ENCUESTA DE MICROEMPRENDIMIENTO 2017

DISEÑO MUESTRAL

INSTITUTO NACIONAL DE ESTADÍSTICAS

Diciembre /2017

SUBDIRECCIÓN TÉCNICA
DEPARTAMENTO DE INVESTIGACIÓN Y DESARROLLO
SECCIÓN DE ESTADÍSTICAS SOCIALES
DEPARTAMENTO DE ESTUDIOS LABORALES

Encuesta de Microemprendimiento 2017

Diseño Muestral

Instituto Nacional de Estadísticas.

Diciembre / 2017.

ÍNDICE

ÍNDICE DE CUADROS

INTRODUCCIÓN

El presente documento describe las características del diseño muestral, así como la metodología de cálculo de los factores de expansión de la Quinta Encuesta de Microemprendimiento (V EME). En los primeros dos capítulos se presentan los aspectos relacionados con el diseño muestral, exponiéndose los detalles e insumos necesarios para la determinación del tamaño muestral, las unidades muestrales, así como también las características del marco y unidades seleccionadas. El tercer capítulo está focalizado en el desarrollo y construcción del factor de expansión. En él se detallan las probabilidades de selección, el ponderador base (inverso de las probabilidades de selección), el ajuste por falta de respuesta y la calibración a stock de total de microempendedores según macrozona y sexo[1], además se detalla el procedimiento de suavizamiento de los ponderadores. Finalmente, en el cuarto capítulo, se especifica la forma de utilizar las variables que definen el diseño muestral en la estimación y respectivos errores.

[1] Ver más detalles en capítulo 3.

1. ANTECEDENTES DEL DISEÑO MUESTRAL

La EME está dirigida a hogares en donde reside un dueño de un microemprendimiento, tiene carácter bienal y es realizada desde el año 2013 por el Instituto Nacional de Estadísticas (INE) en conjunto con el Ministerio de Economía, Fomento y Turismo, convirtiéndose en el instrumento oficial que permite caracterizar la heterogénea realidad de los microemprendimientos en Chile, aportando información para la elaboración, seguimiento y evaluación de políticas públicas en este ámbito.

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, ya sea formales o informales, pertenecientes a todos los sectores.

A continuación, se exponen los objetivos del estudio, población objetivo, unidad de información y nivel de estimación, utilizados para definir la estrategia de muestreo.

1.1. Objetivo General

- Lograr, a través de la implementación de una encuesta a hogares, una caracterización de la heterogénea realidad de los microemprendimientos del país, sus dueños y trabajadores, y su evolución en el tiempo.

- Desarrollar un mejor análisis de los microemprendimientos, la formalidad de ellos, el acceso que tienen a financiamiento, el nivel de capital humano y los determinantes y/o motivaciones que configuran el inicio de una actividad económica.

1.2. Objetivos Específicos

- Identificar y caracterizar la situación de formalidad bajo distintas dimensiones (Registros contables, inscripción en servicios de impuestos internos, declaración de impuestos, organización jurídica, generación de empleo formal e informal, etc.) y sus determinantes.
- Indagar acerca de la relación que tiene el negocio con el sistema financiero, a través del acceso y trabas al financiamiento, sus características y usos del mismo.

- Estudiar la motivación y las razones del surgimiento de los microemprendimientos. Si éstos son motivados por necesidad, por oportunidad o bien, causados por situaciones del entorno.
- Identificar los obstáculos que dificultan el desarrollo de las unidades productivas, tales como las restricciones en materia de acceso a tecnología, capacitación, financiamiento, entre otros. Conocer la situación laboral actual del microemprendedor, así como sus experiencias o fracasos anteriores.
- Conocer el nivel educacional con que cuentan los microemprendedores, además de las áreas más importantes donde han recibido capacitación en los últimos tres años.
- Realizar una recopilación de datos que permita comparar los resultados con estadísticas internacionales sobre industrias y microemprendimiento.

1.3. Población Objetivo

El estudio está enfocado a las unidades productivas de menor tamaño, es decir, al microemprendedor tradicional, que es por lo general informal y más precario, que puede ser captado mediante una encuesta a hogares, en contraposición de un emprendedor de alto impacto que puede ser captado por otras fuentes.

Debido a que no existe un consenso entre los especialistas en emprendimiento sobre una definición de quiénes son microemprendedores, la Subsecretaría de Economía en conjunto con el INE ha optado por definir como población objetivo a todos quienes se hayan clasificados en la Encuesta Nacional de Empleo (ENE) como “Trabajadores por Cuenta Propia” o “Empleadores dueños de una empresa con hasta 10 trabajadores (incluyendo al dueño)”, en el período de levantamiento referencial, lo cual está en línea con los estándares internacionales promovidos por OIT[2], con la clasificación nacional dispuesta en el estatuto de las PYME[3] y la experiencia de este tipo de encuestas en la región.

En este contexto, la población objetivo son todos los trabajadores por cuenta propia y empleadores con hasta 10 trabajadores (incluyendo al dueño), denominados Microemprendedores, que residen en viviendas particulares ocupadas del territorio nacional.

[2] Ver manual estadístico sobre el sector informal y el empleo informal publicado por OIT en el año 2013; http://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_222986.pdf

[3] Para mayor información revisar el siguiente link: https://www.bcn.cl/leyfacil/recurso/estatuto-de-las-pymes

1.4. Unidad de información

La unidad de información es el microemprendedor que reside en la vivienda particular y que haya sido entrevistado en la Encuesta Nacional de Empleo, y clasificado en dicha categoría laboral. El cuestionario es respondido de manera directa por el informante seleccionado.

1.5. Nivel de estimación

Se entiende por nivel de estimación aquellas desagregaciones geográficas o características sociodemográficas, para las cuales se desean obtener estimaciones con márgenes de error adecuados[4] y buena cobertura geográfica.

La muestra de la V EME fue seleccionada aleatoriamente a fin de representar tanto las áreas urbanas y rurales de las 15 regiones del país, el diseño muestral fue concebido con la finalidad de obtener estimaciones a nivel regional, y por lo tanto para mayores desagregaciones no garantiza buenos márgenes de error.

[4] Adecuado, en el contexto de que el error relativo, por ejemplo, en los niveles de estimación y otros definidos por el estudio, no supere cierto umbral establecido de acuerdo a criterios históricos y consensuados.

2. DISEÑO MUESTRAL

La quinta versión de la Encuesta de Microemprendimiento, posee un diseño muestral bifásico, en que la primera fase corresponde a un muestreo probabilístico, estratificado y bietápico, donde las unidades primarias corresponden a manzanas en el área urbana y secciones en el área rural; mientras que las unidades de segunda etapa son las viviendas particulares. Las unidades primarias fueron seleccionadas en forma proporcional al tamaño, mientras que las unidades de segunda etapa se seleccionaron de forma sistemática y con igual probabilidad. Así, las unidades seleccionadas y encuestadas en la Encuesta Nacional de Empleo del periodo MAM 2017 fueron utilizadas como marco de muestreo[5] para la V EME, pues permitió identificar las viviendas donde residen Microemprendedores (según la clasificación en la ENE).

En la segunda fase, se clasificó las viviendas en dos grupos, de acuerdo a si éstas contenían o no, en el período de referencia, al menos un Microemprendedor. Las viviendas que no poseen microemprendedores fueron descartadas, formando el marco de muestreo con sólo aquellas viviendas con unidades elegibles. Este listado fue revisado en profundidad para descartar todas las unidades que no cumplían 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.

Posteriormente a la depuración del listado, se seleccionaron con igual probabilidad y de forma sistemática las viviendas a formar parte de la muestra. Luego, se listaron todos los microemprendedores al interior de la vivienda y del hogar y se seleccionaron de forma aleatoria tantos trabajadores como tipos de actividad distinta (al interior del hogar) que éstos desempeñaban. Si al interior de algún hogar dentro de la vivienda existía más de un microemprendedor desempeñando una misma actividad económica, entonces sólo se seleccionó a uno de ellos para efectos de no redundar en la misma información[6].

---

[5] Los muestreos en fases operan de esta forma. Se levanta una gran encuesta en primera fase para capturar una primera población objetivo, y que sirve para identificar a los individuos que entrarán en segunda fase (ya que, de otra manera, si se decidiera a levantar una muestra independiente para una segunda población objetivo, habría que levantar una muestra tan grande como la primera para capturar a los individuos objeto de estudio, encareciendo en demasía los costos). La primera encuesta (primera fase) entra a operar como si fuera un marco de muestreo (un sub-marco) para la selección de los individuos en segunda fase.

[6] En efecto, al interior del hogar, los microemprendedores que desempeñan una misma actividad económica, generalmente comparten muchas características comunes en cuanto al microemprendimiento que realizan.

En las siguientes secciones se describen las características de los marcos de muestreo de ambas fases, y la estimación y distribución del tamaño muestral.

2.1. Características del Marco Muestral

A continuación, se describen las características del marco muestral a partir del cual se seleccionó la muestra de la V EME. Como las unidades seleccionadas en la EME proceden desde la ENE, se deben revisar las características del marco de muestreo asociados a la fase 1 (ENE) y la fase 2 (EME).

2.1.1. Cobertura geográfica

La cobertura es una propiedad estadística asociada al marco muestral y un indicador de la calidad de la encuesta, que se utiliza para la selección de la muestra. Así, el ámbito geográfico de la cobertura muestral, comprende el área urbana y rural del país. Sin embargo, se deben hacer algunas especificaciones de ciertas áreas que no cubre la encuesta.

La V 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 2017); y segundo el marco utilizado para la selección de los “Microemprendedores” para la V EME.

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 como: hogares de ancianos, hospitales, cárceles, conventos, etc.; ni tampoco 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 catalogadas por el INE como áreas de difícil acceso (ADA) o alto costo (que corresponden al 0,3% del total viviendas)[7]. Por otro lado, para optimizar el trabajo de campo y dadas las características de las unidades muestrales del área urbana

[7] Ver Anexo N°1

(manzanas) se descartan del Marco muestral, previo a la selección de las unidades, las viviendas en manzanas con 7 o menos viviendas. En total, el marco de la ENE excluye alrededor del 1,03% de las viviendas del país, según el Censo de Población y Vivienda del año 2002.

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

2.1.2. Estratificación del Marco Muestral

El Marco de Muestreo de la ENE fue estratificado según su condición geográfica (División Político Administrativa) y según el número de viviendas y población que contenían al CENSO 2002, además de una segregación dependiendo de la actividad económica preponderante en el área.

La estratificación del Marco de la ENE da origen a los siguientes estratos:

- Ciudades o grandes Centros Urbanos (CD): Conformadas por ciudades o conjuntos de ciudades adyacentes con 40.000 ó 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 (según información del Censo de Población y Vivienda del año 2002) dedicada a actividades primarias[8].

En la segunda fase, la V EME tiene cobertura geográfica del área urbana y rural del país, estratificada de forma natural de acuerdo a las 15 regiones que posee el país.

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 1 se detalla la composición de cada macrozona.

[8] Se entiende por Actividad Primaria a toda aquella actividad relacionada con la extracción de recursos naturales (agricultura, caza, pesca, minería, etc.).

Cuadro 1. Composición de macrozonas

Macrozona	Región
Norte	Arica y Parinacota (15)
	Tarapacá (1)
	Antofagasta (2)
	Atacama (3)
	Coquimbo (4)
Centro	Valparaíso (5)
	Libertador General Bernardo O'Higgins (6)
	Maule (7)
	Biobío (8)
Sur	La Araucanía (9)
	Los Ríos (14)
	Los Lagos (10)
	Aysén del General Carlos Ibáñez del Campo (11)
	Magallanes y La Antártica Chilena (12
Metropolitana	Metropolitana de Santiago (13)

Fuente: Elaboración propia

2.1.3. Depuración del listado de Independientes.

En correspondencia con el diseño muestral de la V EME, se elaboró un listado de unidades que permitiera la identificación de los microemprendedores, este proceso lo realiza el Departamento de Estudios Laborales (DEL) a partir de la información recogida en la Encuesta Nacional de Empleo en el trimestre MAM 2017, se creó un listado de personas, clasificadas como independientes, el cual fue utilizado como marco muestral para la selección de la muestra en la V EME.

La revisión se hizo a partir de las variables de rama de actividad económica y grupo ocupacional. A partir de esta última, se eliminaron del listado los informantes que tienen un alto grado de subordinación o dependencia que en muchos casos son considerados como independientes debido a que se encuentran en una situación intermedia entre ser Asalariado o dueño de su propio negocio (zonas grises de asalariados encubiertos)[9]. Específicamente se borraron del listado los independientes que realizaban actividades relacionadas con el “cuidado de personas” (de niños, ancianos, enfermos, etc.), actividades de servicio doméstico en hogares (limpieza de

[9] Actualmente la OIT se encuentra en un proceso de revisión al Clasificador Internacional de Situación en el Empleo del año 1993.

hogares, de pisos, de vidrios, planchadores, lavanderos, jardineros que trabajaban en un hogar, etc.) o independientes con altos grados de dependencia (personal de apoyo, ayudantes, cargadores, empaquetadores, reponedores, meseros, vendedores por catálogo, captadores de clientes, etc.). Además, fueron eliminadas del listado algunas actividades temporales o esporádicas como los temporeros agrícolas o el personal que trabaja de forma esporádica en la realización de eventos. Esto, porque la ENE es contestada por un informante idóneo (proxy), quien responde por él y por todos los integrantes de su hogar, lo que constituye una fuente de error no muestral de clasificación, propio de las encuestas a hogares, según los cuales una persona pudiera ser clasificada como independiente en la ENE, pero que en la realidad no lo sea, y viceversa. Una vez realizada esta depuración permanecen solo aquellos casos que cumplen con la condición de microemprendedor, es decir, todos aquellos trabajadores por cuenta propia y empleadores con hasta 10 trabajadores incluyéndose, que realizan una actividad económica de producción de bienes y servicios orientado al mercado y que recibe a cambio de eso una ganancia, especies y/o ingreso monetario.

En el cuadro 2, se presentan las variables “Total Microemprendedores ENE”, correspondiente al total de personas clasificadas en la ENE como microemprendedores, en el período MAM 2017; 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 V EME. En total, la depuración del marco corresponde a 12,7%[10] de casos descartados por ser potenciales unidades no elegibles[11], observándose los mayores cambios en la región de Aysén (17,1%) y los menores en la región del Maule y Magallanes, ambas con un 6,9%.

[10] $\frac{11.536-10.068}{11.536} = 0{,}127$

[11] En la EME, se entiende por 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, en el periodo de referencia de la ENE no era un microemprendedor, o también que haya cambiado de estado (dejó de ser microemprendedor).

Cuadro 2. Distribución del total de microemprendedores muestrales según ENE MAM2017 y según Marco EME

Macrozona	Región	Total microemprendedores ENE	Total microemprendedores EME
Total		**11.536**	**10.068**
Norte	Arica y Parinacota	480	442
	Tarapacá	444	382
	Antofagasta	312	269
	Atacama	295	260
	Coquimbo	799	689
Total Norte		**2.330**	**2.042**
Centro	Valparaíso	1.461	1.222
	Libertador Gral. Bernardo O'Higgins	576	514
	Maule	713	664
	Biobío	1.197	1.057
Total Centro		**3.947**	**3.457**
Sur	La Araucanía	854	763
	Los Ríos	367	316
	Los Lagos	940	819
	Aysén del Gral. Carlos Ibáñez del Campo	315	261
	Magallanes y Antártica Chilena	130	121
Total Sur		**2.606**	**2.280**
Metropolitana	Metropolitana de Santiago	2.653	2.289
Total Metropolitana		**2.653**	**2.289**

Fuente: Elaboración propia

2.2. Estimación y Distribución del tamaño muestral

La V EME al poseer un diseño bifásico, considera que sus unidades serán seleccionadas a partir de otra encuesta o listado, en particular de la ENE. En este contexto, los parámetros a utilizar para la determinación del tamaño muestral fueron extraídos de la ENE, para la subpoblación específica de "Microemprendedores".

2.2.1. Tamaño de la muestra

La estimación del tamaño muestral, se obtuvo a partir de un muestreo aleatorio simple en cada nivel de estimación, al cual se le aplican principalmente tres correcciones: la primera da cuenta del diseño muestral a partir de un estadígrafo denominado efecto del diseño (deff); la segunda da cuenta que la población en estudio es finita; y la tercera, corrige el tamaño para compensar la falta de respuesta, pérdida usual en este tipo de estudios.

El parámetro de estudio o variable de interés (pivote) para el cual se necesita obtener estimaciones precisas en la población U o nivel de estimación (regional), es una razón entre dos variables:

$$R_{Y/X} = \frac{N°\ Trabajadores\ por\ cuenta\ propia}{Total\ de\ Microemprendedores} = \frac{Y}{X} = \frac{\sum_{k\in U} y_k}{\sum_{k\in U} x_k} \qquad (1)$$

El método a utilizar para estimar un tamaño muestral adecuado en términos de precisión de acuerdo a los requerimientos, se basa en la relación entre el error estándar[12] y el tamaño de muestra empleado para obtenerlo.

El Error Estándar SE del estimador $\hat{P}$ en relación al porcentaje de individuos con cierta característica, en el contexto de un muestreo polietápico, está dado aproximadamente por la expresión:

$$V(\hat{P}) = SE_{\hat{P}}^2 \approx \left(1 - \frac{m}{M}\right)\frac{S_{\hat{P}}^2 \cdot Deff_{\hat{P}}}{m} \qquad (2)$$

En esta expresión, $Deff_{\hat{P}}$ es el efecto del diseño[13], $f = \frac{m}{M}$ es la fracción de muestreo y $cpf = 1 - f = \left(1 - \frac{m}{M}\right)$ es la corrección por finitud o factor de corrección de la varianza en muestreo de poblaciones finitas, siendo m el número de viviendas a encuestar y M el número de viviendas en la población del nivel de estimación requerido.

[12] El error estándar de la estimación es simplemente la raíz cuadrada de la varianza de la estimación, esto es: $SE_{\hat{P}} = \sqrt{V(\hat{P})}$, o alternativamente, la varianza es igual al cuadrado del error estándar, $V(\hat{P}) = SE_{\hat{P}}^2$

[13] Se puede interpretar como el aumento o disminución en la varianza, debido a considerar un muestreo complejo (es decir. estratificado, bietápico, por conglomerados) en vez de un muestreo aleatorio simple de viviendas. Aproximadamente, es el cociente entre la varianza de un muestreo multietápico y la de un muestreo aleatorio simple de viviendas.

El error absoluto de la estimación del parámetro P, denotado como $E_A(\hat{P})$, está relacionado con la varianza de esta misma estimación por la expresión:

$$E_A(\hat{P}) = Z_{1-\frac{\alpha}{2}} \cdot SE_{\hat{P}} = Z_{1-\frac{\alpha}{2}} \cdot \sqrt{V(\hat{P})} \quad (3)$$

Siendo $Z_{1-\alpha/2}$ el percentil $1-\alpha/2$ de la distribución Normal Estándar, asociada a una estimación por intervalos de $1-\alpha$ de nivel de confianza. Por lo general, se usa un nivel de confianza del 95%, por lo cual el percentil equivale al 97,5% y el valor usado es entonces. $Z_{1-\alpha/2} = 1{,}96$

Luego, para determinar el tamaño muestral se deben fijar ciertos parámetros, como: la tasa de no respuesta (Tnr), el error absoluto $E_A(\hat{P}) = e_0$, y el nivel de confianza $1-\alpha$.

