METODOLÓGICO

Documentos de trabajo

Estratificación socioeconómica del marco muestral de viviendas 2017 (MMV 2017)

Autores:
Miguel Guerrero
Julio Guerrero
Andrea Marín
Matías Pizarro

Nº.11, Octubre de 2019

Instituto Nacional de Estadísticas · Chile

INSTITUTO NACIONAL DE ESTADÍSTICAS

Morandé 801, piso 22, Santiago, Chile

Teléfono: 562 3246 1010

Correo: ine@ine.cl

Facebook: @ChileINE

Twitter: @INE_Chile

Miguel Guerrero

Julio Guerrero

Andrea Marín

Matías Pizarro

Departamento de Metodologías e Innovación Estadística

Los autores agradecen la colaboración y comentarios de los distintos involucrados en cada una de las etapas realizadas para la estratificación del marco muestral, particularmente a Beatriz Salinas, Pablo Sánchez, María Mercedes Jeria y Denisse López quienes participaron activamente en discusiones técnicas y estratégicas para dotar al marco de información útil para los diseños muestrales en encuestas de hogares. También, agradecemos la valiosa contribución en los análisis preliminares y de preprocesamiento de datos por parte de Klaus Lehmann, Vania Martínez, Diego Rodríguez, Iván Touron, Patricia Mauna, Alfredo Heufemann y Victoria Herrera. Finalmente, como equipo de trabajo agradecemos de manera especial la colaboración y guía de la Unidad de Estadísticas Sociales, por medio del experto regional en estadísticas sociales el sr Andrés Gutiérrez, así como al sr Álvaro Fuentes, estadístico de la unidad.

Los Documentos de Trabajo del Instituto Nacional de Estadísticas están dirigidos a investigadores, académicos, estudiantes y público especializado en materias económicas, y tienen como objetivo proporcionar un análisis exhaustivo sobre aspectos conceptuales, analíticos y metodológicos claves de los productos estadísticos que elabora la institución y, de esta forma, contribuir al intercambio de ideas entre los distintos componentes del Sistema Estadístico Nacional.

Las interpretaciones y opiniones que se expresan en los Documentos de Trabajo pertenecen en forma exclusiva a los autores y colaboradores y no reflejan necesariamente el punto de vista oficial del INE ni de la institución a la que pertenecen los colaboradores de los documentos.

El uso de un lenguaje que no discrimine ni marque diferencias entre hombres y mujeres ha sido una preocupación en la elaboración de este documento. Sin embargo, y con el fin de evitar la sobrecarga gráfica que supondría utilizar en castellano “o/a” para marcar la existencia de ambos sexos, se ha optado por utilizar -en la mayor parte de los casos- el masculino genérico, en el entendido de que todas las menciones en tal género representan siempre a hombres y mujeres, abarcando claramente ambos sexos.

Estratificación socioeconómica del marco muestral de viviendas 2017 (MMV 2017).

Resumen

La estratificación socioeconómica del marco muestral o la clasificación de las unidades primarias de muestreo según el nivel socioeconómico de las personas, hogares y viviendas que la componen, permite aumentar la eficiencia de la inferencia en las encuestas de hogares que utilizan este marco.

En este documento se presenta una breve descripción de la metodología usada para la estratificación socioeconómica del marco muestral de viviendas derivado del Censo de Población y Vivienda del 2017. Basándonos en la experiencia nacional y regional, así como en la literatura estadística, la metodología escogida abarca distintos métodos estadísticos (principalmente algoritmos de k-medias, PRINCALS y análisis en componentes principales) y consideraciones sobre la información analizada. Posteriormente, para elegir cuál de las estratificaciones obtenidas era la más apropiada, se implementó un método de evaluación, basado en que el objetivo de la estratificación es contribuir en el desarrollo de estrategias muestrales que induzcan una mayor eficiencia y precisión a la hora de la estimación de las estadísticas oficiales, y adicionalmente teniendo en cuenta el carácter multipropósito de la estratificación del marco, en el sentido que será usada por diferentes encuestas sociales. Bajo este criterio y algunas consideraciones adicionales, el método escogido es el que utiliza el análisis en componentes principales (primera componente) y el algoritmo de estratificación óptima para tres grupos.

Abstract

The socioeconomic stratification of the sampling frame or the classification of the primary sampling units according to the socioeconomic level of the people, households and dwellings that compose it, allows to increase the efficiency of inference in household surveys that use this frame.

This document presents a brief description of the methodology used for the socioeconomic stratification of the sampling frame derived from the 2017 Census. Based on national and regional experience, as well as statistical literature, the chosen methodology covers different

statistical methods (mainly k-means algorithms, PRINCALS, Principal Component Analysis) and considerations on the analyzed information.

Subsequently, to choose which of the stratifications obtained was the most appropriate, an evaluation method was implemented, based on the objective of the stratification is to contribute to the development of sample strategies that induce greater efficiency and precision at the time of estimation of official statistics, and additionally taking into account the multipurpose nature of the stratification of the framework, in the sense that it will be used by different social surveys.

Under this criterion and some additional considerations, the method chosen is the one that uses Principal Component Analysis (first component) and the optimal stratification algorithm for three groups.

Palabras clave: Estratificación, Marco Muestral, Censo 2017, k-medias, Análisis de Componentes Principales.

Índice

1 Antecedentes

En abril del año 2017 se realizó el XIX Censo Nacional de Población y VIII de Vivienda o Censo de Población y Vivienda 2017, el cual utilizó la metodología de un censo tradicional de hecho y su carácter fue abreviado. Por tanto, se delimitaron sus objetivos y, en consecuencia, se redujeron los contenidos del cuestionario censal (21 preguntas). El censo de población y vivienda constituye el principal insumo para la elaboración del marco muestral[1] de viviendas y, por tanto, el principal insumo para la elaboración de las encuestas a hogares.

El marco muestral de viviendas derivado de este censo (MMV 2017), es uno bietápico que en la primera etapa, es un marco de áreas delimitadas y cartografiadas (basadas en los sectores de empadronamiento censal) que cubren todo el país, denominadas Unidades Primarias de Muestreo (UPMs). En la segunda etapa, corresponde a un marco de lista, donde para cada UPM se listan todas las viviendas que ésta contiene. Las UPMs se caracterizan por ser unidades homogéneas en términos de número de viviendas[2], tanto en el área urbana como rural. El rango aproximado para del tamaño de las UPMs en el área urbana es de [160.240] viviendas y en el área rural es de [70.110] viviendas.

La estratificación socioeconómica del MMV o la clasificación de las UPMs de acuerdo al nivel socioeconómico permite aumentar la eficiencia de la inferencia en las encuestas de hogares que utilizan este marco. Tanto a nivel nacional como regional se pueden encontrar diversas experiencias para la clasificación de las UPMs del marco. En el Instituto Nacional de Estadísticas (INE) de Chile, por ejemplo, existe una experiencia previa de estratificación socioeconómica, implementada en el marco muestral de manzanas[3], utilizado el análisis de componentes principales no lineal. El INE de Bolivia utiliza una combinación entre el método de componentes principales y el algoritmo de k-medias. El Instituto Brasileño de Geografía y Estadística (IBGE) en Brasil estratifica su marco a partir de los algoritmos de optimización que son variantes del algoritmo k-medias desarrollados por Montenegro y Brito (Montenegro & Brito, 2006). El Instituto Nacional de Estadísticas y Censos (INEC) en Ecuador utiliza para la estratificación el algoritmo de k-medias.

