VI Encuesta de Microemprendimiento
EME 2019

Documento metodológico de diseño muestral

INSTITUTO NACIONAL DE ESTADÍSTICAS

2019

DEPARTAMENTO DE METODOLOGÍAS E INNOVACIÓN ESTADÍSTICA
SUBDEPARTAMENTO DISEÑO DE MARCOS Y MUESTRAS
UNIDAD DE ESTADÍSTICAS SOCIALES

VI Encuesta De Microemprendimiento EME 2019.
Documento metodológico de diseño muestral.
Entrega N°02 / Versión N°01.

Instituto Nacional de Estadísticas, Chile.
2019.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE CUADROS

PRESENTACIÓN

La Subsecretaría de Economía y Empresas de Menor Tamaño, ha solicitado al Instituto Nacional de Estadísticas (INE) la aplicación, desde el año 2013, de la "Encuesta de Microemprendimiento" (EME). Esta encuesta es una herramienta de enorme valor estadístico para el país, puesto que es el único estudio de este tipo que se realiza a lo largo de todo Chile, abarcando unidades económicas pequeñas, formales e informales, pertenecientes a todos los sectores económicos.

El objetivo principal de la EME, es lograr una caracterización profunda de los microemprendimientos que se desarrollan a nivel nacional, permitiendo conocer las limitantes y los elementos facilitadores que tienen las unidades económicas de menor tamaño para llevar a cabo sus actividades dentro del mercado laboral. Para esto, se toma una muestra representativa de viviendas particulares ocupadas a nivel nacional y regional.

La muestra de la VI EME, se obtiene a partir de los datos levantados en la "Encuesta Nacional de Empleo" (ENE) del trimestre móvil marzo, abril, mayo (MAM 2019) por lo que el diseño considera una estrategia de muestreo bifásico, donde la primera fase corresponde al levantamiento de la ENE en el trimestre señalado. Por su parte, la segunda fase corresponde a todas las viviendas, que se identificaron en la primera fase, que contienen al menos un microemprendedor[1]; formando así el marco de muestreo.

En el cálculo del tamaño muestral se considera sobremuestreo, que consiste en enviar a levantamiento una cantidad de viviendas, que incluye la posible pérdida de algunas de ellas por falta de respuesta o por problemas de desactualización del marco; estrategia que se encuentra en línea con las recomendaciones internacionales (ONU, 2009, pág. 124).

Es importante mencionar que, a partir de esta versión, se incorpora la región de Ñuble, al igual que en todas las encuestas que realiza el INE desde el año 2018.

El presente informe, contiene una descripción metodológica del diseño muestral, que incluye las características del marco utilizado, el cálculo del tamaño muestral, los métodos de selección de las unidades y la forma de expandir los datos.

[1] Todos los trabajadores por cuenta propia y empleadores con hasta 10 trabajadores, incluyéndose.

I. DISEÑO MUESTRAL

I.1. Objetivos

I.1.1. Objetivo general de la encuesta

Caracterizar los microemprendimientos que se desarrollan a nivel nacional, permitiendo conocer las limitantes y los elementos facilitadores que tienen las unidades económicas de menor tamaño, para llevar a cabo sus actividades dentro del mercado laboral, a partir de una muestra representativa de viviendas particulares que contienen al menos a un microemprendedor, a nivel nacional y regional.

I.1.2. Objetivos específicos de la encuesta

1. Observar las formas en las que operan las unidades económicas.
2. Estimar la productividad y los ingresos de las Microempresas.
3. Estudiar el grado de formalidad de los micronegocios.
4. Cuantificar el acceso al sistema financiero y cómo se vinculan con éste.
5. Identificar sus recursos productivos.
6. Cuantificar el empleo generado por las microempresas y las características de éste.

I.1.3. Objetivo del diseño muestral

Obtener estimaciones del parámetro de interés "Proporción de microemprendedores por cuenta propia sobre el total de microemprendedores" según los niveles de precisión establecidos[2] a nivel nacional y regional.

I.2. Población objetivo

La población objetivo está conformada por todos los trabajadores por cuenta propia y empleadores con hasta 10 trabajadores incluyéndose, denominados Microemprendedores, que residen en viviendas particulares del territorio nacional.

[2] Los niveles de precisión son: error absoluto de 1,3% y error relativo de 1,5%, a nivel nacional.
Para las regiones, el error absoluto máximo es 8,2% y, el error relativo no supera el 11,0%, correspondiente a la Región de Magallanes.

I.3. Unidad de información

La unidad de información es el Microemprendedor perteneciente a la vivienda particular, que fue clasificado como trabajador independiente en la Encuesta Nacional de Empleo y que mantiene su condición.

I.4. Cobertura geográfica y niveles de estimación (dominios de estudio)

La cobertura geográfica contempla la población residente en todo Chile con exclusión de las ADA´s y de las manzanas con 7 o menos viviendas, debido a potenciales problemas operativos y los costos que implica recolectar información en unidades de muy poca densidad poblacional. En el Cuadro I.1 se listan las áreas geográficas excluidas.

Cuadro I.1. Áreas geográficas excluidas del Marco de Muestreo del INE, clasificadas como ADA's.

Región	Nombre Provincia	Nombre Comuna	Total Viviendas Censo 2002
Total Viviendas ADA's			**16.046**
Arica y Parinacota	Parinacota	General Lagos	447
Tarapacá	Tamarugal	Colchane	1.395
Antofagasta	El Loa	Ollagüe	287
Valparaíso	Valparaíso	Juan Fernández	257
	Isla de Pascua	Isla de Pascua	1.416
Los Lagos	Llanquihue	Cochamó	1.676
	Palena	Chaitén	2.305
		Futaleufú	853
		Hualaihué	2.553
		Palena	760
Aysén del General Carlos Ibáñez del Campo	Coyhaique	Lago Verde	590
	Aysén	Guaitecas	463
	Capitán Prat	O'Higgins	249
		Tortel	187
Magallanes y de La Antártica Chilena	Magallanes	Laguna Blanca	267
		Río Verde	197
		San Gregorio	603
	Antártica Chilena	Cabo de Hornos (Ex - Navarino)	626
		Antártica	24
	Tierra el Fuego	Primavera	459
		Timaukel	172
	Última Esperanza	Torres del Paine	260

Elaborado por el Instituto Nacional de Estadísticas.

En relación a los niveles de estimación, los tamaños muestrales son determinados para obtener representatividad estadística en la estimación del parámetro de interés "Proporción de microemprendedores por cuenta propia sobre el total de microemprendedores" en los siguientes niveles:

- Nacional
- Regional

I.5. Período de referencia y periodicidad

La información se levanta en el trimestre de mayo, junio y julio 2019 (MJJ 2019), con periodicidad bienal. Mientras que, el periodo de referencia corresponde al trimestre MAM 2019 en el cual se realizó la primera fase de la encuesta (levantamiento de la ENE).

I.6. Marco Muestral

Un marco muestral se define como un conjunto de unidades en que todos sus elementos son inequívocamente identificables, mediante algún procedimiento o listado, a los que es posible asociarles una probabilidad de selección distinta de cero de acuerdo con la estrategia de muestreo.

El marco muestral de manzanas que utiliza el INE para las zonas geográficas urbanas está conformado por unidades geográficas con límites fijos, denominadas manzanas (o conglomerados). El INE posee información cartográfica y del número de viviendas que contiene cada manzana según Censo de Población y Vivienda del año 2002, donde el 80% de los conglomerados sufrió una actualización en el total de viviendas, según información del Precenso de 2016. Este marco es utilizado para seleccionar muestras de las principales encuestas de hogares que se realizan en el país, tal como la Encuesta de Caracterización Socioeconómica Nacional (Casen).

Hasta el año 2009, la actualización del marco se realizaba mediante registros administrativos, provenientes de los certificados de recepción final de las nuevas construcciones otorgados mensualmente, por las direcciones de obras de las municipalidades, capturadas en el Formulario Único de Edificación del INE, lo que permitía mantener actualizado el número total de viviendas en las manzanas e incorporar aquellas creadas con posterioridad al Censo de Población y Vivienda 2002.

A fines de 2018, se realiza una actualización tanto al Marco de secciones 2002 como al Marco de manzanas, a partir de los totales de viviendas provenientes del Precenso del año 2016. Estos marcos, que son los que actualmente mantiene vigentes el INE, se denominan: **Marco muestral de manzanas 2016** (MMM2016) para el área Urbana y **Marco muestral de secciones** (MMS2016) para las áreas Rural y Resto de Área Urbana (RAU).

A continuación, se describen las características del marco muestral utilizado para la selección muestral de la VI EME. Dado que, las unidades seleccionadas en la VI EME provienen de la ENE, se deben revisar las características del marco de muestreo asociado a la fase 1 (ENE) y a la fase 2 (EME).

I.6.1. Construcción del marco muestral

El marco muestral se encuentra organizado en forma jerárquica, de acuerdo con la división político-administrativa del territorio. Por su parte, las unidades se ordenan en forma descendente según región, provincia y comuna. Al interior de cada comuna, se conforma la división censal que da origen a las áreas geográficas urbanas y rurales.

I.6.2. Estratificación del marco muestral

La estratificación, previa a la selección de la muestra, corresponde al proceso de agrupar los elementos de la población según su homogeneidad respecto a características determinadas. Con el objeto de mejorar la precisión estadística de las estimaciones, esta agrupación debe contener unidades homogéneas internamente y heterogéneas entre sí, además de ser mutuamente excluyentes y colectivamente exhaustivas, es decir, cada elemento en la población debe ser asignado solo a un estrato y ningún elemento de la población puede quedar excluido.

El marco muestral posee una estratificación geográfica según la división político-administrativa que, en conjunto con la división censal, según Censo 2002, da origen a las áreas urbanas y rurales.

I.6.2.1. Estratificación geográfica

El INE cuenta con dos listados de áreas geográficas conformadas a partir del Censo 2002 que, en conjunto, forman el marco muestral, que es la base para la seleccionar viviendas en las diferentes encuestas de hogares.

Los listados contienen las unidades geográficas que cubren todo el territorio nacional, organizadas de forma jerárquica de acuerdo con la división político-administrativa: región, provincia y comuna.

La estratificación del marco de la ENE, correspondiente a la primera fase, da origen a los siguientes estratos:

- **Ciudad o grandes Centros Urbanos (CD)**: Conformadas por ciudades o conjuntos de ciudades adyacentes con 40.000 o más habitantes.

- **Resto de Área Urbana (RAU)**: Conformadas por conjuntos de Centros Urbanos con menos de 40.000 habitantes.

- **Área Rural (R)**: Conformado por el conjunto de entidades clasificadas como rurales de acuerdo a un tamaño poblacional menor a 1.000 habitantes, o entre 1.001 y 2.000 habitantes con predominio de población económicamente activa dedicada a actividades primarias[3].

En la segunda fase, la VI EME está estratificada de forma natural de acuerdo con las 16 regiones que posee el país.

I.6.3. Cobertura

La cobertura es una propiedad estadística que caracteriza al marco muestral. La falta de cobertura constituye el error de no incluir ciertos elementos (o unidades completas) de la población objetivo o de estudio según el marco muestral que se ha definido. Este error, en ocasiones no es planeado por el investigador. Un ejemplo de esto es la exclusión de ciertas unidades en el proceso de conteo e identificación de las viviendas, previo a la selección.

I.6.3.1. Cobertura geográfica

La VI EME, posee un diseño muestral bifásico, por lo tanto, comparte las propiedades de cobertura de dos marcos muestrales, primero el utilizado para la selección de las viviendas de la ENE (período MAM 2019); y segundo el marco utilizado para la selección de los "Microemprendedores".

El marco muestral del INE, utilizado como base para la ENE y todas las encuestas de hogares que se levantan en la Institución, cubre sólo a la población que reside en viviendas particulares ocupadas y, por lo tanto, excluye a la población que habita en viviendas colectivas, tales como: hogares de ancianos, hospitales, cárceles, conventos, entre otros; y a la población que reside en la calle. Sin embargo, se incluye a los hogares de personas que habitan y trabajan dentro de dichos centros, como porteros, conserjes y otros.

Además, el marco muestral de la ENE excluye las viviendas ubicadas en las 22 áreas geográficas definidas por el INE como áreas de difícil acceso (ADA'S) o de alto costo. Por otro lado, para optimizar el trabajo de campo y dadas las características de las unidades muestrales del área urbana (manzanas) se descartan del marco muestral, previo a la selección de las unidades, las manzanas que contienen 7 o menos viviendas.

Finalmente, en la elaboración del marco muestral de la VI EME, se excluyen intencionadamente todas las viviendas que no poseen un "Microemprendedor", es decir, que no poseen unidades elegibles.

[3] Toda aquella actividad relacionada con la extracción de recursos naturales (agricultura, caza, pesca, minería, entre otras).

Cabe señalar que, para fines de análisis y ajustes de los factores de expansión, las regiones fueron agrupadas en cuatro macrozonas: Norte, Centro, Sur, y Región Metropolitana. En el Cuadro I.2 se detalla la composición de cada macrozona.

Cuadro I.2. Composición de macrozonas.

Macrozona	Región
Norte	Arica y Parinacota
	Tarapacá
	Antofagasta
	Atacama
	Coquimbo
Centro	Valparaíso
	O'Higgins
	Maule
	Ñuble
	Biobío
Sur	La Araucanía
	Los Ríos
	Los Lagos
	Aysén
	Magallanes
Metropolitana	Metropolitana

Elaborado por el Instituto Nacional de Estadísticas.

I.7. Estrategia muestral

El diseño muestral de la VI EME es bifásico, es decir, se levanta una gran encuesta (ENE) para capturar información de una primera población objetivo, que sirve para identificar a los individuos que formarán parte de la segunda fase. La primera encuesta (primera fase) sirve de marco de muestreo, para la selección de unidades en la segunda fase.

La primera fase del diseño posee un muestreo probabilístico, estratificado y bietápico, donde las unidades primarias de muestreo corresponden a manzanas en el área urbana y secciones en el área rau - rural; mientras que, las unidades secundarias de muestreo son las viviendas particulares ocupadas.

Las unidades primarias se seleccionan en forma proporcional al tamaño dentro de cada estrato de muestreo, mientras que, las unidades secundarias se seleccionan de forma sistemática y con igual probabilidad. Así, las unidades seleccionadas y encuestadas en la ENE para el periodo MAM 2019, son

utilizadas como marco de muestreo para la VI EME pues, permite identificar las viviendas donde reside al menos un microemprendedor.

En la segunda fase, se clasifican las viviendas en dos grupos, de acuerdo a si éstas contienen o no, en el período de referencia, al menos un microemprendedor. Las viviendas que no poseen microemprendedores se descartan, formando el marco de muestreo con sólo aquellas viviendas con presencia de microemprendedores.

Debido a que en la primera fase, la ENE es contestada por un informante idóneo (proxy), quien responde por él y por todos los integrantes de su hogar, constituye una fuente de error no muestral de clasificación, propio de las encuestas de hogares, donde una persona pudiera ser clasificada como independiente en la ENE pero, que en la realidad no lo sea, o viceversa. Es por esto, que este listado se revisa en profundidad para descartar todas las unidades que no cumplen los criterios técnicos para ser clasificadas efectivamente como un microemprendimiento, mitigando así los problemas posteriores de levantamiento debido a la existencia de casos fuera de muestra.

Posterior a la depuración del listado, se seleccionan con igual probabilidad y de forma sistemática, las viviendas que formarán parte de la muestra. Luego, se listan todos los microemprendedores al interior de la vivienda y del hogar; y se seleccionan de forma aleatoria a un microemprendedor por tipo de actividad económica[4], al interior del hogar. Si al interior de algún hogar dentro de la vivienda existe más de un microemprendedor que desempeña una misma actividad económica, entonces sólo se selecciona a uno de ellos para efectos de no redundar en la misma información.

En la Tabla I.1, se presentan las variables “Total microemprendedores ENE”, correspondiente al total de personas clasificadas en la ENE como microemprendedores, en el período MAM 2019; junto con la variable “Total microemprendedores EME”, la cual hace referencia al universo de microemprendedores luego de la depuración de la base de la ENE, utilizado para la selección de la muestra en la VI EME. Además de la variable “Porcentaje depuración”, correspondiente al porcentaje de microemprendedores que fueron depurados en la base de la ENE para crear el Marco de selección para la EME.