Finalmente, el tamaño muestral se determina mediante la siguiente fórmula,

$$m = \frac{Z_{1-\alpha/2}^2 \cdot S_{\hat{P}}^2 \cdot Deff_{\hat{P}}}{e_0^2 + \frac{Z_{1-\alpha/2}^2 \cdot S_{\hat{P}}^2 \cdot Deff_{\hat{P}}}{M}} \cdot \frac{1}{(1-Tnr)} \quad (4)$$

2.2.2. Estimación del Tamaño Muestral

De acuerdo a lo señalado anteriormente, primero se determinó el tamaño muestral a nivel nacional bajo las dos primeras correcciones: el efecto del diseño y por finitud. De acuerdo a esto, el tamaño muestral resultante, es de 6.799 viviendas, tamaño determinado con un nivel de confianza del 95%, y un error absoluto nacional de 1,17%.

Cuadro 3. Total de viviendas a encuestar sin considerar corrección de no respuesta.

	Parámetros Obtenidos EME 2015		Tamaño sin Tnr		
Nivel de Estimación	**Estimación P[14]**	**Deff**	**N° viviendas Esperado**	**Error Absoluto E_A**	**Error Relativo E_R**
Nacional	86,4%	2,58	6.799	1,17%	1,35%

Fuente: Elaboración propia

Todas las encuestas de hogares sufren la pérdida de unidades debido al agotamiento del informante, o unidades no elegibles debido a desactualización del marco de muestreo, rechazos, etc. En encuestas donde el diseño muestral es bifásico, dicho problema puede acrecentarse debido a que la condición que hace la unidad elegible puede cambiar en el tiempo. En el caso de la V EME, la condición de "microemprendedor" puede cambiar de un período a otro, por lo tanto, es más probable obtener un menor número de unidades con información al finalizar el proceso de levantamiento.

Al considerar una tasa de no respuesta esperada de alrededor del 15%, el total de viviendas a seleccionar y enviar a terreno es de 8.062, de las cuales se espera obtener información de al menos 6.799 unidades.

Cuadro 4. Tamaño muestral (Total de viviendas) determinado según la proporción de cuenta propia sobre microemprendedores.

Nivel de Estimación	**Estimación P[15]**	**Deff**	**N° viviendas seleccionar**
Nacional	86,4%	2,58	8.199

Fuente: Elaboración propia

[14] P corresponde a la razón entre el total de trabajadores por cuenta propia y el total de microemprendedores en el período de referencia.

[15] P corresponde a la razón entre el total de trabajadores por cuenta propia y el total de microemprendedores en el período de referencia.

El tamaño de la muestra esperada[16] es de 8.199 viviendas aproximadamente, sujeto a un nivel de estimación nacional y regional y error absoluto fijo de 1,17%. Dichas unidades fueron distribuidas de forma proporcional en las 15 regiones del país, de acuerdo a la estructura observada en la ENE para el trimestre de referencia, teniendo la encuesta como niveles de estimación, nivel nacional y regional.

2.2.3. Distribución de la muestra entre regiones según submuestra.

Una vez obtenido este tamaño muestral requerido de acuerdo a los objetivos de precisión a nivel nacional - 8.199 viviendas - se distribuyeron éstas en los distintos subniveles de desagregación en forma proporcional al tamaño, según la cantidad de microemprendedores reportados en la ENE, debido que, al momento de diseñar la muestra, aún no se contaba con el total de microemprendedores del periodo MAM 2017.

Como la muestra de la ENE está subdividida en tres meses o períodos de levantamiento, con el objetivo de disminuir el tiempo transcurrido entre el levantamiento de información de la ENE y la EME y con ello tener una menor atrición[17], se distribuyó la muestra de la V EME independientemente en tres meses de levantamiento: mayo, junio y julio, de acuerdo al mes de levantamiento de la ENE, marzo, abril y mayo, respectivamente.

La distribución regional se realizó de forma proporcional en cuanto al total de viviendas con al menos un microemprendedor reportado en MAM 2017. Es decir, en aquellas regiones donde se observó un mayor número de microemprendedores se le asignó un mayor número de viviendas a encuestar. Posteriormente, al interior de cada región la muestra fue subdividida en tres partes iguales, cuando ello fuera posible, según el mes de levantamiento. Así, las viviendas a encuestar en el mes de mayo en la V EME deberán ser aquellas viviendas que fueron entrevistadas en marzo 2017 en la ENE.

A continuación, se ilustra la distribución de la muestra según mes de levantamiento y región.

---

[16] Cabe mencionar que los errores efectivos se calculan con la muestra efectivamente lograda en terreno, ante lo cual los errores pueden ser mayores a los esperados. Este tamaño corresponde al obtenido de las simulaciones adicionales, considerando una tasa de no-respuesta del 15%, aproximadamente.

[17] La atrición, entendida como la acumulación de pérdida de información por la no respuesta que se presenta en estudios sobre unidades en el tiempo por parte de los participantes.

Cuadro 5. **Total de viviendas seleccionadas según región y mes de levantamiento.**

Macrozona	Región	Mes Levantamiento V EME			Total
		Mayo	Junio	Julio	
Total		**2.683**	**2.765**	**2.751**	**8.199**
Norte	Arica y Parinacota	122	113	113	348
	Tarapacá	110	98	99	307
	Antofagasta	75	86	80	241
	Atacama	57	83	72	212
	Coquimbo	194	198	197	589
Total Norte		**558**	**578**	**561**	**1.697**
Centro	Valparaíso	331	327	328	986
	Libertador Gral. Bernardo O'Higgins	136	148	143	427
	Maule	174	174	175	523
	Biobío	300	298	298	896
Total Centro		**941**	**947**	**944**	**2.832**
Sur	La Araucanía	176	176	176	528
	Los Ríos	91	91	91	273
	Los Lagos	211	211	211	633
	Aysén del Gral. Carlos Ibáñez del Campo	70	70	70	210
	Magallanes y Antártica Chilena	33	34	35	102
Total Sur		**581**	**582**	**583**	**1.746**
Metropolitana	Metropolitana de Santiago	603	658	663	1.924
Total Metropolitana		**603**	**658**	**663**	**1.924**

Fuente: Elaboración propia

2.3. Selección de Unidades

La Encuesta Nacional de Empleo registra para cada miembro del hogar de 15 o más años, la información necesaria para caracterizarlos de acuerdo a si éstos pertenecen o no a la Fuerza de Trabajo. Además de ello, registra información que permite la categorización de las personas "ocupadas" según la Clasificación Internacional de la Situación de Empleo (CISE), lo que permite identificar la población objetivo, es decir, "los microemprendedores". Esta variable es la que permite la construcción del Marco de Muestreo de la EME, a partir del cual se seleccionaron las viviendas y personas (microemprendedores). En el cuadro 6 se presenta la distribución del total de viviendas y personas según región.

Cuadro 6. Total de viviendas y personas Marco EME

Macrozona	Región	ENE MAM 2017		Selección V EME	
		Total viviendas	Total Micro-emprendedores	Total viviendas	Total Micro-emprendedores
Total		**8.873**	**10.068**	**8.199**	**8.820**
Norte	Arica y Parinacota	384	442	348	376
	Tarapacá	331	382	307	338
	Antofagasta	241	269	241	261
	Atacama	229	260	212	230
	Coquimbo	623	689	589	628
Total Norte		**1.808**	**2.042**	**1.697**	**1.833**
Centro	Valparaíso	1.094	1.222	986	1.050
	Libertador Gral. Bernardo O'Higgins	456	514	427	464
	Maule	586	664	523	568
	Biobío	957	1.057	896	950
Total Centro		**3.093**	**3.457**	**2.832**	**3.032**
Sur	Araucanía	645	763	528	558
	Los Ríos	283	316	273	287
	Los Lagos	712	819	633	674
	Aysén del Gral. Carlos Ibáñez del Campo	238	261	210	230
	Magallanes y Antártica Chilena	107	121	102	109
Total Sur		**1.985**	**2.280**	**1.746**	**1.858**
Metropolitana	Metropolitana de Santiago	1.987	2.289	1.924	2.097
Total Metropolitana		**1.987**	**2.289**	**1.924**	**2.097**

Fuente: Elaboración propia

La selección se realizó en dos etapas, primero sobre las viviendas y luego en su interior a los microemprendedores. Las viviendas fueron seleccionadas con igual probabilidad, de forma sistemática al interior de cada región. Las unidades seleccionadas fueron ordenadas previamente de acuerdo a las variables que identifican la división política administrativa (región, provincia, comuna, distrito censal, zona censal, manzana) y área (urbano-rural). De esta manera se garantiza que estén representadas todas las áreas geográficas en la misma medida como éstas se encuentran en el marco de muestreo.

Posteriormente, en cada vivienda seleccionada y hogar al interior de esta, se identificaron a todos los microemprendedores y las actividades económicas en que éstos se desenvuelven. Luego, se seleccionaron de forma aleatoria y con igual probabilidad, tantos microemprendedores como actividades distintas identificadas, es decir, en caso de encontrar más de un microemprendedor en el hogar ejecutando la misma actividad económica, se tomó el resguardo de seleccionar sólo a un representante por actividad dentro del hogar.

3. FACTORES DE EXPANSIÓN

La muestra de la V EME fue diseñada para lograr representatividad a nivel nacional y regional. En atención a los errores que se desea alcanzar, al presupuesto disponible y para compensar las pérdidas asociadas a la no respuesta, la muestra objetivo fue sobre-dimensionada aproximadamente en un 15%, determinándose como tamaño óptimo la recolección de 8.199 viviendas. Este sobredimensionamiento, tuvo algunas excepciones en regiones como Antofagasta y Metropolitana, donde se aplicó un sobremuestreo del 23% y 17%, respectivamente, debido a que su tasa de no respuesta en la versión anterior fue más alta que el resto de las regiones.

Los factores de expansión se obtienen como el inverso de las probabilidades de selección, además de la aplicación de diversos ajustes. En este caso, las probabilidades de selección asociados a los microemprendedores tienen varias componentes:

1. Probabilidad de que la vivienda hubiera sido seleccionada y contestado la ENE periodo MAM 2017.
    - Probabilidad de seleccionar el conglomerado de pertenencia.
    - Probabilidad de seleccionar la vivienda dado que el conglomerado al que pertenece fue seleccionado.
    - Probabilidad de responder ENE.
2. Probabilidad de seleccionar una vivienda para EME, dado que la vivienda posee microemprendedores.
3. Probabilidad de seleccionar un microemprendedor, dado que su vivienda fue seleccionada.

Mientras que respecto a los ajustes que se deben realizar, éstos son:

1. Ajuste por falta de respuesta (probabilidad de que el microemprendedor participe en EME V).
2. Ajuste a un stock poblacional dado un periodo de referencia.

Respecto a los elementos utilizados en los cálculos del factor de expansión, se puede especificar que:

1. Lo referido a las probabilidades de selección de la primera fase, se extraen directamente de la Encuesta Nacional de Empleo, ya que son éstas las

utilizadas en el factor de expansión de la ENE (previo a la post-estratificación por sexo y tramo de edad).

2. Tanto las probabilidades de selección de las viviendas y de las personas se extraerán directamente desde el Marco de la V EME.
3. Respecto al ajuste por falta de respuesta, se deben utilizar los grupos o "celdas de ajuste", creadas a partir de la información existente, tanto de los que responden como los que no, en la V EME.
4. Calibración al stock poblacional, el cual fue creado a partir de los datos recogidos en la ENE en el periodo de referencia donde se seleccionó la muestra (MAM 2017), pero ajustados al crecimiento poblacional estimado a partir de las proyecciones poblacionales de junio del 2017, según macrozona y sexo.

En los apartados siguientes se detalla el proceso de cálculo de las probabilidades de selección, así como también de los factores. Se hablará indistintamente de factores de expansión y de ponderadores.

3.1. Ponderador Base

El ponderador base se define como el factor de expansión obtenido sólo con las probabilidades de selección, sin ajustes ni correcciones.

En la V EME, las personas seleccionadas, corresponden a un subconjunto de personas que participaron durante el proceso de encuestaje del trimestre MAM 2017 de la ENE. Por lo tanto, uno de los insumos fundamentales del ponderador base, son los factores de expansión de vivienda de la ENE, que dan cuenta de la probabilidad de que una vivienda haya sido seleccionada y entrevistada en la ENE. La sección 3.1.1 expone las probabilidades de selección y respuesta de la ENE; la sección 3.1.2 expone la formula explicita de la probabilidad condicional de selección de un microemprendedor; finalmente, en la sección 3.1.3 se expone la fórmula matemática del ponderador base.

3.1.1. Probabilidad de selección y entrevista de las viviendas en la muestra de la ENE –MAM 2017.

El diseño muestral de la Encuesta Nacional de Empleo, corresponde a un muestreo probabilístico, estratificado y bietápico, donde las unidades primarias corresponden a manzanas en el área urbana y secciones en el área 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. 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 expresión que se detalla a continuación fue extraída desde el documento "Manual conceptual y Metodológico del diseño muestral de la ENE[18]", que corresponde al ponderador inicial o teórico corregido por no respuesta.

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

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 responde en la ENE en el periodo MAM 2017.

En consecuencia, la probabilidad de haber sido seleccionada y entrevistada la vivienda j, del conglomerado i, en el estrato h en el trimestre móvil MAM 2017 en la ENE, está dado por

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

[18] http://www.ine.cl/canales/chile_estadistico/mercado_del_trabajo/empleo/metodologia/pdf/031110/manual_metodologico031110.pdf

3.1.2. Probabilidad de selección de los microemprendedores

La selección de los microemprendedores se realizó en dos etapas. Primero, se seleccionaron con igual probabilidad viviendas que contenían al menos un microemprendedor, según mes de levantamiento al interior de cada región.

Así, la probabilidad de selección de una vivienda que posee al menos un microemprendedor está dada por:

$$p_{Rj}^{v} = \frac{m_R^{micro}}{M_R^{micro}}$$

Donde:

R: Subíndice que representa la región de pertenencia. $R = 1, \dots, 15$.

j: Subíndice que representa la vivienda j.

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

M_R^{micro}: Corresponde al total de viviendas con al menos un microemprendedor en la región R, de acuerdo a la clasificación de la ENE (en el marco de la ENE).

m_R^{micro}: Corresponde al 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, puede ser aproximada por:

$$p_{Rjklm}^{micro|v} = \frac{S_{Rjlm}^{micro}}{T_{Rjlm}^{micro}}$$

Donde:

T_{Rjlm}^{micro}: Corresponde al 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}: Corresponde al 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, puede ser aproximada por la siguiente expresión:

$$p_{Rjk}^{micro} = p_{Rj}^{v} \cdot p_{Rjklm}^{micro|v}$$

En el cuadro 7 se observa que, en general, la probabilidad de seleccionar microemprendedores, condicional a que la vivienda fue seleccionada en la EME, oscila entre 20% y 100%.

Las altas probabilidades de selección de viviendas en la EME que se presentan en las distintas macrozonas, se explican principalmente por la selección, en algunas regiones, de un gran número de microemprendedores en comparación al total de unidades disponibles para seleccionar[19], casi abarcando en su totalidad el universo disponible de unidades del listado de microemprendedores obtenidos de la ENE.

Cuadro 7. Estadísticas descriptivas de la probabilidad de selección de las viviendas con al menos un microemprendedor y probabilidad condicional de microemprendedores (condicional[20] a la actividad dentro del hogar) según macrozona.

Estadísticas Descriptivas	Macrozona							
	Norte		Centro		Sur		Región Metropolitana	
	Probabilidad de selección vivienda(1)	Probabilidad de selección personas(2)	Probabilidad de selección vivienda(1)	Probabilidad de selección personas(2)	Probabilidad de selección vivienda(1)	Probabilidad de selección personas(2)	Probabilidad de selección vivienda(1)	Probabilidad de selección personas(2)
Recuento	1.833	1.833	3.032	3.032	1.858	1.858	2.097	2.097
Moda	1,00	1,00	1,00	1,00	0,89	1,00	0,97	1,00
Mínimo	0,79	0,25	0,83	0,25	0,77	0,20	0,94	0,33
Percentil 05	0,79	1,00	0,83	1,00	0,77	0,50	0,94	0,50
Percentil 25	0,90	1,00	0,88	1,00	0,86	1,00	0,94	1,00
Mediana	0,95	1,00	0,91	1,00	0,89	1,00	0,97	1,00
Percentil 75	1,00	1,00	0,96	1,00	0,91	1,00	1,00	1,00
Percentil 95	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Percentil 99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Máximo	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Media	0,94	0,98	0,92	0,98	0,89	0,96	0,97	0,97

Fuente: Elaboración propia

(1) Probabilidad de selección de una vivienda en el marco EME (marco con todas las viviendas de la ENE que contienen al menos a un microemprendedor).

(2) Probabilidad de selección de un microemprendedor (persona) condicional a que la vivienda fue seleccionada en el marco EME.

[19] Nótese que, a nivel nacional, la probabilidad de selección de una vivienda con al menos un microemprendedor del marco EME es aproximadamente 92% (8.199 viviendas a seleccionar sobre 8.873 viviendas del marco EME. Ver cuadro 6, página 15).

[20] Esta probabilidad condicional es, en la mayoría de los casos, igual a uno, dado que la mayoría de las viviendas contiene un solo hogar y, a su vez, en la mayoría de los hogares existe un solo emprendedor que desarrolla una actividad.

Cuadro 8. Estadísticas descriptivas de la probabilidad de selección condicional de los microemprendedores, según rama de actividad económica.

Estadísticas Descriptivas	B14. Rama de Actividad Económica de la Empresa Donde Trabaja, CIIU Revisión 4.cl – 1 Dígito, según CAENES[21]																		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**16**	**17**	**18**	**19**	**Total**
Recuento	1.405	27	1.070	3	31	1.118	2.407	699	447	68	19	63	398	95	108	178	130	554	8.820
Moda	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Mínimo	0,20	0,50	0,50	1,00	1,00	0,33	0,33	0,33	0,33	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,20
Percentil 05	0,50	1,00	1,00	1,00	1,00	1,00	0,50	1,00	1,00	1,00	0,50	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,50
Percentil 25	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Mediana	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Percentil 75	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Percentil 95	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Percentil 99	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Máximo	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Media	0,95	0,98	0,98	1,00	1,00	0,99	0,97	0,98	0,98	0,98	0,97	0,99	0,98	0,99	0,99	0,99	0,98	0,99	0,97

Fuente: Elaboración propia

[21] 1 Agricultura, ganadería, silvicultura y pesca; 2 Explotación de minas y canteras; 3 Industrias manufactureras; 4 Suministro de electricidad, gas, vapor y aire acondicionado; 5 Suministro de agua; evacuación de aguas residuales, gestión de desechos y descontaminación; 6 Construcción; 7 Comercio al por mayor y al por menor; reparación de vehículos automotores y motocicletas; 8 Transporte y almacenamiento; 9 Actividades de alojamiento y de servicio de comidas; 10 Información y comunicaciones; 11 Actividades financieras y de seguros; 12 Actividades inmobiliarias; 13 Actividades profesionales, científicas y técnicas; 14 Actividades de servicios administrativos y de apoyo; 15 Administración pública y defensa; planes de seguridad social de afiliación obligatoria; 16 Enseñanza; 17 Actividades de atención de la salud humana y de asistencia social; 18 Actividades artísticas, de entretenimiento y recreativas; 19 Otras actividades de servicios; 20 Actividades de los hogares como empleadores; actividades no diferenciadas de los hogares como productores de bienes y servicios; 21 Actividades de organizaciones y órganos extraterritoriales.

