METODOLÓGICO

Documentos de trabajo

Estratificación socioeconómica del marco muestral de viviendas 2017 (MMV 2017)-Implementación en R

Autores:
Julio Guerrero
Andrea Marín

Nº.17, junio de 2020

Instituto Nacional de Estadísticas · Chile

INSTITUTO NACIONAL DE ESTADÍSTICAS

Morandé 801, Santiago de Chile
Teléfono: 562 3246 1000
Correo: ine@ine.cl
Facebook: @ChileINE
Twitter: @INE_Chile

Julio Guerrero
Andrea Marín

Departamento de Metodologías e Innovación Estadística
Subdepartamento de Investigación Estadística

Los autores agradecen la valiosa contribución en los análisis preliminares y de preprocesamiento de datos realizado por Klaus Lehmann, Patricia Mauna y Alfredo Heufemann. Finalmente, como equipo de trabajo agradecemos de manera especial la colaboración y guía de la Unidad de Estadísticas Sociales de la CEPAL, representada por el Sr. Andrés Gutiérrez, experto regional en estadísticas sociales, y por el 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) – Implementación en R

Resumen

La estratificación socioeconómica del marco de muestreo derivado del Censo de Población y Vivienda de 2017 (MMV 2017), permite aumentar la eficiencia de la inferencia en las encuestas de hogares que utilizan este marco.

En este documento presentamos la metodología utilizada para el proceso de estratificación y brindamos información sobre su implementación mediante el *software* estadístico R, un sistema altamente extensible para computación estadística y gráficos, distribuido en la red. Basados en la experiencia nacional y regional, así como en la literatura estadística, la metodología elegida cubre diferentes métodos estadísticos no supervisados, y su implementación se llevó a cabo sobre un conjunto de indicadores socioeconómicos que miden bienestar, llamado matriz de información. Consideramos dos escenarios principales para propósitos de estratificación: el univariante (sobre una medida resumida de la matriz de información), donde se implementaron métodos como Lavallée e Hidiroglou o la estratificación óptima; y el multivariante (sobre todas o algunas de las variables de la matriz de información), donde se implementaron diferentes variantes del algoritmo k-medias.

Posteriormente, para elegir cuál de las estratificaciones obtenidas era la más apropiada, se implementó como método de evaluación, el cálculo del efecto de diseño generalizado *G(S)*. Este método tiene en cuenta tanto el objetivo de la estratificación de un marco muestral como su carácter multipropósito, pues será usada en los diseños muestrales de diferentes encuestas sociales. Bajo este criterio, el método elegido para la estratificación del MMV 2017 fue clasificación mediante el algoritmo de estratificación óptima sobre la primera componente principal para tres grupos.

Abstract

The socioeconomic stratification of the sampling frame derived from the Population and Housing Census of 2017 (MMV 2017), allows to increase the efficiency of inference in household surveys that use this frame.

In this document we present the methodology used for the stratification process and give insights into its implementation by statistical software R, a highly extendable system for statistical computing and graphics, distributed over the net. Based on national and regional experience, as well as statistical literature, the chosen methodology covers different statistical unsupervised methods, and their implementation took place on a set of socioeconomic indicators that measure well-being, called information matrix. We considered two major scenarios for stratification purposes: the univariate (on a summary measure of the information matrix); where methods as Lavallée and Hidiroglou or optimal stratification was implemented; and the multivariate (on all or some of the variables of the information matrix), where different variants of k-medias algorithm was implemented.

Subsequently, to choose which of the stratifications obtained was the most appropriate, the calculation of the generalized design effect *G (S)* was implemented as an evaluation method. This method takes into account both the objective of the stratification of a sampling frame and its multipurpose character, in the sense that it will be used by different social surveys. Based on this criterion, the method chosen for the stratification of the MMV 2017 uses the first principal component for classification by optimal stratification algorithm for three groups.

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

Índice

1 Introducción

La estratificación socioeconómica del marco muestral 2017 (MMV 2017) o la clasificación de las unidades primarias de muestreo (UPMs) 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. Así, se garantiza la homogeneidad entre grupos y se disminuye la incertidumbre de la estimación.

En este trabajo presentamos los procedimientos computacionales necesarios para implementar mediante el *software* estadístico `R` (R Core Team, 2019), la metodología utilizada por el Instituto Nacional de Estadísticas (INE) para la estratificación del MMV 2017, que se encuentra descrita en el documento Estratificación socioeconómica del marco muestral 2017 (MMV 2017).[1]

Este documento está dividido en las siguientes secciones principales:

1. La sección **Metodología de Estratificación** presenta un esquema de la metodología implementada por el INE y describe los escenarios propuestos para la estratificación.

2. La sección **Información a nivel de UPM** describe el conjunto de datos utilizados para realizar la estratificación, criterios de construcción de indicadores y consideraciones para el procesamiento de la información.

3. La sección **Algoritmo de k-medias** describe el enfoque de una estrategia de estratificación multivariada, basada en la implementación de diferentes variantes del algoritmo de k-medias. Esta sección está enriquecida con instrucciones de código en `R` y resultados obtenidos.

4. La sección **Análisis de componentes principales y algoritmos de estratificación univariada** describe el enfoque de una estrategia de estratificación univariada, la que se basa en la información contenida en la primera componente principal obtenida a partir de la matriz de indicadores socioeconómicos. Esta sección está enriquecida con instrucciones de código en `R` y resultados obtenidos.

5. La sección **Evaluación** presenta los criterios de evaluación de los resultados de estratificación.

6. La sección **Discusión** presenta una discusión de los resultados obtenidos, procedimientos aplicados y sus conclusiones.

[1] Documento disponible en https://www.ine.cl/inicio/documentos-de-trabajo/documento/estratificaci%C3%B3n-socioecon%C3%B3mica-del-marco-muestral-de-viviendas-2017.

2 Objetivo

Este documento tiene por objetivo describir los procedimientos computacionales necesarios para implementar mediante el *software* estadístico `R`, la metodología utilizada por el Instituto Nacional de Estadísticas (INE) para la estratificación del MMV 2017.

3 Metodología de estratificación

La teoría estadística ha definido que la mejor estratificación es aquella que minimiza los errores de muestreo de los estimadores, expresados en forma de varianzas o errores estándar. Existe evidencia empírica de que la mayoría de los fenómenos que se observan en las encuestas de hogares están supeditados a la distribución de la población en las UPMs. Lo anterior, conlleva a afirmar que existe una alta correlación entre la UPM que se habita y la incidencia de los fenómenos sociales y económicos. Por tanto, los ejercicios de estratificación que consideren el uso de información a nivel de UPM tendrán una alta consistencia interna, de tal manera que, al escoger la mejor estratificación se garantiza que el INE dispondrá de una clasificación óptima en el período intercensal para todas las encuestas de hogares que se ejecuten.

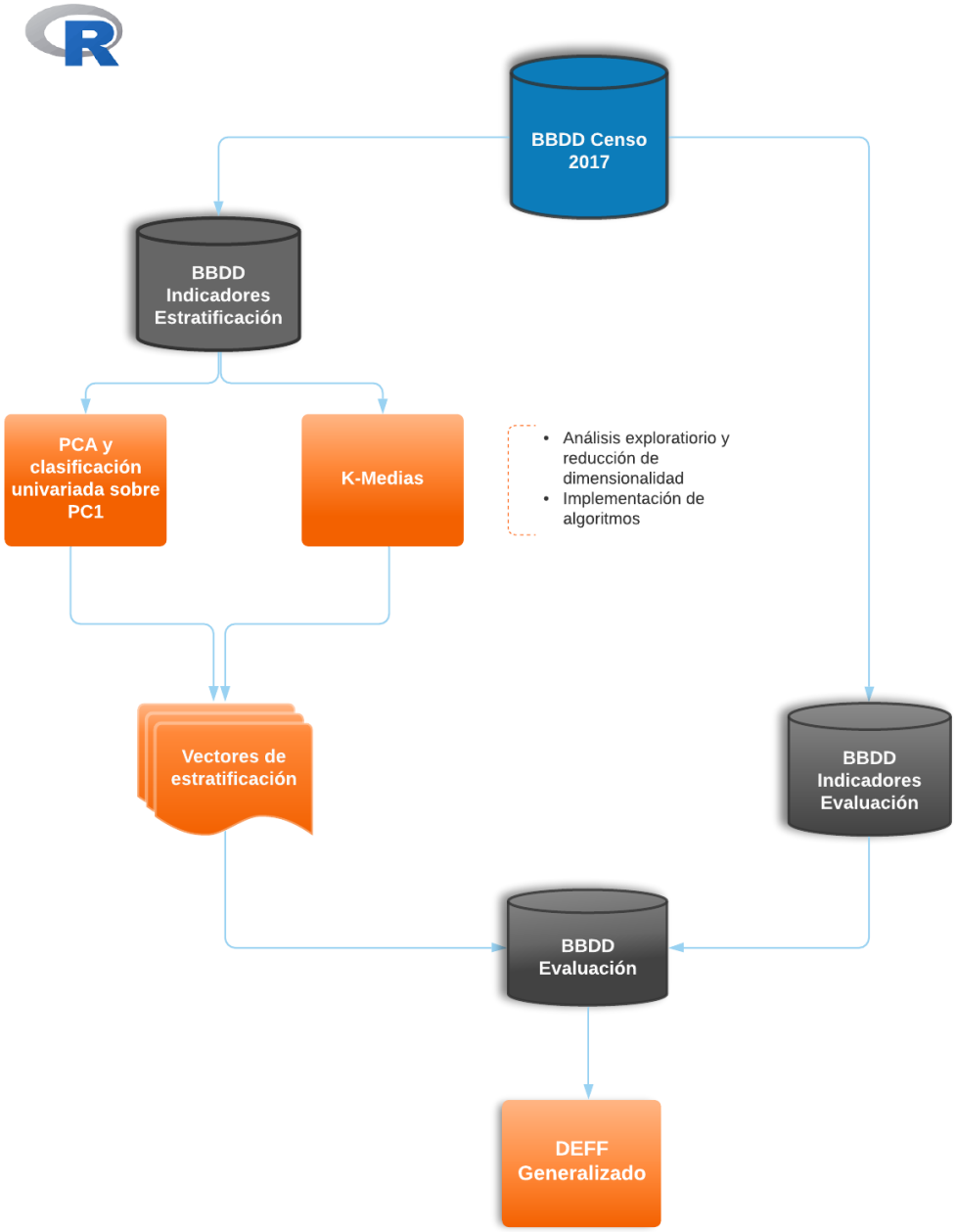
A modo general, la metodología para la estratificación socioeconómica del MMV 2017 se basó en el uso de indicadores socioeconómicos a nivel de UPM, construidos a partir de la base del XIX Censo Nacional de Población y VIII de Vivienda levantado en abril 2017. Estos indicadores forman parte de lo que llamaremos matriz de información X, que serán descritos en la sección **Información a nivel de UPM**.