En total, la depuración del marco corresponde a 8,8% de casos descartados por ser potenciales unidades no elegibles[5], observándose los mayores cambios en la región de Los Lagos (12,2%) y, los menores, en la región de Magallanes con un 5,6%.

[4] En la submuestra 1 se consideran como actividades distintas, todas las ramas clasificadas en grupos distintos a 4 dígitos, salvo la rama de agricultura. En la submuestra 2 y 3, la selección de informantes en actividades distintas se realiza a partir de caenes a 2 dígitos; debido a que, desde abril se deja de generar la variable caenes a 4.

[5] En la EME son unidades no elegibles aquellos individuos que en la ENE fueron clasificados como microemprendedores, según información proporcionada por informante proxy, sin embargo, al momento de realizar el trabajo de campo se observa que la persona seleccionada no era un microemprendedor, o también en el caso que el individuo haya cambiado de estado (dejó de ser microemprendedor).

Tabla I.1. Total de microemprendedores según ENE trimestre MAM 2019 y según Marco VI EME 2019

Macrozona	Región	Total Microemprendedores ENE	Total Microemprendedores EME	Porcentaje depuración
Total Nacional		**12.223**	**11.143**	**8,8%**
Norte	Arica y Parinacota	465	436	6,2%
	Tarapacá	453	420	7,3%
	Antofagasta	321	286	10,9%
	Atacama	365	333	8,8%
	Coquimbo	843	774	8,2%
Total Norte		**2.447**	**2.249**	**8,1%**
Centro	Valparaíso	1.461	1.339	8,4%
	O'Higgins	597	545	8,7%
	Maule	734	679	7,5%
	Ñuble	277	256	7,6%
	Biobío	1.037	947	8,7%
Total Centro		**4.106**	**3.766**	**8,3%**
Sur	La Araucanía	696	641	7,9%
	Los Ríos	480	448	6,7%
	Los Lagos	1.028	903	12,2%
	Aysén	466	423	9,2%
	Magallanes	251	237	5,6%
Total Sur		**2.921**	**2.652**	**9,2%**
Metropolitana	Metropolitana	2.749	2.476	9,9%
Total Metropolitana		2.749	2.476	**9,9%**

Elaborado por el Instituto Nacional de Estadísticas.

I.8. Cálculo y distribución del tamaño muestral

El tamaño muestral es concebido para obtener estimaciones del parámetro de interés con niveles de precisión establecidos para esta versión, que busca representatividad a nivel nacional y regional.

El cálculo del tamaño muestral se realiza con la información de la ENE levantada en el trimestre marzo, abril y mayo de 2018 (MAM 2018). Utiliza como parámetro de interés la proporción de microemprendedores por cuenta propia sobre el total de microemprendedores y sus estadísticos asociados.

Se trabajaron escenarios con similares errores de estimación, donde finalmente, se elige el escenario con un error absoluto de 1,3% y un error relativo de 1,5% a nivel nacional. Con esto, se obtiene una muestra objetivo de 7.086 viviendas que, considerando la tasa de no respuesta observada en la V EME 2017, da como resultado un tamaño con sobremuestreo de 8.469 viviendas.

Para obtener los tamaños regionales, se asigna el 70,0% de las viviendas disponibles en el marco de selección, donde reside al menos un microemprendedor, exceptuando las regiones de: Antofagasta, Ñuble y Magallanes, donde se asigna el 85,0% con el objetivo de disminuir el error muestral.

I.8.1. Consideraciones para el cálculo del tamaño muestral

El cálculo del tamaño muestral utiliza como parámetro de interés la proporción de microemprendedores por cuenta propia sobre el total de microemprendedores. En el Cuadro I.3, se presentan definiciones y consideraciones utilizadas en la obtención del tamaño muestral.

Cuadro I.3. Criterios utilizados en el cálculo del tamaño muestral

Parámetro	Descripción
Variable de diseño	I : Variable Bernoulli que considera los siguientes valores $I = \begin{cases} 1; & Si\ el \text{ microemprendedor es cuenta propia} \\ 0; & \text{En otro caso} \end{cases}$
Parámetro asociado	Proporción de microemprendedores por cuenta propia sobre el total de microemprendedores en la región r (p_{r0})
Estimador asociado	Estimador de razón: $r = \frac{Estimación\ del\ número\ de\ microemprendedores\ por\ cuenta\ propia\ en\ la\ región}{Estimación\ del\ número\ de\ microemprendedores\ en\ la\ región}$
Niveles de estimación	Nacional Regional
Errores de muestreo	- Nacional: El error absoluto no debe superar 1,3% y; el error relativo 1,5% - Regional: Los errores absolutos no deben superar 8,2% y; los errores relativos 11,0%

Elaborado por el Instituto Nacional de Estadísticas.

Sean:

r	: Subíndice que identifica a la región
p_{r0}	: Parámetro de interés en la región r
$\hat{p}_{r0}$	: Estimación del parámetro de interés en la región r
$\bar{m}_{r0}$	: Número promedio de viviendas a encuestar por unidad primaria de muestreo en la región r en la ENE trimestre MAM 2018
n_{r0}	: Número de manzanas o secciones levantadas en trimestre MAM 2018 (ENE) en la región r
$n_{r0} \cdot \bar{m}_{r0} = m_{r0}$	: Número de viviendas logradas en trimestre MAM 2018 (ENE) en la región r
λ	: Porcentaje de selección asignado para cada región (70,0% o 85,0%)

$SE(\hat{p}_{r0})$: Error estándar de la estimación de la proporción de microemprendedores por cuenta propia en la región r. Corresponde a la raíz cuadrada de la varianza de la estimación

$Z_{1-\alpha/2}$: Percentil de nivel $(1-\alpha/2)$ de la distribución Normal, correspondiente a una estimación intervalar de $(1-\alpha)$ de confianza

A continuación, se describe en detalle cada una de las etapas realizadas para definir el número de viviendas a encuestar en cada región r.

Etapa 1:

Utilizando la base de datos del levantamiento de la ENE del trimestre MAM 2018, se obtienen las estimaciones a nivel regional de: parámetro de interés regional ($\hat{p}_{r0}$), error estándar ($SE(\hat{p}_{r0})$) y total de unidades logradas (m_{r0}).

Etapa 2:

Se calculan los errores absolutos regionales (ea_{r0}) a partir de los errores estándar obtenidos en MAM 2018 ($SE(\hat{p}_r)$) estableciendo un nivel de confianza de 95% bajo una distribución normal ($Z_{1-\alpha/2}$). El cálculo se observa en la ecuación (1):

$$ea_{r0} = 1{,}96 \cdot SE(\hat{p}_{r0}) \tag{1}$$

Etapa 3

El tamaño muestral de viviendas a nivel regional (m_{r1}) es equivalente a 70,0% de las viviendas disponibles en el marco de muestreo, exceptuando las regiones de: Antofagasta, Ñuble y Magallanes, donde el tamaño corresponde a 85,0% de lo disponible producto de la alta variabilidad de la variable de interés. El cálculo del tamaño se observa en la ecuación (2):

$$m_{r1} = m_{r0} \cdot \lambda^{6} \tag{2}$$

Etapa 4

Una vez obtenidos los tamaños regionales, se aplica un factor que corresponde a la tasa de no respuesta (tnr) obtenida en la V EME, con la finalidad de salvaguardar la precisión de la estimación del parámetro de interés, debido a la posibilidad de no lograr el total de unidades a encuestar, por diversas razones, tales como: rechazos, moradores ausentes, edificaciones que no forman parte de la población objetivo, entre otras.

[6] Porcentaje de selección asignado, según se haya definido para cada región.

Para esta versión de la encuesta, se utiliza una tasa de no respuesta de 15,0% a nivel regional (según tasas obtenidas en la V EME) salvo para la Región Metropolitana, donde se utiliza 20,0%[7]. Luego, se obtiene el tamaño muestral con sobremuestreo, o ajustado por no respuesta, para cada región como se precisa en la ecuación (3).

$$m_{r2} = \frac{m_{r1}}{1 - tnr} \tag{3}$$

I.8.2. Errores esperados

La Tabla I.2 presenta los errores esperados, asociados a la estimación del parámetro de interés obtenido en la V EME, para un tamaño muestral objetivo de 7.086 viviendas[8] a nivel nacional.

Tabla I.2. Errores esperados asociados al parámetro de interés según tamaño muestral objetivo

Nivel	Estimación V EME	Error absoluto propuesto VI EME	Error relativo propuesto VI EME	Tamaño objetivo VI EME	Tamaño con sobremuestreo VI EME
Nacional	**86,0%**	**1,3%**	**1,5%**	**7.086**	**8.469**
Arica y Parinacota	85,1%	5,2%	6,1%	274	322
Tarapacá	93,7%	3,3%	3,5%	248	292
Antofagasta	81,3%	7,6%	9,4%	212	249
Atacama	87,1%	5,9%	6,8%	206	242
Coquimbo	87,3%	4,4%	5,0%	494	581
Valparaíso	82,1%	3,4%	4,1%	928	1.092
Metropolitana	86,9%	2,4%	2,8%	1.619	2.036
O´Higgins	82,1%	5,4%	6,5%	360	424
Maule	85,0%	4,7%	5,6%	430	506
Ñuble	87,4%	5,9%	6,7%	155	182
Biobío	88,4%	3,3%	3,7%	644	758
La Araucanía	87,7%	4,5%	5,2%	468	551
Los Ríos	86,3%	5,6%	6,5%	236	278
Los Lagos	83,9%	4,2%	5,0%	519	611
Aysén	83,7%	5,7%	6,9%	188	221
Magallanes	74,7%	8,2%	11,0%	105	124

Elaborado por el Instituto Nacional de Estadísticas.

[7] Para esta región se utiliza una tasa más alta debido a que, en el levantamiento de la V EME, en las primeras 2 Submuestras se observaron niveles de no respuesta mayores al 20%.

[8] La estimación de los errores se obtiene a partir del tamaño objetivo y no del tamaño con sobremuestreo, pues dadas las pérdidas naturales de unidades muestrales, se espera obtener un tamaño de 7.086 viviendas.

I.8.3. Distribución de la muestra según submuestra

Como la muestra ENE está subdividida en tres meses o períodos de levantamiento, se debe disminuir el tiempo transcurrido entre el levantamiento de la ENE y de la EME para así tener una menor atrición[9] de la muestra.

Una vez obtenido el tamaño muestral nacional y regional, la muestra de la VI EME, se distribuyó en tres partes iguales, o lo más similares posibles, en los siguientes meses de levantamiento: mayo, junio, julio. La Tabla I.3 muestra la distribución de la muestra según mes de levantamiento y región.

Tabla I.3. Total de viviendas seleccionadas según región y mes de levantamiento

	Mes de levantamiento VI EME			
Región	**Mayo**	**Junio**	**Julio**	**Viviendas seleccionadas**
Nacional	**2.819**	**2.819**	**2.831**	**8.469**
Arica y Parinacota	107	108	107	322
Tarapacá	97	99	96	292
Antofagasta	84	80	85	249
Atacama	80	82	80	242
Coquimbo	193	194	194	581
Valparaíso	364	364	364	1.092
Metropolitana	675	674	687	2.036
O´Higgins	142	141	141	424
Maule	168	169	169	506
Ñuble	61	60	61	182
Biobío	253	253	252	758
La Araucanía	183	179	189	551
Los Ríos	93	93	92	278
Los Lagos	204	205	202	611
Aysén	73	75	73	221
Magallanes	42	43	39	124

Elaborado por el Instituto Nacional de Estadísticas.

[9] Acumulación de pérdida de información por la no respuesta que se presenta en estudios sobre unidades en el tiempo por parte de los participantes (Jones, Koolman, & Rice, 2006).

I.8.4. Errores observados

La Tabla I.4 presenta los errores observados, a nivel nacional y regional, asociados a la prevalencia del parámetro de interés, para un tamaño muestral efectivo de 7.301 viviendas, según levantamiento de la VI EME.

Tabla I.4. Errores observados asociados al parámetro de interés según número de viviendas que responden[10]

Región	Estimación VI EME	Tamaño efectivo	Error absoluto	Error relativo
Nacional	**85,5%**	**7.301**	**1,2%**	**1,4%**
Arica y Parinacota	85,6%	278	5,6%	6,5%
Tarapacá	88,3%	281	5,3%	5,9%
Antofagasta	85,3%	233	9,1%	10,7%
Atacama	87,3%	227	4,8%	5,5%
Coquimbo	85,3%	538	3,6%	4,3%
Valparaíso	86,5%	946	3,0%	3,5%
Metropolitana	84,9%	1.583	2,3%	2,7%
O'Higgins	88,8%	368	4,3%	4,8%
Maule	88,3%	458	2,8%	3,2%
Ñuble	81,0%	154	4,5%	5,6%
Biobío	81,3%	662	5,9%	7,2%
La Araucanía	89,6%	483	2,7%	3,0%
Los Ríos	87,7%	253	5,6%	6,4%
Los Lagos	82,7%	535	2,7%	3,2%
Aysén	80,9%	196	6,0%	7,5%
Magallanes	76,3%	106	7,1%	9,3%

Elaborado por el Instituto Nacional de Estadísticas.

I.9. Selección de unidades muestrales

La selección de unidades muestrales se realiza en 2 etapas. Primero sobre las viviendas que contienen al menos un microemprendedor y luego, al interior se seleccionan los microemprendedores.

I.9.1. Selección de viviendas

La selección de viviendas se realiza de forma sistemática con igual probabilidad de selección, para todas las viviendas al interior de cada región.

[10] La estimación de los errores se obtiene a partir del tamaño efectivo, utilizando como variable de estratificación pseudo estratos (Varstrat) y variable de conglomeración pseudo conglomerados (Varunit),

La selección es implementada en el *software* estadístico SPSS, en el módulo de análisis "Muestras Complejas" específicamente, en el procedimiento "seleccionar una muestra", fijando una semilla aleatoria, a fin de que pueda ser replicable en cualquier momento.

Sean:

m_r : Número de viviendas a seleccionar en la región r.

M_r : Número de viviendas que contiene la región r.

Para la selección de m_r viviendas, el *software* ejecuta los siguientes pasos:

Paso 1

En primera instancia, se ordenan en forma ascendente todas las viviendas según región, estrato, área, comuna, sección, distrito censal, zona censal, número de manzana y número de orden de vivienda dentro de la manzana.

Paso 2

Al interior de cada región se calcula el período (k) que corresponde a:

$k = M_r/m_r$.

Notar que "k" puede ser un número real, no entero (puede tener decimales).

Paso 3

Luego se determina el arranque "A" o primera selección, que corresponde a una semilla aleatoria propia para la encuesta.

Paso 4

Posteriormente se suma sucesivamente el período "k" al arranque "A" para obtener distintos valores, los que dan origen a la selección de unidades de la siguiente forma: "A", "$A + k$", "$A + 2k$", "$A + 3k$", …, "$A + (m_r - 1)k$".

La primera vivienda seleccionada es "A" y es un número entero, la segunda es el redondeo de "$A + k$", la tercera es el redondeo de "$A + 2k$" y así sucesivamente, hasta la m_r selección, dada por el redondeo de "$A + (m_r - 1)k$".

Luego, la probabilidad de inclusión de la $j - ésima$ vivienda dentro de la $r - ésima$ región, es igual a:

$$P_r(j|r) = \frac{m_r}{M_r} \quad (4)$$

I.9.2. Selección de microemprendedores

Una vez seleccionadas las viviendas, dentro de cada una de ellas y al interior de sus hogares, se identifica a los microemprendedores y las actividades económicas en que estos se desenvuelven. Luego, se seleccionan de forma aleatoria y con igual probabilidad, tantos microemprendedores como actividades distintas identificadas dentro del hogar, es decir, en caso de encontrar más de un microemprendedor en el hogar ejecutando la misma actividad económica, se debe seleccionar sólo a un representante por actividad dentro del hogar.

II. DESARROLLO DE FACTORES DE EXPANSIÓN

El factor de expansión se interpreta como la cantidad de unidades en la población que representa una unidad de la muestra, y es calculado como el inverso de la probabilidad de selección de las unidades de muestreo. Atendiendo al diseño muestral de la VI EME, las probabilidades de selección asociadas a los microemprendedores tienen varias componentes:

1. Probabilidad de que la vivienda haya sido seleccionada y contestara en la ENE en el trimestre MAM 2019.
2. Probabilidad de seleccionar una vivienda para la EME, dado que la vivienda posee al menos un microemprendedor.
3. Probabilidad de seleccionar un microemprendedor, dado que su vivienda fue seleccionada.