3.1.3. Inverso de las probabilidades de selección o Ponderador Base

El ponderador base, es aquel que da cuenta de las probabilidades de selección de las viviendas en la fase 1, y las probabilidades de selección de los microemprendedores en la fase 2, condicional a que la vivienda de residencia fue seleccionada en la ENE y que éstas participaran en el periodo MAM 2017.

Así, calculadas las probabilidades de selección y participación de una vivienda en la ENE en el trimestre MAM 2017 y la probabilidad de seleccionar un microemprendedor desde la EME, el ponderador base se calcula como:

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

En el cuadro 9 se observa que tanto las ramas de Industria Manufactureras, Comercio y Servicios, presentan ponderador base sobre 4.000, seguido de las ramas Construcción y Transporte y Almacenamiento con valores sobre 3.000, pero entre ellos, similares distribuciones y variabilidad de sus ponderadores. Según el gráfico 1, el mayor valor del ponderador se observa en la rama de Industrias Manufactureras y también en Servicios, siendo hasta siete veces más grande que los valores extremos de las restantes ramas.

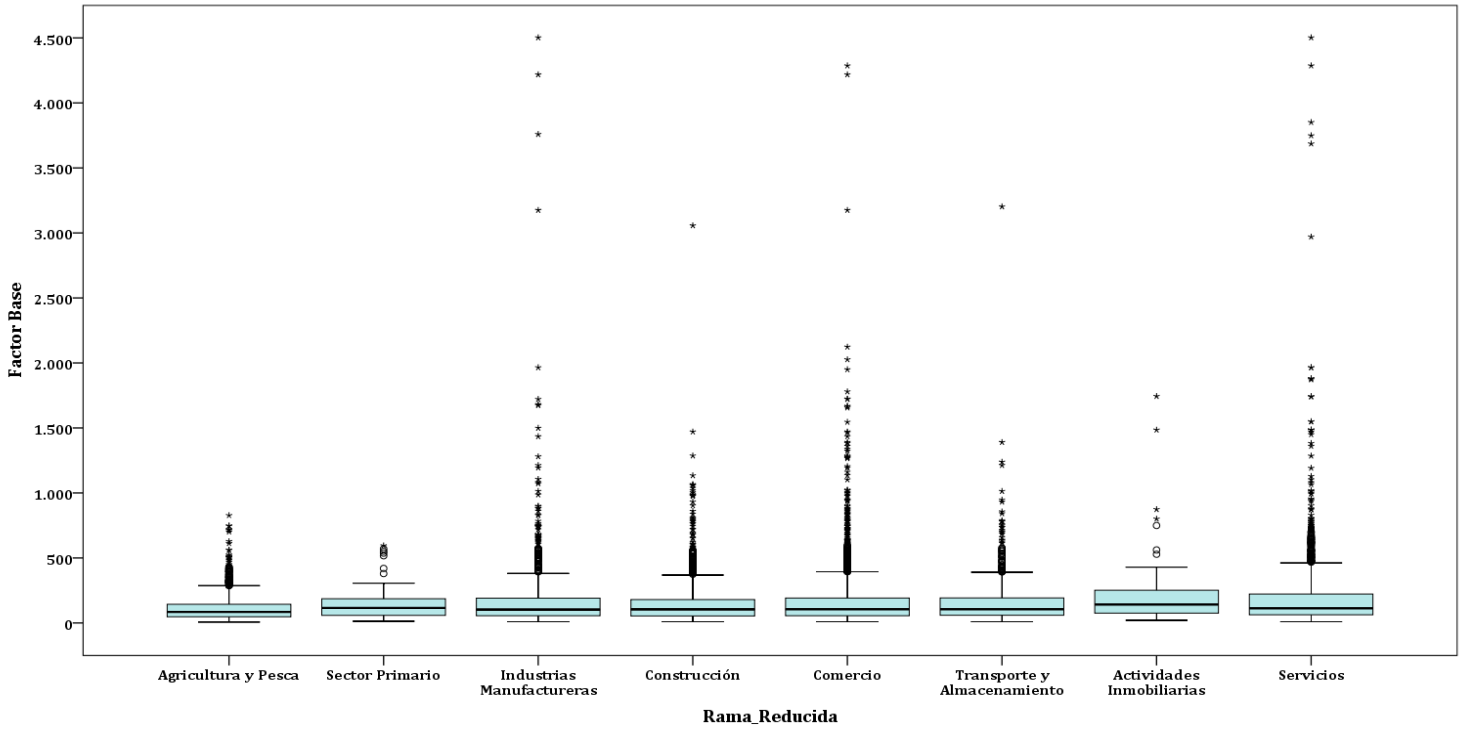
Cuadro 9. Estadísticas descriptivas del ponderador base

Estadísticas Descriptivas	Rama_Reducida							
	1 Agricultura y Pesca	2 Sector Primario	3 Industrias Manufactureras	4 Construcción	5 Comercio	6 Transporte y Almacenamiento	7 Actividades Inmobiliarias	11 Servicios
Recuento	1.405	61	1.070	1.118	2.407	699	63	1.997
Moda	69,5	70,44[a]	745,9	67,63[a]	176,1	35,5	38,66[a]	159,8
Mínimo	6,7	13,2	9,5	9,5	9,5	9,5	20,0	9,6
Percentil 05	18,4	27,2	25,0	25,1	25,0	25,7	38,1	28,1
Percentil 25	47,6	57,8	55,6	54,4	56,2	59,0	75,5	62,7
Mediana	84,7	115,7	102,6	104,8	105,3	105,8	142,1	112,6
Percentil 75	144,0	185,3	190,7	180,2	191,7	194,7	258,5	222,4
Percentil 95	315,9	537,2	569,8	509,5	540,0	520,7	800,7	609,2
Percentil 99	509,5	594,9	1.280,2	985,6	1.273,3	931,6	1.744,0	1.450,1
Máximo	827,8	594,9	4.502,6	3.056,5	4.286,7	3.202,7	1.744,0	4.502,6
Media	113,4	155,0	182,1	160,3	174,7	164,6	238,1	194,9
Error estándar de la media	2,75	19,03	9,50	5,86	5,11	7,83	39,21	6,65
Suma	159.351,8	9.458,0	194.799,0	179.248,7	420.595,9	115.030,0	14.998,0	389.242,9

a. Existen múltiples modos. Se muestra el valor más pequeño

Fuente: Elaboración propia

Gráfico 1. Ponderador base según rama de actividad económica reducida



Fuente: Elaboración propia

También existen valores extremos en el resto de las ramas. Sin embargo, los casos más preocupantes son los considerados “casos influyentes”, pues las características de un individuo pueden representar hasta 4.502 personas, es decir al 2,3% de la población estimada (suma del ponderador) de la rama de Industrias Manufactureras y un 1,1% para la rama de Servicios. Para minimizar el efecto en las estimaciones de ponderadores de esta magnitud, se implementó un método de verificación de valores extremos y suavizamiento de los mismos.

En el siguiente apartado se revisa la pertinencia de suavizar el ponderador base y el método de suavizamiento.

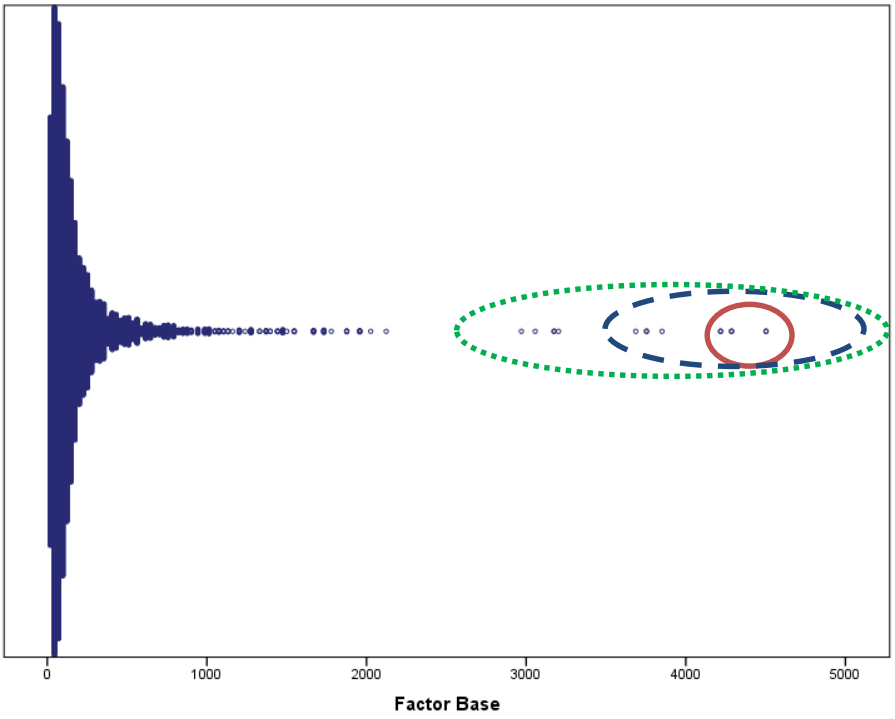
3.1.4. Suavizamiento de Ponderador Base

En la etapa de construcción del ponderador base, se observaron 3 casos con valores mayores a las 4.000 unidades. Las restantes observaciones poseen ponderadores inferiores o iguales a 3.250. Para identificar la presencia de casos influyentes y reducir su impacto, se implementó un procedimiento de suavizamiento de los factores de expansión que puede ser resumido en 5 pasos:

i. Inspeccionar la existencia de valores extremos en la distribución del ponderador,
ii. Determinar puntos de corte a partir de los cuales realizar el suavizamiento,
iii. Suavizar los valores extremos identificados,
iv. Estimar el error cuadrático medio (ECM) para los distintos puntos de corte,
v. Elegir la opción de corte que minimice el ECM,

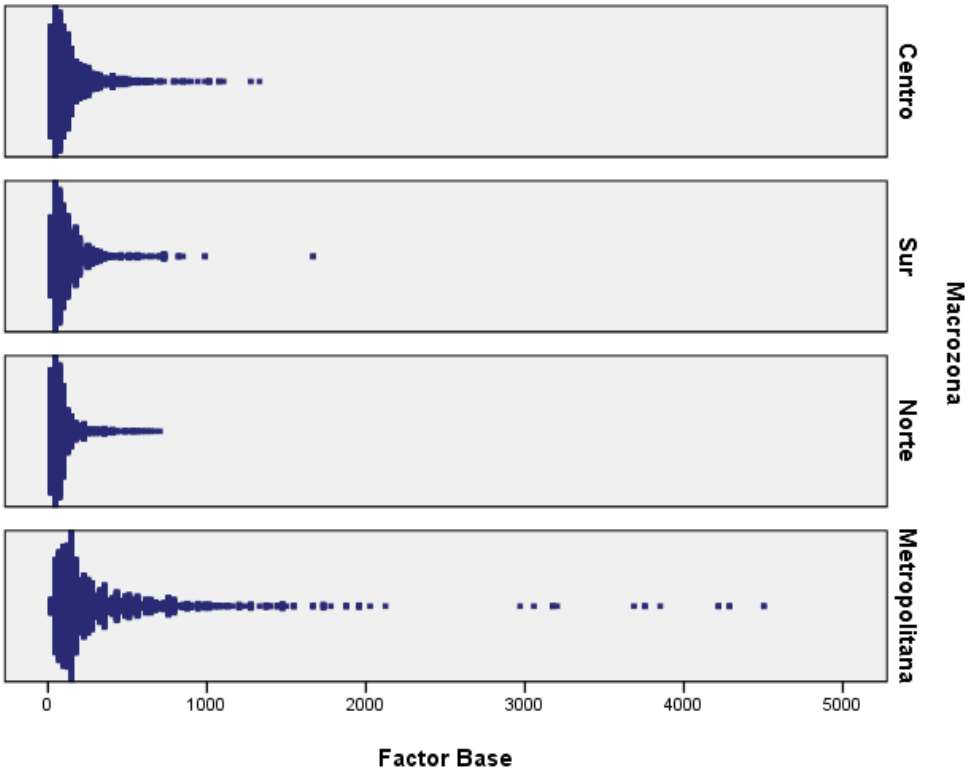
El gráfico 2, que muestra los factores de expansión base de forma ordenada, permite identificar al menos tres puntos de discontinuidad del ponderador base: tres casos extremos (los que se encuentran al interior de la elipse continua) que superan las 4.000 unidades, seis ponderadores, que se encuentran al interior de la elipse semi-continua, que poseen valores superiores a 3.800 y aquellos casos que superan 2.900 unidades. Sin embargo, al realizar este mismo análisis según macrozona, se pueden identificar discontinuidades en valores más pequeños. Por ejemplo, en la macrozona Centro y Sur, se observa una discontinuidad importante a partir del valor 1.400. De igual forma, se realiza el análisis según rama de actividad económica reducida, donde se identifican cortes en la mayoría de las ramas de actividad económica con al menos 3 cortes en la rama de Industria Manufactureras y también en la rama de servicios.

Gráfico 2. Dispersión del Factor de expansión base o inicial



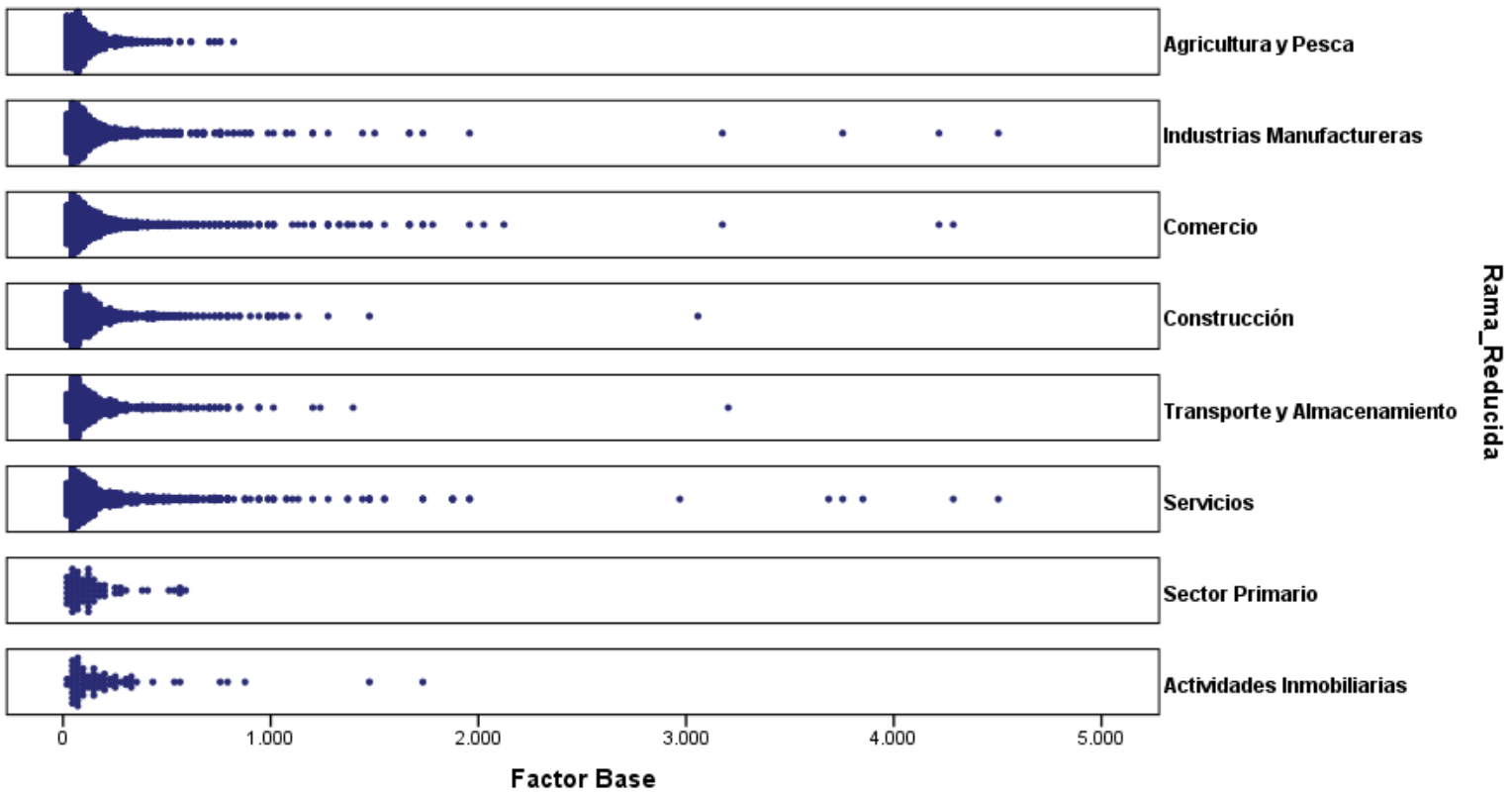
Fuente: Elaboración propia

Gráfico 3. Dispersión del Factor de expansión base o inicial, según macrozona



Fuente: Elaboración propia

Gráfico 4. Dispersión del Factor de expansión base o inicial, según rama reducida



Para inspeccionar la existencia de valores extremos, se utilizaron dos estrategias: (1) identificar discontinuidades, de forma visual, en la distribución del ponderador base; (2) identificar valores extremos a partir de una distancia determinada entre el ponderador promedio y cada valor del ponderador al interior de cada rama de actividad económica reducida[22].

Considerando lo anterior, se analizaron 7 puntos de corte distintos definidos como sigue:

$$\beta_r = r * \overline{F} \ , \quad con \ \overline{F} = \ \bar{F}^{base}_{Rjk,g} \ , \ con \ r \ = \ 4,5,6,7,8,9,10$$

De otra forma, los 7 puntos de corte son:

$$\beta_r = \begin{cases} \beta_4 \ = \ 4 * \overline{F} \\ \beta_5 \ = \ 5 * \overline{F} \\ \beta_6 \ = \ 6 * \overline{F} \\ \beta_7 \ = \ 7 * \overline{F} \\ \beta_8 \ = \ 8 * \overline{F} \\ \beta_9 \ = \ 9 * \overline{F} \\ \beta_{10} = \ 10 * \overline{F} \end{cases}$$

Por otro lado, para realizar el suavizamiento se procede a truncar aquellos ponderadores identificados como valores extremos de la siguiente forma,

$$T_{Rjk,g} = \begin{cases} F^{base}_{Rjk} & si \ \ F^{base}_{Rjk} \leq \beta_r \\ \beta_r & si \ F^{base}_{Rjk} > \beta_r \end{cases}$$

Donde,

g: Subíndice de la Rama de Actividad Reducida, de procedencia de los microemprendedores.