Existen dos grandes escenarios que hemos considerado al momento de proponer una estratificación: el univariado (sobre una medida de resumen de la matriz de información), que tuvo lugar mediante la implementación de Análisis de Componentes Principales y Clasificación Univariada; y el multivariado (sobre todas o algunas de las variables de la matriz de información), que tuvo lugar mediante la implementación del algoritmo de k-medias. Cabe señalar que, para efectos de la metodología descrita en el documento Estratificación socioeconómica del marco muestral 2017 (MMV 2017), tuvo lugar la implementación del algoritmo PRINCALS (Gifi, 1985); sin embargo, la implementación de dicho algoritmo no fue realizada en `R` y, por lo tanto, no será descrita en este documento.

En el Grafico I.1 presentamos el esquema general de la metodología de estratificación, sobre el cual desarrollaremos el paso a paso en este trabajo, enriquecido con códigos en `R`. Cabe señalar que, no realizaremos una impresión y análisis exhaustivo de los resultados

obtenidos, sino que nuestro objetivo es que el lector comprenda la metodología de estratificación y su implementación.

Gráfico I.1.Esquema general de metodología de estratificación



Fuente: Instituto Nacional de Estadísticas (INE).

4 Información a nivel de UPM

La información a nivel de UPM sobre la cual procederán los distintos esquemas de estratificación mencionados en la sección anterior, se basa en un conjunto de datos compuesto por una serie de indicadores socioeconómicos. Estos 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.

Debido a que las UPMs tienen, en estricto rigor, tamaños diferentes, la escala y el nivel en el que se midan los indicadores puede afectar los procesos de clasificación. Luego, si la matriz de información con la cual se realiza la estratificación se construye con base en el número de personas (con determinadas características) dentro de la UPM, al no tener en cuenta el tamaño de esta, es muy probable que las metodologías de estratificación no logren agruparlas de forma homogénea. Por ende, definir la matriz de información en términos relativos (porcentaje de ocurrencia de cada variable) es una mejor alternativa para que el agrupamiento esté controlado por el tamaño de la UPM y supeditado únicamente a cambios estructurales en los constructos de medición del censo.

Todos los indicadores, ya sea referidos a personas, hogares o viviendas, construidos a partir de la información del XIX Censo Nacional de Población y VIII de Vivienda levantado en abril de 2017, representan el porcentaje de personas, hogares o viviendas dentro de la UPM que tienen determinada característica que indica bienestar.

En consecuencia, se debe contar con una matriz de información $\mathbf{X}$ compuesta por P variables, y n filas; en donde cada fila de la matriz de información representará la observación de las UPMs a nivel censal para cada una de las P variables. En la Tabla V.1 se muestran los indicadores construidos a nivel de UPM a partir de la base censal. Para efectos de este documento, se entrega el identificador (ID) de cada indicador, los criterios adoptados y descripción de cada uno que pueden consultarse en Estratificación socioeconómica del marco muestral 2017 (MMV 2017).

Tabla V.1.Indicadores para estratificación a nivel de UPM

ID	Indicador / Título del recurso
educ_sup	Proporción de personas en la UPM con educación superior.
t_ocupa	Proporción de personas en la UPM ocupadas.
t_ocupa_h	Proporción de hombres en la UPM ocupados.
t_ocupa_m	Proporción de mujeres en la UPM ocupados.
t_desocupa	Proporción de personas en la UPM desocupadas.
t_desocupa_h	Proporción de hombres en la UPM desocupados.
t_desocupa_m	Proporción de mujeres en la UPM desocupados.
t_part	Proporción de personas en la UPM con participación en el mercado laboral.
t_part_h	Proporción de hombres en la UPM con participación en el mercado laboral.
t_part_m	Proporción de mujeres en la UPM con participación en el mercado laboral.
r_depe	Proporción de personas en la UPM que son dependientes.
gen_ingre	Proporción de personas en la UPM que son generadores de ingreso.
r_mhij	Proporción de mujeres >=15 años en la UPM con 2 o menos hijos nacidos vivos.
no_hacina	Proporción de viviendas en la UPM sin hacinamiento.
viv_muro	Proporción de viviendas en la UPM con muros en condición de bienestar [2].
acc_agua	Proporción de viviendas en la UPM con acceso al agua por red pública.
ind_mat	Proporción de viviendas en la UPM con índice de materialidad alto.
q45_ingre[3]	Proporción de hogares en la UPM que se encuentran en el cuarto o quinto quintil de ingresos.

Fuente: Instituto Nacional de Estadísticas (INE).

Conjuntamente, es necesario realizar un proceso de refinamiento sobre esta matriz para eliminar aquellas variables que puedan estar altamente correlacionadas con el resto de las variables o que puedan expresarse como combinación lineal de otras variables. De esta manera, se evitan los problemas de multicolinealidad y se asegura una estratificación parsimoniosa.

El primer paso es explorar los datos. Para analizar los datos en R, primero debemos leer los datos en `R`. En este caso, los datos están guardados en un archivo *.csv* (comma-separated values, por sus siglas en ingles). El código `R` en el *Listing 1* para este estudio de caso demuestra cómo hacer esto.

[2] Para el área urbana se consideró como material de bienestar muros de hormigón armado o albañilería, mientras que para el área rural se consideró como material de bienestar muros de hormigón armado, albañilería o tabique forrado por ambas caras.

[3] La estimación de los ingresos y su clasificación en quintiles en la base censal fue realizada mediante un modelo predictivo para los ***ingresos autónomos per cápita*** de los hogares de Chile, desarrollado a partir de la base Casen 2017. La descripción de dicho modelo va más allá del alcance de este documento y no forma parte de ninguna publicación oficial del INE, por lo que, para efectos de este trabajo, solo se hará referencia al indicador generado.

Después de configurar el directorio de trabajo donde se almacena el archivo *.csv*, cargamos los datos en el objeto llamado *UPM_data*. Todos los resultados, a menos que se especifique lo contrario, se guardan en el directorio de trabajo.

Listing 1. Carga de datos

```
setwd("../Estratificacion/BBDD")
UPM_data <- read.csv2("UPM_data.csv", encoding="UTF-8")
```

Verificamos la cantidad de variables, la cantidad de observaciones y los nombres de las variables, como se muestra en el *Listing 2*.

Listing 2. Número de observaciones y de variables y nombres de variables

```
dim(UPM_data)
vars<-gsub(",", "", names(UPM_data))
vars
```

El conjunto de datos tiene 35.085 observaciones (UPM en el MMV) y 32 variables. Las primeras 13 variables corresponden a identificadores territoriales como región, provincia, comuna, etc., mientras que las restantes corresponden a los indicadores socioeconómicos descritos en la Tabla 1.

Es importante señalar que cada UPM debe quedar clasificada en tan solo un grupo socioeconómico y, por ende, es necesario que nos aseguremos que los ID de las UPMs contenidos en la variable "*id_upm*" sean únicos. Además, es necesario garantizar que la información contenida en **X** sea completa, es decir, que no se observen valores faltantes (NA para efectos de `R`). El código `R` en el *Listing 3* demuestra cómo realizar la revisión de valores faltantes para cada una de las variables y la revisión de identificadores únicos para las UPM en el conjunto de datos.

Listing 3. Revisión de valores faltantes e identificadores únicos

```
# Revision de NA
na       <- apply(UPM_data,2,function(x) 100*(sum(is.na(x)))/length(x));na
# Revision de ID de UPM unicos
id_uni   <- length(UPM_data$id_upm)==length(unique(UPM_data$id_upm))

id_uni
```

Como resultado se obtuvo ausencia de valores faltantes para todas las variables en el conjunto de datos e ID únicos para las UPMs.