La metodología de cálculo de los factores de expansión consiste en la aplicación secuencial de cinco ponderadores o ajustes:

1. Ponderador Base.
2. Suavizamiento del ponderador base.
3. Ajuste por no respuesta.
4. Suavizamiento del ajuste por no respuesta.
5. Calibración.

II.1. Ponderador Base

El ponderador Base se define como el factor de expansión obtenido de las probabilidades de selección, sin ajustes ni correcciones, de las viviendas en la fase 1, y la selección de los microemprendedores en la fase 2, condicional a que la vivienda de residencia fue seleccionada en la ENE y que éstas participaron en el trimestre MAM 2019.

En la VI EME, las personas seleccionadas corresponden a un subconjunto de personas que participaron durante el proceso de encuestaje del trimestre MAM 2019 de la ENE. Por lo tanto, uno de los insumos fundamentales del ponderador base, son los factores de expansión de las viviendas de la ENE, que dan cuenta de la probabilidad de que una vivienda haya sido seleccionada y entrevistada en la encuesta. En la sección II.1.1 expone las probabilidades de selección y respuesta de la ENE; mientras que la sección II.1.2 expone la formula explicita de la probabilidad condicional de selección de un microemprendedor.

II.1.1. Probabilidad de selección y entrevista de las viviendas en la ENE- Trimestre MAM 2019

El diseño muestral de la Encuesta Nacional de Empleo corresponde a un muestreo probabilístico, estratificado y bietápico, donde las unidades primarias son las manzanas en el área urbana y secciones en el área rau - rural; mientras que, las unidades de segunda etapa son las viviendas particulares.

Las unidades primarias fueron seleccionadas en forma proporcional al tamaño, mientras que al interior de cada manzana o sección las unidades de segunda etapa se seleccionaron de forma sistemática y con igual probabilidad de selección. El factor de expansión de la ENE posee un ajuste por no respuesta implícito, es decir, el peso de las unidades que no responden es distribuido en el resto de las viviendas del conglomerado al cual pertenecen. La ecuación (5) corresponde al ponderador inicial o teórico corregido por no respuesta de la ENE.

$$F_{hij}^{1} = \underbrace{\left(\frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}^{T}}\right)}_{\substack{Factor\ de\ expansión \\ teórico}} \cdot \overbrace{\frac{m_{hi}^{T}}{\left(m_{hi}^{T} - m_{hi}^{NR}\right)}}^{\substack{Ajuste \\ no\ respuesta}} = \frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}} \qquad (5)$$

Donde:

h : Subíndice que representa el estrato de muestreo ENE

i : Subíndice que representa el conglomerado i

j : Subíndice que representa la vivienda j

M_h : Total de viviendas en el estrato h, según el Marco de muestreo de la ENE

n_h : Total de conglomerados seleccionados en el estrato h en la ENE

M_{hi} : Total de viviendas particulares que contiene el conglomerado i del estrato h, según información del Marco muestral

M'_{hi} : Total de viviendas particulares que contiene el conglomerado i del estrato h , según información recogida en enumeración

m_{hi}^{T} : Total de viviendas seleccionadas en el conglomerado i del estrato h

m_{hi}^{NR} : Total de viviendas seleccionadas en el conglomerado i del estrato h que no responden

m_{hi} : Total de viviendas que responden en la ENE en el trimestre MAM 2019

En consecuencia, la probabilidad de haber sido seleccionada y entrevistada la $j-ésima$ vivienda, del conglomerado i, en el estrato h en el trimestre móvil MAM 2019 en la ENE, se representa mediante la siguiente ecuación (6):

$$P_{hij}^{v} = \frac{1}{F_{hij}^{1}} \tag{6}$$

II.1.2. Probabilidad de selección de los microemprendedores

La selección de los microemprendedores se realiza en dos etapas. Primero, se seleccionan con igual probabilidad las viviendas que contienen al menos un microemprendedor, según submuestra (mes de levantamiento ENE) al interior de cada región.

Así, la probabilidad de seleccionar una vivienda que posee al menos un microemprendedor viene dada por la ecuación (7):

$$p_{Rj}^{v} = \frac{m_{R}^{micro}}{M_{R}^{micro}} \tag{7}$$

Donde:

R : Subíndice que representa la región de pertenencia. $R = 1, \dots, 16$.

j : Subíndice que representa la vivienda j

p_{Rj}^{v} : Probabilidad de seleccionar la vivienda j perteneciente a la región R, según el listado de viviendas del marco EME, que poseen al menos un microemprendedor

M_{R}^{micro} : Total de viviendas seleccionadas del marco EME con al menos un microemprendedor en la región R

m_{R}^{micro} : total de viviendas seleccionadas del marco de la ENE con al menos un microemprendedor en la región R

Luego, una vez seleccionada la vivienda se seleccionan los microemprendedores. La probabilidad de seleccionar al microemprendedor k al interior de la vivienda j, del hogar l y rama de actividad m, perteneciente a la región R, dado que la vivienda fue seleccionada, está determinada por la ecuación (8):

$$p_{Rjklm}^{micro|v} = \frac{S_{Rjlm}^{micro}}{T_{Rjlm}^{micro}} \quad (8)$$

Donde:

T_{Rjlm}^{micro} : Total de microemprendedores identificados en la EME, en la vivienda j, hogar l, rama de actividad m, perteneciente a la región R

S_{jlm}^{micro} : Total de microemprendedores seleccionados, en la vivienda j, hogar l, rama de actividad m, perteneciente a la región R

Luego la probabilidad condicional de seleccionar el microemprendedor k, en la vivienda j, de la región R, viene dada por la ecuación (9):

$$p_{Rjk}^{micro} = p_{Rj}^{v} \cdot p_{Rjklm}^{micro|v} \quad (9)$$

Así, calculadas las probabilidades de selección y participación de una vivienda en la ENE en el trimestre MAM 2019 y, la probabilidad de seleccionar un microemprendedor desde la EME, el ponderador base está definido por la ecuación (10):

$$F_{Rjk}^{base} = \left(\frac{1}{P_{hij}^{v}}\right) \cdot \left(\frac{1}{p_{Rjk}^{micro}}\right) \quad (10)$$

En la Tabla II.1, se observa que tanto las ramas de Comercio y Servicios presentan ponderador base sobre 8.000, seguido de la rama de Transporte y Almacenamiento con valores sobre 4.000.

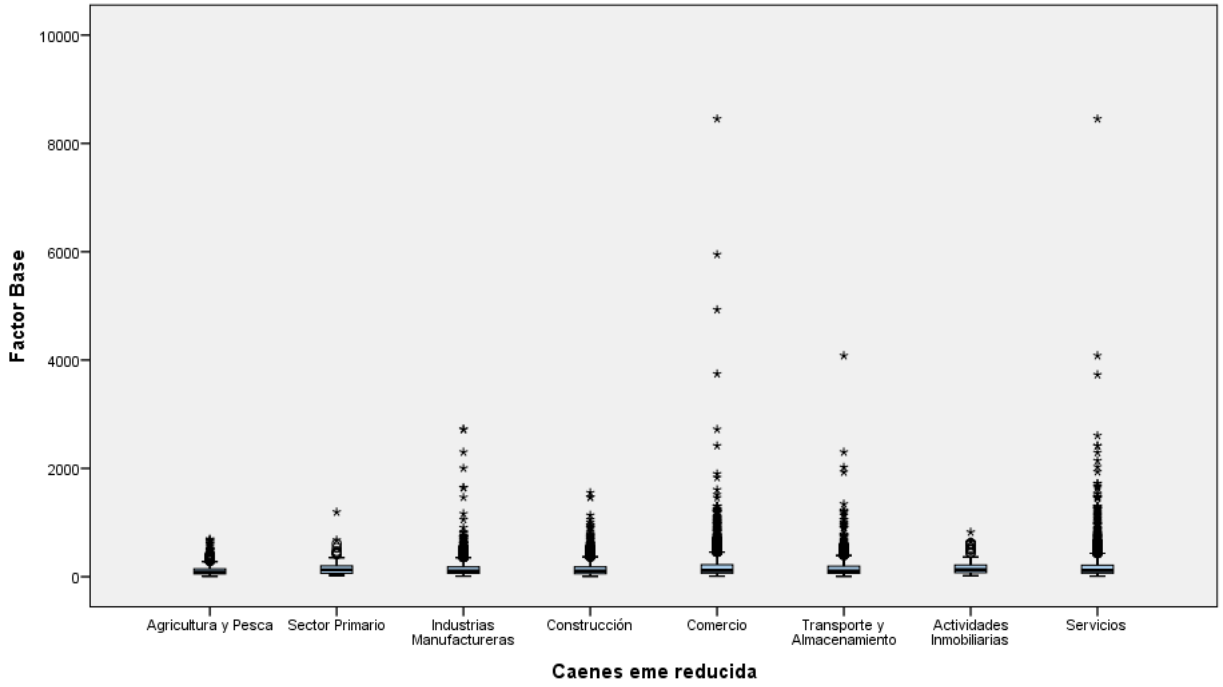
Tabla II.1. Estadísticas descriptivas del ponderador base según rama de actividad reducida[11]

Estadísticas Descriptivas	Rama de actividad reducida								
	Agricultura y Pesca	Sector Primario	Industrias Manufactureras	Construcción	Comercio	Transporte y Almacenamiento	Actividades Inmobiliarias	Servicios	Total
Recuento	1.295	57	1.132	1.115	2.367	829	89	2.343	9.227
Moda	69,5	137,2	733,4	33,0	41,4	60,7	90,7	202,4	77,3
Mínimo	7,0	24,7	10,4	7,9	9,1	7,9	19,7	9,1	7,0
Percentil 05	19,7	32,2	27,8	25,0	26,9	27,0	38,0	27,5	26,2
Percentil 25	53,9	62,6	61,9	57,1	64,2	60,7	73,8	66,1	61,1
Mediana	84,8	128,0	104,0	105,9	115,0	105,6	128,4	113,0	107,0
Percentil 75	144,7	200,6	182,1	183,0	221,8	194,1	215,0	213,3	192,7
Percentil 95	264,1	609,0	483,3	501,3	611,7	544,5	611,7	616,5	520,2
Percentil 99	501,3	1.194,3	815,6	895,3	1.197,6	1.166,8	824,0	1.299,1	1.095,2
Máximo	694,8	1.194,3	2.721,5	1.552,6	8.453,9	4.082,2	824,0	8.453,9	8.453,9
Media	110,2	188,9	162,7	158,1	196,7	177,7	195,1	196,6	173,9
Error estándar de la media	2,5	27,3	6,3	5,2	6,7	8,9	18,8	6,5	2,8
Suma	142.685,8	10.767,2	184.223,2	176.228,0	465.502,9	147.319,2	17.364,5	460.653,6	1.604.744,3

Elaborado por el Instituto Nacional de Estadísticas.

En la Figura II.1, también se logra observar que las dos ramas antes mencionadas tienen el ponderador más alto, siendo hasta diez veces más grande que los valores extremos de las ramas restantes.

Figura II.1 Ponderador base según rama de actividad económica reducida



Elaborado por el Instituto Nacional de Estadísticas.

[11] Agrupación de las categorías de baja prevalencia en la rama de actividad económicas (caenes_1d_eme). Revisar Tabla III.3.

II.2. Suavizamiento del Ponderador Base

Debido a la gran dispersión de los valores del ponderador, es necesario la incorporación de un procedimiento de suavizado de los factores. Dicho procedimiento consiste en truncar los factores de expansión que se ubican por encima de un umbral definido (M_ha), para luego redistribuir la diferencia resultante entre los factores asociados a las unidades que se encuentran por debajo de este umbral de manera proporcional. Para ello, se requiere como insumo el principal parámetro de interés a medir, que permita identificar el umbral de suavizamiento a través de la minimización de su error cuadrático medio.

En este contexto, la variable de interés corresponde a "Microemprendedor por cuenta propia" que se obtiene directamente de los resultados de la VI EME 2019.

Una vez obtenida la estimación, se procede a realizar el suavizamiento del factor a partir de la implementación de las siguientes etapas:

1. Se inspecciona la existencia de valores extremos de éste, al interior de cada una de las ramas de actividad reducida.
2. Se determina puntos de corte a partir de los cuales realizar el suavizamiento.
3. Se suaviza los valores extremos identificados.
4. Se estima el error cuadrático medio (ECM) para el parámetro de interés para los distintos puntos de corte.
5. Se elige la opción de corte que minimiza dicho error[12].

Considerando lo anterior, se analizan 7 puntos de cortes distintos, definidos en la ecuación (11):

$$\beta_c = c\,.\,\overline{F}\,,\ \ con\ \overline{F} = \overline{F}_{Rjk}^{base}\,, para\ c\ =\ \ 4,5,6,7,8,9,10. \tag{11}$$

Donde:

β_c : Producto entre el punto de corte y la media del ponderador base a nivel de rama reducida.

c : Punto de corte en el cual se prueba el suavizamiento.

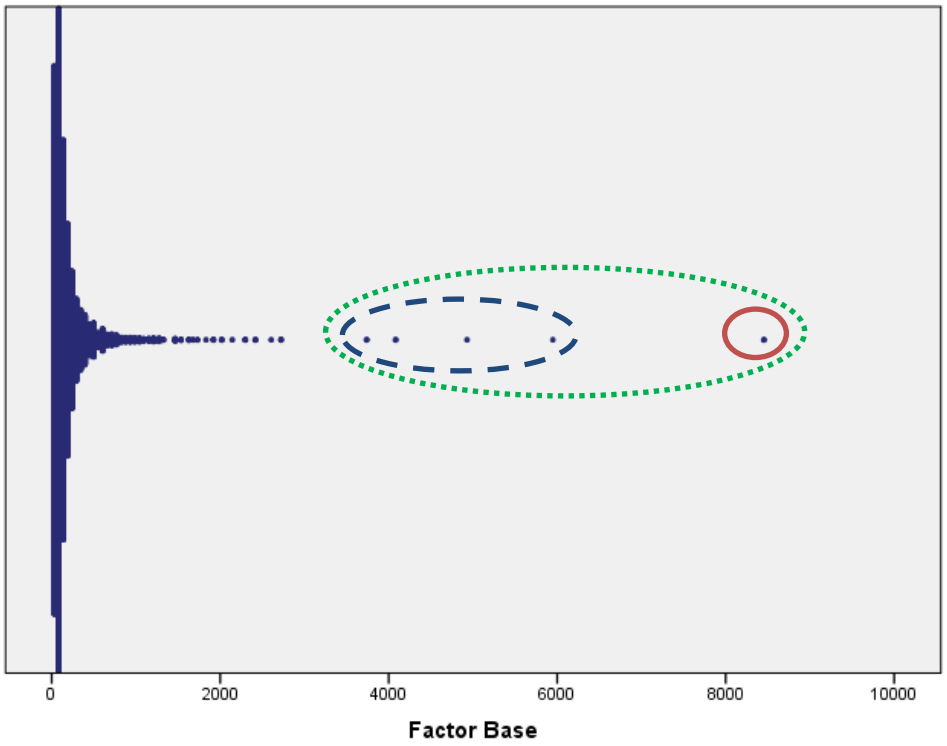
$\overline{F}$: Promedio del ponderador base en la rama reducida *h*

La Figura II.2 muestra la dispersión del ponderador base, donde se observan los factores de expansión base de forma ordenada, permite identificar al menos cinco puntos de discontinuidad del ponderador base: un caso extremo (que se encuentra al interior de la elipse continua) que supera las 8.000 unidades

[12] En el caso que el ECM sea igual para más de un corte, se procederá a elegir el corte que trunque más valores y, que al mismo tiempo, luego del suavizamiento, tenga menos valores que sobrepasen el umbral establecido (M_ha).

y cuatro ponderadores, que se encuentran al interior de la elipse semi-continua, que poseen valores superiores a 3.800.