[1] Según la Organización para la Cooperación y el Desarrollo Económicos (OCDE), un Marco Muestral es "una lista de todos los miembros de una población usada como base para el muestreo" (OCDE, 2007, pág. 694).
[2] Se refiere a viviendas particulares excluyendo de temporada.
[3] El marco muestral de viviendas vigente está basado en el Censo de Población del 2002. En el año 2006 en el área urbana, se redefinió la unidad primaria de muestreo, pasando de secciones a manzanas.

Basándonos en la experiencia nacional y regional, así como en la literatura estadística, la metodología escogida para la estratificación socioeconómica del MMV 2017, abarca distintos métodos estadísticos y consideraciones sobre la información analizada. El método de evaluación se basa por un lado, en que el objetivo de la estratificación es contribuir en el desarrollo de estrategias muestrales que induzcan una mayor eficiencia y precisión a la hora de la estimación de las estadísticas oficiales, y por otro lado en el carácter multipropósito de la estratificación del marco, en el sentido que será usada por diferentes encuestas sociales que buscan indagar sobre fenómenos que se relacionan con la situación socioeconómica de los hogares y las personas que los componen.

2 Objetivo

Generar una estratificación socioeconómica de las UPMs del MMV 2017 que permita aumentar la eficiencia de los diseños muestrales.

3 Metodología de estratificación

3.1 Esquema general

A modo general, la metodología para la estratificación del MMV 2017 se basó en el uso de métodos multivariados y algoritmos de clasificación, utilizando como principal insumo la información de la base del XIX Censo Nacional de Población y VIII de Vivienda levantado en abril 2017[4].

Si bien el objetivo era estratificar las UPMs, la unidad de análisis podía ser la persona, el hogar, la vivienda o la UPM, en cualquier caso, se requería agregar la información, ya fuese antes de aplicar los métodos multivariados (UPM como unidad de análisis) o después (persona, hogar o vivienda como unidad de análisis). La figura I. muestra que en la primera fase se decidió trabajar con dos unidades de análisis diferentes: UPM y persona. En cada caso se seleccionaron las variables del Censo que ingresarían en los distintos métodos de estratificación y posteriormente se realizaron algunos tratamientos pertinentes, como por ejemplo, construcción de nuevos indicadores a partir de las variables originales, eliminación de la información de personas que no son residentes habituales, así como quienes tenían datos faltantes para las variables del análisis.

En la segunda fase, contando por un lado con la información a nivel de UPM se seleccionaron distintos sets de variables y se realizó en cada caso un análisis de componentes principales (ACP), posteriormente se aplicaron diferentes algoritmos de estratificación univariantes (considerando 3 y 4 estratos) sobre los resultados de la primera componente principal. Con base en el conjunto de indicadores a nivel de UPM, también se aplicaron variantes del algoritmo k-medias, haciendo particiones en 3, 4 y 5 estratos. Por otra parte, a partir de la información a nivel de personas se seleccionaron distintos sets de variables y se aplicó el algoritmo PRINCALS (Principal Components Analysis by means of Alternating Least Squares), luego el puntaje obtenido sobre la primera componente principal fue agregado a nivel de UPM y finalmente se realizaron particiones a través de terciles, cuartiles y quintiles.

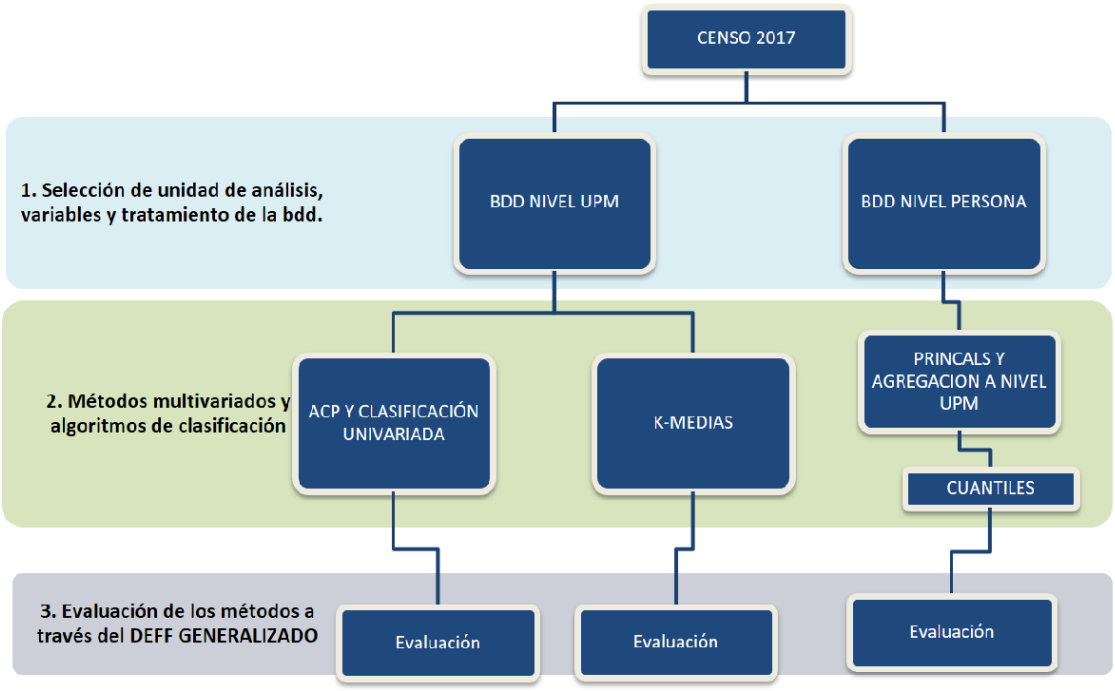
Una vez obtenidos las diferentes estratificaciones basadas en las metodologías mencionadas, se debió elegir cuál de ellas era la más conveniente para el MMV 2017 (Figura I. Fase 3). En

[4] En adelante, nos referiremos a esta base como Censo o base censal.

este punto, es importante destacar que el objetivo de la estratificación de un marco de muestreo es aumentar la eficiencia de los diseños muestrales, reduciendo los errores de estimación, al proporcionar información que permite realizar selecciones independientes en cada grupo o estrato. Con la estratificación, se espera que las UPMs dentro de un mismo estrato sean similares (disminución de la intra-varianza) y diferentes a las de otro estrato (aumento de la inter-varianza).

A partir de esto, se concluyó que la mejor estratificación para el MMV era la que asegurara obtener varianzas mínimas para un conjunto de indicadores de interés calculados a partir de la base censal. La medida que se utilizó para la evaluación es el *efecto de diseño generalizado*, el cual explicaremos en detalle más adelante.

Figura I.Esquema general de la metodología de estratificación del MMV



Fuente: Instituto Nacional de Estadísticas (INE).

3.2 Selección de variables y tratamiento de la información faltante

Base a nivel de UPM

Selección de variables

La unidad de análisis coincide con la unidad de estratificación. Los indicadores se construyeron bajo la premisa que la mayoría de los fenómenos estudiados por las encuestas sociales a hogares guardan relación con el grado de bienestar de la población.

Tabla V.1.Indicadores construidos a nivel de UPM a partir de la base censal

Nivel Base Censal	Indicador	Descripción
Viviendas	Materialidad de la vivienda	Hormigón armado o albañilería (para viviendas del área urbana). Hormigón armado, albañilería o tabique forrado por ambas caras (para viviendas del área rural).
		Casa o departamento -muros exteriores donde el material predominante es hormigón armado o Albañilería (para viviendas del área urbana). Casa, departamento o vivienda tradicional - muros exteriores donde material predominante es hormigón armado, albañilería o tabique forrado por ambas caras (para viviendas del área rural)
		Índice de materialidad alto
	Acceso al agua	Acceso a agua a través de la red pública
	Sin hacinamiento	Razón de personas por habitación con uso exclusivo para dormitorio < 2.5
Personas	Escolaridad	Población >=25 años con nivel de escolaridad alto[5]
	Tasa de ocupados	Ocupados/Población en edad de trabajar
	Tasa de no desocupación	1 - Desocupados/Población económicamente activa
	Tasa de participación	Población económicamente activa / Población en edad de trabajar
	Razón de dependencia	(Población < 15 años + Población > 64 años)/ Población entre 15 y 64 años.
	Capacidad de generar ingresos	(Población <15 años + Población > 64 años)/Población Ocupada
	Indicador sobre total de hijos nacidos	Población de mujeres con un total de 2 o menos hijos nacidos vivos/Número de mujeres >=15 años

Fuente: Instituto Nacional de Estadísticas (INE).