En consecuencia, disponemos de un conjunto de datos para realizar la estratificación socioeconómica de las UPMs.

5 Algoritmo de k-medias

El algoritmo de k-medias es un algoritmo de clasificación no supervisada (clusterización) que agrupa *n* objetos en *k* grupos basándose en sus características. El agrupamiento se realiza minimizando la suma de distancias entre cada objeto y el centroide de su grupo o *cluster*. Se suele usar la distancia cuadrática. El algoritmo consiste en tres pasos:

1. **Inicialización:** una vez escogido el número de grupos, *k*, se establecen *k* centroides en el espacio de los datos, por ejemplo, escogiéndolos aleatoriamente.

2. **Asignación de objetos a los centroides:** cada objeto de los datos es asignado a su centroide más cercano.

3. **Actualización de los centroides**: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los objetos pertenecientes a dicho grupo.

Se repiten los pasos 2 y 3 hasta que los centroides no se mueven, o se mueven por debajo de una distancia umbral en cada paso.

Hay un buen número de versiones del algoritmo de 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 hemos considerado cinco variaciones del algoritmo: Forgy (Forgy, 1965), MacQueen (MacQueen, 1967), Lloyd (Lloyd, 1982), Hartigan-Wong (Hartigan & Wong, 1979) y Jarque (Jarque, 1981).

En esta sección se describe el procedimiento utilizado para la clasificación de las 35.085 UPM en *k* grupos socioeconómicos. El procedimiento consiste en dos pasos:

1. **Reducción de dimensionalidad:** proceso que se refiere al refinamiento de la matriz *X* mediante la eliminación de información redundante.
2. **Implementación:** implementación del algoritmo de k-medias.

5.1 Reducción de dimensionalidad

Como se mencionó anteriormente, es necesario realizar un proceso de refinamiento sobre la matriz **X** para eliminar aquellas variables que puedan estar altamente correlacionadas con el resto de las variables o que puedan expresarse como combinación lineal de otras variables. Así, se elimina información redundante y se logra una estratificación más parsimoniosa. Un procedimiento útil para la reducción de dimensionalidad y selección de variables es el *cluster* de variables, que consiste en organizar las variables en grupos homogéneos, es decir, grupos de variables que están fuertemente relacionadas entre sí y, por lo tanto, brindan la misma información. La librería *ClustOfVar* (Saracco, 2015) ha sido desarrollada específicamente para este propósito.

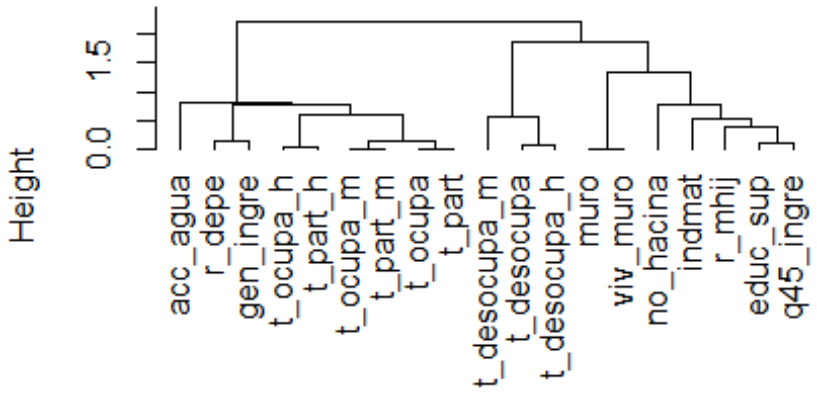
El *Listing 4* muestra cómo realizar el cluster de variables con la base de UPM, como recomendación, debemos estar seguros de tener instaladas las librerías necesarias para realizar cada procedimiento, de no ser así, debemos instalarlas antes de intentar ejecutar el código. Ya con las librerías instaladas, las cargamos en nuestra sesión mediante la función *library()*.

Listing 4. Cluster de variables

```
library(ClustOfVar)
# Seleccion de variables
xquant <- UPM_data[,c(14:32)]

# Hierarchal clustering
tree <- hclustvar(xquant)
plot(tree)
```

Gráfico I.2.Cluster de Variables



Fuente: Instituto Nacional de Estadísticas (INE).

En el Gráfico I.2 se observa que los niveles de agregación sugieren elegir entre 5 y 8 grupos de variables, dependiendo el nivel de agrupamiento que se elija. Se observa que las agrupaciones muestran coherencia en términos del contexto del problema, como por ejemplo, las variables relacionadas con ocupación y participación en la fuerza laboral, las variables relacionadas con desocupación, o las variables relacionadas con estudios superiores y niveles de ingresos altos. Es importante que el lector tenga en mente que el dendrograma no indica el signo de estas relaciones, por lo que un cuidadoso estudio de estas puede realizarse mediante un análisis de correlaciones.

El usuario puede usar la función *stability()* para tener una idea de la estabilidad de las particiones del dendrograma anterior. En el *Listing 5* se muestra cómo hacer esto.

Listing 5. Estimación del número de clusters

```
stab <- stability(tree, B=50)
stab$meanCR
stab$matCR
boxplot(stab$matCR, main="Dispersion of the ajusted Rand index")
```

A través de la media (sobre 50 muestras *bootstrap*) del índice Rand ajustado (Rand, 1971) obtenido a partir del procedimiento anterior, establecemos en cinco el número de *clusters* de variables. En el *Listing 6* se muestra el código `R` que permite obtener la composición e información referente a los *clusters*, en este caso con k=5.

Listing 6. Partición para clustering k-means

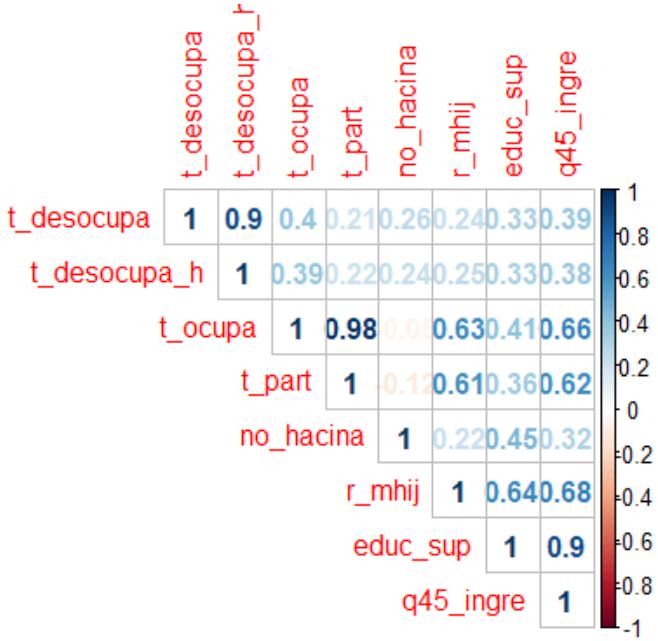
```
k.means <- kmeansvar(xquant, init=5)
summary(k.means)
```

En complemento al análisis cluster, se realizó un análisis de correlaciones. El *Listing 7* muestra cómo realizarlo.

Listing 7. Matriz de correlaciones

```
library (corrplot)
corr<-cor(round(UPM_data[,14:32],2));corrplot(corr, method="number", type="upper")
library(PerformanceAnalytics)
chart.Correlation(UPM_data[,14:32], histogram = T, method="pearson", pch = 19)
```

Gráfico I.3.Matriz de correlaciones de los indicadores.



Fuente: Instituto Nacional de Estadísticas (INE).

Para efectos de ilustración y mejor visualización de los resultados, en el Gráfico I.3 se muestra una parte de la matriz de correlaciones de los indicadores. Claramente se observa

que existen altas correlaciones entre variables, por ejemplo, las variables relacionadas con ocupación y participación en la fuerza laboral (0.98) y estudios superiores e ingresos altos (0.90), lo que es consistente con lo observado en el análisis *cluster*.

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 las siguientes variables: *"viv_muro", "educ_sup", "t_desocupa", "no_hacina"* y *"t_ocupa"*. 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.

5.2 Implementación

5.2.1 Estrategia 1: mejores variables

La primera estrategia consistió en implementar el algoritmo para cada una de las cinco variantes mencionadas, usando solo las variables seleccionadas en la sección anterior, a saber*: "viv_muro", "educ_sup", "t_desocupa", "no_hacina"* y *"t_ocupa"*. Para efectos de la ilustración, solo se mostrará el procedimiento para la variante de "Hartigan-Wong", que mostró los mejores resultados. No obstante, instruiremos cómo proceder para obtener los resultados bajo las otras variantes del algoritmo.