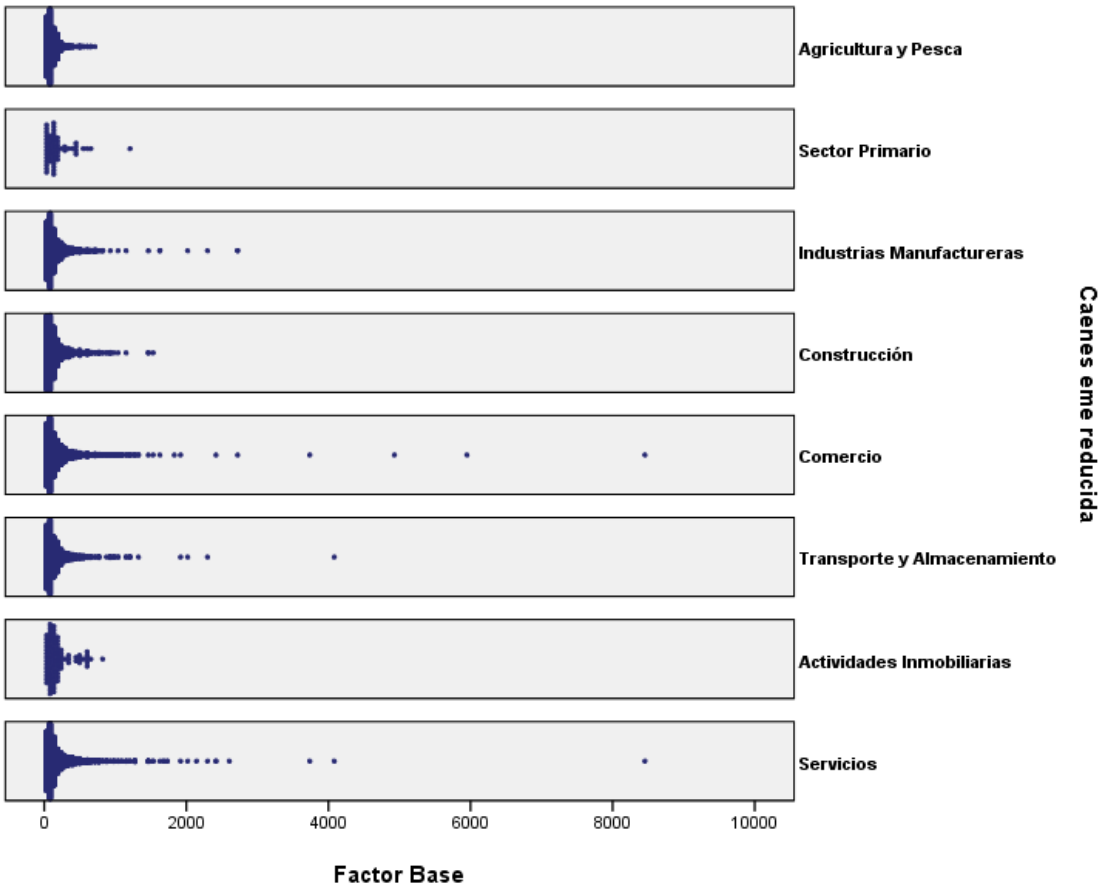
Figura II.2 Dispersión del ponderador base o inicial



Elaborado por el Instituto Nacional de Estadísticas

De igual forma, se realiza el análisis según rama de actividad económica reducida. En la Figura II.3, se identifican discontinuidades en las ramas de Comercio, con al menos 3 cortes; en la rama de Servicios, seguidos de Transporte y Almacenamiento; y también en la rama de Industria Manufactureras.

Figura II.3 Dispersión del Factor de expansión base o inicial, según rama reducida



Elaborado por el Instituto Nacional de Estadísticas

Luego, para realizar el suavizamiento se procede a truncar aquellos ponderadores identificados como valores extremos según la siguiente regla de decisión:

$$T_{Rjk,g} = \begin{cases} F_{Rjk}^{base} & si \;\; F_{Rjk}^{base} \leq \beta_c \\ \beta_c & si \;\; F_{Rjk}^{base} > \beta_c \end{cases} \tag{12}$$

Al sumar todos los valores $T_{Rjk,g}$, se obtiene un total de unidades estimadas inferior que al sumar los ponderadores base, por lo tanto se debe distribuir la diferencia faltante en el resto de los ponderadores que no fueron truncados. Los pesos fueron distribuidos al interior de cada rama reducida tal como indica la ecuación (13):

$$F_{Rjk}^{Sr} = \begin{cases} F_{Rjk}^{base} \cdot \dfrac{\left(\sum_{k \in g} F_{Rjk}^{base} - \sum_{\in g \cap F_{Rjk}^{base} > \beta_c} \beta_c\right)}{\sum_{k \in g \cap F_{Rjk}^{base} \leq \beta_c} F_{Rjk}^{base}} & , \; si \;\; F_{Rjk}^{base} \leq \beta_c \\ \beta_c & , si \;\; F_{Rjk}^{base} > \beta_c \end{cases} \tag{13}$$

Donde:

$F_{Rjk,g}^{Sr}$: Factor de expansión suavizado del individuo k de la vivienda j en la región R de la Rama de Actividad Económica Reducida g

Esto es, aquellos ponderadores identificados como valores extremos son truncados al valor máximo establecido (β_c), mientras que el peso "sobrante" de los ponderadores truncados es distribuido sobre el resto de los ponderadores.

Luego, para determinar el punto de corte donde se realiza finalmente el suavizamiento, se calcula un estadígrafo que da cuenta del sesgo y de la variabilidad. Este estadígrafo corresponde al Error Cuadrático Medio (ECM) asociado al parámetro de interés. De esa forma el sesgo y el ECM, respectivamente, se calculan como:

$$sesgo\left(\hat{P}_{c_p}\right) = P_{base} - \hat{P}_{c_p} \tag{14}$$

$$ECM\left(\hat{P}_{c_p}\right) = Sesgo^2\left(\hat{P}_{c_p}\right) + Var\left(\hat{P}_{c_p}\right) \tag{15}$$

Donde:

P_{base} : La proporción de microemprendedores por cuenta propia en la rama reducida g obtenida con el factor de expansión sin truncar F_{Rjk}^{base}.

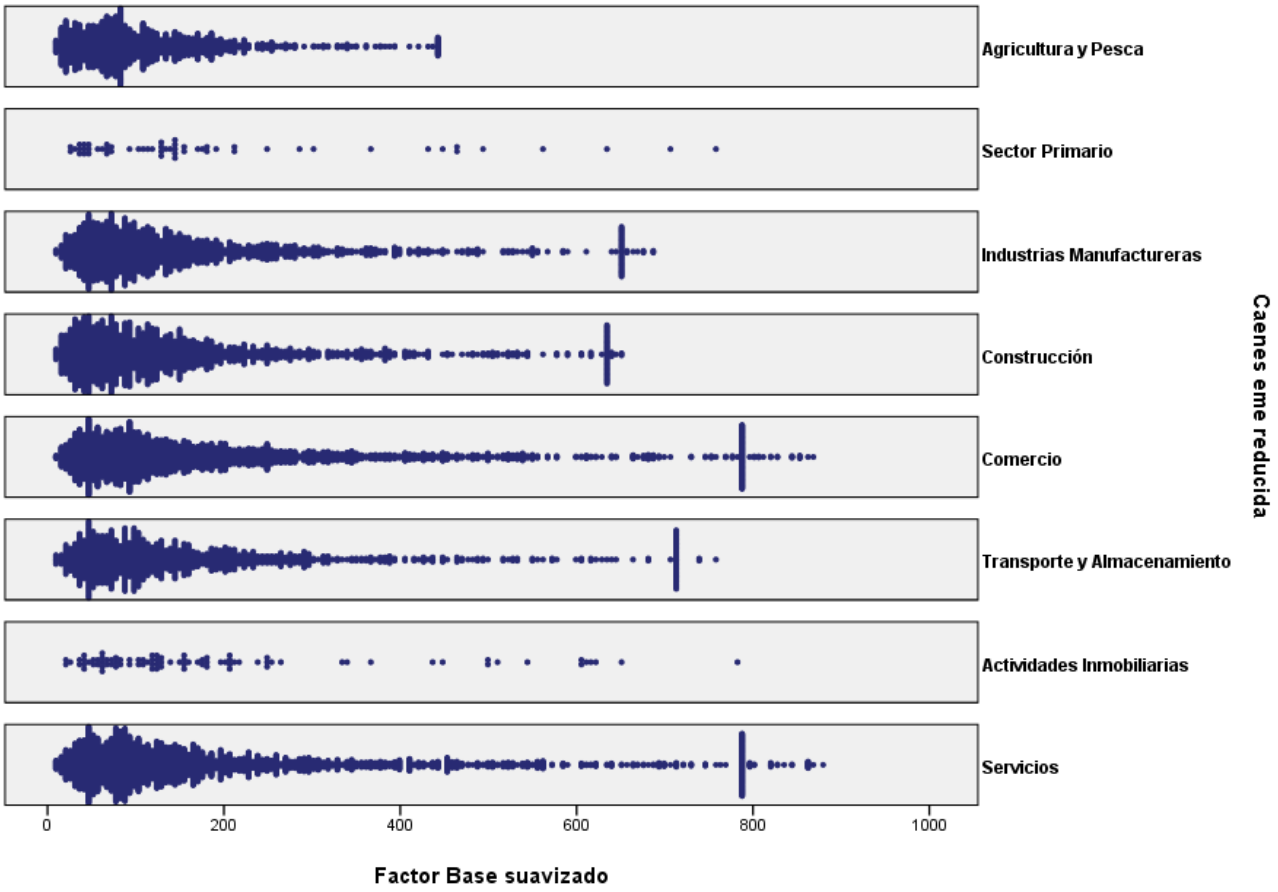
$\hat{P}_{c_p}$: La proporción de microemprendedores por cuenta propia en la rama reducida g obtenida con el factor suavizado en el umbral a.

Para entender el corte elegido al que se suavizan los factores de expansión, se procede de la siguiente forma:

1. Se calcula la mediana de los errores cuadráticos medios obtenidos con todos los cortes, desde $c4$ a $c10$.
2. Se identifica el mínimo de estas medianas entre todos los cortes.
3. El corte que contenga la mediana con valor más bajo es el que se utiliza para realizar el suavizamiento. Resultando para este caso, el corte más conveniente, c $= 4$.

En la Figura II.4, se observa cómo quedan los factores en las ramas de actividad, luego del suavizamiento.

Figura II.4 Dispersión del suavizamiento del Factor de expansión base o inicial, según rama reducida



Elaborado por el Instituto Nacional de Estadísticas

En la Tabla II.2 se observa que, en términos de distribución, al comparar el ponderador base versus el ponderador suavizado, las ramas que presentaban valores extremos fueron suavizados. El mayor cambio se encuentra en las ramas de Comercio y en la Rama de Servicios, ambas con un valor máximo de 8.453 disminuyendo a 867 y 879 unidades respectivamente. Seguidos de la rama de Transporte y Almacenamiento con un valor máximo de 4.082 disminuyendo a 760 unidades.

Tabla II.2. Estadísticas descriptivas del ponderador base y ponderador suavizado, según rama reducida

Estadísticas Descriptivas	Rama reducida																	
	Agricultura y Pesca		Sector Primario		Industrias Manufactureras		Construcción		Comercio		Transporte y Almacenamiento		Actividades Inmobiliarias		Servicios		Total	
	Factor Base	Factor $r4$	Factor Base	Factor $r4$	Factor Base	Factor $r4$	Factor Base	Factor $r4$	Factor Base	Factor $r4$	Factor Base	Factor $r4$	Factor Base	Factor $r4$	Factor Base	Factor $r4$	Factor Base	Factor r4
Recuento	1.295	1.295	57	57	1.132	1.132	1.115	1.115	2.367	2.367	829	829	89	89	2.343	2.343	9.227	9.227
Moda	69,5	440,7	137,2	143,5	733,4	651,0	33,0	632,2	41,4	786,7	60,7	710,8	90,7	90,9	202,4	786,4	77,3	786,7
Mínimo	7,0	7,1	24,7	25,9	10,4	11,3	7,9	8,3	9,1	10,2	7,9	8,8	19,7	19,8	9,1	10,2	7,0	7,1
Percentil 05	19,7	20,0	32,2	33,6	27,8	30,1	25,0	26,1	26,9	30,0	27,0	30,1	38,0	38,1	27,5	30,9	26,2	28,3
Percentil 25	53,9	54,8	62,6	65,5	61,9	66,9	57,1	59,8	64,2	71,6	60,7	67,8	73,8	74,0	66,1	74,4	61,1	66,4
Mediana	84,8	86,3	128,0	133,9	104,0	112,5	105,9	111,0	115,0	128,4	105,6	117,9	128,4	128,7	113,0	127,2	107,0	115,4
Percentil 75	144,7	147,2	200,6	209,8	182,1	196,9	183,0	191,7	221,8	247,5	194,1	216,6	215,0	215,6	213,3	240,1	192,7	208,5
Percentil 95	264,1	268,5	609,0	636,9	483,3	522,6	501,3	525,1	611,7	682,5	544,5	607,8	611,7	613,3	616,5	694,1	520,2	564,8
Percentil 99	501,3	440,7	1.194,3	755,6	815,6	651,0	895,3	632,2	1.197,6	786,7	1.166,8	710,8	824,0	780,4	1.299,1	795,7	1.095,2	786,7
Máximo	694,8	440,7	1.194,3	755,6	2.721,5	685,1	1.552,6	652,5	8.453,9	867,9	4.082,2	760,4	824,0	780,4	8.453,9	879,5	8.453,9	879,5
Media	110,2	110,2	188,9	188,9	162,7	162,7	158,1	158,1	196,7	196,7	177,7	177,7	195,1	195,1	196,6	196,6	173,9	173,9
Error estándar de la media	2,5	2,3	27,3	23,8	6,3	4,4	5,2	4,4	6,7	3,9	8,9	5,8	18,8	18,6	6,5	3,9	2,8	1,8
Suma	142.685,8	142.685,8	10.767,2	10.767,2	184.223,2	184.223,2	176.228,0	176.228,0	465.502,9	465.502,9	147.319,2	147.319,2	17.364,5	17.364,5	460.653,6	460.653,6	1.604.744,3	1.604.744,3

Elaborado por el Instituto Nacional de Estadísticas.

Posteriormente, utilizando como insumo el ponderador base suavizado, se realiza el ajuste por falta de respuesta, el cual se detalla en el siguiente apartado.

II.3. Ponderador ajustado por falta de respuesta

En esta etapa solo son consideradas las viviendas elegibles y que cumplen con las características de pertenecer a la población objetivo. Sin embargo, se presentan algunos casos, en los que residentes de las viviendas consideradas elegibles manifiesten su voluntad de no participar en la encuesta o que existan algunas viviendas en que no se encuentren moradores al momento del levantamiento, entre otras cosas, dando origen a la no respuesta.

En la VI EME la información recabada, corresponde a los microemprendedores, por lo tanto, la ausencia de sus respuestas debe ser corregida con la finalidad de reducir sesgos provocados por este tipo de errores no muestrales. Sin embargo, se debe señalar que, la ausencia de información se corrige sólo para algunos casos, es decir cuando, el informante rechazó la entrevista; la vivienda de residencia del informante se encuentra sin moradores presentes en todas las visitas efectuadas; entre otras.

Según lo anterior, surge la necesidad de aplicar un ajuste, con el objetivo de lograr que las unidades que no responden sean representadas por las que sí, previendo no introducir sesgo debido a la posibilidad de que exista relación entre la no respuesta y la variable de interés. De un total de 8.469 viviendas seleccionadas, se seleccionaron 9.227 microemprendedores. De éstos, 8.857 fueron clasificados como elegibles (96,0%), de los cuales 7.808 respondieron la encuesta[13]. Por lo tanto, la tasa de respuesta de la EME, ajustada por elegibilidad, es de 88,2%.

El método a implementar para compensar la falta de respuesta es el método de estratificación mediante "propensity score". De acuerdo, a lo indicado por Valliant (Valliant, Dever, & Kreuter, 2013), este método consiste en modelar la probabilidad de respuesta en la VI EME como la realización de un proceso de variables latentes ($R_i^* = x_i^T\beta + u_i$), es decir, un conjunto de variables que inciden en la "motivación" (R^*) de participar de una unidad (de responder). Así, mediante un conjunto de variables conocidas para quienes responden y quienes no responden se busca estimar la probabilidad de responder en la encuesta ($P(R_i^* > \theta)$).