Dado que la información en la base censal estaba a nivel de viviendas, hogares y personas, en primera instancia se eligió qué indicadores construir a partir de las variables originales, teniendo en cuenta además que las UPMs no tienen estrictamente la misma cantidad de viviendas ni personas, por lo que se calcularon todos los indicadores en términos relativos (porcentaje). De esta forma, los indicadores representan el porcentaje de personas, hogares o viviendas dentro de la UPM que tienen determinada característica que indica bienestar. En la tabla V.1 se muestran los indicadores construidos, en el caso del nivel persona, los indicadores también fueron calculados por sexo. Si bien algunos indicadores están altamente correlacionados como, por ejemplo, la tasa de ocupación y la tasa de participación,

[5] Se considera nivel de escolaridad alto estudios medios o superiores en el caso de las personas de 65 años y más, y nivel superior para las personas entre 25 y 64 años.

el objetivo era determinar cuáles de ellos permitían obtener los mejores resultados, en términos de la varianza explicada (ACP o k-medias) como en el método de evaluación que se explicará más adelante.

En cuanto al tratamiento de la información faltante para las variables de interés, se decidió utilizar información completa. De esta forma se eliminaron aquellos registros que tenían al menos un dato faltante en las variables insumo para el cálculo de los indicadores presentados en la tabla V.1. Las tablas V.2 y V.3 muestran la distribución de la información omitida según área. Adicionalmente, se calcularon los porcentajes de omisión a nivel regional, comunal y de UPM con el objetivo de cerciorarnos que la información omitida no se concentrara en ninguna de estas áreas.

Tabla V.2.Porcentaje de viviendas excluidas según área

Información	Área	
	Urbano	Rural
Completa	94,9	95,1
Missing	5,1	4,9
Total	**100**	**100**

Fuente: Instituto Nacional de Estadísticas (INE).

Tabla V.3.Porcentaje de personas excluidas según área

Información	Área	
	Urbano	Rural
Completa	91,4	91,2
Missing	8,6	8,8
Total	**100**	**100**

Fuente: Instituto Nacional de Estadísticas (INE).

Base a nivel persona

Selección de variables

Para la selección de las variables se escogieron aquellas que, de acuerdo a varios estudios[6] nacionales e internacionales, constituían factores correlacionados con el constructo *bienestar socioeconómico,* entre las que destacan el nivel de educación, condición de actividad económica, rama de actividad económica, características de la vivienda (material

[6] Entre los estudios que vinculan el bienestar socioeconómico con variables como ingresos, nivel de educación, ocupación, se tiene: - "La distribución del ingreso en Chile 1987-2006: Análisis y consideraciones de política", (Solimano, Andrés; Torche, 2008). – "Enfoques para la medición de la pobreza", (Féres Juan Carlos, 2001). – "Estratificación Socioeconómica en encuestas de hogares", (Instituto Nacional de Estadísticas de Chile (INE), 2012). – "Distribución del ingreso en Chile. Nueve hechos y algunos mitos", (Contreras Dante, 1998). – "Mapa socioeconómico de Chile", (Adimark, 2002).

de construcción predominante en los muros exteriores, en el techo y piso). Dado el contexto de un censo abreviado, se excluyeron variables como la ocupación u oficio, tanto del jefe de hogar como de los demás integrantes y además se debe tener presente que el Censo no indaga por el nivel de ingresos de las personas ni del hogar, variable altamente correlacionada con el bienestar socioeconómico.

A nivel de persona no hay muchas elecciones de variables del censo abreviado que pudieran explicar el bienestar socioeconómico de las unidades de las personas (ver tabla V.4).

Tabla V.4.Variables consideradas para el método PRINCALS

	Variables consideradas inicialmente				Variables consideradas finalmente		
X	Viv_Techo	:	Material en la cubierta del techo	X	Viv_Techo	:	Material en la cubierta del techo
	Viv_Tipo	:	Tipo de vivienda		Viv_Tipo	:	Tipo de vivienda
X	Viv_Muros	:	Material de los muros exteriores	X	Viv_Muros	:	Material de los muros exteriores
X	RAMA	:	Rama de actividad económica		RAMA	:	Rama de actividad económica
X	Nivel_Educ	:	Nivel educacional	X	Nivel_Educ	:	Nivel educacional
	Viv_Dormitorios	:	Número de dormitorios		Viv_Dormitorios	:	Número de dormitorios
	Viv_Piso	:	Material de construcción del piso		Viv_Piso	:	Material de construcción del piso
X	Viv_Agua	:	Origen del Agua		Viv_Agua	:	Origen del Agua
	Condición_Actividad	:	Situación en el mercado laboral		Condición_Actividad	:	Situación en el mercado laboral
X	Area	:	Urbano - Rural		Area	:	Urbano - Rural
	Condición_Actividad	:	Situación en el mercado laboral		Condición_Actividad	:	Situación en el mercado laboral
	Comuna	:	Comuna de división político-adm.	X	Comuna	:	Comuna de división político-adm.

Fuente: Instituto Nacional de Estadísticas (INE).

Finalmente, marcadas con 'X' en la tabla V.4. están las variables que se utilizaron para generar la estratificación en quintiles, cuartiles y terciles, que se listan a continuación:

- Material de construcción predominante en muro de la vivienda
- Material de construcción predominante en techo de la vivienda
- Nivel educacional
- Comuna

Manejo de información faltante y "No aplica"

Aparte de los valores missing que se encontraron, existen muchos valores "No aplica" que bien pueden ser considerados como otra categoría. Con el objetivo de potenciar la aplicación del algoritmo PRINCALS, aumentando la correlación entre las variables, se procedió a identificarlos y asignarles las categorías correspondientes a otras variables.

Tabla V.5.Variable Rama y Condición de actividad económica homologadas

	RAMA**	Frecuencia		Condición_Actividad	Frecuencia
Ocupado	1 A : Agricultura, ganadería, silvicultura y pesca	451.196			
	2 B : Explotación de minas y canteras	105.434			
	:	:			
	:	:		1 Ocupado	7.442.563
	:	:			
	21 U : Actividades de organizaciones y órganos ..	2.061			
	22 Z : Rama no declarada	1.082.199			
No Aplica	23 Desocupado	566.158		2 Desocupado	566.158
	24 Estudiante	1.392.727		3 Estudiante	1.392.727
	25 Quehaceres del hogar	1.656.542		4 Quehaceres hogar	1.656.542
	26 Jubilados, pensionados	1.396.247		5 Jubilados, pensionados	1.396.247
	27 Otra situación de Inactivo	492.357		6 Otra situación	492.357
	28 Menor de edad	3.396.290		98 No aplica	3.396.290
Perdidos		288.436		99 Missing	288.436
Total		**16.631.320**		**Total**	**16.631.320**

Fuente: Instituto Nacional de Estadísticas (INE).