$\bar{F}^{base}_{Rjk,g}$: Corresponde al ponderador base promedio en Rama g

$T_{Rjk,g}$: Ponderador base truncado de la Región R, vivienda j persona k, perteneciente a la Rama g.

Si se suman todos los valores $T_{Rjk,g}$, se obtiene un total de unidades estimadas inferior que al sumar los ponderadores base, por lo tanto se deben distribuir los pesos

[22] Al chequear el gráfico 4, se observó que el comportamiento de los ponderadores es distinto al interior de cada rama, por lo tanto, se determinó realizar el suavizamiento de forma independiente al interior de cada una de estas.

faltantes en el resto de los ponderadores que no fueron truncados. Los pesos fueron distribuidos al interior de cada Rama de la siguiente forma:

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

Donde $F_{Rjk,g}^{SR}$ es el factor 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 ($\beta_r = r * \overline{F}$), mientras que el peso "sobrante" de los ponderadores truncados es distribuido sobre el resto de los ponderadores.

En el cuadro 10 se exponen las estadísticas descriptivas de cada uno de los ponderadores base suavizados según cada uno de los cortes establecidos. Se observa que el promedio del ponderador no sufre cambios. Sin embargo, el error estándar asociado decrece. Por otro lado, respecto a los estadísticos relacionados a la forma de la distribución, el coeficiente de asimetría disminuye respecto al obtenido con el ponderador base (la distribución es más simétrica mientras más bajo en valor absoluto es el coeficiente de asimetría) en cada uno de los suavizamientos. Asimismo, se aprecia una mejora importante en el coeficiente de curtosis, pues este estadígrafo se reduce tres cuartas partes, o más, con cada uno de los distintos puntos de suavizamiento. En términos generales, se observa que, mientras más exigente[23] es el punto de corte determinado, el número de unidades suavizadas es mayor y, por tanto, los estadígrafos tienen un mejor comportamiento (se reducen los valores extremos, se reduce la variabilidad, mejora el coeficiente de asimetría y curtosis, etc.).

[23] Un punto de suavizamiento es más exigente cuanto más cerca de la media se encuentra. Es decir, es más exigente cuando, por ejemplo, se considera suavizar a aquellos factores que se encuentran a una distancia de 4 veces la media en vez de 5 veces o alguna distancia mayor.

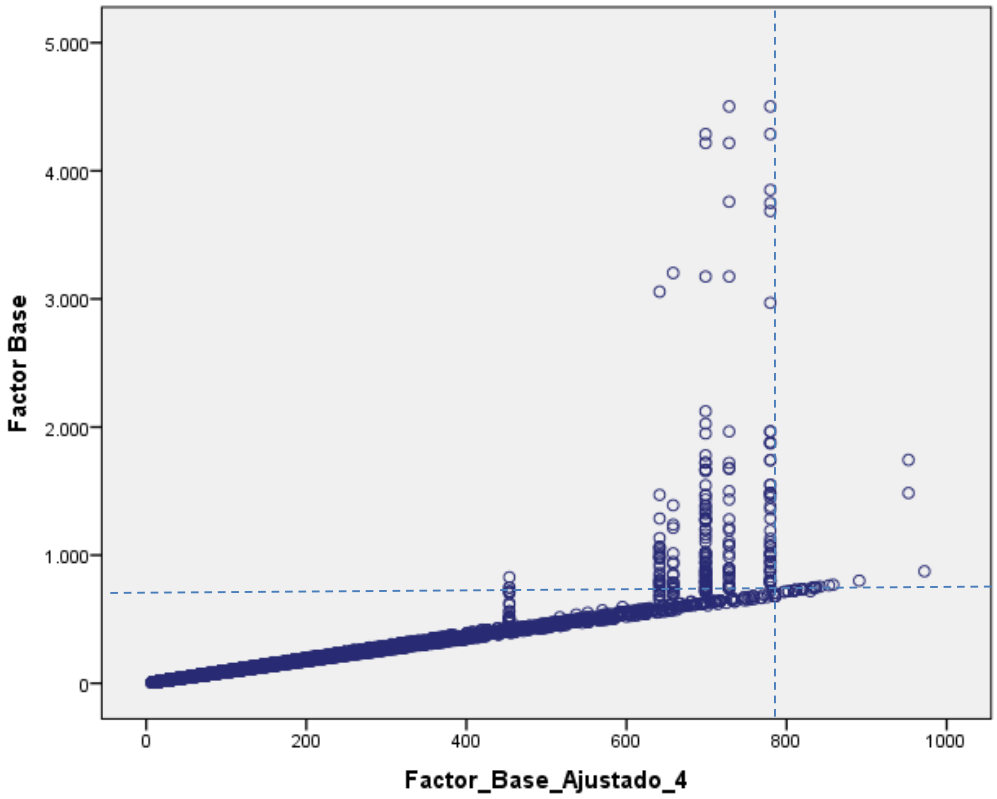
Cuadro 10. Estadísticas descriptivas del ponderador base y ponderador base suavizados en distintos puntos de corte

Estadístico		Ponderador Base	Ponderador Suavizado K4	Ponderador Suavizado K5	Ponderador Suavizado K6	Ponderador Suavizado K7	Ponderador Suavizado K8	Ponderador Suavizado K9	Ponderador Suavizado K10
Rango		4.496	965	1.184	1.422	1.660	1.737	1.801	1.942
Mínimo		7	7	7	7	7	7	7	7
Máximo		4.503	972	1.190	1.428	1.666	1.744	1.808	1.949
Suma		1.482.724	1.482.724	1.482.724	1.482.724	1.482.724	1.482.724	1.482.724	1.482.724
Media	**Estimación**	168	168	168	168	168	168	168	168
	Error estándar	2,615	1,750	1,882	1,979	2,060	2,130	2,181	2,226
Desviación estándar		246	164	177	186	193	200	205	209
Varianza		60.308	27.023	31.237	34.559	37.423	40.002	41.954	43.694
Asimetría	**Estimación**	8	2	2	3	3	3	4	4
	Error estándar	0,026	0,026	0,026	0,026	0,026	0,026	0,026	0,026
Curtosis	**Estimación**	96	4	6	9	12	15	19	22
	Error estándar	0,052	0,052	0,052	0,052	0,052	0,052	0,052	0,052

Fuente: Elaboración propia

Pese a lo indicado anteriormente, se debe revisar el comportamiento de aquellos valores del ponderador que siendo “grandes”, no caen en la categoría de valores extremos, pues al momento de redistribuir los pesos “sobrantes”, es probable que superen el umbral establecido. Esto puede ser visualizado en el gráfico 5, al observar que ciertos ponderadores base, con valores iguales o mayores a 800, luego de ser suavizados, superan el umbral establecido (línea vertical segmentada), lo que significa que el umbral establecido no es óptimo. Ante esto se descartaron inmediatamente dos de los 7 puntos establecidos (k_4 y k_5).

Gráfico 5. Dispersión del ponderador base versus ponderador suavizado en corte igual a k_4.



Fuente: Elaboración propia

Luego, para determinar el punto de corte donde se realizará finalmente el suavizamiento, se calculó un estadígrafo que diera cuenta del sesgo y de la variabilidad. Para esto se obtuvo el Error Cuadrático Medio (ECM) asociado a la variable de interés. Como en esta encuesta se pretende caracterizar los microemprendedores, se estableció analizar la estructura de la variable "Rama de actividad" (reducida a aquellas categorías más importantes[24]) y sobre estas categorías se calculó el sesgo utilizando la siguiente fórmula:

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

Tras calcular el sesgo de cada categoría, se calculó el efecto sobre la variable completa, a través de la suma del valor absoluto del sesgo de cada categoría. En el cuadro 11 se aprecia que los puntos de corte k7, K8, K9 y K10, son los que poseen menor sesgo en todas las categorías, y por tanto a nivel agregado también son los menos sesgados.

[24] Ver más detalles en capítulo 4.2

Cuadro 11. Estimación del sesgo de la estructura de la rama de actividad económica.

	Sesgo						
Rama Actividad Reducida	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
1 Agricultura y pesca	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2 Sector primario	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3 Industrias manufactureras	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
4 Construccion	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
5 Comercio	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
6 Transporte y almacenamiento	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
7 Actividades inmobiliarias	0,0011	0,0006	0,0002	0,0000	0,0000	0,0000	0,0000
11 Servicios	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Total	0,0013	0,0008	0,0004	0,0001	0,0000	0,0000	0,0000

Fuente: Elaboración Propia.

Finalmente, se calcula el ECM para cada categoría como:

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

Al revisar el cuadro 12, se observa que en términos del ECM, el valor mínimo de la mediana de los cortes, por cada rama de actividad reducida, corresponde al valor 0,000162, asociado a “Agricultura y pesca”. El primer corte asociado a este valor es el corte K7 y es el que se aplica a todas las ramas para el suavizamiento del factor de expansión base. Este corte es uno de los que posee menor sesgo y también menor ECM.

Cuadro 12. Estimación del ECM de la estructura de la rama de actividad económica.

Rama Actividad Reducida	**ECM**							**Mediana**
	k4	**k5**	**k6**	**k7**	**k8**	**k9**	**k10**	
1 Agricultura y pesca	0,000148	0,000154	0,000160	**0,000162**	0,000162	0,000162	0,000162	**0,000162**
2 Sector primario	0,005216	0,005216	0,005216	0,005216	0,005216	0,005216	0,005216	0,005216
3 Industrias manufactureras	0,000401	0,000399	0,000411	0,000433	0,000464	0,000504	0,000534	0,000433
4 Construccion	0,000323	0,000347	0,000366	0,000375	0,000384	0,000384	0,000384	0,000375
5 Comercio	0,000203	0,000198	0,000187	0,000177	0,000178	0,000182	0,000194	0,000187
6 Transporte y almacenamiento	0,000410	0,000476	0,000472	0,000460	0,000452	0,000447	0,000444	0,000452
7 Actividades inmobiliarias	0,006527	0,007280	0,008447	0,009805	0,010306	0,010306	0,010306	0,009805
11 Servicios	0,000202	0,000214	0,000236	0,000245	0,000249	0,000258	0,000261	0,000245

Fuente: Elaboración Propia.

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

1. Por cada rama de actividad reducida, se elige la mediana de los errores cuadráticos medios obtenidos con todos los cortes, desde K4 a K10.
2. Se identifica al mínimo de estas medianas y la rama de actividad reducida asociada
3. Al interior de la rama reducida identificada en el paso anterior, se identifica el primer corte en forma ascendente (k4, k5, ⋯ , k10) que coincide con este mínimo de medianas de ECM.

En el cuadro 13 se observa que, en términos de distribución, al comparar el ponderador base versus el ponderador suavizado no se registran cambios de los factores de la Rama Sector Primario. Sin embargo, en las restantes ramas los valores extremos fueron suavizados. El mayor cambio se encuentra en la Rama Industrias Manufactureras y en la Rama Servicios, ambas con un valor máximo de 4.503 disminuyendo a 1.306 y 1.437 unidades respectivamente.

Cuadro 13. Estadísticas descriptivas del ponderador base y ponderador suavizado.

Estadísticas Descriptivas	Rama_Reducida																	
	Agricultura y Pesca		Sector Primario		Industrias Manufactureras		Construcción		Comercio		Transporte y Almacenamiento		Actividades Inmobiliarias		Servicios		Total	
	Factor Base	Factor k7	Factor Base	Factor k7	Factor Base	Factor k7	Factor Base	Factor k7	Factor Base	Factor k7	Factor Base	Factor k7	Factor Base	Factor k7	Factor Base	Factor k7	Factor Base	Factor k7
Recuento	1.405	1.405	60	60	1.070	1.070	1.118	1.118	2.408	2.408	699	699	63	63	1.997	1.997	8.820	8.820
Moda	70	70	70	70	746	1.274	68	69	176	1.223	35	36	39	39	160	1.364	35	1.223
Mínimo	7	7	13	13	9	10	9	10	9	10	9	10	20	20	10	10	7	7
Percentil 05	18	18	26	26	25	27	25	25	25	26	26	26	38	38	28	30	25	25
Percentil 25	48	48	58	58	56	60	54	55	56	58	59	60	75	76	63	66	56	58
Mediana	85	85	116	116	103	110	105	106	105	110	106	108	142	143	113	119	102	106
Percentil 75	144	144	186	186	191	205	180	183	192	199	195	199	258	260	222	235	186	193
Percentil 95	316	316	544	544	570	614	509	517	540	562	521	532	801	805	609	644	521	543
Percentil 99	510	510	595	595	1.280	1.274	986	1.000	1.273	1.223	932	953	1.744	1.666	1.450	1.364	1.038	1.076
Máximo	828	794	595	595	**4.503**	**1.306**	3.057	1.122	4.287	1.254	3.203	1.152	1.744	1.666	**4.503**	**1.437**	4.503	1.666
Media	113	113	157	157	182	182	160	160	175	175	165	165	238	238	195	195	168	168
Error estándar de la media	3	3	19	19	10	7	6	5	5	4	8	7	39	39	7	5	3	2
Suma	159.352	159.352	9.416	9.416	194.799	194.799	179.249	179.249	420.638	420.638	115.030	115.030	14.998	14.998	389.243	389.243	1.482.724	1.482.724

Fuente: Elaboración Propia

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

3.2. Ponderador ajustado por falta de respuesta

En las encuestas de hogares se puede observar falta de respuesta de sus unidades por diversas causas, por ejemplo: no se identifica la dirección, no contacto con el informante, informante cambia de domicilio, informante con dificultad física o mental, rechazo de la entrevista, entre otras incidencias de terreno.

En la V 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; a la fecha de la visita el informante ha fallecido; al momento de la visita el informante se ha cambiado de domicilio o se encuentra fuera del país; al momento de la visita el informante posee dificultades físicas o mentales para contestar la encuesta; el informante no domina el idioma bajo el cual se realiza la encuesta; se impidió el acceso a la vivienda de residencia del informante (administrador, conserjes, etc. niegan el acceso a la vivienda).

Existen otras causas de no respuesta que quedan fuera del ámbito de corrección del factor de expansión, ya que corresponden a viviendas o personas sin encuestar debido a que no debieron pertenecer al marco de muestreo y, por lo tanto, no debieron ser seleccionados para responder la V EME (técnicamente no elegibles). Estos casos incluyen situaciones, donde la vivienda de residencia del informante ha cambiado de estado - colectiva, de uso temporal, desocupada temporalmente, incendiada, demolida etc.- (viviendas no elegibles) o, por otro lado, se identifica que los individuos fueron clasificados erróneamente como microemprendedores en la ENE (individuos no elegibles).

En la V EME, de un total de 8.199 viviendas seleccionadas, se seleccionaron 8.820 microemprendedores. De éstos, 8.444 fueron clasificados como elegibles (95,7%), de los cuales 7.492 respondieron la encuesta[25]. Por lo tanto, la tasa de respuesta de la EME, ajustada por elegibilidad, es de 88,7%.

Es posible que la falta de respuesta afecte sólo la precisión de la estimación. Sin embargo, si existe alguna relación entre las unidades faltantes y la variable de interés,

[25] Mayor detalle ver Anexo N° 2

es posible obtener estimaciones sesgadas. Por lo tanto, es recomendable realizar algún método de ajuste para compensar estas pérdidas, y mitigar dichos problemas.

El método a implementar para compensar la falta de respuesta fue el método de estratificación mediante "propensity score". De acuerdo, a lo indicado por Valliant[26], este método consiste en modelar la probabilidad de respuesta en la V 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 ($P(R_i^* > \theta)$). Dentro de los modelos paramétricos, se utilizan generalmente tres, los que responde a distintas características:

i. **Modelo Probit**. La probabilidad es modelada como si los valores fueran iguales a los de la función de distribución acumulada de la Normal. Por lo tanto, está bajo un supuesto de Normalidad.
ii. **Modelo Logístico**. Si bien modela la probabilidad de responder al igual que un modelo probit, la diferencia fundamental se encuentra en la función de enlace[27] (expresión matemática), que si bien es simétrica, ésta no requiere un supuesto de normalidad.
iii. **Modelo c-log-log.** La probabilidad de responder es modelada bajo la función de enlace de la distribución log-Weibull. El uso de este modelo es equivalente a suponer que el error asociado al proceso de variables latentes (u_i), tiene una distribución de valores extremos.

Cabe mencionar que para implementar el modelo probit se debiera contar con un set de variables latentes que en conjunto tengan distribución normal. Sin embargo, en la V EME, el potencial conjunto de variables (sexo, tramo etario, categoría ocupacional, nivel educacional, etc.) corresponden a variables de tipo categóricas, lo que dificulta el cumplimiento de dicho supuesto. Por otro lado, para utilizar el modelo c-log-log se debiera suponer que en la V EME, el error asociado a la estimación de la probabilidad de responder – a través de un set de variables latentes - estaría explicado por un comportamiento anómalo o difícil de explicar. Como las viviendas y personas

[26] Mayor detalle ver Valliant, R. Drever, J. Kreuter, F. (2013, section 13.5) "Practical Tools for Designing and Weighting Survey Samples", New York. Springer.
[27] Mayor detalle ver Valliant, R. Drever, J. Kreuter, F. (2013, section 13.5, página 323) "Practical Tools for Designing and Weighting Survey Samples", New York. Springer.

seleccionadas ya participaron en la ENE, tanto la respuesta como el rechazo de éstos en la V EME, responden a un comportamiento más bien predecible.

De acuerdo a lo anterior y según el comportamiento de los datos de la V EME, el modelo adecuado a utilizar es el modelo logístico. Así, el método de estratificación para el ajuste de no respuesta puede ser resumido 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 del modelo elegido, calcular la probabilidad de responder de cada una de las unidades elegibles que fueron utilizadas en el modelo.
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[28].

Una vez creadas las celdas de ajustes[29], se procede a estimar el factor de ajuste por falta de respuesta, el cual está dado por la siguiente expresión:

$$\hat{R}_c^{NR} = \frac{\sum_{k \in S_c} F_{Rjk}^{base_{tr}}}{\sum_{k \in S_{c,R}} F_{Rjk}^{base}}$$

Donde:

c: Es el 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 factor de expansión base para la persona *k,* de la vivienda *j,* en la región *R.*

Así, la expresión del ponderador de no respuesta es,

$$F_{Rjk}^{NR} = F_{Rjk}^{base_{tr}} \cdot \hat{R}_c^{NR}$$

[28] Mayor detalle ver Anexo N°2

[29] Las celdas de ajuste son varias agrupaciones de individuos unidos por 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.

Así, de acuerdo a la metodología antes expuesta, son 6 las celdas en las cuales se realizarán los ajustes por falta de respuesta. En el cuadro 14 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 72% aproximadamente.

Cuadro 14. Total unidades elegibles, que responde y tasa de respuesta.