El modelo que se utiliza para realizar el ajuste por no respuesta es el modelo logístico, el cual se resume en los siguientes pasos:

1. Determinar las variables que se incluirán en el modelo de regresión logística con el cual se realizará la predicción de la probabilidad de respuesta de una persona elegible.
2. A través de este modelo, calcular la probabilidad de responder de cada una de las unidades elegibles que fueron utilizadas.

[13] Mayor detalle ver Anexo N° 1

3. Ordenar las unidades de menor a mayor, según la probabilidad estimada.
4. Crear los estratos o "celdas de ajustes" donde se realizarán las correcciones de no respuesta.

Una vez creadas las celdas de ajustes[14], se procede a estimar el ponderador ajustado por falta de respuesta, el cual está dado por la ecuación (16):

$$\widehat{R}_c^{NR} = \frac{\sum_{k \in S_c} F_{Rjk}^{base_{tr}}}{\sum_{k \in S_{c,R}} F_{Rjk}^{base}} \quad (16)$$

Donde:

c : Subíndice de la celda de ajuste por falta de respuesta. $c = 1, \ldots, 6$

S_c : Total de microemprendedores seleccionados y elegibles en la celda c

$S_{c,R}$: Total de microemprendedores seleccionados en la celda c y que responde la encuesta.

F_{Rjk}^{base} : Corresponde al Ponderador base para la persona *k,* de la vivienda *j,* en la región *R.*

Así, la expresión del ponderador ajustado por no respuesta, está dada por la ecuación (17):

$$F_{Rjk}^{NR} = F_{Rjk}^{base_{tr}} \cdot \hat{R}_c^{NR} \quad (17)$$

De acuerdo a la metodología antes expuesta, son 6 las celdas en las cuales se realizan los ajustes por falta de respuesta. En la Tabla II.3 se presentan las tasas de respuesta para cada una de estas celdas, así como también el factor de ajuste por no respuesta ($\hat{R}_c^{NR}$). Se observa que el grupo 1 presenta menor tasa de respuesta, por lo que cada factor base fue "abultado" en 60% aproximadamente.

[14] Las celdas de ajuste son varias agrupaciones de individuos que poseen características similares asociadas a responder la encuesta. En el caso específico de la EME, se generaron 6 celdas de ajuste denominadas "sextiles" de respuesta, mediante un modelo logístico, en que se predice la probabilidad de responder, y en base a esta probabilidad que se ordena en forma ascendente o descendente, se generan seis grupos de igual cantidad de individuos ordenados por esta probabilidad.

Tabla II.3. Total de unidades elegibles, que responde y tasa de respuesta

Celda Ajuste	Total Elegibles	Total Responde	Tasa de Respuesta	$\hat{R}_c^{NR}$
Total	**8.857**	**7.808**	**88,20%**	**1,13**
1	1.476	921	62,40%	1,60
2	1.476	1.244	84,30%	1,19
3	1.476	1.353	91,70%	1,09
4	1.476	1.411	95,60%	1,05
5	1.476	1.436	97,30%	1,03
6	1.477	1.443	97,70%	1,02

Elaborado por el Instituto Nacional de Estadísticas.

La Tabla II.4 presenta las principales estadísticas descriptivas del ponderador ajustado por falta de respuesta, donde el ponderador más alto se encuentra en la rama de Servicios con un valor igual a 1.440,7 unidades y, el mínimo en la rama de Agricultura y Pesca con valor de 7,3 unidades.

Tabla II.4. Estadísticas descriptivas del ponderador ajustado por no respuesta, según rama reducida

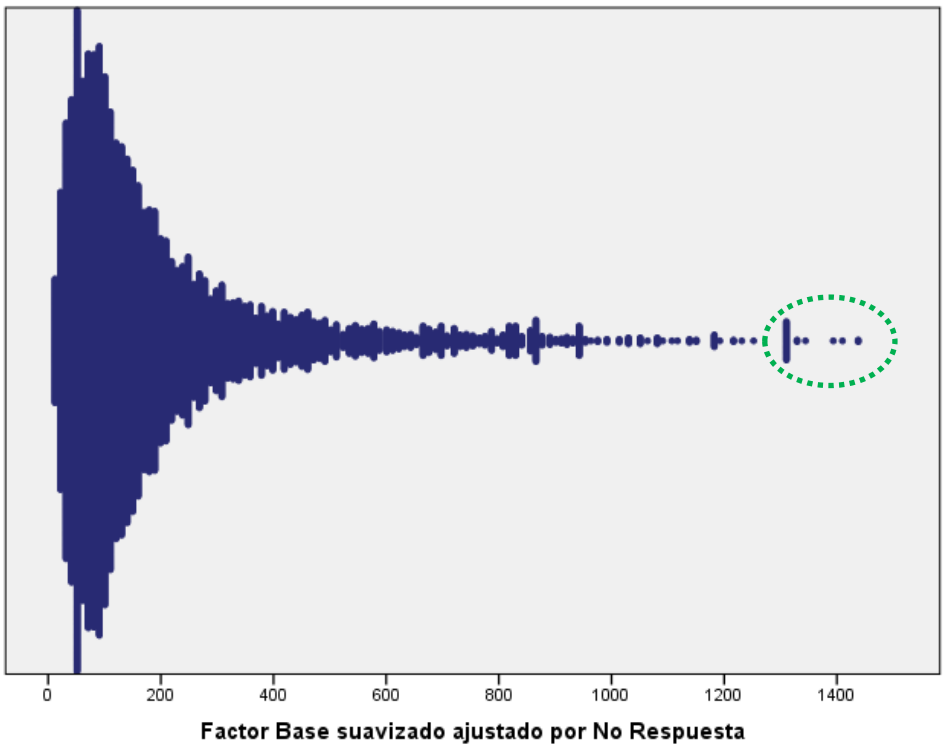
Estadísticas Descriptivas	Rama de actividad reducida								
	Agricultura y Pesca	Sector Primario	Industrias Manufactureras	Construcción	Comercio	Transporte y Almacenamiento	Actividades Inmobiliarias	Servicios	Total
Recuento	1.176	48	1.018	902	1.980	689	69	1.926	7.808
Moda	72,9	69,3	682,9	695,8	865,8	782,3	122,9	1.311,8	1.311,8
Mínimo	7,3	31,1	11,8	9,9	12,1	10,6	31,8	10,6	7,3
Percentil 05	22,6	35,3	32,5	28,8	33,2	34,2	49,5	34,4	30,4
Percentil 25	59,8	58,9	70,4	67,4	77,0	77,2	84,2	80,4	72,0
Mediana	94,6	132,2	120,3	124,0	141,2	134,7	151,0	142,9	127,6
Percentil 75	160,8	240,4	224,6	226,2	272,9	249,5	290,5	274,3	237,2
Percentil 95	305,4	586,8	583,7	611,1	771,5	739,7	833,1	783,7	661,9
Percentil 99	462,4	908,9	798,9	934,4	1.017,3	1.185,7	1.083,9	1.311,8	975,9
Máximo	735,2	908,9	1.142,8	1.086,6	1.409,9	1.231,2	1.083,9	1.440,7	1.440,7
Media	123,2	207,7	180,6	185,5	220,5	210,0	244,0	225,5	197,0
Error estándar de la media	2,8	30,9	5,4	6,2	5,0	8,3	29,1	5,4	2,3
Suma	144.861,8	9.969,3	183.804,5	167.317,0	436.500,5	144.665,5	16.833,0	434.401,1	1.538.352,8

Elaborado por el Instituto Nacional de Estadísticas.

II.4. Suavizamiento del ponderador ajustado por falta de respuesta

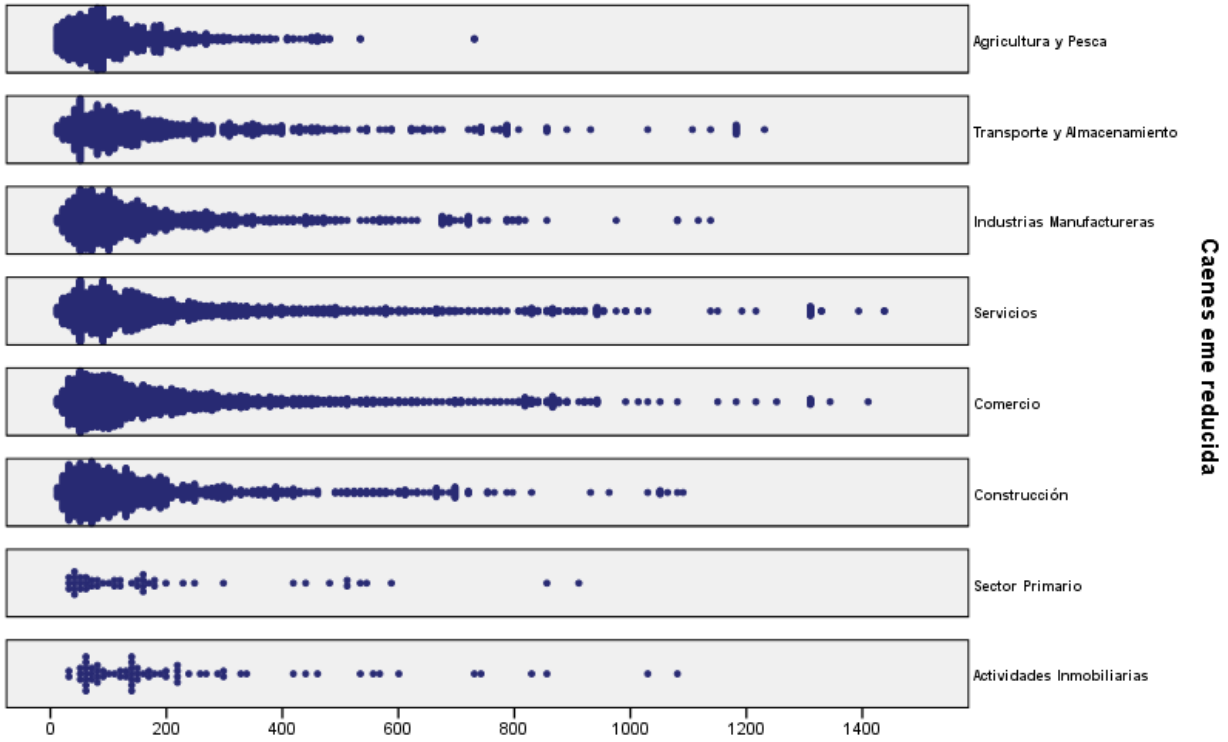
Obtenido el factor de viviendas ajustado por no respuesta, se evaluó la pertinencia de realizar suavizamiento. En la Figura II.5, al observar los factores ordenados, permite identificar valores más altos que el resto de los ponderadores.

Figura II.5 Dispersión del Ponderador ajustado por falta de respuesta



Realizando este mismo ejercicio a nivel de rama reducida, se observa en la Figura II.6, nos permite identificar la discontinuidad del ponderador en la mayoría de las ramas de actividad económicas.

Figura II.6 Dispersión del Ponderador ajustado por falta de respuesta, según rama de actividad



En consecuencia, se decide realizar el suavizamiento del ponderador ajustado por falta de respuesta. Se aplica la misma metodología descrita anteriormente para decidir el punto de corte, dando como resultado el corte $r = 4$. La Tabla II.5 muestra que, en términos de distribución, al comparar el ponderador ajustado por falta de respuesta F_{Rjk}^{NR} versus el ponderador suavizado, todos los valores extremos fueron truncados. El mayor cambio se encuentra en la rama de Servicios donde el valor máximo disminuye de 1440 a 925 unidades.

Tabla II.5. Estadísticas descriptivas del ponderador ajustado por falta de respuesta y su suavizamiento, según rama reducida

Estadísticas Descriptivas	Caenes eme reducida																	
	Agricultura y Pesca		Sector Primario		Industrias Manufactureras		Construcción		Comercio		Transporte y Almacenamiento		Actividades Inmobiliarias		Servicios		Total	
	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$	F_{Rjk}^{NR}	Factor $r4$
Recuento	1.176	1.176	48	48	1.018	1.018	902	902	1.980	1.980	689	689	69	69	1.926	1.926	7.808	7.808
Moda	72,9	73,2	69,3	70,1	682,9	722,2	695,8	742,0	865,8	881,8	782,3	839,9	122,9	124,3	1.311,8	902,2	1.311,8	902,2
Mínimo	7,3	7,3	31,1	31,5	11,8	12,0	9,9	10,2	12,1	12,4	10,6	11,0	31,8	32,1	10,6	10,9	7,3	7,3
Percentil 05	22,6	22,6	35,3	35,7	32,5	33,0	28,8	29,4	33,2	33,8	34,2	35,3	49,5	50,0	34,4	35,4	30,4	31,0
Percentil 25	59,8	60,0	58,9	59,6	70,4	71,6	67,4	68,8	77,0	78,6	77,2	79,7	84,2	85,1	80,4	82,9	72,0	73,2
Mediana	94,6	95,0	132,2	133,8	120,3	122,2	124,0	126,6	141,2	144,0	134,7	139,1	151,0	152,7	142,9	147,4	127,6	130,2
Percentil 75	160,8	161,4	240,4	243,3	224,6	228,2	226,2	230,9	272,9	278,4	249,5	257,7	290,5	293,6	274,3	283,1	237,2	242,0
Percentil 95	305,4	306,6	586,8	593,8	583,7	593,1	611,1	624,0	771,5	787,1	739,7	763,9	833,1	842,1	783,7	808,7	661,9	677,2
Percentil 99	462,4	464,2	908,9	830,8	798,9	722,2	934,4	742,0	1.017,3	883,3	1.185,7	839,9	1.083,9	975,8	1.311,8	902,2	975,9	885,8
Máximo	735,2	492,7	908,9	830,8	1.142,8	731,1	1.086,6	742,0	1.409,9	898,5	1.231,2	839,9	1.083,9	975,8	1.440,7	925,0	1.440,7	975,8
Media	123,2	123,2	207,7	207,7	180,6	180,6	185,5	185,5	220,5	220,5	210,0	210,0	244,0	244,0	225,5	225,5	197,0	197,0
Error estándar de la media	2,8	2,7	30,9	30,2	5,4	5,2	6,2	5,8	5,0	4,7	8,3	7,6	29,1	28,4	5,4	4,9	2,3	2,2
Suma	144.861,8	144.861,8	9.969,3	9.969,3	183.804,5	183.804,5	167.317,0	167.317,0	436.500,5	436.500,5	144.665,5	144.665,5	16.833,0	16.833,0	434.401,1	434.401,1	1.538.352,8	1.538.352,8

Elaborado por el Instituto Nacional de Estadísticas.

Posteriormente, el ponderador ajustado por falta de respuesta suavizado, se utiliza como insumo para realizar la calibración de este factor.

II.5. Calibración

En general, en todas las encuestas de hogares el ponderador final o factor de expansión es calibrado, con el objetivo de alcanzar algún stock poblacional obtenido de una fuente externa a la encuesta. Por ejemplo, los factores de expansión de la Encuesta Nacional de Empleo son calibrados, cada trimestre móvil, al total de población estimada por sexo y tramo de edad (menores de 15 años y 15 o más años) para cada estrato ENE, con fecha 15 de cada mes central del periodo de levantamiento.

En el ejemplo expuesto, la población objetivo corresponde a personas que poseen ciertos atributos demográficos, cuantificados en los Censos de Población y Vivienda, lo que permite obtener una estimación de la población desagregada a esos niveles. Para la EME en cambio, existe un inconveniente, no existe una estimación "oficial" o de referencia, respecto a los "microemprendedores" (formales e informales) a nivel del país.

Por otro lado, la muestra seleccionada en la VI EME, está anclada a la población de referencia del trimestre MAM 2019 de la ENE, lo cual implica que la EME hace un seguimiento a los microemprendedores que se encontraban, en ese período, clasificados como microemprendedores, sin tomar en cuenta los flujos de entrada a esa condición laboral.