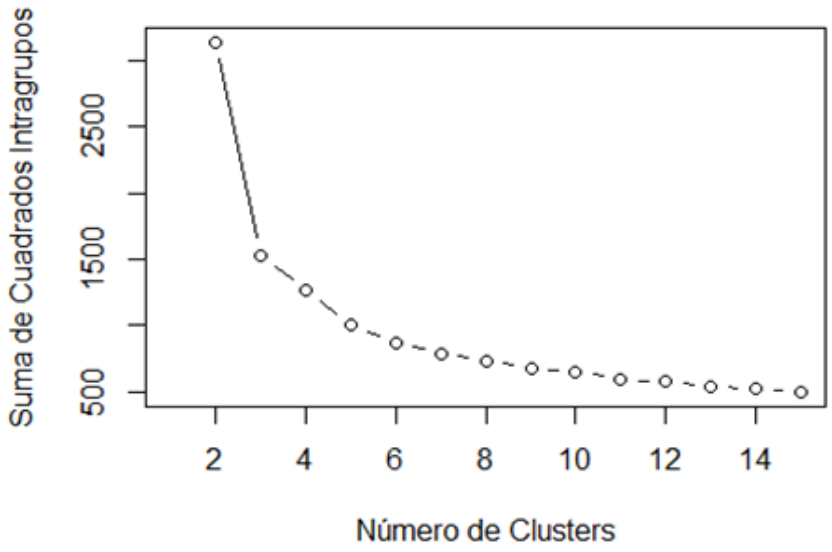
En primer lugar, conviene aclarar que el resultado del procedimiento depende del punto de arranque de este, específicamente en la elección de los centroides (Paso 1), por lo que, para garantizar la reproducibilidad de resultados, es necesario establecer semilla. En `R` esto se hace mediante la función *set.seed()*. En segundo lugar, se debe establecer el número de grupos o *clusters* a considerar. Para hacer esto, se puede emplear un código `R` como el que se muestra en el *Listing 8*.

Listing 8. Selección de "k" para k-medias

```
# fijar Semilla
set.seed(300)

# eleccion de "k" para k-medias
# sc intra
wss <- (nrow(UPM_data)-1)*sum(apply(UPM_data,1,var))
for (i in 2:15) wss[i] <- sum(kmeans(UPM_data[,c("viv_muro","educ_sup","t_desocupa
",
                                                 "no_hacina","t_ocupa")],
                                     centers=i,iter.max = 10,
                                     algorithm ="Hartigan-Wong")$withinss)
plot(1:15, wss, type="b", xlab="Número de Clusters",ylab="Suma de Cuadrados Intrag
rupos")
```

Gráfico I.4. Selección *de k para k-medias.*



Fuente: Instituto Nacional de Estadísticas (INE).

Como se puede apreciar en el Gráfico I.4, la suma de cuadrados intragrupo disminuye conforme aumenta el número de *clusters*. Sin embargo, para fines prácticos y para el objetivo de la estratificación socioeconómica de UPMs, se consideró k=3, 4 y 5.

La función `R` desarrollada para implementar el algoritmo de k-medias es *kmeans()*, y queremos llamar la atención del lector a dos de sus argumentos: *centers* y *algorithm*. Mediante el argumento *centers* el usuario puede especificar el número de grupos o *clusters* (en nuestro caso 3, 4 o 5); el argumento *algorithm* permite especificar la variante que se desea implementar, y como se mencionó anteriormente, se ilustrará el procedimiento para la variante de "Hartigan-Wong". Para especificar una variante distinta, el usuario debe modificar este argumento y usar las variantes "Forgy", "Lloyd" o "MacQueen" para obtener los resultados bajo estos esquemas. Para conseguir resultados bajo el algoritmo de Jarque, se requiere un paso adicional que presentaremos más adelante.

En el *Listing 9* se muestra cómo implementar el algoritmo k-medias versión "Hartigan-Wong" con k=3.

Listing 9. Algoritmo Hartigan-Wong con k=3

```
# Fijar semilla
set.seed(300)
# HW3
clust_HW3<-kmeans(UPM_data[,c("viv_muro","educ_sup","t_desocupa",
                              "no_hacina","t_ocupa")],
                  centers=3,iter.max = 10,
                  algorithm = "Hartigan-Wong");clust_HW3
```

Una serie de resultados se obtienen del procedimiento anterior: el número de UPMs que compone cada uno de los tres grupos, el vector con el *cluster* al que pertenecen las primeras 1.000 UPM, la bondad de ajuste del agrupamiento y los componentes disponibles del objeto *clust_HW3*. Entre los resultados mencionados, invitamos al lector a prestar atención al resultado relacionado con la bondad de ajuste del agrupamiento, cuya salida de `R` es:

```
Within cluster sum of squares by cluster:
[1] 421.5039 733.1063 372.9277
(between_SS / total_SS = 70.8 %)
```

Que resulta moderadamente buena.

En el *Listing 10* se muestra cómo implementar el algoritmo k-medias versión "Hartigan-Wong" con k=4.

Listing 10. Algoritmo Hartigan-Wong con k=4

```
# Fijar semilla
set.seed(300)
# HW4
clust_HW4<-kmeans(UPM_data[,c("viv_muro","educ_sup","t_desocupa",
                              "no_hacina","t_ocupa")],
                  centers=4,iter.max = 10,
                  algorithm = "Hartigan-Wong");clust_HW4
```

La salida es similar a la del caso anterior. Como era de esperarse, la bondad de ajuste del agrupamiento mejoró debido a considerar una cantidad de grupos mayor. La salida de `R` es:

```
Within cluster sum of squares by cluster:
[1] 468.5021 242.9800 159.8711 386.9596
(between_SS / total_SS = 75.9 %)
```

El resultado es moderadamente bueno.

En el *Listing 11* se muestra cómo implementar el algoritmo k-medias versión "Hartigan-Wong" con k=5.

Listing 11. Algoritmo Hartigan-Wong con k=5

```
# Fijar semilla
set.seed(300)
# HW5
clust_HW5<-kmeans(UPM_data[,c("viv_muro","educ_sup","t_desocupa",
                              "no_hacina","t_ocupa")],
                  centers=5,iter.max = 10,
                  algorithm = "Hartigan-Wong");clust_HW5
```

Para k=5, la salida de `R` es:

```
Within cluster sum of squares by cluster:
[1] 247.6333 170.1295 133.1750 214.4581 238.7988
(between_SS / total_SS = 80.8 %)
```

Que resulta en una buena bondad de ajuste.

En consecuencia, el algoritmo de k-medias versión "Hartigan-Wong" entrega los mejores resultados para k=5. Sin embargo, la decisión de "mejor clasificación" no se basa únicamente en los resultados entregados por el algoritmo, sino en criterios estadísticos que guardan relación con los diseños de las encuestas que emplearán el marco estratificado, que serán descritos en la sección **Evaluación**. Por tanto, es necesario guardar todos los resultados obtenidos. Para ello, crearemos una matriz de resultados que llamaremos estratos, la que iremos completando conforme obtengamos resultados de estratificación. Esta matriz nos será de gran utilidad en la etapa de **Evaluación**. En el *Listing 12* se muestra cómo generar la matriz *estratos* y cómo poblarla con los resultados obtenidos.

Listing 12. Guardado de los resultados

```
# Creamos el objeto estratos 'vacio'
estratos <-dplyr::as_tibble(matrix(NA, ncol = 13,nrow = nrow(UPM_data)))
# Asignamos nombre a la columna donde se guardarán los identificadores de las UPMs
names(estratos)[1]<-paste("id_","upm",sep="")
# Llenamos la columna id_upm con el vector de identificador de UPMs
estratos[,1] <- UPM_data$id_upm
# Asignamos nombre a la columna donde se guardarán los resultados de clust_HW3
names(estratos)[2]<-paste("HW",3,sep="")
# Llenamos la columna HW3 con el vector de grupos (estratos) de clust_HW3
estratos[,2] <- clust_HW3$cluster
# Asignamos nombre a la columna donde se guardarán los resultados de clust_HW4
names(estratos)[3]<-paste("HW",4,sep="")
# Llenamos la columna HW4 con el vector de grupos (estratos) de clust_HW4
estratos[,3] <- clust_HW4$cluster
# Asignamos nombre a la columna donde se guardarán los resultados de clust_HW5
names(estratos)[4]<-paste("HW",5,sep="")
```

```
# Llenamos la columna HW5 con el vector de grupos (estratos) de clust_HW5
estratos[,4] <- clust_HW5$cluster
```

Por último, instruimos al lector sobre cómo implementar la versión Jarque. Esta versión corresponde a una extensión de la versión "MacQueen", y requiere una transformación de los datos que se puede lograr mediante el código `R` en el *Listing 13*.

Listing 13. Transformación de datos para implementación de versión Jarque

```
for (i in 14:32)
  UPM_data[,i]=UPM_data[,i]*100

mydataJ = scale(UPM_data[,c(14:32)], center = F,
                scale = apply(UPM_data[,c(14:32)], 2, sd))
```

Una vez transformados (re-escalados) los datos, aplicamos los códigos `R` presentados en los *Listing 9, Listing 10* y *Listing 11,* especificando el argumento algorithm="MacQueen" en cada caso.

5.2.2 Estrategia 2: Primera Componente Principal (PC1) por cada cluster de variables

La segunda estrategia implicó la realización de análisis de componentes principales (PCA, por sus siglas en inglés), que es una técnica estadística utilizada para la reducción de dimensionalidad de un conjunto de datos mediante una transformación lineal que describe dicho conjunto en términos de nuevas variables (componentes) no correlacionadas. Como se vio anteriormente, se tiene cinco *clusters* de variables, y la idea es realizar un PCA para cada uno de ellos, tomando en cada caso la primera componente principal (PC1), que es la que explica la mayor proporción de varianza de los datos (más cuanto más correlacionadas estén las variables originales).