Celda Ajuste	Total Responde	Total Elegibles	Tasa de Respuesta	$\hat{R}_c^{NR}$
Total	**7.492**	**8.444**	**88,7%**	
1	828	1.407	58,8%	1,72
2	1.246	1.407	88,6%	1,18
3	1.332	1.408	94,6%	1,06
4	1.340	1.407	95,2%	1,07
5	1.359	1.407	96,6%	1,04
6	1.387	1.408	98,5%	1,02

Fuente: Elaboración Propia

En el cuadro 15 presenta las principales estadísticas descriptivas del ponderador base suavizado y del ponderador ajustado por falta de respuesta. En promedio, existe un aumento de los ponderadores al ser ajustados de aproximadamente un 12%, observándose en la rama Transporte y Almacenamiento el mayor crecimiento promedio de los ponderadores (19%). Por otro lado, el mayor ponderador se encuentra en la rama Servicios el cual no supera las 2.500 unidades. En el siguiente apartado se revisará la pertinencia de un nuevo suavizamiento de los ponderadores, utilizando las mismas estrategias aplicadas para el ponderador base.

Cuadro 15. Estadísticas descriptivas del ponderador ajustado por falta de respuesta.

Estadísticas Descriptivas			Recuento	Moda	Mínimo	Percentil 05	Percentil 25	Mediana	Percentil 75	Percentil 95	Percentil 99	Máximo	Media	Error estándar de la media	Suma
Rama Reducida	Agricultura y Pesca	Factor Base Suavizado	1.347	70	7	18	48	84	148	312	510	794	114	3	153.138
		F_base_NR	1.240	145	7	21	56	92	162	335	597	901	126	3	156.247
	Sector Primario	Factor Base Suavizado	57	70	13	27	66	116	185	551	595	595	159	20	9.074
		F_base_NR	50	73	16	37	73	124	204	568	950	950	181	26	9.033
	Industrias Manufactureras	Factor Base Suavizado	1.031	1.274	10	27	59	109	205	609	1.274	1.306	180	7	185.639
		F_base_NR	942	15	12	28	64	115	228	751	1.358	2.195	202	8	190.548
	Construcción	Factor Base Suavizado	1.071	69	10	25	55	107	184	521	1.001	1.122	162	5	173.471
		F_base_NR	920	21	12	28	63	119	212	586	1.086	1.628	183	7	168.205
	Comercio	Factor Base Suavizado	2.307	1.223	10	26	58	109	198	562	1.223	1.254	174	4	400.921
		F_base_NR	2.071	2.106	11	26	61	116	216	658	1.303	2.160	196	6	405.215
	Transporte y Almacenamiento	Factor Base Suavizado	665	36	10	26	60	109	199	525	953	1.152	165	7	109.998
		F_base_NR	571	38	11	29	68	127	238	572	1.237	1.984	197	9	112.558
	Actividades Inmobiliarias	Factor Base Suavizado	60	39	20	38	76	140	253	842	1.666	1.666	240	40	14.407
		F_base_NR	53	41	28	41	90	178	359	915	1.592	1.592	283	45	15.001
	Servicios	Factor Base Suavizado	1.906	1.364	10	30	66	118	235	644	1.364	1.437	195	5	371.222
		F_base_NR	1.645	21	11	32	70	129	270	717	1.415	2.475	219	6	361.063
	Total	Factor Base Suavizado	8.444	1.223	7	25	58	106	192	542	1.077	1.666	168	2	1.417.870
		F_base_NR	7.492	145	7	27	63	115	214	619	1.261	2.475	189	3	1.417.870

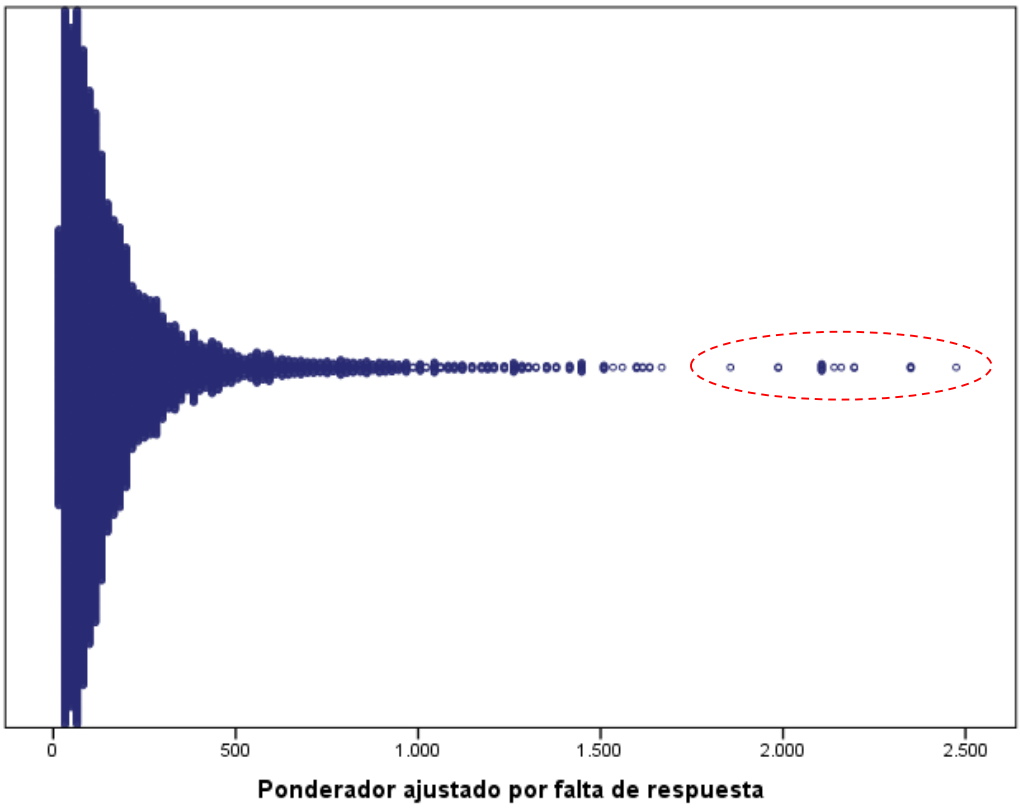
Fuente: Elaboración propia

3.2.1. Suavizamiento del Ponderador ajustado por falta de respuesta

Para los ponderadores ajustados por no respuesta, se evaluó la pertinencia de realizar suavizamiento de acuerdo al último punto de corte o criterio establecido para el ponderador base, k_7.

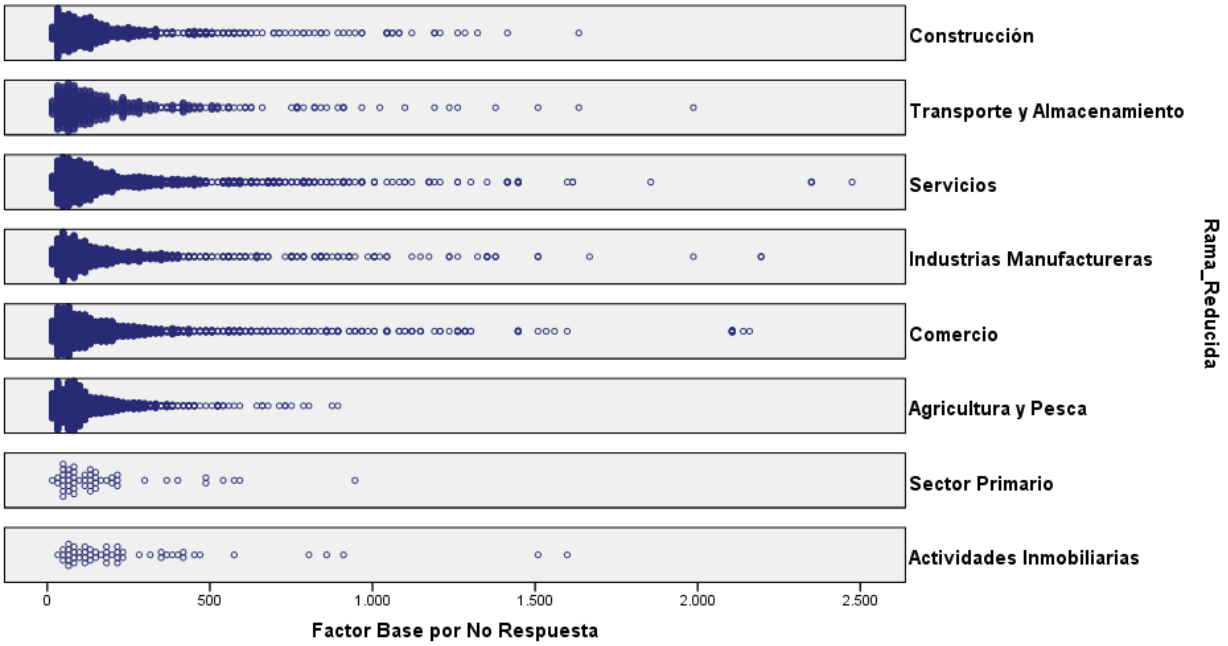
En el gráfico 6 se observa que existen valores grandes para el ponderador ajustado por falta de respuesta, no obstante, se debe evaluar si, de acuerdo a los criterios establecidos, son o no valores extremos.

Gráfico 6. Distribución de Factor ajustado por falta de respuesta.



Fuente: Elaboración propia

Gráfico 7. Distribución del Factor ajustado por falta de respuesta, según rama de actividad



Fuente: Elaboración propia

El gráfico 7, que muestra los factores de expansión base de forma ordenada en cada rama, permite identificar al menos dos puntos de discontinuidad del ponderador ajustado por falta de respuesta tanto en la rama de Servicios, Industrias manufactureras y comercio.

.

Cuadro 16. Estadísticas Descriptivas del Factor ajustado por falta de respuesta y ponderador por falta de respuesta suavizado

Estadístico		Ponderador Base	Ponderador Suavizado K4	Ponderador Suavizado K5	Ponderador Suavizado K6	Ponderador Suavizado K7	Ponderador Suavizado K8	Ponderador Suavizado K9	Ponderador Suavizado K10
Rango		2.467	1.125	1.408	1.584	1.584	1.749	1.968	2.188
Mínimo		7	7	7	7	7	7	7	7
Máximo		2.475	1.132	1.415	1.592	1.592	1.756	1.975	2.195
Suma		1.417.870	1.417.870	1.417.870	1.417.870	1.417.870	1.417.870	1.417.870	1.417.870
Media	**Estimación**	189	189	189	189	189	189	189	189
	Error estándar	2,711	2,163	2,348	2,479	2,564	2,611	2,646	2,681
Desviación estándar		235	187	203	215	222	226	229	232
Varianza		55.054	35.059	41.314	46.027	49.245	51.086	52.460	53.869
Asimetría	**Estimación**	4	2	2	3	3	3	3	4
	Error estándar	0,028	0,028	0,028	0,028	0,028	0,028	0,028	0,028
Curtosis	**Estimación**	19,447	3,537	6,028	8,533	10,805	12,630	14,510	16,969
	Error estándar	0,057	0,057	0,057	0,057	0,057	0,057	0,057	0,057

Fuente: Elaboración propia

En el cuadro 16 se exponen las estadísticas descriptivas de cada uno de los ponderadores por falta de respuesta suavizados según cada uno de los cortes establecidos. Se observa que el promedio del ponderador, no sufre cambios. Sin embargo, el error estándar del promedio asociado decrece. Por otro lado, respecto a los estadísticos relacionados a la forma de la distribución, el coeficiente de asimetría mejora con los primeros cortes (la distribución se hace más simétrica) con cada uno de los suavizamientos. Asimismo, se aprecia una mejora importante en el coeficiente de curtosis en los primeros cortes también, pues este estadígrafo se reduce de 19,447 para el ponderador base, a 3,537 en el primer corte K4 y va aumentando a medida que aumenta el corte de suavizamiento.

Cuadro 17. Estimación del sesgo de la estructura de la rama de actividad económica.

	Sesgo						
Rama Actividad Reducida	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
1 Agricultura y pesca	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2 Sector primario	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3 Industrias manufactureras	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
4 Construcción	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
5 Comercio	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
6 Transporte y almacenamiento	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
7 Actividades inmobiliarias	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
11 Servicios	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Total	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000

Cuadro 18. Estimación del ECM de la estructura de la rama de actividad económica.

Rama Actividad Reducida	**ECM**							**Mediana**
	k4	**k5**	**k6**	**k7**	**k8**	**k9**	**k10**	
1 Agricultura y pesca	0,000169	0,000175	0,000184	0,000189	0,000190	0,000190	0,000190	0,000189
2 Sector primario	0,007489	0,007316	0,007285	0,007285	0,007285	0,007285	0,007285	0,007285
3 Industrias manufactureras	0,000446	0,000426	0,000428	0,000447	0,000463	0,000479	0,000491	0,000447
4 Construccion	0,000377	0,000408	0,000442	0,000457	0,000465	0,000465	0,000465	0,000457
5 Comercio	0,000104	0,000114	0,000127	**0,000137**	0,000139	0,000141	0,000143	**0,000137**
6 Transporte y almacenamiento	0,000572	0,000580	0,000626	0,000668	0,000686	0,000685	0,000684	0,000668
7 Actividades inmobiliarias	0,003408	0,003211	0,003178	0,003178	0,003178	0,003178	0,003178	0,003178
11 Servicios	0,000248	0,000264	0,000273	0,000280	0,000284	0,000288	0,000294	0,000280

En el cuadro 19 se observa que, en términos de distribución, al comparar el ponderador ajustado por falta de respuesta F_{Rjk}^{NR} versus el ponderador suavizado F_{Rjk}^{NRS} de éste, no se registran cambios de los factores[30] en la rama de Sector Primario y Actividades Inmobiliarias. Sin embargo, en las ramas restantes los valores extremos fueron suavizados. El mayor cambio se encuentra en la rama de Servicios donde el valor máximo 2.475 disminuye a 1.536 unidades.

[30] No se registran cambios en el sector primario, por ejemplo, porque el corte de suavizamiento K7 aplicado consiste en truncar todos los factores que están a una distancia de 7 veces el promedio del ponderador base (aproximadamente los que están sobre 1267=181*7) pero resulta que el máximo del ponderador base en este sector es 950, por lo cual no es afectado.

Cuadro 19. Estadísticas Descriptivas del ponderador ajustado por falta de respuesta y suavizamiento

Estadísticas Descriptivas	Rama_Reducida																	
	Agricultura y Pesca		Sector Primario		Industrias Manufactureras		Construcción		Comercio		Transporte y Almacenamiento		Actividades Inmobiliarias		Servicios		Total	
	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}	F_{Rjk}^{NR}	F_{Rjk}^{NRS}
Recuento	1.240	1.240	50	50	942	942	920	920	2.071	2.071	571	571	53	53	1.645	1.645	7.492	7.492
Moda	145	145	73	73	15	1.416	21	21	2.106	1.370	38	38	41	41	21	1.536	145	1.370
Mínimo	7	7	16	16	12	12	12	13	11	11	11	12	28	28	11	11	7	7
Percentil 05	21	21	37	37	28	28	28	28	26	27	29	29	41	41	32	33	27	27
Percentil 25	56	56	73	73	64	65	63	63	61	62	68	69	90	90	70	71	63	64
Mediana	92	92	124	124	115	117	119	120	116	119	127	128	178	178	129	130	115	117
Percentil 75	162	162	204	204	228	232	212	212	216	221	238	240	359	359	270	273	214	217
Percentil 95	335	335	568	568	751	762	586	588	658	673	572	578	915	915	717	725	619	626
Percentil 99	597	597	950	950	1.358	1.377	1.086	1.089	1.303	1.333	1.237	1.249	1.592	1.592	1.415	1.431	1.261	1.271
Máximo	901	882	950	950	2.195	1.416	1.628	1.280	2.160	1.370	1.984	1.386	1.592	1.592	2.475	1.536	2.475	1.592
Media	126	126	181	181	202	202	183	183	196	196	197	197	283	283	219	219	189	189
Error estándar de la media	3	3	26	26	8	8	7	7	6	5	9	9	45	45	6	6	3	3
Suma	156.247	156.247	9.033	9.033	190.548	190.548	168.205	168.205	405.215	405.215	112.558	112.558	15.001	15.001	361.063	361.063	1.417.870	1.417.870

Posteriormente, utilizando como insumo el ponderador ajustado por falta de respuesta suavizado F_{Rjk}^{NRS}, en el corte k7, se realiza la calibración de este factor, el cual se detalla a continuación.

3.3. Ponderador calibrado

En general, en todas las encuestas de hogares el ponderador final o factor de expansión se encuentra 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 estimado[31] 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; mientras que la Encuesta Casen 2015 fue calibrada al stock poblacional residente en viviendas particulares ocupadas según región, con fecha 30 de noviembre.

En los dos ejemplos expuestos anteriormente 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 V EME está anclada a la población de referencia del trimestre MAM 2017 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 decidió utilizar la estimación del total de microemprendedores del trimestre MAM 2017 de la ENE actualizada al período del trabajo de campo de la V EME. Para esto, se utilizó el crecimiento proyectado (crecimiento natural de la población según las estimaciones del CENSO 2002) para el mes central del período de levantamiento de la encuesta, es decir junio 2017. En definitiva, la estimación utilizada en la calibración del ponderador de la EME se obtuvo a través de los siguientes pasos:

1. Primero, se considera toda la información levantada para la ENE en el período MAM 2017, es decir, todos los microemprendedores que fueron clasificados como

[31] Estimaciones realizadas por el departamento de demografía del INE, a partir de información auxiliar.

tales, sin importar que a futuro puedan cambiar de condición. Esta población, expandida a junio de 2017 es la que se toma como referencia para la calibración de los microemprendedores de la V EME.

2. Segundo, se calcula un nuevo factor de expansión, considerando las proyecciones de población a junio de 2017.

En el período MAM 2017, la ENE utilizó el siguiente cálculo:

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

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}: Corresponde al total de personas de sexo y tramo de edad s, en la vivienda j, del conglomerado i, del estrato ENE h.

P_{hs}^{4}: Total de población del sexo s, del estrato ENE h, proyectado al 15 de abril de 2017.

Para obtener la estimación del total de microemprendedores para la V EME, se calculó con la misma fórmula, sin embargo, el stock poblacional utilizado corresponde al proyectado con fecha junio 2017. Es decir,

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

En el cuadro 16, se presenta el total de microemprendedores estimado a partir de la publicación de la ENE, periodo MAM 2017; y según total de personas estimado con la información levantada en MAM 2017, pero con proyecciones actualizadas a la fecha de levantamiento de la EME (en adelante I_{gs}.)

Como se observa en el cuadro 16, el total de “microemprendedores” estimados y publicados oficialmente son 2.097.712 personas. Sin embargo, al actualizar las proyecciones de población este total asciende a 2.103.974, lo que equivale a un incremento del 0,30% a nivel nacional, un 0,40% en la macrozona Norte, un 0,32% en el Centro, un 0,38% en el Sur y un 0,21% en la Región Metropolitana.

