UNIVERSIDAD NACIONAL DE ROSARIO

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

CARRERA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

Tema: Diseño de una Muestra Maestra para la provincia de Santa Fe. MUMSAFE

Autora: Leonor María Maldonado.

Director: Gonzalo Marí

Fecha: abril 2026

Resumen: Esta tesis evalúa los distintos componentes que forman parte del diseño muestral a partir del cual se construirá la Muestra Maestra de la provincia de Santa Fe (MUMSAFE), para determinar aquellos aspectos que brinden estimaciones más eficientes y eficaces. Las oficinas de estadística utilizan entre sus operativos, las encuestas a hogares para realizar mediciones de indicadores específicos de las problemáticas públicas. Las encuestas se apoyan en muestras que son seleccionadas mediante procedimientos probabilísticos para asegurar la representatividad de las mismas respecto a la población, y requieren un marco de muestreo compuesto por el conjunto de elementos que identifican, localizan y acceden a cada unidad de la población. Una opción para construir un marco o Muestra Maestra es seleccionar una muestra grande en una primera fase inicial que sirva de marco para seleccionar sucesivas muestras. Las encuestas por muestreo a menudo utilizan la estratificación y los conglomerados para mejorar la eficiencia estadística; además el muestreo basado en conglomerados puede mejorar cuestiones operativas y de recursos. El Instituto Provincial de Estadística y Censos (IPEC) tiene como objetivo la realización de una Muestra Maestra en la provincia de Santa Fe, la cual estará basada en el diseño de la Muestra Maestra Urbana de Viviendas de la República Argentina (MMUVRA) desarrollada por el Instituto Nacional de Estadística y Censos (INDEC) entre otros. En base a los datos disponibles del Censo Nacional de Población, Hogares y Viviendas 2010, se comparan los resultados considerando en cada dominio de estimación geográfica establecido un diseño con una etapa de muestreo; en esta etapa se evalúan dos componentes. Por un lado, se compara la construcción de conglomerados de viviendas urbanas considerando: unidades de distintos tamaños, de distintas formas de agrupamiento, y con distintas probabilidades de selección. Por otro lado, se compara la estratificación de los conglomerados construidos, considerando distintas variables de clasificación con distintos métodos de estratificación, y distintas cantidades de estratos. Para comparar los diseños planteados, con los datos disponibles, se tienen en cuenta los estimadores de una encuesta que surja de la Muestra Maestra, considerando entonces una segunda etapa de muestreo, donde se seleccionan hogares por un diseño con igual probabilidad dentro de cada unidad primaria de muestreo (UPM) seleccionada. Para evaluar la eficiencia se toman como medidas de precisión el efecto de diseño y el coeficiente de variación, de los estimadores relacionados con el estado ocupacional de las personas; asociados a la principal encuesta socioeconómica planteada aquí dentro del programa de encuestas sugerido. Para evaluar la eficacia se tienen en cuenta medidas de cercanía dentro de los conglomerados e igualdad de tamaño entre ellos. El diseño de la Muestra Maestra quedará constituido, en una etapa, por el conjunto de viviendas de los conglomerados seleccionados.

Finalmente, evaluando los componentes planteados se determina para la provincia de Santa Fe, que el diseño de Muestra Maestra más eficaz resulta de tomar UPM con 2 radios censales agrupadas mediante el algoritmo de Elliott. A su vez el diseño resultante considera formar estratos mediante el algoritmo de Lavallée-Hidiroglou, para estratificar poblaciones asimétricas de dichas unidades primarias con el nivel educativo del jefe/jefa de hogar; y seleccionadas con probabilidad proporcional al tamaño (PPT) de la vivienda habitable.

PALABRAS CLAVE: MUESTRA MAESTRA, DISEÑO MUESTRAL COMPLEJO, EFECTO DE DISEÑO, COEFICIENTE DE VARIACIÓN.

Contenido

1 Introducción 1
2 Objetivo 4
2.1 General 4
2.2 Específicos 4
3 Metodología 5
3.1 Diseño de muestras en varias etapas 5
3.1.1 Muestreo en dos etapas 6
3.1.2 MSA en la primera y segunda etapa. 7
3.1.3 PPT en la primera etapa y MSA en la segunda etapa. 8
3.1.4 Estimadores no lineales en muestreo en dos etapas 10
3.2 Construcción de Conglomerados 14
3.2.1 Agrupamiento geoespacial de igual tamaño. Método de Elliot 15
3.3 Estratificación de conglomerados 18
3.3.1 Variables de estratificación 19
3.3.2 Componentes Principales Robustas y Bootstrap Rápido y Robusto 19
3.3.3 Estratificación de poblaciones asimétricas 25
4 Materiales 30
5 Resultados 30
5.1 Conglomerados 32
5.1.1 Formación de Conglomerados 32
5.1.2 Evaluación de los Conglomerados 36
5.1.3 Elección de Conglomerado 48
5.2 Estratificación de UPM 48
5.2.1 Formación de estratos 49
5.2.2 Evaluación de Estratos 57
5.2.3 Elección de Estratos 65
6 Conclusiones 66
Bibliografía 68
A. Anexo 71
A.1. Selección de Variables para utilizar el algoritmo de Elliott 71
A.2. Coeficientes de Variación según estratificación de conglomerados (de 5.2.2) 74
A.3. Procedimientos computacionales empleados 108

1 Introducción

Las oficinas de estadística utilizan entre sus operativos, las encuestas a hogares para realizar mediciones de indicadores específicos de las problemáticas públicas, con el fin de favorecer la toma de decisiones políticas basadas en evidencia para mejorar la calidad de vida humana en general. Las encuestas se apoyan en muestras que son seleccionadas mediante procedimientos probabilísticos, los cuales aseguran la representatividad de las mismas respecto a la población, brindando además medidas de precisión de los datos que permiten evaluar la calidad de la información relevada y así producir estadísticas confiables.

Las encuestas por muestreo a menudo utilizan la estratificación de unidades para mejorar la eficiencia estadística. El muestreo estratificado es una técnica de muestreo probabilístico para reducir la variancia, que divide o bien aprovecha las agrupaciones conocidas, como una partición de la población en grupos mutuamente excluyentes y homogéneos intragrupo, denominados estratos. Dentro de cada estrato y de manera independiente entre estratos, se selecciona una muestra probabilística. El tamaño de la muestra total se distribuye entre los distintos estratos mediante un criterio de asignación: de igual tamaño, proporcional u óptimo, entre otros. Kish (1975) considera que el muestreo estratificado es una de las formas más efectivas de controlar la variación en la muestra y asegurar la representación adecuada de subpoblaciones importantes (por ejemplo, por edad, género o región).

El muestreo de conglomerados por otro lado, se utiliza para aumentar la eficacia, mejorando cuestiones operativas y económicas, mediante la formación de agrupaciones de unidades preferentemente heterogéneas intragrupo y contiguas. Esto permite disminuir los costos operativos que surgen cuando se requiere la recopilación de datos en áreas grandes mediante encuestas personales, permitiendo a los encuestadores, por ejemplo, reducir los viajes para recolectar una cantidad de encuestas en un área determinada (Elliott M. R., 2011). El muestreo por conglomerado aleatorio simple sin reemplazo en muchas situaciones resulta ineficiente, especialmente si los conglomerados son homogéneos y/o de tamaños desiguales (Särndal, Swensson, & Wretman, 1992). Por lo tanto, se prefiere determinar aquellos conglomerados con mayor eficiencia, es decir heterogéneos y de igual tamaño con respecto a la cantidad de unidades de muestreo.

Para seleccionar una muestra por procedimientos aleatorios se requiere un marco de muestreo, compuesto por el conjunto de elementos que sirvan para identificar, localizar y acceder a cada una de las unidades de la población. El marco debe ser actual y completo, esto es, debe incluir sin omisión ni repetición a todos los elementos de la población. Una opción para la construcción de un marco es la de seleccionar una muestra grande en una primera fase inicial que posteriormente sirva de marco para seleccionar sucesivas muestras. Kish (1975) denomina a este instrumento marco maestro (*master frame*). El mismo incluye listados de elementos físicos y procedimientos de muestreo que, a su vez, permiten obtener muestras de unidades elementales, sin el esfuerzo de haberlas listado a todas, para abastecer múltiples encuestas a lo largo de varios años sin que haya necesidad de entrevistar a los mismos informadores repetidas veces. Las actualizaciones del marco maestro requieren menos esfuerzo que las actualizaciones del marco muestral convencional, como lo es un Censo.

Dicho concepto se aplicó por primera vez en la Muestra Maestra de Agricultura en Estados Unidos (King & Jessen, 1945), una muestra de tamaño extremadamente grande, aplicándose luego en distintos ámbitos de la estadística agropecuaria. El primer antecedente en Argentina de elaboración de un marco de muestreo nacional para encuestas a hogares data de fines de la década del 60 (Zabala & Librandi, 2018). Dicho marco

fue elaborado con el propósito de extraer periódicamente muestras de viviendas para la Encuesta Nacional de Salud, un proyecto conjunto del Gobierno Nacional, la Organización Panamericana de la Salud y la Asociación de Facultades de Medicina. La encuesta se realizó con cobertura nacional urbana y rural desde 1970 a 1972. Luego de esa experiencia, INDEC desarrolló, a partir de 1992, el Marco de Muestreo Nacional Urbano (MMNU) para encuestas a hogares (INDEC, 1999) con el propósito de extraer muestras de viviendas para encuestas con alcance nacional urbano, destinadas a aglomerados de 5000 o más habitantes según el Censo Nacional de Población y Viviendas de 1991 (Censo 1991). A partir del Censo Nacional de Población, Hogares y Viviendas del año 2001 (Censo 2001), se realizó una ampliación del MMNU para posibilitar la extracción de muestras para encuestas con mayor cobertura geográfica, incluyendo a localidades de 2000 y más habitantes, según la información del Censo 1991 y algunas actualizaciones realizadas en el Censo 2001. En 2011, y en base a los datos del Censo Nacional de Población, Hogares y Viviendas 2010 (Censo 2010), se desarrolló la MMUVRA, la cual es usada actualmente para seleccionar muestras para todas las encuestas a hogares que lleva a cabo el INDEC según Documento Interno Metodológico (INDEC, 2012). En la actualidad, se encuentra desarrollando una nueva Muestra Maestra tomando como base el Censo Nacional de Población, Hogares y Viviendas que se llevó a cabo durante el año 2022.

El diseño muestral de la MMUVRA considera una muestra probabilística, polietápica y estratificada de viviendas. La provincia es la primera variable de estratificación. La selección de la muestra se realiza en todas las provincias con un diseño bietápico. En la primera etapa se muestrea, considerando un diseño con PPT y estratificado, UPM definidas como aglomeraciones[1] urbanas con 2000 y más habitantes, residiendo en viviendas particulares al momento del Censo 2010. Luego, dentro de las UPM muestreadas, se seleccionan unidades secundarias de muestreo (USM) definidas a partir de radios o conjuntos de radios censales[2]. Los procedimientos de selección aplicados en esta segunda etapa son los de un muestreo PPT y estratificado. La probabilidad de selección asignada a las unidades de muestreo en cada etapa es proporcional al total de viviendas particulares ocupadas o en construcción según el Censo 2010. Por último, se listan todas las viviendas de las USM muestreadas de cada UPM seleccionada (INDEC, 2019).

El IPEC, basado en el diseño de muestras maestras desarrolladas por INDEC y otras oficinas nacionales de estadística, se propone llevar a cabo una MUMSAFE, con el objetivo de seleccionar muestras probabilísticas para encuestas a hogares en el período intercensal. Entre las características principales de una Muestra Maestra se pueden mencionar, en general:

- Diseño de una muestra grande que evita la situación de agotamiento de viviendas, y permite que no se solapen, ni repitan las viviendas seleccionadas de una encuesta a otra, o en varias rondas de la misma encuesta.
- Automatización de la selección de las muestras, lo que reduce la programación ad-hoc de la selección muestral para cada encuesta.
- Integración del proceso de listado, con la formación de un único marco muestral maestro.

[1] Se entiende por aglomeración a una localidad o conjunto de localidades que por continuidad de edificaciones y calles configuran una misma unidad urbana.

[2] Los radios censales son unidades menores definidas para la asignación de las cargas de trabajo de censistas; en las zonas urbanas generalmente se definen como un conjunto de manzanas o sectores con alrededor de 300 viviendas.

- Reducción del esfuerzo y del costo al momento de la actualización del marco muestral maestro.
- Programación de un Plan de Encuestas, conformado por Encuestas propuestas para relevar datos de interés mediante muestras extraídas de la Muestra Maestra durante el período de vigencia de la misma; asegurando la selección de las muestras requeridas y con posibilidad de agregar Encuestas afines al marco, no incluidas en el Plan.
- Unificación de la cartografía de la Muestra Maestra para todos los operativos del Plan de Encuestas.

Para MUMSAFE, en particular:

- Consideración de múltiples dominios de estimación:
    - Población urbana provincial.
    - Ciudad de Rosario.
    - Ciudad de Santa Fe.
    - Ciudad de Rafaela.
    - Ciudad de Venado Tuerto.
    - Ciudad de Reconquista.
- Construcción del Plan de Encuestas con sus principales indicadores y población objetivo:
    - Encuesta socio-económica por principales ciudades: sus principales indicadores son la tasa de actividad, la tasa de empleo, según jefes o jefas de hogar, sexo o edad; y la tasa de desocupación por sexo. Población objetivo: personas que habiten radios urbanos de 14 años de edad o más que viven en alguna de las 5 principales ciudades de la provincia.
    - Encuesta de Seguridad: o Encuesta Provincial de Victimización. Indicadores principales: tasa de hogares victimizados y porcentaje de hogares victimizados que realizan la denuncia formal. Población objetivo: personas que habiten radios urbanos de localidades con 18 años de edad o más, de la provincia de Santa Fe.
    - Otra encuesta a hogares: según demanda de alguna de las principales ciudades consideradas con indicadores referidos a la población residente en hogares, o bien estudiar características puntuales de los hogares de la población urbana provincial.

En esta tesis, se plantean pruebas de diseño de una Muestra Maestra en una etapa de muestreo: primero se considera evaluar diseños por conglomerados con distintos métodos de selección y distintos tamaños de unidades de muestreo basados en radios censales, y luego se consideran distintos métodos de estratificación con distintas variables, variando la cantidad de estratos. Se pretende encontrar el diseño más conveniente entre los propuestos, en base a la precisión como medida de eficiencia y a la eficacia evaluada a partir de cuestiones operativas como la cercanía entre las unidades, y así aportar metodología y recomendar el diseño de Muestra Maestra más efectivo para la futura ejecución de la MUMSAFE.

La estructura de esta tesis está compuesta en Capítulos: el Capítulo 1 contiene una introducción a la temática, el Capítulo 2 está compuesto por los objetivos general y específicos, en el Capítulo 3 se presentan los métodos aplicados, en el Capítulo 4 se presentan los materiales utilizados, y en el Capítulo 5 los resultados encontrados, consecuentemente en el Capítulo 6 se concluye con la elección del diseño más eficaz. Por último, se presenta la bibliografía consultada y se anexan algunas Figuras, Tablas y Gráficos mencionados en los resultados.

2 Objetivo

2.1 General

El objetivo de esta tesis es establecer lineamientos metodológicos para la conformación de MUMSAFE, que permitirá la selección de las muestras de viviendas de las encuestas a hogares a realizarse en la provincia, durante el período intercensal.

2.2 Específicos

- Elaborar una reseña en base al estudio de las bases metodológicas para la construcción de una Muestra Maestra.
- Plantear distintos métodos de construcción de UPM por conglomerados mediante la agrupación de áreas censales, considerando criterios de tamaño, proximidad geográfica y selección muestral.
- Comparar los métodos de agrupación y elegir el que brinde estimaciones más eficientes en términos de eficacia operativa y eficiencia de las estimaciones de los indicadores de interés.
- Determinar alternativas de estratificación considerando diferentes criterios de formación y cantidad de estratos, contrastando su impacto sobre la precisión de las estimaciones.
- Proponer una configuración de UPM que permita establecer lineamientos para la conformación efectiva de la MUMSAFE, considerando conjuntamente criterios de eficiencia y eficacia.

3 Metodología

El diseño de la muestra tiene impacto directo en las estimaciones que se construyen en función a los valores de las variables obtenidos una vez aplicada la encuesta a las unidades seleccionadas y los elementos que componen el diseño utilizado. Ya que su precisión varía conforme se apliquen diferentes diseños, se realizan pruebas planteando diseños muestrales variando los distintos componentes del mismo para evaluar como impactan en la precisión de los estimadores y, teniendo en cuenta su factibilidad, obtener conclusiones que permitan determinar el diseño más efectivo. En esta Sección se presenta la metodología que se utiliza para lograr dicho propósito.

Principalmente, se plantean varios escenarios de diseños de la Muestra Maestra en una etapa de muestreo. Para evaluar los mismos se comparan los índices de precisión de estimadores que simulen surgir de una encuesta que haya sido seleccionada de la Muestra Maestra, sobre los datos del Censo 2010. Entonces al considerar una segunda etapa de muestreo se presenta en la Sección 3.1 las medidas de precisión de los estimadores resultantes de un muestreo en dos etapas: se plantean distintos métodos de selección en la primera etapa, primero considerando un MSA, y luego un muestreo con PPT. Se tienen en cuenta estimadores lineales como el total y no lineales como la razón, basado en Valliant et. al. (2018). Para obtener las medidas de precisión para éstos últimos, se presenta la técnica de Linealización por series de Taylor según Särndal et. al. (1992).

Además, se plantea en principio la construcción de las unidades de muestreo de la primera etapa como conglomerados de viviendas, y su posterior estratificación. Para la formación de conglomerados, en la Sección 3.2 se presenta un algoritmo de agrupamiento espacial de unidades cercanas que permite la construcción de conglomerados de tamaños aproximadamente iguales, basados en la similitud de un conjunto de variables que las caracterizan, propuesto por Elliott (2011).

En la Sección 3.3 se presentan distintas técnicas para estratificar los conglomerados construidos. Se presenta la selección de variables de estratificación por Naciones Unidas (2010) y como insumo para la estratificación se presenta la estimación de Componentes Principales (CP) robustos utilizando el método de Bootstrap Rápido y Robusto (FRB, por sus siglas en inglés *Fast and Robust Bootstrap*) para definir las cargas significativas y obtener una CP que resuma la información de muchas variables, presentado por Salibian-Barrera et. al. (2006). Por último, se presenta el algoritmo de estratificación para poblaciones asimétricas, propuesto por Lavallée-Hidiroglou (1988) mediante el método iterativo de Sethi, que busca establecer los límites de los estratos que minimizan el tamaño de la muestra, dado un valor del coeficiente de variación.

3.1 Diseño de muestras en varias etapas

Es frecuente aplicar en encuestas a hogares un muestreo utilizando más de una etapa. El uso de muestras seleccionadas en varias etapas a menudo puede ser una solución práctica y económica en situaciones en las que no se dispone de una lista de unidades elementales (o analíticas) para el muestreo directo de las mismas, o aun existiendo dicha lista, la selección y concreción de una muestra traería aparejado un aumento de los costos operativos por la gran dispersión geográfica de la misma.

Para diseñar una muestra en varias etapas se deben considerar las contribuciones de estas a la variancia de un estimador. Estas componentes de la variancia generalmente dependen de la variable de análisis en particular, de la naturaleza del diseño muestral en general y también de la forma del estimador. A continuación, en la Sección 3.1.1 se presentan algunos conceptos básicos para el cálculo de la variancia en un muestreo de dos etapas, en la Sección 3.1.2 se presenta el caso de un muestreo simple al azar sin reemplazo (MSA) en ambas etapas, en la Sección 3.1.3 se presenta el caso de un muestreo de selección PPT en la primera etapa y MSA en la segunda, considerando estimadores lineales en todos los casos, mientras en la Sección 3.1.4 se consideran los diseños de muestreo anteriores con estimadores no lineales, por lo que se incluye en la Sección 3.1.4.1 la aproximación Lineal por Series de Taylor como técnica para aproximar su variancia.

3.1.1 Muestreo en dos etapas

Se considera un diseño de muestra de dos etapas en el que las unidades de la primera etapa se denominan UPM. Se toma en la segunda etapa la selección de elementos mediante MSA sin reemplazo dentro de las UPM seleccionadas. Las fórmulas cuando se seleccionan conglomerados en la segunda etapa son similares. Se define:

U = población de UPM

M = número de UPM en la población

U_i = población de elementos en la UPM i

N_i = número de elementos en la UPM i

$N = \sum_{i \in U} N_i$ es el número total de elementos en la población

π_i = probabilidad de inclusión de la UPM i

π_{ij} = probabilidad de inclusión conjunta de las UPMs i y j

m = tamaño de muestra de UPMs

n_i = tamaño de muestra de elementos en la UPM i

s = conjunto muestra de UPMs

s_i = conjunto de muestra de elementos en la UPM i

y_{ik} = variable de análisis para el elemento k en la UPM i

$\bar{y}_U$ = media de la variable Y en la población

$\bar{y}_{U_i}$ = media de la variable Y en la UPM i

El estimador π del total en la población, $t_U = \sum_{i \in U} \sum_{k \in U_i} y_{ik}$, de una variable de análisis Y viene dado por:

$$\hat{t}_\pi = \sum_{i \epsilon s} \frac{\hat{t}_i}{\pi_i} \tag{3.1}$$

donde $\hat{t}_i = (N_i/n_i)\sum_{k\epsilon s_i} y_{ik}$, es el estimador del total de la UPM i, t_i, bajo MSA. La variancia del estimador del total se puede expresar con dos componentes, una vinculada a la selección de las UPM y la otra relacionada a la selección de las unidades secundarias de muestreo (USM) dentro de las UPM. Las mismas varían dependiendo del método de selección en cada etapa de muestreo. Se plantea a continuación, dos alternativas de selección de UPM: considerar primero una MSA y luego PPT sin reemplazo, fijando la selección MSA en la segunda etapa de muestreo. El cálculo de variancia del primer caso resulta más sencillo con respecto al segundo.

3.1.2 MSA en la primera y segunda etapa.

Se considera una MSA en la primera etapa de m de M UPMs y una MSA en la segunda etapa de n_i elementos seleccionados de la población de N_i elementos en la UPM i. El π estimador para el total de la variable Y para este diseño viene dado por

$$\hat{t}_\pi = \frac{M}{m}\sum_{i\epsilon s}\frac{N_i}{n_i}\sum_{k\epsilon s_i} y_{ik}$$

y la variancia es igual a

$$V(\hat{t}_\pi) = \frac{M^2}{m}\frac{M-m}{M}S_{U1}^2 + \frac{M}{m}\sum_{i\epsilon U}\frac{N_i^2}{n_i}\frac{N_i-n_i}{N_i}S_{U2i}^2 \qquad (3.2)$$

donde $S_{U1}^2 = \frac{\sum_{i\epsilon U}(t_i-\bar{t}_U)^2}{M-1}$ siendo t_i el total poblacional de Y en la UPM i, $\bar{t}_U = \sum_{i\epsilon U} t_i\,/M$ la media de los totales de las UPMs y $S_{U2i}^2 = \sum_{k\epsilon U_i}\left(y_{ik} - \bar{y}_{U_i}\right)^2/(N_i-1))$ la variancia unitaria de Y entre los elementos de la UPM i. La variancia relativa de $\hat{t}_\pi$, $V(\hat{t}_\pi)/t_U^2$, viene dada por

$$V(\hat{t}_\pi)\Big/_{t_U^2} = \frac{1}{m}\frac{M-m}{M}B^2 + \frac{1}{t_U^2}\frac{M}{m}\sum_{i\epsilon U}\frac{N_i^2}{n_i}\frac{N_i-n_i}{N_i}S_{U2i}^2 \qquad (3.3)$$

donde $B^2 = S_{U1}^2/\bar{t}_U^2 = M^2S_{U1}^2/\,t_U^2$ es la variancia relativa unitaria entre los totales de las UPMs.

Si se seleccionan $\bar{n}$ elementos en cada UPM y las fracciones de muestreo de las UPMs y de los elementos dentro de las UPMs son pequeñas, la variancia relativa puede expresarse de manera simplificada como

$$V(\hat{t}_\pi)\Big/_{t_U^2} = \frac{B^2}{m} + \frac{W^2}{m\bar{n}} \qquad (3.4)$$

con $W^2 = M\sum_{i\epsilon U} N_i^2S_{U2i}^2/\,t_U^2$ la componente de variancia dentro de las UPMs[3].

[3] La función *BW2stageSRS* del paquete *PracTools* (Valliant, Dever, & Kreuter, 2022) del programa *R* considera la ecuación (3.4) como versión aproximada de la variancia relativa del estimador del total.

Si todas las UPMs tienen el mismo tamaño, $\bar{N}$, y se seleccionan $\bar{n}$ elementos en cada una, la fracción de muestreo de la segunda etapa es $\bar{n}/\bar{N}$. Esto implica que la muestra es autoponderada, o sea, $\pi_i\pi_{k|i} = m\bar{n}/M\bar{N}$. La variancia relativa expresada en la ecuación (3.3) se simplifica a

$$V(\hat{t}_\pi)\Big/_{t_U^2} = \frac{1}{m}\frac{M-m}{M}B^2 + \frac{1}{m\bar{n}}\frac{\bar{N}-\bar{n}}{\bar{N}}W^2$$

con $W^2 = (1/M\bar{y}_U^2)\sum_{i\epsilon U} S_{U2i}^2$.

Suponiendo que se seleccionan $\bar{n}$ elementos en cada UPM seleccionada, y que m/M y $\bar{n}/N_i$ son pequeños, la forma más general de la variancia relativa, $V(\hat{t}_\pi)/t_U^2$ en la ecuación (3.3), también se puede escribir, de manera aproximada, en términos de una medida de homogeneidad δ:

$$V(\hat{t}_\pi)\Big/_{t_U^2} \cong \frac{\tilde{V}}{m\bar{n}}k[1 + \delta(\bar{n}-1)] \tag{3.5}$$

donde $\tilde{V} = \frac{S_U^2}{\bar{y}_U^2}$, con $S_U^2 = \frac{\sum_{i\in U}\sum_{k\in U_i}(y_{ik}-\bar{y}_U)^2}{N-1}$, $k = \frac{(B^2+W^2)}{\tilde{V}}$, y

$$\delta = B^2/(B^2+W^2) \tag{3.6}$$

La ecuación (3.5) es útil para calcular el tamaño de la muestra, ya que el número de UPMs de la muestra y las unidades de muestra en cada UPM están explícitamente indicados en la fórmula. En el caso que $N_i = \bar{N}$ y tanto M como $\bar{N}$ sean grandes, se puede expresar la variancia relativa de la variable en la población como la suma de las variancias relativas entre y dentro de las UPM, o sea,

$$S_U^2/\bar{y}_U^2 = \sum_{i\in U}\sum_{k\in U_i}(y_{ik}-\bar{y}_U)^2/\left(\bar{y}_U^2(N-1)\right) \cong B^2 + W^2 \tag{3.7}$$

Si $k = 1$, la ecuación (3.5) queda expresada de la manera usual que aparece en la bibliografía clásica de muestreo (ver, por ejemplo, Deming, 1950; Kish, 1975; Cochran, 1977). Sin embargo, cuando el número de elementos por UPM N_i varía, conduce a diferencias en los totales t_i, a un valor de B^2 grande, y el valor de k puede ser muy superior a 1 (ver ejemplo 9.2 de Valliant et.al, 2018). En esos casos, es conveniente utilizar la ecuación (3.5) con una estimación real de k para determinar los tamaños de muestra y calcular estimaciones anticipadas de los coeficientes de variación. La ecuación (3.5) también relaciona la variancia del estimador del total con la variancia que se obtendría de una muestra aleatoria simple, ya que $\tilde{V}/m\bar{n}$ es la variancia relativa del total estimado en una MSA de tamaño $m\bar{n}$ cuando la fracción de muestreo es pequeña. El producto $k[1 + \delta(\bar{n}-1)]$ es un tipo de efecto de diseño. La ecuación (3.5) con $k = 1$ generalmente se utiliza sin considerar el método de selección de las muestras de UPMs y de los elementos dentro de las UPM, ni tampoco tiene en cuenta el tipo de estimador que se utilice. Cabe notar que esta fórmula no es la que corresponde cuando se utilizan métodos de muestreo distintos al MSA en diferentes etapas.

3.1.3 PPT en la primera etapa y MSA en la segunda etapa.

Se considera un diseño de muestra de dos etapas en el que m de M unidades de la primera etapa se seleccionan utilizando un muestreo PPT sin reemplazo; y en la segunda etapa la selección de n_i de N_i elementos mediante MSA sin reemplazo. La variancia de diseño del estimador del total se escribe como la suma de dos componentes

$$V(\hat{t}_\pi) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\frac{t_i}{\pi_i}\frac{t_j}{\pi_j} + \sum_{i\in U}\frac{N_i^2}{\pi_i n_i}(1 - \frac{n_i}{N_i})\,S_{U2i}^2$$

dónde $S_{U2i}^2 = \sum_{k\in U_i}\left(y_{ik} - \bar{y}_{U_i}\right)^2/(N_i - 1)$ es la variancia de Y entre los elementos de la UPM i.

Uno de los inconvenientes que trae aparejada dicha fórmula de variancia para el cálculo del tamaño de la muestra de las UPM es que el mismo no se presenta de manera explícita, resultando complejo a la hora de determinar los tamaños de muestra correspondientes a cada una de las etapas. Además, es común considerar un diseño donde el cálculo de la variancia sea menos complicado que el que realmente se utiliza.

Entonces, se considera la selección de m de M UPM con reemplazo con distinta probabilidad, y la muestra dentro de cada UPM es seleccionada por MSA, seleccionando n_i de N_i elementos. Si bien es cierto que los diseños con reemplazo no se utilizan con frecuencia en la práctica para evitar que las mismas unidades sean seleccionadas más de una vez en la muestra, también lo es que dichos diseños poseen fórmulas de variancia simples, y bajo ciertas condiciones, estas variancias pueden servir de aproximaciones para los diseños sin reemplazo. Asumiendo el diseño con reemplazo en la primera etapa, el estimador pwr del total viene dado por:

$$\hat{t}_{pwr} = \frac{1}{m}\sum_{i\in s}\hat{t}_i/p_i$$

donde $\hat{t}_i = \frac{N_i}{n_i}\sum_{k\in s_i} y_{ik}$ es el total estimado de la UPM i de una muestra aleatoria simple y p_i es la probabilidad de selección de la UPM i. La variancia de $\hat{t}_{pwr}$ viene dada por:

$$V(\hat{t}_{pwr}) = \frac{1}{m}\sum_{i\in U} p_i\left(\frac{t_i}{p_i} - t_U\right)^2 + \sum_{i\in U}\frac{N_i^2}{m p_i n_i}(1 - \frac{n_i}{N_i})S_{U2i}^2.$$

Haciendo la misma suposición anterior que $\bar{n}$ elementos se seleccionan en cada UPM, la variancia se reduce a:

$$V(\hat{t}_{pwr}) = \frac{S_{U1(pwr)}^2}{m} + \frac{1}{m\bar{n}}\sum_{i\in U}\left(1 - \frac{\bar{n}}{N_i}\right)\frac{N_i^2 S_{U2i}^2}{p_i} \tag{3.8}$$

donde $S_{U1(pwr)}^2 = \sum_{i\in U} p_i\left(\frac{t_i}{p_i} - t_U\right)^2$. Dividiendo por t_U^2 y asumiendo que la fracción de muestreo dentro de la UPM i, $\bar{n}/N_i$, es despreciable, se obtiene la variancia relativa aproximada de $\hat{t}_{pwr}$:

$$V(\hat{t}_{pwr})/t_U^2 \cong \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}}k[1 + \delta(\bar{n} - 1)] \tag{3.9}$$

con $\tilde{V}(\hat{t}_{pwr}) = \frac{S_U^2}{\bar{y}_U^2}, k = (B^2 + W^2)/\tilde{V}$, y

$$B^2 = {S_{U1(pwr)}^2}\Big/{t_U^2}, \tag{3.10}$$

$$W^2 = \frac{1}{t_U^2}\sum_{i\in U} N_i^2\frac{S_{U2i}^2}{p_i} \tag{3.11}$$

$$\delta = {B^2}\Big/{(B^2 + W^2).} \tag{3.12}$$

La expresión (3.9) tiene la misma forma que la ecuación (3.5) pero con diferente definición de B^2 y W^2. La expresión (3.9) también tiene la interpretación de una variancia relativa MSA, $\tilde{V}/m\bar{n}$, multiplicada por un efecto de diseño, $k[1 + \delta(\bar{n} - 1)]$ como en la ecuación (3.5). Por lo tanto, se utiliza la ecuación (3.5) y la (3.9) para comparar la precisión de ambos diseños respectivamente, mediante el efecto de diseño $k[1 + \delta(\bar{n} - 1)]$.

3.1.4 Estimadores no lineales en muestreo en dos etapas

En las encuestas muestrales de diseños tanto simples como complejos, a menudo es deseable o necesario emplear estimadores que sean no lineales en las observaciones. Las razones, diferencias de razones, coeficientes de correlación, coeficientes de regresión y medias post estratificadas son ejemplos comunes de tales estimadores. Las expresiones exactas para las variancias de muestreo de estimadores no lineales no existen y, por lo tanto, tampoco estimadores simples e insesgados de la variancia.

Un método útil para aproximar la variancia de un estimador no lineal es aproximar el estimador no lineal por una función lineal de las observaciones mediante el método de Series de Taylor. Es decir, los métodos de Series de Taylor producen una aproximación lineal de la estadística muestral de interés. Luego, se requiere aplicar las fórmulas de variancias, como las presentadas anteriormente en este Capítulo, sobre una función lineal, para estimar la variancia de la aproximación lineal.

En el apartado 3.1.4.1 se presenta el concepto básico de la Linealización por Series de Taylor para aproximar variancia, para luego en el apartado 3.1.4.2 para el diseño muestral específico, aplicar las fórmulas de variancia apropiadas a la aproximación lineal para un estimador de razón.

3.1.4.1 Linealización por Series de Taylor para calcular variancia

La idea básica de la linealización por series de Taylor es aproximar la función no lineal del estimador por una función lineal alrededor de un punto. Esta aproximación permite usar las fórmulas tradicionales de cálculo de variancia de funciones lineales presentadas en secciones anteriores.

Se plantea estimar un parámetro θ poblacional que puede ser expresado como una función de q totales poblacionales, $t_1, \ldots, t_q$,

$$\theta = f(t_1, \ldots, t_q)$$

donde $t_j = \sum_U y_{jk}$ $(j = 1, \ldots, q)$. Se asume que el vector $(y_{1k}, \ldots, y_{jk}, \ldots, y_{qk})'$ puede ser observado para $k \in s$, permitiendo formar estimadores π,

$$\hat{t}_{j\pi} = \sum_s \check{y}_{jk} = \sum_s y_{jk}/\pi_k \; ; \; j = 1, \ldots, q$$

donde $\check{y}_{jk} = y_{jk}/\pi_k$. Por el principio mencionado por Sarndal et. al. (1992, pág. 163) cuando el estimador θ es una función de totales, la estimación se obtiene reemplazando los totales de la función por sus estimadores π, entonces el estimador de θ resulta:

$$\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{q\pi}). \tag{3.13}$$

Analizando las propiedades de $\hat{\theta}$, cuando f es lineal, es decir $\theta = a_0 + \sum_{j=1}^{q} a_j t_j$, con a_j constantes en $j = 0, \ldots, q$, el estimador según la ecuación (3.13) resulta entonces

$$\hat{\theta} = a_0 + \sum_{j=1}^{q} a_j \hat{t}_{j\pi} \tag{3.14}$$

el cuál es insesgado para θ, con variancia

$$V(\hat{\theta}) = V\left(\sum_{j=1}^{q} a_j \hat{t}_{j\pi}\right) = \sum_{j=1}^{q} \sum_{j'=1}^{q} a_j a_{j'} C(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) \tag{3.15}$$

donde las $C(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum \sum_U \Delta_{kl} \breve{y}_{jk} \breve{y}_{j'l}$ son las covariancias entre los totales $\hat{t}_{j\pi}$ y $\hat{t}_{j'\pi}$, $\Delta_{kl} = C(I_k, I_l)$ las covariancias de las funciones indicadoras I_k e I_l, donde $I_g = 1$ cuando el elemento g pertenece a la muestra e $I_g = 0$ en otro caso, para $g = k, l$. La variancia de $\hat{\theta}$ expresada en la ecuación (3.15) se estima, usando las covariancias estimadas

$$\hat{V}(\hat{\theta}) = \sum_{j=1}^{q} \sum_{j'=1}^{q} a_j a_{j'} \hat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) \tag{3.16}$$

con $\hat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \sum \sum_s \check{\Delta}_{kl} \breve{y}_{jk} \breve{y}_{j'l}$ donde $\check{\Delta}_{kl} = \Delta_{kl}/\pi_{kl}$, y cuando $j' = j, \hat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi}) = \hat{V}(\hat{t}_{j\pi})$.

Si f es una función lineal, la variancia dada por la ecuación (3.15), como su estimador indicado en la ecuación (3.16), tienen expresiones alternativas que ofrecen ventajas al momento de calcularlas. Al tomar el estimador de θ e invertir el orden de las sumatorias, entre las unidades de selección y_{jk} y los totales $\hat{t}_{j\pi}$ de la expresión lineal (3.14), recurso propuesto por Woodruff (1971), quedaría:

$$\hat{\theta} = a_0 + \sum_{j=1}^{q} a_j(\hat{t}_{j\pi}) = a_0 + \sum_{j=1}^{q} a_j \left(\sum_s \frac{y_{jk}}{\pi_k} \right) = a_0 + \sum_s \frac{\sum_{j=1}^{q} a_j \; y_{jk}}{\pi_k}$$
$$= a_0 + \sum_s \breve{u}_k$$

donde $\breve{u}_k = u_k/\pi_k$, con $u_k = \sum_{j=1}^{q} a_j y_{jk}$. Entonces la ecuación (3.15) puede expresarse como

$$V(\hat{\theta}) = V(\sum_s \breve{u}_k) = \sum \sum_U \Delta_{kl} \breve{u}_k \breve{u}_l, \tag{3.17}$$

la cual se estima mediante

$$\hat{V}(\hat{\theta}) = V(\sum_s \breve{u}_k) = \sum \sum_U \check{\Delta}_{kl} \breve{u}_k \breve{u}_l. \tag{3.18}$$

El cálculo de $\hat{V}(\hat{\theta})$ según la ecuación (3.18) requiere el cálculo previo de u_k para $k \in s$, una operación bastante simple. Una vez que se dispone de los u_k, se requiere una sola operación de suma doble de la estimación deseada, ecuación (3.18). Ésta resulta computacionalmente más simple que usar la ecuación (3.16), que requiere el cálculo de $q(q+1)/2$ variancias y covariancias estimadas $\hat{C}(\hat{t}_{j\pi}, \hat{t}_{j'\pi})$.

El enfoque en este apartado está puesto cuando $\theta = f(t_1, \ldots, t_q)$ es una función no lineal de los q totales. En este caso, a menudo es imposible obtener resultados exactos sobre el sesgo y la variancia de $\hat{\theta} = f(\hat{t}_{1\pi}, \ldots, \hat{t}_{q\pi})$. Aplicar la técnica de Linealización por Series de Taylor proporciona una expresión aproximada para la variancia de $\hat{\theta}$. Esta técnica aproxima el estimador no lineal $\hat{\theta}$ por la estadística linealizada, denotada con $\hat{\theta}_0$, que es función lineal de $\hat{t}_{1\pi}, \ldots, \hat{t}_{q\pi}$, por lo que es fácil de operar. Se utiliza a su vez, la $V(\hat{\theta}_0)$ como la aproximación de $V(\hat{\theta})$, siendo sencillo encontrar una estimación de la variancia $\hat{V}(\hat{\theta})$.

La técnica para encontrar $\hat{\theta}_0$ consiste en aplicar la aproximación al término de primer orden del desarrollo por Series de Taylor de la función f, expandida alrededor del punto $(t_1, \ldots, t_q)$, y despreciando el término residual. Se obtiene

$$\hat{\theta} \doteq \hat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j \left(\hat{t}_{j\pi} - t_j\right) \tag{3.19}$$

donde las a_j se definen como:

$$a_j = \left.\frac{\partial f}{\partial \hat{t}_{j\pi}}\right|_{(\hat{t}_{1\pi},\ldots,\hat{t}_{q\pi})=(t_1,\ldots,t_q)}. \tag{3.20}$$

Para muestras grandes, cuando $\hat{t}_{1\pi}, \ldots, \hat{t}_{q\pi}$ tienen una alta probabilidad de tomar valores cercanos a $t_1, \ldots, t_q$, el estimador $\hat{\theta}$ funcionará aproximadamente como la variable aleatoria $\hat{\theta}_0$. Entonces, se utiliza la variancia exacta de la estadística linealizada $\hat{\theta}_0$ como aproximación a la del estimador más complejo (no lineal) $\hat{\theta}$. Para facilitar el cálculo, se considera el recurso de Woodruff (1971) ya mencionado. Se fija

$$u_k = \sum_{j=1}^{q} a_j y_{jk} \tag{3.21}$$

y $\breve{u}_k = u_k/\pi_k$. Entonces, como las ecuaciones (3.15) y (3.17) son equivalentes, la variancia aproximada de $\hat{\theta}$ se obtiene a partir de

$$V(\hat{\theta}) \doteq V(\hat{\theta}_0) = V\left(\sum_{j=1}^{q} a_j\, \hat{t}_{j\pi}\right) = V\left(\sum_s \breve{u}_k\right) = \sum\sum_U \Delta_{kl}\breve{u}_k\breve{u}_l \tag{3.22}$$

donde la aproximación a la variancia de $\hat{\theta}$ o su estimación se obtiene de $V(\sum_s \breve{u}_k)$ usando la ecuación (3.17) o (3.18) respectivamente, considerando en cualquier caso la expresión de la variancia correspondiente al diseño muestral utilizando la estadística linealizada u_k definida en (3.21).

3.1.4.2 Estimador de una razón

Se considera un estimador no lineal como la razón $\hat{\theta}$, función de dos totales estimados, $\hat{t}_{1\pi}$ y $\hat{t}_{2\pi}$ debido a que es uno de los estimadores que se va a utilizar en esta tesis:

$$\hat{\theta} = f(\hat{t}_{1\pi}, \hat{t}_{2\pi}) = \hat{t}_{1\pi}/\hat{t}_{2\pi}$$

donde $\hat{t}_{j\pi} = \sum_{i\in s}\sum_{k\in s_i} y_{jk}/\pi_k$ (j = 1,2), y π_k es la probabilidad de inclusión del elemento k.

Mediante el desarrollo por Series de Taylor, se puede escribir la aproximación lineal de $\hat{\theta}$ como

$$\hat{\theta} - \theta \doteq \sum_{j=1}^{q} a_j \left(\hat{t}_{j\pi} - t_j\right)$$

donde θ es la razón poblacional a estimar. Se obtiene u_k usando (3.20) y (3.21):

$$u_k = \sum_{j=1}^{2} a_j y_{jk} \qquad \text{con}$$

$$a_1 = \left.\frac{\partial \hat{\theta}}{\partial \hat{t}_{1\pi}}\right|_{(\hat{t}_{1\pi},\hat{t}_{2\pi})=(t_1,t_2)} = \left.\frac{1}{\hat{t}_{2\pi}}\right|_{(\hat{t}_{1\pi},\hat{t}_{2\pi})=(t_1,t_2)} = \frac{1}{t_2}$$

$$a_2 = \left.\frac{\partial \hat{\theta}}{\partial \hat{t}_{2\pi}}\right|_{(\hat{t}_{1\pi},\hat{t}_{2\pi})=(t_1,t_2)} = -\left.\frac{\hat{t}_{1\pi}}{\hat{t}_{2\pi}^2}\right|_{(\hat{t}_{1\pi},\hat{t}_{2\pi})=(t_1,t_2)} = -\frac{t_1}{t_2^2}.$$

Luego,

$$u_k = (y_{1k} - \theta y_{2k})/t_2 \tag{3.23}$$

El término u_k se denomina estadística linealizada. La variancia de $\hat{\theta}$ se puede aproximar calculando la variancia de $\sum_{i\in s}\sum_{k\in s_i} u_k/\pi_k$.

En el caso de MSA en ambas etapas, $\frac{1}{\pi_k} = \frac{M}{m}\frac{N_i}{n_i}$. La razón se puede aproximar como

$$\hat{\theta} \doteq \frac{M}{m}\sum_{i\in s} \hat{t}_{ui}.$$

donde $\hat{t}_{ui} = N_i/n_i \sum_{k\in s_i} u_{ik}$, con u_{ik} valor que toma la variable linealizada para el elemento k en la UPM i. Por tanto, la variancia aproximada de la razón de $\hat{\theta}$ se puede expresar de la misma forma que la variancia del $\hat{t}_\pi$, estimador π del total de la variable linealizada u de la ecuación (3.1). En consecuencia, la variancia de $\hat{\theta}$ se puede aproximar utilizando la ecuación (3.2), asumiendo que $n_i = \bar{n}$:

$$V(\hat{\theta}) \doteq V(\hat{t}_\pi) = \frac{M^2}{m}\frac{M-m}{M}S_{U1}^2 + \frac{M}{m}\sum_{i\in U}\frac{N_i^2}{n_i}\frac{N_i - n_i}{N_i}S_{U2i}^2$$

donde los componentes de variancia S_{U1}^2 y S_{U2i}^2 se escriben en términos de u_{ik} en lugar de y_{ik}. Específicamente:

$$S_{U1}^2 = \frac{\sum_{i\in U}(t_{ui} - \bar{t}_{Uu})^2}{M-1}$$

con $t_{ui} = \sum_{k\in U_i} u_{ik}$, y $\bar{t}_{Uu} = \frac{\sum_{i\in U} t_{ui}}{M}$

$$S_{U2i}^2 = \frac{\sum_{k\in U_i}(u_{ik} - \bar{u}_{U_i})^2}{N_i - 1}$$

con $\bar{u}_{U_i} = \sum_{k\in U_i} u_{ik}/N_i$.

Análogamente, en el caso de muestreo con PPT en la primera etapa y MSA en la segunda, se puede utilizar la ecuación (3.8) en términos del total estimado de la estadística linealizada y suponiendo que se seleccionan $\bar{n}$ elementos en cada UPM, la variancia aproximada de $\hat{\theta}_{pwr}$ se puede expresar:

$$V(\hat{\theta}_{pwr}) \doteq \frac{S^2_{U1(pwr)}}{m} + \frac{1}{m\bar{n}} \sum_{i \in U} (1 - \frac{\bar{n}}{N_i}) \frac{N_i^2 S^2_{U2i}}{p_i} \tag{3.24}$$

donde $S^2_{U1(pwr)} = \sum_{i \in U} p_i (\frac{t_{ui}}{p_i} - t_{U_u})^2$.

Otros estimadores no lineales pueden resolverse con este mismo método. Por ejemplo, una media estimada como $\hat{\bar{y}} = \hat{t}_\pi \,/\, \hat{N}_\pi$ o una razón de odds en una tabla de 2×2 se pueden linealizar y escribir, aproximadamente, como un total estimado de sus estadísticas linealizadas (Valliant, Dever, & Kreuter, 2018).

3.2 Construcción de Conglomerados

El muestreo de conglomerados se utiliza para aumentar la eficacia, mejorando cuestiones operativas y económicas, mediante la formación de agrupaciones de unidades heterogéneas intragrupo y preferentemente contiguas. Esto permite disminuir los costos operativos que surgen cuando se requiere la recopilación de datos en áreas grandes mediante encuestas personales, permitiendo a los encuestadores, por ejemplo, reducir los viajes para recolectar una cantidad de encuestas en un área determinada (Elliott M. R., 2011). El muestreo por conglomerado aleatorio simple sin reemplazo en muchas situaciones resulta ineficiente, especialmente si los conglomerados son homogéneos y/o de tamaños desiguales (Särndal, Swensson, & Wretman, 1992). Por lo tanto, se prefiere seleccionar aquellos conglomerados con mayor eficiencia, es decir heterogéneos y de igual tamaño con respecto a la cantidad de unidades de muestreo. Cabe destacar que también resulta favorable aplicar dichos criterios dentro del plan de la Muestra Maestra que considera seleccionar sucesivas muestras sobre unidades primarias con tamaños similares, ya que éstas se reducirán homogéneamente, evitando el reemplazo de unidades primarias pequeñas por agotamiento y controlando el presupuesto a la hora de actualizar el listado de las unidades primarias excesivamente grandes.

En ocasiones los conglomerados suelen derivarse de entidades físicas o políticas ya existentes, como sectores censales. Cuando es posible construir conglomerados como parte del diseño muestral, se abre la oportunidad de incorporar las características deseadas al diseño (de manera similar a la estratificación tradicional) buscando en este caso construir conglomerados heterogéneos de igual tamaño que contengan unidades cercanas entre sí. Esto permite usar el conocimiento sobre la estructura de la variabilidad en la población para reducir la misma en la muestra. Existen muchas técnicas disponibles para encontrar estructuras simples en conjuntos de datos complejos. Estos incluyen:

- Enfoques algorítmicos no paramétricos como dendrogramas o árboles de clasificación y regresión (LaVarnway, 1988).
- Procedimientos de agregación a través de algoritmos de k-medias (MacQueen, 1967).
- Métodos estadísticos paramétricos como los modelos mixtos Gaussianos (MacLachen & Peel, 2000).