Así, tal como para la estrategia 1, tendremos cinco variables de entrada. La diferencia está en que para la estrategia 1 se seleccionó solo una variable de cada *cluster*, mientras que, bajo esta estrategia, seleccionamos una variable (PC1) que resume la información contenida en las variables que componen cierto *cluster*.

En el *Listing 14* se muestra cómo realizar el análisis de componentes principales para cada *cluster* de variables.

Listing 14. Análisis de componentes principales

```
acp1=princomp(UPM_data[,c("t_part","t_ocupa","gen_ingre","t_ocupa_h",
                          "t_part_h","t_part_m","t_ocupa_m","r_depe","r_mhij",
                          "acc_agua")], cor = TRUE) #3
```

```
acp2=princomp(~ ., data = UPM_data[,c("t_desocupa","t_desocupa_h","t_desocupa_m")]
,
                           cor = TRUE) #2
acp3=princomp(~ ., data = UPM_data[,c("viv_muro","muro","indmat")], cor = TRUE) #2
acp5=princomp(~ ., data = UPM_data[,c("educ_sup","q45_ingre")], cor = TRUE) #1
```

A partir de los resultados obtenidos del PCA, construimos un nuevo conjunto de datos para la implementación del algoritmo de k-medias. Este nuevo conjunto está compuesto por los *scores* obtenidos en cada PCA aplicado, el procedimiento se muestra en el *Listing 15*.

Listing 15. Construcción de conjunto de datos para implementación de algoritmo de k-medias

```
mydata=cbind(UPM_data$id_upm,acp1$scores,acp2$scores,acp3$scores,UPM_data$no_hacin
a,
             acp5$scores)
mydata=as.data.frame(mydata)
names( mydata )[ 1:20 ] <- c("id_upm","x1","x2","x3","x4","x5","x6","x7","x8","x9"
,"x10",
                             "x11","x12",
                             "x13","x14","x15","x16","x17","x18","x19")
```

Luego, seleccionamos el identificador de las UPMs *("id_upm")* y las variables que se requieren para implementar el algoritmo de k-medias. Se realiza un resumen de las variables mediante la función *summary()* para acceder a los principales estadígrafos. Este paso se realiza a través del código `R` que se muestra en el *Listing 16*.

Listing 16. Selección de variables para implementación de algoritmo de k-medias

```
mydata<-dplyr::select(mydata,id_upm,pca_1=x1,pca_2=x11,pca_3=x14,pca_5=x18)
summary(mydata)
```

Para efectos de la ilustración, solo se mostrará el procedimiento para la variante de "Hartigan-Wong", que mostró los mejores resultados. No obstante, instruiremos al lector cómo proceder para obtener los resultados bajo las otras variantes del algoritmo.

En el *Listing 17* se muestra el código `R` que permite la implementación del algoritmo k-medias versión "Hartigan-Wong" para k=3, 4 y 5. Cabe recordar, que primero debemos establecer una semilla mediante *set.seed()* para asegurar la reproducibilidad de los resultados. Para obtener resultados bajo otra versión del algoritmo, se debe modificar en la función *kmeans()* el argumento *algorithm* por alguna de las siguientes opciones*: "Forgy", "Lloyd"* o *"MacQueen"*. Para obtener resultados bajo la versión Jarque, como indicamos en la sección anterior, es necesario realizar una transformación en los datos usando un código como el que se muestra en el *Listing 13,* y luego, aplicar el algoritmo con la base transformada, usando la versión "MacQueen".

Listing 17. Algoritmo de Hartigan-Wong con k=3,4 y 5

```
set.seed(300)
clustPCA_HW3<-kmeans(mydata[,-1], centers=3,iter.max = 10,
                     algorithm = "Hartigan-Wong");clustPCA_HW3
clustPCA_HW4<-kmeans(mydata[,-1], centers=4,iter.max = 10,
                     algorithm = "Hartigan-Wong");clustPCA_HW4
clustPCA_HW5<-kmeans(mydata[,-1], centers=5,iter.max = 10,
                     algorithm = "Hartigan-Wong");clustPCA_HW5
```

La interpretación de los resultados es análoga a la comentada anteriormente en la estrategia 1. Cabe destacar, que para k=5 se logró la mejor bondad de ajuste (61.9%), que resulta moderada.

En consecuencia, se logra un mejor agrupamiento de las UPMs usando el algoritmo de "Hartigan-Wong", con k=5, bajo la estrategia 1.

Como se mencionó anteriormente, la "mejor clasificación" o más precisamente, estratificación socioeconómica de las UPMs se decide mediante criterios estadísticos que van más allá de las métricas que proporciona el algoritmo, por lo que es necesario conservar todos los resultados obtenidos. Por ello, añadimos los resultados de las estratificaciones obtenidas bajo la estrategia 2 a la matriz *estratos* generadas mediante el *Listing 12*. El *Listing 18* muestra cómo hacer esto.

Listing 18. Guardado de resultados

```
# Asignamos nombre a la columna donde se guardarán los resultados de clustPCA_HW3
names(estratos)[5]<-paste("PCA_HW",3,sep="")
# Llenamos la columna PCA_HW3 con el vector de grupos (estratos) de clustPCA_HW3
estratos[,5] <- clustPCA_HW3$cluster
# Asignamos nombre a la columna donde se guardarán los resultados de clustPCA_HW4
names(estratos)[6]<-paste("PCA_HW",4,sep="")
# Llenamos la columna PCA_HW4 con el vector de grupos (estratos) de clustPCA_HW4
estratos[,6] <- clustPCA_HW4$cluster
# Asignamos nombre a la columna donde se guardarán los resultados de clustPCA_HW5
names(estratos)[7]<-paste("PCA_HW",5,sep="")
# Llenamos la columna PCA_HW5 con el vector de grupos (estratos) de clustPCA_HW5
estratos[,7] <- clustPCA_HW5$cluster
```

6 Análisis de componentes principales y algoritmos de estratificación univariada

6.1 Análisis de componentes principales

De forma muy general podemos definir el análisis de componentes principales (PCA) como un tipo de transformación lineal aplicada a un conjunto de datos multivariantes (donde todas las variables son cuantitativas) habitualmente correlacionados entre sí, que permite expresar la información contenida en un menor número de variables o componentes principales no correlacionadas entre sí. El PCA 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. A partir del conjunto de datos *UPM_data se* realizaron ocho PCA, con diferentes combinaciones del set de indicadores (básicamente se inició con un gran conjunto de indicadores que se iban eliminando progresivamente del análisis según los resultados de la matriz de correlaciones o las bajas contribuciones a la primera componente principal). Como podemos apreciar en la Tabla V.2, en la mayoría de los casos, el porcentaje de varianza retenida por la primera componente principal (PC1) fue superior al 50%.

Tabla V.2.PCA implementados

PCA	ID Indicadores	% Var. Exp.PC1
PCA-1	*educ_sup, t_ocupa, t_desocupa, r_depe, r_mhij, no_hacina, acc_agua, indmat, q45_ingre*	46.68
PCA-2	*educ_sup, t_ocupa, r_depe, r_mhij, no_hacina, indmat, q45_ingre*	55.20
PCA-3	*educ_sup, t_ocupa, r_depe, r_mhij, indmat, q45_ingre*	62.63
PCA-4	*educ_sup, t_ocupa, r_depe, r_mhij, muro_viv, q45_ingre*	58.27
PCA-5	*educ_sup, t_ocupa, r_mhij, indmat, q45_ingre*	68.38
PCA-6	*educ_sup, t_ocupa, t_desocupa, r_depe, r_mhij, no_hacina, acc_agua, indmat*	42.84
PCA-7	*educ_sup, t_ocupa, r_depe, r_mhij, no_hacina, indmat*	51.43
PCA-8	*educ_sup, t_ocupa, r_mhij, indmat*	65.11

Fuente: Instituto Nacional de Estadísticas (INE).

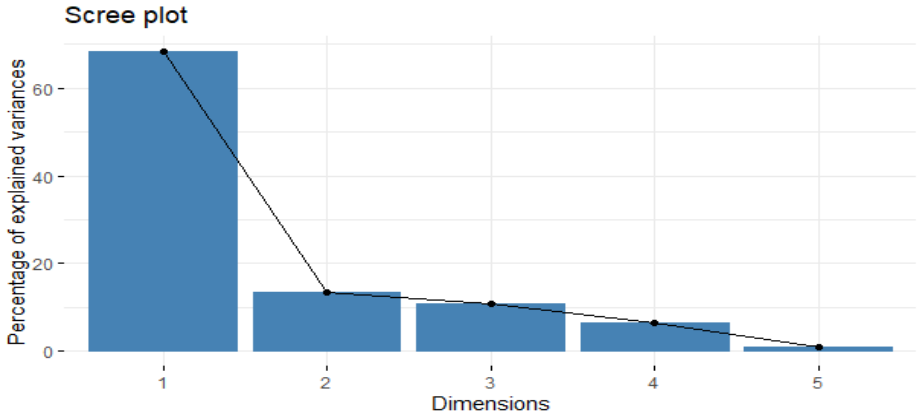
El *Listing 19* muestra el código para generar los PCA-5 y PCA-8. La función *prcomp()* de la librería *factoextra* (Mundt, 2020) implementa el PCA, se debe declarar el conjunto de datos para el análisis (en el ejemplo, solo difieren por la variable *q45_ingre*, que ocupa la posición 32 en nuestra BBDD) e indicar en el parámetro booleano *scale*, si se quiere o no, un escalamiento de los datos, es decir, una transformación sobre las variables, para que su desviación estándar sea uno. Adicionalmente, la función *fviz_pca_var,* permite la visualización del círculo de correlaciones sobre el primer plano factorial, indicando los porcentajes de varianza retenidos por cada componente.