Cuadro 20. Total de microemprendedores estimados a partir de la ENE- Período MAM 2017

		Total Microemprendedores	
Macrozona	**Sexo**	**Factor Expansión Oficial ENE - MAM**	**Factor Expansión Información ENE - ajustado a Junio**
Total	Hombre	1.278.697	1.282.662
	Mujer	819.015	821.312
	Total	**2.097.712**	**2.103.974**
Norte	Hombre	154.956	155.590
	Mujer	101.294	101.673
	Total	**256.251**	**257.263**
Centro	Hombre	381.851	383.133
	Mujer	214.081	214.730
	Total	**595.933**	**597.863**
Sur	Hombre	258.838	259.836
	Mujer	137.126	137.643
	Total	**395.964**	**397.479**
Metropolitana	Hombre	483.052	484.103
	Mujer	366.513	367.265
	Total	**849.565**	**851.368**

Fuente: Elaboración propia

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 2017. Es decir, se estimó el total de microemprendedores a través de la utilización del ponderador de no respuesta, tal como se muestra a continuación:

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

Donde:

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 a fuentes externas (ENE), y la estimación de la encuesta obtenida en el paso (1):

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

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.

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

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 2017, actualizado al crecimiento poblacional de junio 2017 - mes central de levantamiento de EME.

Cuadro 21. Estadísticas descriptivas del ponderador ajustado por falta de respuesta suavizado y calibrado a stock de microemprendedores, según sexo.

Estadísticas descriptivas	Sexo					
	Hombre		Mujer		Total	
	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}	F_{Rjk}^{NR}	F_{gjs}^{cal}
Recuento	4.495	4.495	2.997	2.997	7.492	7.492
Moda	1.370	2.041	38	57	1.370	2.041
Mínimo	8	12	7	10	7	10
Percentil 05	28	41	26	40	27	41
Percentil 25	65	99	61	91	64	97
Mediana	119	177	114	169	117	174
Percentil 75	217	331	217	325	217	329
Percentil 95	622	930	629	906	626	924
Percentil 99	1.279	1.860	1.232	1.722	1.271	1.813
Máximo	1.536	2.342	1.592	2.225	1.592	2.342
Media	190	285	188	274	189	281
Error típico de la media	3	5	4	6	3	4
Suma	855.445	1.282.662	562.425	821.312	1.417.870	2.103.974

Fuente: Elaboración propia

En el cuadro 22 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 Sur un microemprendedor, hombre, representaba a 1.536 personas, sin embargo, al ajustar según sexo y macrozona, esta persona representa 2.342 individuos. Cabe señalar, que en el caso de una mujer de la misma macrozona su ponderador cambió de 1.592 a 2.225, lo cual puede ser revisado en el cuadro 21. En todo caso, los valores extremos dados al comienzo, ya han sido suavizados (valores sobre 4.000 ahora apenas sobrepasan los 2.000). 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.

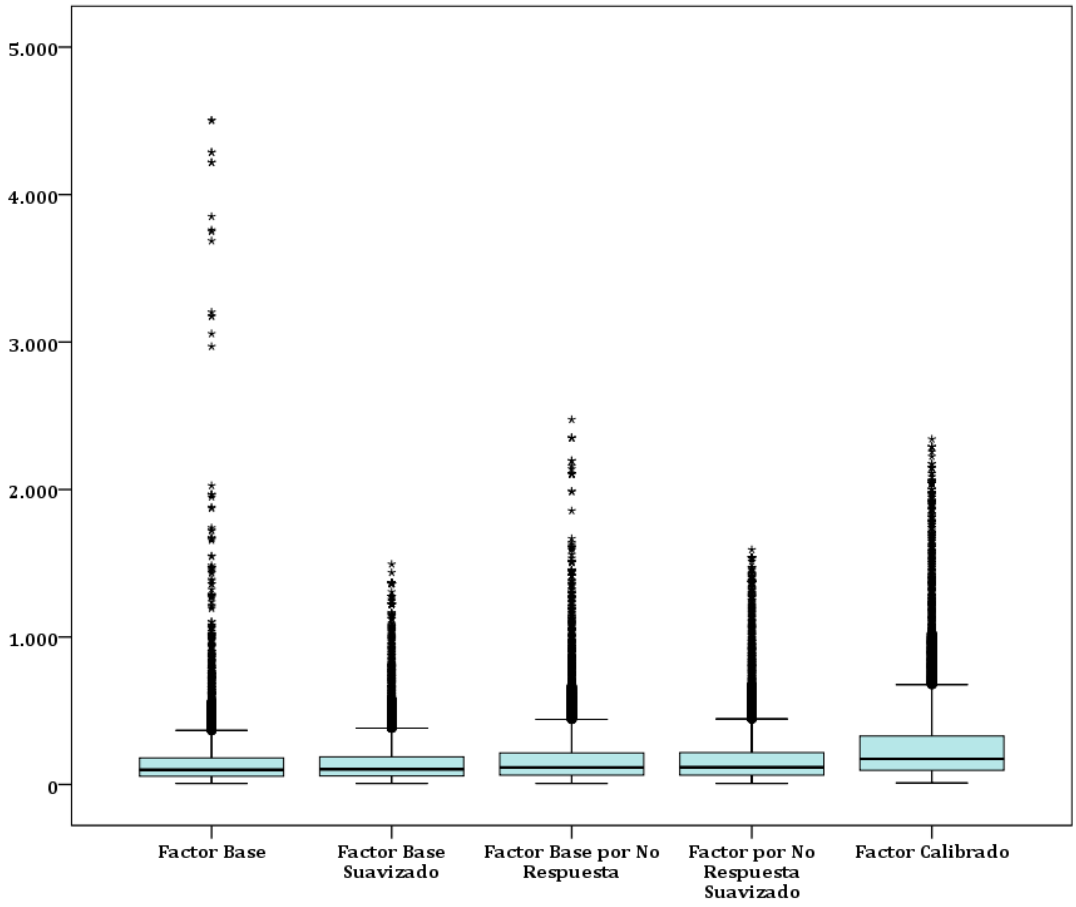
Cuadro 22. Estadísticas descriptivas del ponderador ajustado por falta de respuesta suavizado F_{Rjk}^{NRS} y calibrado a stock de microemprendedores, según macrozona.

Estadísticas descriptivas	Macrozona									
	Norte		Centro		Sur		Región Metropolitana		Total	
	F_{Rjk}^{NRS}	F_{gjs}^{cal}	F_{Rjk}^{NRS}	F_{gjs}^{cal}	F_{Rjk}^{NRS}	F_{gjs}^{cal}	F_{Rjk}^{NRS}	F_{gjs}^{cal}	F_{Rjk}^{NRS}	F_{gjs}^{cal}
Recuento	1.585	1.585	2.596	2.596	1.651	1.651	1.660	1.660	7.492	7.492
Moda	145	226	219	301	80	136	1.370	2.041	1.370	2.041
Mínimo	7	10	11	15	12	21	17	23	7	10
Percentil 05	17	27	27	39	30	49	56	80	27	41
Percentil 25	44	67	60	85	61	100	129	188	64	97
Mediana	77	117	114	162	100	167	231	330	117	174
Percentil 75	128	195	197	279	187	307	445	647	217	329
Percentil 95	316	483	470	668	386	636	1.161	1.676	626	924
Percentil 99	507	789	890	1.236	773	1.252	1.431	2.056	1.271	1.813
Máximo	980	1.524	1.416	1.947	1.370	2.342	1.592	2.290	1.592	2.342
Media	106	162	162	230	145	241	354	513	189	281
Error típico de la media	3	4	3	5	3	6	8	12	3	4
Suma	168.678	257.263	421.550	597.863	240.059	397.479	587.583	851.368	1.417.870	2.103.974

Fuente: Elaboración Propia

En el gráfico 8, se puede observar visualmente todos los ponderadores desde el Factor Base hasta el Factor Calibrado

Gráfico 8. Gráfico de caja para el ponderador base, base suavizado, ajustado por no respuesta, ajustado por no respuesta suavizado y Calibrado.



Se puede apreciar en el gráfico los distintos ajustes que son aplicados al factor base. El primer suavizamiento elimina todos los valores extremos, aquellos que sobrepasaban 7 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, este se muestra 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 2017).

4. ESTIMACIÓN DE VARIANZA

De acuerdo a lo descrito en los apartados anteriores, el diseño muestral de la V EME es bifásico y complejo, por lo tanto, las probabilidades de selección de los individuos son desiguales. Así, cualquier análisis que se desee realizar a partir de la V EME, se debe hacer utilizando el factor de expansión. Si no se usa el factor de expansión se obtendrán estimaciones sesgadas y que sólo darán cuenta del comportamiento de las unidades seleccionadas, pero no de la población total.

Por otro lado, al momento de realizar un estudio, se sugiere a los analistas, estimar la variabilidad muestral asociada a la estimación puntual. Para ello, existen diversos paquetes estadísticos en STATA (svyset), SPSS (csplan, en muestras complejas), R (Survey, svydesign), SAS (PROC surveyfreq, Proc surveymeans), etc. que utilizan fórmulas convencionales (aun cuando muchas veces éstas no tienen una fórmula explícita) para la estimación de las varianzas, bajo un supuesto de muestreo aleatorio con reemplazo y con pesos, lo que facilita los cálculos.

Para la utilización de los paquetes estadísticos de forma apropiada, se requiere identificar las variables que definen el diseño muestral de la encuesta. 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 principal la estructura de la rama de actividad de los microemprendedores. Como existen algunas categorías como pesca; electricidad, gas y agua; entre otras, en las cuales se observa una baja prevalencia, se crea una variable más agregada denominada “rama reducida”. Es sobre esta variable que en la sección 4.2 se realizan las estimaciones de los errores.

4.1. Variables que identifican el diseño

El diseño muestral de la V EME posee las características de un diseño muestral bifásico y complejo. La primera fase se caracteriza por poseer un diseño muestral complejo, pues es estratificado geográficamente y la selección de las viviendas que participan en la ENE se realizó en dos etapas, seleccionando en primera instancia los conglomerados de forma sistemática y con probabilidad proporcional al tamaño, mientras que las viviendas en su interior fueron seleccionadas de forma sistemática pero con igual probabilidad. La segunda fase se caracteriza porque las viviendas se seleccionaron a partir de un listado de viviendas de la ENE tal que en su interior residen al menos un microemprendedor, luego en su interior se seleccionaron

sistemáticamente, tantos microemprendedores como actividades únicas se identifican en su interior.

En este contexto las variables que identifican el diseño muestral de ambas fases, corresponden a una variable llamada “Estrato” que identifica los estratos geográficos de la ENE; y una variable “IdDirectorio” que corresponde a una variable ficticia que identifica de forma única los conglomerados en la ENE. Las variables que identifican el diseño de la V EME, corresponden a las mismas de la ENE, ya que la selección de las unidades se realizó al interior de cada región de forma independiente, por lo tanto, como por construcción, los estratos de la ENE son interiores a las regiones, entonces el cruce Región versus Estrato da como resultado los mismos estratos de la ENE.

En general, cuanto más complejo es el diseño muestral bajo el cual se implementa una encuesta, más complejo se vuelve la forma de determinar los errores muestrales. Tanto así, que no existen fórmulas exactas y/o explícitas para esto. Sin embargo, paquetes estadísticos en software especializados, facilitan los cálculos a través de aproximaciones realizadas mediante distintos modelos o métodos de estimación, para lo cual se debe identificar las variables que definen el diseño muestral (estratos, conglomerados) y el factor de expansión apropiado (considerando todos los ajustes pertinentes).

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 sólo una unidad primaria con unidades elegibles; variabilidad de los tamaños de los conglomerados.

En el caso de la V EME, se observan principalmente tres dificultades: (1) Diseño muestral bifásico y complejo; (2) existen estratos de muestreo (los de la ENE) que poseen solo un conglomerado (manzana o sección); (3) el número de unidades seleccionadas y que responde en cada conglomerado es desigual y muy variable. A fin de minimizar los problemas señalados anteriormente, y siguiendo las recomendaciones internacionales[32], los errores fueron estimados a partir de modelos que buscan dar cuenta, lo más fielmente posible del diseño muestral. Para ello se agruparon estratos y conglomerados a fin de que estos nuevos pseudo-estratos y pseudo-conglomerados, garanticen la estimación de varianzas en cada nuevo estrato,

[32] Ver Capítulo 15.5 en Valliant *et al.* (2013).

y de ésta forma no subestimar los errores. A continuación, se detallan los procedimientos y criterios utilizados en la creación de dichas variables.

4.1.1. 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), esto es subestimar 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;

i. Primero se contabiliza, al interior de cada estrato original, el total de unidades[33] que participa en la encuesta. Si el estrato contiene menos de 30 (2•15) unidades, entonces deberá ser colapsado con otro.
ii. Se ordenan todos los estratos, geográficamente, de acuerdo a la división político administrativa en urbanos y rurales, y luego al interior de cada región según ordenamiento del estrato.
iii. 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 las 60 unidades.

De un total de 160 estratos que posee la ENE, en la V EME se seleccionaron unidades de todos los estratos, de los cuales 2 de estos estratos contienen a un solo conglomerado. Por otra parte, existen 60 estratos con 30 o menos unidades (microemprendedores). Así el total de pseudo-estratos creados desciende a 126.

[33] Cada estrato debe contener al menos 30 unidades o viviendas, que es equivalente, y en forma aproximada, a contar con 30 microemprendedores por estrato. A su vez, cada conglomerado debe contar aproximadamente con 15 viviendas, que al igual que antes, es equivalente a contar en forma aproximada con 15 microemprendedores.

Cuadro 23. Total de estratos y de pseudo-estratos, según macrozona.

Macrozona	Estratos	Pseudo-estrato
Total	**126**	**160**
Norte	22	25
Centro	48	63
Sur	21	25
Metropolitana	35	47

Fuente: Elaboración propia

4.1.2. 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 (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 que estos en conjunto reunieran entre 12 y 18 unidades (en promedio 15 viviendas).

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

i. 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.
ii. 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.
iii. Finalmente, al interior de cada pseudo-estrato se colapsan aquellos estratos 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 4.162 conglomerados, en 3.175 de estos conglomerados se seleccionaron microemprendedores, los que se transformaron en 532 pseudo-conglomerados.

En el cuadro 24 se expone el total de conglomerados y pseudo-conglomerados según macrozona.

Cuadro 24. Total de conglomerados y de pseudo-conglomerados, según macrozona

Macrozona	Conglomerados	Pseudo-Conglomerados
Total	**3.175**	**532**
Norte	621	110
Centro	1.257	182
Sur	602	114
Metropolitana	695	126

Fuente: Elaboración propia

4.2. Estimación de variables y varianzas en Spss y Stata

Diversos paquetes estadísticos poseen algoritmos que permiten la estimación de los errores muestrales bajo diseños muestrales 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 realizar algunos supuestos: se asume que la selección de las unidades, en las distintas etapas, se realizó 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 V EME son:

i. F_Calibrado: 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.
ii. VarStrat: variable que identifica el estrato de muestreo, tal que éste contiene al menos dos conglomerados, para garantizar la estimación de la varianza.

iii. VarUnit: variable que identifica al conglomerado, tal que éste contiene 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:

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

Considerando la estructura de la rama de actividad económica (CIIU rev. 3) para los microemprendedores como la variable de interés -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 1.

Tabla 1. Rama de actividad económica según CIIU Rev 3. 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

Fuente: Elaboración Propia

A continuación se presenta un resumen con la estimación de la rama de actividad reducida, en la cual fueron clasificados los microemprendedores, según las estimaciones realizadas en Spss[34].

Cuadro 25. Estructura de la Actividad económica en la cual se desenvuelven los microemprendedores- estimación realizada en SPSS

Rama de actividad económica	Estimación	Error típico	Intervalo de confianza al 95%		Coeficiente de variación	Efecto del diseño
			Inferior	Superior		
Agricultura y Pesca	11,5%	0,4%	10,7%	12,3%	0,036	1,265
Sector Primario	0,6%	0,1%	0,4%	0,9%	0,194	1,806
Industrias Manufactureras	13,4%	0,7%	12,1%	14,7%	0,049	2,768
Construcción	12,0%	0,6%	10,9%	13,1%	0,047	2,225
Comercio	28,3%	0,8%	26,8%	29,9%	0,028	2,382
Transporte y Almacenamiento	8,0%	0,5%	7,0%	9,1%	0,066	2,823
Actividades Inmobiliarias	1,0%	0,2%	0,7%	1,5%	0,188	2,778
Servicios	25,2%	0,9%	23,6%	27,0%	0,034	2,954
Total	100,0%	0,0%	100,0%	100,0%	0,000	

Fuente: Elaboración Propia

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 del 53% de los microemprendedores.

Respecto a la precisión de las estimaciones, se puede constatar que todas las ramas de actividad económica reducida presentan errores relativos[35] aceptables (menores de 30%), a excepción de las ramas ‘Sector primario’ y ‘Actividades inmobiliarias’, cuyos errores relativos son de aproximadamente 39% y 37%, respectivamente.

[34] Ver Anexo N°4

[35] El error relativo es aproximadamente el doble del coeficiente de variación.

BIBLIOGRAFÍA

1. Valliant, R. Drever, J. Kreuter, F. (2013). Practical Tools for Designing and Weighting Survey Samples", Springer, New York.

2. Heeringa, S., West, B., and Berglund, P. (2010). Applied Survey Data Analysis. Chapman and Hall, CRC Press, Boca Raton, Florida

3. Dobson, A. (2002) An Introduction to Generalized Linear Models. CRC Press.

4. Burgueño, M; García-Bastos, J; González-Buitrago, J.(1993). Las curvas ROC en la evaluación de pruebas diagnósticas. Medicina Clínica Vol. 104. Núm. 17.1.995. España.

5. Montgomery. D; Peck, E; Vining, G. (2006). Introducción al Análisis de Regresión Lineal. 1era Edición español. Cía. Editorial continental. México.

6. The American Association for Public Opinion Research (2011). Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys.

ANEXOS

1. Anexo N°1. Áreas de Difícil acceso o Alto Costo

Cuadro 26. Áreas geográficas excluidas del Marco de Muestreo del INE, clasificadas como ADA's.

Región	Nombre Provincia	Nombre Comuna	Total Viviendas Censo 2002
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
Aisén del General Carlos Ibáñez del Campo	Coihaique	Lago Verde	590
	Aisén	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
Total Viviendas ADA's			**16.046**

Fuente: Elaboración propia

2. Anexo N°2. Códigos de disposición última visita

En el cuadro 28, aparece el código de disposición de las viviendas en su última visita. Así, 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.