Sin embargo, se encuentra en menor medida en la literatura de muestreo información sobre la construcción de tales agrupaciones bajo las características buscadas de contigüidad e igual tamaño; por ejemplo Cantwell (1990) describe el método de la Oficina del Censo de EE.UU. para construir conglomerados con características iguales (ECC por sus siglas en inglés) en el que la variancia entre conglomerados se reduce ordenando los valores observados de la covariable X vinculada con la variable de interés Y. Las unidades de muestreo con los valores de X más grande y más pequeño se unen, luego se juntan las unidades con el segundo valor más grande y el segundo más pequeño, y así sucesivamente; este proceso se repite $M-1$ veces, para formar conglomerados de tamaño 2^M. Este procedimiento disminuye la variancia entre grupos B^2 y da valores pequeños de $\delta = B^2/(B^2+W^2)$, la correlación intra-grupo que determina la eficiencia estadística del diseño. Dado que los valores más grandes de δ se asocian con una peor eficiencia, el método ECC debería mejorar la eficiencia, aunque en la práctica las ganancias parecen ser muy limitadas (Cantwell, 1991). Por otra parte, Elliott (2011) propone un algoritmo que agrupa unidades bajo las condiciones de igual tamaño, contigüidad geográfica y maximizando (o minimizando) la variancia dentro de los grupos para las variables de interés según se formen conglomerados o estratos respectivamente, el cual se presenta a continuación. Cabe aclarar que para la formación de estratos con la propuesta de Elliott no se considerará el requisito de contigüidad geográfica, al carecer de sentido.

3.2.1 Agrupamiento geoespacial de igual tamaño. Método de Elliot

El algoritmo desarrollado por Elliott (2011) es un procedimiento iterativo que construye grupos aproximadamente del mismo tamaño probando con cada una de las unidades a agrupar como semillas iniciales, alrededor de las cuales se unen las unidades más cercanas hasta obtener el tamaño adecuado. El algoritmo prueba diferentes puntos de inicio, algunos cercanos entre sí y otros más alejados, para generar distintas configuraciones. Esto permite que los conglomerados finales no sean siempre iguales ni demasiado similares, y se pueda explorar diferentes posibilidades de agrupamiento sin tener que revisar todas las combinaciones posibles (lo cual sería computacionalmente muy costoso). Se elige un conjunto de conglomerados óptimo en función del análisis de la variancia sobre las variables de interés teniendo en cuenta como factor las agrupaciones, los criterios de mínima distancia total entre unidades para todos los conglomerados y mínimo coeficiente de variación de los tamaños de los mismos. El procedimiento[4] está desarrollado bajo el programa *R*.

El algoritmo requiere el posicionamiento geoespacial de las unidades dado por latitud y longitud, una medida de tamaño y variables con información de interés de las unidades para armar los grupos. Sea K el número de conglomerados deseados de un conjunto de r unidades geográficas conocidas (radios, segmentos, etc.) con un tamaño total de N (población, viviendas, etc.). El algoritmo consta de las siguientes etapas:

1) Ordenar las unidades geográficas de 1 al r (cualquier método de ordenación es aceptable).
2) Calcular una matriz de distancia $r * r$ que contenga en la celda (i, j) la distancia entre las unidades i y j usando puntos de latitud y longitud del centroide de las unidades, a partir de la distancia euclídea bidimensional (se define más adelante).

[4] Código obtenido a partir de una comunicación con el autor.

3) Formar el primer grupo: Utilizar la primera unidad de la lista ordenada obtenida en 1), como semilla para construir el primer grupo. Agregar las unidades más cercanas a la unidad semilla teniendo en cuenta la matriz de distancia calculada en 2) hasta que el tamaño del grupo sea mayor al criterio de mínimo tamaño por conglomerado $\lfloor N/K - N/2r \rfloor$, donde $\lfloor x \rfloor$ es la parte entera del número x.
4) Formar el segundo grupo candidato: Elegir la unidad semilla para el segundo grupo llamado candidato como la siguiente unidad más cercana al primer grupo semilla que no estaba incluida previamente en el mismo. Agregar unidades que aún no estén asignadas al primer grupo a este segundo grupo en el orden de su distancia desde la segunda unidad semilla, hasta que el tamaño de este segundo grupo sea mayor que $\lfloor N/K - N/2r) \rfloor$
5) Repetir el segundo grupo candidato con cada unidad que no esté en el primer grupo semilla: repetir 4), esta vez elegir como la unidad semilla para el segundo grupo candidato a la segunda unidad más cercana al primer grupo semilla, luego tomando la semilla como la tercera unidad más cercana, y así sucesivamente, a través de todas las unidades no incluidas en el primer grupo semilla, generando tantos segundos grupos candidatos, como unidades que no están incluidas previamente a un grupo.

   Seleccionar un grupo candidato: Para cada uno de todos los pares de grupos formados entre el grupo semilla y el grupo candidato realizar un análisis de variancia, utilizando los dos grupos como factores y las variables de interés como explicativas, tratando a los grupos restantes como un tercer grupo residual. Seleccionar el segundo grupo candidato con la unidad semilla que produzca el valor $\hat{\sigma}^2$ máximo o mínimo si se va a maximizar o minimizar la variancia dentro del agrupamiento, según se quiera formar un conglomerado o un estrato, respectivamente. Se selecciona el de mayor (menor) medida de heterogeneidad para construir el grupo como conglomerado (estrato). Se define la Suma de Cuadrados Intra-grupo o residual $SCW \cong \hat{\sigma}^2$, como: $SCW = \sum_{k=1}^{K} \frac{x_{..}}{x_{.k}} \sum_{j=1}^{N_k} \frac{x_{jk}}{x_{.k}} \left(\frac{x_{.k}}{x_{jk}} y_{jk} - y_{.k} \right)^2$, donde N_k es el número de unidades en el k-ésimo grupo; x_{jk} es la medida de tamaño de la j-ésima unidad en el k-ésimo grupo, $x_{.k} = \sum_{j=1}^{N_k} x_{jk}$; $x_{..} = \sum_{k=1}^{K} \sum_{j=1}^{N_k} x_{jk}$; y_{jk} es el valor observado de la variable para la cual se evalúa la heterogeneidad de la j-ésima unidad en el k-ésimo grupo, e $y_{.k} = \sum_{j=1}^{N_k} y_{jk}$. En el caso de utilizar más de una variable informativa para formar grupos: $y_{1jk}, y_{2jk}, \ldots, y_{ijk}, \ldots y_{Ijk}$ se toma como medida de evaluación a $\hat{\sigma}^2 = \frac{1}{r-K} \sum_i w_i \hat{\sigma}_i^2$, como la combinación de los residuales de los modelos de cada variable y_{ijk},. $\hat{\sigma}_i^2$, y su ponderación w_i(por defecto es 1).
6) Formar el tercer grupo hasta el K-ésimo grupo: Repetir los pasos 4) y 5) para los grupos tercero a $K - 1$. Las unidades no asignadas se convierten en el K-ésimo grupo, quedando así formado el primer agrupamiento candidato o conjunto de grupos, es decir todas las unidades pertenecen a un grupo, donde la primera unidad es la semilla del primer grupo.
7) Repetir el agrupamiento con cada unidad dada, como semilla del primer grupo: Repetir los pasos 3) a 6), comenzando con la segunda unidad en el ordenamiento definido en 1) como semilla para la construcción del primer grupo, luego se toma la tercera unidad ordenada como semilla para la construcción del primer grupo, y así sucesivamente a lo largo de todo el proceso, con las

r unidades. Consecuentemente, se crean r agrupamientos candidatos, cada uno corresponde a cada unidad tomada como semilla para formar el primer grupo.

8) Generar medidas de evaluación para los agrupamientos: Realizar r análisis de variancia utilizando los r agrupamientos candidatos definidos en 7) y las variables de interés, generando r valores de $SCW \cong \hat{\sigma}^2$, es decir $\hat{\sigma}_r^2$.
Calcular la media de las distancias al cuadrado entre todas las unidades en cada uno de los r agrupamientos generados en 7): $CMD_r = \frac{\sum_k \sum_{i,k \in k, j>i} d_{ij}^2}{\sum_k N_k}$,donde k indica la asignación de grupo para las i-ésima y j-ésima unidad y d_{ij}^2 la distancia euclídea entre las unidades i y j.
Calcular también, las medidas de comparación de homogeneidad de tamaños por grupos como el coeficiente de variación del tamaño de los grupos en cada uno de los r agrupamientos $CV_r = \frac{\sum_{k=1}^{K}(x_{.k}-x_{..})^2}{\sum_{k=1}^{K} x_{.k}}$

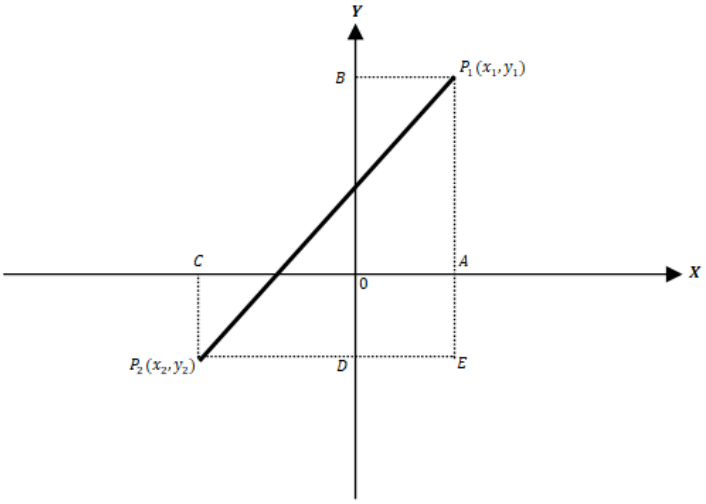
9) Elegir un agrupamiento: A partir de 8), elegir el agrupamiento generado por la unidad semilla del primer grupo que proporcione el mayor valor de $\hat{\sigma}_r^2$, combinado con el valor más pequeño de CMD_r y de CV_r.

Distancia euclídea

Para definir la distancia euclídea bidimensional d_{12} entre P_1 y P_2, siendo $P_1(x_1, y_1)$ y $P_2(x_2, y_2)$ dos puntos dados cualesquiera (Figura 1), se traza la recta que pasa por P_1 y el pie de la perpendicular al eje x. Análogamente, se traza la recta que pasa por P_2 y el pie de la perpendicular al eje y. La interSección de ambas rectas es el punto E. Luego, se aplica el teorema de Pitágoras al triángulo rectángulo $\overset{\triangle}{ABC}$ resultando:

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \qquad (3.25)$$

Figura 1: Distancia Euclídea entre $P_1(x_1, y_1)$ y $P_2(x_2, y_2)$



Notar que d_{ij} se define en el espacio euclídeo bidimensional que es un plano. Los puntos de un plano euclidiano cumplen los axiomas de la geometría de Euclides; los más significativos (Lehmann, 1989) para esta tesis son:

- Por dos puntos pasa una sola recta.
- Tres puntos sobre el plano forman un triángulo cuyos ángulos internos siempre suman 180º.
- En un triángulo rectángulo el cuadrado de la hipotenusa es igual a la suma de los cuadrados de sus catetos.

No todos los espacios bidimensionales cumplen la geometría euclidiana. La superficie de una esfera es un espacio bidimensional, los ángulos de un triángulo sobre una superficie esférica no suman 180º y con ello no se cumple el teorema de Pitágoras, por tanto, una superficie esférica no cumple los axiomas de Euclides (Lifeder, 2019).

Dado que la tierra es mayormente redonda, hay medidas de distancia que tienen en cuenta que las unidades no se encuentran sobre una superficie plana, sino curva. Además, el planeta tierra no tiene un radio constante, ya que se encuentra girando sobre su propio eje a una velocidad tan rápida, que se achata levemente. Existen funciones de distancia que contemplan el achatamiento y curvatura de la tierra y calcula las distancias de una manera acorde[5].

Para simplicidad de cómputo aquí se trabaja con las distancias euclídeas tal como lo propone el algoritmo de Elliott según la fórmula (3.25). La pérdida de precisión es despreciable cuando los puntos están cercanos en la tierra o en la misma zona geográfica como es el caso de Santa Fe, y no justifica utilizar la fórmula de cálculo de distancia dada por la ley esférica de los cosenos[6].

Por último, cabe aclarar que al utilizar las coordenadas de un punto en grado angular (latitud y longitud) según el Sistema Geodésico Mundial 1984 WGS84 (por sus siglas en inglés de World Geodetic System 84), la distancia euclídea entre puntos resulta en grados angulares, no en la unidad de distancia real en metros o kilómetros; proporcionando un indicador aproximado e igualmente útil para comparar las distancias entre puntos.

3.3 Estratificación de conglomerados

Las encuestas por muestreo a menudo utilizan la estratificación de unidades para mejorar la eficiencia estadística. La estratificación es una herramienta que crea o bien aprovecha las agrupaciones conocidas, como una partición de la población en grupos mutuamente excluyentes y homogéneos intragrupo para reducir la variancia. Se van a presentar técnicas de formación de estratos que sirven para cualquier unidad de muestreo, ya sea económica, poblacional, como también aplicables a conjuntos de unidades, o conglomerados, siendo

[5] Equatorial Radius of the Earth. Recuperado de http://mgaia.usno.navy.mil/NSFA/NSFA_cbe.html#EarthRadius2009.

[6] Geographic distance can be simple and fast Recuperado de: http://jonisalonen.com/2014/computing-distance- betweencoordinates-can-be-simple-and-fast/.

este último el caso que se plantea en esta tesis en particular. En la Sección 3.3.1 se presenta una clasificación de variables para construir estratos, en la Sección 3.3.2 como variable de insumo para la estratificación se presenta la estimación de CP robustos utilizando el método FRB para definir las cargas significativas, y así obtener una CP que resuma la información de las variables consideradas. Finalmente, en la Sección 3.3.3 se presentan técnicas para estratificar poblaciones asimétricas.

3.3.1 Variables de estratificación

En general, las variables utilizadas para la estratificación dependen del objetivo principal de la encuesta que se desee realizar. Por ello, además de la estratificación primaria: geográfica o administrativa, los siguientes criterios pueden resultar útiles para la estratificación ulterior de los marcos de las encuestas sobre hogares. Cuando se dispone de una serie de características o variables de las unidades censales que sirvan como posibles variables para una estratificación, deben elegirse entre ellas las más útiles con respecto a los estimadores buscados en la encuesta (por ejemplo, ingresos de los hogares, incidencia de las unidades del sector informal, características de las viviendas, ocupación del jefe de hogar, etc.). El primer paso consiste en examinar el intervalo de variación (en términos relativos) o el coeficiente de variación de cada característica, por separado para cada estrato primario, lo que implica examinar las distribuciones de frecuencia y los coeficientes de variación cuando así convenga (pues la unidad de análisis son las unidades del censo u otro tipo de elemento que sirva como UPM) de cada característica importante dentro de cada estrato primario. Las características que presenten altos coeficientes o intervalos de variación son candidatas para ser utilizadas como variables de estratificación. Por lo general resulta más eficaz utilizar una serie de variables de forma simultánea, que utilizar clases de una variable única. Sin embargo, no es conveniente incluir todas las variables si éstas presentan una correlación elevada entre ellas. Por este motivo es necesario examinar la correlación entre aquellas con altos intervalos o coeficientes de variación (Naciones Unidas, 2010).

3.3.2 Componentes Principales Robustas y Bootstrap Rápido y Robusto

Una alternativa para construir estratos con muchas variables que resultan de interés, y estén posiblemente correlacionadas, consiste en tomar las primeras dimensiones relevantes que resultan de aplicar métodos de análisis multivariado de reducción de dimensiones, generando índices independientes con combinaciones de dichas variables como posible clasificación de estratos. Aquí se emplea en particular la técnica de CP para combinar un conjunto de las variables de interés y reducirlo a un índice dedicado a la formación de estratos. Para la estimación de las CP se utilizan los estimadores multivariantes robustos (MM) al conocer que las variables estudiadas presentan observaciones muy alejadas, denominado CP robustas.

En general, la inferencia basada en los estimadores robustos para regresión lineal o modelos multivariados de posición y forma se basa en la distribución asintótica de estos estimadores. Esta distribución asintótica ha sido estudiada bajo el supuesto que la distribución de los datos es simétrica o elíptica. Para estimar las CP robustas se puede usar, por ejemplo, la aproximación normal y una estimación de la matriz de covariancia asintótica apropiada. Desafortunadamente estos supuestos no se mantienen en la mayoría de las situaciones prácticas, cuando serían recomendados los métodos robustos, por ejemplo, cuando hay presencia de outliers en los datos.

La inferencia basada en las variancias asintóticas aún puede arrojar resultados razonables para pequeñas cantidades de contaminación cuando el tamaño de la muestra es grande. Un enfoque alternativo para realizar inferencias basadas en estimadores robustos viene dado por el método bootstrap (Efron, 1979) que generalmente no requiere supuestos distribucionales rigurosos. En la mayoría de los casos prácticos, el bootstrap se basa en simular la distribución del estimador de interés generando una gran cantidad de nuevas muestras extraídas aleatoriamente de los datos originales, denominadas muestras bootstrap. El bootstrap proporciona una alternativa con uso intensivo computacional que puede funcionar mejor para tamaños de muestra más pequeños y para desviaciones mayores del modelo Gaussiano. Además, dado que el bootstrap puede estimar la distribución muestral del estimador de interés, ofrece una gama más amplia de aplicaciones de inferencia y permite evaluar la estabilidad de los resultados de CP. Sin embargo, aplicar bootstrap a estimadores robustos como el estimador MM plantea algunas dificultades. Un problema grave es el alto costo computacional de estos estimadores. De hecho, calcular el estimador MM (particularmente el estimador S inicial) es una tarea que requiere mucho tiempo. Otro problema típico que surge es la inestabilidad del procedimiento bootstrap clásico. Si hay observaciones atípicas presentes en la muestra original, estas también podrán ocurrir en las muestras bootstrap, por lo tanto, la distribución bootstrap simulada puede verse gravemente afectada por muestras bootstrap con una mayor proporción de valores atípicos que en el conjunto de datos original. Incluso si el estimador MM de los datos originales produce una solución robusta de CP, podría fallar en muestras bootstrap con muchos valores atípicos. En otras palabras, la inferencia basada en el bootstrap clásico puede ser menos robusta que el propio estimador MM.

Salibian-Barrera y Zamar (2002) plantean un método bootstrap para estimadores robustos que tiene la ventaja de ser al mismo tiempo rápido de calcular y resistente a valores atípicos en la muestra, es decir, puede superar los dos problemas mencionados arriba. En principio, el método se puede utilizar para cualquier estimador que pueda escribirse como la solución de un sistema de ecuaciones suavizadas de punto fijo (Salibián-Barrera, Van Aelst, & Willems, 2006), entre los cuales se encuentran los estimadores multivariantes S y MM planteados en el principio de este apartado.

A continuación, se introducen brevemente los conceptos que forman parte del método FRB aplicado a CP robustas, para luego desarrollar el método completo. En el apartado 3.3.2.1 se define la técnica de CP, en el siguiente 3.3.2.2 se definen los estimadores robustos multivariantes MM, luego en 3.3.2.3 se presenta el método de estimación por FRB y en el último apartado 3.3.2.4, finalmente se presenta dicho método FRB para CP robustas.

3.3.2.1 Componentes Principales

Sea una muestra $\mathbf{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^p$ de una distribución de $\mathbf{p}$ variables desconocida $\mathbf{F}$ con centro $\boldsymbol{\mu}$ y matriz de covariancia $\boldsymbol{\Sigma}$ de dimensión pxp. El análisis de CP tiene como objetivo explicar la estructura de covariancia de $\mathbf{F}$. Se puede usar para reducir la dimensión de los datos sin demasiada pérdida de información, proyectando las observaciones en un pequeño número de componentes principales r que son combinaciones lineales de las p variables originales ($r < p$). A nivel poblacional, las componentes principales vienen dados por los autovectores de $\boldsymbol{\Sigma}$. En CP clásico, las componentes se estiman mediante los autovectores de la matriz de covariancia de la muestra o la matriz de forma $\boldsymbol{\Gamma}_n$. Los autovalores miden la cantidad de variancia explicada por las componentes correspondientes (Van Aelst & Willems, 2013).

La primera componente principal, $\boldsymbol{v_1}$, es la combinación lineal de las variables originales que resulta de la proyección que tiene mayor variabilidad de manera tal que:

$$Var\left(\boldsymbol{v_1'X_n}\right) = max \text{ } dado\ que \|\boldsymbol{v_1}\| = 1 \tag{3.26}$$

siendo $\boldsymbol{v_1'X_n}$ la forma que toma la combinación lineal. La segunda componente principal $\boldsymbol{v_2}$, se obtiene de buscar otro vector con el mismo procedimiento que se halla $\boldsymbol{v_1}$ y además que dichos vectores sean ortogonales entre sí: $\boldsymbol{v_2'v_1} = 0$. Similarmente, las componentes principales subsiguientes resultan los autovectores de $\boldsymbol{\Sigma_n}$ asociados a los autovalores en forma decreciente. El número total de componentes es $\boldsymbol{p}$, los autovalores de la matriz $\boldsymbol{\Sigma_n}$ son $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ y sus respectivos autovectores son $\boldsymbol{v_1, v_2, \ldots, v_p}$. La variancia de cada componente principal $\boldsymbol{j}$ esta dada por:

$$Var\left(\boldsymbol{v_j'X_n}\right) = \lambda_j.$$

Por lo tanto, el autovalor en sí mismo, mide la cantidad de información que puede ser explicada por esa proyección $\boldsymbol{j}$. El número de componentes puede ser seleccionado a través del criterio de porcentaje de variancia explicada:

$$\frac{\sum_{i=1}^{r} \lambda_j}{\sum_{i=1}^{p} \lambda_j}.$$

Sin embargo, la matriz de covariancia muestral es conocida por ser sensible a los valores atípicos. Por lo tanto, en la literatura se han propuesto varias versiones robustas para el análisis de CP clásico. Aquí se considera el análisis de CP basado en estimaciones robustas de forma. En particular, se utilizan los autovectores y los autovalores de los MM estimadores de forma multivariados introducidos por Tatsuoka y Tyler (2000). Los MM estimadores están diseñados para ser altamente robustos contra valores atípicos y altamente eficientes para datos normales.

3.3.2.2 MM-Estimadores Multivariante

Los estimadores MM de posición y forma multivariada se basan en dos funciones de pérdida $\boldsymbol{\rho}$, las cuales cumplen las siguientes condiciones de regularidad:

$(R1)$ $\boldsymbol{\rho}$ es real, simétrica, doblemente diferenciable y $\boldsymbol{\rho(0) = 0}$.

$(R2)$ $\boldsymbol{\rho}$ estrictamente creciente en $[0, c_0]$ y constante en $[c_0, \infty)$ para alguna constante finita $\boldsymbol{c_0}$.

En primer lugar, se estima la escala mediante un estimador S robusto y luego los estimadores MM multivariados estiman la posición y la forma mediante una función $\boldsymbol{\rho}$ diferente que produzca una mejor eficiencia. A continuación, se definen los estimadores MM multivariados de posición, forma y covariancia:

Definición: Sea $\mathbf{X_n} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\} \subset \mathbb{R}^p$ con $n \geq p + 1$. Sean ρ_0 y ρ_1 dos funciones de pérdida que satisfacen $(R1)$ y $(R2)$ y sean los estimadores S multivariados $(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$, es decir que minimizan $|\boldsymbol{C}|$ sujeto a:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left([(\boldsymbol{x_i} - \boldsymbol{T})^t \boldsymbol{C}^{-1}(\boldsymbol{x_i} - \boldsymbol{T})]^{1/2}\right) = b$$

entre todos los $\boldsymbol{T} \in \mathbb{R}^p$ y $\boldsymbol{C} \in \boldsymbol{PDS}(p)$, donde $\boldsymbol{PDS}(p)$ indica el conjunto de matrices simétricas definidas positivas de orden $p\ x\ p$. Sea $\hat{\sigma}_n := \left|\widetilde{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}$ ($:=$ igual por definición), entonces el estimador MM multivariado de posición y forma $(\widehat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Gamma}}_n)$ minimiza:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left([(\boldsymbol{x}_i - \boldsymbol{T})^t \boldsymbol{G}^{-1}(\boldsymbol{x}_i - \boldsymbol{T})]^{1/2}/\hat{\sigma}_n\right)$$

entre todos $\boldsymbol{T} \in \mathbb{R}^p$ y $\boldsymbol{G} \in \boldsymbol{PDS}(p)$ con $|\boldsymbol{G}| = 1$. El estimador MM para la matriz de covariancia es $\widehat{\boldsymbol{\Sigma}}_n = \hat{\sigma}_n^2 \hat{\boldsymbol{\Gamma}}_n$. Entonces, la matriz $\boldsymbol{G}$ se considera como la matriz de forma o estimador de forma que se corresponde a la matriz de covariancia (o estimador) $\mathbf{C}$, mediante $\mathbf{G} = |\mathbf{C}|^{1/p}\mathbf{C}$.

Se considera como funciones de pérdida a la familia de funciones bi-ponderadas de Tukey, dadas por

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c_0^2} + \dfrac{x^6}{6c_0^4}, & |x| \le c_0 \\ \dfrac{c_0^2}{6} & |x| \ge c_0 \end{cases}$$

donde $c_0 > 0$ es la constante de ajuste elegida por el usuario. Notar que el estimador S puede ser visto como un caso especial del estimador MM, ya que elegir $\rho_0 = \rho_1$ produce el estimador S inicial.

Si $\boldsymbol{\Gamma}_n$ tiene autovalores distintos $\lambda_1 \succcurlyeq \lambda_2 \succcurlyeq \cdots \succcurlyeq \lambda_p > 0$ con sus correspondientes autovectores $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$, el método de CP robusto basado en el estimador MM consiste esencialmente en estimar estos autovalores y autovectores mediante los autovalores y autovectores del estimador MM de la matriz de forma $\hat{\boldsymbol{\Gamma}}_n$. En principio, realizar CP basado en $\widehat{\boldsymbol{\Sigma}}_n$ en lugar de $\hat{\boldsymbol{\Gamma}}_n$ produciría el mismo análisis, ya que $\widehat{\boldsymbol{\Sigma}}_n$ y $\hat{\boldsymbol{\Gamma}}_n$ tienen los mismos autovalores y autovectores proporcionales. Sin embargo, se va a trabajar principalmente con la matriz de forma $\hat{\boldsymbol{\Gamma}}_n$ buscando evitar el potencial sesgo del componente $\widehat{\boldsymbol{\Sigma}}_n$ que generaría si las variables tuvieran escalas de medidas muy disimiles.

3.3.2.3 Bootstrap Rápido y Robusto

Se considera una muestra $\boldsymbol{X}_n$ y sea $\widehat{\boldsymbol{\theta}}_n \in \mathbb{R}^p$ las estimaciones robustas de los parámetros de interés ($\widehat{\boldsymbol{\theta}}_n$ puede incluir estimaciones para los coeficientes de regresión y la escala en el modelo lineal, o para la posición y la matriz de dispersión en el modelo multivariado). Se supone ahora que $\widehat{\boldsymbol{\theta}}_n$ se puede representar como una solución de ecuaciones de punto fijo:

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{g}_n(\widehat{\boldsymbol{\theta}}_n) \tag{3.27}$$

donde la función $\boldsymbol{g}_n: \mathbb{R}^p \to \mathbb{R}^p$ generalmente depende de la muestra $\boldsymbol{X}_n$. Dada una muestra bootstrap $\boldsymbol{X}_n^*$, la estimación recalculada $\widehat{\boldsymbol{\theta}}_n^*$ resuelve:

$$\widehat{\boldsymbol{\theta}}_n^* = \boldsymbol{g}_n^*(\widehat{\boldsymbol{\theta}}_n^*) \tag{3.28}$$

donde la función $\boldsymbol{g}_n^*$ ahora depende de $\boldsymbol{X}_n^*$. Como se explicó anteriormente, calcular $\widehat{\boldsymbol{\theta}}_n^*$ para cada muestra bootstrap $\boldsymbol{X}_n^*$ puede ser una tarea computacionalmente costosa para estimadores robustos como los S o MM.

Además, aunque se puede suponer que la solución a (3.27) resulta resistente a los valores atípicos en $\boldsymbol{X}_n$, esto no garantiza que se obtendrá una solución igualmente resistente a (3.28) ya que $\boldsymbol{g}_n^*$ se puede ver potencialmente más afectado por los valores atípicos que $\boldsymbol{g}_n$. Entonces, en lugar de calcular $\widehat{\boldsymbol{\theta}}_n^*$, se puede calcular la aproximación:

$$\widehat{\boldsymbol{\theta}}_n^{1*} := \boldsymbol{g}_n^*(\widehat{\boldsymbol{\theta}}_n) \tag{3.29}$$

como una estimación en un solo paso de $\widehat{\boldsymbol{\theta}}_n^*$ con valor inicial $\widehat{\boldsymbol{\theta}}_n$. Se puede demostrar que, bajo ciertas condiciones, la distribución de $\widehat{\boldsymbol{\theta}}_n^*$ estima consistentemente la distribución muestral de $\widehat{\boldsymbol{\theta}}_n$. Sin embargo, es intuitivamente claro que la distribución de $\widehat{\boldsymbol{\theta}}_n^{1*}$ puede no tener esta propiedad en general. De hecho, generalmente $\widehat{\boldsymbol{\theta}}_n^{1*}$ subestima la variabilidad de $\widehat{\boldsymbol{\theta}}_n$, principalmente porque cada muestra bootstrap utiliza el mismo valor inicial en la aproximación de un solo paso. Para remediar esto, se puede aplicar una corrección lineal, usando un suavizado de $\boldsymbol{g}_n$, con una expansión por Series de Taylor de $\widehat{\boldsymbol{\theta}}_n$, alrededor del valor límite $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{g}_n(\boldsymbol{\theta}) + \boldsymbol{\nabla g}_n(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) + R_n \tag{3.30}$$

donde R_n es el término remanente y $\boldsymbol{\nabla g}_n(.) \epsilon\ \mathbb{R}^{pxp}$ es la matriz de derivadas parciales. Cuando el término remanente es insignificante, la ecuación (3.30) se puede reescribir como:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \dot{\sim} [\boldsymbol{I} - \boldsymbol{\nabla g}_n(\boldsymbol{\theta})]^{-1}\sqrt{n}(\boldsymbol{g}_n(\boldsymbol{\theta}) - \boldsymbol{\theta})$$

donde $\dot{\sim}$ indica que ambos lados tienen la misma distribución límite. Bajo ciertas condiciones se tiene que $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n) \dot{\sim} \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ y $\sqrt{n}(\boldsymbol{g}_n^*(\boldsymbol{\theta}) - \boldsymbol{\theta}) \dot{\sim} \sqrt{n}(\boldsymbol{g}_n^*(\widehat{\boldsymbol{\theta}}_n) - \widehat{\boldsymbol{\theta}}_n)$. Si además se aproxima $[\mathbf{I} - \boldsymbol{\nabla g}_n(\boldsymbol{\theta})]^{-1}$ por $\left[\mathbf{I} - \boldsymbol{\nabla g}_n(\widehat{\boldsymbol{\theta}}_n)\right]^{-1}$ se obtiene:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n) \dot{\sim} \left[\boldsymbol{I} - \boldsymbol{\nabla g}_n(\widehat{\boldsymbol{\theta}}_n)\right]^{-1}\sqrt{n}(\boldsymbol{g}_n^*(\widehat{\boldsymbol{\theta}}_n) - \widehat{\boldsymbol{\theta}}_n). \tag{3.31}$$

Se define entonces la versión corregida linealmente de la aproximación de un paso (3.29) como:

$$\widehat{\boldsymbol{\theta}}_n^{R*} := \widehat{\boldsymbol{\theta}}_n + \left[\boldsymbol{I} - \boldsymbol{\nabla g}_n(\widehat{\boldsymbol{\theta}}_n)\right]^{-1}(\widehat{\boldsymbol{\theta}}_n^{1*} - \widehat{\boldsymbol{\theta}}_n). \tag{3.32}$$

Cuando (3.31) efectivamente se cumple, $\widehat{\boldsymbol{\theta}}_n^{R*}$ estimará la misma distribución límite que los cálculos bootstrap reales $\widehat{\boldsymbol{\theta}}_n^*$ (Salibián-Barrera, Van Aelst, & Willems, 2008).

A continuación, se aplica dicho estimador, $\widehat{\boldsymbol{\theta}}_n^{R*}$ al análisis de CP que resulta mucho más fácil de calcular que $\widehat{\boldsymbol{\theta}}_n^*$, y también resulta más estable numéricamente (más robusto), por lo cual el método de inferencia resultante se denomina bootstrap rápido y robusto o FRB según Salibian-Barrera y Zamar (2002).

3.3.2.4 CP robusto con Bootstrap Rápido y Robusto

Los MM estimadores multivariados ya definidos, se pueden escribir como un sistema de ecuaciones de punto fijo de la siguiente manera:

$$\widehat{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i/|\widehat{\boldsymbol{\Sigma}}_n|^{1/(2p)}\right)}{d_i}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i/|\widehat{\boldsymbol{\Sigma}}_n|^{1/(2p)}\right)}{d_i}\,\boldsymbol{x}_i\right) \tag{3.33}$$

$$\widehat{\boldsymbol{\Gamma}}_n = Z\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i/|\widehat{\boldsymbol{\Sigma}}_n|^{1/(2p)}\right)}{d_i}\,(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)^t\right) \tag{3.34}$$

$$\widetilde{\boldsymbol{\Sigma}}_n = \frac{1}{nb}\left(\sum_{i=1}^{n} p\,\frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}\,(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)^t + \left(\sum_{i=1}^{n} \widetilde{\omega}_i\right)\widetilde{\boldsymbol{\Sigma}}_n\right) \tag{3.35}$$

$$\widetilde{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}\,\boldsymbol{x}_i\right) \tag{3.36}$$

donde $\mathrm{Z}(\mathbf{A}) = |\mathbf{A}|^{-1/\mathrm{p}}\mathbf{A}$ para las matrices $\mathbf{A}$ de dimensión $p\mathrm{x}p$, $\mathrm{d_i} = \left[(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)^{\mathrm{t}}\widehat{\boldsymbol{\Gamma}}_{\mathrm{n}}^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)\right]^{1/2}$, $\tilde{\mathrm{d}}_{\mathrm{i}} = \left[(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)^{\mathrm{t}}\widetilde{\boldsymbol{\Sigma}}_{\mathrm{n}}^{-1}(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)\right]^{1/2}$ y $\widetilde{\omega}_{\mathrm{i}} = \rho_0(\tilde{\mathrm{d}}_{\mathrm{i}}) - \rho_0'(\tilde{\mathrm{d}}_{\mathrm{i}})\tilde{\mathrm{d}}_{\mathrm{i}}$.

Este sistema de ecuaciones permite aplicar el procedimiento FRB, utilizando las ecuaciones para calcular aproximaciones rápidas a las estimaciones MM en cada muestra bootstrap. En particular, dada una muestra bootstrap $\boldsymbol{X}_n^* = \{\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*\}$, una forma de obtener rápidamente cálculos aproximados es de la siguiente manera:

$$\widehat{\boldsymbol{\mu}}_n^* = \left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i^*/|\widehat{\boldsymbol{\Sigma}}_n|^{1/(2p)}\right)}{d_i^*}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i^*/|\widehat{\boldsymbol{\Sigma}}_n|^{1/(2p)}\right)}{d_i^*}\,\boldsymbol{x}_i^*\right) \tag{3.37}$$

$$\widehat{\boldsymbol{\Gamma}}_n^* = G\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i^*/|\widehat{\boldsymbol{\Sigma}}_n|^{1/(2p)}\right)}{d_i^*}\,(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)^t\right) \tag{3.38}$$

$$\widetilde{\boldsymbol{\Sigma}}_n^* = \frac{1}{nb}\left(\sum_{i=1}^{n} p\,\frac{\rho_0'(\tilde{d}_i^*)}{\tilde{d}_i^*}\,(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)^t + \left(\sum_{i=1}^{n} \widetilde{\omega}_i^*\right)\widetilde{\boldsymbol{\Sigma}}_n\right) \tag{3.39}$$

$$\widetilde{\boldsymbol{\mu}}_n^* = \left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i^*)}{\tilde{d}_i^*}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i^*)}{\tilde{d}_i^*}\,\boldsymbol{x}_i^*\right) \tag{3.40}$$

donde $\mathrm{d_i^*} = \left[(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)^{\mathrm{t}}\widehat{\boldsymbol{\Gamma}}_{\mathrm{n}}^{-1}(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)\right]^{1/2}$, $\tilde{\mathrm{d}}_{\mathrm{i}}^* = \left[(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)^{\mathrm{t}}\widetilde{\boldsymbol{\Sigma}}_{\mathrm{n}}^{-1}(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)\right]^{1/2}$ y $\widetilde{\omega}_{\mathrm{i}}^* = \rho_0(\tilde{\mathrm{d}}_{\mathrm{i}}^*) - \rho_0'(\tilde{\mathrm{d}}_{\mathrm{i}}^*)\tilde{\mathrm{d}}_{\mathrm{i}}^*$. Tener en cuenta que dado que se mantienen los estimadores $\widehat{\boldsymbol{\mu}}_n$, $\widehat{\boldsymbol{\Gamma}}_n$, $\widetilde{\boldsymbol{\Sigma}}_n$ y $\widetilde{\boldsymbol{\mu}}_n$ fijos en el lado derecho de (3.37) a (3.40), estas aproximaciones probablemente subestimarán la variabilidad del estimador MM. Para remediar esto se aplica la corrección lineal $\widehat{\boldsymbol{\theta}}_n^{R*}$, presentada en (3.32), de la siguiente manera:

$$\widehat{\boldsymbol{\theta}}_n^{R*} := \widehat{\boldsymbol{\theta}}_n + \left[\boldsymbol{I} - \boldsymbol{\nabla g}_n(\widehat{\boldsymbol{\theta}}_n)\right]^{-1}(\boldsymbol{g}_n^*(\widehat{\boldsymbol{\theta}}_n) - \widehat{\boldsymbol{\theta}}_n).$$

Luego, el FRB se obtiene calculando el lado derecho en lugar del lado izquierdo para cada muestra de bootstrap, donde $\widehat{\boldsymbol{\theta}}_n$ indican las estimaciones iniciales (3.33) a (3.36), $\boldsymbol{g}_n^*$ representa las estimaciones bootstrap (3.37) a (3.40) de los respectivos estimadores $\widehat{\boldsymbol{\mu}}_{\mathrm{n}}^*$, $\widehat{\boldsymbol{\Gamma}}_{\mathrm{n}}^*$, $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{n}}^*$ y $\widetilde{\boldsymbol{\mu}}_{\mathrm{n}}^*$, y la matriz de derivadas parciales $\boldsymbol{\nabla g}_n$ proporciona la corrección lineal. Se denomina a las aproximaciones obtenidas de $\widehat{\boldsymbol{\theta}}_n^{R*}$ con $\widehat{\boldsymbol{\mu}}_{\mathrm{n}}^{\mathrm{R}*}$, $\widehat{\boldsymbol{\Gamma}}_{\mathrm{n}}^{\mathrm{R}*}$,

$\widetilde{\boldsymbol{\Sigma}}_{\mathrm{n}}^{\mathrm{R}*}$y $\widetilde{\boldsymbol{\mu}}_{\mathrm{n}}^{\mathrm{R}*}$ respectivamente. Entonces, las estimaciones de los autovalores y autovectores de la matriz de forma $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{n}}$ usando el boostrap rápido y robusto, resultan los autovalores y autovectores de $\hat{\boldsymbol{\Gamma}}_{\mathrm{n}}^{\mathrm{R}*}$, tomando a los autovalores y autovectores de estas estimaciones como versiones recalculadas de los autovalores y autovectores de la matriz de forma (Salibián-Barrera, Van Aelst, & Willems, 2006).

3.3.3 Estratificación de poblaciones asimétricas

Se debe tener en cuenta a la hora de estratificar una población, la forma de la distribución de la variable o indicador bajo estudio. Existen en la literatura un conjunto de métodos de estratificación que son válidos para distribuciones simétricas, tales como Ekman (1949), o Dalenius y Hodges (1959). El problema de estos métodos es que no brindan soluciones precisas cuando la distribución de la variable bajo estudio es asimétrica. A continuación, se presentan métodos que permiten estratificar poblaciones basándose en variables con distribución asimétrica.

Hidiroglou (1986) proporciona un algoritmo para dividir la población en un estrato *take-all* (donde se seleccionan todas las unidades en ese conjunto) y un estrato *take-some* (donde se seleccionan algunas unidades de ese conjunto) para minimizar el tamaño total de la muestra dado un nivel de precisión, asumiendo un MSA en el estrato *take-some*.

Lavallee-Hidiroglou (1988) presentan un método iterativo para estratificar poblaciones marcadamente asimétricas en un estrato *take-all* y varios estratos *take-some*. El objetivo es minimizar el tamaño de muestra general dado el coeficiente de variación del estimador y dado un esquema de asignación para la selección de los estratos *take-some*. Los límites de los estratos pueden ser calculados en términos de una variable auxiliar que se encuentre relacionada con la variable bajo estudio de la encuesta. Para una encuesta multipropósito, como los límites de los estratos se utilizan usando una variable auxiliar relacionada a la variable principal, la optimalidad de los límites disminuirá para otras variables que no se encuentren tan relacionadas con la misma. El algoritmo es una modificación del método de Sethi (1963) para estratificar poblaciones. Los límites resultan óptimos por proporcionar el tamaño de muestra mínimo requerido.

Kozak (2004) presenta un método de optimización numérica que brinda límites de los estratos cercanos al óptimo para poblaciones asimétricas, considerando también un estrato *take-all*. Se busca de manera aleatoria los límites que minimicen el tamaño de la muestra, en función a la variable en estudio, para obtener un nivel de precisión dado, bajo dos condiciones que aseguran que todos los estratos tengan al menos dos unidades y que a su vez de cada estrato se seleccione un mínimo de dos unidades.

La estratificación Geométrica es un método aproximado desarrollado por Gunning y Horgan (2004) 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 ellos. La misma se basa en una observación de Cochran (1961) que, con límites de estratos cercanos a los óptimos, los coeficientes de variación resultan aproximadamente iguales en todos los estratos. Este procedimiento resulta ser el de mayor facilidad de aplicación debido a que no usa ningún proceso iterativo, como los mencionados.

Hidiroglou y Kozak (2018) evaluaron una propuesta integral para calcular los límites de estratificación, dado un esquema de asignación. Se comparan los métodos aproximados de la acumulada de las raíces de la función

de distribución (Dalenius & Hodges, 1957) y geométrico con aquellos basados en la optimización como Lavallee-Hidiroglou (1988) y Kozak (2004). Entre algunas de las conclusiones del artículo, demuestran que la estratificación basada en optimización es más eficiente que los métodos aproximados. En cuanto a los esquemas de asignación, los métodos X-proporcional y Neyman resultan los más eficientes, mientras que la asignación N-proporcional es ineficiente si no se incluye un estrato *take-all*.

Por lo tanto, se desarrolla a continuación el método que se propone utilizar en esta tesis, el de estratificación para poblaciones asimétricas propuesto por Lavallée-Hidiroglou mediante el algoritmo iterativo de Sethi, que busca establecer los límites de los estratos que minimizan el tamaño de la muestra, dado un valor del coeficiente de variación.

3.3.3.1 Algoritmo de estratificación por Lavallée-Hidiroglou

Para ilustrar el método, los autores eligen un esquema de asignación de potencia. El algoritmo, que es de naturaleza recursiva, es simple de programar y converge rápidamente a los puntos límites óptimos.

Se considera una población finita ordenada de N unidades: $y_{(1)}, y_{(2)}, \dots, y_{(N)}$, con $y_{(i)} \leq y_{(i+1)}$, para $i = 1, 2, \dots, N-1$. Se desea estratificar en L estratos esta población. El número de unidades en cada estrato se indica con N_h , $h = 1, 2, \dots, L$. Se indica con n_h, la cantidad de unidades a seleccionar en cada estrato *take-some* ($h = 1, 2, \dots, L-1$) usando un MSA sin reemplazo, y $n_L = N_L$ correspondiente al estrato *take-all*. Si bien hasta el momento se consideró la estimación del total, el método está planteado para el estimador de la media. La media a estimar es:

$$\bar{Y} = \sum_{h=1}^{L} \sum_{j=M_{h-1}+1}^{M_h} \frac{y_{(j)}}{N} \tag{3.41}$$

donde $M_h = \sum_{i=1}^{h} N_i$ para $h = 1, 2, \dots, L$ y $M_0 = 0$.

Para un conjunto dado, el estimador de la media poblacional $\bar{Y}$ es:

$$\hat{\bar{Y}} = \frac{\left[\sum_{h=1}^{L-1} \frac{N_h}{n_h} \sum_{j=m_{h-1}+1}^{m_n} z_j + \sum_{j=M_{L-1}+1}^{N} y_{(j)}\right]}{N} \tag{3.42}$$

donde $y_{M_{h-1}} \leq z_j \leq y_{M_h}$ para $j = m_{h-1} + 1, \dots, m_h$, $m_h = \sum_{i=1}^{h} n_i$ para $h = 1, 2, \dots, L$ y $m_0 = 0$.

Se toma c como el nivel de precisión deseado para la estimación de la media (coeficiente de variación) y a_h como la proporción de unidades muestrales a ser asignadas a cada uno de los primeros L-1 estratos donde $\sum_{h=1}^{L-1} a_h = 1$. El término a_h se usa para representar cualquier tipo de asignación para estratos. Por ejemplo, en el caso de asignación de potencia N-proporcional:

$$a_h = \frac{N_h^P}{\sum_{h=1}^{L-1} N_h^P}$$

con $h = 1, 2, \dots, L-1$, y en el caso de asignación de potencia Y-proporcional:

$$a_h = \frac{Y_h^p}{\sum_{h=1}^{L-1} Y_h^p}$$

donde $0 < p < \infty$. La asignación de potencia bajo supuestos relativamente simples y para una elección adecuada de p, tiene la propiedad buscada, que los CV tienden a igualarse para los estratos *take-some*, sin un aumento significativo en el CV general.

En la práctica, el valor de p se elige frecuentemente 0,5 o 0,33. Un valor pequeño de p (por ejemplo, cercano a 0) usualmente produce CV similares, mientras un valor más grande incrementa la discrepancia entre los CV, pero también aumenta la precisión estimada general.

La variancia de $\hat{\bar{Y}}$ es:

$$V\left(\hat{\bar{Y}}\right) = \frac{1}{N^2}\sum_{h=1}^{L-1}\frac{N_h}{n_h}(N_h - n_h)S_h^2 \tag{3.43}$$

donde S_h^2 es la variancia poblacional de la variable Y en cada estrato h. En términos del nivel deseado del coeficiente de variación c, $V\left(\hat{\bar{Y}}\right)$ se puede expresar como $V\left(\hat{\bar{Y}}\right) = c^2\bar{Y}^2$. Sustituyendo $n_h = (n - N_h)a_h$ y $V\left(\hat{\bar{Y}}\right) = c^2\bar{Y}^2$ en (3.43) y despejando n se obtiene:

$$n = N_L + \frac{\sum_{h=1}^{L-1}\frac{N_h^2 S_h^2}{a_h}}{(Nc\bar{Y})^2 + \sum_{h=1}^{L-1}N_h S_h^2}. \tag{3.44}$$

El problema consiste en encontrar los límites $b_{(1)}, b_{(2)}, \dots, b_{(L-1)}$ (donde $y_{(1)} < b_{(1)} < \cdots < b_{(L-1)} < y_{(N)}$) de manera tal que minimice el tamaño de la muestra total, dado el nivel de precisión y un esquema de asignación.

El enfoque presentado a continuación, para obtener los límites de estratificación mediante un nivel de precisión deseado, ha sido utilizado por primera vez por Dalenius (1950) en el caso de límites de estratificación para un tamaño de muestra determinado.

El algoritmo supone en primer lugar, que el muestreo se realiza a partir de una población cuya distribución de frecuencia puede representarse con suficiente precisión mediante una función de densidad continua $f(y)$. Entonces, para un conjunto dado de límites $b_{(1)}, b_{(2)}, \dots, b_{(L-1)}$, se definen las siguientes cantidades:

$$W_h = \int_{b_{(h-1)}}^{b_{(h)}} f(y)dy$$

$$\mu_h = \int_{b_{(h-1)}}^{b_{(h)}} yf(y)dy / W_h$$

$$\sigma_h^2 = \int_{b_{(h-1)}}^{b_{(h)}} y^2 f(y)dy / W_h - \mu_h^2$$

para $h = 1, 2, \dots, L$ con $b_{(0)} = -\infty$ y $b_{(L)} = +\infty$.