En cuanto al tratamiento de los valores perdidos, se deja al propio algoritmo PRINCALS que los excluya del análisis cuando corresponda, pero no se imputan. El porcentaje de valores perdidos por área urbana y rural, para cada variable utilizada en el modelo se presenta en la siguiente tabla V.6.

Tabla V.6.Porcentaje de valores perdidos por área y variable utilizada en el algoritmo PRINCALS

Nivel educacional	Urbano	Rural	Total	Frec.
Información completa	98,0	97,6	98,0	16.295.859
Missing	2,0	2,4	2,0	335.461
Total	**100,0**	**100,0**	**100,0**	**16.631.320**

Rama	Urbano	Rural	Total	Frec.
Válidos	38,9	33,4	38,2	6.360.364
Rama no declarada	6,5	6,2	6,5	1.082.199
No Aplica	52,9	58,4	53,5	8.900.321
Missing	1,7	2,0	1,7	288.436
Total	**100,0**	**100,0**	**100,0**	**16.631.320**

Condición de Actividad	AREA			
	Urbano	Rural	Total	Frec.
Información completa	98,3	98,0	98,3	16.342.884
Missing	1,7	2,0	1,7	288.436
Total	**100,0**	**100,0**	**100,0**	**16.631.320**

Material de los muros exteriores	AREA			
	Urbano	Rural	Total	Frec.
Información completa	99,6	99,7	99,6	16.567.196
Missing	0,4	0,3	0,4	64.124
Total	**100,0**	**100,0**	**100,0**	**16.631.320**

Material en la cubierta del techo	AREA			
	Urbano	Rural	Total	Frec.
Información completa	99,3	99,5		16.524.253
Missing	0,7	0,5		107.067
Total	**100,0**	**100,0**		**16.631.320**

Material de construcción del piso	AREA			
	Urbano	Rural	Total	Frec.
Información completa	99,2	99,4	99,2	16.501.040
Missing	0,8	0,6	0,8	130.280
Total	**100**	**100**	**100**	**16.631.320**

Dormitorios exclusivos para dormir	AREA			
	Urbano	Rural	Total	Frec.
Información completa	96,8	97,1	96,8	16.099.062
Missing	3,2	2,9	3,2	532.258
Total	**100,0**	**100,0**	**100,0**	**16.631.320**

Origen del Agua	AREA			
	Urbano	Rural	Total	Frec.
Información completa	99,3	99,4	99,3	16.523.137
Missing	0,7	0,6	0,7	108.183
Total	**100,0**	**100,0**	**100,0**	**16.631.320**

Fuente: Instituto Nacional de Estadísticas (INE).

Se aprecia que el porcentaje de valores perdidos no supera el 3% y se distribuye en forma homóloga en el área urbana y en el área rural, por lo que la pérdida de información no es un tema relevante que implique un menoscabo en la calidad del análisis.

3.3 Métodos de estratificación

3.3.1 Algoritmos de k-medias

Descripción General:

Dado un conjunto de observaciones de la matriz de información $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ a nivel de UPM, donde cada observación es un vector real de d dimensiones, k-medias construye una partición de las observaciones en k conjuntos ($k \leq n$) a fin de minimizar la suma de los cuadrados dentro de cada grupo $S = \{S_1, \ldots, S_i\}$

$$f = arg\min_{S} \sum_{i=1}^{I} \sum_{x_j \in S_i} \|x_j - \bar{x}_i\|^2 \qquad \textbf{(1)}$$

En donde $\bar{x}_i$ corresponde al centroide de puntos en S_i.

El algoritmo más común de k-medias usa una técnica de refinamiento iterativo. Dado un conjunto inicial de k centroides $m_1^{(1)}, \ldots, m_k^{(1)}$ (ver más abajo), el algoritmo continúa alternando entre dos pasos (MacKay, 2003):

Paso de asignación: asigna cada observación al grupo con la media más cercana (es decir, la partición de las observaciones de acuerdo con el diagrama de Voronoi (Aurenhammer, 1991) generado por los centroides

$$S_i^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\| \leq \left\|x_p - m_j^{(t)}\right\| \ \forall\, 1 \leq j \leq k\right\}$$

Donde cada x_p va exactamente dentro de un $S_i^{(t)}$.

Paso de actualización: calcular los nuevos centroides como el centroide de las observaciones en el grupo

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

El algoritmo se considera que ha convergido cuando las asignaciones ya no cambian.

En este trabajo se ha considerado $k = 3{,}4\ y\ 5$ y la distancia euclideana como métrica de distancia.

El algoritmo está orientado a proporcionar soluciones de calidad de manera bastante rápida, a través de una búsqueda en un subconjunto muy restringido de todas las posibles soluciones.

Variantes del algoritmo de las k-medias:

Hay un buen número de versiones del algoritmo de las k-medias. La diferencia entre estos algoritmos está asociada básicamente a la forma de definir las agrupaciones iniciales, la regla de movimiento en los grupos (intercambio de objetos) y la regla de actualización (definición de los centroides). Los criterios de parada de estos algoritmos normalmente se asocian a un tiempo máximo de la ejecución y la verificación de la diferencia entre soluciones obtenidas en dos iteraciones consecutivas del algoritmo. Para este trabajo se han considerado cinco variaciones del algoritmo: Forgy (Forgy, 1965), MacQueen (MacQueen, 1967), Lloyd (Lloyd, 1982), Hartigan-Wong (Hartigan & Wong, 1979) y Jarque (Jarque, 1981).

Previo a la implementación de los algoritmos, un análisis exploratorio, consistente en un análisis cluster de variables y matriz de correlaciones, fue realizado. En ambos casos, se observaron fuertes relaciones entre las variables y, por lo tanto, se debió reducir la dimensionalidad de $\boldsymbol{X}$. El índice Rand ajustado (Rand, 1971) sugirió la elección de cinco clusters de variables.

En base a estos resultados, dos estrategias se emplearon para abordar la reducción de dimensionalidad: selección de una variable por cada cluster (cinco variables para análisis) y aplicación de análisis de componentes principales. En el primer caso, la selección de mejor variable por cluster dio como resultado realizar la estratificación de las UPMs a través de: proporción de personas con estudios superiores, tasa de ocupados, tasa de desocupados, bienestar de vivienda-muro y acceso a agua. Por su parte, el análisis de componentes principales fue realizado para cada cluster de variables, donde la primera componente

principal en cada caso fue seleccionada como variable de entrada para realizar la estratificación de las UPMs.

En total, bajo el enfoque de k-medias se realizaron 45 estratificaciones (30 bajo el enfoque de selección de variables y 15 bajo el enfoque de componentes principales). Los mejores resultados se obtuvieron bajo el enfoque de selección de variables a través de los algoritmos de Hartigan-Wong y MacQueen, cuyas bondades de ajustes fueron 88,3%, 92,6% y 94,7% considerando 3, 4 y 5 estratos, respectivamente.