Dado lo anterior, se decide utilizar la estimación del total de microemprendedores del trimestre MAM 2019, obtenido de la ENE, actualizada al período del trabajo de campo de la VI EME. Para esto, se utiliza el crecimiento proyectado para el mes central del período de levantamiento de la encuesta, es decir junio 2019. En definitiva, la estimación utilizada en la calibración del ponderador de la EME se obtiene a través de los siguientes pasos:

1. Primero, se considera toda la información levantada para la ENE en el período MAM 2019, es decir, todos los microemprendedores que fueron clasificados como tales, sin importar que, a futuro puedan cambiar de condición. Esta población, expandida a junio de 2019 es la que se toma como referencia para la calibración de los microemprendedores de la VI EME.
2. Segundo, se calcula un nuevo factor de expansión, considerando las proyecciones de población a junio de 2019.
   En el período MAM 2019, la ENE utilizó el siguiente cálculo:

$$F_{hij}^{2} = \frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}} \cdot \frac{P_{hs}^{4}}{\hat{P}_{hs}} = F_{hij}^{1} \cdot \frac{P_{hs}^{4}}{\hat{P}_{hs}} \qquad (18)$$

Donde:

$F_{hij}^{1} = \frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}}$ es el factor teórico inicial de la ENE

$$\hat{P}_{hs} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} F_{hij}^{1} \cdot p_{hijs}$$

p_{hijs} : Total de personas de sexo y tramo de edad, en la vivienda j, del conglomerado i, del estrato ENE h

P_{hs}^{4} : Total de población de sexo y tramo de edad, del estrato ENE h, proyectada al 15 de abril de 2019

Para obtener la estimación del total de microemprendedores para la VI EME, se calcula con la misma fórmula, sin embargo, el stock poblacional utilizado corresponde al proyectado con fecha junio 2019, que viene dada por la ecuación (19):

$$F_{hij}^{2} = \frac{M_h}{n_h \cdot M_{hi}} \cdot \frac{M'_{hi}}{m_{hi}} \cdot \frac{P_{hs}^{6}}{\hat{P}_{hs}} \qquad (19)$$

La Tabla II.6, presenta el total de microemprendedores estimado a partir de la publicación de la ENE, periodo MAM 2019; y según total de personas estimadas con la información levantada en MAM 2019, pero con proyecciones actualizadas a la fecha de levantamiento de la EME (en adelante I_{gs}.). En esta tabla, se observa que, el total de "microemprendedores" estimados y publicados oficialmente son 2.191.076 personas. Sin embargo, al actualizar las proyecciones de población este total asciende a 2.197.540, lo que equivale a un incremento de 0,3% a nivel nacional; 0,4% en la macrozona Norte; 0,4% en el Centro; 0,4% en el Sur y 0,2% en la Región Metropolitana.

Tabla II.6. Total de microemprendedores estimados a partir de la ENE – periodo MAM 2019

		Total microemprendedores	
Macrozona	**Sexo**	**Factor Expansión Oficial ENE - MAM 2019**	**Factor Expansión Información ENE - ajustado a Junio 2019**
Total	Hombre	1.335.972	1.340.021
	Mujer	855.104	857.519
	Total	**2.191.076**	**2.197.540**
Norte	Hombre	182.146	182.879
	Mujer	121.529	121.977
	Total	**303.674**	**304.856**
Centro	Hombre	388.891	390.367
	Mujer	228.024	228.772
	Total	**616.915**	**619.139**
Sur	Hombre	246.809	247.681
	Mujer	127.550	127.997
	Total	**374.360**	**375.678**
Metropolitana	Hombre	518.126	519.095
	Mujer	378.001	378.774
	Total	**896.126**	**897.868**

Elaborado por el Instituto Nacional de Estadísticas.

Finalmente, el ponderador calibrado, se le asigna a cada una de las personas entrevistadas en la EME. El procedimiento de cálculo de este ponderador se resume en tres pasos:

1. Estimar el total de microemprendedores según sexo, para cada macrozona a partir de la EME 2019. Es decir, se estimó el total de microemprendedores a través de la utilización del ponderador de no respuesta, tal como se muestra en la ecuación (20):

$$\hat{P}_{gs} = \sum_{j=1}^{m_g} \sum_{k=1}^{p_g} F_{Rjk}^{NR} \cdot p_{jks} \quad \begin{matrix} g = 1,2,3,4 \\ s = 1,2 \end{matrix} \tag{20}$$

g : Subíndice de la macrozona de procedencia de las unidades

p_g : Número de personas entrevistadas en la vivienda g

m_g : Número de viviendas entrevistadas en la macrozona g

$$p_{jks} = \begin{cases} 1, & si\ persona\ k\ es\ sexo\ s \\ 0, & en\ otro\ caso \end{cases}$$

2. Construir el ajuste a la población total, mediante la razón entre la estimación del total de microemprendedores de acuerdo con fuentes externas (ENE), y la estimación de la encuesta obtenida en el paso (1):

$$\hat{R}_{gs} = \frac{I_{gs}}{\hat{P}_{gs}} \tag{21}$$

3. Construir el Factor de Expansión final, o Ponderador Calibrado, como el producto entre el ponderador ajustado por falta de respuesta con el ajuste a la población total, calculado en el paso (2)

$$F_{gjs}^{cal} = F_{Rjk}^{NR} \cdot \hat{R}_{gs} \tag{22}$$

Al usar el ponderador calibrado, se debe tener en consideración que éste expande al total de "microemprendedores", de sexo s y residentes en la macrozona g, estimados a partir de la Encuesta Nacional de Empleo, en el trimestre móvil MAM 2019, actualizado al crecimiento poblacional de junio 2019 - mes central de levantamiento de EME.

En la Tabla II.7 se observa un incremento en los ponderadores, y por tanto un aumento en los casos más extremos, aunque mucho más acotado las fases anteriores. Por ejemplo, en la macrozona Centro un microemprendedor representaba a 902 personas, sin embargo, al ajustar según macrozona, esta persona representa 1.194 individuos. En todo caso, los valores extremos dados al comienzo ya han sido suavizados (valores sobre 8.000 ahora apenas sobrepasan los 1.500). También hay que considerar que, si se vuelve a suavizar este ponderador calibrado, se descalibraría y la suma de él no llegaría a los stocks de microemprendedores por sexo y tramo etario inducidos por las proyecciones de población, por tanto el criterio estadístico es utilizar como factor final el resultante de la calibración sin aplicar ninguna metodología de suavizamiento adicional.

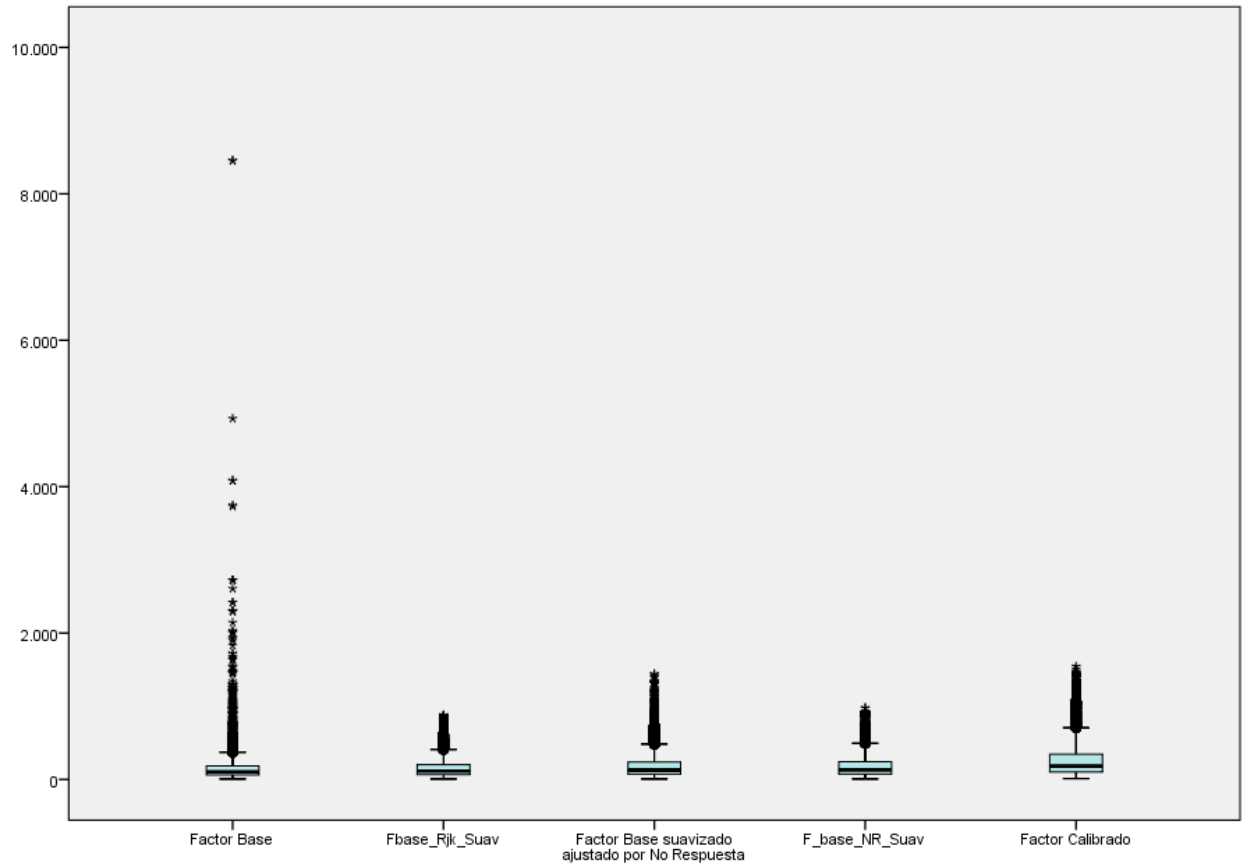
Tabla II.7. Estadísticas descriptivas del ponderador ajustado por falta de respuesta suavizado y calibrado al stock de microemprendedores

Estadísticas Descriptivas	Macrozona									
	Norte		Centro		Sur		Metropolitana		Total	
	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}
Recuento	1.698	1.698	2.747	2.747	1.672	1.672	1.691	1.691	7.808	7.808
Moda	73	32	84	111	88	147	902	1.291	902	1.291
Mínimo	7	10	10	13	11	18	19	27	7	10
Percentil 05	22	29	31	41	31	49	60	94	31	43
Percentil 25	58	78	68	90	68	107	153	230	73	103
Mediana	97	131	126	166	107	170	265	396	130	183
Percentil 75	149	198	221	290	176	278	507	763	242	344
Percentil 95	415	554	478	630	380	585	883	1.291	677	967
Percentil 99	678	961	726	961	683	1.103	902	1.428	886	1.294
Máximo	897	1.136	902	1.194	925	1.508	976	1.545	976	1.545
Media	134	180	172	225	142	225	356	531	197	281
Error estándar de la media	3	4	3	4	3	5	6	9	2	3
Suma	226.759	304.856	472.066	619.139	237.610	375.678	601.919	897.868	1.538.353	2.197.540

Elaborado por el Instituto Nacional de Estadísticas.

En la Figura II.7, se puede observar visualmente todos los ponderadores desde el Factor Base hasta el Factor Calibrado y su comportamiento después de cada ajuste. El primer suavizamiento elimina todos los valores extremos, aquellos que sobrepasaban 4 veces la media en la rama reducida. El ajuste por no respuesta vuelve a generar algunos valores extremos, que nuevamente son suavizados, disminuyendo la dispersión. Finalmente, la calibración del factor aumenta levemente las cotas superiores, pero evidentemente, con menor dispersión respecto al factor base (el aumento de las cotas superiores se debe en parte, al aumento de las proyecciones de población de abril a junio de 2019).

Figura II.7 Gráfico de caja para los diferentes ponderadores.



Elaborado por el Instituto Nacional de Estadísticas

III. ESTIMACIÓN DE LA VARIANZA

El diseño muestral de la VI EME, como ya fue mencionado, contempla un diseño muestral complejo. En general, cuanto más complejo es el diseño muestral bajo el cual se implementa una encuesta, más compleja se vuelve la forma de determinar los errores muestrales. Tanto así que, no existen ecuaciones exactas y/o explícitas para esto. Sin embargo, paquetes estadísticos en softwares especializados, facilitan los cálculos a través de aproximaciones realizadas mediante distintos modelos o métodos de estimación, donde se deben identificar las variables que definen el diseño muestral (estratos, conglomerados) y el factor de expansión apropiado (considerando todos los ajustes pertinentes).

En este contexto, en los siguientes apartados se exponen las variables que identifican el diseño muestral, así como también su implementación en Spss y Stata. Para ello, se definió como variable de análisis la estructura de la rama de actividad de los microemprendedores. Como existen algunas categorías en las cuales se observa una baja prevalencia, tales como: pesca, electricidad, gas y agua; entre otras, se crea una variable más agregada denominada “rama reducida”, sobre la cual se realizan las estimaciones.

En ocasiones, pueden existir algunas dificultades en la implementación de la estimación de los errores mediante un paquete estadístico, originadas por las características del diseño muestral, por ejemplo: más de una fase de muestreo; muestreo multietápico de las unidades muestrales, selección de unidades sin reemplazo, estratos de muestreo con solo una unidad primaria con unidades elegibles, variabilidad de los tamaños de los conglomerados, etc.
En esta encuesta se observan principalmente tres dificultades:

- Diseño muestral complejo
- Existencia de estratos de muestreo (los de la ENE) que poseen sólo un conglomerado (manzana o sección)
- El número de viviendas que responde en cada conglomerado es desigual y muy variable

A fin de minimizar los problemas señalados anteriormente y, siguiendo las recomendaciones internacionales (Valliant, Dever, & Kreuter, 2013) los errores son estimados a partir de modelos que buscan dar cuenta, lo más fielmente posible del diseño muestral. Para ello, se formaron pseudo-estratos y pseudo-conglomerados a través de la agrupación de los estratos y conglomerados originales. El objetivo de esta agrupación es garantizar la estimación de varianzas en los nuevos estratos y, de esta forma, no subestimar los errores.

III.1. Creación de Pseudo-estratos

Los estratos ficticios o pseudo-estratos son construidos con el objetivo de corregir los problemas generados por la existencia de estratos con solo un conglomerado (estratos unitarios), que subestiman la varianza de cualquier variable de interés.

Los pseudo-estratos son construidos a través de la agrupación de dos o más estratos originales, los que pueden ser unitarios o no, de acuerdo a un patrón u ordenamiento jerárquico de variables geográficas o de tamaño, de modo que estos contengan al menos dos conglomerados, los que a su vez deberán contener aproximadamente al menos 15 unidades que responden en su interior.

A continuación, se detalla el procedimiento de construcción de los pseudo-estratos:

1. Primero se contabiliza, al interior de cada estrato original, el total de unidades que participa en la encuesta. Si el estrato contiene menos de 30 (2•15) unidades, entonces deberá ser colapsado con otro.
2. Se ordenan todos los estratos, geográficamente, de acuerdo con la división político-administrativa en el área urbana y rural y luego, al interior de cada región según ordenamiento del estrato.
3. Finalmente, al interior de la misma área geográfica y región se colapsan aquellos estratos con menos de 30 unidades, lo más cercano geográficamente, pero sin que en conjunto estos superen 60 unidades.

De un total de 160 estratos que posee la ENE, en la VI EME se seleccionaron unidades de todos los estratos, de los cuales 4 de estos estratos contienen a un solo conglomerado. Por otra parte, existen 65 estratos con 30 o menos unidades (viviendas). Por lo que, el total de pseudo-estratos creados desciende a 126.

Tabla III.1. Total de estratos y de pseudo-estratos, según macrozona.

Macrozona	Estratos	Pseudo-estrato
Total	**160**	**126**
Norte	25	22
Centro	63	48
Sur	25	23
Metropolitana	47	33

Elaborado por el Instituto Nacional de Estadísticas.

III.2. Creación de Pseudo-conglomerados

Los conglomerados ficticios o pseudo-conglomerados fueron construidos con el objetivo de reducir los problemas generados a causa de la diversidad de tamaños de los conglomerados originales (número de unidades que participa en ellos), pues a mayor variabilidad en el tamaño de los conglomerados, la varianza de los estimadores tiende a incrementarse y volverse más inestable.

Los pseudo-conglomerados fueron creados a partir de un ordenamiento jerárquico, según comuna y total de unidades que responde, al interior de cada pseudo-estrato. Luego, se unieron los conglomerados a fin de que estos en conjunto reunieran entre 12 y 18 unidades aproximadamente (en promedio 15 viviendas).