Listing 19. PCA-5 y PCA-8, círculo de correlaciones

```
library(factoextra)
# PCA-5
res.PCA5 <- prcomp(UPM_data[,c(14,15,26,31,32)], scale = TRUE)
fviz_pca_var(res.PCA5, title= "PCA5: Círculo de correlaciones")

# PCA-8
res.PCA8 <- prcomp(UPM_data[,c(14,15,26,31)], scale = TRUE)
fviz_pca_var(res.PCA8, title= "PCA8: Círculo de correlaciones")
```

Gráfico I.5. PCA5: histograma de valores propios.



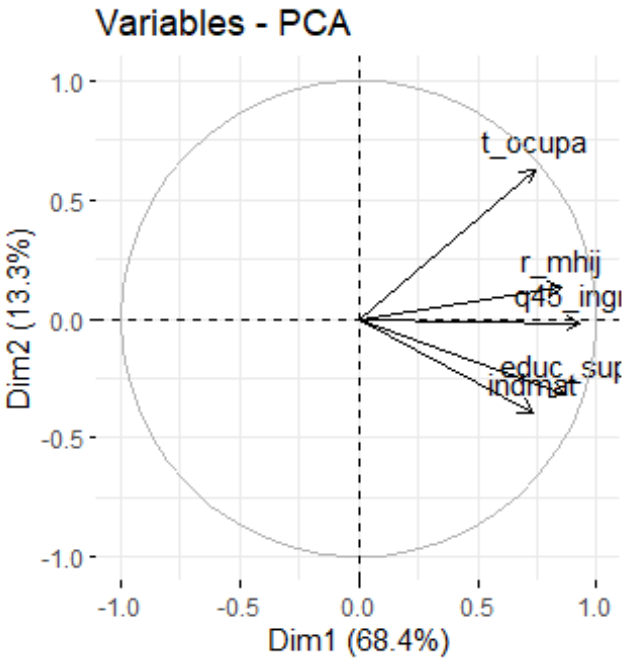
Fuente: Instituto Nacional de Estadísticas (INE).

En el Gráfico I.5, se observa el histograma de valores propios (también llamados valores característicos o raíces latentes) del PCA-5, donde asociado a cada componente se indica el porcentaje de varianza explicada de la misma, esta representación es una ayuda visual para determinar el número de componentes a seleccionar para los posteriores análisis. Se observa un decremento importante al pasar de la primera a la segunda componente, por tanto, una buena alternativa sería retener las dos primeras componentes.

Para algunos de los PCA implementados, todos los indicadores estaban correlacionados positivamente con la PC1, situándose de un solo lado del plano factorial, constituyendo de esta forma a la PC1 como un *factor tamaño* (Véase Gráfico I.6 y Gráfico I.7).

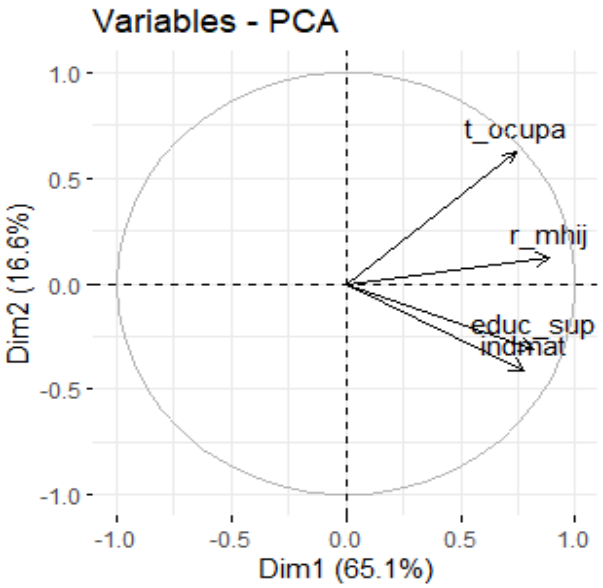
Los cinco primeros análisis incluyeron la variable *q45_ingre* y de estos, el PCA-5 retiene el mayor porcentaje de varianza en el primer plano factorial (81.7%, que se obtiene sumando la varianza retenida por la primera y segunda componente principal). Los últimos tres análisis no incluyen la variable relacionada con el ingreso, siendo el PCA-8 el que arroja el mejor resultado, 65.1% de varianza retenida en la PC1. Se destaca, analizando el círculo de correlaciones correspondiente al PCA-5, que la variable *q45_ingre* está altamente correlacionada con la PC1 (Ver Gráfico I.6) y, por tanto, se podría suprimir del análisis, ya que redunda sobre PC1, al ser también una combinación lineal de indicadores sociodemográficos de la base censal.

Gráfico I.6. PCA-5: círculo de correlaciones.



Fuente: Instituto Nacional de Estadísticas (INE).

Gráfico I.7. PCA-8: círculo de correlaciones.



Fuente: Instituto Nacional de Estadísticas (INE).

6.2 Clasificación a partir de la primera componente principal

Los PCA previamente realizados, se constituyen en un pretratamiento que entrega en cada análisis la variable latente PC1, la cual se convierte en el insumo para la aplicación de diversos algoritmos de clasificación univariada no supervisados. Es necesario entonces, guardar la primera componente de cada PCA realizado y posteriormente realizar un re-escalamiento, para asegurar que todos los valores de los vectores sobre los que se implementan los algoritmos de clasificación sean mayores a cero. En el *Listing 20* se muestra este proceso para el PCA-8:

Listing 20. Preparación de los datos del PCA-8

```
ind_8 <-get_pca_ind(res.PCA8)$coord[,1]
# Re-escalamiento
ind_8r <- (ind_8 - min(ind_8))+1
```

La metodología aplicada, probó k=3, 4, 5 particiones, a nivel nacional y regional, generando de esta forma, más de cien vectores de estratificación para las UPMs del MMV. A continuación, se listan los diferentes métodos utilizados para clusterizar y su respectiva implementación.

6.2.1 Cuantiles

Bajo este método la PC1 se divide en cuartiles (cuatro grupos) o terciles (tres grupos), concentrando cada uno el 25% o el 33% de las UPMs del MMV, respectivamente. Para la implementación en R, utilizamos para fines ilustrativos una partición en terciles (i=3) sobre la PC1 del PCA-8, que previamente ha sido re-escalada. Hay que recordar que, tal como en los ejercicios realizados con k-medias, debemos guardar todos los resultados de estratificación para su posterior evaluación, por lo que incluimos la instrucción para poblar la matriz *estratos* con el resultado de esta estratificación. Asimismo, procederemos para guardar los resultados obtenidos bajo los algoritmos que se presentan a continuación. En el *Listing 21* mostramos la implementación.

Listing 21. Partición en terciles.

```
library(Hmisc)
i=3
# Asignamos nombre a la columna donde se guardarán los resultados de cuantil3
names(estratos)[8]<- paste("Q3" ,i, sep="")
# Llenamos la columna cuantil3 con el vector de grupos (estratos) de cuantil3
estratos[,8] <- cut2(ind_8r, g=i)
```

6.2.2 Método de la raíz de la frecuencia acumulada

El método de la raíz de la frecuencia acumulada (Dalenius & Hodges, 1959) consiste en la formación de estratos de manera que la varianza obtenida de una variable cuantitativa (en este caso PC1) sea mínima para el estrato. Se basa en la acumulación de la raíz cuadrada de las frecuencias acumuladas de la PC1 sobre las UPMs. Es una técnica exacta y no requiere de algún procedimiento iterativo. La librería *stratification* (Baillargeon, 2017) contiene varias funciones para realizar estratificación univariada, que incluye este método y los que se verán a continuación.

En el *Listing 22* se muestra el código para su implementación usando la función *strata.cumrootf()*. Son especificados los parámetros, *Ls* (número de estratos), *CV* (coeficiente de variación de la muestra), además al especificar *alloc = (0.5, 0, 0)*, se asume que la asignación de la muestra dentro de los estratos es de tipo proporcional.

Listing 22. Dalenius-Hodges.

```
library(stratification)
i=3; cv= 0.1
# Asignamos nombre a la columna donde se guardarán los resultados de DH3
names(estratos)[9]<- paste("DH" ,i, sep="")
# Llenamos la columna cuantil3 con el vector de grupos (estratos) de DH3
estratos[,9]<- strata.cumrootf(ind_8r,
                               Ls=i,
                               alloc=c(0.5,0,0),
                               CV=cv)$stratumID
```

6.2.3 Estratificación óptima

La estratificación óptima (Lavallée & Hidiroglou, 1988) 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. Kozak (2004) desarrolló un método de estratificación basado en optimización numérica, que brinda límites de estratificación cercanos al óptimo cuando se cuenta con variables de estratificación con distribución asimétrica. La implementación se presenta en el *Listing 23,* los parámetros especificados para la función *strata.LH()* son los mismos que en el *Listing 22,* generando particiones para tres y cuatro estratos.