3.3.2 Análisis de componentes principales y algoritmos de estratificación univariada

Descripción general:

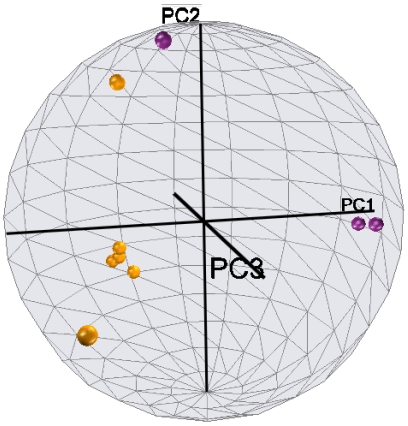
El análisis de componentes principales (ACP) es una metodología inventada por Karl Pearson como parte del análisis de factores (Pearson, 1901), sin embargo, no fue sino hasta 1939 cuando Hotelling hizo una presentación más formal e incorporó el término de componente principal (Hotelling, 1933). De forma muy general podemos definir el ACP como un tipo de transformación lineal aplicada a un conjunto de datos multivariantes habitualmente correlacionados entre sí, que permite expresar la información contenida en un menor número de variables o componentes principales no correlacionadas entre sí (ver Gráfico I.1). El ACP se caracteriza por extraer la información más importante de un conjunto de datos multivariantes, manteniendo solo la información que se considere importante (reducir la dimensionalidad de los datos), para simplificar la descripción y análisis de la estructura de las observaciones y de las variables (Abdi & Williams, 2010). Como medida de la cantidad de información incorporada en una componente se utiliza su varianza. Es decir, cuanto mayor sea su varianza, mayor es la cantidad de información que lleva incorporada dicha componente. Por esta razón, se selecciona como primera componente aquella que tenga mayor varianza, mientras que la última componente es la de menor varianza. En general, la extracción de componentes principales se efectúa sobre variables tipificadas para evitar problemas derivados de la escala, aunque también se puede aplicar sobre variables expresadas en desviaciones respecto a la media. El nuevo conjunto de variables que se obtiene por el método de componentes principales es igual en número al de las variables

originales. Es importante destacar que la suma de sus varianzas es igual a la suma de las varianzas de las variables originales.

En términos notacionales podemos resumir el ACP como sigue:

Sea $\boldsymbol{X} = [X_1,\ldots, X_p\,]$ una matriz de datos multivariantes, en donde las filas representan los individuos y las columnas p **variables continuas** analizadas. Las componentes principales son variables compuestas $Y_1 = Xt_1,\ Y_2 = Xt_2,\ \ldots,\ Y_p = Xt_p,$ tales que $Var(Y_1) = máx(Var(Y_1), \ldots, Var(Y_p))$ condicionado a $t_1' t_1 = 1$ y donde $\boldsymbol{T} = [t_1, \ldots, t_p]$ es la matriz pxp cuyas columnas son los vectores que definen las componentes principales. De esta forma la transformación lineal $X \rightarrow T$ se denomina la transformación por componentes principales y se tiene $Y = XT$.

Gráfico I.1.Análisis en componentes principales (tres dimensiones de análisis)



Fuente: http://www.genecodes.com/codelinker/visualizations/clustering/pca

A partir de los indicadores a nivel de UPM presentados anteriormente se realizaron ocho ACP, producto de las diferentes combinaciones del set de indicadores (básicamente se inició con un gran conjunto de indicadores y se fueron eliminando del análisis teniendo en cuenta sus correlaciones y las bajas contribuciones a la primera componente principal). En la mayoría de los casos el porcentaje de varianza retenida era superior al 50% y todas las variables estaban correlacionadas positivamente con la primera componente, lo que la constituía en un factor tamaño.

Clasificación a partir de la primera componente principal

Para el proceso de clasificación o en este caso de estratificación, el ACP previo se constituye en un pretratamiento, que transforma los datos originales en variables continuas no correlacionadas (Pardo & Del Campo Neira, 2007). La primera componente, al ser una función lineal de las variables del análisis y la que recoge más varianza, es la principal medida de resumen y sobre ella se pueden aplicar diversos algoritmos de clasificación. Otra alternativa es no restringir la estratificación a la primera componente, sino aplicar algoritmos de clasificación multivariada sobre varias componentes reteniendo así un mayor porcentaje de varianza.

Para este caso se decidió utilizar solamente la primera componente, la cual es interpretada a partir de los resultados como una medida de bienestar de la UPM. Así, se aplicaron diferentes tipos de particiones considerando tres y cuatro grupos o estratos, para el conjunto total de las UPMs del MMV, y adicionalmente se probó generando las particiones para cada una de las regiones**. Por tanto, bajo esta metodología se obtuvieron 128 vectores de estratificación del marco.**

A continuación, se resumen las diferentes estratificaciones utilizadas[7]:

- **Cuantiles:** con este método la primera componente principal se divide en cuartiles (cuatro grupos) o terciles (tres grupos) concentrando cada uno el 25% o el 33% de las UPMs del MMV respectivamente.

- **Método de la raíz de la frecuencia acumulada (Dalenius & Hodges, 1959):** este método consiste en la formación de estratos de manera que la varianza obtenida de una variable cuantitativa (en este caso la primera componente principal) sea mínima para el estrato.

- **Estratificación óptima (Lavallée & Hidiroglou, 1988):** este método consiste en un algoritmo iterativo para poblaciones sesgadas (asimétricas), tal que el tamaño de la muestra se minimiza para un nivel dado de precisión expresado en términos del coeficiente de variación. Para la implementación se utiliza el algoritmo de Kozak[8].

[7] Todos los algoritmos fueron implementados a través de R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[8] Este algoritmo desarrolla un método de estratificación que brinda límites cercanos al óptimo cuando se cuenta con variables de estratificación con distribución asimétrica.(Kozak, 2004).

- **Estratificación geométrica (Gunning & Horgan, 2004):** el objetivo de este método es que los coeficientes de variación de la medida de resumen tiendan a ser iguales dentro de los estratos.

Finalmente, los mejores resultados en términos del ACP se obtuvieron al considerar las variables *tasa de ocupación, índice de materialidad de la vivienda, educación e indicador sobre el total de hijos nacidos*. Cabe mencionar que a partir de un trabajo que realizamos basado en el análisis de la Encuesta de Caracterización Socioeconómica Nacional del año 2017 (CASEN 2017), se experimentó una metodología que permitía imputar el *ingreso autónomo per cápita* a los hogares del Censo, esta variable también fue incorporada en los ACP, mostrando que en muchos casos potenciaba los resultados, sin embargo no fue incorporada en los resultados definitivos dadas las limitaciones que se identificaron así como la necesidad de robustecer las conclusiones obtenidas a partir de otras fuentes. En un próximo documento abordaremos esta metodología.

3.3.3 PRINCALS y cuantiles

Descripción general:

PRINCALS (Gifi, 1985) es un acrónimo para Principal Components analysis by means of Alternating Least Squares. PRINCALS se refiere a un algoritmo que incluye toda una familia de soluciones. Una de ellas es la solución por Análisis de Componentes Principales (ACP), solución que requiere que todas las variables sean tratadas como numéricas. Otra opción requiere que todas las variables sean tratadas como nominales múltiples, donde PRINCALS se convierte entonces en lo mismo que HOMALS (Leeuw & Mair, 2009).

PRINCALS puede ser introducido de diferentes formas. Gifi (1985) lo introduce como una extensión de ACP, en el sentido de que además de soportar variables numéricas, también permite el tratamiento tanto de variables nominales como ordinales.

El insumo para PRINCALS es una matriz de datos $\boldsymbol{X}$, con n filas para objetos y p columnas para variables. Los objetos son las unidades de observación, como personas, hogares, viviendas o UPM. Las variables ordenan los objetos en categorías. Por ejemplo, personas con el mismo nivel educacional quedarán en la misma categoría. Cuando una variable continua

es medida, como la edad, las mediciones se redondean a 'número de años'. En la práctica, por lo tanto, esta variable también ordena objetos mediante un limitado número de categorías.

El algoritmo PRINCALS comprende dos etapas: una etapa de **cuantificación** y otra de **ortogonalización**. La etapa de cuantificación consiste en la asignación de valores numéricos a las diferentes categorías de cada variable. Esto da lugar a una matriz de datos cuantificados $\boldsymbol{Q_{n\times p}}$ y columnas de *scores* de los n objetos $\boldsymbol{x}$, donde PRINCALS puede entregar un número diferente de soluciones para $\boldsymbol{Q}$ y $\boldsymbol{x}$, lo que se conoce como *dimensiones*. La cuantificación tiene la característica de maximizar la distancia entre objetos y maximizar la correlación entre variables. Por su parte, la ortogonalización se refiere a que los *scores* calculados deben ser ortogonales (i.e. independientes) para cada dimensión.

