1_1_2	9.938	1,92
00_1_2_1	1.167	8,16
00_2_2_2	3.740	6,05


INEC
Buenas cifras,
mejores vidas
@ecuadorencifras
@ecuadorencifras
@InecEcuador
t.me/ecuadorencifras
INEC/Ecuador
INECEcuador
INEC Ecuador