Listing 23. Lavallée-Hidiroglou.

```
for (i in c(3:4)){
  k=7+i
  # Asignamos nombre a la columna donde se guardarán los resultados de LH3
  names(estratos)[k]<- paste("LH" ,i, sep="")
  estratos[,k]<- strata.LH(ind_8r,
                           Ls=i,
                           alloc=c(0.5,0,0),CV=cv,
                           algo="Kozak")$stratumID
                }
```

6.2.4 Estratificación geométrica

Gunning & Horgan (2004) desarrollaron un método aproximado para estratificar poblaciones asimétricas. Esta estrategia recurre a una progresión geométrica para determinar los límites de los estratos, bajo el supuesto de obtener coeficientes de variación *(CV)* aproximadamente iguales en cada uno de los estratos. La misma se basa en una observación de Cochran (1961), según la cual, con límites de estratos cercanos a los óptimos, los coeficientes de variación resultan aproximadamente iguales en todos los estratos. La implementación se presenta en el *Listing 24,* y tal como en el caso anterior, los parámetros especificados para la función *strata.geo()* son los mismos que en el *Listing 22*.

Listing 24. Gunning-Horgan.

```
for (i in c(3:4)){
k=i+9
names(estratos)[k]<- paste("GH" ,i, sep="")
estratos[,k]<- strata.geo(ind_8r,
                          Ls=i,
                          alloc=c(0.5,0,0),
                          CV=cv)$stratumID}
```

7 Evaluación

Como se mencionó, el objetivo de la estratificación del marco muestral es su uso en los diseños muestrales para disminuir la incertidumbre de las estimaciones, por tanto, la metodología de evaluación se basó en la reducción de la varianza para un conjunto de indicadores a nivel de UPM calculados a partir de la base censal. Específicamente, la medida de calidad multivariada empleada para la evaluación fue el efecto de diseño generalizado (*G(S)*) propuesto por Jarque (1981).

Las técnicas de estratificación descritas en las secciones anteriores arrojaron como resultado 180 posibles estratificaciones, las que fueron evaluadas mediante *G(S)* calculado a partir del efecto de diseño (DEFF) de un conjunto de 24 indicadores que se encuentran descritos en la Tabla V.3.

Tabla V.3.Indicadores para evaluación

ID	Indicador / Título del recurso
ocupa_h	Proporción de hombres en la UPM ocupados.
ocupa_m	Proporción de mujeres en la UPM ocupadas.
desocupa_h	Proporción de hombres en la UPM desocupados.
desocupa_m	Proporción de mujeres en la UPM desocupadas.
inac_h	Proporción de hombres en la UPM inactivos.
inac_m	Proporción de mujeres en la UPM inactivas.
ftp_h	Proporción de hombres en la UPM en la fuerza de trabajo primaria con respecto a la PEA[4].
ftp_m	Proporción de mujeres en la UPM en la fuerza de trabajo primaria con respecto a la PEA.
esc_prim_h	Proporción de hombres en la UPM con educación primaria.
esc_prim_m	Proporción de mujeres en la UPM con educación primaria.
esc_sec_h	Proporción de hombres en la UPM con educación secundaria.
esc_sec_m	Proporción de mujeres en la UPM con educación secundaria.
esc_sup_h	Proporción de hombres en la UPM con educación terciaria (superior).
esc_sup_m	Proporción de mujeres en la UPM con educación terciaria (superior).
A_h	Proporción de hombres en la UPM que están ocupados en agricultura.
B_h	Proporción de hombres en la UPM que están ocupados en pesca.
F_h	Proporción de hombres en la UPM que están ocupados en construcción.
H_h	Proporción de hombres en la UPM que están ocupados en transporte y almacenamiento.
G_m	Proporción de mujeres en la UPM que están ocupadas en comercio.
I_m	Proporción de mujeres en la UPM que están ocupadas en alojamiento y servicios de comidas.
P_m	Proporción de mujeres en la UPM que están ocupadas en enseñanza.
Q_m	Proporción de mujeres en la UPM que están ocupadas en actividades de salud.
p_ext	Proporción de extranjeros en la UPM.
h_uni	Proporción de hogares en la UPM que son unipersonales.

Fuente: Instituto Nacional de Estadísticas (INE).

El procedimiento de evaluación de las estratificaciones está compuesto por los siguientes pasos:

- Cálculo de DEFF para cada indicador.
- Cálculo de DEFF generalizado *(G(S))*.

7.1 Cálculo de DEFF para cada indicador

Un elemento importante de comparación entre dos estrategias muestrales[5] 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. Una estrategia es mejor que otra si

[4] Población económicamente activa.
[5] Se denomina estrategia muestra a la combinación entre el diseño y el estimador.

la varianza del estimador es menor o análogamente su DEFF es menor (Bautista, 1988). De esta forma, el DEFF de una variable x_p, se define con la siguiente expresión:

$$DEFF_p = \frac{Var_{ST}(\overline{x}_p)}{Var_{SI}(\overline{x}_p)}$$

Donde, $Var_{ST}(\overline{x}_p)$ y $Var_{SI}(\overline{x}_p)$ denotan la varianza del diseño estratificado y la varianza de un muestreo aleatorio simple para la media poblacional (porcentaje) de la p-ésima variable de la matriz de información. Por otro lado, Gutiérrez (2016, pág. 184) demuestra que, cuando la asignación es proporcional, esta relación se puede expresar así:

$$DEFF_p = \frac{\sum_{h=1}^{H} W_h \, S^2_{x_{hp}}}{S^2_{x_p}} \qquad p = 1, \dots, P.$$

En donde, para cada estrato $h = 1, \dots, H$, se tiene que $S^2_{x_p}$ es la varianza de la variable x_p en la población y $S^2_{x_{hp}}$ es la varianza de la variable x_p supeditada al estrato h **y, por tanto, bajo este supuesto, podemos realizar los cálculos directamente sobre el set de indicadores de evaluación de la base censal.**

Para el cálculo de los DEFF, en primer lugar, debemos cargar el archivo de datos que contiene los 24 indicadores construidos para la evaluación, el cual se encuentran en formato. *RData*. Finalmente, este archivo debe combinarse con el conjunto de datos *estratos*, que contiene los vectores de estratificación. En el *Listing 25* se muestra cómo cargar y fundir estos dos conjuntos de datos.

Listing 25. Carga y Merge de archivos de datos necesarios para el cálculo de DEFF

```
# Carga de archivo con indicadores para calculo de DEFF
load("indi_eval.RData")
# Merge entre vectores de estratificacion y basa con indicadores de evaluacion
UPM_data2 <- merge(estratos,indi_eval, by.x="id_upm",by.y="id_upm",
                   all.x=TRUE,all.y=FALSE)
```

El procedimiento para el cálculo del DEFF es análogo para cada indicador y estratificación obtenida, y por tanto, para efectos de la ilustración, solo mostraremos el código `R` para el cálculo en el indicador de escolaridad primaria en hombres (*esc_prim_h*), en cada uno de los doce vectores de estratificación obtenidos de los pasos anteriores.

Listing 26. Cálculo de DEFF

```
deff <- dplyr::as_tibble(matrix(NA,ncol =12,nrow = 1))
for (i in 1:(length(names(estratos))-1)){
aux  <- UPM_data2 %>%
  group_by_at(names(estratos)[i]) %>%
  summarise(var=var(esc_prim_h),nh=n(),wh=nh/(nrow(UPM_data2)),
            var_pond=var*wh,deff=var_pond/var(UPM_data2$esc_prim_h))
  names(deff)[i]<-paste("DEFF_",names(estratos)[i],sep="")
deff[1,i] <- sum(aux$deff)
}
t(deff)[order( t(deff)),]
```

Finalmente, podemos ver en la Tabla V.4, que, para el indicador de escolaridad primaria en hombres, el mejor resultado en términos de minimización del efecto de diseño se obtiene cuando se particiona en cuatro estratos con el método de estratificación de Lavallée-Hidiroglou sobre la primera componente principal.

Tabla V.4.DEFF para escolaridad primera en hombres

LH4	Q3	DH3	LH3	HW4	GH4	HW5	PCA_HW5	PCA_HW4	PCA_HW3	HW3	GH3
0,338	0,380	0,387	0,399	0,420	0,468	0,486	0,508	0,556	0,567	0,574	0,620

Fuente: Instituto Nacional de Estadísticas (INE).

7.2 Cálculo de DEFF generalizado (*G(S)*)

Jarque (1981) propone la siguiente medida de calidad, definida como el efecto de diseño generalizado $G(S)$ sobre todas las variables que conforman el set de evaluación:

$$G(S) = \sum_{p=1}^{P} D\,EFF_p = \sum_{p=1}^{P} \frac{1}{S_{x_p}^2} \sum_{h=1}^{H} W_h\, S_{x_{hp}}^2$$

Ante una estratificación pertinente, se esperaría que $Var_{ST}(\bar{x}_p) < Var_{SI}(\bar{x}_p)$, por lo tanto, $0 < DEFF_p < 1$, lo que conlleva a que $0 < G(S) < P$. **Luego, se debería escoger el escenario para el cual $G(S)$ fuera mínimo**.