Una vez que el proceso converge, como resultado se obtiene un *score* estandarizado (con media igual a cero y varianza unitaria) por objeto para cada dimensión. Si el análisis se realiza con base en la primera dimensión (que explica el mayor porcentaje de varianza), el *score* estandarizado produce un ordenamiento de los objetos de acuerdo al sentido de la correlación de las variables que tienen mayor correlación con esta dimensión. Por ejemplo, si el ingreso y el nivel educacional son las variables que están más correlacionadas con la primera dimensión y la correlación es positiva, entonces significa que a mayor nivel en educación e ingreso, el *score* será mayor.

Como las componentes son combinaciones lineales de las variables en estudio e incluidas en el algoritmo, la misma combinación lineal representa la posición de los individuos en esta escala o puntaje ordenado, donde individuos con el mismo puntaje serán más similares en cuanto a las variables o constructo que se pretende explicar, como bienestar socioeconómico, vulnerabilidad u otro.

Una vez obtenido el *score* para cada objeto, en una o más dimensiones, existen varias alternativas para realizar una clasificación de los objetos, ya sea de manera homogénea a través de cuartiles, quintiles o terciles; o a través de métodos basados en la optimización de una función objetivo, como k-medias y método de la raíz de frecuencia acumulada, entre otros.

El *score* generado para los *n* objetos, es una medida de escalamiento de un constructo o variable latente, es decir, no observada directamente y su ordenamiento no tiene por qué coincidir con un ordenamiento de alguna variable en particular, sino que obedece simultáneamente tanto a un ordenamiento multidimensional como multifactorial.

Implementación sobre datos censales:

El potencial del método PRINCALS es la incorporación tanto de variables numéricas como categóricas en el algoritmo, y esto puede ser aprovechado en la implementación del algoritmo sobre la base censal a nivel de personas, es decir, los *n* objetos de la matriz de información inicial corresponderán a personas. Las variables consideradas en el análisis, que incluyen *condiciones de la materialidad de la vivienda (techo y muro), el nivel educacional y la comuna de residencia*, fueron tratadas a nivel de hogares. El procedimiento realizado fue el siguiente:

i. Se seleccionaron 25 muestras aleatorias a nivel de hogares correspondientes al 20% de la base censal de hogares, usando semillas de aleatorización diferentes para cada una.

ii. Se aplicó el algoritmo PRINCALS usando set de variables diferentes, a fin de maximizar la varianza explicada y el alpha de Cronbach.

iii. Una vez obtenidos los *scores* en dos o tres dimensiones, el resultado se agregó a nivel de UPM mediante el promedio de los *scores* sobre las 25 muestras seleccionadas. El resultado fue una base de 35.094 UPM.

iv. La estratificación de las UPMs se realizó mediante la aplicación de cuantiles (terciles, cuartiles y quintiles) de los *scores* contenidos en la primera dimensión.

La desventaja del método PRINCALS respecto a otros métodos con matriz de información a nivel de UPM, es que depende fuertemente de las variables consideradas a niveles inferiores de agregación. Es decir, si falta una o más variables relevantes correlacionadas con el bienestar socioeconómico, el método de estratificación es menos eficiente.

4 Evaluación

A partir de las técnicas descritas en la sección anterior se obtuvieron 180 vectores de estratificación que consideraban entre tres y cinco particiones o estratos. Como se mencionó anteriormente, dado que el objetivo de la estratificación del marco muestral es su uso en los diseños muestrales para disminuir la incertidumbre de las estimaciones, la metodología de evaluación debe basarse en **la reducción de la varianza para un conjunto de indicadores a nivel de UPM calculados a partir de la base censal**. Un elemento importante de comparación entre dos estrategias muestrales[9] es el efecto de diseño (*DEFF*), que compara la varianza de una estrategia muestral con la varianza del estimador en un diseño por muestreo aleatorio simple. Desde esta perspectiva, una estrategia es mejor que otra si la varianza del estimador es menor o análogamente su DEFF es menor (Bautista, 1988). Por otro lado, Gutiérrez (2016, pág. 184) demuestra que, cuando la asignación es proporcional el *DEFF* puede escribirse en términos de las varianzas poblacionales, por tanto, bajo este supuesto podemos realizar los cálculos directamente sobre la base censal. Adicionalmente, dado el carácter multipropósito de la estratificación, este criterio puede extenderse a un conjunto de *P* variables, estableciendo una medida de calidad multivariada denominada efecto de generalizado $G(S)$ propuesta por Jarque (1981):

$$G(S) = \sum_{p=1}^{P} D\,EFF_p$$

Donde $DEFF_p$ hace referencia al DEFF calculado para el p-ésimo indicador de un set de P indicadores. Finalmente, la mejor estratificación será aquella en donde se minimice G(S).

A continuación, se lista el conjunto de 24 indicadores para los cuales se calcula el DEFF en cada una de las 180 posibles estratificaciones:

[9] Se denomina estrategia muestra a la combinación entre el diseño y el estimador.

- *Tasa de ocupación por sexo.*
- *Tasa de desocupación por sexo.*
- *Tasa de inactividad por sexo.*
- *Porcentaje de extranjeros.*
- *Porcentaje de personas dentro de la fuerza de trabajo primaria respecto de la PEA por sexo.*
- *Porcentaje de hogares unipersonales.*
- *Porcentaje de población femenina ocupada por rama*. La rama de más alta participación femenina: comercio, alojamiento-servicio-comida, enseñanza y actividades de salud.
- *Porcentaje de población masculina ocupada por rama*. La rama de más alta participación masculina: minería, construcción, transporte y agricultura y pesca.
- *Porcentaje de personas según nivel educacional por sexo.*

El esquema de la matriz de evaluación basándose en los vectores de estratificación y 24 indicadores de evaluación se muestra en la tabla V.7.

Tabla V.7.Matriz de evaluación para las estratificaciones del MMV 2017

DEFF	Est_1	...	Est_j	...	Est_{180}
Ind_1	$DEFF_{1,1}$		$DEFF_{1,j}$	...	$\boldsymbol{DEFF_{1,180}}$
Ind_i	$DEFF_{i,1}$	...	$DEFF_{i,j}$	...	$\boldsymbol{DEFF_{i,180}}$

$\mathbf{Ind_{24}}$	$DEFF_{24,1}$	…	$DEFF_{24,j}$	…	$\boldsymbol{DEFF_{24,180}}$
G(S)	$\mathbf{G_1}(s)$	…	$\mathbf{G_j}(s)$	…	$\mathbf{G_{180}}(s)$

Fuente: Instituto Nacional de Estadísticas (INE). Censo-2017 y ENE MAM 2017.

Finalmente, los mejores métodos de estratificación (menor $G(S)$), si se restringe la evaluación a tres y cuatro estratos, son los derivados de la estratificación de la primera componente principal mediante los algoritmos de estratificación óptima y el método de la raíz de la frecuencia acumulada.

5 Resultados sobre el MMV 2017

El método elegido para la estratificación del MMV 2017 fue el de estratificación óptima de la primera componente principal del ACP que considera los indicadores de *porcentaje de personas en la educación superior, tasa de ocupación, porcentaje de viviendas con índice de materialidad alto y el indicador de total de hijos nacidos vivos*. La elección del número de estratos (tres o cuatro), se basó en el análisis de las distribuciones de las UPMs según comuna área. Según esto, dada la baja prevalencia de UPM en los estratos extremos principalmente en el área rural, resultaba más conveniente considerar 3 estratos. La tabla V.8 muestra la distribución a nivel nacional y regional según área de las UPMs y estrato socioeconómico. En el gráfico I.2, se muestra la distribución espacial de las UPMs para el área urbana conurbada del Gran Santiago.