A continuación, se detalla el procedimiento de construcción de los pseudo-conglomerados;

1. Primero se contabiliza, al interior de cada conglomerado original, el total de individuos que participa en la encuesta. Si el conglomerado contiene menos de 15 unidades entonces deberá ser colapsado con otro.
2. Se ordenan todos los conglomerados geográficamente según área (urbana o rural); región, provincia y comuna (RPC); y total de unidades que responde, al interior de cada pseudo-estrato.
3. Finalmente, al interior de cada pseudo-estrato se colapsan aquellos conglomerados con menos de 15 unidades, los más cercanos geográficamente, pero sin que en conjunto estos superen las 30 unidades.

La ENE posee un total de 4.496 conglomerados, en 3.411 se seleccionaron microemprendedores, los que se transformaron en 560 pseudo-conglomerados.

En la Tabla III.2 se expone el total de conglomerados y pseudo-conglomerados según macrozona.

Tabla III.2. Total de conglomerados y de pseudo-conglomerados, según macrozona

Macrozona	Conglomerados	Pseudo-Conglomerados
Total	**3.411**	**560**
Norte	635	110
Centro	1.289	198
Sur	776	117
Metropolitana	711	135

Elaborado por el Instituto Nacional de Estadísticas.

III.3. Estimación de variables y varianzas en Spss y Stata

Diversos paquetes estadísticos poseen algoritmos que permiten la estimación de los errores muestrales bajo diseños complejos a través de métodos como, el método de linearización de Taylor; métodos de replicación repetido (Jackknife, Bootstrap), entre otros. Sin embargo, para que éstos sean más simples de implementar, se deben considerar algunos supuestos: se asume que la selección de las unidades, en las distintas etapas, se realiza de forma independiente y con reemplazo (esto simplifica los cálculos y las expresiones matemáticas); por otro lado, aun cuando el diseño muestral de la encuesta posea muchas etapas, sólo se da cuenta de la primera etapa, pues es esta la que aporta la mayor variabilidad al error total.

En Spss, previo a la estimación de la variable en estudio y los errores asociados a ella, se debe definir el diseño muestral bajo el cual se realizarán las estimaciones. Las variables, que se encuentran en la base de datos y que definen el diseño muestral de la VI EME son:

1. Factor_EME: corresponde al factor de expansión que da cuenta de las probabilidades de selección, de la fase 1 y 2, ajuste por falta de respuesta y calibración.
2. VarStrat: variable que identifica el estrato, estos contienen al menos dos conglomerados, para garantizar la estimación de la varianza.
3. VarUnit: variable que identifica al conglomerado, estos contienen entre 12 y 18 unidades aproximadamente.

Así, para revisar la estructura de la actividad en la cual se desenvuelven los microemprendedores, previamente, el investigador debiera hacer lo siguiente:

1. Determinar y construir la variable de interés, si ésta no está definida.
2. Especificar las variables que definen el diseño complejo
3. Realizar la estimación correspondiente

Considerando la estructura de la rama de actividad económica (caenes_1d_eme) para los microemprendedores como la variable de análisis -se observa la existencia de categorías en las que la proporción de microemprendedores observados es pequeña, lo que conlleva a obtener estimaciones con gran variabilidad o error muestral-, por lo cual se agruparon las categorías de baja prevalencia en dos grandes grupos, dando origen a una nueva variable denominada “rama de actividad reducida”, según como se indica en la Tabla III.3.

Tabla III.3. Rama de actividad económica según caenes_1d_eme. vs Rama de actividad reducida

Rama de actividad Económica	Rama de actividad Económica Reducida
1 Agricultura, ganadería, silvicultura y pesca	1 Agricultura y Pesca
2 Explotación de minas y canteras	2 Sector Primario
3 Industrias manufactureras	3 Industrias Manufactureras
4 Suministro de electricidad, gas, vapor y aire acondicionado	2 Sector Primario
5 Suministro de agua; evacuación de aguas residuales, gestión de desechos y descontaminación	2 Sector Primario
6 Construcción	4 Construcción
7 Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas	5 Comercio
8 Transporte y almacenamiento	6 Transporte y Almacenamiento
9 Actividades de alojamiento y de servicio de comidas	11 Servicios
10 Información y comunicaciones	11 Servicios
11 Actividades financieras y de seguros	11 Servicios
12 Actividades inmobiliarias	7 Actividades Inmobiliarias
13 Actividades profesionales, científicas y técnicas	11 Servicios
14 Actividades de servicios administrativos y de apoyo	11 Servicios
16 Enseñanza	11 Servicios
17 Actividades de atención de la salud humana y de asistencia social	11 Servicios
18 Actividades artísticas, de entretenimiento y recreativas	11 Servicios
19 Otras actividades de servicios	11 Servicios

Elaborado por el Instituto Nacional de Estadísticas.

A continuación, se presenta un resumen con la estimación de la rama de actividad reducida, en la que fueron clasificados los microemprendedores, según las estimaciones realizadas en Spss[15]. Respecto a la estructura de la rama de actividad de los microemprendedores, se observa que éstos se concentran principalmente, en Comercio, seguido de Servicios, actividades que, en conjunto, reúnen a más de 56,4% de los microemprendedores.

Respecto a la precisión de las estimaciones, se puede constatar que todas las ramas de actividad económica reducida presentan errores relativos aceptables (menores de 30%), a excepción de las ramas 'Sector primario' y 'Actividades inmobiliarias', cuyos errores relativos son de aproximadamente 40% y 32%, respectivamente.

[15] Ver Anexo N°3

Tabla III.4. Estructura de la actividad económica de los microemprendedores- estimación realizada en SPSS

Rama de actividad económica	Estimación	Error estándar	Intervalo de confianza al 95%		Coeficiente de variación	Efecto de diseño
			Inferior	Superior		
Total	**100,0%**	**0,0%**	**100,0%**	**100,0%**	**0,0**	
Agricultura y Pesca	9,6%	0,4%	9,0%	10,4%	0,0	1,2
Sector Primario	0,6%	0,1%	0,4%	1,0%	0,2	2,1
Industrias Manufactureras	11,8%	0,6%	10,8%	13,0%	0,0	2,3
Construcción	10,9%	0,5%	9,9%	11,9%	0,0	2,1
Comercio	28,1%	0,7%	26,8%	29,5%	0,0	1,9
Transporte y Almacenamiento	9,5%	0,4%	8,6%	10,4%	0,0	1,8
Actividades Inmobiliarias	1,1%	0,2%	0,8%	1,5%	0,2	2,3
Servicios	28,3%	0,8%	26,7%	29,9%	0,0	2,6

Elaborado por el Instituto Nacional de Estadísticas.

ANEXOS

Anexo N°1. Códigos de disposición final de casos. Última visita

Esta tabla muestra, las categorías que en la variable “elegible” dice sí, corresponden a unidades elegibles y sobre las cuales se realizan los ajustes por falta de respuesta. Las restantes unidades fueron clasificadas como no elegibles. Cabe señalar que aquellas unidades no elegibles, no son contabilizadas en el ajuste por falta de respuesta.

Código de disposición de la última visita a la vivienda	Frecuencia	Porcentaje	Elegible
Total	**9.227**	**100,0%**	**8.857**
Encuesta completa	7.808	84,6%	SI
Informante de la vivienda rechazó la entrevista	107	1,2%	SI
Informante directo rechazó la entrevista	233	2,5%	SI
Se interrumpió entrevista al informante directo	12	0,1%	SI
Se impidió el acceso a la vivienda	6	0,1%	SI
Vivienda ocupada sin moradores presentes	361	3,9%	SI
No se logra contacto con el informante directo	304	3,3%	SI
Casos especiales	17	0,2%	SI
Problemas de idioma	0	0,0%	SI
Encuesta anulada por falsificación	0	0,0%	SI
Otra razón elegible (especifique en observaciones)	9	0,1%	SI
No se envió a terreno	0	0,0%	NO
Área peligrosa o de difícil acceso	5	0,1%	NO
No fue posible localizar la dirección	10	0,1%	NO
Otra razón de elegibilidad desconocida	15	0,2%	NO
Empresa, oficina de gobierno u otra organización	0	0,0%	NO
Viviendas colectivas o Instituciones	0	0,0%	NO
Vivienda en demolición, incendiada, destruida o erradicada	2	0,0%	NO
Vivienda particular desocupada	13	0,1%	NO
Vivienda de temporada	1	0,0%	NO
Muerte del informante	5	0,1%	NO
Cambio de domicilio	205	2,2%	NO
Informante fuera de marco	110	1,2%	NO
Otra razón no elegible (especifique en observaciones)	4	0,0%	NO

Elaborado por el Instituto Nacional de Estadísticas.

Anexo N°2. Regresión logística implementada en la construcción de celdas para ajustes de no respuestas

Para la selección del mejor modelo que permita estimar la probabilidad de responder de un microemprendedor seleccionado para participar en la VI EME, se consideraron dos herramientas: 1) Modelamiento y 2) Sensibilidad del modelo. Para el modelamiento de la variable de respuesta, se seleccionaron un conjunto de variables que permitieran ajustar mejor la respuesta de interés, para así llegar a la selección del modelo ideal. En la etapa de Sensibilidad del modelo se determina qué "tan bueno" es nuestro ajuste, a través de la Curva ROC.**Regresión Logística**

Dado que nuestra variable de interés tiene dos categorías provenientes de una respuesta binaria (Responde vs No responde), se utiliza un modelo que considera esta característica a medir. Los modelos ampliamente usados para estudiar este fenómeno están dentro de una clase mayor de modelos llamados modelos lineales generalizados. Primeramente, se define la variable aleatoria binaria como:

$$Y_i = \begin{cases} 1, & si\ la\ i-ésima\ persona\ responde\ dado\ que\ pertenece\ a\ una\ unidad\ elegible \\ 0, & si\ la\ i-ésima\ persona\ no\ responde\ dado\ que\ pertenece\ a\ una\ unidad\ elegible \end{cases}$$

Con $P(Y=1)=\pi$ y con $P(Y=0)=1-\pi$. Si hay n variables aleatorias $Y_1, Y_2, \dots, Y_n$, independientes entre sí, con $P(Y_i=1)=\pi_i$, $\forall\ i=1,\dots,n$, entonces su función de probabilidad conjunta es:

$$\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = exp\left[\sum_{i=1}^{n} y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} log(1-\pi_i)\right]$$

Función perteneciente a la familia exponencial.

Al considerar la siguiente función de enlace[16]:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^t\beta \quad (29)$$

[16] Nuestro interés es modelar $E(Y_i)=\pi_i$ con, $\pi_i \in [0,1]$, a través, de $\boldsymbol{x}_i^t\beta$. Sin embargo, no existe una relación lineal entre π_i y $\boldsymbol{x}_i^t\beta$, tal que $E(Y_i)=\pi_i=\boldsymbol{x}_i^t\beta$, por lo general esta relación es de tipo no lineal. Para resolver esto, se necesita una función g que relacione la respuesta media con los regresores a estimar, es decir, $g(\pi_i)=\boldsymbol{x}_i^t\beta$, de tal forma que, $E(Y_i)=\pi_i=g^{-1}(\pi_i)$, entonces, se dice que g es una función de enlace. Ahora bien, si Y_i se puede expresar de forma general como $f(y;\pi)=\exp[a(y)b(\pi)+c(\pi)+d(y)]$, se dice que Y_i pertenece a la familia exponencial. Además, si $a(y)=y$ se dice que la distribución es de la forma canónica (o, estándar) y $b(\pi)$ se llama el parámetro natural de la distribución. Nuestra variable de interés sigue una distribución binomial, es decir, $Y_i \sim Binomial(1,\pi_i)$, se sabe que esta variable aleatoria pertenece a la familia exponencial con parámetro natural $b(\pi_i)=\log(\pi_i/1-\pi_i)$ y eso nos permite tomar este parámetro natural como función de enlace para $\boldsymbol{x}_i^t\beta$, de tal forma que, $\log(\pi_i/1-\pi_i)=\boldsymbol{x}_i^t\beta$. Finalmente, nuestro modelo a estimar es $Y_i \sim Binomial\left(1,\frac{exp(\boldsymbol{x}_i^t\beta)}{1+exp(\boldsymbol{x}_i^t\beta)}\right)$. [Para mayor detalle consultar Dobson (2002)]

Con $\boldsymbol{x}_i^t = \left(1, x_1, x_2, \dots, x_p\right)^t$ y $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$, tal que, $\boldsymbol{x}_i^t\beta = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_p\beta_p$

Se tiene que la probabilidad del suceso es:

$$\pi_i = \frac{exp\left(\boldsymbol{x}_i^t\beta\right)}{1 + exp\left(\boldsymbol{x}_i^t\beta\right)} = P\left(Y_i = 1/\boldsymbol{x}_i^t\beta\right)$$

1.2. Indicadores estadísticos para evaluar el desempeño de un procedimiento diagnóstico

1.2.1. Sensibilidad y especificidad

La **sensibilidad** y la **especificidad** son las medidas tradicionales y básicas del valor diagnóstico de un modelo. Miden la discriminación diagnóstica de un modelo en relación a un criterio de referencia, que se considera la verdad.

La **sensibilidad** (S) indica la capacidad del modelo para detectar a un sujeto que responde, es decir, expresa cuan "sensible" es la prueba a la presencia de personas que responden. Para cuantificar su expresión se utilizan términos probabilísticos: si la persona responde, ¿cuál es la probabilidad de que el resultado sea positivo?

La **especificidad** (E) indica la capacidad que tiene el modelo para identificar a las personas que no responden cuando efectivamente no responden.

Considerando un espacio de unidades elegibles y las personas que responden la encuesta versus las que no, se definen los siguientes cuantificadores para la variable de respuesta:

VP: Verdaderos positivos, número de personas que respondieron la encuesta y fueron diagnosticados como positivos por el modelo.

FP: Falsos positivos, número de personas que no respondieron y fueron diagnosticados como positivos por el modelo.
FN: Falsos negativos, números de personas que respondieron y fueron diagnosticado como negativos por el modelo.
VN: Verdaderos negativos, número de personas que no respondieron y fueron diagnosticado como negativos por el modelo.

Con estos términos, la Matriz de confusión puede expresarse así:

		Criterio de Verdad		Total
		Responden	No responden	
Prueba Diagnóstica	**Positivos**	VP	FP	VP+FP
	Negativos	FN	VN	FN+VN
	Total	VP+FN	FP+VN	N=(VP+FP+FN+VN)

Fuente: Elaboración propia, INE.

$$Sensibilidad(S) = \frac{Verdaderos\ positivos}{Total\ de\ Responden} = \frac{VP}{VP + FN}$$

$$Especificidad(E) = \frac{Verdaderos\ negativos}{Total\ de\ No\ responden} = \frac{VN}{VN + FP}$$

1.2.2. Valores predictivos

A pesar de que S y E se consideran las características operacionales fundamentales de una prueba diagnóstica, en la práctica su capacidad de cuantificación de la incertidumbre es limitada. Se necesita más bien evaluar la medida en que sus resultados modifican realmente el grado de conocimiento que se tenía sobre el estado de la persona. Concretamente, le interesa conocer la probabilidad de que un individuo para el que se haya obtenido un resultado positivo sea efectivamente una persona que responde; y lo contrario, conocer la probabilidad de que un individuo con un resultado negativo este efectivamente libre no responder. Las medidas o indicadores que responden a estas interrogantes se conocen como **valores predictivos.**

El **valor predictivo de una prueba positiva** equivale a la probabilidad condicional de que los individuos con una prueba positiva realmente respondan:

$$VP(+) = P(Resp/T+)$$

El **valor predictivo de una prueba negativa** es la probabilidad condicional de que los individuos con una prueba negativa realmente no respondan:

$$VP(-) = P(No\ Resp/T-)$$

Mediante la tabla de 2×2 que se introdujo antes se puede ilustrar también como se estiman los valores predictivos (suponiendo que esta tabla se conforma seleccionando una muestra al azar de tamaño N de la población, y luego se clasifiquen los sujetos de la muestra en los cuatro grupos posibles según la prueba diagnóstica y el criterio de verdad) a través de:

$$Valor\ predictivo\ positivo = \frac{Verdaderos\ positivos}{Total\ de\ positivos} = \frac{VP}{VP + FP}$$

$$Valor\ predictivo\ negativo = \frac{Verdaderos\ negativos}{Total\ de\ negativos} = \frac{VN}{VN + FN}$$

1.2.3. Curva ROC

Para la elección entre dos o más modelos, se recurre a las curvas ROC, ya que es una medida global e independiente del punto de corte (o umbral).