Si bien se requiere algún conocimiento de la densidad aproximada de $f(y)$, como la población se considera finita, es posible superar esta condición reemplazando las cantidades anteriores W_h, y de los parámetros de

la población infinita por la expresión correspondiente basada en la población finita. Cochran (1977) plantea los homólogos de la población infinita, quedando

$$W_h = \frac{N_h}{N} \tag{3.45}$$

$$\bar{Y}_h = \frac{1}{N_h}\sum_{j=b_{(h-1)}+1}^{b_{(h)}} y_{(j)} \tag{3.46}$$

$$S_h^2 = \frac{1}{N_{h-1}}\sum_{j=b_{(h-1)}+1}^{b_{(h)}} y_{(j)}^2 - N_h\bar{Y}_h^2 \tag{3.47}$$

para $h = 1, \dots, L$ con $b_{(0)} = y_{(1)}$ y $b_{(L)} = y_{(N)}$.

La ecuación (3.44) se puede escribir:

$$n = NW_L + \frac{N(\sum_{h=1}^{L-1} W_h^2 S_h^2/a_h)}{Nc^2\bar{Y}^2 + \sum_{h=1}^{L-1} W_h S_h^2} \tag{3.48}$$

donde $\bar{Y} = \sum_{h=1}^{L} \bar{Y}_h /N$.

El a_h también se puede representar usando las cantidades (3.45), (3.46) y (3.47). En el caso de la asignación de potencia N-proporcional, se obtiene

$$a_h = \frac{W_h^P}{\sum_{h=1}^{L-1} W_h^P} \text{ para } h = 1, \dots, L-1. \tag{3.49}$$

Para la asignación de potencia Y-proporcional, se tiene que

$$a_h = \frac{(W_h\bar{Y}_h)^p}{\sum_{h=1}^{L-1}(W_h\bar{Y}_h)^p} \tag{3.50}$$

donde $0 < p < \infty$.

A partir de aquí, se desarrolla considerando solamente la asignación de potencia Y-proporcional, pero los cálculos también pueden aplicarse en la asignación de potencia N-proporcional, o a cualquier asignación a_h que cumpla $\sum_{h=1}^{L-1} a_h = 1$. Se reemplaza la ecuación (3.50) en la (3.48) y se obtiene:

$$n = NW_L + \frac{N[\sum_{h=1}^{L-1}(W_h S_h)^2 (W_h\bar{Y}_h)^{-p}][\sum_{h=1}^{L-1}(W_h\bar{Y}_h)^p]}{Nc^2\bar{Y}^2 + \sum_{h=1}^{L-1} W_h S_h^2}. \tag{3.51}$$

Para encontrar los límites óptimos $b_{(1)}, \dots, b_{(L-1)}$ de manera tal que el tamaño de la muestra n sea mínimo, se deriva n de la ecuación (3.51) con respecto a $b_{(1)}, \dots, b_{(L-1)}$ respectivamente y se igualan a cero. Resultan las siguientes ecuaciones para $h = 1, \dots, \mathrm{L}-2$:

$$\begin{gathered}[FT_h - FT_{h+1}]b_{(h)}^2 + \\ [FK_h - 2\mu_h FT_h - FK_{h+1} + 2\mu_{h+1}FT_{h+1} + 2\mu_h AB - 2\mu_{h+1}AB]b_{(h)} + \\ [FT_h\mu_h^2 + FT_h\sigma_h^2 - FT_{h+1}\mu_{h+1}^2 - FT_{h+1}\sigma_{h+1}^2 - AB\mu_h^2 + AB\mu_{h+1}^2] = 0\end{gathered} \tag{3.52}$$

y para $h = L-1$

$$[FT_{L-1} - AB]b_{(L-1)}^2 +$$

$$[FK_{L-1} - 2\mu_{L-1}FT_{L-1} + 2\mu_{L-1}AB]b_{(h)} + [FT_{L-1}\mu_{L-1}^2 + FT_{L-1}\sigma_{L-1}^2 - AB\mu_{L-1}^2 - F] = 0, \quad (3.53)$$

donde

$$A = \sum_{h=1}^{L-1}(W_h\bar{Y}_h)^p$$

$$B = \sum_{h=1}^{L-1}(W_hS_h)^2(W_h\bar{Y}_h)^{-p}$$

$$F = Nc^2\bar{Y}^2 + \sum_{h=1}^{L-1}W_h S_h^2$$

$$K_h = Bp(W_h\bar{Y}_h)^{p-1} - Ap(W_hS_h)^2(W_h\bar{Y}_h)^{-p-1}$$

$$T_h = AW_h(W_h\bar{Y}_h)^{-p}.$$

Llamando al coeficiente de $b_{(h)}^2$ como α_h, al coeficiente de $b_{(h)}$ como β_h y a los términos remanentes como γ_h, las ecuaciones (3.52) y (3.53) se pueden representar con una ecuación cuadrática de la forma $\alpha_h b_{(h)}^2 + \beta_h b_{(h)} + \gamma_h = 0$. Como señala Sethi (1963), los términos α_h , β_h y γ_h son funciones de los límites $b_{(1)}, \dots, b_{(L-1)}$ a través de las ecuaciones (3.45), (3.46) y (3.47). Usando el algoritmo de Sethi, las ecuaciones (3.52) y (3.53) se pueden resolver usando el siguiente modelo iterativo:

Paso 1. Establecer límites arbitrarios $b_{(1)}^{'} < \cdots < b_{(L-1)}^{'}$.

Paso 2. Calcular las proporciones $W_h^{'}$, las medias $\bar{Y}_h^{'}$ y las variancias $S_h^{2'}$ en base a las ecuaciones (3.45), (3.46) y (3.47) respectivamente, basándose en los límites del paso anterior, para $h = 1, \dots, L-1$. También calcular los coeficientes de la función cuadrática $\alpha_h^{'}$, $\beta_h^{'}$ y $\gamma_h^{'}$, en base a las ecuaciones (3.52) y (3.53) mediante los valores recién calculados ($W_h^{'}$, $\bar{Y}_h^{'}$ y $S_h^{2'}$), y el valor del coeficiente de variación c fijado.

Paso 3. Reemplazar los límites anteriores por cada $b_{(1)}^{''}, \dots, b_{(L-1)}^{''}$ que minimiza[7] n con la función cuadrática obtenida de los coeficientes del paso anterior $\alpha_h^{'}$, $\beta_h^{'}$ y $\gamma_h^{'}$, donde:

$$b_{(h)}^{''} = \frac{-\alpha_h^{'} + \sqrt{\beta_h^{'2} - 4\alpha_h^{'}\gamma_h^{'}}}{2\alpha_h^{'}}, h = 1, \dots, L-1.$$

Paso 4. Repetir los pasos 2 y 3 hasta que dos conjuntos consecutivos de límites sean idénticos o que difieran en una cantidad muy pequeña (insignificante) por ejemplo

$$\max_{h=1}^{L-1} \left| b_{(h)}^{''} - b_{(h)}^{'} \right| < \varepsilon \text{ para algún } \varepsilon > 0.$$

[7] Notar que el signo antes de la raíz cuadrada es positivo ya que puede demostrarse que $b_{(h)}^{'}$ se encuentra entre $\mu_h^{'}$ y $\mu_{h+1}^{'}$.

4 Materiales

Se trabaja con las bases de datos disponibles del Censo 2010 (por falta de disponibilidad del Censo 2022 al inicio del desarrollo de esta tesis) que comprenden una base de datos de 3 194 537 personas con 98 variables, una base de 1 023 777 hogares con 80 variables y una base de 1 145 270 viviendas con 23 variables. El conjunto de variables se corresponde con los datos recolectados a partir de los cuestionarios del Censo 2010[8]. Además, se utiliza la base de datos de las unidades geográficas[9] de la Provincia de Santa Fe del Censo 2010, con la georreferencia correspondiente a los radios acompañada de su código de identificación e información estadística asociada; en cada uno de estos registros se calculó el centroide de cada radio con el programa QGIS obteniendo para cada uno de los 3 774 radios dos coordenadas: latitud y longitud. De aquí en adelante se utiliza el software R para procesar todas las bases de datos. Entonces, mediante el código de identificación de radio se agregó a las bases de datos del Censo 2010 la latitud y longitud del centroide del radio correspondientemente a las personas, hogares y viviendas de 3 714 radios con datos censales. La diferencia de 60 radios corresponde a radios que no cuentan con datos de viviendas particulares. Luego de una depuración de dichas bases, tomando los radios con más de 10 viviendas, para realizar las comparaciones entre los diseños propuestos aquí, quedan 3 148 735 personas, 1 018 574 hogares y 1 044 905 viviendas habitables, divididas geográficamente en 3 532 radios en toda la provincia de Santa Fe. A partir de estas bases se generan las bases que se utilizan en las distintas secciones de esta tesis, utilizando las variables latitud, longitud, máximo nivel de instrucción alcanzado, asistencia escolar, sexo, edad, condición de actividad, relación de parentesco con el jefe o jefa del hogar de la base de personas. De la base de hogares se utiliza hacinamiento por habitación y necesidades básicas insatisfechas total, y de la base de viviendas tipo de vivienda, si está habitada, revestimiento del suelo, si posee agua corriente y si posee sanitarios con cisterna de agua.

5 Resultados

En este Capítulo se presentan los resultados obtenidos de las pruebas de diseño realizadas, basadas en la MMUVRA de la cual se seleccionan las encuestas a hogares que desarrolla el INDEC. En la misma, la provincia se toma como estratificación geográfica. En cada una de las provincias se seleccionan muestras en dos etapas, considerando en la primera de ellas la selección de localidades. En la segunda etapa, se seleccionan radios o conjuntos de radios denominados áreas dentro de cada una de las localidades seleccionadas. En ambas etapas se estratifican las unidades de muestreo en base al nivel educativo del jefe/a de hogar (JPI) y selección PPT al total de viviendas particulares en el Censo 2010.

[8] Disponible en https://www.indec.gob.ar/ftp/cuadros/poblacion/cuestionario_basico_2010.pdf.

[9] INDEC, 2015. Unidades Geoestadísticas. Cartografía y códigos geográficos del Sistema Estadístico Nacional. Buenos Aires: INDEC, [Consulta: 14/11/2022]. Disponible en http://www.indec.gov.ar/codgeo.asp.

Se plantea el análisis de la provincia de Santa Fe a partir de las bases del Censo 2010 presentadas anteriormente, en el Capítulo 4. Se traslada la estructura muestral de INDEC aplicada en el país, a una estructura más desagregada: la provincia. Entonces, en vez de tomar como estratificación geográfica a las provincias, se toma como estratos los siguientes dominios de estimación constituidos por las principales ciudades de la provincia consideradas (Rafaela, Reconquista, Rosario, Santa Fe y Venado Tuerto), y un estrato, constituido por los radios con más de 10 viviendas del resto de localidades de la provincia (Resto provincial) que resulta de excluir éstas 5 localidades. Esto implicara obtener estimaciones de cada uno de los dominios por separado, como también al juntarlos permitiría alcanzar una estimación provincial. En cada uno de los dominios de estimación se seleccionan muestras probabilísticas; en una primera etapa se seleccionan radios o conjuntos de radios y luego dentro de los mismos, viviendas habitables[10].

Se plantean diferentes interrogantes con respecto al diseño de muestreo de la primera etapa tanto para la formación de conglomerados (por agrupamiento de radios) como en la formación de estratos. En principio en la Sección 5.1 se presentan los resultados de los siguientes interrogantes sobre la agrupación de radios censales o formación de conglomerados:

¿Qué unidad usar en la primera etapa de muestreo, UPM, dentro de los estratos geográficos o dominios de estimación?

- Radios.
- Conjunto de 2 radios.
- Conjunto de 3 radios.

¿Con que criterio unir radios?

- Unión consecutiva, por definición de INDEC.
- Método de Elliott, unión por cercanía geográfica en grupos heterogéneos de igual tamaño.

¿Cómo seleccionar esas unidades?

- MSA.
- PPT.

Luego de formado los conglomerados, se responde en la Sección 5.2 las siguientes cuestiones acerca de construcción de estratos de las UPM (conglomerados):

¿Qué características deben utilizarse para la estratificación?

- Criterio de INDEC, nivel educativo del jefe/a de hogar (JPI).
- Indicador multivariado.

¿Cuántos estratos deben crearse?

- Tres.
- Cuatro.
- Cinco.

¿De qué forma deben determinarse los límites entre estratos?

[10] Se define como viviendas habitables, a todas aquellas viviendas que al momento del censo están habitadas o son potencialmente habitables, incluye las categorías de viviendas con personas presentes y ausentes, como también las viviendas en venta/alquiler y en construcción.

- Lavallée-Hidiroglou, para poblaciones asimétricas.
- Método de Elliott, unión por cercanía geográfica en grupos homogéneos de igual tamaño.

5.1 Conglomerados

En esta Sección se plantea la selección del diseño con respecto a la formación de conglomerados en la UPM. En el caso de la provincia de Santa Fe las UPM resultan ser los radios o conjunto de los mismos, dentro de cada uno de los 6 dominios de estimación, es decir, las 5 principales ciudades y el resto provincial. En principio, en la Sección 5.1.1, se presentan los diferentes diseños planteados: cantidad de radios, método de unión y de selección, que combinados entre sí determinan distintas maneras de formar conglomerados. En la Sección siguiente, 5.1.2, se evalúan los diseños planteados a partir de analizar la efectividad de las tres características deseadas en los conglomerados: eficiencia por precisión y por igual tamaño, y eficacia por cercanía entre radios. Por último, en la Sección 5.1.3 se determina cuál de los conglomerados construidos presenta la mejor efectividad. Tomando un diseño efectivo según Drucker[11] (Drucker, 1989) donde se busca el punto de equilibrio ideal entre el diseño más eficiente y el más eficaz.

5.1.1 Formación de Conglomerados

Se plantean las siguientes UPM, tomando la unidad como la división censal de radios, como también la unión contigua de los mismos para evaluar si es conveniente trabajar con radios o áreas más grandes:

- Radios (R), división censal: conjunto de 300 viviendas en promedio.
- 2R, unión de 2 radios.
- 3R, unión de 3 radios.

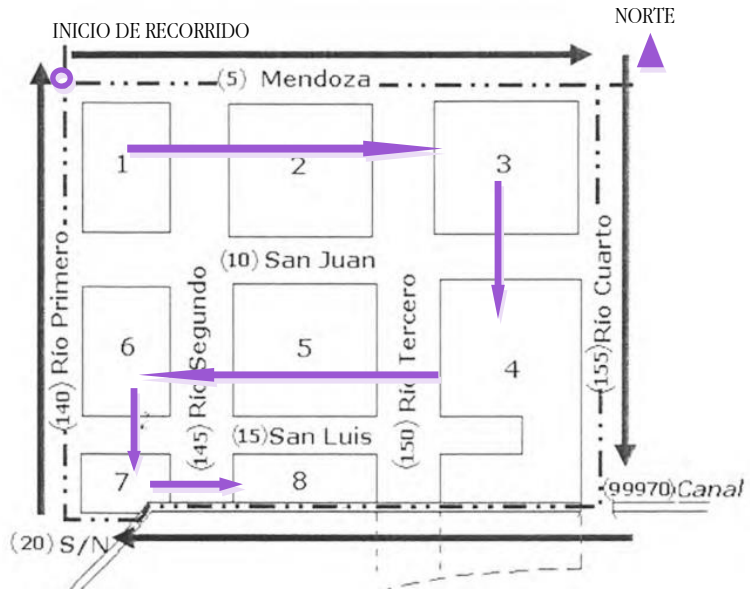
Con respecto a la unión de radios[12] se plantean dos opciones. Por un lado, se forman los conjuntos de radios mediante la unión consecutiva (UC) de acuerdo con su numeración censal. La numeración censal se inicia por la unidad ubicada en el extremo noroeste del área asignada, se continúa con la próxima inmediata, siguiendo el recorrido en forma de "serpentina" de Oeste a Este y de Norte a Sur hasta completar la numeración de todas las unidades en el área. Dicha numeración se aplica en las fracciones por departamento, en los radios dentro de cada fracción y en las manzanas dentro de cada radio (ver procedimiento

[11] Eficiencia: **hacer bien las cosas**. Es decir, realizar una tarea buscando la mejor relación posible entre los recursos empleados y los resultados obtenidos. La eficiencia tiene que ver con el «cómo».
Eficacia: **hacer las cosas correctas**. Es decir, llevar a cabo tareas de la mejor manera, que conduzcan a la consecución de los resultados. Tiene que ver con «qué» cosas se hacen. Eficacia es hacer lo necesario para alcanzar o lograr los objetivos deseados o propuestos.
Efectividad: **hacer bien las cosas correctas**. Es decir, que las tareas que se lleven a cabo se realicen de manera eficiente y eficaz. Tiene que ver con «qué» cosas se hacen y «cómo».
La efectividad que plantea Drucker, **quiere encontrar un término medio con sentido, un punto de equilibrio ideal entre eficiencia y eficacia**. La búsqueda de la efectividad impide que un exceso de foco en la eficiencia provoque que no se alcance el resultado deseado o que no se haga a tiempo. De la misma manera, impide que un exceso de foco en la eficacia pueda dar con el punto de equilibrio al traste con la rentabilidad del resultado, descompensando el esfuerzo. Cuando trabajamos de forma efectiva, el foco se mantiene en el punto óptimo entre lo más eficiente y lo más eficaz.

[12] Los radios con 500 viviendas o más se consideran en forma independiente.

computacional en Anexo A.3). A continuación, se presenta en la Figura 2 el diagrama de numeración censal con un ejemplo de manzanas dentro de radios (INDEC, 2010).

Figura 2: Diagrama de numeración censal. Censo 2010



Por otro lado, se juntan los radios por cercanía geográfica en grupos heterogéneos de igual tamaño mediante el algoritmo propuesto por Elliott. El método presenta requisitos de posicionamiento geoespacial de las unidades dado por latitud y longitud, de la cantidad de grupos a formar, K, de una medida de tamaño y de variables con información de interés de las unidades para armar los grupos. Para medir la cercanía se utiliza la latitud y longitud de los centroides de los radios. Se agrupan los radios de a dos (2R) o de a tres (3R) asignando a K como el número entero que resulta de dividir por 2 o 3 la cantidad de radios de cada dominio de estimación respectivamente. Se toma la cantidad de viviendas habitables por radio como la medida de tamaño a igualar por grupo, y a las siguientes variables (seleccionadas en pruebas previas, mencionadas más adelante) para generar los grupos heterogéneos: proporción de desocupados/as, proporción de mujeres económicamente activas y de jefes o jefas de hogar con secundario incompleto por radio.

Cabe destacar que el método de Elliott es un algoritmo originado para agrupar cualquier número de unidades. Si bien el autor presenta una aplicación con $r = 105$ unidades, agrupadas en $K = 10\ a\ 15$ estratos, la cual se ejecuta en pocos minutos, a la hora de aplicar el mismo en localidades grandes, como Rosario con $r = 1.066$ radios a agrupar en $K = 533(2R)$ ó $355(3R)$ conglomerados, el algoritmo no converge luego de más de 30 horas. Por lo tanto, se adaptó el mismo para aliviar la carga de las iteraciones. Además, se utilizó una primera subdivisión en $K = 5$ grupos para la ciudad de Rosario, y dentro de cada subdivisión se agruparon a su vez en 2 y en 3 radios, acortando el tiempo de procesamiento a 2 horas aproximadamente en cada ejecución del algoritmo (ver procedimiento computacional en Anexo A.3). Para la ciudad de Santa Fe con 398 radios, se realizó una subdivisión de $K = 2$ grupos y para el Resto provincial con 1.845 radios, se tomó a cada uno de los 19 departamentos como subdivisiones.

Se realizaron pruebas previas con el algoritmo de Elliott, para seleccionar variables que formaran grupos con características deseables en el sentido de formar conglomerados más heterogéneos y que resultaran geográficamente válidas en términos operativos de cercanía. Se probaron con todas las combinaciones de las

siguientes variables: proporción de PEA, tasa de desocupación, proporción de PEA de mujeres, proporción de jefes o jefas de hogar con primario incompleto y proporción de jefes o jefas de hogar con secundario incompleto. Las mismas resultan seleccionadas a su vez, por explicar la mayor variabilidad de los datos (45%) según las tres primeras dimensiones del análisis de correspondencias múltiples realizado sobre la base de datos del Censo 2010 por radio, utilizando el paquete *FactoMineR* del software R (ver procedimiento computacional en Anexo A.3).

Se toma como medida de evaluación de grupos heterogéneos al valor $SCW = \hat{\sigma}^2$ obtenido de la ejecución del algoritmo de Elliott, al realizar el análisis de variancia utilizando los agrupamientos como factores y cada una de las variables estudiadas como explicativas (ver Tabla 5.1). Además, se considera como medida geográfica válida al análisis gráfico de los centroides de los radios censales identificados según grupos por cada combinación de las variables estudiadas, (ver Figura 5.2 y Figuras A.1 a A.6 en Anexo A.1). Se considera aquella combinación de variables que presente el máximo valor entre los SCW y que genere a su vez el gráfico de centroides por grupos más compactos.

Para reducir los tiempos de procesamiento en esta instancia previa, se aplicó el algoritmo en la ciudad de Rosario, tomando $K = 5$ grupos. Luego se aplica el resultado obtenido en el resto de los dominios en estudio, bajo el supuesto que las variables seleccionadas para Rosario son válidas para el resto de los dominios. Por lo tanto, en base a la máxima medida de heterogeneidad según Tabla 5.1 y al gráfico de centroides según grupos más compacto (Figura 5.2), el conjunto de variables que se van a utilizar para ejecutar el algoritmo de Elliott es: proporción de PEA mujer, proporción de desocupados y proporción de jefes o jefas de hogar con secundario incompleto, las cuales se utilizarán en cada dominio de estimación.

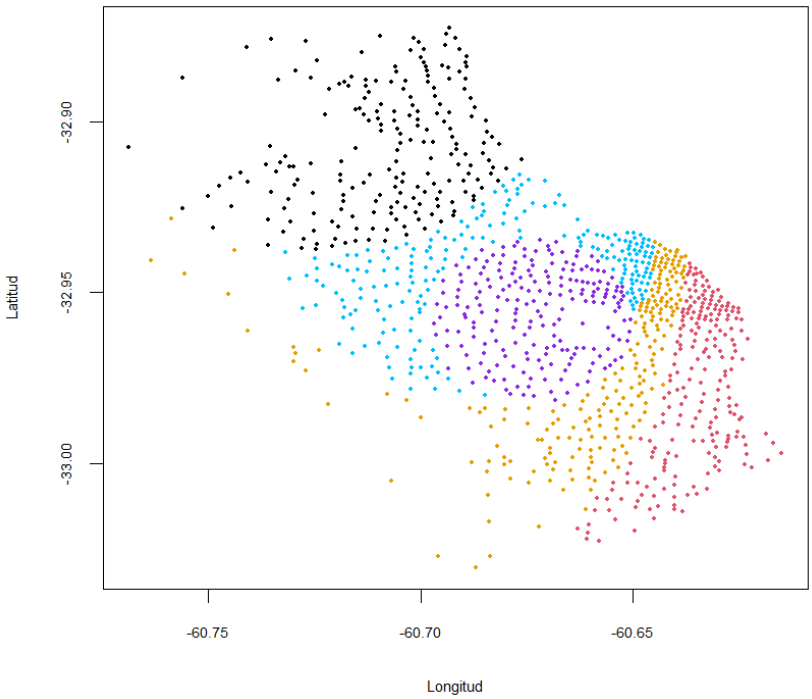
Por último, con respecto al método de selección, en la primera etapa se consideran dos tipos de muestreo: MSA sin reemplazo y muestreo PPT de viviendas habitables con[13] reemplazo. Mientras que en la segunda etapa se seleccionan hogares como unidad secundaria de muestreo, dentro de cada UPM seleccionada, a través de una MSA.

Tabla 5.1: Medidas de máxima heterogeneidad de las agrupaciones aplicando el algoritmo Elliott, según los escenarios estudiados. Rosario. Censo 2010

Escenarios (Variables estudiadas)	Máxima SCW
Tasa de desocupación	0,025
Proporción de población económicamente activa (PEA)	0,062
Proporción de PEA mujer	0,080
Proporción de PEA mujer, Tasa desocupado	0,084
Proporción de jefes/as con secundario incompleto	0,248
Proporción de PEA, de desocupado, y de jefes/as con secundario incompleto	0,257
Proporción de PEA mujer, de desocupado, y jefes/as con secundario incompleto	0,262

[13] Los diseños con reemplazo no se utilizan generalmente en la práctica, para evitar que las mismas unidades sean seleccionadas más de una vez en la muestra. Sin embargo, dichos diseños poseen fórmulas de variancia simples que bajo ciertas condiciones pueden servir de aproximaciones para diseños sin reemplazo.

Figura 5.2: *Centroides de los radios censales, identificados en K=5 grupos, según el algoritmo de Elliott mediante tasa de desocupación, proporción de PEA de mujeres y de jefes o jefas de hogar con hasta secundario incompleto por radio. Rosario. Año 2010.*



La cantidad de radios censales del 2010, varía según los dominios de estimación planteados; la ciudad de Reconquista, Venado Tuerto y Rafaela presentan menos de 100 radios cada una, le sigue la ciudad de Santa Fe con 398 radios y luego se encuentran Rosario y Resto provincial contando con una gran cantidad de 1.066 y 1.845 radios respectivamente, sumando un total de 3.532 radios en la provincia de Santa Fe. Entonces, una vez aplicados los distintos diseños planteados se obtienen, para los mismos radios, las distintas combinaciones de formación de conglomerados. Por lo tanto, se presenta en la Tabla 5.2 la cantidad resultante de UPM para cada método de agrupamiento en los dominios de estimación.

Tabla 5.2: Cantidad de UPM según método de agrupamiento por cantidad de radios, para cada dominio de estimación. Provincia de Santa Fe. Año 2010

	Método de agrupamiento (cantidad de radios)				
Número de UPM	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
Rafaela	85	45	43	32	31
Reconquista	60	33	33	24	25
Rosario	1066	551	554	380	375
Santa Fe	398	209	203	146	140
Venado	78	42	45	30	31
Resto Provincial	1845	967	1003	674	696

Notar que la cantidad de UPM de ambos tipos de agrupamientos, ya sea de a 2 como de a 3 radios, no es exactamente la cantidad de radios dividido dos o tres, ya que los radios de 500 viviendas habitables o más no se han considerado en la unión. Además, para el método de Elliott se agrega la condición de igual tamaño por UPM, lo que implica que podría llegar a variar la cantidad de radios dentro de las UPM si los mismos no cumplen con el mínimo de viviendas requeridas, que el algoritmo genera automáticamente en función de las características de los propios datos (para más información ver paso 3 del algoritmo en Metodología, Sección

3.2.1). Sin embargo, a pesar de dichos puntos, la cantidad de unidades en la primera etapa de muestreo según los métodos de agrupamiento resultan similares, por ejemplo, en la ciudad de Rosario los 1.066 radios, quedaron con 551 UPM agrupadas de a 2 radios por unión consecutiva y con 554 UPM agrupadas por Elliott, quedando para cada diseño un total de conglomerados diferente, por ejemplo, para el agrupamiento de a 2 radios según Elliott se formarían en total 1881 conglomerados o UPM para la Muestra Maestra de la provincia de Santa Fe.

5.1.2 Evaluación de los Conglomerados

Para evaluar la eficiencia de los diseños planteados, en el apartado 5.1.2.1 se compara la precisión de las estimaciones mediante el efecto de diseño (deff). Bajo el supuesto que se toman m de las M UPMs, $\bar{n} = 10$ hogares en cada UPM seleccionada, y que las fracciones de muestreo de las UPM (m/M) y dentro de las UPM ($\bar{n}/N_i$) son pequeñas, se calcula el efecto de diseño a partir de las ecuaciones (3.5) y (3.9) según el diseño considerado, para cada una de las distintas definiciones de UPM respectivamente, correspondiente a cada estimador:

$$deff = k[1 + \delta(\bar{n} - 1)]. \tag{5.1}$$

Se considera la estimación de la proporción de PEA, de desocupación, de la población por sexo; y la de la población por sexo y edad, utilizando el paquete *PracTools* del software R (ver procedimiento computacional en Anexo A.3). Estos estimadores están asociados a la encuesta socio-económica planteada dentro del programa de encuestas sugerido (ver Capítulo 1).

Para evaluar la eficacia de los diseños se comparan tanto la igualdad de tamaño de los grupos, que se presenta en el apartado 5.1.2.2, medida a través de la cantidad de viviendas habitables por UPM, como la cercanía de las mismas, que se presenta en el apartado 5.1.2.3, evaluada a partir de la distancia entre radios dentro de las UPM. Notar que las viviendas por grupo se vinculan, por un lado, al deff, ya que k depende del tamaño de los grupos y por otro a cuestiones operativas (por ejemplo, a la hora de listar grupos y mantener actualizados los mismos). Las distancias se relacionan también con la operatividad del trabajo de campo y los recursos de traslado del personal, entre otros.

5.1.2.1 Efecto diseño: deff

En esta Sección se analiza el efecto de diseño en los Gráficos 5.1 a 5.6 y las Tablas 5.3 a 5.8, para los 6 dominios de estimación: Rafaela, Reconquista, Rosario, Santa Fe, Venado Tuerto y Resto provincial respectivamente. Se toma como medida de precisión el efecto diseño de los estimadores de las proporciones y de las tasas mencionadas en la Sección anterior, calculado para cada una de las cinco uniones de radios planteadas y según el método de selección: con igual probabilidad de inclusión para las unidades de primera etapa utilizando MSA o un muestreo con probabilidades de inclusión distintas considerando un PPT con reemplazo de viviendas habitables. Las uniones de radios consideradas surgen de la combinación entre el tipo de agrupamiento, UC y unión por cercanía geográfica en grupos heterogéneos de igual tamaño (Elliott), y la cantidad de radios agrupados: dos radios, tres radios, o sin agrupación que es el caso de los radios individuales (R). En el final del apartado se encuentra el análisis de la eficiencia, presentado aquí, para todos los dominios de estimación en general.

Rafaela

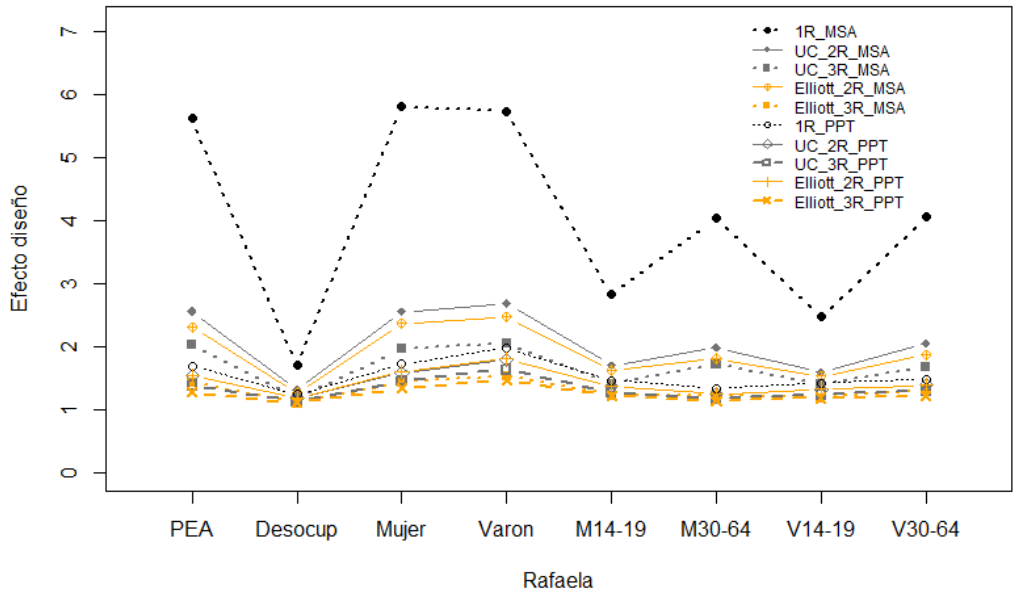
Se presenta en la Tabla 5.3 y el Gráfico 5.1 los efectos de diseño de los estimadores variando UPM, combinando método de agrupamiento y tipo de selección para la ciudad de Rafaela.

Tabla 5.3: Efecto de diseño de los estimadores en estudio, según método de agrupamiento por cantidad de radios y tipo de selección en la primera etapa de muestreo. Rafaela. Año 2010

Rafaela	Método de agrupamiento (cantidad de radios)				
Selección MSA/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	5,62	2,56	2,31	2,03	1,42
Desocupados/as	1,70	1,31	1,27	1,19	1,10
Mujer	5,80	2,55	2,36	1,96	1,44
Varón	5,73	2,68	2,47	2,04	1,53
Mujer 14-29 años	2,83	1,70	1,62	1,41	1,22
Mujer 30-64 años	4,03	1,98	1,81	1,72	1,22
Varón 14-29 años	2,47	1,60	1,53	1,39	1,18
Varón 30-64 años	4,05	2,04	1,86	1,68	1,31

Rafaela	Método de agrupamiento (cantidad de radios)				
Selección PPT/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	1,69	1,53	1,54	1,38	1,28
Desocupados/as	1,24	1,19	1,18	1,14	1,11
Mujer	1,72	1,58	1,60	1,45	1,35
Varón	1,98	1,79	1,81	1,64	1,46
Mujer 14-29 años	1,46	1,36	1,37	1,27	1,22
Mujer 30-64 años	1,33	1,25	1,25	1,18	1,14
Varón 14-29 años	1,42	1,31	1,32	1,24	1,18
Varón 30-64 años	1,48	1,39	1,39	1,29	1,21

Gráfico 5.1: Efecto de diseño de los estimadores en estudio (perfiles) por cada método de agrupamiento según cantidad de radios y tipo de selección en la primera etapa de muestreo. Rafaela. Año 2010



Reconquista

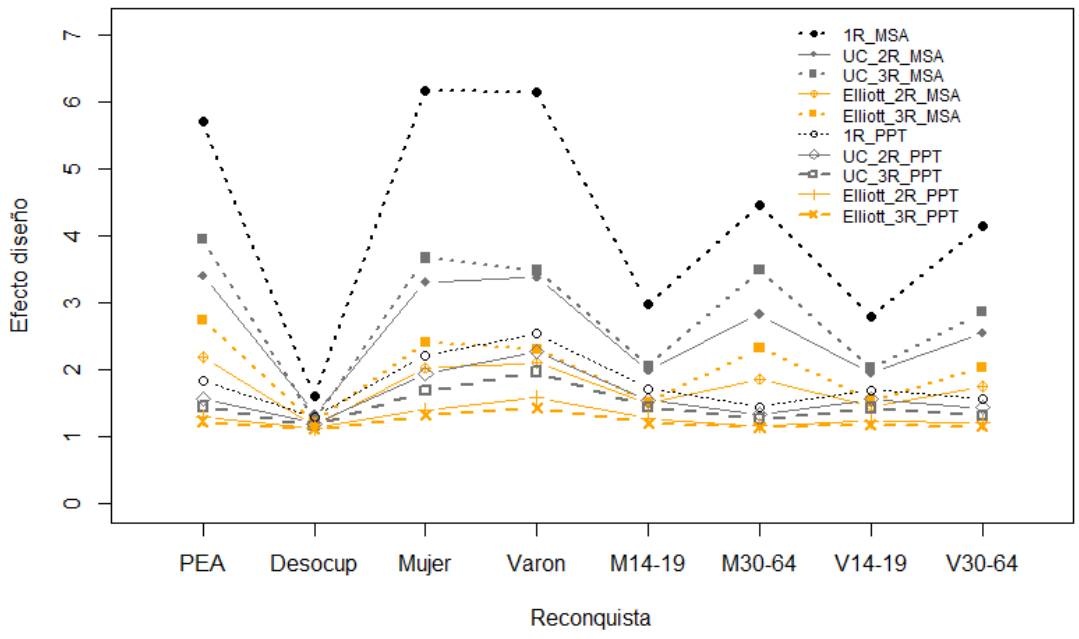
A continuación, en la Tabla 5.4 y el Gráfico 5.2 se presenta la comparación de la precisión de los estimadores en estudio a través del efecto de diseño planteado para la ciudad de Reconquista.

Tabla 5.4: Efecto de diseño de los estimadores en estudio, según método de agrupamiento por cantidad de radios y tipo de selección en la primera etapa de muestreo. Reconquista. Año 2010

Reconquista Selección MSA/MSA	Método de agrupamiento (cantidad de radios)				
	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	5,71	3,40	2,19	3,94	2,73
Desocupados/as	1,60	1,33	1,17	1,23	1,22
Mujer	6,16	3,31	2,02	3,66	2,40
Varón	6,13	3,38	2,10	3,47	2,30
Mujer 14-29 años	2,97	1,98	1,49	2,05	1,51
Mujer 30-64 años	4,46	2,83	1,86	3,49	2,32
Varón 14-29 años	2,79	1,95	1,45	2,02	1,53
Varón 30-64 años	4,14	2,54	1,75	2,86	2,02

Reconquista Selección PPT/MSA	Método de agrupamiento (cantidad de radios)				
	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	1,84	1,57	1,28	1,45	1,23
Desocupados/as	1,29	1,19	1,13	1,16	1,11
Mujer	2,21	1,93	1,40	1,68	1,32
Varón	2,53	2,27	1,58	1,96	1,43
Mujer 14-29 años	1,71	1,54	1,26	1,43	1,19
Mujer 30-64 años	1,44	1,33	1,16	1,26	1,13
Varón 14-29 años	1,69	1,55	1,23	1,42	1,17
Varón 30-64 años	1,57	1,43	1,20	1,31	1,15

Gráfico 5.2: Efecto de diseño de los estimadores en estudio (perfiles) por cada método de agrupamiento según cantidad de radios y tipo de selección en la primera etapa de muestreo. Reconquista. Año 2010



Rosario

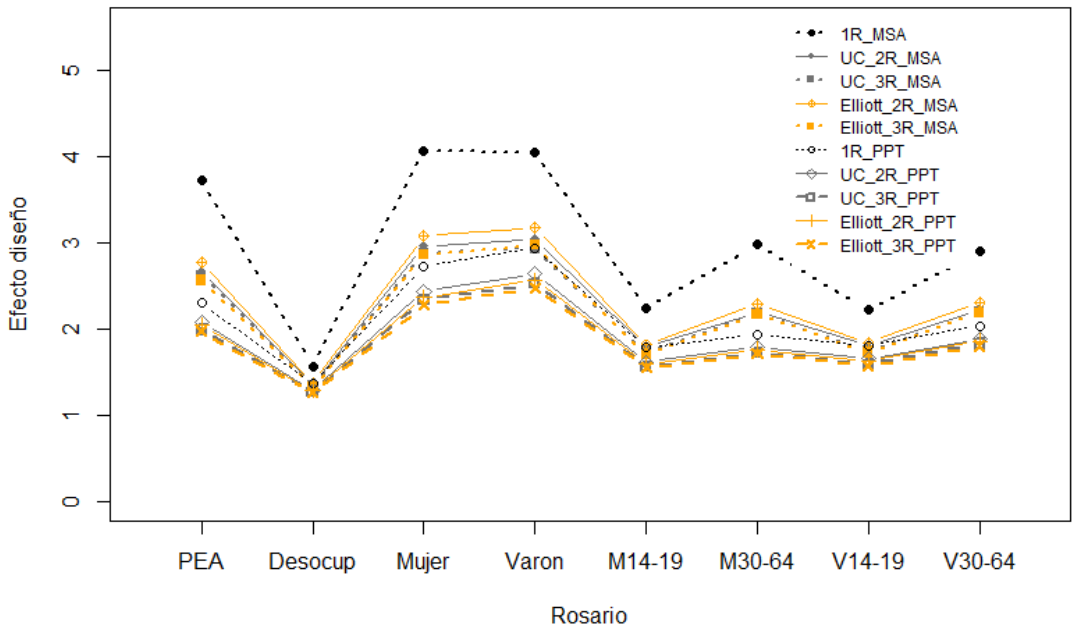
En la Tabla 5.5 y el Gráfico 5.3 se aprecia el efecto de diseño de los estimadores en estudio de la ciudad de Rosario, para cada diseño planteado según método de agrupamiento y de selección en estudio.

Tabla 5.5: Efecto de diseño de los estimadores en estudio, según método de agrupamiento por cantidad de radios y tipo de selección en la primera etapa de muestreo. Rosario. Año 2010

Rosario	Método de agrupamiento (cantidad de radios)				
Selección MSA/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	3,72	2,67	2,77	2,62	2,56
Desocupados/as	1,57	1,36	1,37	1,34	1,32
Mujer	4,07	2,96	3,08	2,88	2,86
Varón	4,04	3,05	3,18	2,93	2,97
Mujer 14-29 años	2,24	1,79	1,81	1,72	1,70
Mujer 30-64 años	2,98	2,20	2,29	2,18	2,17
Varón 14-29 años	2,22	1,80	1,84	1,74	1,73
Varón 30-64 años	2,90	2,23	2,31	2,20	2,19

Rosario	Método de agrupamiento (cantidad de radios)				
Selección PPT/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	2,30	2,08	2,05	1,99	1,97
Desocupados/as	1,36	1,29	1,28	1,27	1,26
Mujer	2,73	2,44	2,37	2,34	2,28
Varón	2,94	2,64	2,57	2,53	2,48
Mujer 14-29 años	1,79	1,62	1,60	1,56	1,55
Mujer 30-64 años	1,94	1,79	1,76	1,72	1,71
Varón 14-29 años	1,81	1,66	1,63	1,60	1,57
Varón 30-64 años	2,04	1,89	1,86	1,83	1,80

Gráfico 5.3: Efecto de diseño de los estimadores en estudio (perfiles) por cada método de agrupamiento según cantidad de radios y tipo de selección en la primera etapa de muestreo. Rosario. Año 2010



Santa Fe

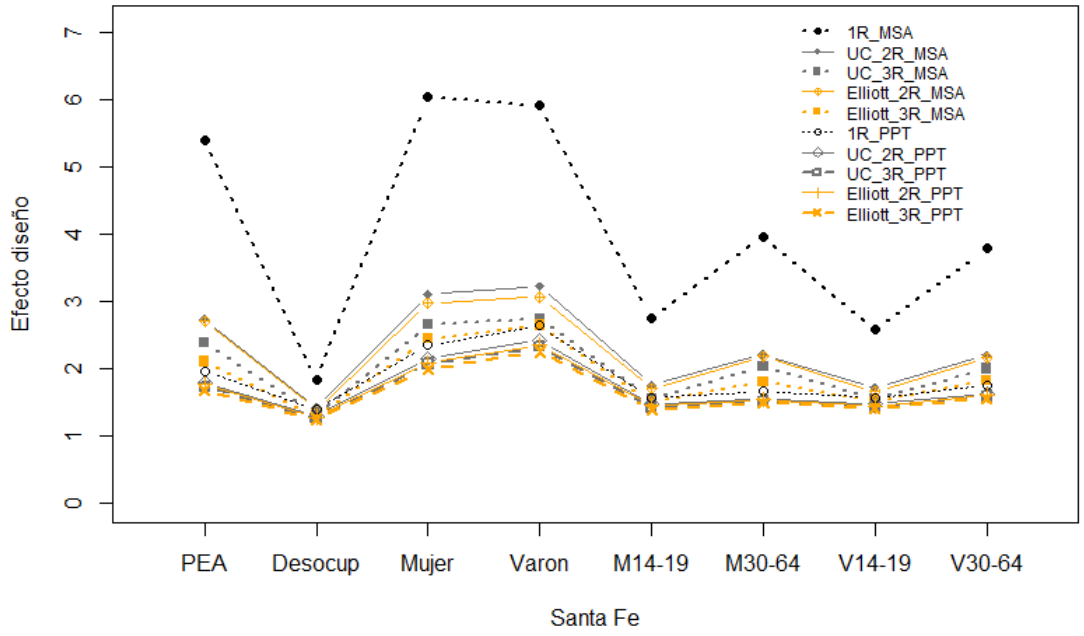
En la ciudad de Santa Fe se observa en la Tabla 5.6 y el Gráfico 5.4 el efecto diseño para los estimadores considerados variando el método de agrupamiento por cantidad de radios y el método de selección.

Tabla 5.6: Efecto de diseño de los estimadores en estudio, según método de agrupamiento por cantidad de radios y tipo de selección en la primera etapa de muestreo. Santa Fe. Año 2010

Santa Fe	Método de agrupamiento (cantidad de radios)				
Selección MSA/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	5,38	2,73	2,70	2,38	2,10
Desocupados/as	1,84	1,41	1,39	1,32	1,29
Mujer	6,03	3,10	2,97	2,65	2,44
Varón	5,90	3,22	3,06	2,73	2,64
Mujer 14-29 años	2,74	1,75	1,70	1,55	1,50
Mujer 30-64 años	3,96	2,21	2,18	2,03	1,80
Varón 14-29 años	2,58	1,72	1,65	1,55	1,50
Varón 30-64 años	3,79	2,20	2,16	2,00	1,83

Santa Fe	Método de agrupamiento (cantidad de radios)				
Selección PPT/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	1,95	1,78	1,76	1,73	1,67
Desocupados/as	1,39	1,30	1,28	1,26	1,25
Mujer	2,35	2,14	2,08	2,07	1,99
Varón	2,64	2,42	2,34	2,33	2,25
Mujer 14-29 años	1,57	1,47	1,45	1,42	1,39
Mujer 30-64 años	1,66	1,55	1,54	1,52	1,48
Varón 14-29 años	1,56	1,47	1,45	1,42	1,40
Varón 30-64 años	1,74	1,63	1,59	1,59	1,55

Gráfico 5.4: Efecto de diseño de los estimadores en estudio (perfiles) por cada método de agrupamiento según cantidad de radios y tipo de selección en la primera etapa de muestreo. Santa Fe. Año 2010



Venado Tuerto

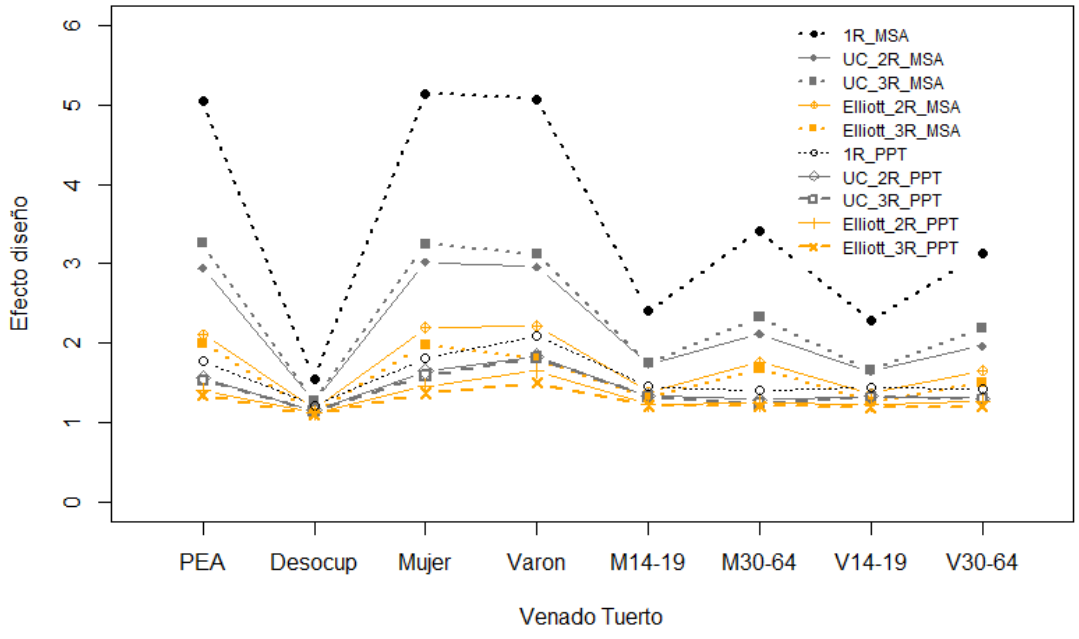
Se exhibe en la Tabla 5.7 y el Gráfico 5.5 el efecto diseño para todos los estimadores planteados de la ciudad de Venado Tuerto.