Cabe mencionar que estos resultados se presentan para 35.090 UPM mientras que el MMV 2017 cuenta con 35.149 UPM. Esto se debe principalmente a que al cruzar el MMV con la información de la base Censal hay UPMs cuyas viviendas no cuentan con información, ya que, por ejemplo, en el momento del levantamiento censal no eran viviendas particulares ocupadas.

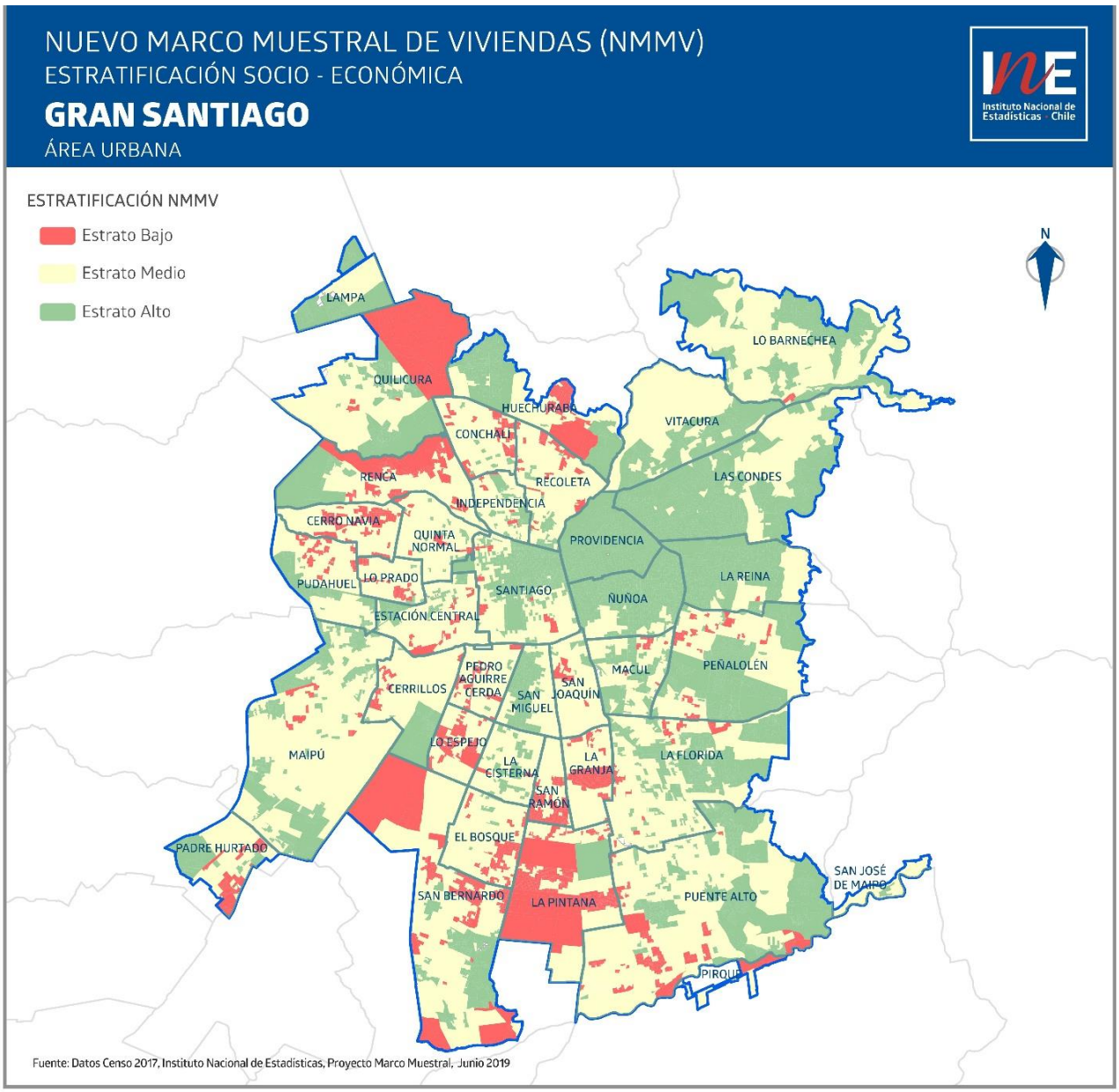
Tabla V.8.Distribución de la estratificación socioeconómica en el MMV 2017 según área y región

Región	Área	Total		Estrato					
				1		2		3	
		Abs.	Rel.%	Abs.	Rel.%	Abs.	Rel.%	Abs.	Rel.%
Total Nacional	**Total**	**35,090**	**100.0%**	**12,728**	**36.3%**	**15,518**	**44.2%**	**6,844**	**19.5%**
	Urbano	26,147	74.5%	5,846	45.9%	13,611	87.7%	6,690	97.7%
	Rural	8,943	25.5%	6,882	54.1%	1,907	12.3%	154	2.3%
Arica y Parinacota	**Total**	**405**	**100.0%**	**178**	**44.0%**	**182**	**44.9%**	**45**	**11.1%**
	Urbano	331	81.7%	108	60.7%	178	97.8%	45	100.0%
	Rural	74	18.3%	70	39.3%	4	2.2%	0	0.0%
Tarapacá	**Total**	**567**	**100.0%**	**193**	**34.0%**	**252**	**44.4%**	**122**	**21.5%**
	Urbano	502	88.5%	133	68.9%	248	98.4%	121	99.2%
	Rural	65	11.5%	60	31.1%	4	1.6%	1	0.8%
Antofagasta	**Total**	**964**	**100.0%**	**254**	**26.3%**	**498**	**51.7%**	**212**	**22.0%**
	Urbano	894	92.7%	203	79.9%	481	96.6%	210	99.1%
	Rural	70	7.3%	51	20.1%	17	3.4%	2	0.9%
Atacama	**Total**	**607**	**100.0%**	**350**	**57.7%**	**180**	**29.7%**	**77**	**12.7%**
	Urbano	462	76.1%	226	64.6%	161	89.4%	75	97.4%
	Rural	145	23.9%	124	35.4%	19	10.6%	2	2.6%
Coquimbo	**Total**	**1,738**	**100.0%**	**946**	**54.4%**	**622**	**35.8%**	**170**	**9.8%**
	Urbano	1,092	62.8%	356	37.6%	568	91.3%	168	98.8%
	Rural	646	37.2%	590	62.4%	54	8.7%	2	1.2%
Valparaíso	**Total**	**3,680**	**100.0%**	**1,075**	**29.2%**	**2,048**	**55.7%**	**557**	**15.1%**
	Urbano	3,000	81.5%	671	62.4%	1,780	86.9%	549	98.6%
	Rural	680	18.5%	404	37.6%	268	13.1%	8	1.4%
Metropolitana	**Total**	**12,015**	**100.0%**	**1,497**	**12.5%**	**6,250**	**52.0%**	**4,268**	**35.5%**
	Urbano	11,100	92.4%	1,159	77.4%	5,726	91.6%	4,215	98.8%
	Rural	915	7.6%	338	22.6%	524	8.4%	53	1.2%
O'Higgins	**Total**	**2,150**	**100.0%**	**1,102**	**51.3%**	**901**	**41.9%**	**147**	**6.8%**
	Urbano	1,196	55.6%	356	32.3%	696	77.2%	144	98.0%
	Rural	954	44.4%	746	67.7%	205	22.8%	3	2.0%