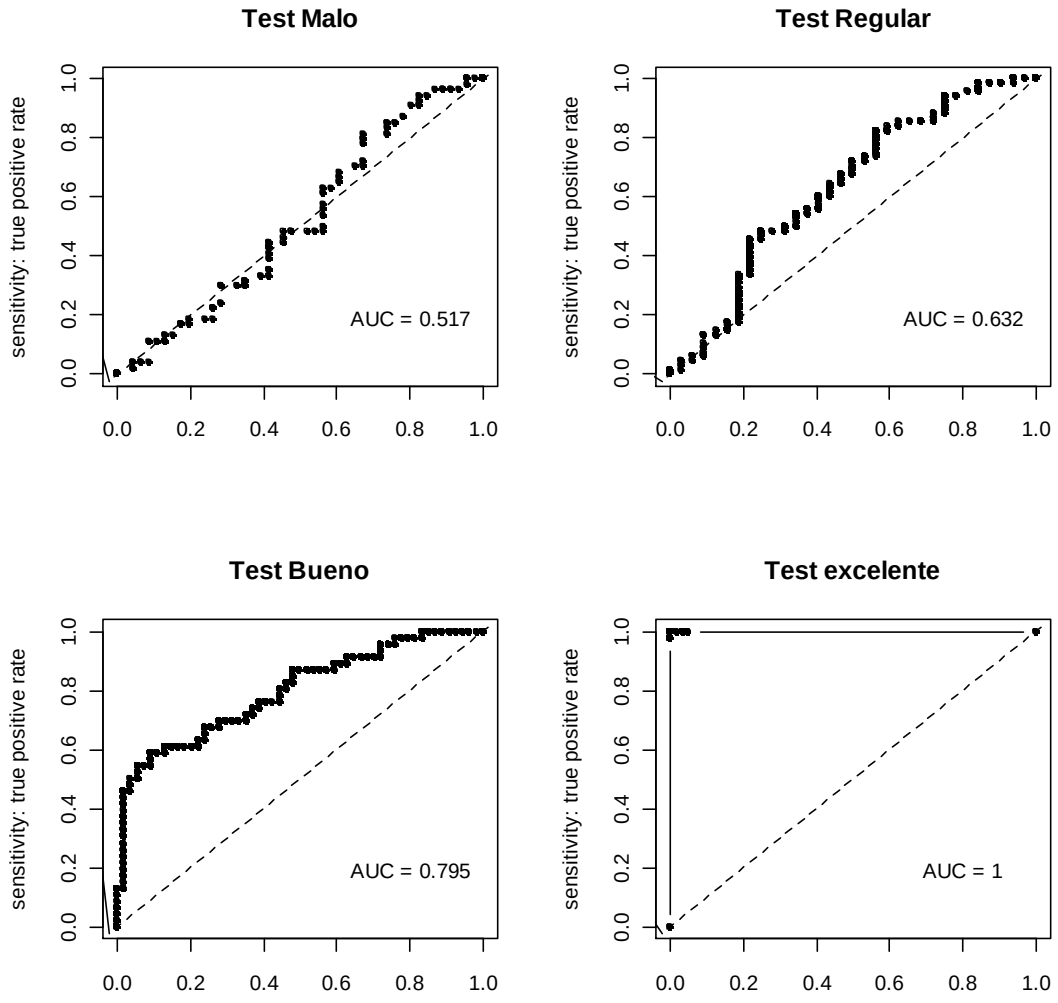
Tradicionalmente cuando se tiene un test cuantitativo, se escoge el cut-off o punto de corte más adecuado, que combine mejor la sensibilidad y especificidad del test (es decir, mayor rendimiento). Habitualmente deberían estar con sensibilidad de 85 %, con especificidad de 74 % o cercanos a estos valores.

La elección se realiza mediante la comparación del área bajo la curva (AUC, de su acrónimo en inglés Area Under the Curve) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Por ejemplo, si el AUC para una prueba diagnóstica médica es 0,8 significa que existe un 80% de probabilidad de que el diagnóstico realizado a un enfermo sea más correcto que el de una persona sana escogida al azar. Por esto, siempre se elige la prueba diagnóstica que presente una mayor área bajo la curva.

A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC:

- [0,5 - 6): Test malo.
- [0,6 - 0,75): Test regular.
- [0,75 - 0,9): Test bueno.
- [0,9 - 0,97): Test muy bueno.
- [0,97 - 1] Test excelente.

Figura III.1 Diferentes curvas ROC



Elaborado por el Instituto Nacional de Estadísticas.

1.3. Análisis de Elegibilidad

Para modelar la probabilidad de que una persona conteste la encuesta de la VI EME dado que pertenece a una unidad elegible, se analiza primeramente la operacionalización de la variable "Código de disposición de la última visita al hogar" reportado por el encuestador en la hoja de ruta.

1.3.1. Operacionalización de variables

Código de disposición de la última visita a la vivienda	Frecuencia	Porcentaje	Elegible	La persona responde
Total	**9.227**	**100,0%**	**8.857**	**7.808**
Encuesta completa	7.808	84,6%	SI	SI
Informante de la vivienda rechazó la entrevista	107	1,2%	SI	NO
Informante directo rechazó la entrevista	233	2,5%	SI	NO
Se interrumpió entrevista al informante directo	12	0,1%	SI	NO
Se impidió el acceso a la vivienda	6	0,1%	SI	NO
Vivienda ocupada sin moradores presentes	361	3,9%	SI	NO
No se logra contacto con el informante directo	304	3,3%	SI	NO
Casos especiales	17	0,2%	SI	NO
Problemas de idioma	0	0,0%	SI	NO
Encuesta anulada por falsificación	0	0,0%	SI	NO
Otra razón elegible (especifique en observaciones)	9	0,1%	SI	NO
No se envió a terreno	0	0,0%	NO	-
Área peligrosa o de difícil acceso	5	0,1%	NO	-
No fue posible localizar la dirección	10	0,1%	NO	-
Otra razón de elegibilidad desconocida	15	0,2%	NO	-
Empresa, oficina de gobierno u otra organización	0	0,0%	NO	-
Viviendas colectivas o Instituciones	0	0,0%	NO	-
Vivienda en demolición, incendiada, destruida o erradicada	2	0,0%	NO	-
Vivienda particular desocupada	13	0,1%	NO	-
Vivienda de temporada	1	0,0%	NO	-
Muerte del informante	5	0,1%	NO	-
Cambio de domicilio	205	2,2%	NO	-
Informante fuera de marco	110	1,2%	NO	-
Otra razón no elegible (especifique en observaciones)	4	0,0%	NO	-

Elaborado por el Instituto Nacional de Estadísticas.

En base a este cuadro se dividen las unidades elegibles (8.857) de las que no (370) y dentro de las unidades elegibles clasificamos las personas que responden (7.808) versus las que no (1.049).

1.4. Aplicación Regresión logística

El principio básico en la inclusión de variables está basado en un modelo simple con un número de variables restringido sobre el total de variables existentes. Se probaron varios modelos, sin embargo, el que mejor cumple las condiciones, es el que contiene las siguientes variables explicativas; edad de la persona, macrozona de pertenencia del hogar, grupo ocupacional, área geográfica, número de visitas, proveedor principal y sexo de la persona. La siguiente tabla, muestra los parámetros estimados para este modelo, de acuerdo a las categorías que son estadísticamente significativas (p-value) de cada variable explicativa.

Los **Odd Ratios** $e^{\widehat{\beta}_1}$ se pueden interpretar como el aumento estimado en la probabilidad de éxito asociado con un cambio unitario en el valor de la variable predictora. En general, el aumento estimado está asociado con un cambio de d unidades en la variable predictora, es decir, $e^{d\cdot\widehat{\beta}_1}$.

Estimaciones de parámetro

				95% de intervalo de confianza			95% de intervalo de confianza para Exp(B)	
Responde		B	Desv. Error	Inferior	Superior	Exp(B)	Inferior	Superior
SÍ	(Intersección)	-1,708	,343	-2,380	-1,035	,181	,093	,355
	[Area=1]	1,346	,265	,827	1,865	3,843	2,287	6,457
	[Area=2]	,000[a]				1,000		
	[Macrozona=1]	1,752	,311	1,142	2,361	5,765	3,134	10,605
	[Macrozona=2]	1,236	,288	,672	1,801	3,443	1,957	6,054
	[Macrozona=3]	1,144	,289	,577	1,710	3,138	1,781	5,530
	[Macrozona=4]	,000[a]				1,000		
	[sexo=1]	-,547	,085	-,714	-,379	,579	,490	,684
	[sexo=2]	,000[a]				1,000		
	[proveedor=0]	-,200	,083	-,362	-,038	,819	,696	,963
	[proveedor=1]	,000[a]				1,000		
	[Nivel_colap=1]	,155	,125	-,091	,401	1,168	,913	1,493
	[Nivel_colap=2]	,046	,099	-,148	,240	1,047	,862	1,271
	[Nivel_colap=3]	,000[a]				1,000		
	[ciuo_1d_eme=1]	,168	,214	-,252	,588	1,183	,777	1,801
	[ciuo_1d_eme=2]	,313	,176	-,032	,657	1,367	,968	1,930
	[ciuo_1d_eme=3]	,524	,201	,130	,917	1,688	1,139	2,503
	[ciuo_1d_eme=4]	-,621	,529	-1,658	,416	,538	,191	1,516
	[ciuo_1d_eme=5]	,507	,140	,233	,780	1,660	1,262	2,182
	[ciuo_1d_eme=6]	,734	,175	,391	1,076	2,083	1,479	2,933
	[ciuo_1d_eme=7]	,501	,138	,231	,771	1,651	1,260	2,162
	[ciuo_1d_eme=8]	,524	,168	,194	,854	1,689	1,214	2,350
	[ciuo_1d_eme=9]	,000[a]				1,000		
	[visitas_colap=1]	2,796	,115	2,572	3,021	16,386	13,091	20,509
	[visitas_colap=2]	,961	,105	,755	1,167	2,614	2,127	3,213
	[visitas_colap=3]	,000[a]				1,000		
	edad	,005	,003	,000	,011	1,005	1,000	1,011
	[Area=1] * [Macrozona=1]	-,653	,343	-1,326	,020	,520	,265	1,020
	[Area=1] * [Macrozona=2]	-1,116	,303	-1,710	-,521	,328	,181	,594
	[Area=1] * [Macrozona=3]	-,598	,316	-1,218	,022	,550	,296	1,023
	[Area=1] * [Macrozona=4]	,000[a]				1,000		
	[Area=2] * [Macrozona=1]	,000[a]				1,000		
	[Area=2] * [Macrozona=2]	,000[a]				1,000		
	[Area=2] * [Macrozona=3]	,000[a]				1,000		
	[Area=2] * [Macrozona=4]	,000[a]				1,000		

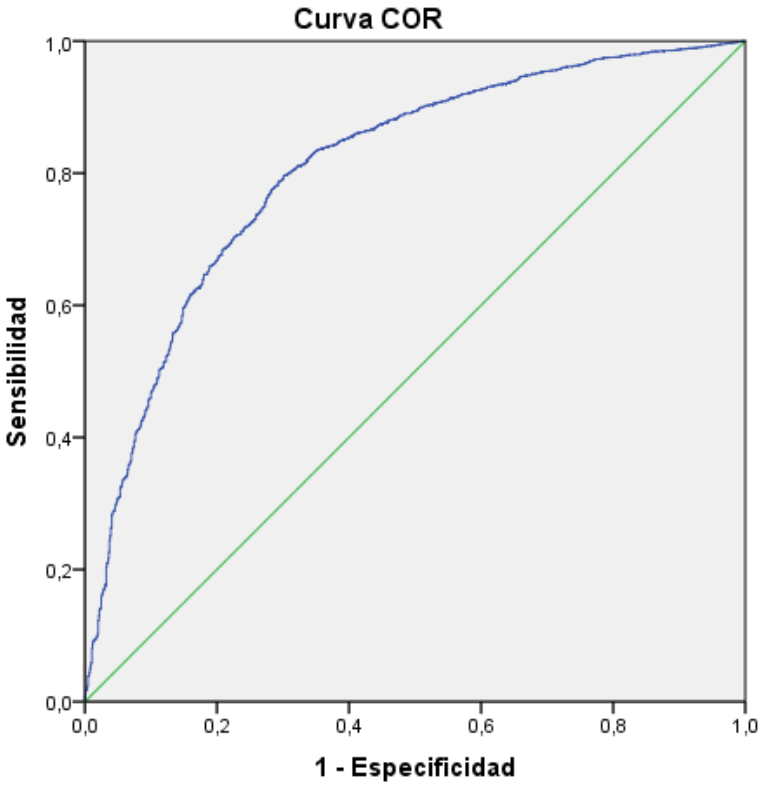
Variable dependiente: Responde (categoría de referencia = NO)
Modelo: (Intersección), Area, Macrozona, sexo, proveedor, Nivel_colap, ciuo_1d_eme, visitas_colap, edad, Area * Macrozona
a. Definido en cero porque este parámetro es redundante.

La interpretación de los coeficientes de regresión en el modelo de regresión logístico múltiple se parece al caso en el que el predictor lineal sólo contiene un regresor, que nos indica que la cantidad $e^{\widehat{\beta}_1}$ es el cociente de ventaja para la covariable x_j, suponiendo que las demás variables predictoras son constantes.

1.4.1. Análisis de Resultados

En base al modelo estimado se puede observar que el área bajo la curva (AUC) es de 0,807, que indica que está dentro **de una categoría de "Test bueno".** O bien, se puede decir que el modelo tiene una capacidad de predicción del 80,7% de los casos.

Figura III.2 Probabilidad estimada de responder para cada una de las personas que pertenecen a la unidad elegible



Elaborado por el Instituto Nacional de Estadísticas.

En la siguiente tabla, se observa cómo queda conformada la tabla de confusión.

Clasificación			
	Pronosticado		
Observado	NO	SÍ	Porcentaje correcto
NO	146	903	13,9%
SÍ	124	7.684	98,4%
Porcentaje global	3,0%	97,0%	88,4%

Variable dependiente: Responde (categoría de referencia = NO)

Modelo: (Intersección), Area, Macrozona, sexo, proveedor, Nivel_colap, ciuo_1d_eme, visitas_colap, edad, Area * Macrozona

Finalmente, la sensibilidad y especificidad calculada es:

$$Sensibilidad = \frac{7.684}{(7.684+903)} = 0{,}895 \qquad Especificidad = \frac{146}{(146+124)} = 0{,}541$$

Anexo N°3. Estimación de varianzas

```
* Plan de muestreo.

CSPLAN ANALYSIS
  /PLAN FILE='Plan_EME.csaplan'
  /PLANVARS ANALYSISWEIGHT=Factor_EME
  /SRSESTIMATOR TYPE=WOR
  /PRINT PLAN
  /DESIGN STRATA=VarStrat CLUSTER=VarUnit
  /ESTIMATOR TYPE=WR.


* Estimación de frecuencias en Spss.

CSTABULATE
  /PLAN FILE='Plan_EME.csaplan'
  /TABLES VARIABLES=caenes_eme_red
  /CELLS TABLEPCT
  /STATISTICS SE CV CIN(95) DEFF
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```

BIBLIOGRAFÍA

Cochran, W. (1998). *Técnicas de Muestreo.* México D.F.: Compañía Editorial Continental, S.A. de C.V.

Jones, A., Koolman, X., & Rice, N. (2006). Health Related Non Response in The British Household Panel Survey and European Community Household Panel:Using Inverse Probability Weighted Estimators in NonLinear Models. *Journal of The Royal Statistical Society. Series A (Statistics in Society), Vol. 169, No. 3*.

Kish, L. (1963). Changing strata and selection probabilities. *Proceedings of the Social Statistics Section, American Statistical Association*, 124-131.

ONU. (2009). *Diseño de muestras para encuestas de hogares. Directrices prácticas.* Recuperado el 15 de Noviembre de 2017, de https://unstats.un.org/unsd/publication/seriesf/Seriesf_98s.pdf

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples.* New York: Springer.