Tabla 5.7: Efecto de diseño de los estimadores en estudio, según método de agrupamiento por cantidad de radios y tipo de selección en la primera etapa de muestreo. Venado Tuerto. Año 2010

Venado Tuerto	Método de agrupamiento (cantidad de radios)				
Selección MSA/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	5,04	2,94	2,11	3,26	2,00
Desocupados/as	1,54	1,27	1,17	1,27	1,12
Mujer	5,14	3,02	2,20	3,25	1,98
Varón	5,07	2,96	2,22	3,12	1,80
Mujer 14-29 años	2,40	1,74	1,38	1,75	1,31
Mujer 30-64 años	3,42	2,12	1,76	2,33	1,68
Varón 14-29 años	2,28	1,64	1,36	1,66	1,26
Varón 30-64 años	3,13	1,97	1,66	2,19	1,50

Venado Tuerto	Método de agrupamiento (cantidad de radios)				
Selección PPT/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	1,77	1,57	1,40	1,53	1,34
Desocupados/as	1,22	1,16	1,12	1,15	1,10
Mujer	1,80	1,64	1,46	1,59	1,37
Varón	2,09	1,84	1,65	1,82	1,50
Mujer 14-29 años	1,46	1,34	1,23	1,33	1,21
Mujer 30-64 años	1,40	1,29	1,24	1,24	1,21
Varón 14-29 años	1,45	1,33	1,23	1,32	1,20
Varón 30-64 años	1,42	1,31	1,27	1,30	1,20

Gráfico 5.5: Efecto de diseño de los estimadores en estudio (perfiles) por cada método de agrupamiento según cantidad de radios y tipo de selección en la primera etapa de muestreo. Venado Tuerto. Año 2010



Resto Provincial

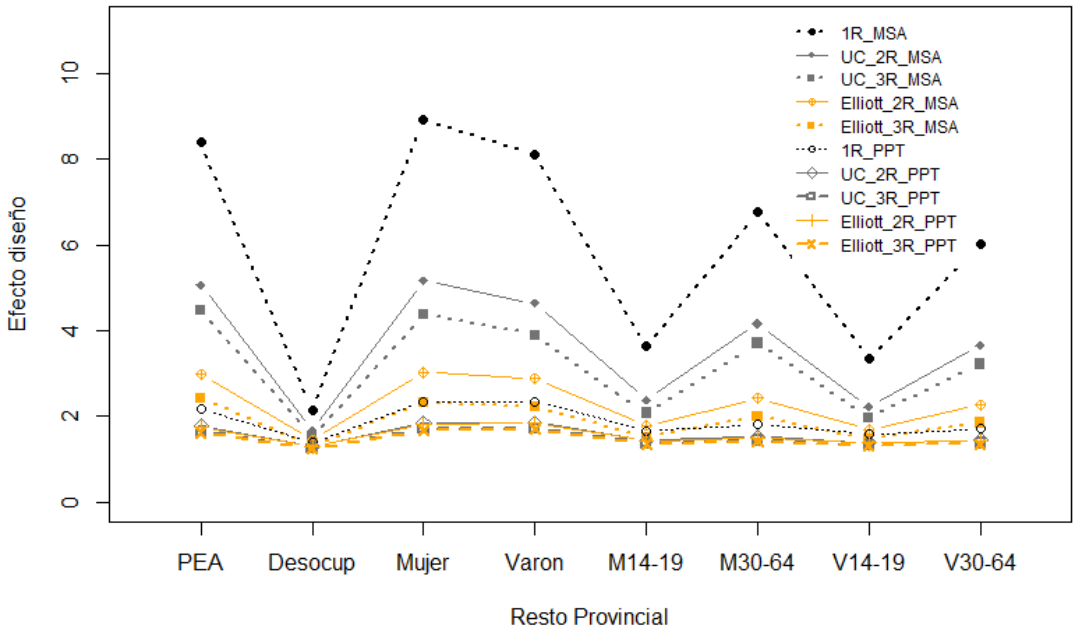
Por último, se presenta la comparación del efecto diseño de los estimadores estudiados en el resto de la provincia de Santa Fe en la Tabla 5.8 y el Gráfico 5.6.

Tabla 5.8: Efecto de diseño de los estimadores en estudio, según método de agrupamiento por cantidad de radios y tipo de selección en la primera etapa de muestreo. Resto Provincial. Año 2010

Resto Provincial	Método de agrupamiento (cantidad de radios)				
Selección MSA/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	8,38	5,06	2,99	4,46	2,42
Desocupados/as	2,14	1,66	1,45	1,54	1,32
Mujer	8,92	5,16	3,03	4,37	2,33
Varón	8,10	4,64	2,89	3,90	2,24
Mujer 14-29 años	3,65	2,38	1,79	2,08	1,52
Mujer 30-64 años	6,76	4,16	2,43	3,71	2,00
Varón 14-29 años	3,36	2,22	1,70	1,96	1,47
Varón 30-64 años	6,02	3,66	2,26	3,22	1,86

Resto Provincial	Método de agrupamiento (cantidad de radios)				
Selección PPT/MSA	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
PEA	2,17	1,77	1,75	1,67	1,63
Desocupados/as	1,39	1,30	1,30	1,26	1,26
Mujer	2,34	1,86	1,83	1,72	1,68
Varón	2,35	1,86	1,84	1,73	1,70
Mujer 14-29 años	1,66	1,45	1,44	1,37	1,36
Mujer 30-64 años	1,81	1,53	1,50	1,45	1,41
Varón 14-29 años	1,59	1,39	1,38	1,33	1,32
Varón 30-64 años	1,71	1,45	1,43	1,38	1,36

Gráfico 5.6: Efecto de diseño de los estimadores en estudio (perfiles) por cada método de agrupamiento según cantidad de radios y tipo de selección en la primera etapa de muestreo. Resto Provincial. Año 2010



Análisis General

Según las tablas de efecto de diseño presentadas, el mínimo valor de deff por dominio de estimación resulta de 1,10 en la ciudad de Rafaela, Reconquista y Venado Tuerto (Tablas 5.3, 5.5 y 5.7 respectivamente), mientras que Rosario, Santa Fe y Resto presentan el mínimo deff alrededor de 1,26 (Tablas 5.5, 5.6 y 5.8 respectivamente). Con respecto al máximo valor de deff se presentan en orden creciente: Rosario con deff de 4,07; Venado Tuerto de 5,14; Rafaela de 5,80; Santa Fe de 6,03; Reconquista de 6,16 y por último el Resto provincial con deff de 8,92 (Tablas 5.5, 5.7, 5.3, 5.6, 5.4 y 5.8 respectivamente).

En relación con la precisión de las estimaciones en estudio para cada diseño planteado (perfiles) se observan comportamientos similares en la mayoría de los dominios de estimación, con pocas excepciones. En general para todos los dominios de estimación, la diferencia más clara se observa en los valores de deff pequeños, es decir con mejor precisión, para los perfiles considerando la selección PPT en vez de MSA en la primera etapa de muestreo. El único caso contrario se observa en la ciudad de Rafaela, que presenta una leve diferencia en el mínimo deff: 1,11 vs 1,10 (Tabla 5.3).

Comparando los deff con respecto a la cantidad de radios que conforman las UPM se obtiene para todos los dominios de estimación que la peor precisión la presentan los perfiles para las UPM tomadas de 1 radio. Las UPM con 2 radios presentan valores de deff intermedios, mientras que las UPM con 3 radios presentan la mejor precisión. La excepción se presenta en Reconquista, Venado Tuerto y el Resto provincial (sólo para MSA) que presentan deff más pequeños tomando de a 2 radios según Elliott respecto a 3 radios por UC (Tablas 5.4, 5.7 y 5.8 respectivamente).

Los valores del efecto diseño para la unión consecutiva resultan más grandes que los valores por la agrupación por Elliott, en todos los dominios de estimación, excepto para la agrupación de 2 radios en la ciudad de Rafaela con selección PPT y en Rosario con MSA (Tablas 5.3 y 5.5).

En los Gráficos 5.1 a 5.6, en general se aprecia también para todos los dominios de estimación como los perfiles del efecto de los diseños con MSA se encuentran muy por encima de los perfiles con PPT (más precisos) además que los de MSA presentan mayor separación entre ellos, mientras los diseños con PPT se encuentran mucho más cercanos entre sí.

Cabe destacar que para todos los dominios, tanto en las Tablas 5.3 a 5.8 como en los Gráficos 5.1 a 5.6, si bien se prefiere la selección PPT a la MSA, la agrupación por Elliott a la unión consecutiva y formar UPM de a 3 radios, algunos perfiles con selección PPT: Elliot_3R vs UC_3R vs Elliot_2R, se encuentran muy cercanos por lo que podría considerarse cualesquiera de estos perfiles como semejantes, teniendo en cuenta otros factores a la hora de plantear el diseño sin perder tanta precisión, ya que la diferencia de los efectos de diseños resulta muy similar entre ellos.

Notar que, en todos los dominios de estimación, el perfil con la mejor precisión se observa para el diseño PPT con UPM agrupadas de a 3 radios según algoritmo de Elliott (Elliot_3R_PPT), mientras que la peor precisión la presentan las UPM tomadas de 1 radio y seleccionadas mediante muestra simple al azar sin reemplazo (1R_MSA).

Conclusiones Generales

En síntesis, se observa que, principalmente, se mejora la precisión de las estimaciones cuando el muestreo considerado es un diseño PPT en vez de un MSA. Con respecto a la cantidad de radios considerados a la hora de formar UPM, las de 1 radio son las de peor precisión, luego la unión de 2 radios, mientras que las unidades de mayor tamaño, 3 radios, presentan la mejor precisión. Finalmente, los deff según el método de agrupamiento de radios por Elliott presenta mejor precisión que por unión consecutiva.

Considerando la marcada diferencia que se presenta entre el diseño con probabilidades de inclusión distintas, PPT, para las unidades de primera etapa respecto a la precisión comparado al efecto de diseño con igual probabilidad como el MSA, esta puede deberse a la heterogeneidad entre los tamaños de las unidades, medida que determina la probabilidad de inclusión de las mismas. Por tanto, se presenta a continuación el estudio del tamaño de viviendas habitables de las UPM, medida utilizada para la selección PPT.

5.1.2.2 Tamaño de viviendas por UPM

Continuando con el análisis de eficiencia, se presenta la comparación de la cantidad de viviendas habitables por UPM o del tamaño de viviendas por UPM para evaluar la heterogeneidad de las unidades respecto al tamaño. Cuando se forman conglomerados del mismo tamaño el diseño resulta más preciso que con tamaños diferentes (Ver Sección 3.2). En el Gráfico 5.7 se muestra la distribución del tamaño de viviendas para los 5 tipos de UPM, con sus escalas respectivas y la Tabla 5.9 contiene su coeficiente de variación.

En general, la cantidad de viviendas habitables por UPM formadas por 1 radio presenta una gran amplitud en todos los dominios de estimación, variando desde 23 a 1.260 viviendas por radio. Mientras que el promedio de viviendas por radio en cada dominio de estimación resulta cercano a 300 viviendas, su amplitud es mucho más heterogénea de la que se esperaría al tomar las UPM de 1 radio (según definición de INDEC alrededor de 300 viviendas por radio). Con relación a las UPM con 2 o 3 radios, se observa en el Gráfico 5.7 que los promedios de viviendas resultan cercanos para las UPM con la misma cantidad de radios, y obviamente aumentan conforme aumenta la cantidad de radios agrupados en las UPM.

Además, se aprecia que la dispersión del tamaño de viviendas por UPM según método de agrupamiento es mayor para las UPM por unión consecutiva, seguidas por las UPM con 1 radio. En la mayoría de los dominios de estimación, el menor rango de tamaño lo presenta la unión de radios por Elliott, en particular para las UPM agrupadas de a 2 radios.

Para poder comparar la variabilidad de los tamaños, según viviendas habitables por UPM con distinta cantidad de radios, se presenta el CV de las mismas en Tabla 5.9. Se observa, en dicha tabla, en correspondencia con los deff y el tamaño de UPM por viviendas estudiadas anteriormente, un ordenamiento de todos los dominios de estimación del CV de la cantidad de viviendas habitables por UPM según el método de agrupación, donde la mayor variación la presentan las UPM sin agrupación, luego le sigue la unión consecutiva de radios y la menor variación en tamaño de UPM la presenta la agrupación según el algoritmo de Elliott. A su vez dentro del método de agrupación se ordena la variabilidad promedio de las viviendas por UPM de a 2 radios y luego las de a 3 radios; es decir, el menor CV de tamaño lo presenta la UPM con agrupación de a 3 radios por Elliott.

En general se observa que la distribución más dispersa la presentan las UPM constituidas por 1 radio, ya que se distinguen muchos valores extremos, formando una marcada asimetría en su distribución, mientras que

las UPM con radios agrupados, presentan una distribución de tamaños de viviendas más simétrica. La variación en el tamaño de los conglomerados tiene un efecto en los factores, como δ, necesarios para diseñar una muestra, como también el factor k, los cuales intervienen en el cálculo del efecto diseño, la cual es una medida de la eficiencia.

En síntesis, se observa para todos los dominios de estimación en estudio, valores medios de viviendas por UPM similares cuando se agrupa la misma cantidad de radios, mientras que la unión consecutiva de radios presenta mayor dispersión de tamaño de UPM por cantidad de viviendas que la agrupación por Elliott.

A continuación, se evalúa otro factor: la cercanía de los radios entre UPM, que permite analizar una nueva característica del diseño.

Gráfico 5.7: Cantidad de viviendas habitables según método de agrupamiento por cantidad de radios en la primera etapa de muestreo, para cada dominio de estimación. Provincia de Santa Fe. Año 2010

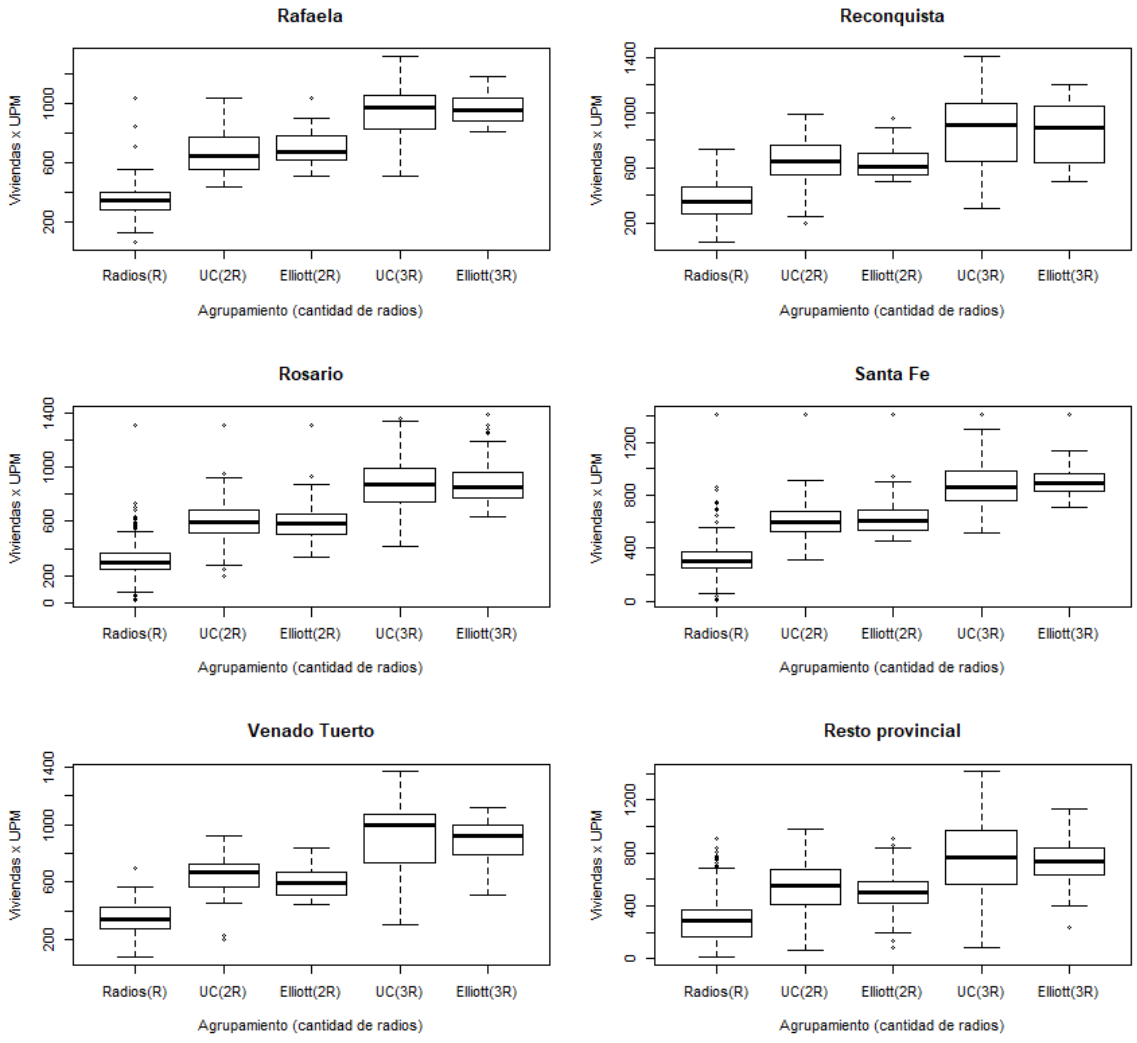


Tabla 5.9: Coeficiente de Variación de la cantidad de viviendas habitables, por método de agrupamiento y cantidad de radios en la primera etapa de muestreo, para cada dominio de estimación. Provincia de Santa Fe. Año 2010

	Método de agrupamiento (cantidad de radios)				
CV	Radios(R)	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
Rafaela	0,39	0,20	0,17	0,21	0,10
Reconquista	0,40	0,27	0,21	0,35	0,12
Rosario	0,33	0,21	0,18	0,21	0,15
Santa Fe	0,39	0,21	0,19	0,21	0,11
Venado Tuerto	0,31	0,22	0,19	0,27	0,20
Resto Provincial	0,52	0,38	0,25	0,36	0,19

5.1.2.3 Cercanía: distancia entre radios por UPM

Se presenta el análisis de la eficacia, para mejorar cuestiones económicas y operativas en el sentido que las UPM estén agrupadas por radios contiguos o muy cercanos entre ellos. Al momento de relevar datos mediante encuestas personales en las UPM se optimizaría el recorrido del encuestador cuanto más concentrada se encuentre su tarea en dicha área de trabajo. Esto reduciría los viajes y el tiempo de traslado para recolectar una cantidad de encuestas en cada área determinada. Se evalúa mediante el estudio de las distancias entre los centroides de los radios, dentro de cada UPM planteada para cada uno de los 6 dominios de estimación. Tomando la distancia euclídea con latitud y longitud de cada centroide, la unidad de medida de distancia se expresa en grados angulares (Ver 3.2.1). En el Gráfico 5.8 se presenta la distribución de las distancias entre radios dentro de cada unidad de muestreo y en la Tabla 5.10 un resumen de la suma total por dominio de estimación de las mismas. Notar que dichas distancias son nulas para las UPM con 1 radio (1R) por lo que no figuran en esta instancia.

La distribución de las distancias en todos los dominios de estimación según el Gráfico 5.8, muestran una menor amplitud para las UPM de a 2 radios con respecto a las UPM de a 3 radios, como es de esperar. El efecto visual de los outliers genera que se vean muy comprimidas las cajas de los boxplots y dificulta la comparación particularmente para 2R; para 3R, sin tener en cuenta los outliers, se observa que Elliott presenta menor amplitud que UC (excepto en Rafaela).

En la Tabla 5.10 se presenta la suma total de las distancias entre los centroides de los radios de cada UPM, para cada dominio de estimación. Se aprecia nuevamente una menor distancia total para las UPM tomadas de a 2 radios. En particular, para las UPM con 2 radios la distancia total mediante la unión consecutiva resulta más pequeña en comparación con la agrupación geoespacial de Elliott (excepto en Rafaela y Santa Fe). Mientras que para la agrupación de a 3 radios, esta relación se invierte, resultando menor la distancia total para el método de agrupación por Elliott (excepto en Reconquista y Venado Tuerto).

Gráfico 5.8: Distancia entre los centroides de los radios dentro de cada UPM en grados angulares, según método de agrupamiento por cantidad de radios en la primera etapa de muestreo, para cada dominio de estimación. Provincia de Santa Fe. Año 2010

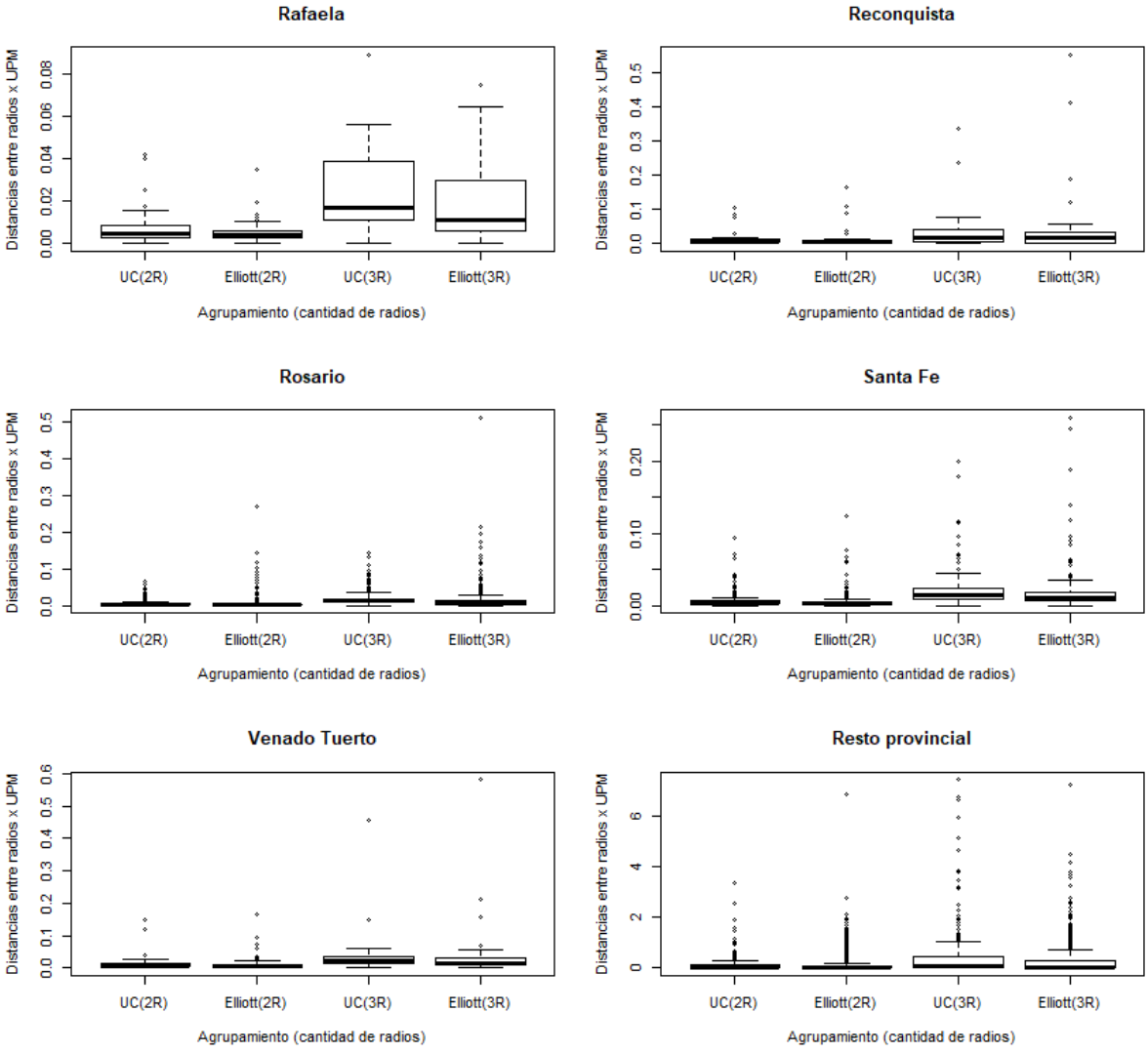


Tabla 5.10: Distancia total entre los centroides de los radios dentro de cada UPM en grados angulares, según método de agrupamiento por cantidad de radios en la primera etapa de muestreo y para cada dominio de estimación. Provincia de Santa Fe. Año 2010

	Método de agrupamiento (cantidad de radios)			
Distancia total	UC(2R)	Elliott(2R)	UC(3R)	Elliott(3R)
Rafaela	0,35	0,24	0,75	0,62
Reconquista	0,44	0,65	1,04	1,53
Rosario	2,89	3,03	6,72	6,17
Santa Fe	1,43	1,30	3,19	2,99
Venado Tuerto	0,63	0,65	1,25	1,45
Resto Provincial	92,13	121,48	228,21	212,56

5.1.3 Elección de Conglomerado

Para determinar la forma de construir conglomerados se busca un diseño efectivo tomando el punto de equilibrio ideal entre el diseño más eficiente (deff e igual tamaño de UPM) y el más eficaz (distancia).

En síntesis, para todos los dominios de estimación si se evalúan todas las componentes de las agrupaciones analizadas según Tabla 5.11, se prefiere el método de selección que considera probabilidades de inclusión distintas comparado a un diseño con igual probabilidad (MSA), ya que brinda más eficiencia respecto a la precisión, en base al efecto de diseño. Con respecto a la cantidad de radios, las unidades de muestreo con 3 radios presentan mejor eficiencia según la precisión en términos del efecto de diseño y de igualdad de tamaño de UPM en relación a la cantidad de viviendas. Las UPM con 2 radios presentan mayor eficacia que las UPM de 3 radios respecto a la cercanía entre radios dentro de la agrupación. Por otro lado, se prefiere el método de agrupación de radios según el algoritmo de Elliott por su mayor eficiencia, en precisión y en igualdad de tamaño de las unidades primarias, mientras que para la eficacia no se presentan preferencias con la agrupación consecutiva de radios. Teniendo en cuenta tanto la eficiencia como la eficacia, la agrupación de 2 radios por Elliott con selección PPT resulta la más efectiva, por lo que ésta se utiliza a partir de aquí para realizar las subsiguientes pruebas de diseño según estratificación y a partir de ahora se lo denomina UPM.

Tabla 5.11: Orden de preferencias de los componentes de las agrupaciones en la primera etapa de muestreo. Todos los dominios de estimación. Provincia de Santa Fe. Año 2010

Componente de UPM Preferido	**Eficiencia** Precisión (deff)	**Eficiencia** Tamaño (Cantidad de viviendas)	**Eficacia** Cercanía de radios (Distancia de UPM)	**Efectividad**
Método de Selección	PPT > MSA	-	-	PPT
Cantidad de Radios	3 radios > 2 radios > 1 radio	3 radios > 2 radios > 1 radio	2 radios > 3 radios	2 radios
Unión de Radios	Elliott > UC	Elliott > UC	Elliott = UC	Elliott

5.2 Estratificación de UPM

El propósito de este apartado es determinar la estrategia a utilizar en la formación de estratos en la primera etapa de muestreo de las UPM elegidas como conglomerados en la Sección anterior (5.1.3). A continuación, se presenta a lo largo de la Sección 5.2.1 los pasos seguidos para llegar a formar los distintos tipos de los estratos, planteando la combinación de: variables de estratificación, método de estratificación y cantidad de estratos. En la Sección 5.2.2 se evalúan los estratos planteados a partir de analizar la eficiencia según la precisión, en base al coeficiente de variación de los parámetros de estudio. Por último, en la Sección 5.2.3 se elige cuál de los tipos de estratos construidos presenta la mejor eficiencia, en los dominios de estimación.

5.2.1 Formación de estratos

Se plantean distintas alternativas para la formación de estratos. Con respecto a las variables de estratificación, se propone en primer lugar, la utilizada en la MMUVRA llevada a cabo por INDEC, basada en el nivel educativo del jefe o jefa de hogar (jefes/as de hogar con primario incompleto, JPI), tomando la proporción por UPM. También se proponen estudiar 10 variables indicadoras de carencias críticas provenientes del Censo 2010, publicados por IPEC[14], para caracterizar las UPM respecto a las condiciones socio-económicas de la población que las habita. Una alternativa para construir estratos con muchas variables que resultan de interés, y estén posiblemente correlacionadas, consiste en tomar las primeras dimensiones relevantes que resultan de aplicar métodos de análisis multivariado de reducción de dimensiones, generando índices independientes con combinaciones de dichas variables como posible clasificación de estratos. En base a los indicadores de carencia se plantea, por un lado, tomar el índice agregado de carencias que resulta de la primera CP robusta como variable de estratificación (Bussi, Marí, & Méndez, 2017). Por otro lado, se forman estratos tomando todas las variables indicadoras de carencias como variables de estratificación utilizando el algoritmo de Elliott[15], ya que también permite generar estratos de igual tamaño y homogéneos dentro.

Debido a la asimetría de las variables de estratificación, JPI e índice multivariado de CP robusta, se plantea estratificar mediante un algoritmo propuesto por Lavallée-Hidiroglou (L-H).

A su vez, se estratifica por distinta cantidad de estratos, considerando 3, 4 y 5 estratos y utilizando la asignación proporcional para adjudicar la muestra de UPM entre los mismos, según el tamaño de viviendas habitables.

A continuación, se presentan los pasos seguidos para llegar a formar las distintas estratificaciones combinando las variables de estratificación propuestas con los métodos de estratificación y la cantidad de estratos planteados. Para poder llevar a cabo esta evaluación, se debe contar con un tamaño de muestra que contemple los requisitos planteados de antemano para la estimación de parámetros de interés en cada dominio de estimación. En la Sección 5.2.1.1 se presenta el cálculo del tamaño muestral de UPM a seleccionar en cada dominio de estimación para poder realizar la asignación muestral en cada estrato. En la Sección 5.2.1.2 se presentan los indicadores de carencias propuestos para utilizarlos como clasificadores; en la Sección 5.2.1.3 se presentan los resultados del análisis de CP robustas por FRB y luego en la Sección 5.2.1.4 se analiza la forma de la distribución de las variables de estratificación planteadas.

---

[14] http://www.santafe.gov.ar/index.php/web/Estructura-de-Gobierno/Ministerios/Economia/Secretaria-de-Planificacion-y-Politica-Economica/Direccion-Provincial-del-Instituto-Provincial-de-Estadistica-y-Censos-de-la-Provincia-de-Santa-Fe/Temas-Especificos/Datos-Estadisticos/Poblacion/Censo-Nacional-Poblacion-Hogares-y-Viviendas-2010/Todos-los-Distritos-de-la-Provincia/Estadisticas/CARENCIAS/Indicadores-de-Carencias-Criticas-segun-Censo-Nacional-de-Poblacion-2010.-Provincia-Santa-Fe.

[15] Notar que Elliott se utilizó en el Capítulo anterior para formar conglomerados, aquí se invierte el criterio de selección del análisis de la variancia para que genere estratos, es decir grupos homogéneos dentro (en vez de heterogéneos) tomando como estratos los conjuntos de UPM generados por la unidad semilla con mínima suma de cuadrados intra-grupo $SCW \cong \hat{\sigma}^2$ (en lugar de máxima). Se mantiene el criterio de igual tamaño de grupo y se omite la condición de mínima distancia dentro de los grupos en el último paso.

5.2.1.1 Cálculo del tamaño muestral

Se estableció un tamaño de muestra de UPM a seleccionar de cada dominio de estimación, en base a la siguiente fórmula del tamaño de muestra de viviendas (Naciones Unidas, 2010):

$$n_v = (z^2)\ (r)\ (1-r)\ (f)(h)\ /(p)\ (\bar{y})\ (e^2)$$

siendo

- z: el percentil normal del 95% de confianza, de 1,96
- r: parámetro de referencia, se toma la tasa de desocupación, y la estimación de la Encuesta Permanente de Hogares llevada a cabo por INDEC (EPH) como el promedio entre Gran Rosario y Gran Santa Fe durante 2010: de 0,09.
- f: el efecto diseño de los conglomerados específicos calculados en la Sección anterior 5.1.2.1 (mencionados adelante).
- h: factor de corrección que depende de la tasa de no respuesta de 10%, como 1,1 excepto para Santa Fe con 20%, de 1,2 y para Rosario con 30%, de 1,3.
- p: la proporción de la población total a la que representa la población objetivo, la tasa de actividad y su estimación de EPH: de 0,452 (salvo Rosario de 0,47).
- $\bar{y}$: el número promedio de personas por viviendas específicos según Censo 2010 (Tabla 5.12).
- e: la precisión relativa como $e = z * \sigma_r = z * (CV * r)$, con el coeficiente de variación (CV) como 15%: de $z * 0{,}15 * r$.

Una vez calculada la cantidad de viviendas n_v, se obtiene el total de UPM a seleccionar, m, dividiendo n_v por la cantidad de viviendas a seleccionar dentro de cada UPM. Entonces, para Rosario, Santa Fe y Resto provincial se toma el m teniendo en cuenta una selección de 10 viviendas promedio por UPM como máximo (cantidad establecida en la Sección anterior, ver fórmula 5.1), ya que tienen una gran cantidad total de UPM (M iguales a 554, 203 y 1.003 respectivamente) la cual permite tomar una cantidad considerable de UPM con 10 viviendas en cada una de las UPM seleccionadas. Rafaela, Reconquista y Venado Tuerto presentan poca cantidad de UPM (M iguales a 43, 33 y 65 respectivamente), por lo cual resulta conveniente tomar más viviendas por UPM, con lo cual se considera la selección de 15 viviendas por UPM, quedando así los siguientes tamaños de muestra de UPM por dominio de estimación.

Tabla 5.12: Tamaño de muestra de UPM, de vivienda, tamaño total de vivienda y cantidad promedio de personas por hogar según dominio de estimación. Año 2010

	m	n_v	N_v	$\bar{y}$
Rafaela	31	469	30.016	3,3
Reconquista	30	453	21.338	3,9
Rosario	72	720	328.516	3,3
Santa Fe	60	604	126.873	3,4
Venado Tuerto	34	517	26.891	3,0
Resto provincial	56	372	511.271	3,1

5.2.1.2 Indicadores de carencias

A continuación, se describen los 10 indicadores de carencias educativas, de servicios básicos y de calidad de la vivienda utilizados para formar estratos de UPM, considerando la primera componente principal robusta y el algoritmo de Elliott.

- JSA: Jefes o jefas de hogar sin asistencia escolar, sobre jefes o jefas de hogar.
- JPI: Jefes o jefas de hogar con educación primaria incompleta, sobre jefes o jefas de hogar.
- JASA: Mujeres jefas de hogar sin asistencia escolar, sobre jefas de hogar.
- P1419: Población de 14 a 19 años que asiste a la escuela primaria, sobre población entre 14 y 19 años.
- PNINI: Población de 15 a 19 años que no estudia ni trabaja, sobre población entre 15 y 19 años.
- HH: Hogares con hacinamiento por habitación, sobre cantidad de hogares.
- HPT: Hogares en viviendas con suelo sin revestir, sobre cantidad de hogares.
- HSA: Hogares en viviendas sin agua corriente, sobre cantidad de hogares.
- HSR: Hogares en viviendas sin sanitarios con cisterna de agua, sobre cantidad de hogares.
- HNBI: Hogares con necesidades básicas insatisfechas, sobre cantidad de hogares.

En el Gráfico 5.9 se presenta la distribución de los indicadores de carencias estudiados por UPM con sus escalas respectivas, según dominios de estimación. Se aprecian distintos niveles de carencias a través de los dominios de estimación. Apartando los valores atípicos, Rafaela presenta los menores valores de carencias, inferiores a 20%, le sigue Venado Tuerto, con un máximo porcentaje de carencias por UPM de 30%; mientras que el resto de los dominios muestran valores mayores: Reconquista presenta UPM con hasta 50% de algunas carencias, Rosario y Santa Fe UPM con hasta 60% de carencias y el Resto provincial UPM con hasta 80% de algunas carencias. A su vez, se observan carencias que se encuentran más frecuentemente a lo largo de toda la provincia, principalmente los hogares en viviendas sin agua corriente (HSA), en correspondencia, los hogares en viviendas sin sanitarios con cisterna de agua (HSR), y los jefes o jefas de hogar con primaria incompleta (JPI). Teniendo en cuenta que se observan valores atípicos en todos los indicadores univariados estudiados de cada dominio de estimación, se analiza a continuación cómo se comportan los valores atípicos multivariantes de forma gráfica (Gráfico 5.10).

En el Gráfico 5.10 se presenta el diagnóstico por dominio de estimación, mostrando para cada UPM la distancia robusta y la medida empírica de influencia conjunta de los indicadores de carencias en estudio. La distancia robusta indica que tan lejos está una observación del centro robusto del conjunto de datos; valores por encima del límite horizontal muestran outliers alejados. Mientras la influencia empírica es una medida que indica cuanto influye globalmente cada observación en la construcción del análisis de CP clásico; valores a la derecha del límite vertical indican observaciones influyentes. Todos los dominios de estimación presentan UPM con valores por encima del límite de la distancia robusta y de la influencia empírica. En Rafaela se observan 4 outliers alejados e influyentes, en Venado Tuerto 5, en Reconquista 6 y en Rosario, Santa Fe y Resto provincial se observa una gran cantidad de UPM que distorsionarían el CP clásico. Por este motivo, se considera aplicar un análisis de componentes principales robusto.

Gráfico 5.9: Indicadores de carencias por UPM según dominios de estimación. Año 2010

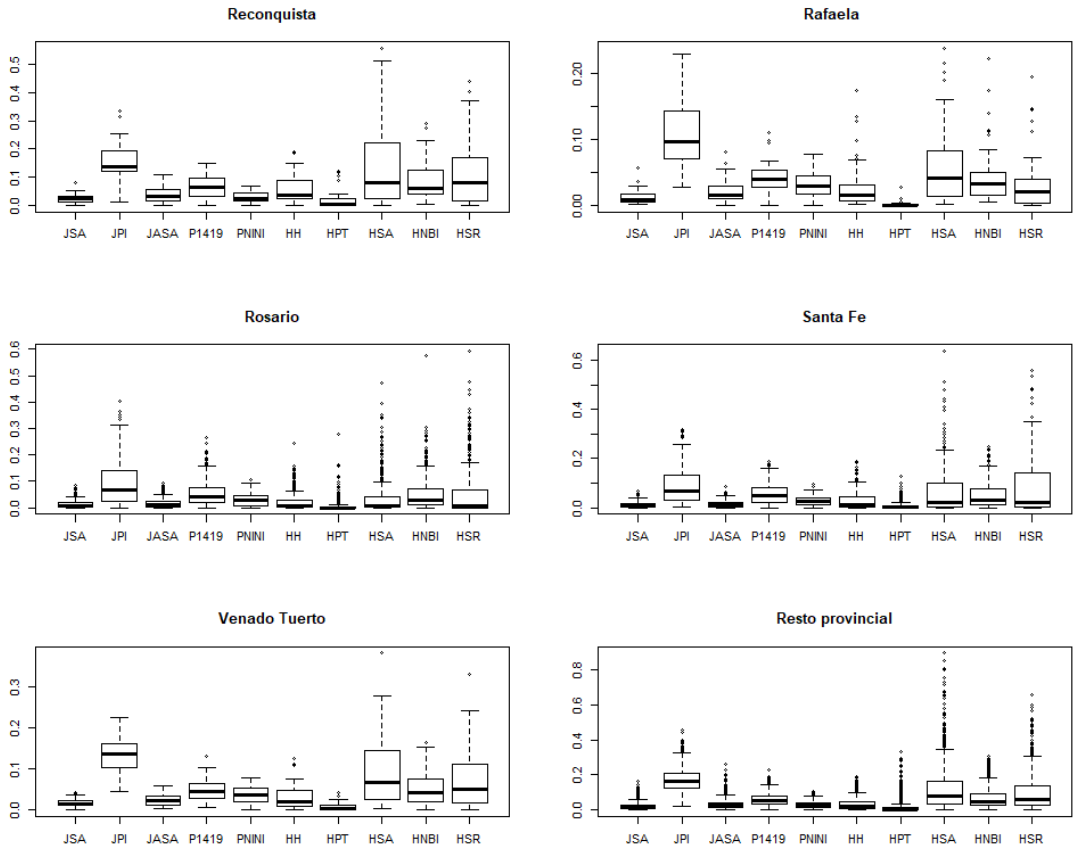
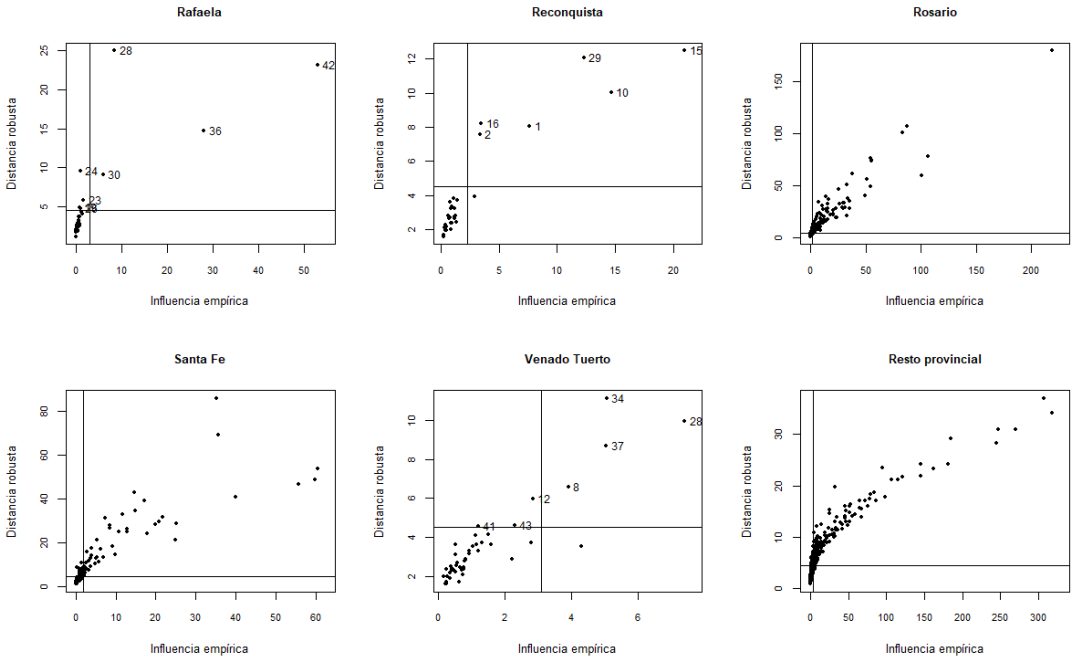


Gráfico 5.10: Gráficos de diagnóstico de influencia empírica robusta de los estimadores sobre los indicadores de carencias por UPM según dominios de estimación. Año 2010



5.2.1.3 Análisis de Componentes Principales Robustos

Como se indicó, en el Gráfico 5.9, se aprecia que la distribución de los indicadores de carencias expone en todos los dominios de estimación presencia de valores extremos con una marcada asimetría hacia la derecha, es decir falta de distribución normal. Como el análisis de CP es sensible a valores extremos, presentes en el Gráfico 5.10, se considera aplicar un CP robusto con estimadores MM. Además, al momento de evaluar los coeficientes significativos y no presentar normalidad se considera también utilizar el método Bootstrap Rápido y Robusto para estimar las CP robustas.

En la construcción de los límites de confianza de las cargas se emplea intervalos bootstrap corregidos por sesgo y aceleración, BCa (por sus siglas en ingles *bias corrected and accelerated*), los cuales establecen una mejora respecto a los intervalos percentiles bootstrap básicos, ya que le incorporan correcciones tanto por sesgo como por asimetría en la distribución bootstrap del estimador. El método ajusta los cuantiles utilizados para construir el intervalo mediante dos correcciones: de sesgo que refleja el desplazamiento de la distribución bootstrap respecto al estimador original, y de aceleración, que captura cambios en la variancia del estimador en función del valor del parámetro (Davison & Hinkley, 1997, pág. 204).

A continuación, se analizan los resultados de aplicar CP robustas con Bootstrap Rápido y Robusto, utilizando el paquete FRB del software R, a cada dominio de estimación, tomando solamente las dos primeras dimensiones (ver procedimiento computacional en Anexo A.3).

En el Gráfico 5.11 se presentan las cargas de la primera componente principal por indicador (PC1) con sus correspondientes límites de confianza de 95% según el intervalo bootstrap BCa, para cada uno de los dominios de estimación. Además, en dicho gráfico se presenta al lado del nombre del dominio de estimación el porcentaje que representa el estimador MM del primer autovalor, es decir el porcentaje que explica esta primera dimensión PC1 con respecto a la variabilidad total del conjunto de datos. Se aprecia en todos los dominios como la PC1 comprende al menos el 70% de la variabilidad total. Esto implica que más de dos tercios del comportamiento total del conjunto de los 10 indicadores se encuentran reflejados con la primera componente principal robusta. Con respecto a las cargas de cada indicador se observa que todos aportan a la formación de la PC1, ya que el límite de confianza del 95% del intervalo BCa no cubre el cero, para todos los dominios de estimación, a excepción de Rafaela que presenta 4 indicadores con cargas alrededor de cero y de Rosario con 1 indicador.

En cambio, en el Gráfico 5.12 el porcentaje de variabilidad total explicada por la segunda componente principal robusta estimada para todos los dominios de estimación se encuentra alrededor del 10%, y además los límites de confianza BCa del 95% de las cargas calculadas con el método FRB cubren el cero para la mitad de los indicadores o más, en todos los dominios de estimación.

Por este motivo, no se considera la PC2 ya que no aporta a la variabilidad total en comparación con el porcentaje de la variabilidad explicada por la PC1. Por lo tanto, se considera una buena representación a la PC1, en adelante CP.