Maule	**Total**	**2,544**	**100.0 %**	**1,482**	**58.3 %**	**880**	**34.6 %**	**182**	**7.2%**
	Urbano	1,373	54.0%	459	31.0%	735	83.5%	179	98.4%
	Rural	1,171	46.0%	1,023	69.0%	145	16.5%	3	1.6%
Ñuble	**Total**	**1,225**	**100.0 %**	**804**	**65.6 %**	**345**	**28.2 %**	**76**	**6.2%**
	Urbano	591	48.2%	225	28.0%	295	85.5%	71	93.4%
	Rural	634	51.8%	579	72.0%	50	14.5%	5	6.6%
Biobío	**Total**	**3,121**	**100.0 %**	**1,568**	**50.2 %**	**1,119**	**35.9 %**	**434**	**13.9%**
	Urbano	2,369	75.9%	887	56.6%	1,053	94.1%	429	98.8%
	Rural	752	24.1%	681	43.4%	66	5.9%	5	1.2%
La Araucanía	**Total**	**2,397**	**100.0 %**	**1,542**	**64.3 %**	**676**	**28.2 %**	**179**	**7.5%**
	Urbano	1,209	50.4%	476	30.9%	558	82.5%	175	97.8%
	Rural	1,188	49.6%	1,066	69.1%	118	17.5%	4	2.2%
Los Ríos	**Total**	**977**	**100.0 %**	**607**	**62.1%**	**296**	**30.3 %**	**74**	**7.6%**
	Urbano	495	50.7%	191	31.5%	236	79.7%	68	91.9%
	Rural	482	49.3%	416	68.5%	60	20.3%	6	8.1%
Los Lagos	**Total**	**2,095**	**100.0 %**	**1,051**	**50.2 %**	**849**	**40.5 %**	**195**	**9.3%**
	Urbano	1,099	52.5%	368	35.0%	563	66.3%	168	86.2%
	Rural	996	47.5%	683	65.0%	286	33.7%	27	13.8%
Aysén	**Total**	**272**	**100.0 %**	**61**	**22.4 %**	**179**	**65.8 %**	**32**	**11.8%**
	Urbano	153	56.3%	17	27.9%	121	67.6%	15	46.9%
	Rural	119	43.8%	44	72.1%	58	32.4%	17	53.1%
Magallanes	**Total**	**333**	**100.0 %**	**18**	**5.4%**	**241**	**72.4 %**	**74**	**22.2%**
	Urbano	281	84.4%	11	61.1%	212	88.0%	58	78.4%
	Rural	52	15.6%	7	38.9%	29	12.0%	16	21.6%

Fuente: Instituto Nacional de Estadísticas (INE).

Gráfico I.2.Estratificación socioeconómica del Gran Santiago (área urbana)



Fuente: Instituto Nacional de Estadísticas (INE).

6 Conclusiones

La metodología desarrollada para la estratificación del MMV 2017 consideró 180 escenarios basados en los métodos de análisis multivariado (ACP y PRINCALS) y algoritmos de clasificación (k-medias y sus variantes, raíz de la frecuencia acumulada, estratificación óptima y estratificación geométrica). El insumo principal fue la base de datos del Censo de Población y Vivienda 2017, a partir de la cual se calcularon indicadores a nivel de UPM expresados en términos de acceso al bienestar y que están directamente relacionados con los fenómenos que se estudiarán por las diferentes encuestas de hogares que harán uso del MMV 2017 y que expresan el grado de bienestar.

Dado que el objetivo de la estratificación del MMV es permitir que los diseños muestrales induzcan una mayor eficiencia y precisión a la hora de la estimación de las estadísticas oficiales, la elección de la mejor estratificación para el marco utilizó como criterio la reducción de la varianza para un conjunto de 24 indicadores a nivel de UPM calculados a partir de la base censal. Específicamente, la medida utilizada fue el efecto de diseño generalizado $G(S)$ propuesta por Jarque (1981). Bajo este criterio, el método elegido es **estratificación óptima de la primera componente principal del ACP que considera los indicadores de porcentaje de personas en la educación superior, tasa de ocupación, porcentaje de viviendas con índice de materialidad alto y el indicador de total de hijos nacidos**. La elección del número de estratos (tres o cuatro), se basó en el análisis de las distribuciones de las UPMs según comuna y área. Según esto, dada la baja prevalencia de UPM en los estratos extremos, principalmente en el área rural, era más conveniente considerar **3 estratos.**

Finalmente, y teniendo en cuenta los procesos de actualización en el MMV 2017 durante el período intercensal, se debe estudiar el impacto que tendrán dichas actualizaciones sobre los resultados del marco (estratificación socioeconómica, áreas de levantamiento especial[10]) y la pertinencia de desarrollar una metodología que permita la actualización de dichos resultados, garantizando en el caso de la estratificación socioeconómica las bondades vinculadas con los diseños muestrales.

[10] Las áreas geográficas de tratamiento especial, basadas en el Censo 2017, son áreas pobladas que por su localización geográfica, presentan problemas operativos para el levantamiento de la información. Estas dificultades están asociadas al grado de accesibilidad que presentan algunos territorios, dados por la lejanía a los centros operativos, la falta de caminos, el tiempo requerido, el aislamiento, el clima y la altura, entre los más importantes.

7 Referencias

Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. *Wiley Interdisc. Rev.*, 433-459.

Aurenhammer, F. (1991). Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys*, *23*(3):345 - 405.

Bautista, L. (1988). *Diseño de muestreo estadístico*. Bogotá: Universidad Nacional de Colombia. Departamento de Matemáticas y Estadística. Unidad de extensión y asesoría.

Dalenius, T., & Hodges, J. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association 54 No. 285*, 88-101.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. *Biometrics*, *21*, 768 – 769.

Gifi, A. (1985). *PRINCALS*, Department of Data Theory, Leiden

Gunning, P., & Horgan, J. M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, 159 - 166.

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

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, *28*, 100–108. doi:10.2307/2346830.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables. *Journal of Educational Psychology*, 417-441.

Jarque, C. M. (1981). «A Solution to the Problem of Optimum Stratification in Multivariate Sampling». *Journal of the Royal Statistical Society. Series C* (Applied Statistics) *30* (2): 163 - 69. https://doi.org/10.2307/2346387.

Kozak, M. (2004). Optimal Stratification Using Random Search Method in Agricultural Surveys. *Statistic in Transition, 6*(5), 797-806.

Lavallée, P., & Hidiroglou, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology , 14*(1), 33-43.

Leeuv, J. & Mair, P. (2009). Gifi Methods for Optimal Scaling in R: The package homals, *Journal of statistical software*, *31*(4): 1 – 21.

Lloyd, S. P. (1957, 1982). Least squares quantization in PCM. Technical Note, Bell Laboratories. Published in 1982 in *IEEE Transactions on Information Theory*, *28*, 128–137.

MacKay, D. (2003). Chapter 20. An Example Inference Task: Clustering. Information Theory, Inference and Learning Algorithms. Cambridge University Press. 284 – 292. ISBN 0-521-64298-1. MR 2012999.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281 – 297. Berkeley, CA: University of California Press.

Montenegro, F. M. T., & Brito, J. A. D. M. (2006). Um Algoritmo Genético para o Problema de Agrupamento. *Anais do XXVIII SOBRAPO-Simpósio Brasileiro de Pesquisa Operacional-Gôiania–GO*.

Pardo, C. E., & Del Campo Neira, P. (2007). Combination of Factorial Methods and Cluster Analysis in R: The Package FactoClass. 30. . *Revista Colombiana de Estadística 30*, 231-245.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points. Philosophical Magazine.