Cuadro 27. Códigos de disposición final de la última visita a la vivienda

Código de disposición de la última visita a la vivienda	Frecuencia	Porcentaje	Elegible
110 Encuesta completa	7.492	84,94	SÍ
211 Informante de la vivienda rechazó la entrevista	66	0,75	SÍ
212 Informante directo rechazó la entrevista	172	1,95	SÍ
213 Se interrumpió la entrevista al informante directo	8	0,09	SÍ
223 Se impidió el acceso a la vivienda	2	0,02	SÍ
224 Vivienda ocupada sin moradores presentes	325	3,68	SÍ
225 Informante directo no ubicable o no puede atender	362	4,10	SÍ
232 Informante con dificultad física, mental o cognitiva para contestar	5	0,06	SÍ
236 Anulación por falsificación	10	0,11	SÍ
290 Otra razón elegible	2	0,02	SÍ
317 Área peligrosa o de difícil acceso	9	0,10	NO
318 No fue posible localizar la dirección	14	0,16	NO
390 Otra razón de elegibilidad desconocida	1	0,01	NO
454 Vivienda en demolición, incendiada, destruida o erradicada	5	0,06	NO
461 Vivienda particular desocupada (en arriendo, en venta, otro.)	23	0,26	NO
462 Vivienda de veraneo o de uso temporal	4	0,05	NO
471 Muerte del informante	4	0,05	NO
472 Cambio de domicilio	169	1,92	NO
473 Informante fuera de marco	145	1,64	NO
490 Otra razón no elegible	2	0,02	NO
Total	**8.820**	**100,00**	**8.444**

Fuente: Elaboración propia

3. Anexo N°3. 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 V EME, se consideraron tres análisis de elegibilidad; 1) Descriptivo, 2) Modelación y 3) Sensibilidad del modelo. El objetivo del análisis descriptivo fue tener una primera aproximación de las variables que influyeron en la respuesta de las personas y de esta manera entender de forma intuitiva nuestro fenómeno de estudio. Luego, para la modelación 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. Finalmente, en la etapa de Sensibilidad del modelo se determinará qué "tan bueno" es nuestro ajuste, específicamente a través de la Curva ROC.

3.1. 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:

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

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:

$$\prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} = exp\left[\sum_{i=1}^{n} y_i log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} log(1-\pi_i)\right] \quad (2)$$

La cual es miembro de la familia exponencial.

Al considerar la siguiente función de enlace[36]:

[36] 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

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^t\beta \quad (3)$$

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:

$$\pi_i = \frac{exp\left(\boldsymbol{x}_i^t\beta\right)}{1 + exp\left(\boldsymbol{x}_i^t\beta\right)} = P\left(Y_i = 1/\boldsymbol{x}_i^t\beta\right) \quad (4)$$

La estimación de los parámetros se realiza mediante un proceso iterativo que aproxima la log-verosimilitud mediante el algoritmo de Newton Raphson o por aproximación Scoring de Fisher. A continuación, se detallan los pasos para la obtención de estos parámetros.

3.2. Estimación de Parámetros

3.2.1. Estimación Máxima verosimilitud

Sean $Y_1, \dots, Y_n$ n variables aleatorias independientes, es decir, cada una con función de densidad de probabilidad $f_i(y_i;\theta)$ donde el vector de parámetro $\theta = \left(\theta_1, \dots, \theta_p\right)^t$ es un elemento del espacio paramétrico Ω que comprende todos los valores a priori admisibles.

La distribución de densidad conjunta de n observaciones independientes $\boldsymbol{y} = (y_1, \dots, y_n)^t$ es:

---

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)]

$$f(y;\theta) = \prod_{i=1}^{n} f_i(y_i;\theta) = L(\theta, \boldsymbol{y}). \quad (5)$$

La expresión $L(\theta, \boldsymbol{y})$ es vista como una función del vector de parámetro desconocido θ dada la muestra $\boldsymbol{y}$ (o datos), denominada función de verosimilitud. A menudo, se trabaja con el logaritmo natural de la función de verosimilitud, llamada función de log - verosimilitud:

$$logL(\theta;\boldsymbol{y}) = \sum_{i=1}^{n} log\, f_i(y_i;\theta). \quad (6)$$

Para encontrar el conjunto de soluciones para el vector de parámetro θ, dada la muestra $\boldsymbol{y}$**,** que maximice la función de verosimilitud o log verosimilitud, consideramos el principio de máxima verosimilitud que postula la elección de $\hat{\theta}$ perteneciente al espacio paramétrico Ω que maximice la función de log-verosimilitud. Para esto, se define el estimador máximo verosímil como $\hat{\theta}$ tal que:

$$logL\left(\hat{\theta};\boldsymbol{y}\right) \geq L(\theta, \boldsymbol{y}), \forall\, \theta. \quad (7)$$

3.2.2. Vector Score

Una forma clásica de encontrar los estimadores máximo verosímil es derivar la función de log – verosimilitud respecto a θ. El procedimiento que calcular la primera derivada de la función log-verosimilitud es llamada la función score de Fisher y es denotada por:

$$u(\theta) = \frac{\partial}{\partial\theta} logL(\theta, \boldsymbol{y}). \quad (8)$$

Se debe notar que el vector de score es un vector de las primera derivada parcial, para cada uno de los elementos de θ[37].

[37] Dado que la transformación logarítmica es una función monótona, esta es apropiada para maximizar $L(\theta, \boldsymbol{y})$ en lugar de $L(\theta, \boldsymbol{y})$. [Para mayor detalle consultar Dobson (2002)]

Para encontrar el estimador máximo verosímil el vector score se iguala a cero, y se resuelve el sistema de ecuaciones[38]:

$$u(\hat{\theta}) = \mathbf{0}. \quad (9)$$

Siendo $\mathbf{0}$ el vector de ceros.

3.2.3. Matriz de información

Una propiedad estadística del vector aleatorio score es que el valor verdadero del parámetro θ tiene media cero.

$$E[u(\theta)] = \mathbf{0}, \qquad (10)$$

La matriz de covarianza del vector $u(\theta)$ nos da la matriz de información:

$$Var[u(\theta)] = E[u(\theta)u^t(\theta)] = \boldsymbol{I}(\theta). \quad (11)$$

Bajo ciertas condiciones de regularidad, la matriz de información puede ser obtenida como el valor negativo del valor esperado de la segunda derivada de la log-verosimilitud:

$$\boldsymbol{I}(\theta) = -E\left[\frac{\partial^2 logL(\theta)}{\partial\theta\partial\theta^t}\right]. \quad (12)$$

La matriz negativa de las segundas derivadas es llamada la matriz de información observada.

3.2.4. Newton-Raphson y Fisher Scoring

El cálculo del estimador máximo verosímil requiere de un proceso iterativo que considere expandir la función score, evaluando en el estimador máximo verosímil $\hat{\theta}$ en torno a un valor θ_0 usando una serie de Taylor de primer orden, tal que:

$$u(\hat{\theta}) \approx u(\theta_0) + \frac{\partial u(\theta)}{\partial\theta}(\hat{\theta} - \theta_0). \quad (13)$$

[38] La primera derivada de la función log-verosimilitud es necesariamente un punto crítico (máximo, mínimo o inflexión). Y si la segunda derivada es menor a cero (cóncava) o si θ es un vector del Hessiano de tal forma que éste definido no negativo, se trata de un máximo. [Para mayor detalle consultar Dobson (2002)]

Dado el Hessiano denotado por $\boldsymbol{H}$ o matriz de segundas derivadas de la función log-verosimilitud, representado por:

$$\boldsymbol{H}(\theta) = \frac{\partial^2 L}{\partial\theta\partial\theta^t} = \frac{\partial u(\theta)}{\partial\theta}. \quad (14)$$

Se considera la expresión (13) y se multiplica $\boldsymbol{H^{-1}}$ por la izquierda, obteniendo lo siguiente:

$$\mathbf{0} = \boldsymbol{H^{-1}}(\theta_0)u(\theta_0) + \left(\hat{\theta} - \theta_0\right), \quad (15)$$

Despejando se tiene:

$$\hat{\theta} = \theta_0 - \boldsymbol{H^{-1}}(\theta_0)u(\theta_0). \quad (16)$$

Este resultado proporciona la base para un enfoque iterativo para el cálculo de la estimación máxima verosimilitud conocida como la técnica de Newton-Raphson. Teniendo en cuenta un valor de prueba θ_0, usando la ecuación (16) para obtener una estimación mejorada y repetir el proceso hasta que las diferencias entre las estimaciones sucesivas son lo suficientemente próximas a cero (o hasta que los elementos del vector de primeras derivadas son lo bastante cercanos a cero).

Un procedimiento alternativo sugerido por Fisher es reemplazar $-\boldsymbol{H^{-1}}(\theta_0)$ por su valor esperado, la matriz de información $-\boldsymbol{I^{-1}}(\theta_0)$. El procedimiento resultante es una estimación mejorada, denotada por,

$$\hat{\theta} = \theta_0 + \boldsymbol{I^{-1}}(\theta_0)u(\theta_0). \quad (17)$$

Este resultado es conocido como Scoring de Fisher.

3.3. Test de Hipótesis

A continuación, se presentan algunos elementos que se necesitan para realizar pruebas de hipótesis.

3.3.1. Test de Wald

Bajo ciertas condiciones de regularidad, el estimador máximo verosimilitud $\hat{\theta}$ tiene una distribución aproximadamente p –normal con vector de media θ y matriz de covarianza dada por la matriz de información inversa $\boldsymbol{I^{-1}}(\theta)$, de modo que:

$$\widehat{\theta} \sim N_p\left(\theta, \boldsymbol{I^{-1}}(\theta)\right) \quad (18)$$

Dentro de las condiciones de regularidad, se debe considerar que el parámetro a estimar pertenezca al espacio paramétrico, la función de log-verosimilitud debe ser tres veces diferenciable y delimitada.

Este resultado proporciona una base para la construcción de pruebas de hipótesis e intervalos de confianza. Por ejemplo, consideremos la siguiente hipótesis:

$$H_0: \theta = \theta_0$$

Para un vector con valor fijo θ_0, la forma cuadrática es:

$$W = \left(\hat{\theta} - \theta_0\right)^t \boldsymbol{I^{-1}}(\theta)\left(\hat{\theta} - \theta_0\right), \quad (19)$$

Bajo H_0, es aproximadamente chi-cuadrado con p grados de libertad. Por otro lado, cuando se requiera evaluar o docimar un parámetro en particular, es decir $H_0: \theta_j = 0$, el estadístico de prueba se construye entre el cociente del valor estimado $\hat{\theta}_j$ y el elemento j –ésimo de la diagonal de la matriz de información inversa en raíz cuadrada. Para este caso el estadístico de Wald es:

$$z = \frac{\hat{\theta}_j}{\sqrt{Var(\hat{\theta}_j)}} \sim N(0,1). \quad (20)$$

Denominado estadístico z.

3.3.2. AIC

Para la selección del modelo más parsimonioso existen varios métodos, destacando entre ellos los criterios de información. Para el caso de la V EME se utilizará el criterio de Akaike (AIC)[39], el cual toma un valor igual a 2 veces la función de log-verosimilitud penalizado por el número de parámetros a estimar, dado por:

$$AIC = -2\left[\log L\left(\hat{\theta}, \boldsymbol{y}\right) + p\right]. \qquad (21)$$

Luego, se elige el modelo que tenga el menor AIC.

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

3.4.1. 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.

---

[39] Los criterios de información fueron construidos como estimadores aproximadamente insesgados de la log-verosimilitud esperada $E_{G(z)}\left(\ln f\left(Y, \hat{\theta}\right)\right)$, o, equivalentemente, de la discrepancia de la Información de Kullback – Leibler entre la verdadera distribución $g(z)$ y un modelo estadístico $f\left(Y, \hat{\theta}\right)$, desde un punto de vista predictivo. En la actualidad estos criterios de información son ampliamente utilizados para la selección de modelo estadístico, en la literatura se pueden encontrar otros criterios de información como por ejemplo: el Criterio con enfoque Bayesiano de Swarchz (BIC), denotado por, $BIC = -2\log L\left(\hat{\theta}, \boldsymbol{y}\right) + \ln(n)p$, donde penaliza el número de parámetros p con $\ln(n)$. También se puede considerar el Criterio de Hannan-Quinn $HQIC = -2\log L\left(\hat{\theta}, \boldsymbol{y}\right) + 2\ln(\ln(n))p$ como una variante del BIC con una pequeña penalización de la magnitud del tamaño muestral. La utilización del modelo AIC se utilizó para fines prácticos bajo el principio de parsimonia que establece que *todo modelo debe ser más simple que los datos en los que se basa.* [Para mayor detalle consultar Rao (2008). McCullagh(1989) y Caballero (2011) entre otros]

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

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

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

3.4.2. Valores predictivos

A pesar de que la S y la 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:

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

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

3.4.3. 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).

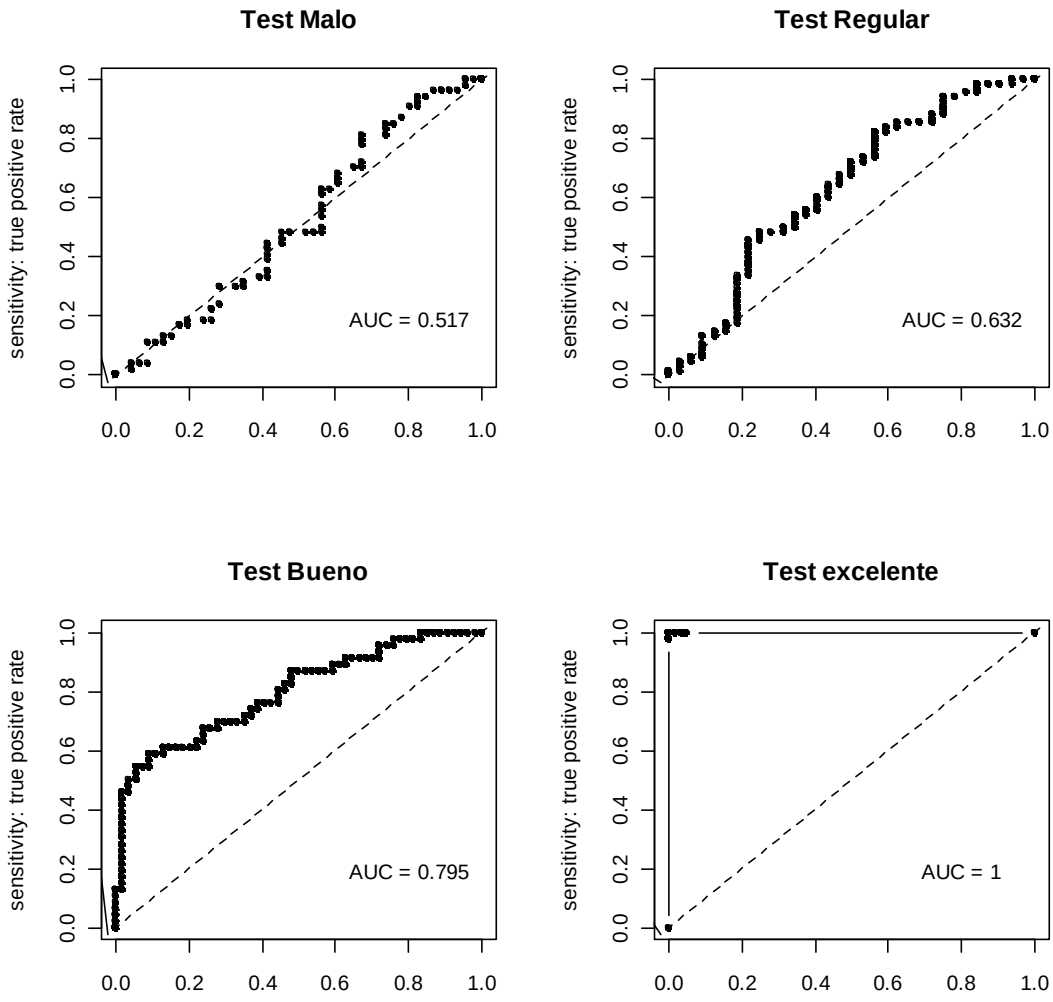
Tradicionalmente cuando se tiene un test cuantitativo, se escoge el cut-off o punto de corte más adecuado, que combine mejor la sensibilidad y especificidad del test (es decir, mayor rendimiento). Habitualmente deberían estar con sensibilidad de 85 %, con especificidad de 74 % o cercanos a estos valores.

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.

Gráfico 9. Diferentes curvas ROC



Fuente: Elaboración propia

3.5. Análisis de Elegibilidad

Para modelar la probabilidad de que una persona conteste la encuesta de la III 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.

3.5.1. Operacionalización de variables

Cuadro 28. Distribución de personas clasificadas según el código de disposición de la última visita al hogar

Código de disposición de la última visita al hogar	Frecuencia	Porcentaje	Elegible	La persona responde
110 Encuesta completa	7.492	84,94	SÍ	SÍ
211 Informante de la vivienda rechazó la entrevista	66	0,75	SÍ	NO
212 Informante directo rechazó la entrevista	172	1,95	SÍ	NO
213 Se interrumpió la entrevista al informante directo	8	0,09	SÍ	NO
223 Se impidió el acceso a la vivienda	2	0,02	SÍ	NO
224 Vivienda ocupada sin moradores presentes	325	3,68	SÍ	NO
225 Informante directo no ubicable o no puede atender	362	4,10	SÍ	NO
232 Informante con dificultad física, mental o cognitiva para contestar	5	0,06	SÍ	NO
236 Anulación por falsificación	10	0,11	SÍ	NO
290 Otra razón elegible	2	0,02	SÍ	NO
317 Área peligrosa o de difícil acceso	9	0,10	NO	-
318 No fue posible localizar la dirección	14	0,16	NO	-
390 Otra razón de elegibilidad desconocida	1	0,01	NO	-
454 Vivienda en demolición, incendiada, destruida o erradicada	5	0,06	NO	-
461 Vivienda particular desocupada (en arriendo, en venta, otro.)	23	0,26	NO	-
462 Vivienda de veraneo o de uso temporal	4	0,05	NO	-
471 Muerte del informante	4	0,05	NO	-
472 Cambio de domicilio	169	1,92	NO	-
473 Informante fuera de marco	145	1,64	NO	-
490 Otra razón no elegible	2	0,02	NO	-
Total	**8.820**	**100,00**	**8.444**	**7.492**

Fuente: Elaboración propia

En base a este cuadro se dividen las unidades elegibles (8.444) de las que no (376) y dentro de las unidades elegibles clasificamos las personas que responden (7.492) versus las que no (952).

3.5.2. Análisis Descriptivo

En esta sección se realiza un estudio descriptivo exploratorio para ver la relación de forma empírica entre algunas variables que pueden ingresar a nuestro modelo y ver su influencia en la variable de interés. Dada la característica de nuestra variable de interés (la persona que pertenece a una unidad elegible, responde sí o no), se realizan principalmente cruces con variables socio-demográficas. En este sentido se inspeccionarán las distribuciones relativas y marginales (perfil fila y columna) de las siguientes variables: Nivel Educacional, Estado Conyugal y Cantidad de Visitas Colapsado.

Para simplificar el análisis de las distribuciones marginales, se divide la muestra en las personas que responde versus las que no de manera independiente. Digamos las personas que no responden pertenecen al Grupo 1 y las personas que responden al Grupo 2.

La variable "**Nivel educacional Colapsado**" corresponde a una simplificación de la variable nivel educacional, en donde la categoría Básica, incluye aquellas personas que declararon su nivel educacional con los códigos 000, 01, 02, 03 (Nunca asistió, Sala Cuna/Jardín Infantil, Kínder/Pre-Kínder, Básica o primaria) respectivamente. La categoría Media comprende los códigos 04, 05, 06 (Media común, Media Técnico Profesional, Humanidades) respectivamente. Finalmente, en la categoría Superior se encuentran los códigos 07, 08, 09, 10, 11, 12, 14 (Centro de formación técnica, Instituto Profesional, Universidad, Postítulo, Magíster, Doctorado y Normalista) respectivamente.