Gráfico 5.11: Carga de la primera componente principal robusta estimada, límites de confianza BCa del 95% calculados con el método FRB, según dominio de estimación con su porcentaje de variabilidad explicada. Año 2010

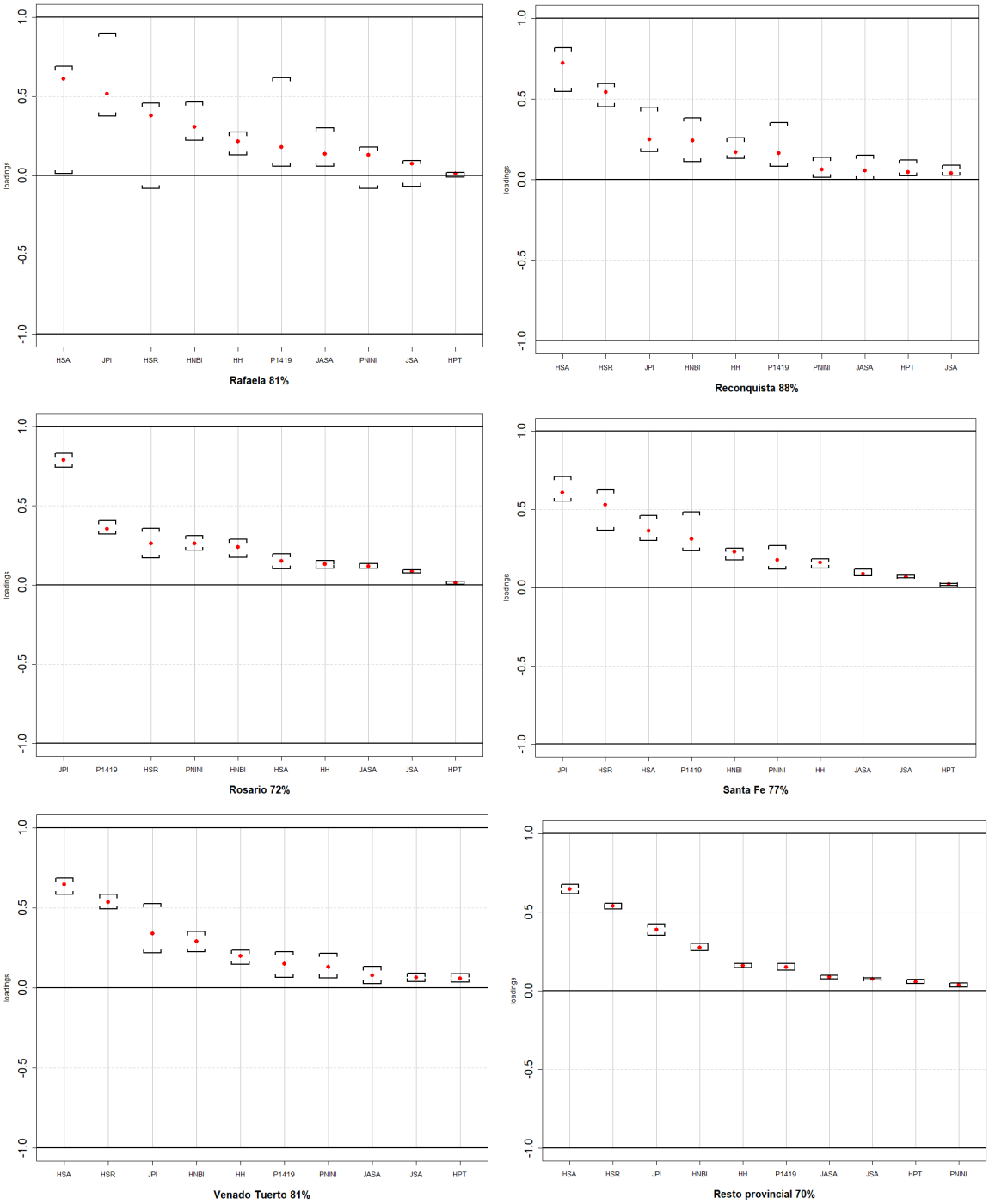
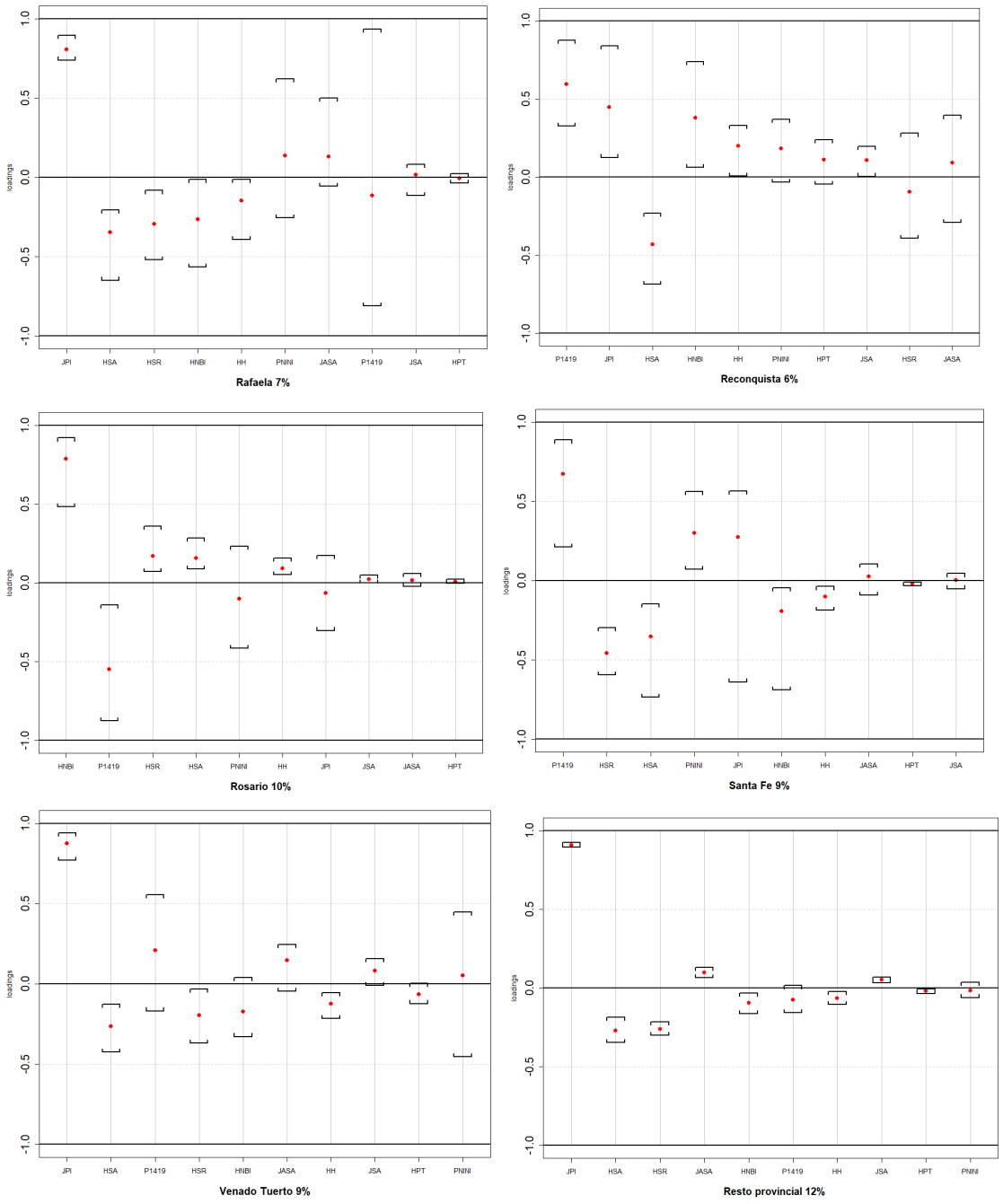


Gráfico 5.12: Carga de la segunda componente principal estimada con el método FRB (PC2), límites de confianza del 95% y porcentaje de la variabilidad explicada, según dominio de estimación. Año 2010



Con relación a la interpretación de la CP seleccionada, se puede pensar como una combinación lineal de los indicadores de carencias, mediante el valor de la carga de la CP mostrada en el Gráfico 5.10. Notar que dichas cargas resultan ser positivas para todos los indicadores de carencia, lo que refleja un índice global de carencia, es decir cuanto mayor es el porcentaje de carencias, mayor será el valor de la CP. En relación a los indicadores que más aportan en la construcción del índice, se aprecia una carga de sustancial influencia en HSA en Rafaela, Reconquista, Venado Tuerto y el Resto Provincial, mientras que JPI es el indicador con mayor influencia en Rafaela, Rosario y Santa Fe. Luego en menor medida, aportan todos los indicadores para todos los dominios de estimación, excepto HSR, PNINI, JSA y HPT en Rafaela y HPT en Rosario.

5.2.1.4 Poblaciones asimétricas

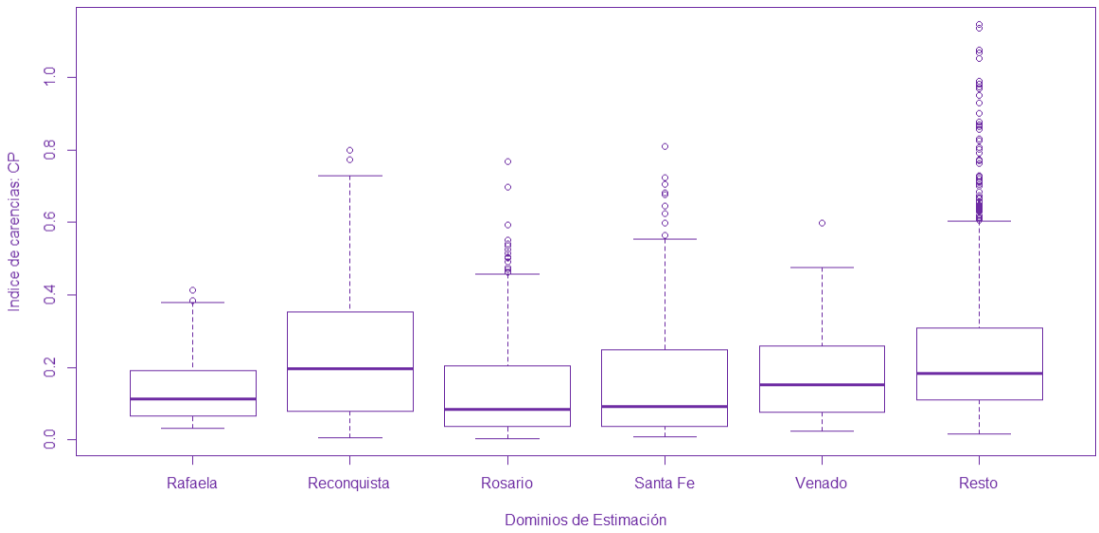
Se estudia a continuación la forma de la distribución de la variable de clasificación propuesta: univariada JPI mediante el Gráfico 5.9 y CP mediante el Gráfico 5.13.

Se observa en el Gráfico 5.9, que la distribución de la variable JPI presenta asimetría en todos los dominios de estimación; hay tres dominios que resultan marcadamente asimétricos, mientras que otros presentan una leve asimetría como Rafaela, Reconquista y Venado Tuerto.

Además, se observa en el Gráfico 5.13, que la distribución del índice CP sobre el conjunto de datos presenta asimetría para todos los dominios de estimación estudiados, en algunos más marcada como en el Resto Provincial y Reconquista, y en otros menos como en Rafaela y Venado Tuerto.

Por lo tanto, en ambos casos, se aplica el método Lavallée-Hidiroglou (L-H) (1988) para la formación de estratos.

Gráfico 5.13: Valor de UPM que toma la primera componente principal obtenida con el método FRB (CP), por dominio de estimación. Año 2010



Se aplica el método L-H utilizando el paquete *stratification* del software R, utilizando el procedimiento de Sethi (1963), hasta obtener límites que minimicen el tamaño de muestra sujeto al nivel de precisión dado, de 0,10. Si bien el método determina un estrato de inclusión forzosa *take-all*, para resolver la formación de estratos con variables asimétricas, aquí se toman todos los estratos resultantes como *take-some*, de los cuales se seleccionan muestras (ver procedimiento computacional en Anexo A.3).

Se presenta en la Tabla 5.13 la distribución de las UPM por estratificación de L-H con algoritmo de Sethi, para ambas variables de estratificación (JPI y CP) y para la estratificación de Elliott (utilizando las 10 indicadoras de carencias como variables de clasificación) en cada dominio, según se formen 3, 4 o 5 estratos. Todos los estratos construidos superan las 2 UPM, lo cual asegura extraer una cantidad de UPM mínima por estrato que permita la estimación de variancia.

Tabla 5.13: Cantidad de UPM según método de estratificación y cantidad de estratos, por dominio de estimación. Provincia de Santa Fe. Año 2010

									Nh						
h	**1**	**2**	**3**	**4**	**5**	***h***	**1**	**2**	**3**	**4**	**5**	***h***	**1**	**2**	**3**	**4**	**5**
	Rafaela					***Reconquista***						***Rosario***					
CP	15	17	11			**CP**	12	10	11			**CP**	305	156	93		
JPI	12	17	14			**JPI**	17	8	8			**JPI**	250	172	132		
Elliott	15	14	14			**Elliott**	10	11	12			**Elliott**	213	172	169		
CP	11	10	11	11		**CP**	9	7	11	6		**CP**	226	162	109	57	
JPI	11	13	9	10		**JPI**	5	12	8	8		**JPI**	208	145	128	73	
Elliott	10	11	12	10		**Elliott**	8	8	8	9		**Elliott**	160	141	123	130	
CP	11	8	11	6	7	**CP**	7	6	7	7	6	**CP**	178	140	92	102	42
JPI	9	7	13	5	9	**JPI**	5	8	7	6	7	**JPI**	163	103	102	113	73
Elliott	9	9	8	8	9	**Elliott**	7	6	6	7	7	**Elliott**	128	120	105	96	105
	Santa Fe					***Venado Tuerto***						***Resto Provincial***					
CP	104	55	44			**CP**	16	15	14			**CP**	554	327	122		
JPI	81	76	46			**JPI**	14	18	13			**JPI**	330	464	209		
Elliott	72	66	65			**Elliott**	15	15	15			**Elliott**	316	335	352		
CP	60	54	48	41		**CP**	12	10	12	11		**CP**	419	296	201	87	
JPI	49	69	50	35		**JPI**	10	11	13	11		**JPI**	245	336	292	130	
Elliott	56	51	46	50		**Elliott**	11	11	11	12		**Elliott**	236	282	231	254	
CP	54	50	43	38	18	**CP**	9	9	10	6	11	**CP**	284	284	219	139	77
JPI	49	32	39	48	35	**JPI**	8	10	9	10	8	**JPI**	145	217	286	230	125
Elliott	46	42	39	37	39	**Elliott**	8	9	9	9	10	**Elliott**	187	231	203	183	199

Se observa en la Tabla 5.13 similar cantidad de UPM entre los estratos formados por Elliott para todos dominios de estimación. En cambio, para los estratos formados por L-H, es decir JPI y CP, para Rosario, Santa Fe y Resto Provincial se observa que la cantidad de UPM disminuye cuando aumenta el estrato, más marcadamente para CP con respecto a JPI.

5.2.2 Evaluación de Estratos

Para determinar la forma de construir estratos más eficientes se compara la precisión de las estadísticas más relevantes de la EPH mediante el CV para los estratos planteados. Se calcula el CV de cada tasa de los principales indicadores del mercado de trabajo que resultan de considerar en la primera etapa de muestreo estratos clasificados mediante los tres métodos de estratificación planteados (JPI, CP y Elliott) formados a su vez con distinta cantidad de estratos (3, 4 o 5 estratos). Para el cálculo del CV se considera un muestreo con selección PPT con cantidad de viviendas habitables como medida de tamaño y una asignación proporcional. En la segunda etapa de muestreo se seleccionan viviendas por MSA. Se plantean las estadísticas para la población de 14 años y más: la tasa específica de actividad, de empleo y de desocupación según jefe de hogar, sexo, edad y sexo; y la composición de la tasa de empleo según sexo y edad, relación de parentesco y por nivel educativo (ver procedimiento computacional en Anexo A.3).

Para evaluar la eficiencia se compara el CV de las tasas para la población total y las tasas para poblaciones específicas. Se calcula la variancia del estimador como la suma de la variancia de cada estrato poblacional en base a la fórmula (3.8) o (3.24) según corresponda, donde m_h se toma como la asignación proporcional de

las UPM con respecto al tamaño de muestra total m calculado en Tabla 5.12 y $\bar{n}$ es igual a 10 o 15 viviendas por UPM de acuerdo con el dominio, para cada uno de los distintos métodos de estratificación.

A continuación, en la Tabla 5.14 y en los Gráficos 5.14 a 5.16, se presentan para Rafaela, a modo de ejemplo, los valores de CV, para cada uno de los distintos métodos de estratificación, de las tasas específicas de actividad, de empleo y de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo; y en la Tabla 5.15 y en el Gráfico 5.17, los CV de la tasa empleo por nivel de instrucción, por sexo y edad; y por relación de parentesco con el jefe o jefa del hogar. Para los demás dominios de estimación se presentan en el Anexo A.2 los valores de CV respectivos en las Tablas A.1 a A.10 y en los Gráficos A.1 a A.20.

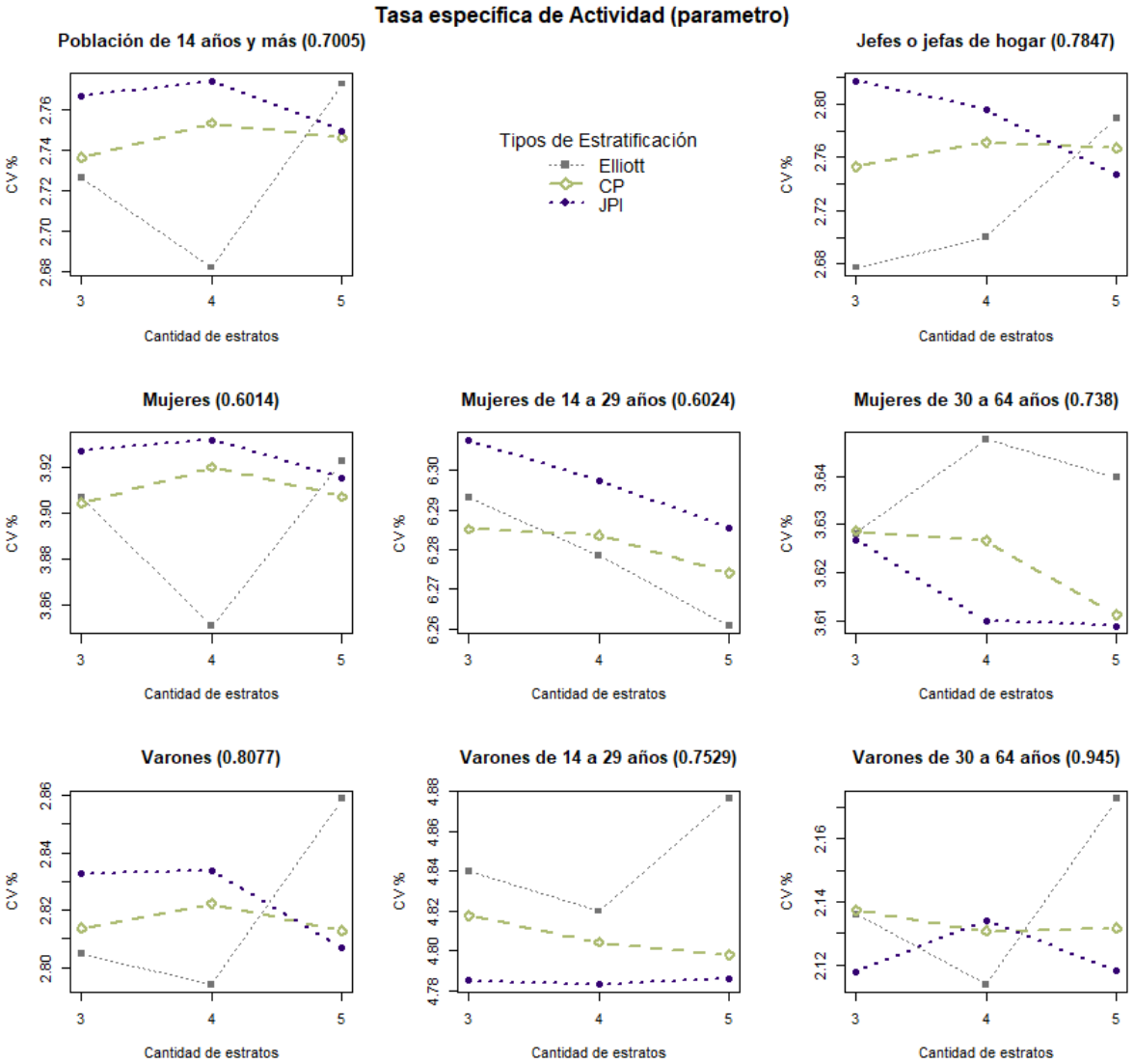
Tabla 5.14: Principales indicadores del mercado de trabajo, parámetro y CV según método de estratificación y cantidad de estratos en la primera etapa de muestreo. Rafaela. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		n°de estratos			n°de estratos			n°de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Actividad										
Tasa de la población total	*54,1*	4,51	4,47	4,50	**4,44**	4,45	4,46	4,55	4,53	4,45
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*70,0*	2,73	**2,68**	2,77	2,74	2,75	2,75	2,77	2,77	2,75
Mujeres	*60,1*	3,91	**3,85**	3,92	3,90	3,92	3,91	3,93	3,93	3,92
Varones	*80,8*	2,80	**2,79**	2,86	2,81	2,82	2,81	2,83	2,83	2,81
Jefes o jefas de hogar	*78,5*	**2,68**	2,70	2,79	2,75	2,77	2,77	2,82	2,80	2,75
Mujeres de 14 a 29 años	*60,2*	6,29	6,28	**6,26**	6,29	6,28	6,27	6,31	6,30	6,29
Mujeres de 30 a 64 años	*73,8*	3,63	3,65	3,64	3,63	3,63	3,61	3,63	**3,61**	3,61
Varones de 14 a 29 años	*75,3*	4,84	4,82	4,88	4,82	4,80	4,80	**4,78**	4,78	4,79
Varones de 30 a 64 años	*94,5*	2,14	**2,11**	2,17	2,14	2,13	2,13	2,12	2,13	2,12
Empleo										
Tasa de la población total	*51,3*	4,44	4,41	4,46	**4,41**	4,42	4,43	4,51	4,50	4,43
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*66,5*	2,86	**2,82**	2,90	2,87	2,88	2,88	2,89	2,90	2,88
Mujeres	*55,6*	4,24	4,20	4,23	4,22	4,22	**4,20**	4,23	4,22	4,21
Varones	*78,2*	**2,95**	2,95	3,01	2,97	2,97	2,97	2,98	2,99	2,97
Jefes o jefas de hogar	*76,7*	**2,78**	2,80	2,88	2,86	2,87	2,87	2,91	2,90	2,85
Mujeres de 14 a 29 años	*52,1*	7,45	7,45	7,40	7,40	7,40	**7,38**	7,45	7,45	7,41
Mujeres de 30 a 64 años	*70,3*	3,99	4,02	4,00	3,97	3,97	3,95	3,97	**3,94**	3,94
Varones de 14 a 29 años	*70,7*	5,37	5,36	5,41	5,35	5,34	5,34	**5,32**	5,32	5,33
Varones de 30 a 64 años	*92,9*	2,29	2,28	2,33	2,30	2,29	2,29	**2,28**	2,29	2,28
Desocupación										
Tasa de la población total	*5,1*	17,10	17,11	16,94	16,81	16,78	**16,73**	16,86	16,81	16,77
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*5,1*	17,10	17,11	16,94	16,81	16,78	**16,73**	16,86	16,81	16,77
Mujeres	*7,5*	20,14	20,19	19,95	19,73	19,71	**19,69**	19,86	19,83	19,74
Varones	*3,2*	27,43	27,47	27,37	27,38	27,38	**27,33**	27,42	27,39	27,43
Jefes o jefas de hogar	*2,2*	34,72	34,83	34,62	34,65	**34,64**	34,64	34,72	34,64	34,69
Mujeres de 14 a 29 años	*13,5*	24,12	24,22	24,04	23,94	**23,93**	23,95	24,07	24,12	24,07
Mujeres de 30 a 64 años	*4,8*	30,61	30,60	30,39	30,40	**30,39**	30,40	30,51	30,41	30,42
Varones de 14 a 29 años	*6,1*	33,17	33,21	**33,15**	33,21	33,22	33,21	33,23	33,31	33,31
Varones de 30 a 64 años	*1,7*	46,41	46,53	**46,38**	46,39	46,36	46,41	46,51	46,47	46,57

Se considera la precisión máxima aceptable de CV de 15%, valor considerado en el cálculo del tamaño muestral.

Notar que los parámetros tienen dos tipos de estimadores, por un lado, están las proporciones que corresponden a las tasas totales de empleo (para Rafaela Tabla 5.15 y Gráfico 5.17) y por el otro las razones que estiman las tasas específicas (para Rafaela Tabla 5.14 y Gráficos 5.14 a 5.16); por lo tanto, las tasas de empleo, aunque presenten parámetros más pequeños que las tasas específicas, entre 5 y 10%, pueden resultar más precisas, es decir con un valor de CV menor a 15%.

Gráfico 5.14: Coeficiente de variación de la tasa específica de actividad según población de 14 años y más, jefes o jefas de hogar, sexo, y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rafaela. Año 2010



Para Rafaela se observa en Tabla 5.14 y Gráficos 5.14 a 5.16 todas las tasas específicas de actividad y de empleo superiores a 50%, con niveles de precisión aceptables de CV, o sea, inferior a 15% para todas las estratificaciones. Contrariamente las tasas específicas de desocupación en Rafaela presentan poca precisión de CV, superior a 15% para todas las estratificaciones.

Gráfico 5.15: Coeficiente de variación de la tasa específica de empleo según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rafaela. Año 2010

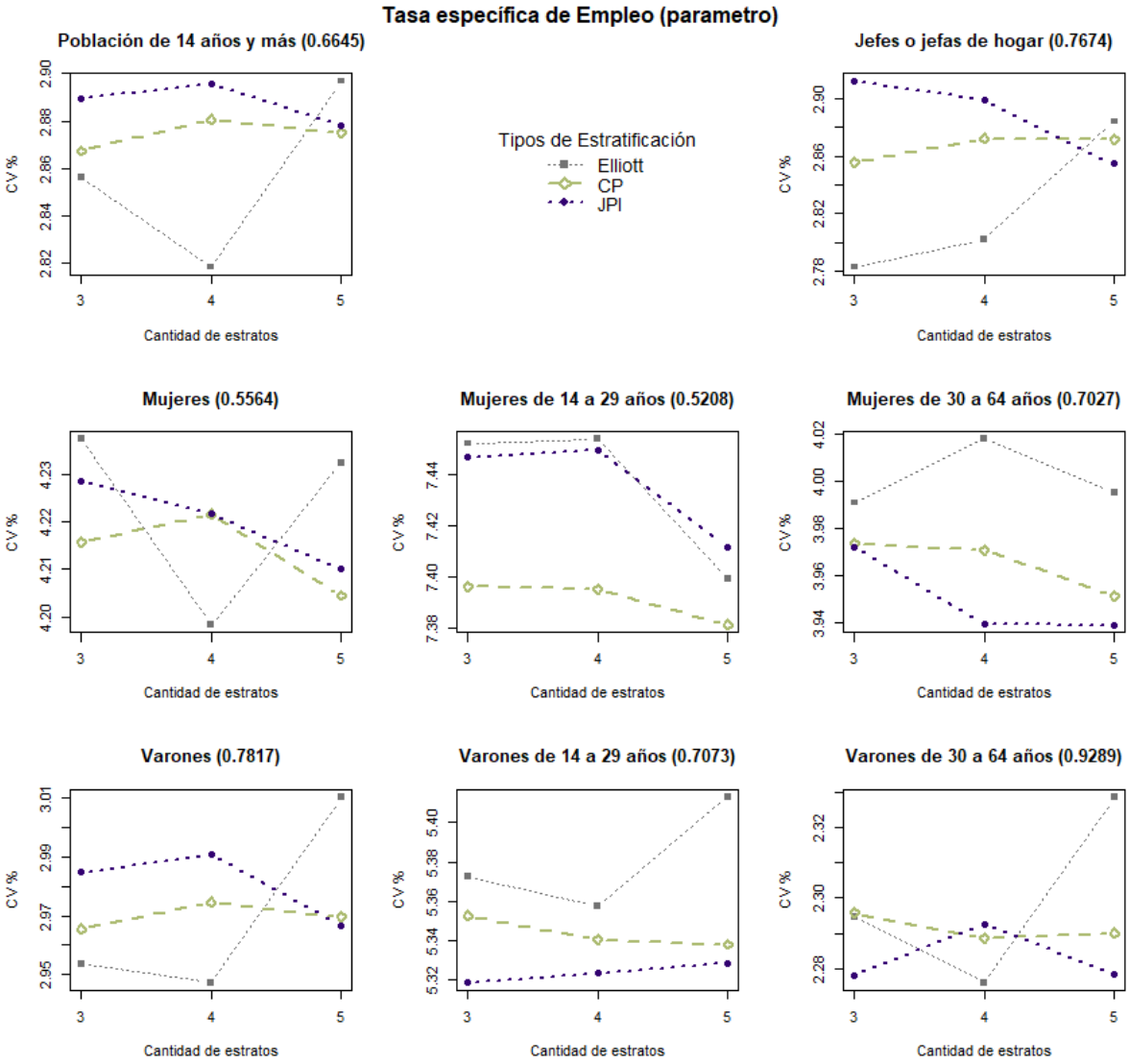


Gráfico 5.16: Coeficiente de variación de la tasa específica de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rafaela. Año 2010

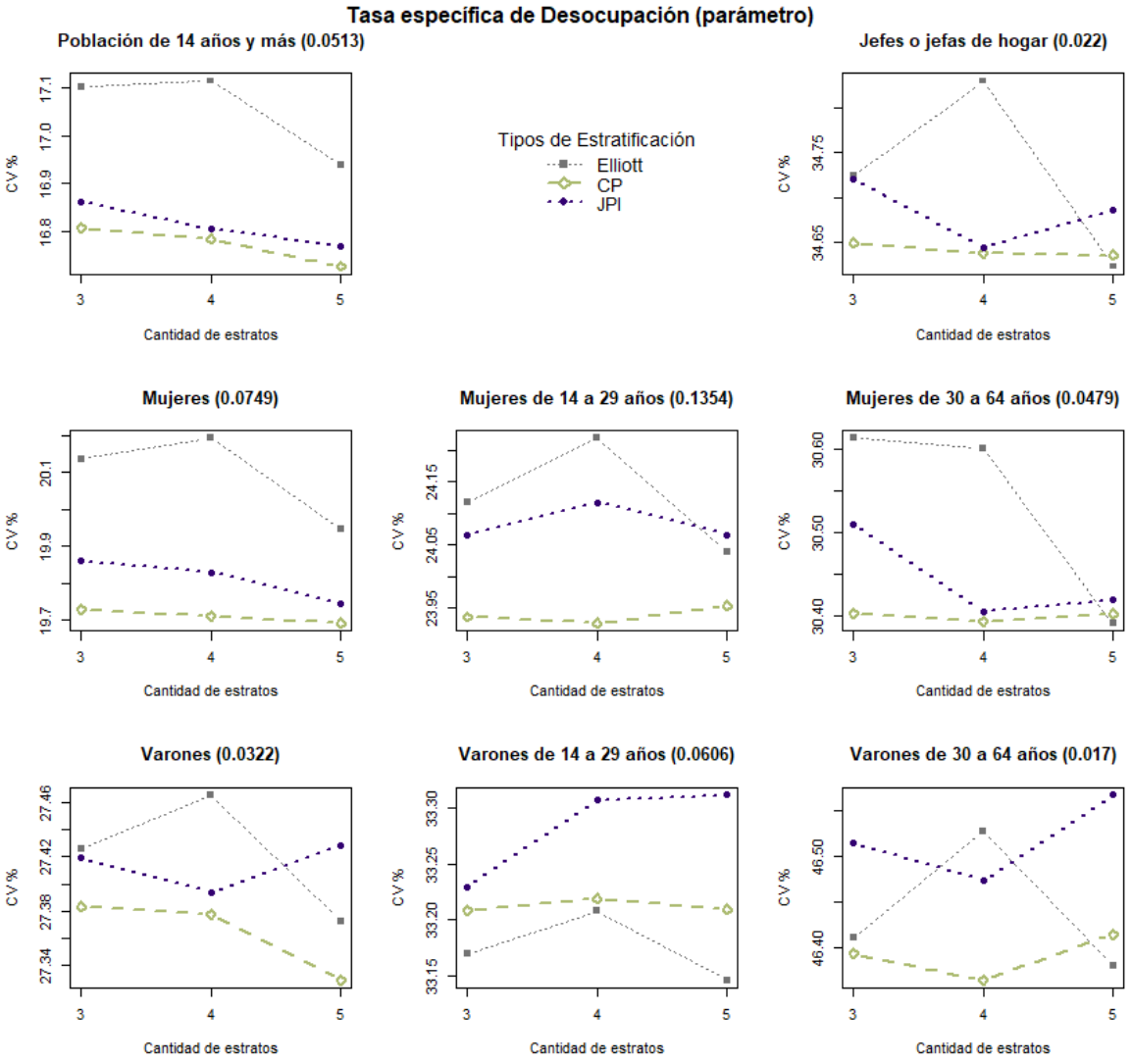


Tabla 5.15: Composición de la tasa de empleo, parámetro y CV según sexo y edad, relación de parentesco con el jefe o jefa de hogar y nivel de instrucción, por método de estratificación y cantidad de estratos en la primera etapa de muestreo. Rafaela. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		n°de estratos			n°de estratos			n°de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Tasa de empleo	*51,3*	4,44	**4,41**	4,46	4,41	4,42	4,43	4,51	4,50	4,43
Sexo y grupos de edad										
Mujeres hasta 29 años	*6,9*	10,69	10,69	10,71	10,67	10,67	10,68	10,69	10,65	**10,64**
Mujeres de 30 a 64 años	*14,2*	**5,99**	5,97	6,02	6,01	6,01	6,01	6,04	6,06	6,02
Mujeres de 65 años y más	*1,3*	23,75	23,47	24,09	23,95	23,77	23,61	24,08	23,78	23,76
Varones hasta 29 años	*9,6*	10,01	10,10	9,99	9,72	9,68	9,60	9,81	9,67	**9,59**
Varones de 30 a 64 años	*18,0*	5,51	**5,50**	5,64	5,59	5,61	5,61	5,74	5,73	5,61
Varones de 65 años y más	*1,4*	21,83	21,65	22,24	22,14	21,99	21,95	22,14	22,05	21,95
Relación de parentesco										
Jefe o jefa de hogar	*24,5*	**3,95**	3,95	3,98	4,01	4,03	4,04	4,10	4,08	4,00
Cónyuge	*12,7*	**6,60**	6,67	6,79	6,77	6,78	6,78	6,85	6,86	6,77
Hijos	*10,9*	10,23	10,22	10,07	9,97	9,96	**9,94**	9,94	9,87	9,91
Otros componentes	*3,2*	18,53	18,36	18,26	18,15	18,10	18,13	18,17	18,15	18,17
Nivel educativo										
Primario incompleto	*3,6*	20,03	19,80	19,11	18,38	18,34	18,33	19,08	18,58	18,25
Primario completo	*13,0*	10,30	10,28	9,83	8,98	8,85	**8,83**	9,06	9,00	8,76
Secundario incompleto	*9,8*	10,31	10,33	10,36	9,98	10,02	**9,94**	10,13	10,18	10,04
Secundario completo	*11,7*	8,43	8,36	8,25	8,10	8,08	**8,05**	8,20	8,19	8,16
Superior y universitario incompleto	*5,0*	12,94	13,08	12,75	12,42	12,39	**12,32**	12,45	12,45	12,33
Superior y universitario completo	*7,6*	12,76	13,00	12,58	11,51	11,13	10,91	10,62	10,70	**10,59**
Sin instrucción	*0,5*	42,00	41,54	40,90	40,92	40,91	40,90	41,65	41,29	41,12

Gráfico 5.17: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rafaela. Año 2010. (Continúa)

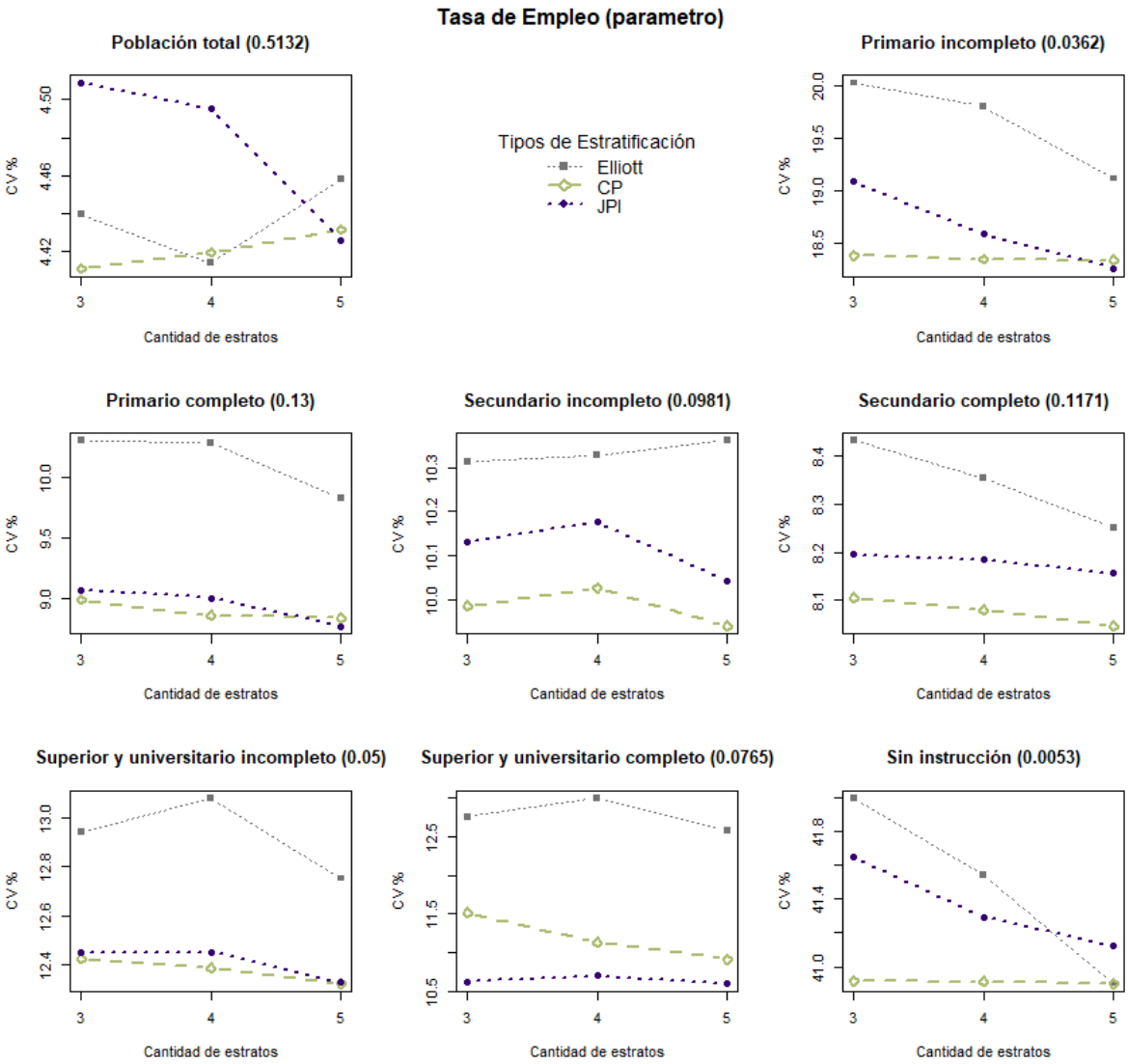
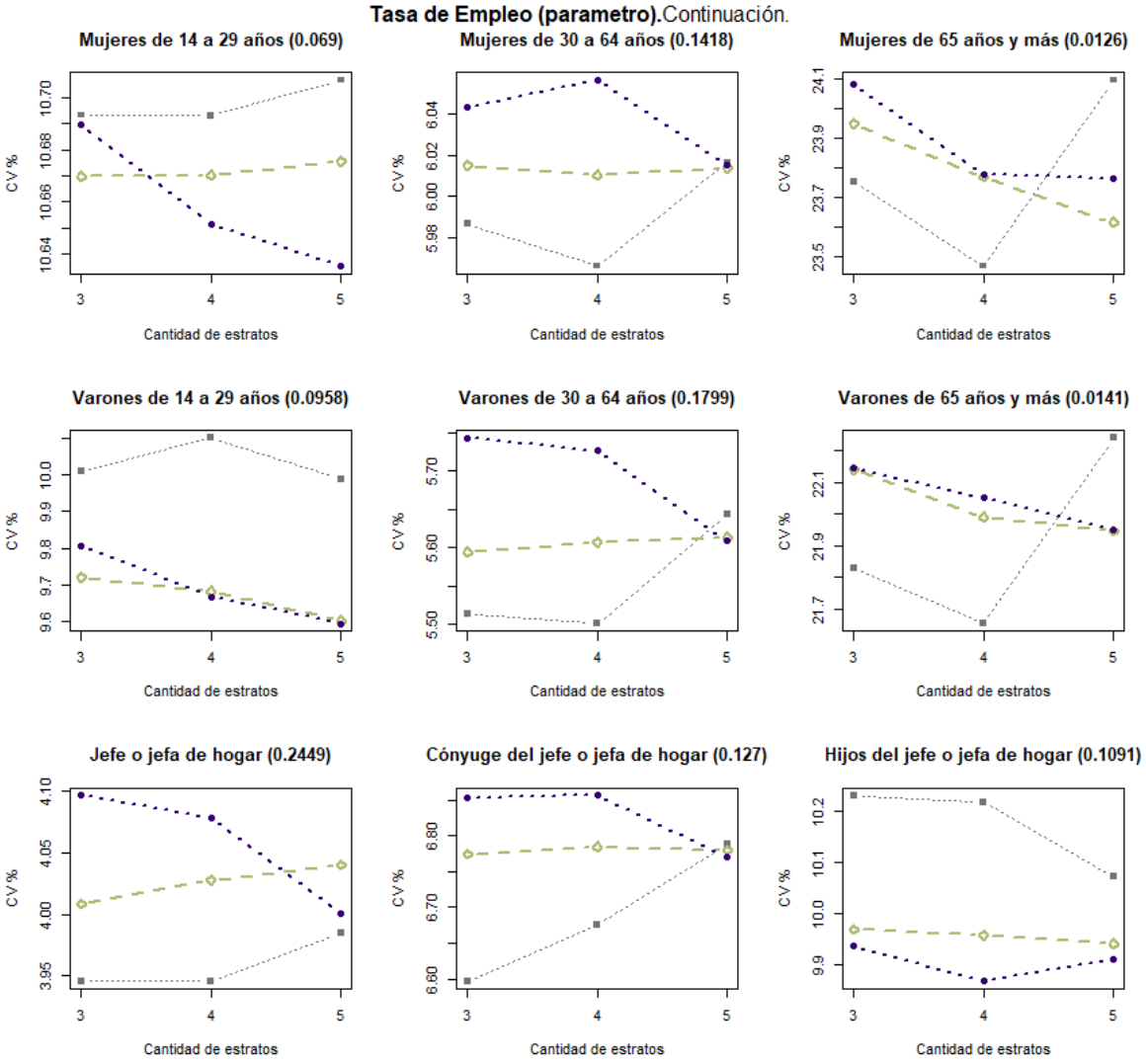


Gráfico 5.17: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rafaela. Año 2010. (Continuación)



En Rafaela en la Tabla 5.15 y el Gráfico 5.17, de las 17 tasas de composición de empleo la mayoría presentan estimadores con un nivel de precisión aceptable (excepto 5) para todas las estratificaciones.

En general en todos los dominios de estimación, se aprecia, como es de esperar, que a medida que disminuyen los valores de los parámetros, los CV van aumentando, quedando pocos estimadores con un nivel aceptable de precisión (ver Tablas 5.14, 5.15, y de A.1 a A.10 del Anexo A.2, y Gráficos 5.14 a 5.17 y de A.1 a A.20 del Anexo A.2).

Además, se aprecia en todos los dominios de estimación, que los valores de los CV para una misma tasa son muy similares para los distintos métodos de estratificación utilizados (encontrando diferencias mayormente en los decimales hasta 3 puntos porcentuales) según los parámetros con niveles de precisión aceptable. A su vez los valores de CV entre CP y JPI resultan más similares entre sí.

5.2.3 Elección de Estratos

A continuación, se evalúa la eficiencia para todos los dominios de estimación de las distintas formas de estratificar las UPM mediante la precisión de las tasas analizadas en base al CV según Tabla 5.16. Se analizan las tasas específicas de actividad, de empleo y de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo; y las tasas de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar.

Con respecto a la precisión de las tasas específicas de actividad y de empleo, los CV presentan valores pequeños y similares entre ellos para todos los dominios de estimación, por lo tanto, las tasas resultan precisas para cualquier método de estratificación y cantidad de estratos. Por otra parte, tanto las tasas de empleo como las tasas específicas de desocupación presentan niveles de precisión muy variados. Cuando se presentan diferencias de precisión entre los métodos de estratificación para una misma tasa, resulta más preciso en todos los dominios estratificar mediante el nivel educativo de JPI por UPM o utilizando la estratificación en base a las CP robustas de los indicadores de carencias (CP) con respecto a aplicar el algoritmo de Elliot; y resulta más preciso formar 3 o 5 estratos en lugar de 4, excepto en Rafaela.

En síntesis, en relación a la eficiencia, se prefiere, en la mayoría de los dominios, estratificar por JPI o por CP. Con respecto a la cantidad de estratos formados se prefiere en la mayoría de los dominios formar 3 estratos, y en algunos dominios resulta más eficiente formar 5 estratos, como en Rosario y Venado Tuerto. Teniendo en cuenta tanto la eficiencia como la eficacia, que JPI también participa en la formación de CP y que los valores de CV son muy similares, se considera estratificar mediante JPI, ya que resulta más simple de formar y de interpretar para el usuario común.

Tabla 5.16: Orden de preferencias de los métodos de formación de estratos en la primera etapa de muestreo, por dominios de estimación. Provincia de Santa Fe. Año 2010

Dominios de estimación	Rafaela	Reconquista	Rosario	Santa Fe	Venado Tuerto	Resto provincial
	Orden de precisión (Rango de CV)					
Tasas específicas de Actividad y de Empleo	CP>JPI Cant: 3=4=5 (2%-8%)	JPI> Elt >CP Cant: 3=4=5 (2%-9%)	JPI=CP=Elt Cant: 3=4=5 (2%-6%)	Elt=JPI=CP Cant: 3=4=5 (2%-7%)	CP=JPI= Elt Cant: 5>4>3 (1%-8%)	JPI=CP=Elt Cant: 3=5>4 (2%-11%)
Tasas de Empleo	CP>JPI>Elt Cant: 4=5>3 (3%-40%)	JPI>Elliot>CP Cant:3>4=5 (2%-30%)	JPI>CP> Elt Cant: 5>4>3 (5%-30%)	Elt >JPI>CP Cant: 3=5>4 (3%-30%)	CP>JPI> Elt Cant: 4=5>3 (2%-21%)	JPI>=CP>Elt Cant: 5>4>3 (3%-30%)
Tasas específicas de Desocupación	CP>JPI (16%-30%)	JPI> Elt >CP (16%-39%)	JPI=CP> Elt Cant: 3=4=5 (12%-30%)	Elt >JPI>CP Cant: 3=4=5 (12%-30%)	CP=JPI= Elt (17%-45%)	JPI=CP>Elt Cant: 3=4=5 (12%-30%)
Preferencias Estratificación (n° estratos)	CP (4 estratos)	JPI (3 estratos)	JPI=CP (5 estratos)	Elliott=JPI (3=5 estratos)	CP (5 estratos)	JPI=CP (5 estratos)

6 Conclusiones

El Instituto Provincial de Estadística y Censos de Santa Fe tiene como meta la realización de una Muestra Maestra en la provincia denominada MUMSAFE. El objetivo de esta es que constituya un instrumento a partir de la cual se seleccionen muestras probabilísticas de viviendas para la realización de encuestas a hogares durante el período intercensal. La determinación de un diseño muestral para una Muestra Maestra de este tipo no es simple dado que se tienen que tener en cuenta aspectos como la vida útil del marco y los distintos objetivos que persiguen las encuestas a hogares del plan de encuestas del Instituto, para las cuales se seleccionan las muestras. Esta tesis evaluó algunos componentes que forman parte del diseño muestral probabilístico a partir del cual se construirá la Muestra Maestra, buscando contribuir con la elección de un diseño que sea eficiente y eficaz.

En base a los datos del Censo Nacional de Población, Hogares y Viviendas 2010, se evaluó un diseño con dos etapas de muestreo. Las pruebas que se realizaron respecto al diseño estuvieron centradas en la primera etapa. En la misma, se evaluaron los componentes que forman parte de ella, la construcción de conglomerados y la estratificación de los mismos, considerando la asignación de la muestra entre estratos a partir de la adjudicación proporcional. También se estudiaron distintos métodos de selección de los conglomerados, considerando diseños con igual y distinta probabilidad de selección. En la segunda etapa, se determinó la selección de 10 o 15 hogares por MSA dentro de cada unidad primaria ya seleccionada. La evaluación se llevó a cabo en todos los dominios de estimación, tomados como estratificación geográfica: las principales ciudades de la provincia (Rafaela, Reconquista, Rosario, Santa Fe y Venado Tuerto), y el aglomerado que resulta de excluir éstas 5 localidades de la provincia de Santa Fe (Resto Provincial).

Se planteó la construcción de conglomerados a partir del agrupamiento con distintos tamaños de radios censales: tomando de a uno, de a dos y de a tres radios, combinados con distintas agrupaciones formadas mediante la unión consecutiva por numeración censal y por la unión por cercanía geográfica en grupos de igual tamaño y heterogéneos (con respecto a 3 variables de clasificación) utilizando el método de Elliott (2011); también se tuvieron en cuenta métodos de selección con probabilidades iguales o distintas, considerando en este último caso, el tamaño medido a partir del total de viviendas habitables.

Entre los resultados a destacar, se encontró que el diseño por conglomerados que considera probabilidades de inclusión distintas brinda mejores resultados respecto a la precisión, en base al cálculo del efecto de diseño de un conjunto de estimadores de interés, comparado a un diseño con igual probabilidad. Las unidades de muestreo con mayor cantidad de radios presentaron mejor precisión, siendo la relación entre precisión y cantidad de radios, inversa. El algoritmo de unión geoespacial por tamaños homogéneos y grupos heterogéneos presentó mejor eficiencia, en precisión y en igualdad de tamaño de las unidades primarias, mientras que para la eficacia no se hallaron diferencias con la agrupación por unión de numeración consecutiva de radios censales. En base a los resultados evaluados, y teniendo en cuenta que el objetivo es la determinación de un diseño que sea eficiente y eficaz, se determinó que la mejor opción es la definición de unidades primarias de muestreo a partir de la unión de dos radios mediante el método de Elliot, y la selección de estas unidades considerando un método de selección que considere probabilidades proporcionales a una medida de tamaño.

En segundo lugar, se evaluó la estratificación de los conglomerados construidos con 2 radios censales mediante la unión geoespacial, considerando distintas variables de clasificación. Inicialmente, se consideró

un método univariado a partir de utilizar como variable de estratificación el nivel educativo del jefe o jefa del hogar, método utilizado por el INDEC en la estratificación de áreas de la MMUVRA. Luego, se tuvo en cuenta la utilización de un método que considera un conjunto de variables determinadas por indicadores de carencias socio-económicas, información que surge del Censo 2010. En particular, y para reducir la dimensionalidad, se construyó un índice multivariado en base a la primera dimensión de componentes principales en su versión robusta para contemplar la existencia de valores outliers o influyentes desde un punto de vista univariado y multivariado. Por otra parte, se evaluaron distintos métodos de estratificación: el algoritmo de Lavallée-Hidiroglou (para estratificar poblaciones asimétricas) y el algoritmo geográfico en grupos de igual tamaño y homogéneos (con respecto a los indicadores de carencias) considerando el método de Elliot. Por último, se evaluaron distintas cantidades de estratos, dividiendo la población de unidades primarias de muestreo en 3, 4 o 5 estratos.