Para el cálculo del DEFF generalizado, se ejecuta el *Listing 26* para cada indicador de evaluación y, finalmente, se suma el resultado para cada vector de estratificación. En la Tabla V.5 se muestra el DEFF generalizado para los indicadores de escolaridad primaria, desocupación y fuerza de trabajo primaria en hombres. **Tanto si se consideran tres como cuatro estratos, el mejor método resulta ser Lavallée-Hidiroglou sobre la primera componente principal.**

Tabla V.5.DEFF Generalizado

Indicador	HW3	PCA_HW3	Q3	DH3	LH3	GH3	HW4	PCA_HW4	LH4	GH4
desocupa_h	0,92	0,858	0,9	0,89	0,89	0,97	0,91	0,88	0,89	0,92
esc_prim_h	0,57	0,567	0,38	0,39	0,4	0,62	0,42	0,556	0,34	0,47
ftp_h	0,87	0,75	0,8	0,75	0,71	0,94	0,85	0,708	0,67	0,85
G(S)	2,37	2,175	2,07	2,03	**2,01**	2,52	2,17	2,144	**1,9**	2,24

Fuente: Instituto Nacional de Estadísticas (INE).

7.3 Caracterización de los grupos

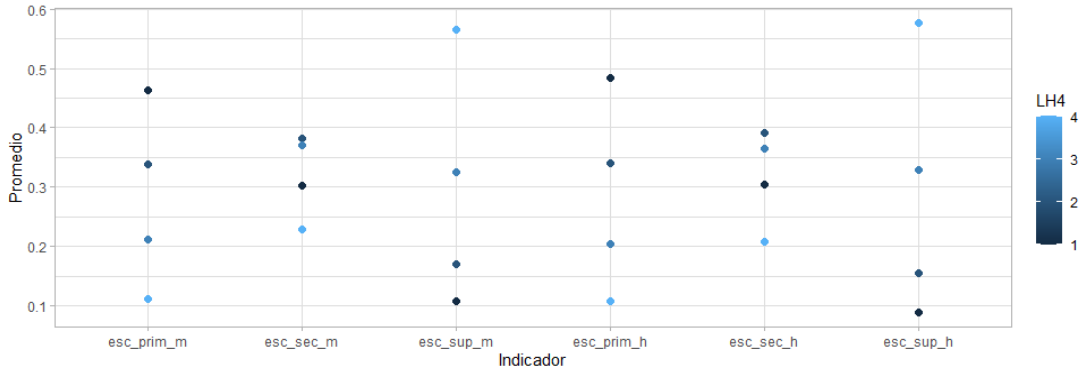
Una vez realizada la clasificación de las UPMs es una buena práctica explorar las características de los grupos en términos de variables (indicadores) de interés, que ayuden a caracterizarlos. Para efectos de la ilustración, se presentará la caracterización de los grupos a través de los resultados reportados por el método de Lavallée-Hidiroglou sobre la primera componente principal, para k=4. En primer lugar, se mostrarán gráficamente las medias de las variables relacionadas con escolaridad según estrato socioeconómico. En el *Listing 27* se muestra el código `R` para calcular la medida de resumen y, posteriormente, realizar el gráfico de medias por estrato.

Listing 27. Representación de medias de variables por grupo

```
resumen = UPM_data2[c(1,11,14:37)]%>%
    group_by(LH4) %>%
    summarise_at(vars(-id_upm), function(x) mean(x,na.rm = TRUE))

library(ggplot2)
data= melt(resumen[c(1,12:17)], id="LH4", variable.name = "Indicador", value.name
= "Promedio")
# Gráfico con medias de indicadores por grupo
data$LH4 <- as.numeric(data$LH4)
ggplot(data, aes(x=Indicador, y=Promedio, color=LH4)) +
  geom_point(size=2) + theme_light ()
```

Gráfico I.8.Resumen de estratificación LH4.



Fuente: Instituto Nacional de Estadísticas (INE).

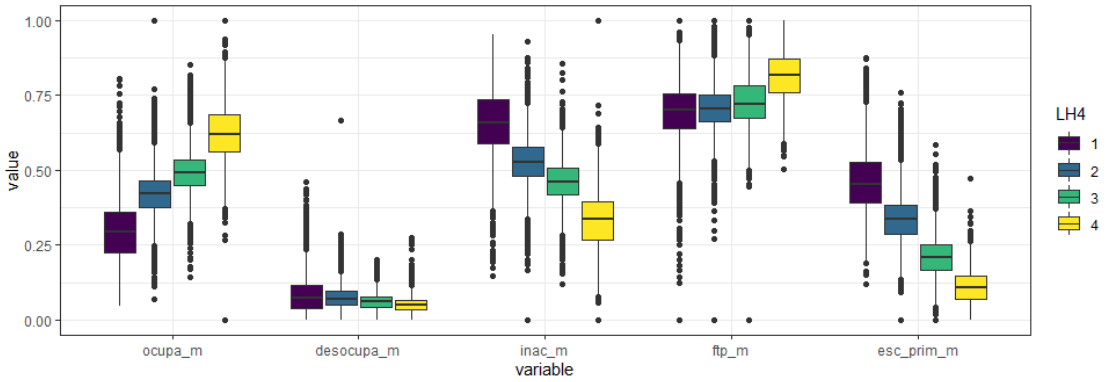
En el Gráfico I.8 se puede apreciar que los indicadores relacionados con proporción de personas con estudios primarios y superiores, tanto en hombres como en mujeres, son los que mejor discriminan entre los distintos grupos socioeconómicos.

Adicionalmente, es posible realizar gráficos para cada variable o indicador de evaluación, representando los valores individuales de las UPMs. En el *Listing 28* se muestra el código `R` que permite realizar un diagrama de caja (*boxplot*) para algunos indicadores.

Listing 28. Representación de UPM por grupo para cada indicador socioeconómico

```
library(ggplot2)
library(viridis)
ggplot(data=melt(UPM_data2[c(11,14:17,24)]), aes(x = variable, y = value)) +
    geom_boxplot(aes(fill = LH4))+scale_fill_viridis(discrete=TRUE) +
    theme_bw()
```

Gráfico I.9.Boxplots para estratificación LH4.



Fuente: Instituto Nacional de Estadísticas (INE).

En el Gráfico I.9 se puede apreciar que los resultados son los esperados, ya que, por ejemplo, el indicador de proporción de mujeres ocupadas y el indicador de mujeres inactivas, muestran un ordenamiento según grupo socioeconómico.

8 Discusión

El INE culminó satisfactoriamente la estratificación del MMV 2017. Durante este proceso fueron considerados 180 escenarios de estratificación junto con varias metodologías apropiadas y acordes, tanto a la experiencia internacional, como a la literatura especializada. **Es importante destacar que la estratificación del MMV 2017 tiene por objetivo contribuir a los equipos responsables de las estrategias de muestreo, en la consecución de diseños de muestreo que induzcan una mayor eficiencia y precisión a la hora de la estimación de las estadísticas oficiales.**

En este trabajo, se ha descrito en detalle la metodología aplicada por el INE y su implementación mediante procedimientos computacionales en el *software* estadístico `R`, que es un sistema altamente extensible para computación estadística y gráficos. Como pudimos ver, la librería *stratification* contiene funciones especializadas para estratificación univariante, como la regla de Dalenius & Hodge, la regla geométrica de Gunning & Horgan y la estratificación óptima de Lavallée & Hidiroglou. Precisamente, bajo el criterio *G(S)* resultó seleccionada la estratificación óptima de la PC1, escogiéndose tres estratos (LH3) por sobre cuatro. Esta decisión se basó en el análisis de las distribuciones de las UPMs para cada combinación de estrato, comuna y área. Según esto, dada la baja prevalencia de UPMs en los estratos extremos, principalmente en el área rural, resulta más conveniente para fines de los diseños muestrales considerar **tres estratos socioeconómicos.**

La estratificación óptima, consideró 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. Teniendo en cuenta que la matriz de información estuvo en escala de porcentajes (siguiendo el enfoque sugerido en la definición de una matriz de bienestar sobre las UPMs), la variabilidad recogida por la medida de resumen (PC1) resultó alta, informativa y sintetizó la información adecuadamente.

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[6]) 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.

9 Referencias

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

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

Chavent M. & Kuentz V. & Liquet B. & Saracco J. (2017). ClustOfVar: *Clustering of Variables*. R package version 1.1. https://CRAN.R-project.org/package=ClustOfVar

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.

[6] 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.

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 - 169. https://doi.org/10.2307/2346387.

Kassambara A. & Mundt F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.

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.

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.

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.

R Core Team. (2019). A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Rand, W (1971). Objective criteria for the evaluation of clustering methods. *Journal of the* American Statistical Association, 66 (336), 846-850.

Rivest L. & Baillargeon S. (2017). *Stratification: Univariate Stratification of Survey Populations*. R package version 2.2-6. https://CRAN.R-project.org/package=stratification