Cuadro 29. Distribución de personas que responden según nivel educacional colapsado y sexo.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No		219	75	294
		398	162	560
		334	140	474
Total No		**951**	**377**	1.328
Sí		1.607	927	2.534
		1.881	1.401	3.282
		1.007	669	1.676
Total Sí		**4.495**	**2.997**	**7.492**
Total general		**5.446**	**3.374**	**8.820**

Fuente: Elaboración propia

El cuadro N°30 muestra cómo se distribuyen los casos muestrales donde por simple inspección se puede apreciar diferencias entre las personas que responden o no, respecto al nivel educacional y sexo.

Con esto se construye la distribución porcentual relativa según nivel educacional y sexo.

Cuadro 30. Distribución porcentual relativa de personas que responden según nivel educacional colapsado y sexo.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	2,5%	0,9%	3,3%
	Media	4,5%	1,8%	6,3%
	Superior	3,8%	1,6%	5,4%
Total No		10,8%	4,3%	15,1%
Sí	Básica	18,2%	10,5%	28,7%
	Media	21,3%	15,9%	37,2%
	Superior	11,4%	7,6%	19,0%
Total Sí		51,0%	34,0%	84,9%
Total general		61,7%	38,3%	100,0%

Fuente: Elaboración propia

En este caso se analiza el aporte de casos, distribuidos en las personas que **Responde**, **Nivel Educacional** y **Sexo**, respecto al total de casos. Donde por ejemplo

el 2,5% de los casos que no respondieron pertenecen al nivel educacional básica y son hombres versus 18,2% de las personas que responden en la misma categoría.

Cuadro 31. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Nivel Educacional con respecto al sexo.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	74,5%	25,5%	100,0%
	Media	71,1%	28,9%	100,0%
	Superior	70,5%	29,5%	100,0%
Total No		71,6%	28,4%	100,0%
Sí	Básica	63,4%	36,6%	100,0%
	Media	57,3%	42,7%	100,0%
	Superior	60,1%	39,9%	100,0%
Total Sí		60,0%	40,0%	100,0%

Fuente: Elaboración propia

En este caso para la distribución marginal Sexo respecto al nivel educacional, se puede decir que dentro de todas las personas que no responden, dado que poseen un nivel educacional básico, el 74,5% de los casos pertenecen al sexo Hombre y el 25,5% Mujer. Para los que pertenecen al nivel educacional medio 71,1% son hombres y 28,9% son mujeres. Finalmente del nivel educacional superior el 70,5% son hombres y el 29,5% son mujeres. De igual forma, en el caso de todas las personas que responden, se tiene que; los que pertenecen al nivel educacional básico el 63,4% son hombres y el 36,6% son mujeres. En Media el 57,3% son hombres y el 42,7% son mujeres. En el caso del nivel educacional superior el 60,1% son hombres y el 39,9% mujeres.

Cuadro 32. Análisis de perfil columna separando la distribución porcentual de personas que responden (si o no). Fijando Sexo con respecto al Nivel Educacional.

		Sexo		
Responde	**Nivel Educacional Colapsado**	**Hombre**	**Mujer**	**Total general**
No	Básica	23,0%	19,9%	22,1%
	Media	41,9%	43,0%	42,2%
	Superior	35,1%	37,1%	35,7%
Total No		100,0%	100,0%	100,0%
Sí	Básica	35,8%	30,9%	33,8%
	Media	41,8%	46,7%	43,8%
	Superior	22,4%	22,3%	22,4%
Total Sí		100,0%	100,0%	100,0%

Fuente: Elaboración propia

Dentro de todas las personas que no responden, las personas que pertenecen al sexo Hombre, el porcentaje que pertenece a educación Básica es de 23%, el que pertenece a la educación Media es 41,9% y a la educación superior es 35,1%. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 19,9% pertenece a la educación Básica, el 43% a la Media y el 37,1% al nivel Superior. Dentro de todas las personas que responden, se puede observar que dado que son hombres; el 35,8% de los casos pertenece al nivel educacional básico, el 41,8% Media y el 22,4% a nivel Superior. En este mismo sentido, en el caso de las mujeres 30,9% pertenece al nivel educacional Básica, 46,7% Media y 22,3% al nivel Superior.

En este contexto, se puede apreciar que el mayor aporte en contestar la encuesta son mujeres que pertenecen al nivel educacional medio. (46,7% versus 41,8% hombres y mujeres respectivamente).

Para la variable "**Estado Conyugal Colapsado**" se realizó una simplificación de la variable Estado Conyugal, en donde la categoría Casado(a) – Conviviente se encuentran los códigos 1 y 2 (Casado y Conviviente) respectivamente. En la categoría otros se encuentran los códigos 3, 4, 5 y 6 (Soltero(a), Viudo(a), separado(a) de hecho anulado(a) y Divorciado(a)) respectivamente. Se constata que las personas que responden la V EME tienen una relación directa con el estado "Casado(a)-conviviente".

Cuadro 33. Distribución de personas que responden según estado conyugal colapsado y sexo.

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	614	187	801
	Otros	337	190	527
Total No		951	377	1328
Sí	Casado(a) - Conviviente	3136	1624	4760
	Otros	1359	1373	2732
Total Sí		4495	2997	7492
Total general		5446	3374	8820

Fuente: Elaboración propia

En base al cuadro anterior se puede construir la frecuencia porcentual relativa de personas que responden según nivel educacional y sexo.

Cuadro 34. Distribución porcentual relativa de personas que responden según nivel educacional colapsado y sexo.

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	7,0%	2,1%	9,1%
	Otros	3,8%	2,2%	6,0%
Total No		10,8%	4,3%	15,1%
Sí	Casado(a) - Conviviente	35,6%	18,4%	54,0%
	Otros	15,4%	15,6%	31,0%
Total Sí		51,0%	34,0%	84,9%
Total general		61,7%	38,3%	100,0%

Fuente: Elaboración propia

Dentro de las personas que no responden el 7% de los hombres y el 2,1% de mujeres, pertenecen a estado conyugal Casado-conviviente representado el 9,1% de los casos. En cambio, dentro de las personas que responden, el 35,6% hombres y 18,4% son mujeres, respecto a la misma categoría.

Por otro lado, si analizamos las distribuciones marginales del estado conyugal con respecto al sexo, se pueden observar pequeñas diferencias porcentuales.

Cuadro 35. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Estado Conyugal con respecto al sexo.

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	76,7%	23,3%	100,0%
	Otros	63,9%	36,1%	100,0%
Total No		71,6%	28,4%	100,0%
Sí	Casado(a) - Conviviente	65,9%	34,1%	100,0%
	Otros	49,7%	50,3%	100,0%
Total Sí		60,0%	40,0%	100,0%

Fuente: Elaboración propia

Dentro de todas las personas que no responden, se puede decir que dado que pertenecen a la categoría Casado - Conviviente, el 76,7% de los casos pertenecen al sexo Hombre y el 23,3% Mujer. Para los que pertenecen a la categoría Otros, el 63,9% son hombres y el 36,1% son mujeres.

Por otro lado, en el caso de todas las personas que responden, se puede decir que dado que pertenecen a la categoría Casado - Conviviente, el 65,9% de los casos pertenecen al sexo Hombre y el 34,1% Mujer. Para los que pertenecen a la categoría Otros, el 49,7% son hombres y el 50,3% son mujeres.

De la misma forma, si analizamos la distribución marginal del estado conyugal fijando el sexo se tiene que:

Cuadro 36. Análisis de perfil columna separando la distribución porcentual de personas que responden (si o no). Fijando Sexo con respecto al Estado conyugal

		Sexo		
Responde	**Estado Conyugal Colapsado**	**Hombre**	**Mujer**	**Total**
No	Casado(a) - Conviviente	64,6%	49,6%	60,3%
	Otros	35,4%	50,4%	39,7%
Total No		100,0%	100,0%	100,0%
Sí	Casado(a) - Conviviente	69,8%	54,2%	63,5%
	Otros	30,2%	45,8%	36,5%
Total Sí		100,0%	100,0%	100,0%

Fuente: Elaboración propia

Dentro de todas las personas que no responden, las personas que pertenecen al sexo Hombre, el porcentaje de estos que pertenecen al estado conyugal Casado – Conviviente es de 64,6%, el 35,4% pertenece a Otros. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 49,6% pertenece a Casado - Conviviente, y 50,4% a Otros. Para las personas que responden, las personas que pertenecen al sexo Hombre, el porcentaje de estos que pertenecen al estado conyugal Casado – Conviviente es de 69,8% y el 30,2% pertenece a Otros. De igual forma dentro de los casos de personas con sexo Mujer se ve que el 54,2% pertenece a Casado - Conviviente, y 45,8% a Otros.

Finalmente, se analiza la variable **cantidad de visitas** que tiene un recorrido de 1 a 12 visitas, para esto se simplificó en tres categorías “1-3”, “4-6” y”7 y más”. Se observa que gran parte de las personas que respondieron la encuesta se encuentra dentro del tramo 1 a 3 visitas al hogar.

Cuadro 37. Distribución de personas que responden según cantidad de visitas Colapsado y sexo.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	393	166	559
	4-6	450	165	615
	7 y más	108	46	154
Total No		951	377	1328
Sí	1-3	3773	2613	6386
	4-6	638	354	992
	7 y más	84	30	114
Total Sí		4495	2997	7492
Total general		5446	3374	8820

Fuente: Elaboración propia

En base a este cuadro se pueden obtener las siguientes frecuencias relativas:

Cuadro 38. Distribución porcentual relativa de personas que responden según Cantidad de Visitas colapsado y sexo.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	4,5%	1,9%	6,3%
	4-6	5,1%	1,9%	7,0%
	7 y más	1,2%	0,5%	1,7%
Total No		10,8%	4,3%	15,1%
Sí	1-3	42,8%	29,6%	72,4%
	4-6	7,2%	4,0%	11,2%
	7 y más	1,0%	0,3%	1,3%
Total Sí		51,0%	34,0%	84,9%
Total general		61,7%	38,3%	100,0%

Fuente: Elaboración propia

Se puede ver que el 42,8% de los casos se concentra en las personas que responden entre “1-3” visitas y son hombres.

Al analizar la distribución marginal de la cantidad de visitas se tiene que:

Cuadro 39. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Sexo con respecto a la cantidad de visitas.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	70,3%	29,7%	100,0%
	4-6	73,2%	26,8%	100,0%
	7 y más	70,1%	29,9%	100,0%
Total No		71,6%	28,4%	100,0%
Sí	1-3	59,1%	40,9%	100,0%
	4-6	64,3%	35,7%	100,0%
	7 y más	73,7%	26,3%	100,0%
Total Sí		60,0%	40,0%	100,0%

Fuente: Elaboración propia

De igual forma se puede obtener la distribución marginal del sexo

Cuadro 40. Análisis de perfil fila separando la distribución porcentual de personas que responden (si o no). Fijando Cantidad de visitas con respecto al sexo.

		Sexo		
Responde	**Cantidad de Visitas Colapsado**	**Hombre**	**Mujer**	**Total**
No	1-3	41,3%	44,0%	42,1%
	4-6	47,3%	43,8%	46,3%
	7 y más	11,4%	12,2%	11,6%
Total No		100,0%	100,0%	100,0%
Sí	1-3	83,9%	87,2%	85,2%
	4-6	14,2%	11,8%	13,2%
	7 y más	1,9%	1,0%	1,5%
Total Sí		100,0%	100,0%	100,0%

Fuente: Elaboración propia

Del total de personas que responde, el 85,2% fue visitada entre 1-3 veces, mientras que sólo el 1,5% fue visitado en 7 o más oportunidades para lograr concretar la entrevista.

Sin embargo, existe una mayor probabilidad de entrevistar a las mujeres en al menos tres visitas, 87,2%.

Para la variable cantidad de visitas, se puede apreciar que el mayor aporte en contestar la encuesta son mujeres que pertenecen a la categoría entre “1-3”. (83,9% hombres versus 87,2% mujeres).

En resumen, se puede observar que existe una relación entre las personas que responden versus nivel educacional, siendo los niveles básico y medio los con mayor participación de personas, como igual a la cantidad de visitas.

3.6. 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. El cuadro 41, 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, $e^{d\cdot\widehat{\beta}_1}$.

Cuadro 41. Parámetros estimados del modelo de regresión logística seleccionado para modelar la respuesta o no de una persona que pertenece a una unidad elegible.

				95% de intervalo de confianza			95% de intervalo de confianza para Exp(B)	
Variables	**B**	**Desv. Error**	**Odd Ratios**	**Inferior**	**Superior**	**Exp(B)**	**Inferior**	**Superior**
(Intersección)	-1,046	0,409		-1,849	-0,243	0,351	0,157	0,784
[Area=1]	0,902	0,324	2,466	0,267	1,538	2,466	1,306	4,656
[Area=2]								
[Macrozona=1]	0,460	0,342	1,584	-0,211	1,131	1,584	0,809	3,100
[Macrozona=2]	0,656	0,348	1,928	-0,026	1,339	1,928	0,974	3,814
[Macrozona=3]	0,914	0,342	2,493	0,244	1,584	2,493	1,276	4,872
[Macrozona=4]								
[sexo=1]	-0,551	0,098	0,576	-0,743	-0,360	0,576	0,476	0,698
[sexo=2]								
[proveedor=0]	-0,100	0,093	1,105	-0,281	0,082	0,905	0,755	1,086
[proveedor=1]								
[Nivel_colap=1]	0,337	0,130	1,401	0,082	0,592	1,401	1,086	1,808
[Nivel_colap=2]	0,223	0,107	1,250	0,014	0,433	1,250	1,014	1,542
[Nivel_colap=3]								
[CIUO88_1=1]	-0,737	0,225	0,479	-1,177	-0,296	0,479	0,308	0,744
[CIUO88_1=2]	-0,583	0,199	0,558	-0,974	-0,192	0,558	0,378	0,825
[CIUO88_1=3]	0,030	0,187	1,030	-0,337	0,396	1,030	0,714	1,487
[CIUO88_1=4]	0,053	0,431	1,054	-0,793	0,898	1,054	0,452	2,456
[CIUO88_1=5]	0,084	0,150	1,088	-0,209	0,377	1,088	0,811	1,458
[CIUO88_1=6]	0,133	0,182	1,143	-0,224	0,491	1,143	0,799	1,634
[CIUO88_1=7]	0,027	0,144	1,027	-0,255	0,308	1,027	0,775	1,361
[CIUO88_1=8]	0,056	0,179	1,058	-0,295	0,407	1,058	0,745	1,502
[CIUO88_1=9]								
[visitas_colap=1]	3,288	0,153	26,791	2,988	3,588	26,791	19,854	36,151
[visitas_colap=2]	0,927	0,146	2,528	0,641	1,214	2,528	1,897	3,368
[visitas_colap=3]								
edad	-0,001	0,003	0,999	-0,007	0,005	0,999	0,993	1,005
[Area=1] * [Macrozona=1]	0,062	0,368		-0,659	0,782	1,064	0,517	2,186
[Area=1] * [Macrozona=2]	-0,330	0,363		-1,042	0,383	0,719	0,353	1,466
[Area=1] * [Macrozona=3]	-0,107	0,375		-0,842	0,628	0,899	0,431	1,874
[Area=1] * [Macrozona=4]								
[Area=2] * [Macrozona=1]								
[Area=2] * [Macrozona=2]								
[Area=2] * [Macrozona=3]								
[Area=2] * [Macrozona=4]								

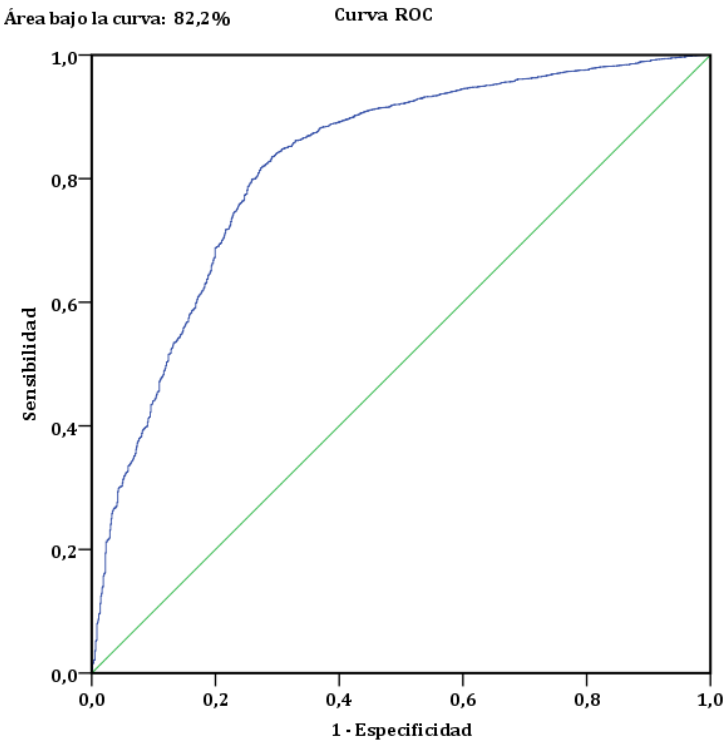
Fuente: Elaboración propia

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.

3.6.1. Análisis de Resultados

En base al modelo estimado se puede observar que el área bajo la curva (AUC) es de 0,822, lo cual está dentro **de una categoría de "Test bueno".** O bien, se puede decir que el modelo tiene una capacidad de predicción del 82,2% de los casos.

Gráfico 10 Probabilidad estimada de responder para cada una de las personas que pertenecen a la unidad elegible



Fuente: Elaboración propia

Cuadro 42. Cuadro de clasificación de los individuos de acuerdo al modelo logístico.

Clasificación

Observado	Pronosticado		
	0 NO	**1 SÍ**	**Porcentaje correcto**
0 NO	156	796	16,4%
1 SÍ	137	7.355	98,2%
Porcentaje global	**3,5%**	**96,5%**	**89,0%**

Variable dependiente: Responde (categoría de referencia = 0 NO)
Modelo: (Intersección), Area, Macrozona, sexo, proveedor, Nivel_colap, CIUO88_1, visitas_colap, edad, Area * Macrozona

Finalmente, la sensibilidad y especificidad calculada es:

$$Sensibilidad = \frac{7.355}{(7.355+796)} = 0{,}902 \qquad Especificidad = \frac{156}{(156+137)} = 0{,}532$$

Anexo N°4. Estimación de varianzas

3.6.2. Determinación del diseño muestral en Spss

***Plan de muestreo**

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

***Estimación de frecuencias en Spss**

```
CSTABULATE
  /PLAN FILE='planEME.csaplan'
  /TABLES VARIABLES=CAENES_1_red CISE_EME
  /CELLS TABLEPCT
  /STATISTICS SE CV CIN(95) DEFF
  /MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.
```