En la elección del método de estratificación, se encontró que formar estratos por el algoritmo que considera poblaciones asimétricas, o sea el método de Lavallée-Hidiroglou, brinda mejores resultados con respecto a la precisión con respecto al algoritmo geoespacial, en base al cálculo del coeficiente de variación de un conjunto de estimadores relacionados con el mercado laboral, que constituye uno de los ámbitos más importantes de las encuestas a hogares que se desarrollan en el ámbito oficial. A su vez, se encontró que usar el nivel educativo del jefe de hogar o la primera dimensión de componentes principales robustas brinda resultados similares respecto a la precisión. Para la cantidad de estratos formados resulta más efectivo formar 3 estratos. Por estos motivos, y por una cuestión de simplicidad a la hora de explicar a usuarios no técnicos el método de estratificación, se determinó considerar la estratificación de unidades primarias de muestreo utilizando tres estratos construidos a partir de la aplicación del método de Lavallée-Hidiroglou sobre el nivel educativo del jefe del hogar.

Mediante la evaluación de las distintas componentes de diseño consideradas se obtuvo que el diseño de Muestra Maestra más efectivo, resulta de tomar, en cada uno de los dominios de estimación establecidos, unidades primarias de muestreo con conglomerados de 2 radios censales, agrupadas mediante el algoritmo de unión geoespacial de igual tamaño y heterogéneos. A su vez considera estratificar los conglomerados, en 3 estratos según el nivel de educación del jefe o jefa del hogar mediante el algoritmo de Lavallée-Hidiroglou, para estratificar poblaciones asimétricas.

Entre los estudios futuros se considera evaluar el diseño determinado con los datos a nivel de radios censales que provengan del último Censo Nacional de Población, Hogares y Viviendas 2022, y realizar los ajustes necesarios en base a los datos más actuales. Por otra parte, se debe considerar en un futuro el ajuste del diseño determinado a las características propias de las encuestas que se consideren realizar en el período intercensal, en particular, aquellas que tengan en cuenta la selección de muestras por panel, como, por ejemplo, las encuestas que miden el estado ocupacional de las personas.

Bibliografía

Bankier, M. D. (1988). Power allocations, determining sample sizes for sub-national areas. *The American Statistician*, 174-177.

Blacona, M. T., Marí, G., & Méndez, F. (2009). Estratificación de Poblaciones Asimétricas. *XIV Jornadas de "Investigación en la Facultad". Facultad de Ciencias Económicas y Estadística .* Rosario: Universidad Nacional de Rosario.

Bussi, J., Marí, G., & Méndez, F. (2017). Stratification of populated areas using a method of robust principal components. *Statistical Journal of the IAOS*, 981-988.

Cantwell, P. J. (1990). Equal Characteristic Clustering. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 231-236.

Cantwell, P. J. (1991). Clustering in Demographic Surveys: Long-Term Results. *Proceeding of the Survey Methods Section, American Statistical Association*, 550-555.

Capretto, T., & Marí, G. (2018). Método de agrupamiento geoespacial para la segmentación de una población de viviendas. *Vigesimoterceras Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística.* Rosario: Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística.

Carrol, J. (1970). Allocation of a sample between States. *Memorandum no publicado.* Australian Bureau of Census ande Statistics.

Cochran, W. G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 345-358.

Cochran, W. G. (1977). *Sampling Techniques, 3rd edition.* New York: John Wiley & Sons, Inc.

Dalenius, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal*, 203-213.

Dalenius, T., & Hodges, J. (1957). The Choice of Stratification Points. *Scandinavian Actuarial Journal*, 198-203.

Dalenius, T., & Hodges, J. (1959). Minimum variance stratification. *Journal of the American Statistical Association,54*, 88-101.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambrigde: Cambridge University Press.

Deming, W. (1950). *Some theory of Sampling.* New York: John Wiley and Sons.

Drucker, P. F. (1989). *El Ejecutivo Eficaz.* Barcelona: Edhasa.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics,7*, 1-26.

Ekman, G. (1949). An Approximation Useful in Univariate Stratification. *The Annals of Mathematical Statistics,30*, 219-229.

Elliott, M. (2011). A Simple Method to Generate Equal-Sized Homogenous Strata or Clusters for Population-Based Sampling. *Annals of Epidemiology, 21*, 290-296.

Elliott, M. R. (2011). A Simple Method to Generate Equal-Sized Homogenous Strata or Clusters for Population-Based Sampling. *Annals of Epidemiology, 21*, 290-296.

Fellegi, I. P. (1981). Should the census counts be adjusted for allocation purposes?-Equity Considerations. *Academic Press*, 47-76.

Glasser, G. J. (1962). On the complete coverage of large units in a statistical study. *International Statistical Review, 30*, 28-32.

Gunning, P., & Horgan, J. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology, 30*, 159-166.

Hidiroglou, M. A. (1986). The construction of self-representing stratum of large units in survey design. *The American Statistican, 40*, 27-31.

Hidiroglou, M. A., & Kozak, M. (2018). Stratification of Skewed Populations: A Comparison of Optimisation-based versus Approximate Methods. *International Statistical Review, 86*, 87-105.

INDEC. (1999). *Marco de Muestreo Nacional Urbano para Encustas a Hogares.* Buenos Aires: INDEC. 12 Metodología.

INDEC. (2010). *Manual del Actualizador del Censo Nacional de Población, Hogares y Viviendas.* Buenos Aires: Documento no publicado.

INDEC. (2012). *Muestra Maestra Urbana de Viviendas de la República Argentina MMUVRA 2011.* Buenos Aires: Documento Interno Metodológico.

INDEC. (2019). *Encuesta de actividad de niñas, niños y adolescentes 2016-2017 : factores de expansión, estimación y cálculo de los errores por muestra para el dominio urbano : nota técnica EANNA / 1a ed.* Ciudad Autónoma de Buenos Aires: INDEC.

King, A., & Jessen, R. (1945). The Master Sample of Agriculture. *Journal of the American Statistical Association, 40*, 38-56.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Kish, L. (1975). *Muestreo de Encuestas.* México: Editorial Trillas.

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

Kozak, M., & Verma, M. (2006). Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency. *Surrvey Methodology, 32*, 157-163.

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

LaVarnway, G. T. (1988). An introduction to CART: Classification and regression trees. *Computer Science and Statistics: Proceedings of the 20th Symposium on the Interface* (págs. 298-301). Alexandria: American Statistical Association.

Lehmann, C. H. (1989). *Geometría Analítica.* México: LIMUSA.

Lifeder. (4 de Diciembre de 2019). *https://www.lifeder.com/*. Obtenido de Distancia euclidiana: concepto, fórmula, cálculo, ejemplo.: https://www.lifeder.com/distancia-euclidiana/

MacLachen, G., & Peel, M. D. (2000). *Finite mixture models.* New York: Wiley.

MacQueen, J. (1967). Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability* (págs. 281-297). Los Angeles: University of California.

Naciones Unidas. (2010). *Medición de la población económicamente activa en los censos de población: Manual.* Nueva York: Naciones Unidas.

Salibian-Barrera, M., & Zamar, R. (2002). Bootstrapping Robust Estimates of Regression. *The Annals of Statistics,30*, 556-582.

Salibián-Barrera, M., Van Aelst, S., & Willems, G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association, 101*, 1198-1211.

Salibián-Barrera, M., Van Aelst, S., & Willems, G. (2008). Fast and Robust Bootstrap. *Statistical Methods and Applications*.

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Sethi, V. K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics,5*, 20-33.

Tatsuoka, K., & Tyler, D. (2000). The Uniqueness of S and M-Functionals Under Non-Elliptical Distributions. *The Annals of Statistics*, 1219-1242.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples. Second Edition.* Cham, Switzerland: Springer International Publishing AG.

Valliant, R., Dever, J., & Kreuter, F. (2022). *PracTools: Tools for Designing and Weighting Survey Samples.* R package version 1.2.6.

Van Aelst, S., & Willems, G. (2013). Fast and Robust Bootstrap for Multivariate. Inference: The R Package FRB. *Journal of Statistical Software*, 1-43.

Woodruff, R. S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association, 66*, 411-414.

Zabala, J. P., & Librandi, J. M. (2018). Medir para planificar: el Estudio sobre Salud y Educación Médica y los límites de la política sanitaria de la Revolución Argentina (1966-1973). *Rev. Cienc. Salud,16*, 550-570.

A. Anexo

A.1. Selección de Variables para utilizar el algoritmo de Elliott

Radios de Rosario, identificados en K=5 grupos según las variables presentadas en la Tabla 5.1.

Figura A.1: *Centroides de los radios censales, identificados en K=5 grupos, según el algoritmo de Elliott mediante tasa de desocupación, proporción de PEA y de jefes o jefas de hogar con hasta secundario incompleto por radio. Rosario. Año 2010.*

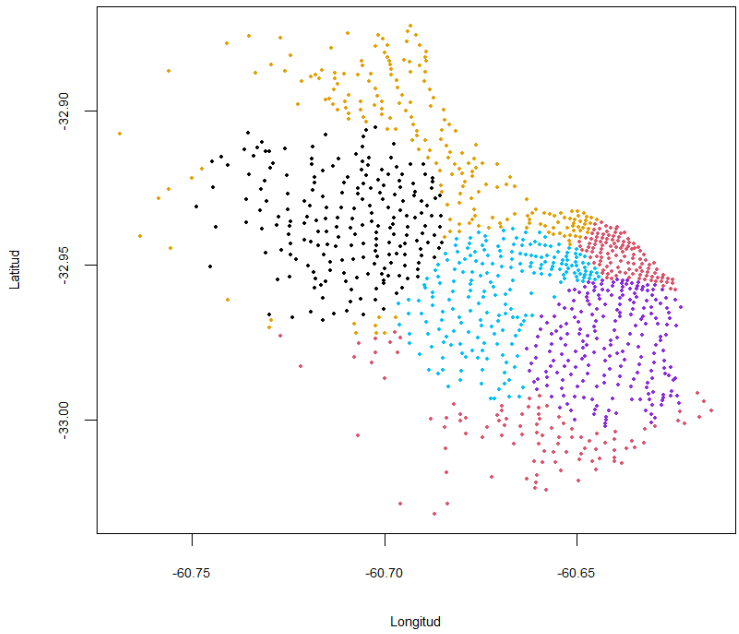


Figura A.2: *Centroides de los radios censales, identificados en K=5 grupos, según el algoritmo de Elliott mediante proporción de jefes o jefas de hogar con hasta secundario incompleto por radio. Rosario. Año 2010.*

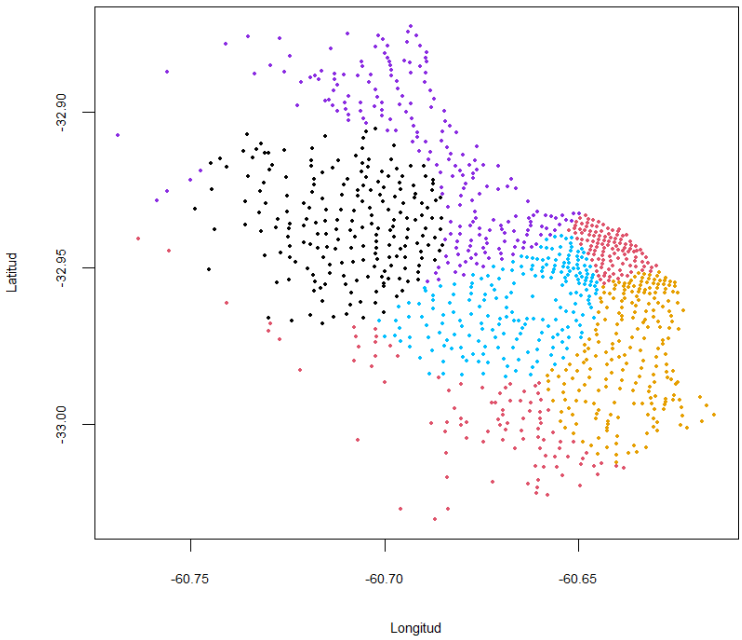


Figura A.3: *Centroides de los radios censales, identificados en K=5 grupos, según el algoritmo de Elliott mediante tasa de desocupación y proporción de PEA de mujeres por radio. Rosario. Año 2010.*

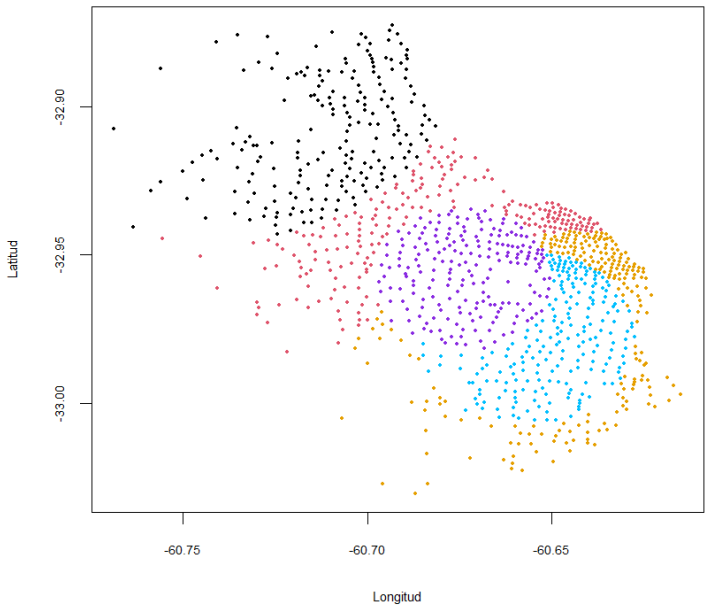


Figura A.4: *Centroides de los radios censales, identificados en K=5 grupos, según el algoritmo de Elliott mediante proporción de PEA de mujeres por radio. Rosario. Año 2010.*

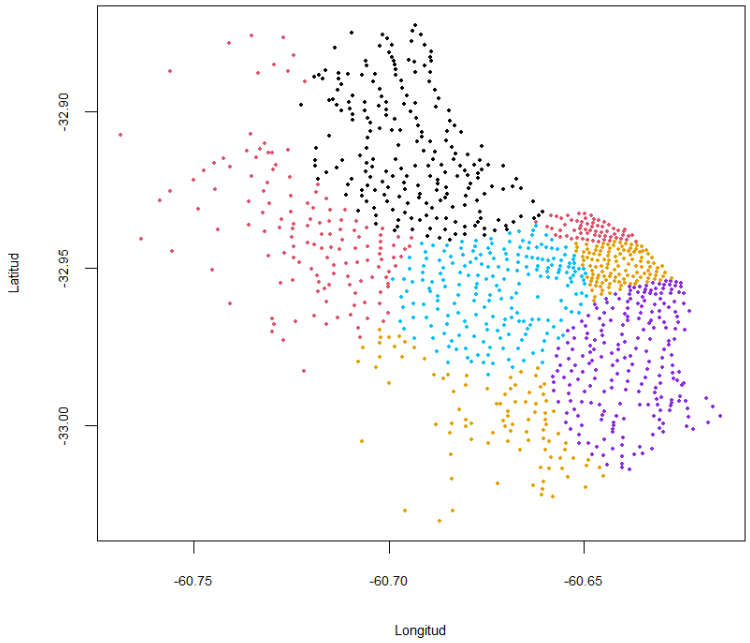


Figura A.5: *Centroides de los radios censales, identificados en K=5 grupos, según el algoritmo de Elliott mediante proporción de PEA por radio. Rosario. Año 2010.*

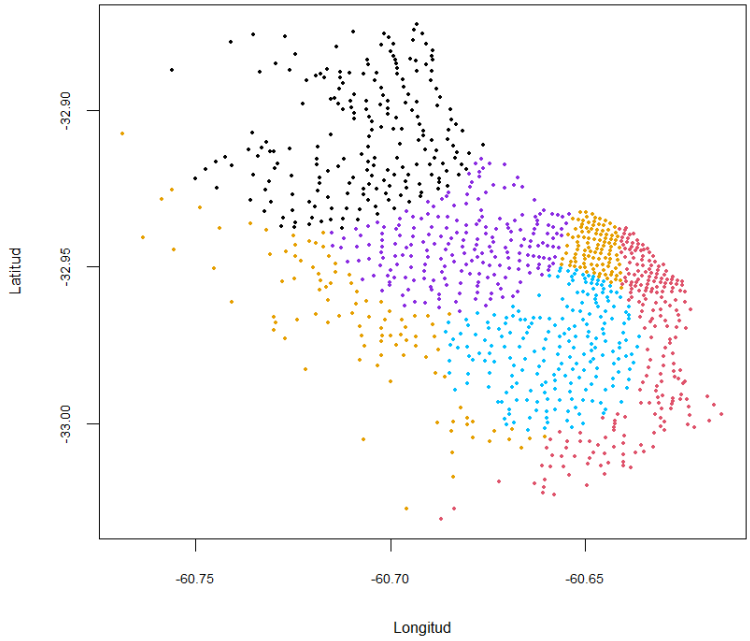
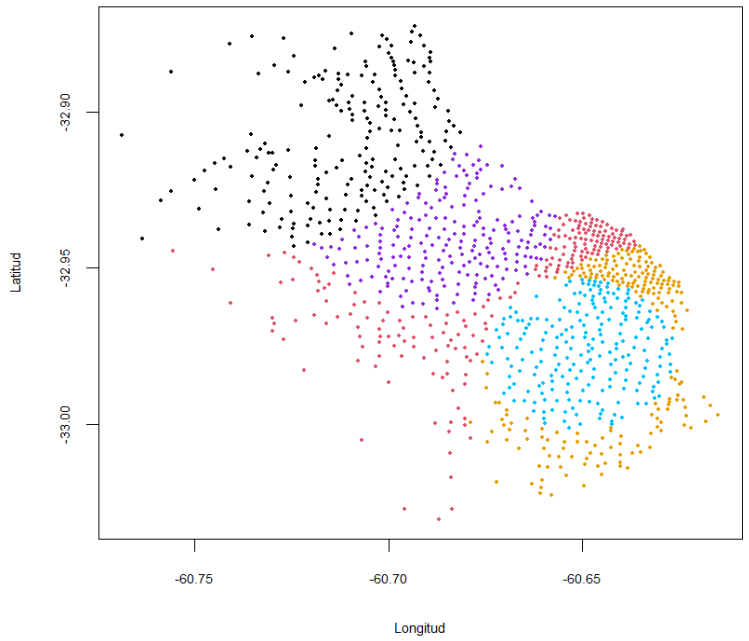


Figura A.6: *Centroides de los radios censales, identificados en K=5 grupos, según el algoritmo de Elliott mediante tasa de desocupación por radio. Rosario. Año 2010.*



A.2. Coeficientes de Variación según estratificación de conglomerados (de 5.2.2)

Reconquista CV(Tasas)

Tabla A.1: Principales indicadores del mercado de trabajo, parámetro y CV según método de estratificación por cantidad de estratos en la primera etapa de muestreo. Reconquista. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		n° de estratos			n° de estratos			n° de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Actividad										
Tasa de la población total	46,7	3,70	3,79	3,79	3,82	3,82	3,79	3,78	3,71	3,70
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	63,4	3,08	3,09	3,10	3,12	3,13	3,15	3,02	3,02	3,01
Mujeres	52,1	4,85	4,83	4,85	5,00	4,98	5,01	4,56	4,56	4,52
Varones	75,8	3,04	3,07	3,08	3,11	3,12	3,13	3,08	3,08	3,08
Jefes o jefas de hogar	76,3	2,75	2,77	2,76	2,82	2,86	2,85	2,87	2,84	2,83
Mujeres de 14 a 29 años	46,1	8,16	8,11	8,12	8,31	8,29	8,33	7,86	7,87	7,83
Mujeres de 30 a 64 años	63,8	5,02	5,03	5,00	5,16	5,13	5,18	4,78	4,75	4,67
Varones de 14 a 29 años	65,0	5,77	5,81	5,83	5,87	5,89	5,87	5,81	5,80	5,83
Varones de 30 a 64 años	91,5	2,44	2,42	2,43	2,43	2,43	2,46	2,41	2,42	2,41
Empleo										
Tasa de la población total	43,7	3,72	3,81	3,80	3,83	3,83	3,81	3,80	3,75	3,75
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	59,3	3,27	3,28	3,28	3,31	3,31	3,34	3,20	3,20	3,18
Mujeres	47,5	5,29	5,28	5,28	5,44	5,43	5,46	4,97	4,95	4,90
Varones	72,3	3,24	3,26	3,27	3,31	3,32	3,33	3,28	3,28	3,29
Jefes o jefas de hogar	74,0	2,91	2,93	2,91	2,98	3,01	3,00	3,02	3,00	3,00
Mujeres de 14 a 29 años	38,5	9,47	9,42	9,41	9,59	9,58	9,61	9,21	9,19	9,15
Mujeres de 30 a 64 años	60,3	5,39	5,40	5,37	5,53	5,50	5,55	5,13	5,08	5,00
Varones de 14 a 29 años	59,4	6,41	6,45	6,47	6,54	6,53	6,52	6,44	6,45	6,48
Varones de 30 a 64 años	89,0	2,65	2,63	2,64	2,65	2,64	2,68	2,63	2,63	2,62
Desocupación										
Tasa de la población total	6,4	16,49	16,31	16,36	16,47	16,48	16,44	16,40	16,25	16,28
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	6,4	16,49	16,31	16,36	16,47	16,48	16,44	16,40	16,25	16,28
Mujeres	8,8	19,72	19,56	19,60	19,69	19,71	19,64	19,53	19,34	19,30
Varones	4,6	24,23	24,02	24,02	24,23	24,13	24,14	24,16	24,12	24,19
Jefes o jefas de hogar	3,1	30,52	30,53	30,31	30,64	30,60	30,56	30,32	30,31	30,27
Mujeres de 14 a 29 años	16,6	23,10	22,91	22,93	23,00	23,03	22,98	22,90	22,86	22,85
Mujeres de 30 a 64 años	5,5	29,91	29,73	29,75	29,78	29,77	29,75	29,65	29,51	29,51
Varones de 14 a 29 años	8,7	29,11	28,85	28,92	29,16	29,00	28,92	28,95	28,96	29,04
Varones de 30 a 64 años	2,6	37,63	37,56	37,41	37,60	37,52	37,47	37,55	37,56	37,55

Gráfico A.1: Coeficiente de variación de la tasa específica de actividad según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y según población total, según nivel de instrucción, según sexo y edad y según relación de parentesco con el jefe o jefa del hogar cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Reconquista. Año 2010

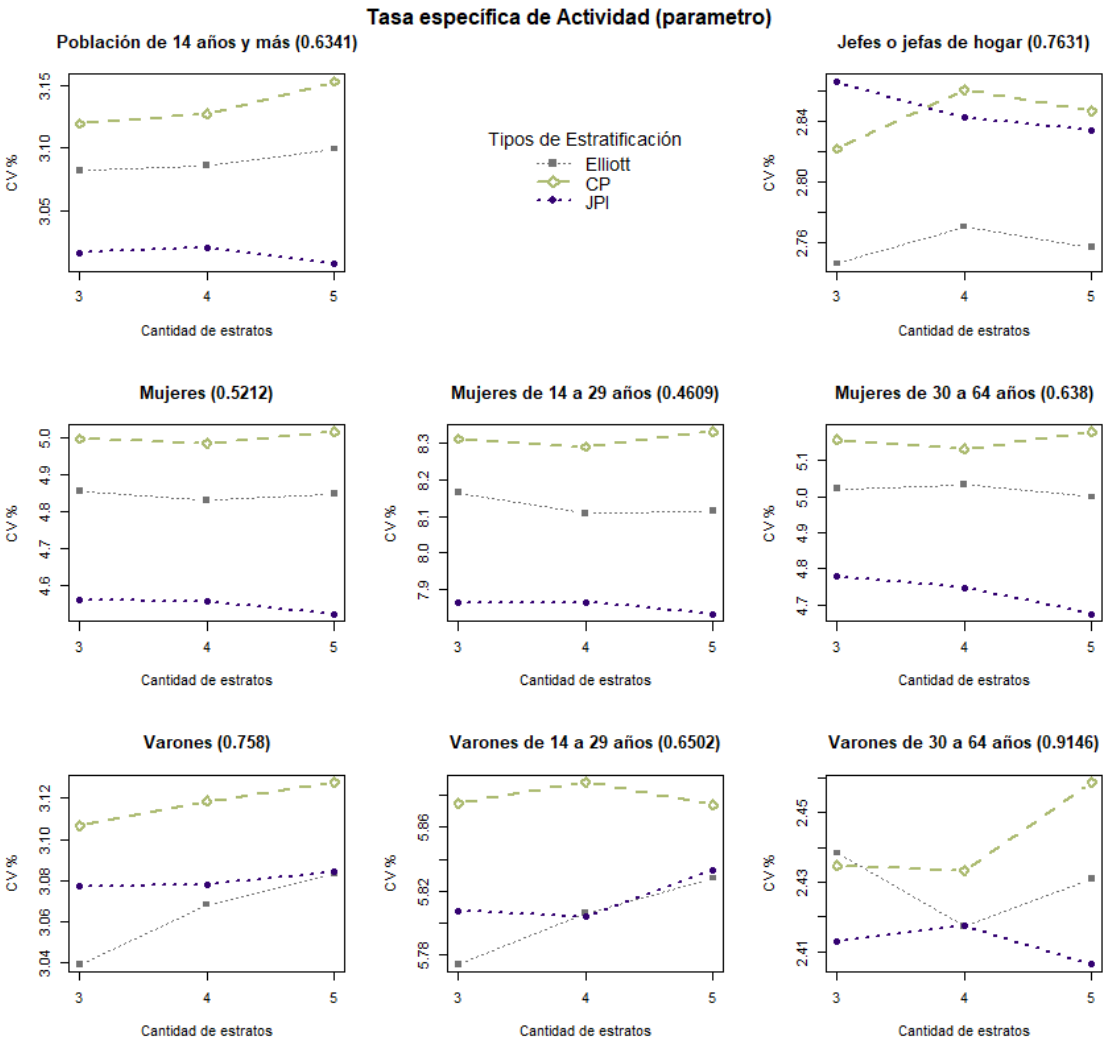


Gráfico A.2: Coeficiente de variación de la tasa específica de empleo según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Reconquista. Año 2010

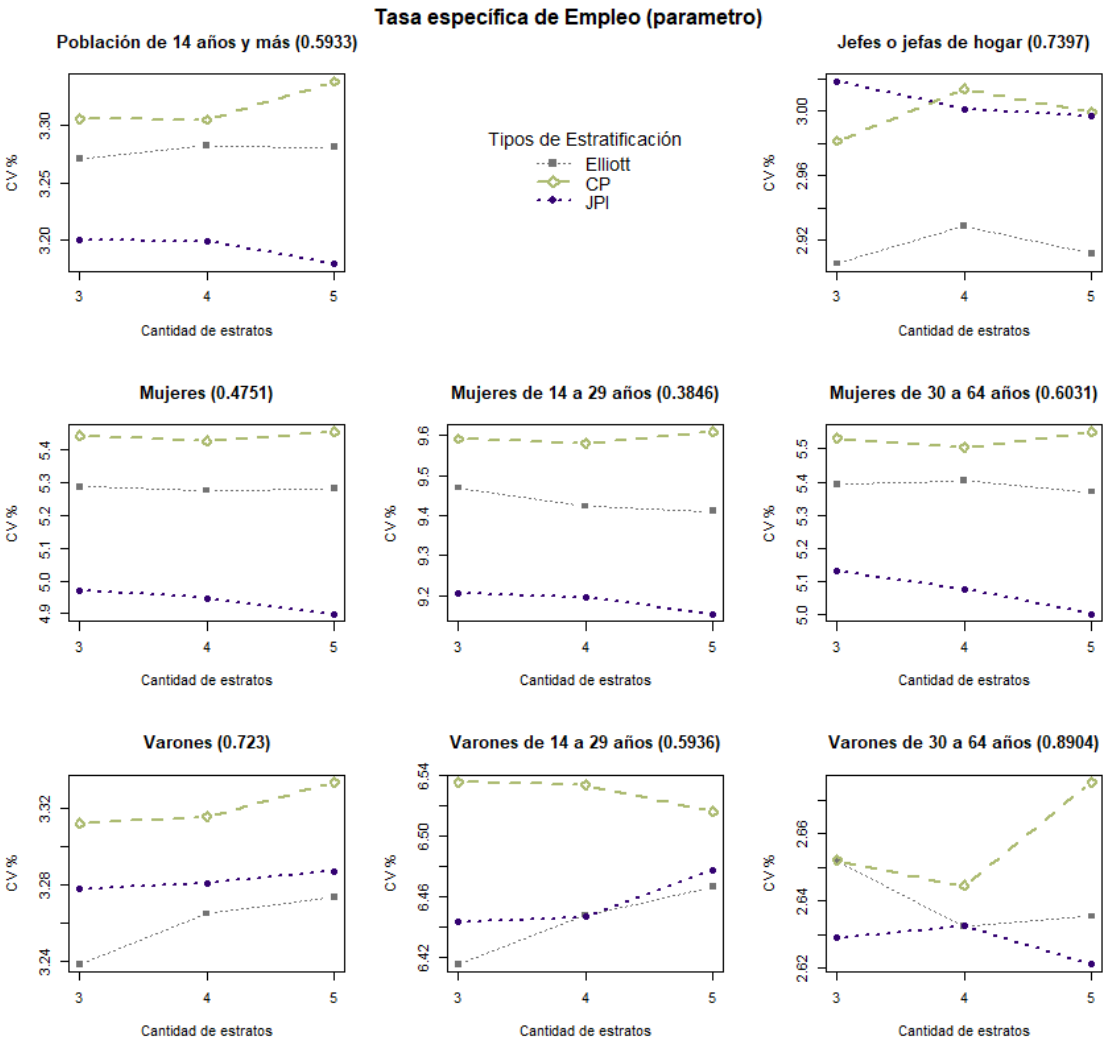


Gráfico A.3: Coeficiente de variación de la tasa específica de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Reconquista. Año 2010

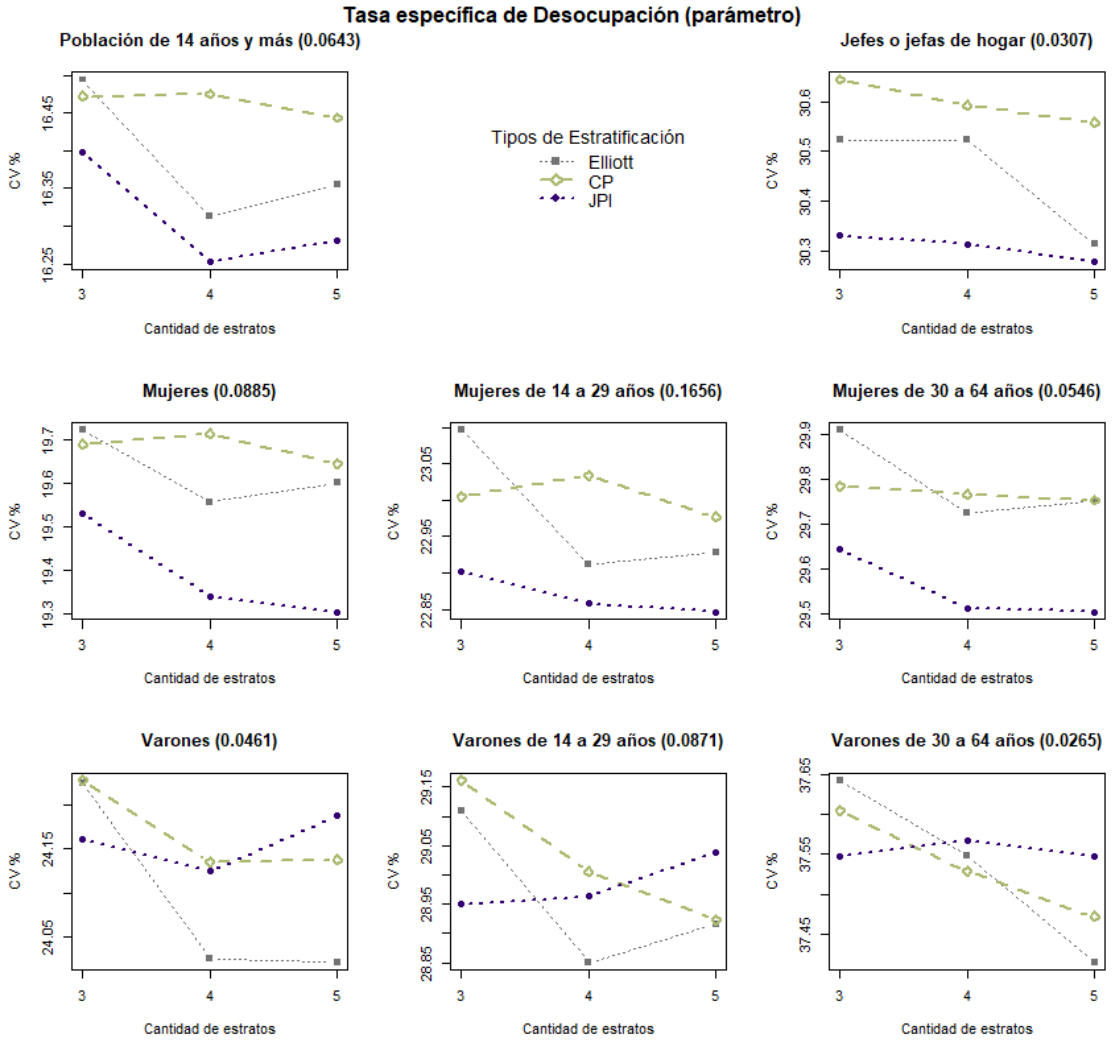


Tabla A.2: Composición de la tasa de empleo, parámetro y CV según sexo y edad, relación de parentesco con el jefe o jefa de hogar y nivel de instrucción, por método de estratificación y cantidad de estratos en la primera etapa de muestreo. Reconquista. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		nº de estratos			nº de estratos			nº de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Tasa de empleo	46,7	3,70	3,79	3,79	3,82	3,82	3,79	3,78	3,71	3,70
Sexo y grupos de edad										
Mujeres hasta 29 años	63,4	3,08	3,09	3,10	3,12	3,13	3,15	3,02	3,02	3,01
Mujeres de 30 a 64 años	52,1	4,85	4,83	4,85	5,00	4,98	5,01	4,56	4,56	4,52
Mujeres de 65 años y más	75,8	3,04	3,07	3,08	3,11	3,12	3,13	3,08	3,08	3,08
Varones hasta 29 años	76,3	2,75	2,77	2,76	2,82	2,86	2,85	2,87	2,84	2,83
Varones de 30 a 64 años	46,1	8,16	8,11	8,12	8,31	8,29	8,33	7,86	7,87	7,83
Varones de 65 años y más	63,8	5,02	5,03	5,00	5,16	5,13	5,18	4,78	4,75	4,67
Relación de parentesco	65,0	5,77	5,81	5,83	5,87	5,89	5,87	5,81	5,80	5,83
Jefe o jefa de hogar	91,5	2,44	2,42	2,43	2,43	2,43	2,46	2,41	2,42	2,41
Cónyuge										
Hijos	43,7	3,72	3,81	3,80	3,83	3,83	3,81	3,80	3,75	3,75
Otros componentes										
Nivel educativo	59,3	3,27	3,28	3,28	3,31	3,31	3,34	3,20	3,20	3,18
Primario incompleto	47,5	5,29	5,28	5,28	5,44	5,43	5,46	4,97	4,95	4,90
Primario completo	72,3	3,24	3,26	3,27	3,31	3,32	3,33	3,28	3,28	3,29
Secundario incompleto	74,0	2,91	2,93	2,91	2,98	3,01	3,00	3,02	3,00	3,00
Secundario completo	38,5	9,47	9,42	9,41	9,59	9,58	9,61	9,21	9,19	9,15
Superior y universitario incompleto	60,3	5,39	5,40	5,37	5,53	5,50	5,55	5,13	5,08	5,00
Superior y universitario completo	59,4	6,41	6,45	6,47	6,54	6,53	6,52	6,44	6,45	6,48
Sin instrucción	89,0	2,65	2,63	2,64	2,65	2,64	2,68	2,63	2,63	2,62

Gráfico A.4: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Reconquista. Año 2010. (Continúa)

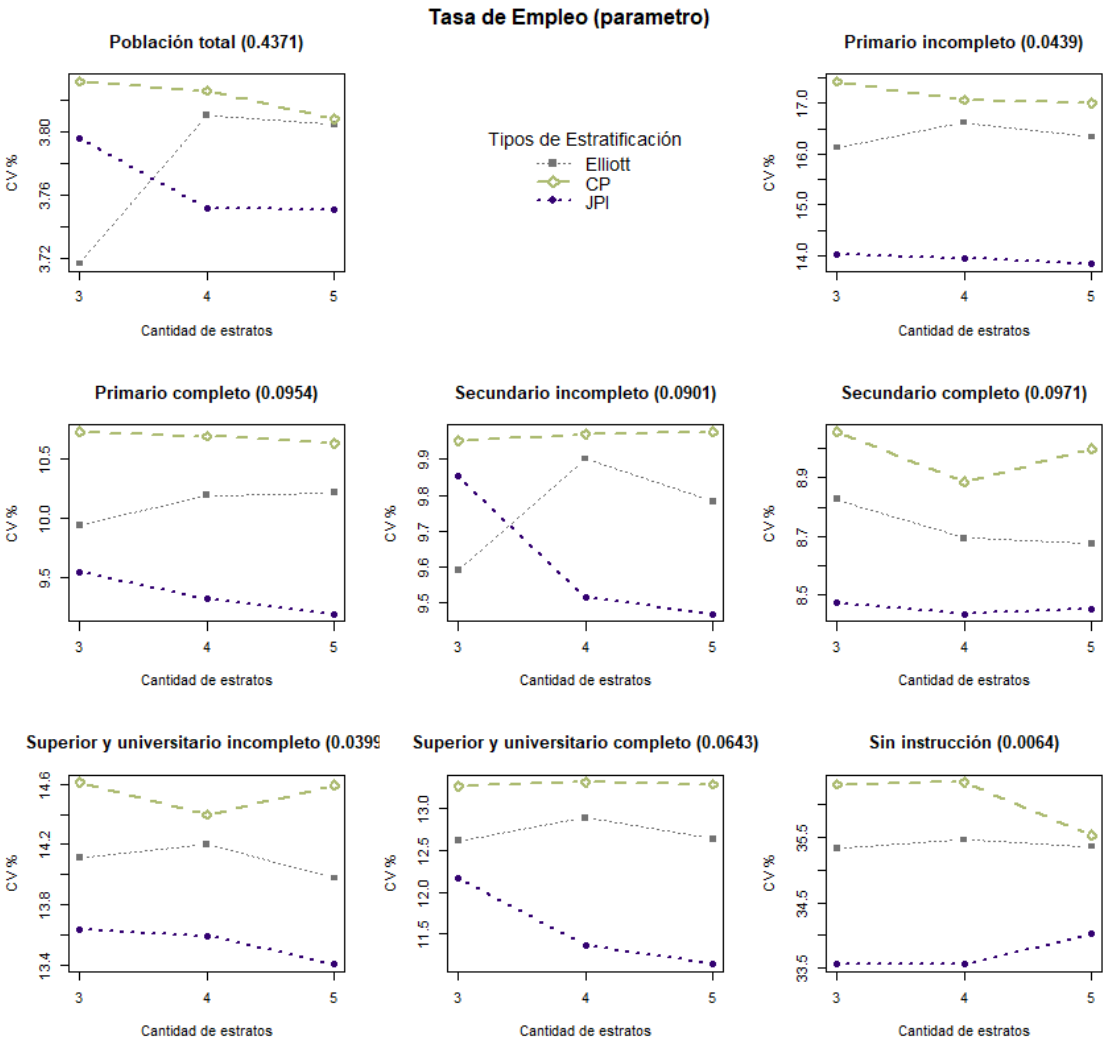
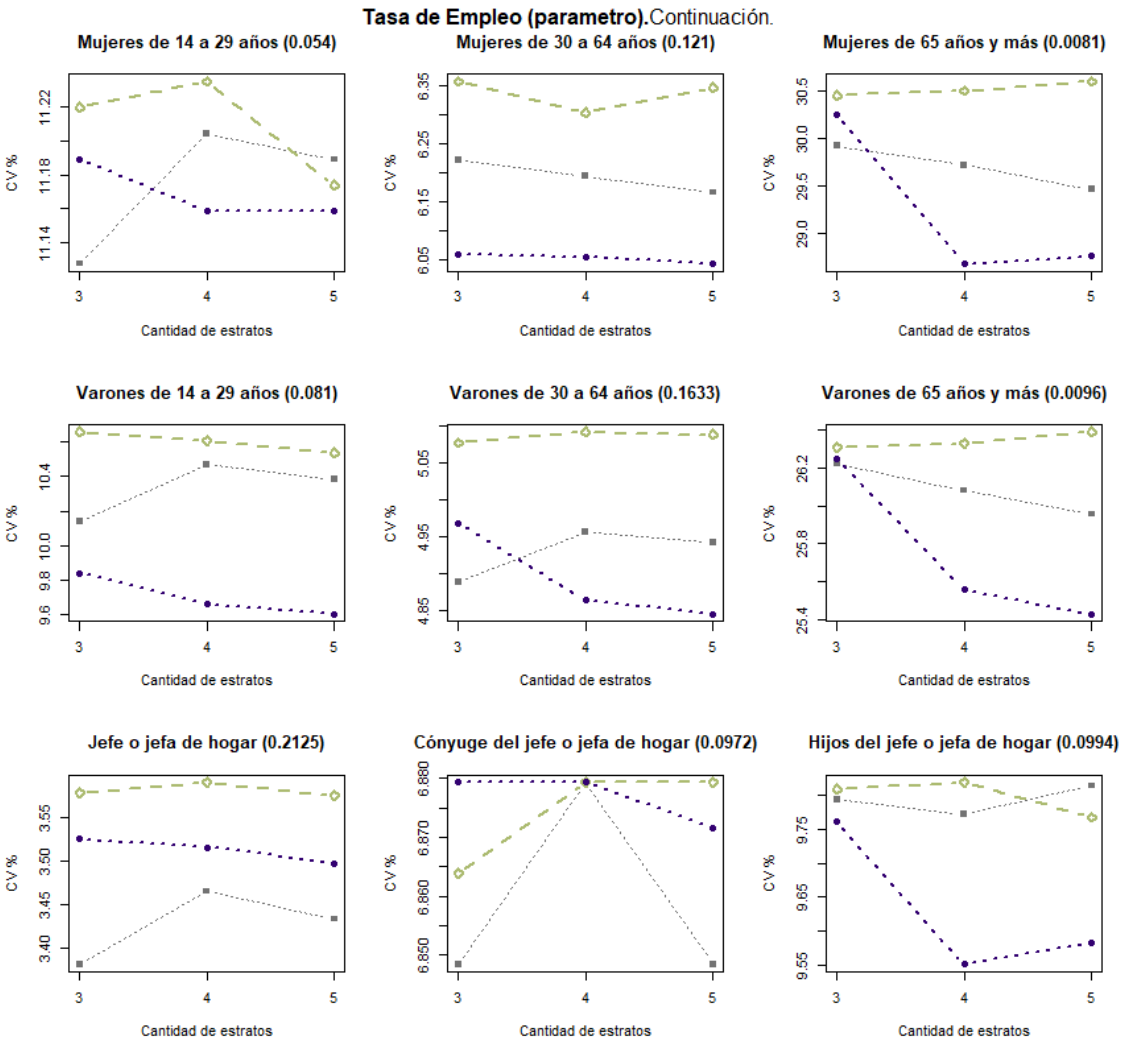


Gráfico A.4: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Reconquista. Año 2010. (Continuación)



Rosario CV(Tasas)

Tabla A.3: Principales indicadores del mercado de trabajo, parámetro y CV según método de estratificación por cantidad de estratos en la primera etapa de muestreo. Rosario. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		n° de estratos			n° de estratos			n° de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Actividad										
Tasa de la población total	54,2	3,44	3,49	3,48	3,37	3,36	3,34	3,35	3,34	3,34
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	66,7	2,69	2,68	2,67	2,70	2,71	2,70	2,71	2,69	2,71
Mujeres	57,5	3,59	3,59	3,57	3,58	3,60	3,59	3,61	3,57	3,60
Varones	77,2	3,09	3,08	3,09	3,12	3,13	3,12	3,13	3,11	3,13
Jefes o jefas de hogar	73,8	2,26	2,26	2,25	2,26	2,28	2,28	2,27	2,26	2,28
Mujeres de 14 a 29 años	57,0	5,67	5,64	5,62	5,58	5,61	5,60	5,63	5,58	5,61
Mujeres de 30 a 64 años	72,2	3,30	3,29	3,27	3,25	3,27	3,27	3,27	3,24	3,26
Varones de 14 a 29 años	70,5	5,21	5,19	5,20	5,25	5,28	5,27	5,26	5,26	5,27
Varones de 30 a 64 años	92,4	2,63	2,62	2,64	2,66	2,66	2,65	2,67	2,64	2,65
Empleo										
Tasa de la población total	50,8	3,46	3,50	3,50	3,42	3,40	3,39	3,40	3,39	3,38
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	62,4	2,82	2,82	2,80	2,83	2,84	2,83	2,84	2,82	2,84
Mujeres	52,6	3,90	3,89	3,86	3,86	3,87	3,86	3,88	3,85	3,87
Varones	73,6	3,23	3,21	3,22	3,26	3,27	3,26	3,27	3,25	3,26
Jefes o jefas de hogar	71,2	2,41	2,41	2,40	2,41	2,43	2,43	2,43	2,41	2,43
Mujeres de 14 a 29 años	48,7	6,66	6,63	6,60	6,53	6,57	6,57	6,59	6,54	6,58
Mujeres de 30 a 64 años	68,0	3,63	3,61	3,58	3,55	3,57	3,56	3,57	3,54	3,56
Varones de 14 a 29 años	64,5	5,68	5,66	5,66	5,72	5,75	5,75	5,73	5,73	5,75
Varones de 30 a 64 años	89,7	2,79	2,78	2,79	2,81	2,81	2,80	2,82	2,79	2,80
Desocupación										
Tasa de la población total	6,4	12,61	12,54	12,50	12,44	12,56	12,56	12,51	12,47	12,55
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	6,4	12,61	12,54	12,50	12,44	12,56	12,56	12,51	12,47	12,55
Mujeres	8,4	15,36	15,27	15,21	15,13	15,28	15,28	15,22	15,16	15,26
Varones	4,7	18,86	18,76	18,73	18,79	18,98	18,99	18,91	18,84	18,97
Jefes o jefas de hogar	3,6	22,58	22,47	22,41	22,45	22,64	22,64	22,56	22,45	22,57
Mujeres de 14 a 29 años	14,6	19,44	19,37	19,32	19,39	19,58	19,60	19,50	19,44	19,56
Mujeres de 30 a 64 años	5,9	22,10	21,95	21,89	21,87	22,08	22,09	22,01	21,90	22,06
Varones de 14 a 29 años	8,5	23,76	23,64	23,60	23,78	24,01	24,01	23,91	23,82	23,97
Varones de 30 a 64 años	2,9	29,74	29,59	29,56	29,61	29,91	29,95	29,79	29,70	29,91

Gráfico A.5: Coeficiente de variación de la tasa específica de actividad según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rosario. Año 2010

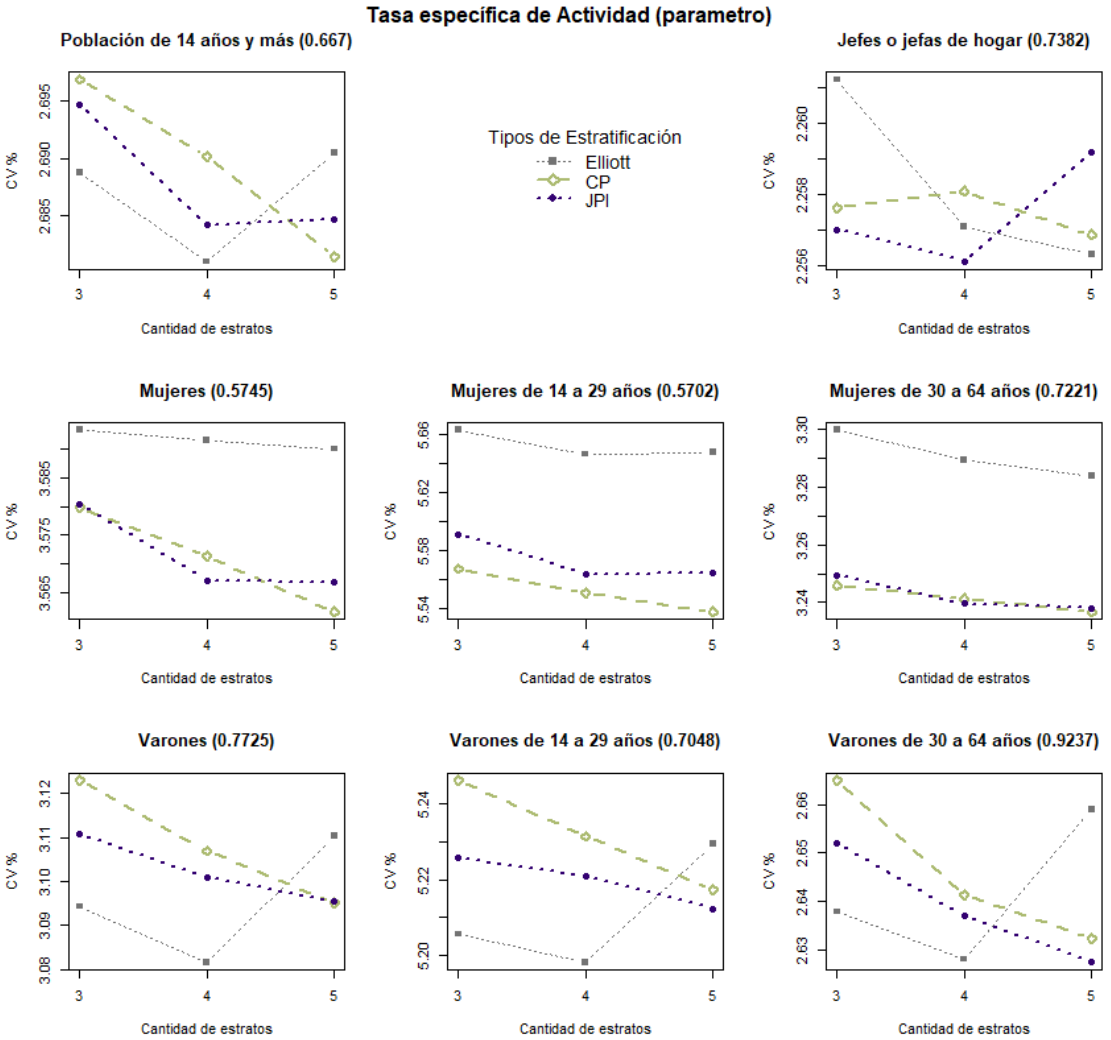


Gráfico A.6: Coeficiente de variación de la tasa específica de empleo según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rosario. Año 2010

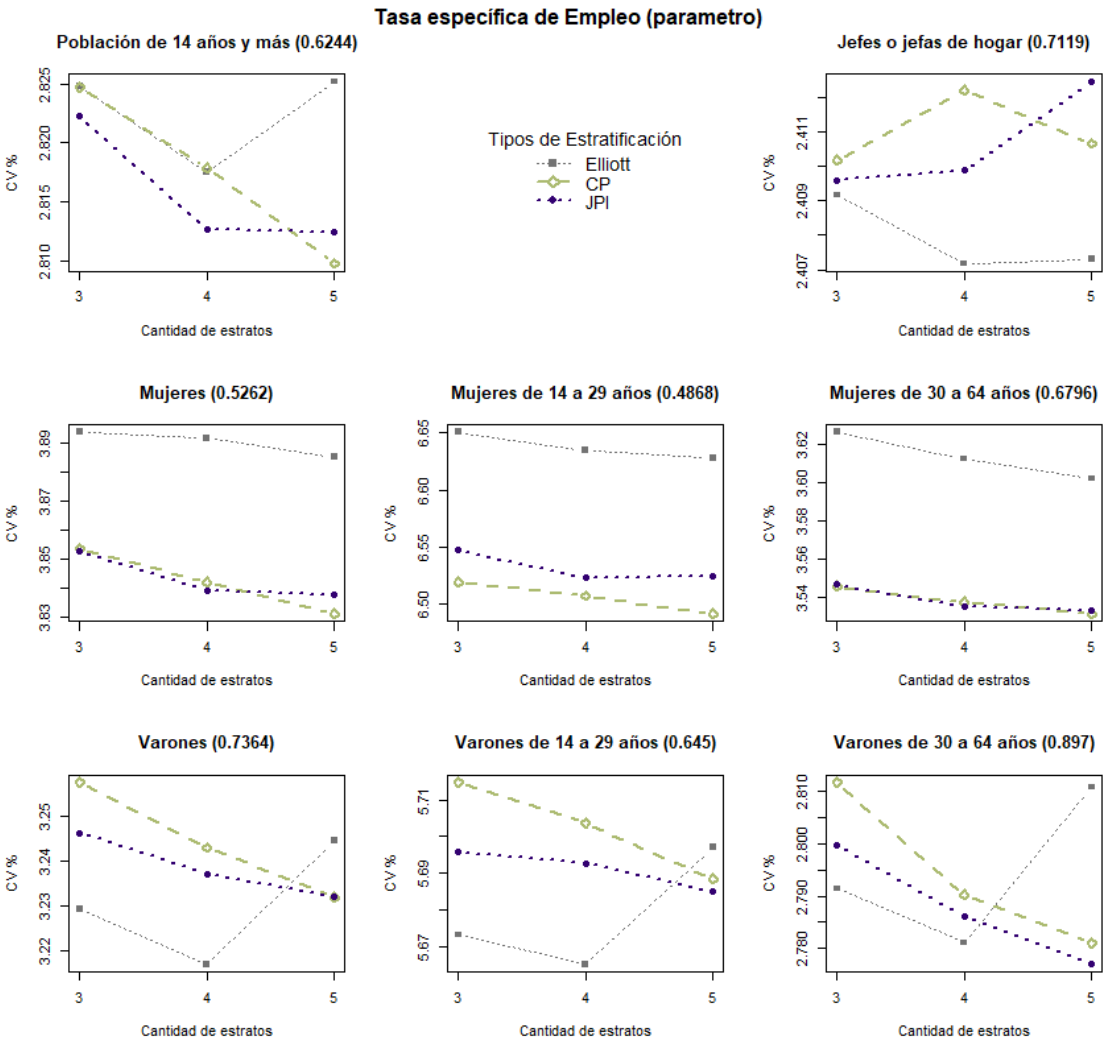


Gráfico A.7: Coeficiente de variación de la tasa específica de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rosario. Año 2010

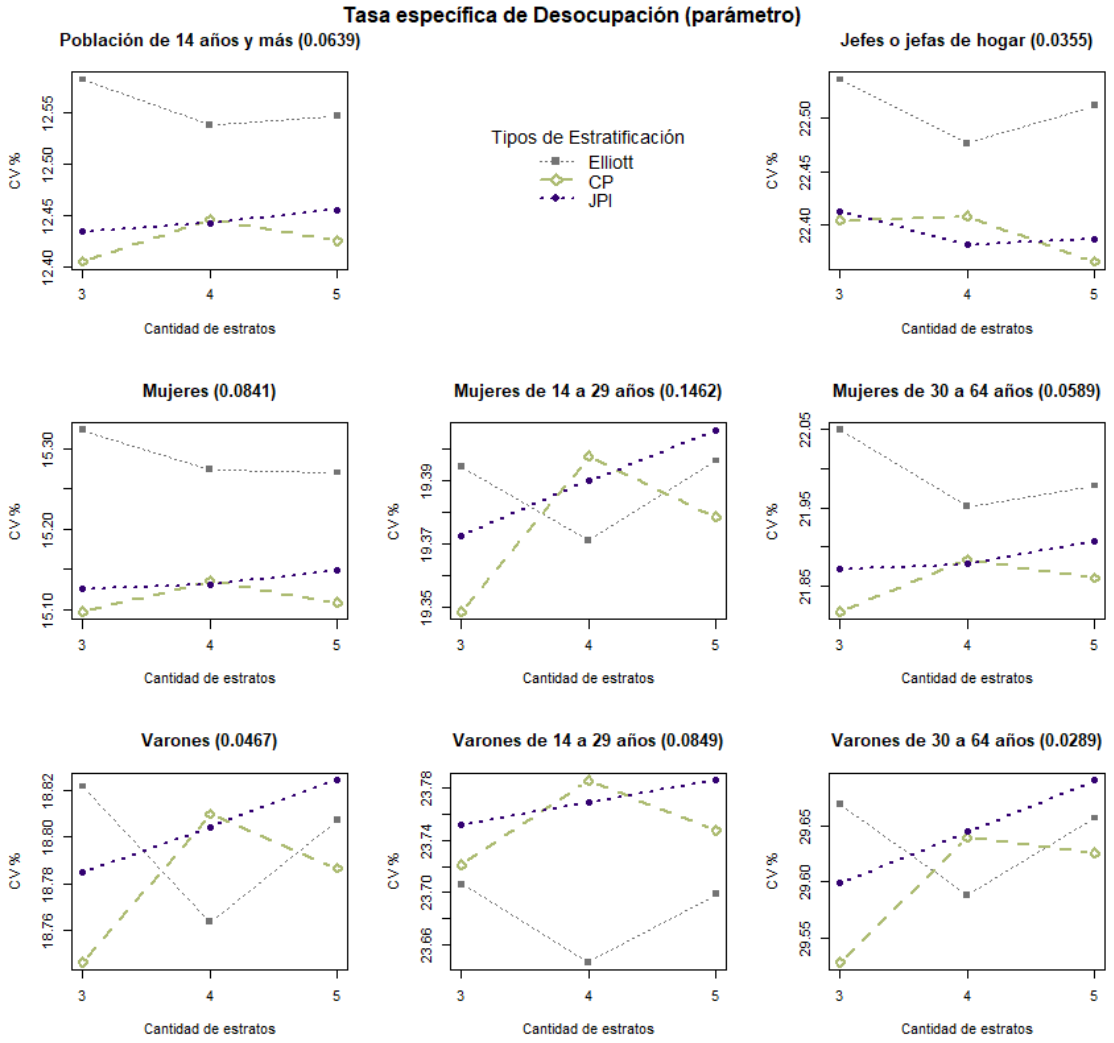


Tabla A.4: Composición de la tasa de empleo, parámetro y CV según sexo y edad, relación de parentesco con el jefe o jefa de hogar y nivel de instrucción, por método de estratificación y cantidad de estratos en la primera etapa de muestreo. Rosario. Año 2010

Tasas		Métodos de estratificación										
		Elliott				CP				JPI		
		n° de estratos				n° de estratos				n° de estratos		
		3	4	5		3	4	5		3	4	5
	parámetro %	CV %										
Tasa de empleo	50,8	3,46	3,50	3,50	0,00	3,42	3,40	3,39	0,00	3,40	3,39	3,38
Sexo y grupos de edad												
Mujeres hasta 29 años	6,7	8,82	8,81	8,93	0,00	8,71	8,71	8,71	0,00	8,70	8,71	8,71
Mujeres de 30 a 64 años	14,6	4,85	4,86	4,92	0,00	4,87	4,86	4,84	0,00	4,86	4,86	4,85
Mujeres de 65 años y más	1,5	17,07	17,03	17,26	0,00	17,01	17,01	16,99	0,00	17,03	17,00	17,00
Varones hasta 29 años	8,7	8,56	8,54	8,60	0,00	8,08	8,06	8,04	0,00	8,08	8,07	8,07
Varones de 30 a 64 años	17,5	4,48	4,54	4,53	0,00	4,52	4,47	4,45	0,00	4,49	4,46	4,43
Varones de 65 años y más	1,7	15,91	15,88	16,11	0,00	15,84	15,84	15,81	0,00	15,87	15,84	15,84
Relación de parentesco												
Jefe o jefa de hogar	24,1	3,17	3,17	3,20	0,00	3,10	3,10	3,11	0,00	3,11	3,11	3,11
Cónyuge	11,4	5,51	5,53	5,59	0,00	5,55	5,53	5,53	0,00	5,54	5,54	5,52
Hijos	11,5	7,78	7,93	7,91	0,00	7,90	7,80	7,72	0,00	7,78	7,72	7,69
Otros componentes	3,7	13,38	13,37	13,56	0,00	13,24	13,22	13,23	0,00	13,19	13,24	13,24
Nivel educativo												
Primario incompleto	3,6	16,95	16,78	16,73	0,00	14,33	13,84	13,83	0,00	14,38	14,07	14,02
Primario completo	8,8	10,16	10,12	10,06	0,00	8,91	8,85	8,69	0,00	8,77	8,70	8,68
Secundario incompleto	8,1	9,17	9,37	9,29	0,00	9,02	8,92	8,81	0,00	8,84	8,80	8,77
Secundario completo	11,5	7,11	7,24	7,23	0,00	7,28	7,07	6,97	0,00	7,13	7,00	6,93
Superior y universitario incompleto	8,3	8,58	8,60	8,62	0,00	8,25	8,18	8,09	0,00	8,16	8,07	8,05
Superior y universitario completo	10,0	8,35	8,50	8,44	0,00	7,87	7,74	7,58	0,00	7,65	7,52	7,48
Sin instrucción	0,5	33,56	33,45	33,86	0,00	32,65	32,33	32,40	0,00	32,39	32,74	32,72

Gráfico A.8: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rafaela. Año 2010. (Continúa)

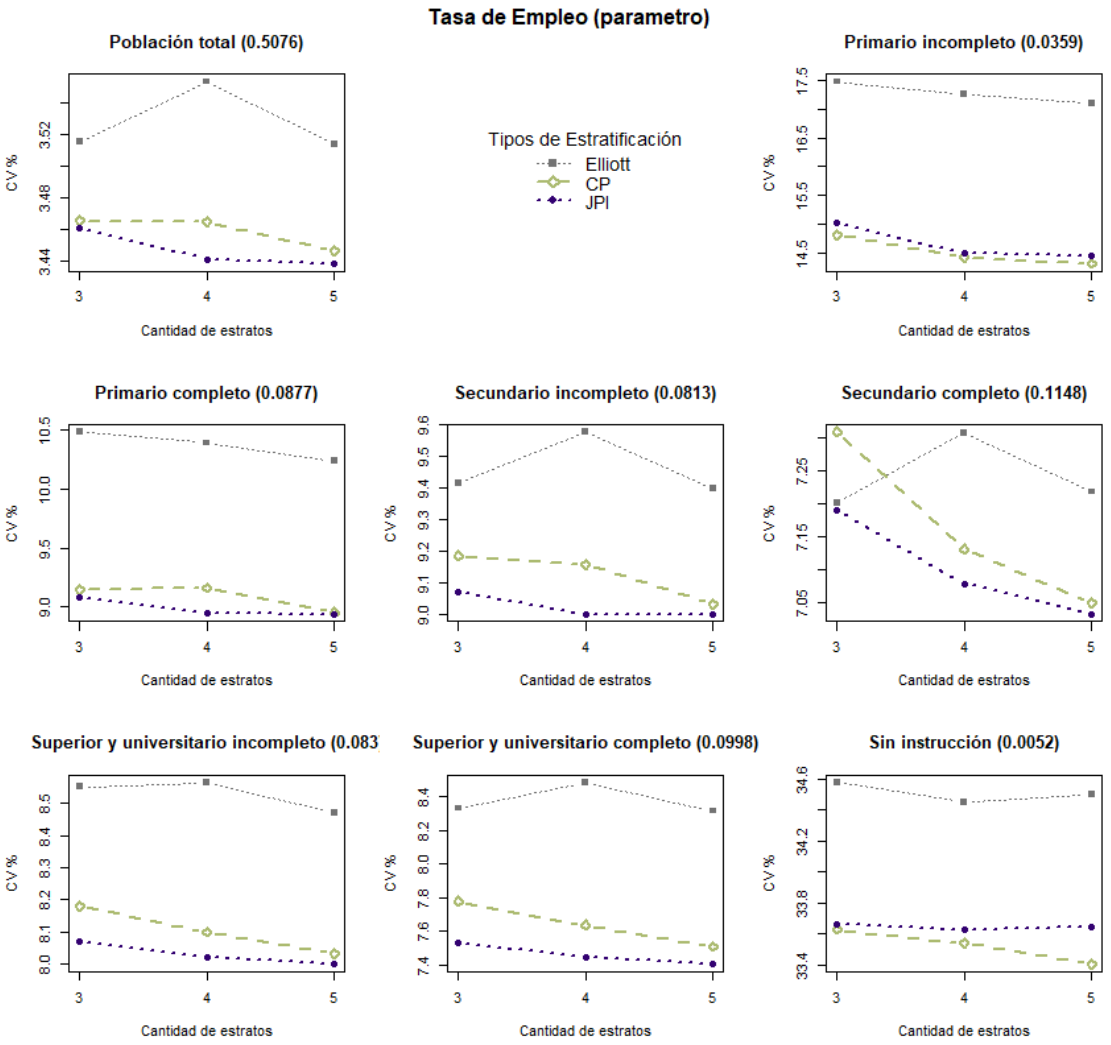
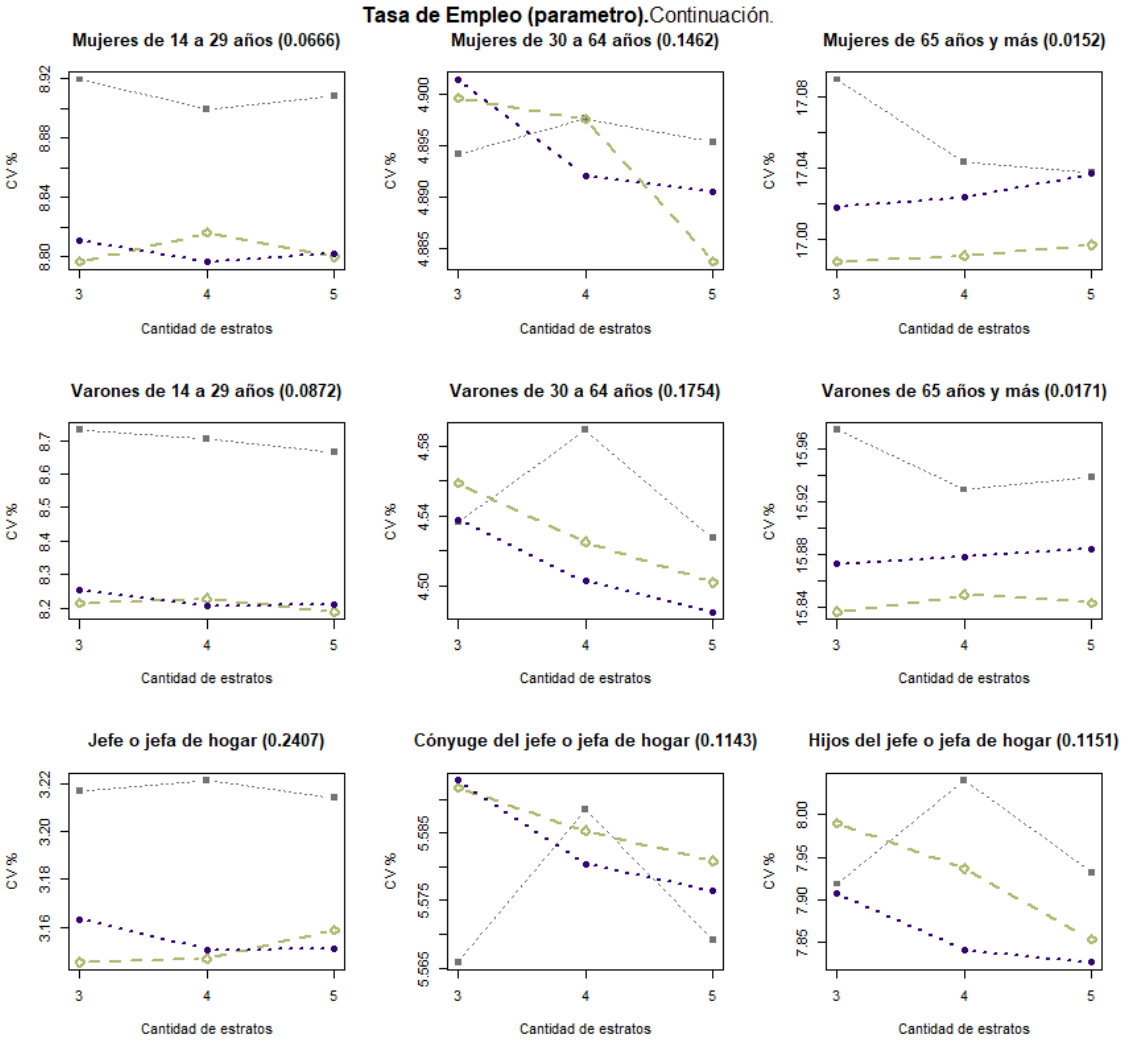


Gráfico A.8: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Rafaela. Año 2010. (Continuación)



Santa Fe CV(Tasas)

Tabla A.5: Principales indicadores del mercado de trabajo, parámetro y CV según método de estratificación por cantidad de estratos en la primera etapa de muestreo. Santa Fe. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		n° de estratos			n° de estratos			n° de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Actividad										
Tasa de la población total	50,1	3,28	3,20	3,25	3,29	3,61	3,25	3,23	3,39	3,30
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	64,4	3,15	3,10	3,09	3,21	3,50	3,17	3,16	3,26	3,21
Mujeres	55,7	3,78	3,72	3,71	3,79	4,03	3,86	3,84	3,91	3,81
Varones	74,4	4,25	4,20	4,19	4,47	5,15	4,20	4,15	4,49	4,48
Jefes o jefas de hogar	73,1	2,33	2,32	2,32	2,33	2,50	2,37	2,37	2,44	2,38
Mujeres de 14 a 29 años	51,5	6,25	6,18	6,15	6,33	6,96	6,40	6,29	6,53	6,43
Mujeres de 30 a 64 años	71,2	3,75	3,65	3,66	3,74	4,01	3,85	3,82	3,89	3,76
Varones de 14 a 29 años	64,7	6,31	6,25	6,24	6,62	7,62	6,29	6,18	6,70	6,66
Varones de 30 a 64 años	90,5	4,14	4,08	4,07	4,35	5,00	4,08	4,03	4,33	4,34
Empleo										
Tasa de la población total	46,2	3,31	3,25	3,28	3,34	3,66	3,31	3,29	3,44	3,36
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	59,4	3,30	3,25	3,24	3,36	3,67	3,32	3,30	3,41	3,37
Mujeres	50,1	4,14	4,07	4,06	4,13	4,43	4,19	4,17	4,27	4,17
Varones	70,1	4,38	4,33	4,32	4,60	5,28	4,33	4,28	4,63	4,61
Jefes o jefas de hogar	70,0	2,51	2,49	2,49	2,52	2,72	2,56	2,54	2,63	2,57
Mujeres de 14 a 29 años	42,0	7,45	7,39	7,37	7,56	8,31	7,62	7,50	7,81	7,69
Mujeres de 30 a 64 años	66,4	4,08	3,97	3,98	4,05	4,36	4,16	4,12	4,21	4,07
Varones de 14 a 29 años	58,1	6,81	6,74	6,73	7,12	8,16	6,82	6,70	7,22	7,17
Varones de 30 a 64 años	87,3	4,26	4,21	4,19	4,47	5,12	4,21	4,15	4,46	4,47
Desocupación										
Tasa de la población total	7,7	12,00	11,93	11,93	12,19	13,66	12,32	12,09	12,68	12,48
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	7,7	12,00	11,93	11,93	12,19	13,66	12,32	12,09	12,68	12,48
Mujeres	10,2	14,29	14,19	14,21	14,45	16,11	14,66	14,45	15,07	14,82
Varones	5,7	17,73	17,67	17,64	18,15	20,29	18,33	17,96	18,91	18,60
Jefes o jefas de hogar	4,2	21,40	21,28	21,25	21,90	24,64	21,94	21,47	22,56	22,38
Mujeres de 14 a 29 años	18,6	17,40	17,37	17,36	17,77	19,75	18,03	17,74	18,49	18,22
Mujeres de 30 a 64 años	6,7	20,93	20,84	20,85	21,20	23,54	21,62	21,32	22,26	21,80
Varones de 14 a 29 años	10,1	22,12	22,05	22,02	22,63	25,13	22,89	22,47	23,56	23,16
Varones de 30 a 64 años	3,5	27,28	27,23	27,15	27,91	31,28	28,24	27,71	29,21	28,70

Gráfico A.9: Coeficiente de variación de la tasa específica de actividad según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Santa Fe. Año 2010

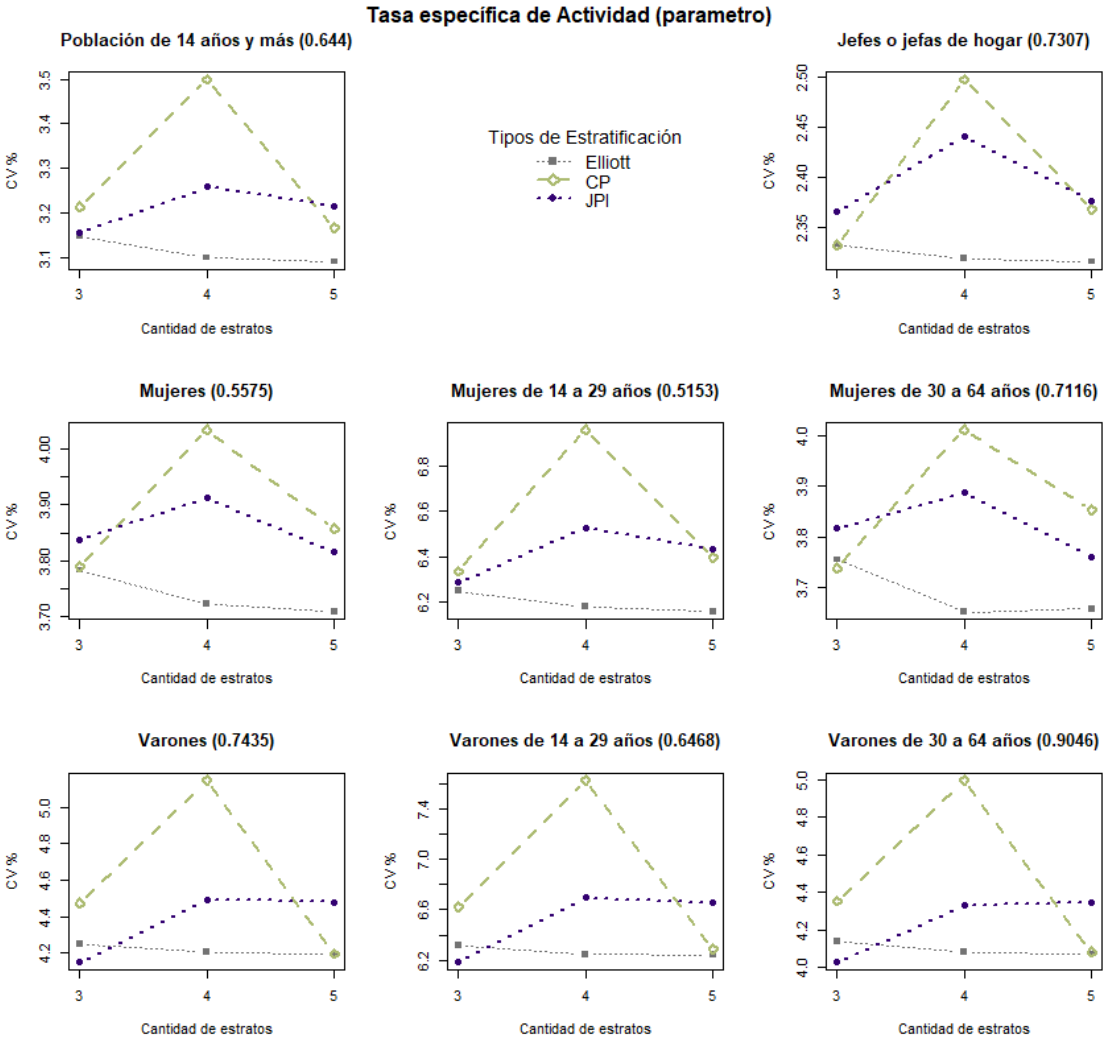


Gráfico A.10: Coeficiente de variación de la tasa específica de empleo según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Santa Fe. Año 2010

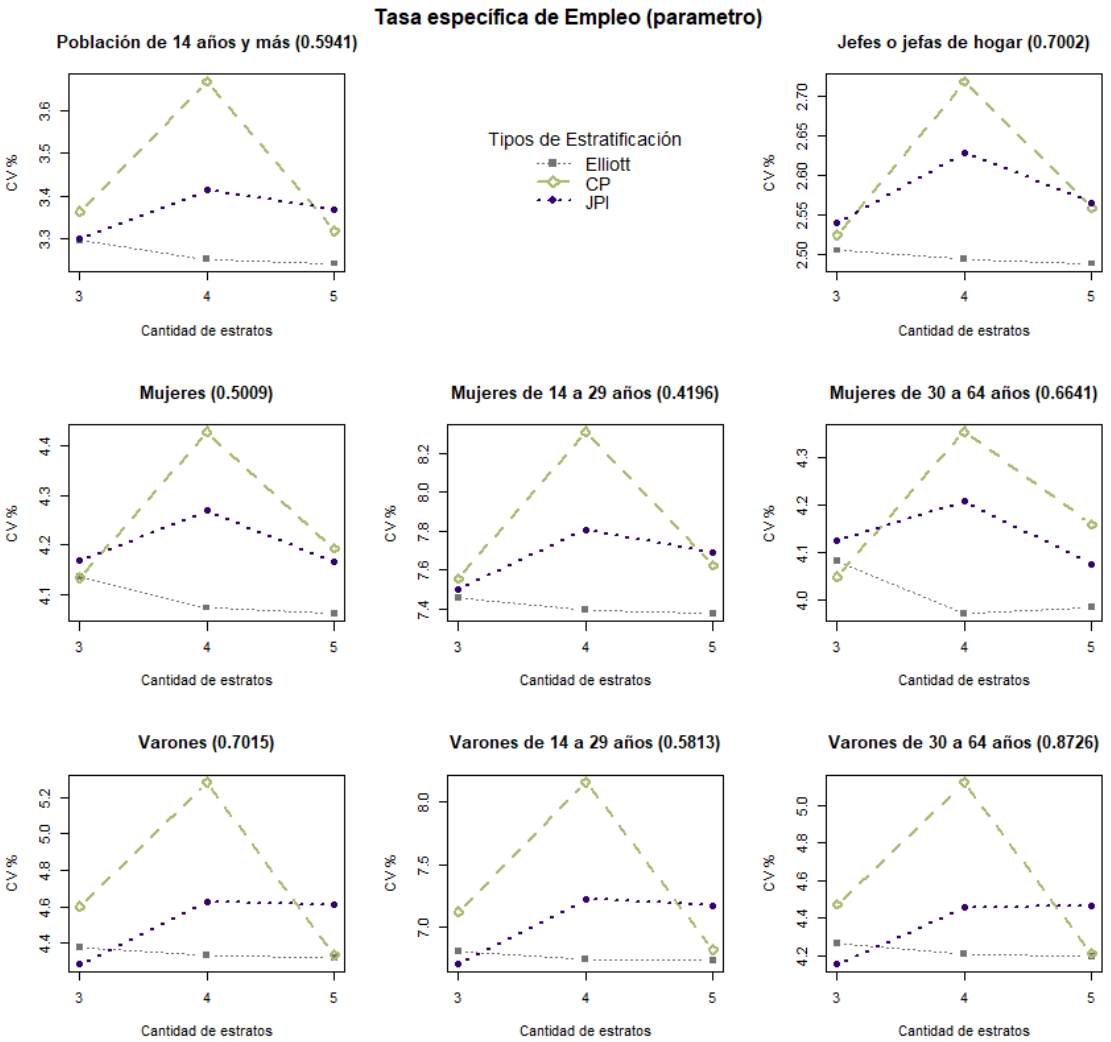


Gráfico A.11: Coeficiente de variación de la tasa específica de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Santa Fe. Año 2010

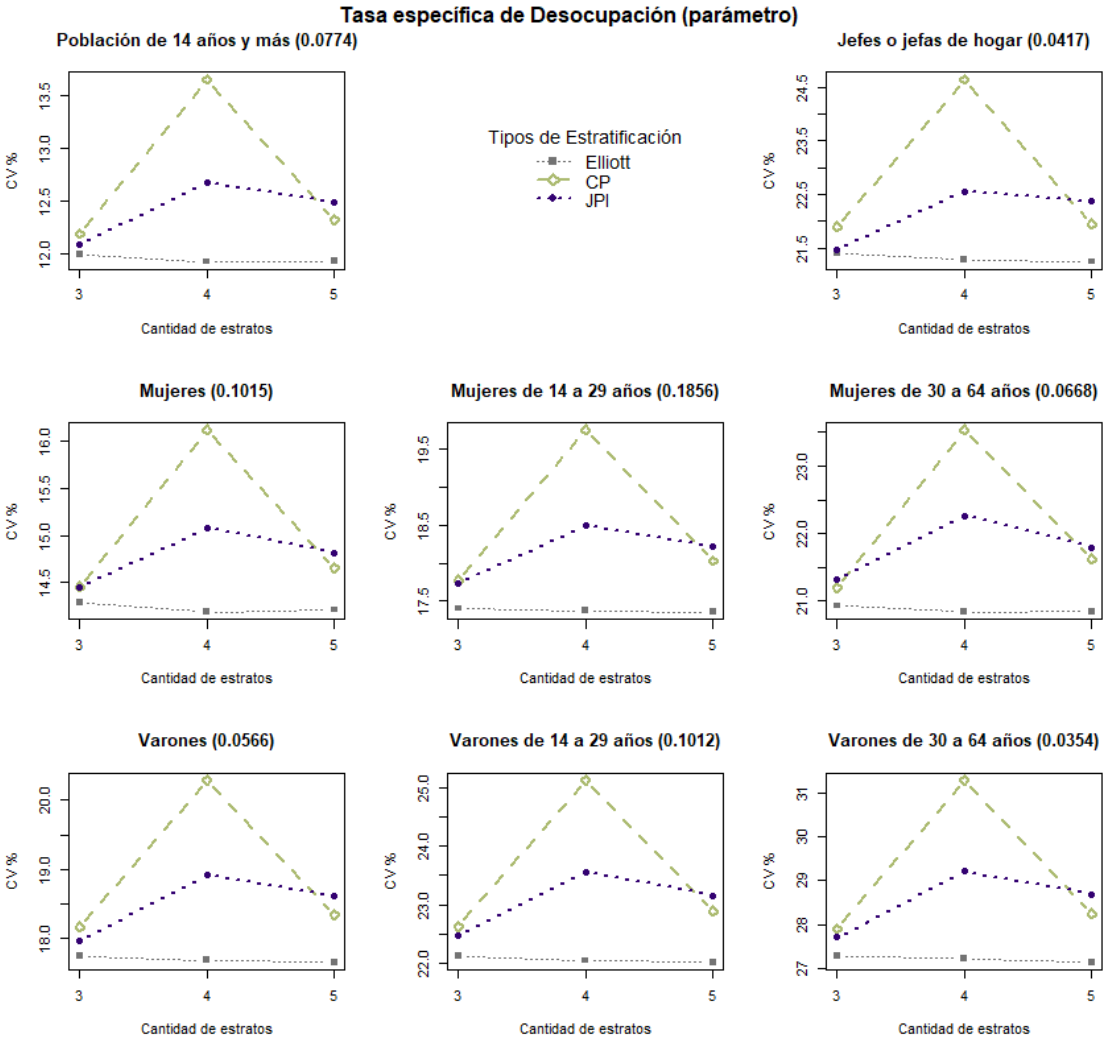


Tabla A.6: Composición de la tasa de empleo, parámetro y CV según sexo y edad, relación de parentesco con el jefe o jefa de hogar y nivel de instrucción, por método de estratificación y cantidad de estratos en la primera etapa de muestreo. Santa Fe. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		nº de estratos			nº de estratos			nº de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Tasa de empleo	46,2	3,31	3,25	3,28	3,34	3,66	3,31	3,29	3,44	3,36
Sexo y grupos de edad										
Mujeres hasta 29 años	5,8	9,26	9,21	9,23	9,34	10,21	9,42	9,32	9,68	9,49
Mujeres de 30 a 64 años	13,8	4,84	4,82	4,81	4,88	5,25	4,88	4,88	5,06	4,93
Mujeres de 65 años y más	1,2	19,14	19,06	19,06	18,97	19,74	19,30	19,43	19,94	19,25
Varones hasta 29 años	8,0	8,41	8,17	8,25	8,22	9,11	8,25	8,19	8,51	8,33
Varones de 30 a 64 años	16,2	4,45	4,37	4,43	4,50	4,85	4,46	4,48	4,62	4,51
Varones de 65 años y más	1,2	18,40	18,34	18,41	18,31	19,16	18,52	18,60	19,11	18,50
Relación de parentesco										
Jefe o jefa de hogar	22,4	3,04	3,00	3,00	3,06	3,37	3,05	3,03	3,18	3,11
Cónyuge	10,4	5,62	5,59	5,61	5,70	6,15	5,75	5,72	5,90	5,77
Hijos	10,4	8,06	7,98	8,06	8,20	8,87	8,21	8,13	8,47	8,23
Otros componentes	3,1	13,68	13,65	13,65	13,87	15,14	14,01	13,87	14,46	14,11
Nivel educativo										
Primario incompleto	3,1	16,40	15,42	15,43	15,28	17,90	14,54	14,61	15,51	15,44
Primario completo	8,4	9,92	9,12	9,60	9,10	10,23	8,98	9,07	9,21	8,98
Secundario incompleto	7,7	9,37	9,02	9,26	9,29	10,20	9,27	9,22	9,58	9,28
Secundario completo	11,0	7,22	7,03	7,05	7,23	7,46	7,06	7,21	7,41	7,13
Superior y universitario incompleto	6,6	9,31	8,96	9,20	8,92	8,95	8,87	9,09	9,03	8,64
Superior y universitario completo	8,8	8,99	8,30	8,75	8,20	8,00	7,95	8,06	7,90	7,48
Sin instrucción	0,5	33,37	32,95	32,80	34,21	39,96	33,62	32,49	35,07	34,90

Gráfico A.12: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Santa Fe. Año 2010. (Continúa)

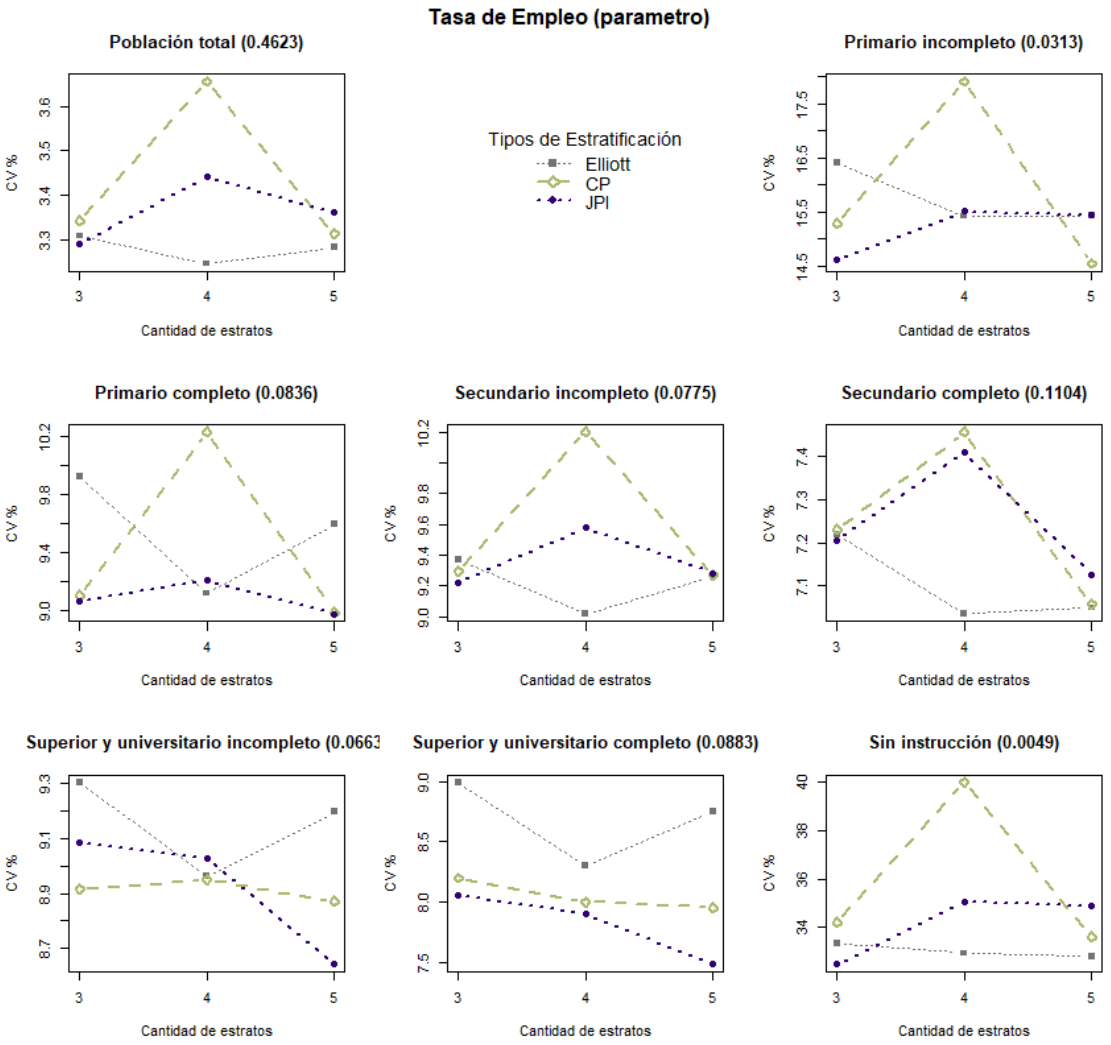
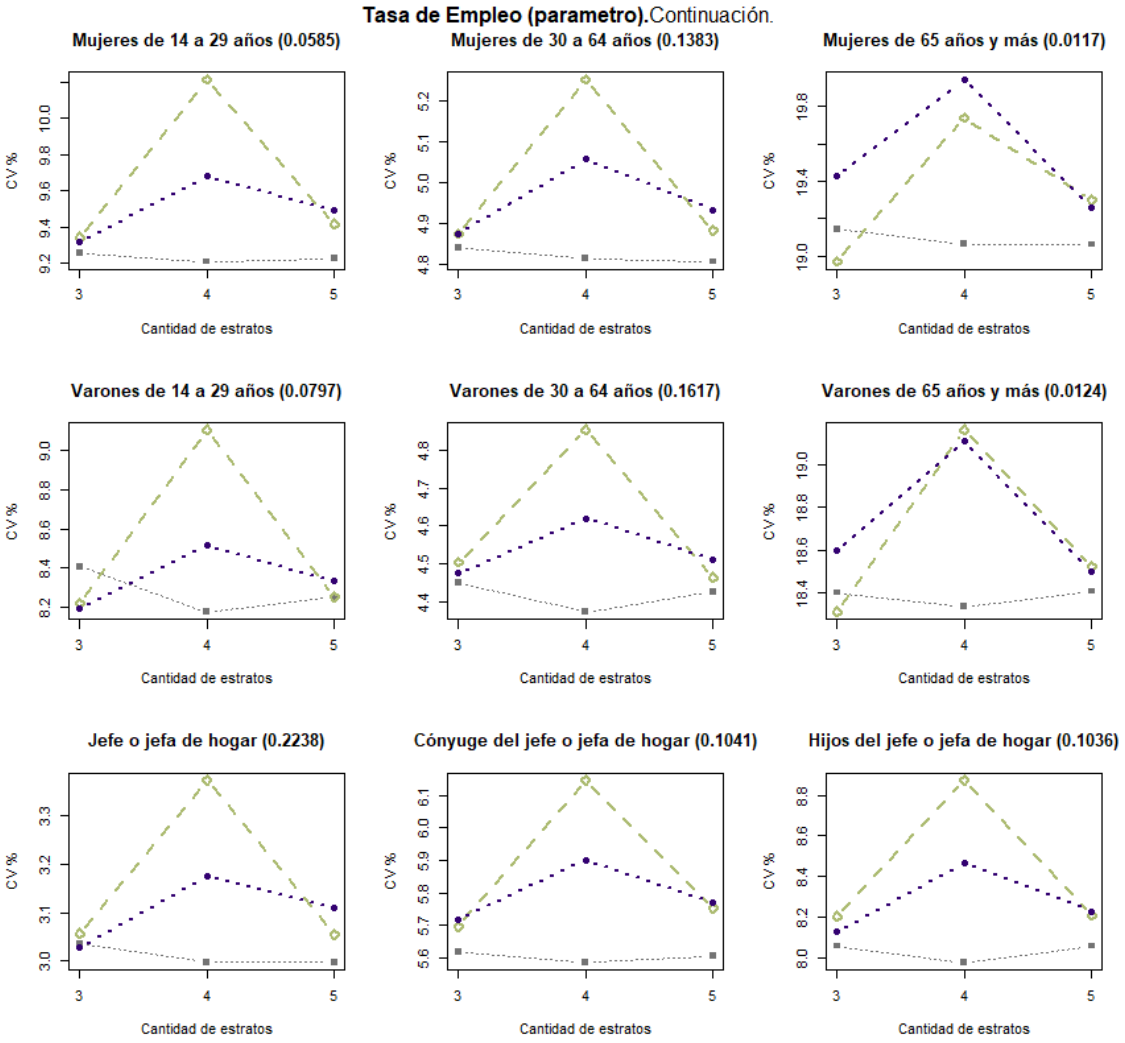


Gráfico A.12: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Santa Fe. Año 2010. (Continuación)



Venado Tuerto CV(Tasas)

Tabla A.7: Principales indicadores del mercado de trabajo, parámetro y CV según método de estratificación por cantidad de estratos en la primera etapa de muestreo. Venado Tuerto. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		n° de estratos			*n° de estratos*			*n° de estratos*		
		3	*4*	*5*	*3*	*4*	*5*	*3*	*4*	*5*
	parámetro %	*CV %*								
Actividad										
Tasa de la población total	*53,9*	3,39	3,37	3,36	3,27	3,26	3,25	3,34	3,32	3,31
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*69,5*	2,58	2,58	2,57	2,58	2,56	2,55	2,57	2,56	2,54
Mujeres	*58,7*	3,93	3,93	3,93	3,93	3,93	3,92	3,93	3,93	3,90
Varones	*81,5*	2,48	2,48	2,47	2,48	2,46	2,46	2,48	2,48	2,46
Jefes o jefas de hogar	*79,1*	2,44	2,43	2,43	2,40	2,38	2,38	2,46	2,46	2,46
Mujeres de 14 a 29 años	*57,2*	6,72	6,67	6,72	6,71	6,69	6,68	6,77	6,81	6,78
Mujeres de 30 a 64 años	*73,0*	3,58	3,57	3,59	3,57	3,57	3,56	3,57	3,59	3,58
Varones de 14 a 29 años	*75,3*	4,67	4,66	4,65	4,65	4,63	4,64	4,63	4,63	4,61
Varones de 30 a 64 años	*95,1*	1,68	1,67	1,66	1,68	1,67	1,67	1,66	1,67	1,66
Empleo										
Tasa de la población total	*51,5*	3,39	3,37	3,37	3,31	3,30	3,29	3,37	3,35	3,33
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*66,4*	2,68	2,68	2,67	2,68	2,67	2,66	2,68	2,67	2,65
Mujeres	*54,9*	4,20	4,20	4,21	4,20	4,19	4,18	4,19	4,20	4,18
Varones	*79,1*	2,61	2,61	2,60	2,61	2,59	2,60	2,61	2,61	2,60
Jefes o jefas de hogar	*77,5*	2,53	2,53	2,53	2,51	2,49	2,49	2,56	2,56	2,56
Mujeres de 14 a 29 años	*49,9*	7,78	7,73	7,77	7,71	7,69	7,68	7,80	7,87	7,84
Mujeres de 30 a 64 años	*70,0*	3,85	3,84	3,85	3,83	3,83	3,82	3,83	3,86	3,84
Varones de 14 a 29 años	*71,4*	5,16	5,15	5,14	5,15	5,13	5,13	5,13	5,13	5,11
Varones de 30 a 64 años	*93,5*	1,84	1,84	1,83	1,84	1,84	1,84	1,83	1,83	1,82
Desocupación										
Tasa de la población total	*4,5*	17,40	17,38	17,31	17,18	17,17	17,17	17,21	17,25	17,25
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*4,5*	17,40	17,38	17,31	17,18	17,17	17,17	17,21	17,25	17,25
Mujeres	*6,5*	21,14	21,11	21,05	20,84	20,79	20,79	20,91	21,01	20,99
Varones	*2,9*	28,02	28,00	27,94	27,94	27,98	27,97	27,91	27,94	27,96
Jefes o jefas de hogar	*2,1*	33,94	33,80	33,79	33,68	33,70	33,73	33,65	33,60	33,64
Mujeres de 14 a 29 años	*12,7*	26,20	26,18	26,16	25,97	25,97	26,00	26,10	26,19	26,13
Mujeres de 30 a 64 años	*4,1*	31,24	31,20	31,13	31,15	31,11	31,13	31,12	31,16	31,14
Varones de 14 a 29 años	*5,3*	36,91	36,84	36,84	36,84	36,92	36,76	36,87	36,96	37,01
Varones de 30 a 64 años	*1,7*	44,86	44,86	44,89	44,88	44,88	44,96	44,69	44,64	44,64

Gráfico A.13: Coeficiente de variación de la tasa específica de actividad según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Venado Tuerto. Año 2010

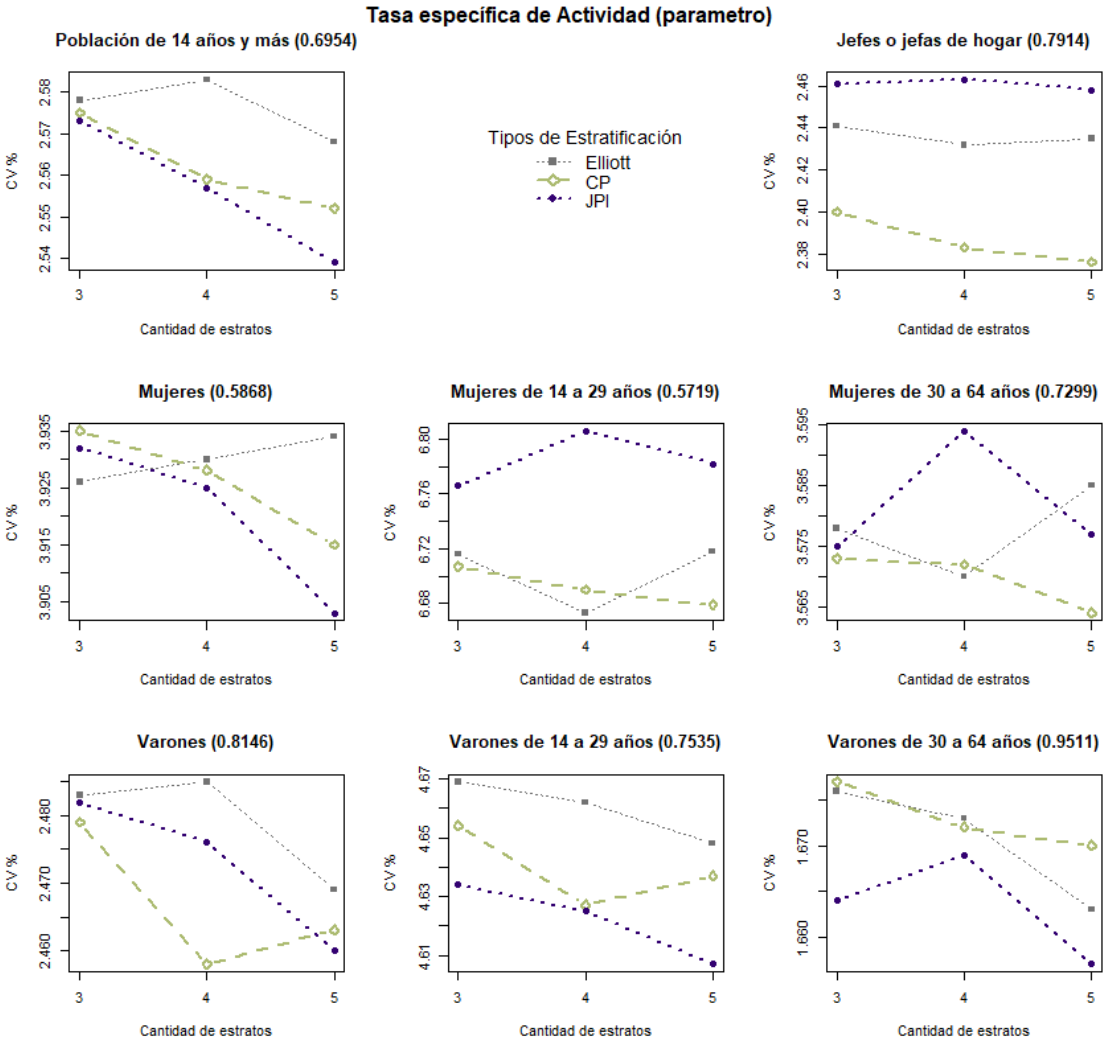


Gráfico A.14: Coeficiente de variación de la tasa específica de empleo según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Venado Tuerto. Año 2010

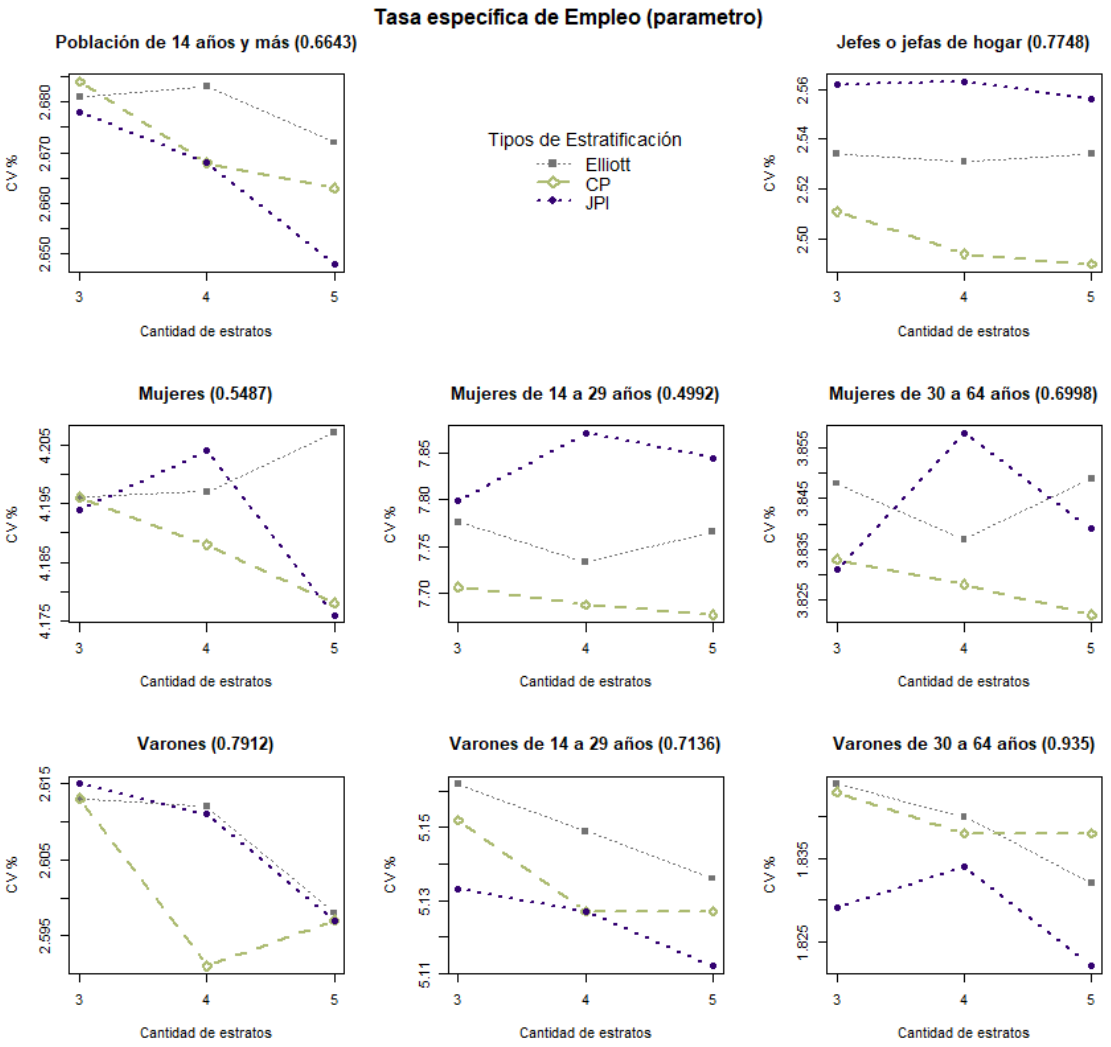


Gráfico A.15: Coeficiente de variación de la tasa específica de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Venado Tuerto. Año 2010

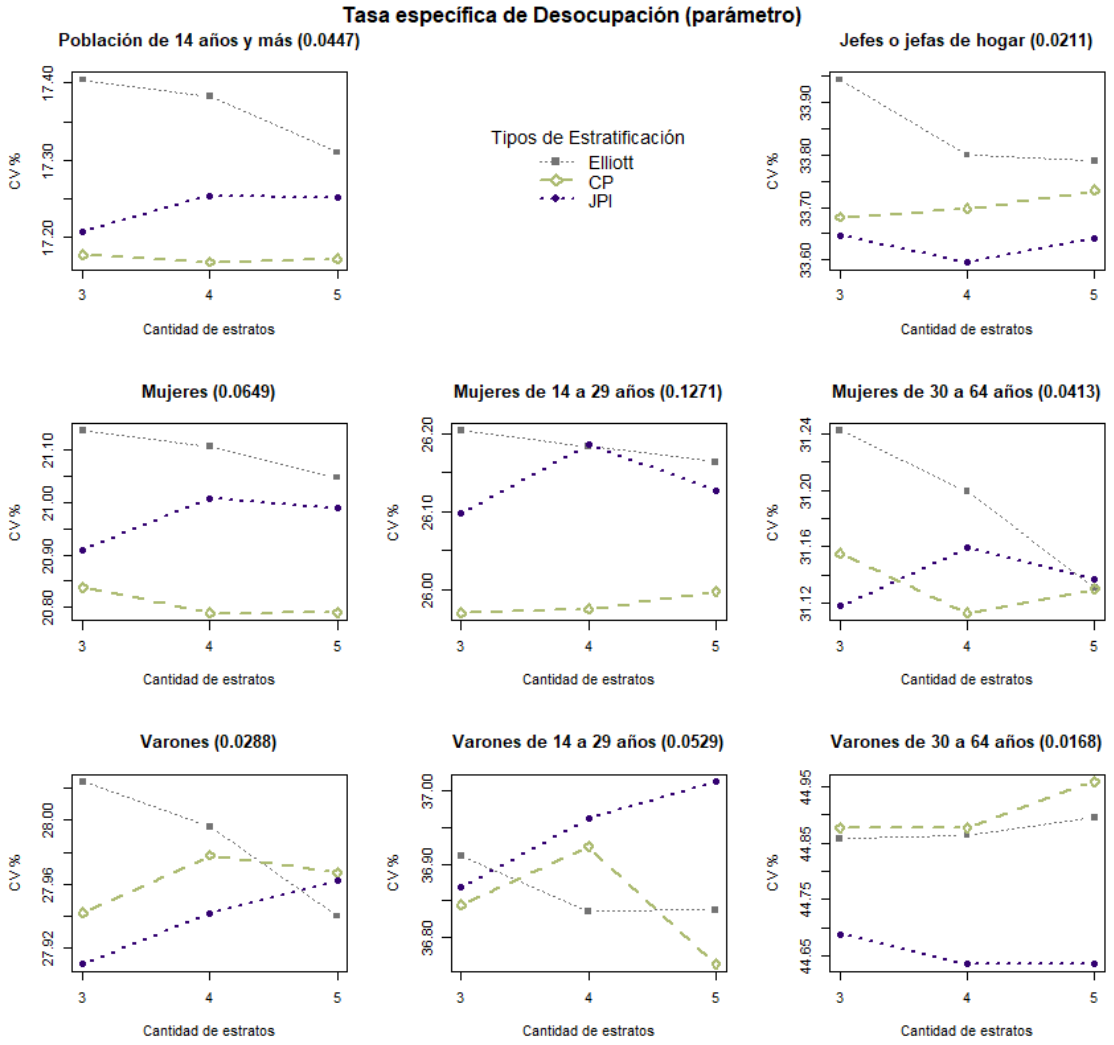


Tabla A.8: Composición de la tasa de empleo, parámetro y CV según sexo y edad, relación de parentesco con el jefe o jefa de hogar y nivel de instrucción, por método de estratificación y cantidad de estratos en la primera etapa de muestreo. Venado Tuerto. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		nº de estratos			nº de estratos			nº de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Tasa de empleo	51,5	3,39	3,37	3,37	3,31	3,30	3,29	3,37	3,35	3,33
Sexo y grupos de edad										
Mujeres hasta 29 años	6,1	10,37	10,37	10,38	10,34	10,32	10,31	10,35	10,34	10,34
Mujeres de 30 a 64 años	14,7	5,38	5,39	5,39	5,38	5,38	5,38	5,39	5,39	5,37
Mujeres de 65 años y más	1,5	21,35	21,35	21,56	21,28	21,25	21,13	21,69	21,76	21,72
Varones hasta 29 años	8,9	9,57	9,47	9,48	9,22	9,18	9,13	9,47	9,48	9,41
Varones de 30 a 64 años	18,4	4,78	4,75	4,75	4,70	4,68	4,67	4,78	4,79	4,76
Varones de 65 años y más	2,0	17,91	17,92	17,97	17,87	17,87	17,81	18,10	18,11	18,02
Relación de parentesco										
Jefe o jefa de hogar	25,6	2,91	2,89	2,91	2,87	2,86	2,85	2,93	2,93	2,92
Cónyuge	12,6	5,88	5,89	5,87	5,88	5,86	5,85	5,91	5,90	5,87
Hijos	10,2	9,51	9,46	9,49	9,34	9,33	9,33	9,33	9,26	9,28
Otros componentes	3,1	17,86	17,76	17,77	17,79	17,79	17,78	17,76	17,72	17,66
Nivel educativo										
Primario incompleto	4,2	15,16	14,56	14,63	13,87	13,83	13,80	13,99	14,19	13,96
Primario completo	13,3	8,83	8,66	8,70	8,14	8,06	8,00	8,36	8,37	8,30
Secundario incompleto	10,5	9,27	9,24	9,22	8,97	8,97	8,93	9,06	8,99	9,05
Secundario completo	10,8	8,20	8,18	8,19	8,10	8,06	8,06	8,19	8,24	8,16
Superior y universitario incompleto	4,5	12,96	12,86	13,00	12,59	12,56	12,54	12,78	12,81	12,82
Superior y universitario completo	7,6	11,91	11,74	11,63	10,99	10,96	10,82	10,84	10,78	10,75
Sin instrucción	0,5	39,87	39,52	39,67	39,22	39,39	39,41	39,34	39,16	39,11

Gráfico A.16: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Venado Tuerto. Año 2010. (Continúa)

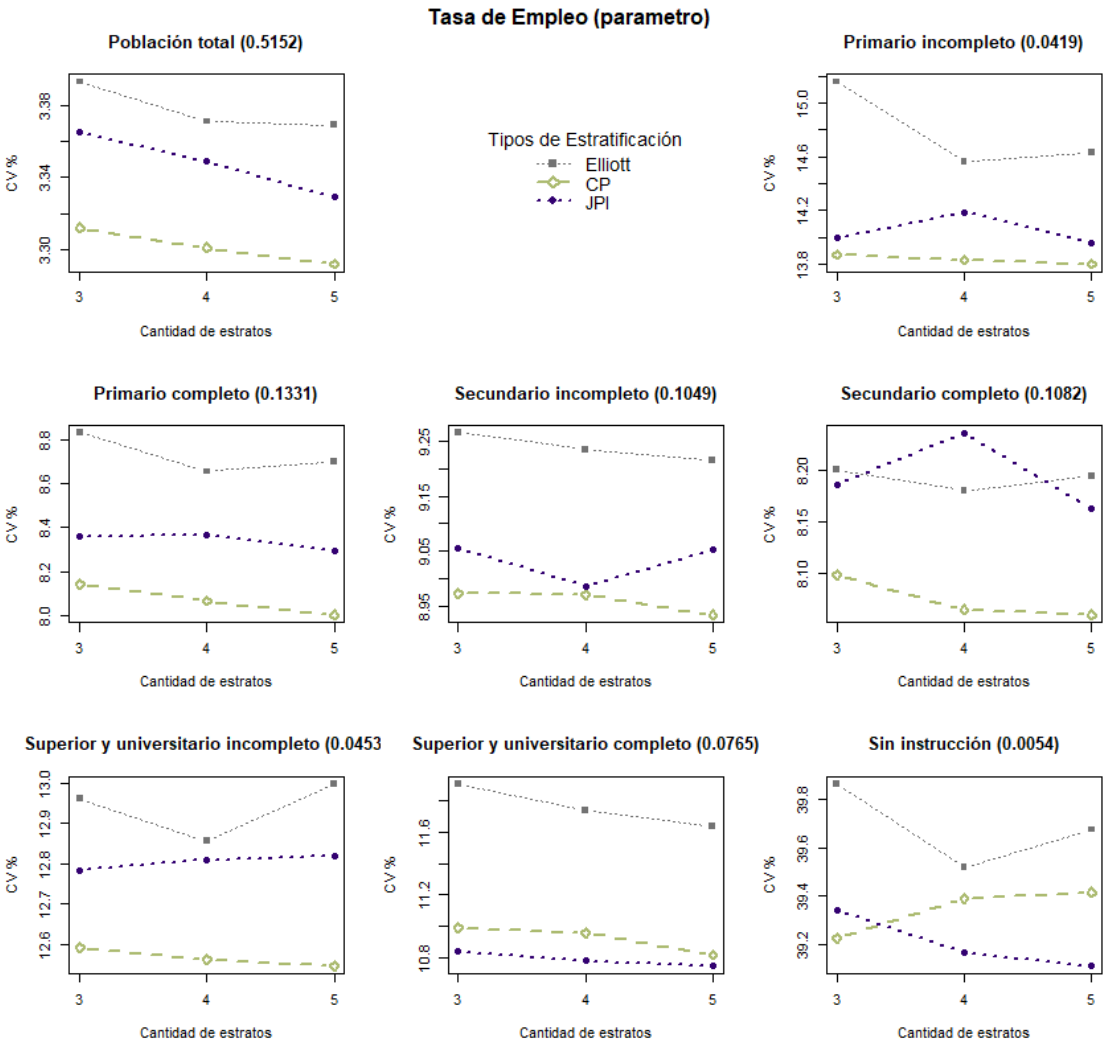
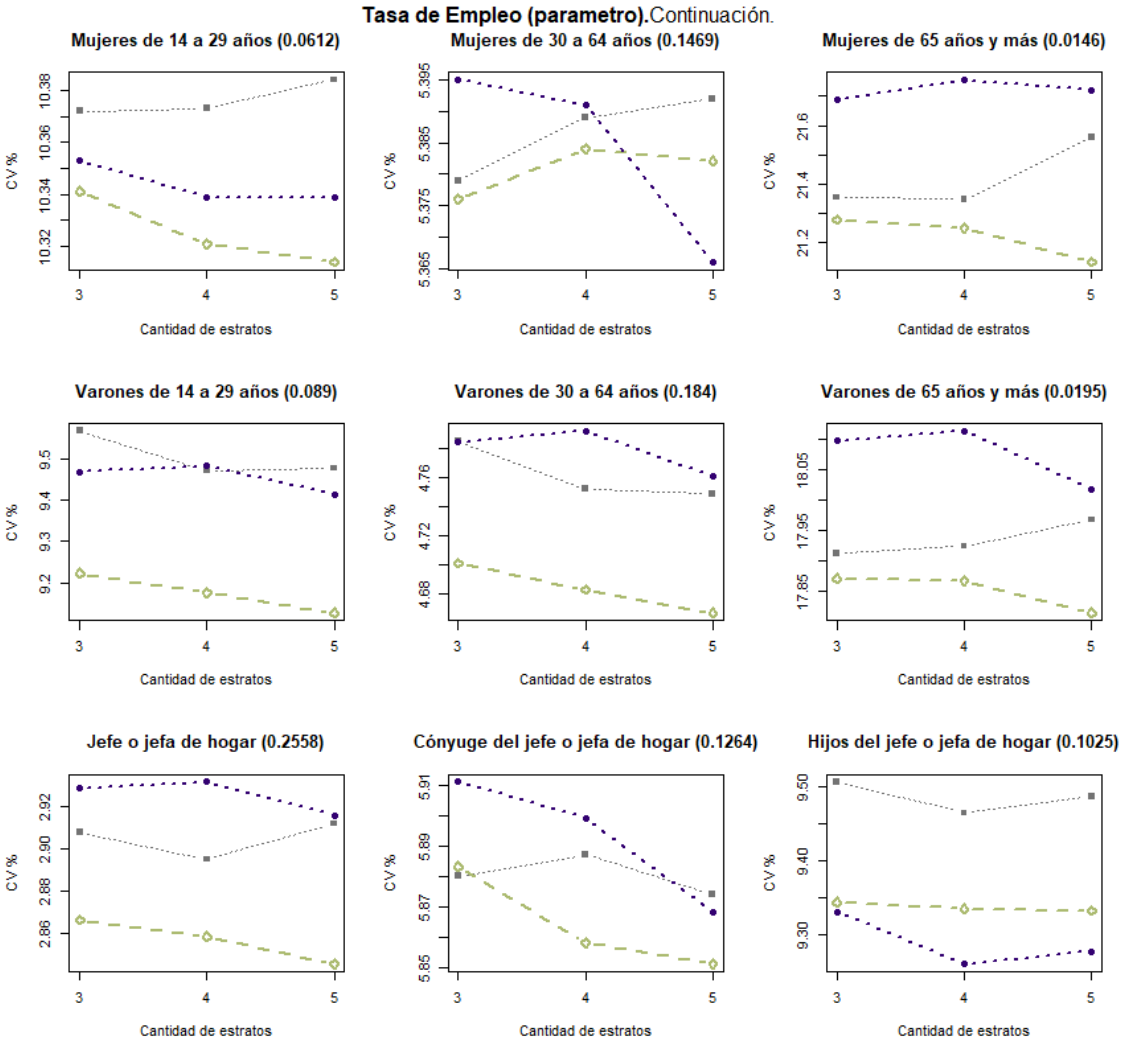


Gráfico A.16: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Venado Tuerto. Año 2010. (Continuación)



Resto Provincial CV(Tasas)

Tabla A.9: Principales indicadores del mercado de trabajo, parámetro y CV según método de estratificación por cantidad de estratos en la primera etapa de muestreo. Resto provincial. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		n° de estratos			*n° de estratos*			*n° de estratos*		
		3	*4*	*5*	*3*	*4*	*5*	*3*	*4*	*5*
	parámetro %	*CV %*								
Actividad										
Tasa de la población total	*48,0*	3,85	3,80	3,76	3,82	3,81	3,81	3,85	3,85	3,85
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*62,3*	3,81	3,87	3,79	3,83	3,82	3,84	3,81	3,81	3,83
Mujeres	*49,4*	4,42	4,41	4,39	4,41	4,41	4,40	4,40	4,39	4,38
Varones	*75,8*	6,04	6,22	6,04	6,07	6,05	6,09	6,01	6,03	6,08
Jefes o jefas de hogar	*72,8*	2,78	2,77	2,76	2,78	2,78	2,77	2,78	2,78	2,78
Mujeres de 14 a 29 años	*47,0*	7,58	7,55	7,54	7,53	7,52	7,52	7,57	7,54	7,54
Mujeres de 30 a 64 años	*62,7*	4,34	4,35	4,34	4,30	4,30	4,30	4,32	4,30	4,29
Varones de 14 a 29 años	*68,9*	10,57	10,87	10,56	10,62	10,59	10,67	10,52	10,55	10,63
Varones de 30 a 64 años	*91,2*	4,58	4,71	4,58	4,59	4,58	4,61	4,55	4,57	4,61
Empleo										
Tasa de la población total	*45,5*	3,85	3,81	3,78	3,83	3,82	3,82	3,85	3,84	3,84
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*59,0*	3,92	3,99	3,91	3,94	3,93	3,95	3,92	3,92	3,94
Mujeres	*45,6*	4,74	4,74	4,73	4,71	4,71	4,71	4,72	4,70	4,69
Varones	*73,1*	6,11	6,29	6,11	6,15	6,13	6,17	6,09	6,11	6,15
Jefes o jefas de hogar	*70,8*	2,91	2,91	2,90	2,92	2,91	2,91	2,92	2,91	2,91
Mujeres de 14 a 29 años	*40,4*	8,63	8,63	8,63	8,57	8,58	8,57	8,63	8,61	8,61
Mujeres de 30 a 64 años	*59,4*	4,67	4,68	4,67	4,62	4,61	4,61	4,64	4,62	4,61
Varones de 14 a 29 años	*64,6*	10,81	11,11	10,81	10,86	10,83	10,91	10,76	10,79	10,87
Varones de 30 a 64 años	*89,1*	4,67	4,80	4,68	4,69	4,67	4,70	4,65	4,67	4,70
Desocupación										
Tasa de la población total	*5,2*	17,02	17,14	17,15	17,07	17,09	17,08	17,17	17,16	17,15
Tasas específicas para la población de 14 años y más										
Tasa de la población de 14 años y más	*5,2*	17,02	17,14	17,15	17,07	17,09	17,08	17,17	17,16	17,15
Mujeres	*7,7*	20,10	20,24	20,27	20,06	20,08	20,08	20,19	20,18	20,17
Varones	*3,5*	25,06	25,29	25,28	25,19	25,21	25,21	25,27	25,24	25,24
Jefes o jefas de hogar	*2,7*	30,04	30,24	30,21	29,90	29,94	29,92	30,13	30,08	30,06
Mujeres de 14 a 29 años	*13,9*	25,06	25,24	25,27	25,07	25,09	25,09	25,13	25,13	25,13
Mujeres de 30 a 64 años	*5,3*	28,53	28,75	28,77	28,51	28,52	28,52	28,60	28,60	28,59
Varones de 14 a 29 años	*6,3*	32,08	32,38	32,39	32,21	32,23	32,24	32,22	32,21	32,22
Varones de 30 a 64 años	*2,3*	37,06	37,32	37,28	37,11	37,14	37,12	37,26	37,22	37,22

Gráfico A.17: Coeficiente de variación de la tasa específica de actividad según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Resto provincial. Año 2010

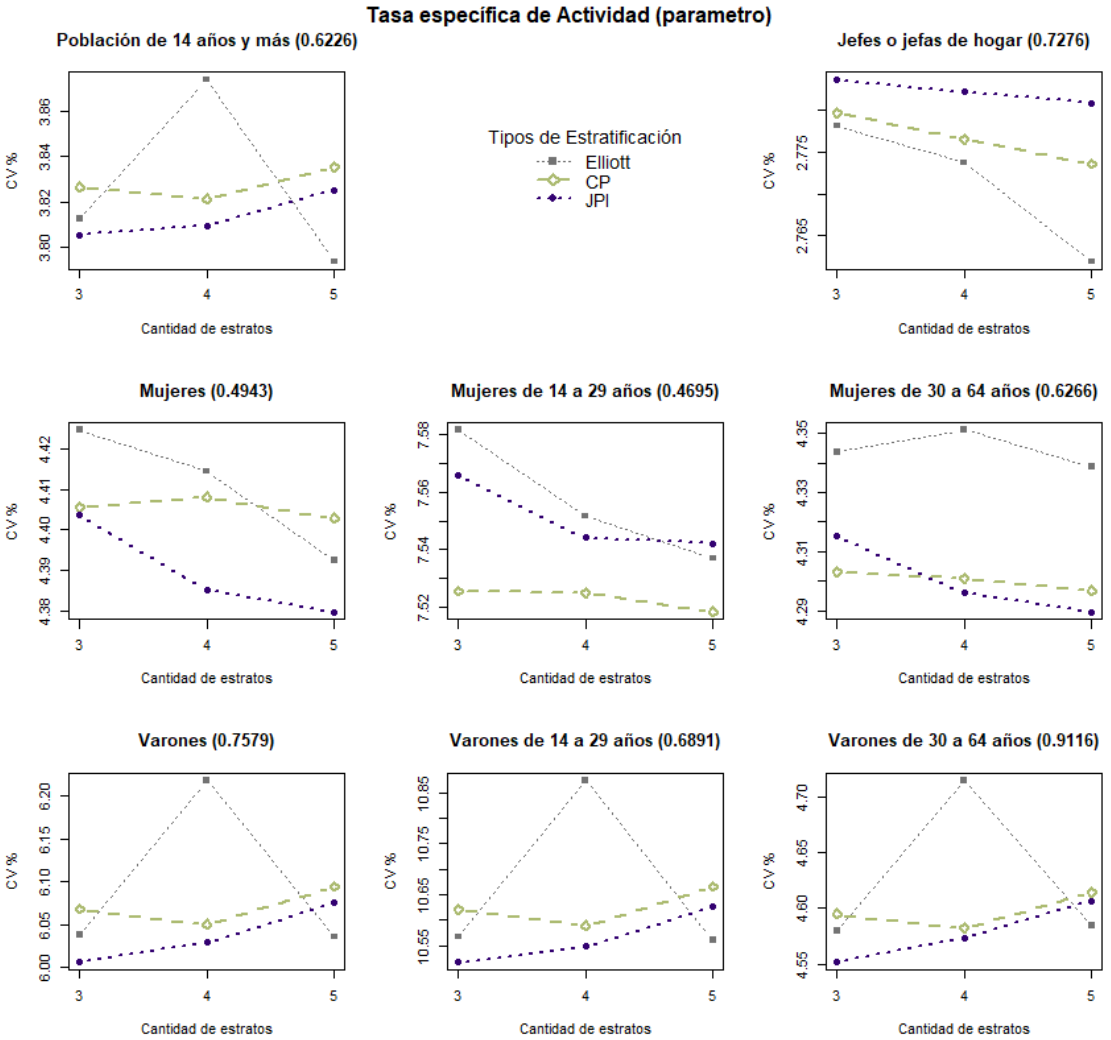


Gráfico A.18: Coeficiente de variación de la tasa específica de empleo según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Resto provincial. Año 2010

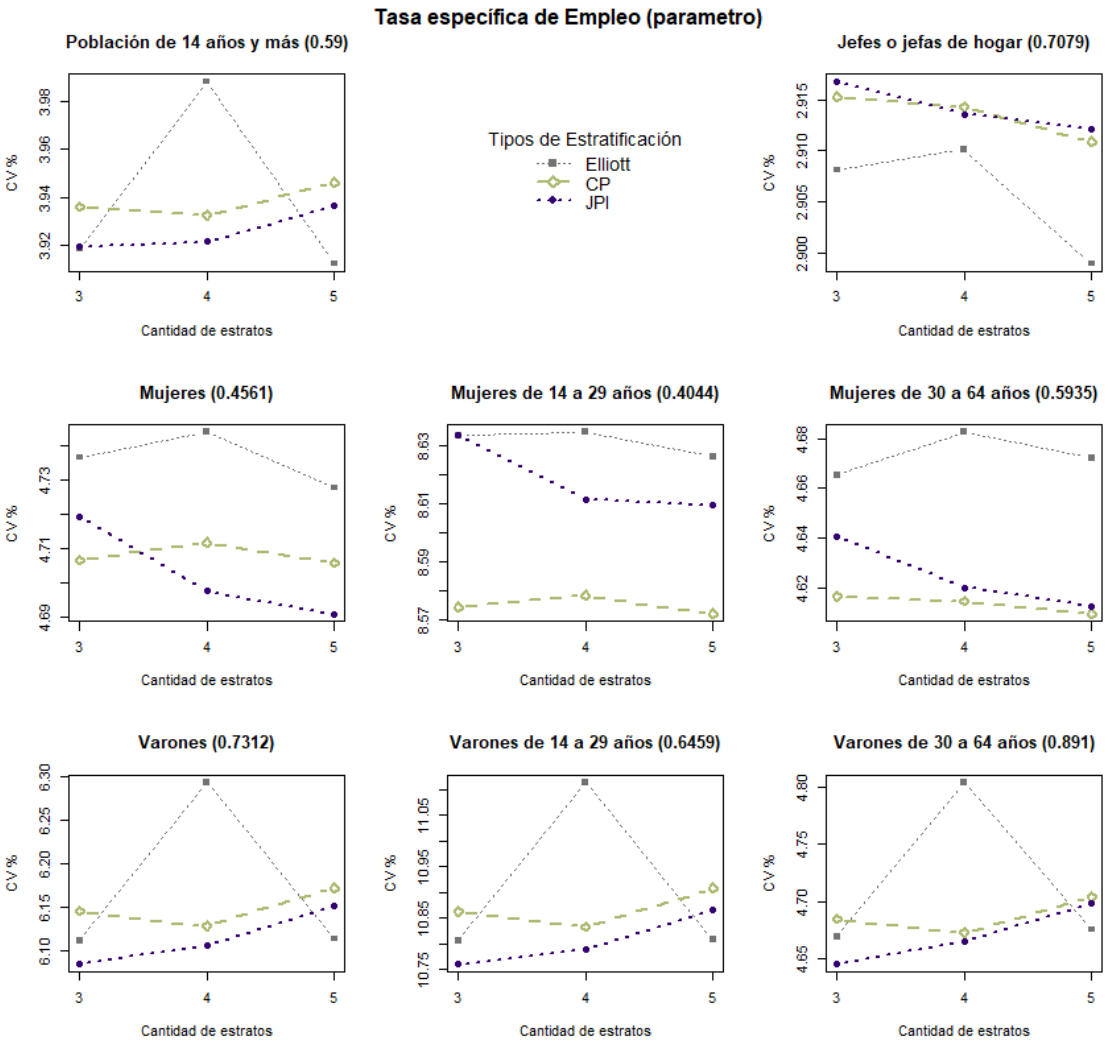


Gráfico A.19: Coeficiente de variación de la tasa específica de desocupación según: población de 14 años y más, jefes o jefas de hogar, sexo; y edad y sexo por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Resto provincial Año 2010

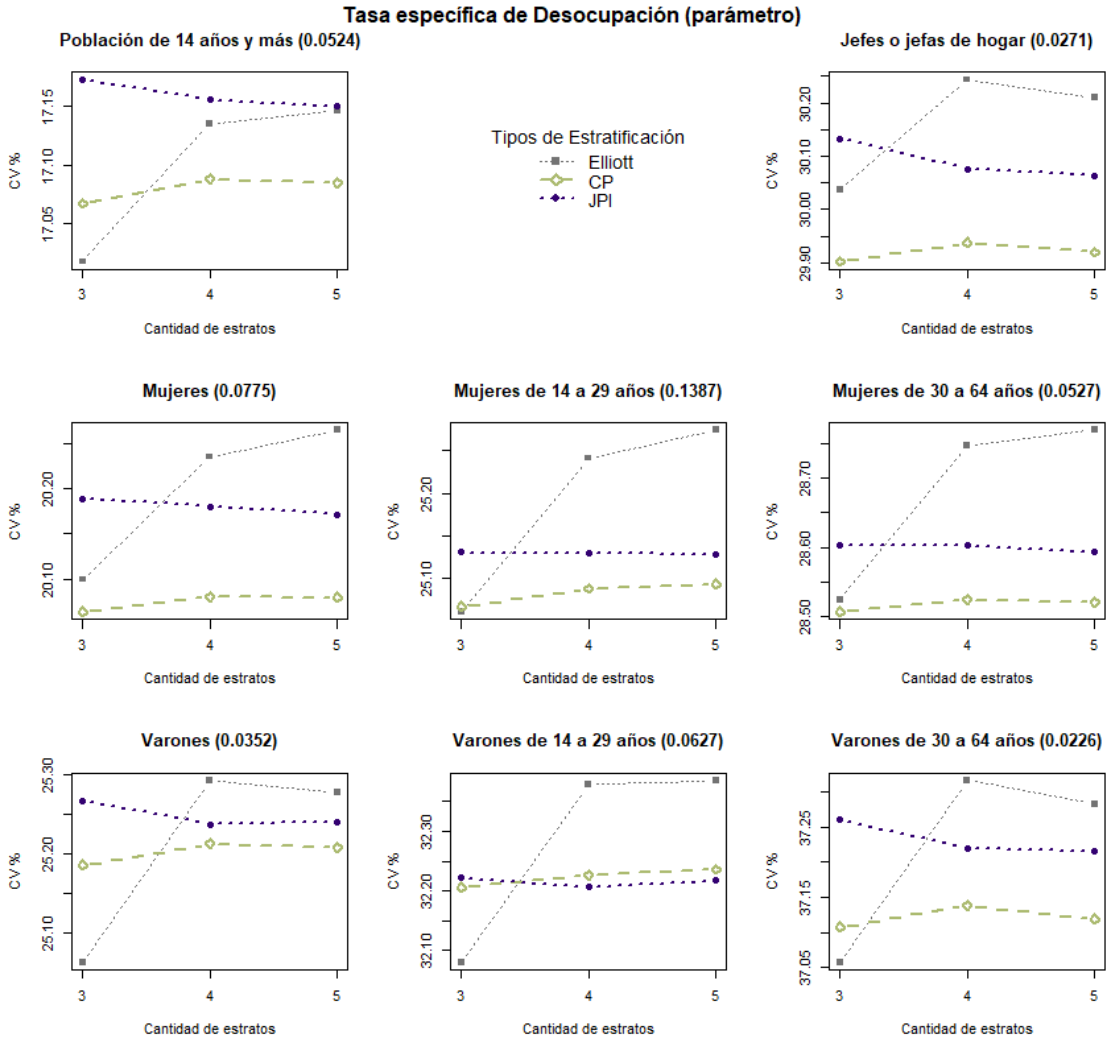


Tabla A.10: Composición de la tasa de empleo, parámetro y CV según sexo y edad, relación de parentesco con el jefe o jefa de hogar y nivel de instrucción, por método de estratificación y cantidad de estratos en la primera etapa de muestreo. Resto provincial. Año 2010

Tasas		Métodos de estratificación								
		Elliott			CP			JPI		
		nº de estratos			nº de estratos			nº de estratos		
		3	4	5	3	4	5	3	4	5
	parámetro %	CV %								
Tasa de empleo	45,5	3,85	3,81	3,78	3,83	3,82	3,82	3,85	3,84	3,84
Sexo y grupos de edad										
Mujeres hasta 29 años	5,2	11,00	11,00	10,99	10,97	10,95	10,96	11,00	11,00	11,00
Mujeres de 30 a 64 años	11,8	6,04	6,03	6,00	5,99	5,98	5,98	5,98	5,96	5,95
Mujeres de 65 años y más	1,0	23,78	23,82	23,84	23,73	23,72	23,71	23,77	23,78	23,78
Varones hasta 29 años	8,5	9,37	9,36	9,36	9,15	9,09	9,09	9,26	9,23	9,22
Varones de 30 a 64 años	17,4	4,64	4,62	4,60	4,62	4,61	4,61	4,63	4,63	4,63
Varones de 65 años y más	1,5	19,00	19,03	19,05	19,05	19,04	19,05	19,09	19,10	19,09
Relación de parentesco										
Jefe o jefa de hogar	22,1	3,48	3,46	3,45	3,46	3,45	3,45	3,48	3,47	3,47
Cónyuge	10,3	6,47	6,45	6,42	6,46	6,46	6,46	6,45	6,44	6,44
Hijos	10,4	8,93	8,91	8,89	8,92	8,91	8,91	8,92	8,92	8,92
Otros componentes	2,7	17,30	17,37	17,35	17,19	17,17	17,17	17,26	17,25	17,25
Nivel educativo										
Primario incompleto	4,7	13,81	13,83	13,80	12,78	12,58	12,54	12,78	12,62	12,58
Primario completo	13,0	7,92	7,95	7,95	7,72	7,67	7,63	7,60	7,56	7,53
Secundario incompleto	8,1	9,77	9,75	9,73	9,71	9,68	9,68	9,77	9,77	9,77
Secundario completo	10,1	8,11	8,09	8,08	7,94	7,92	7,92	7,89	7,86	7,85
Superior y universitario incompleto	3,2	15,67	15,60	15,47	15,42	15,38	15,31	15,00	14,94	14,82
Superior y universitario completo	5,8	11,95	11,96	12,00	11,53	11,44	11,37	11,34	11,27	11,15
Sin instrucción	0,6	33,42	33,40	33,38	32,49	32,41	32,38	33,03	32,83	32,83

Gráfico A.20: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Resto provincial. Año 2010. (Continúa)

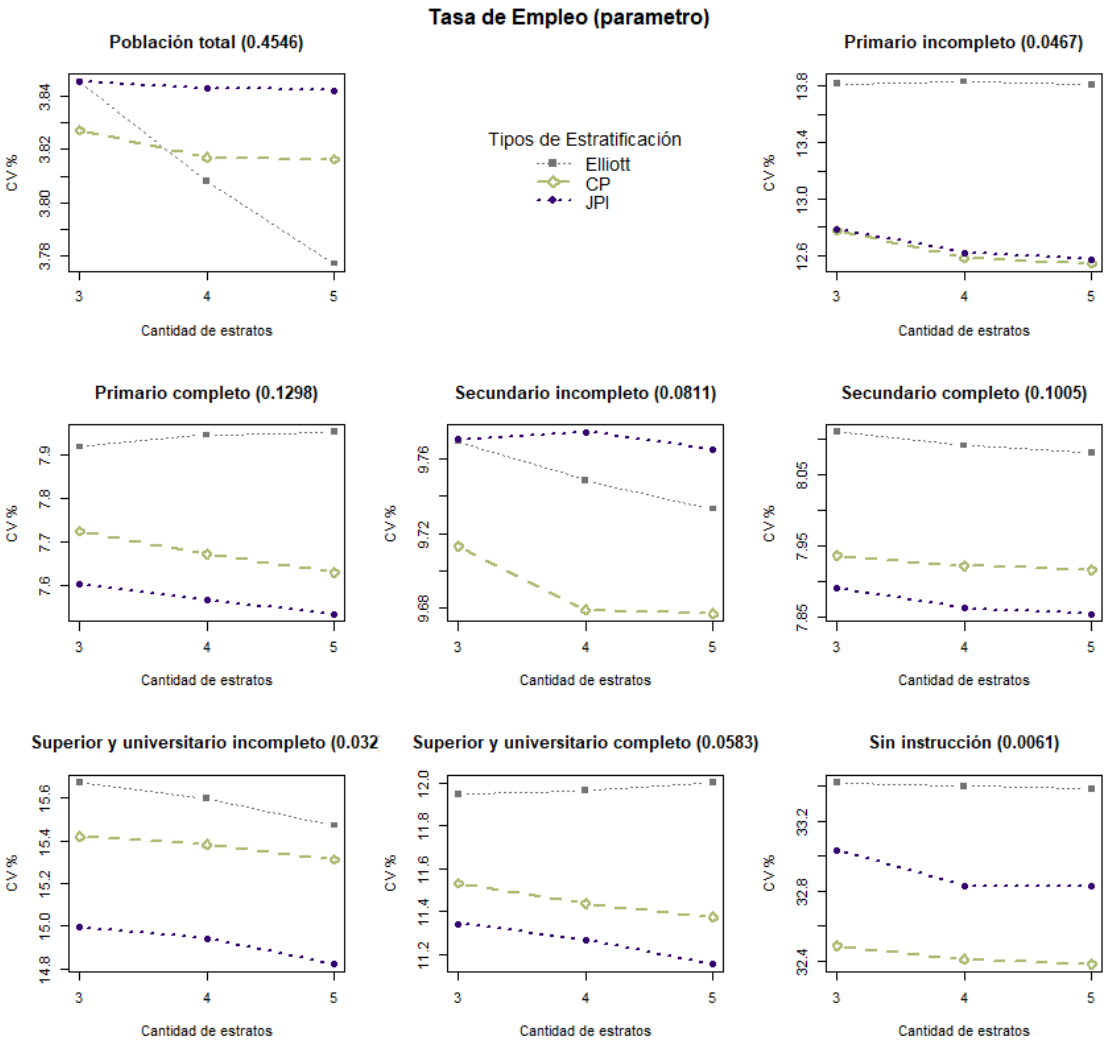
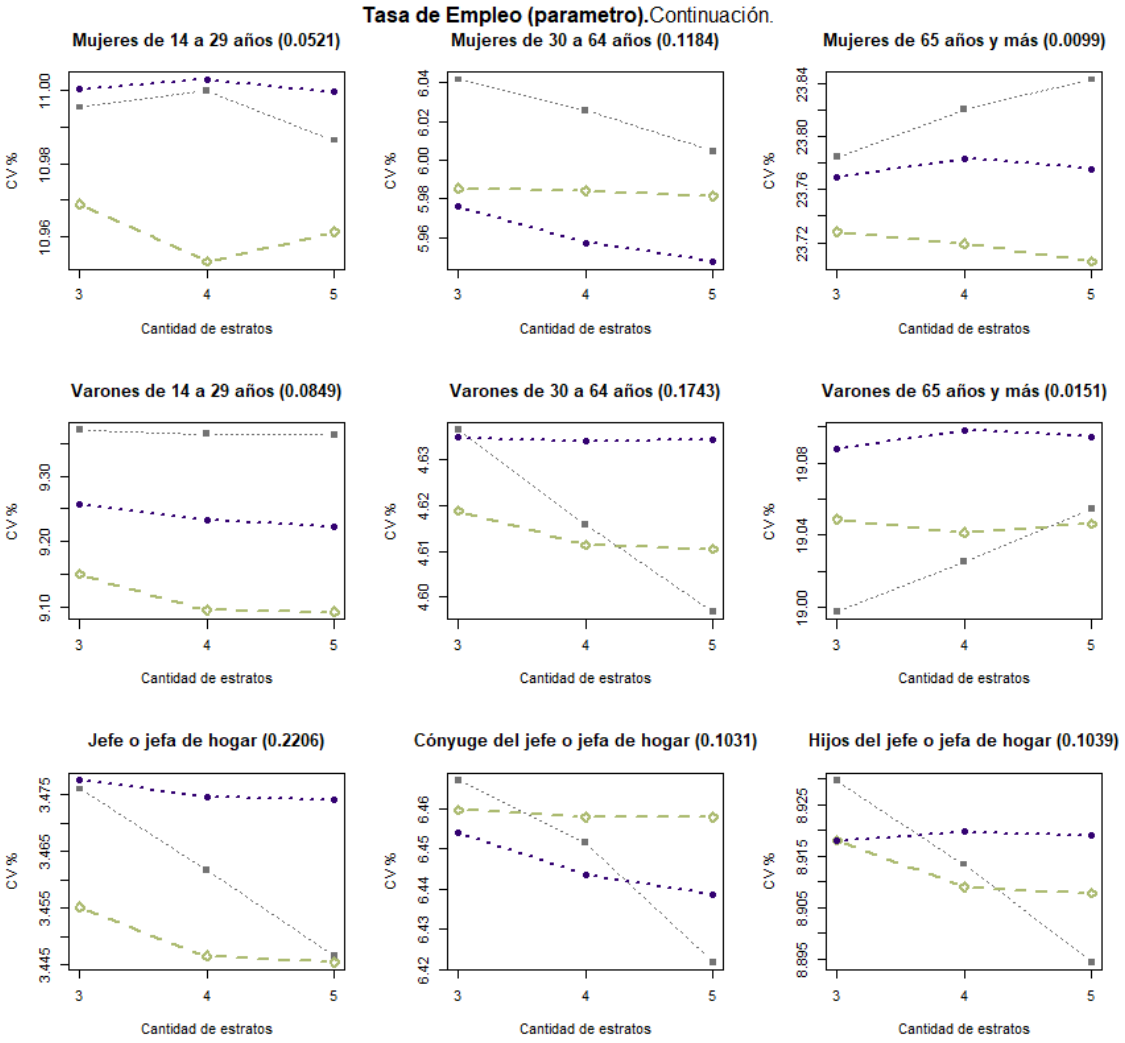


Gráfico A.20: Coeficiente de variación de la tasa de empleo según: población total, nivel de instrucción, sexo y edad; y relación de parentesco con el jefe o jefa del hogar por cantidad de estratos y tipo de estratificación en la primera etapa de muestreo. Resto provincial. Año 2010. (Continuación)



A.3. Procedimientos computacionales empleados

Para favorecer la reproducibilidad computacional del trabajo se presenta, siguiendo el orden de la tesis, las sentencias generadas y los procedimientos empleados bajo el Software R[16], incluyendo el algoritmo de Elliott. Cabe destacar que sería muy repetitivo y excesivo compartir todo el código empleado, por tanto, se presenta a continuación una única sentencia por método empleado, sin mencionar las más simples como, por ejemplo: leer, describir, tabular, filtrar, sumar variables, graficar, guardar o combinar las bases

[16] R *Core Team* (2023). _R: *A Language and Environment for Statistical Computing*_ . *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>.

de datos. A modo genérico se toma como "datos" a la base de datos mencionada en cada Sección, y en cuanto a las variables utilizadas se continúa con la simbología o bien se acortan los nombres, considerando que se pueda interpretar fácilmente.

Unión consecutiva de 3 radios: UC_R3 (menores de 500 viviendas habitables: datosr500) de Sección 5.1.1

```
vivxr500=aggregate(list("viviendas"=viv.datosr500$UR),by=list("radio"=viv.datosr500$U
R),length) #UR:Código de radio
t<- as.integer((nrow(vivxr500))/3)-1       #No quedan redondos radios de a 3
headUC_R3=c(rep(paste("T",1:t),each=3))    #Los primeros t radios se unen de a 3 con T
tailUC_R3=c(rep(paste("A",1:2),each=2))    #Los últimos 4 radios  se unen de a 2 con A
nomUC_R3=c(headUC_R3,tailUC_R3)     # Se genera código único para UC_R3
vivxr500$pop<- as.matrix (table(datosr500$UR))[,1]#Se agrega la población
datosr500$UC_R3<-rep(nomUC_R3,vivxr500$pop)       #Se pega UC_R3 en la población
vivxr500$UC_R3<-nomUC_R3     #Se pega UC_R3 en las viviendas
datos<-rbind(datosR500, datosr500) #Se juntan ambas bases de población r500 y R500
viv.datosR500$UC_R3<-viv.datosR500$UR # La UC_R3 con más de 500 viv.=código de radio
viv.datosr500$UC_R3<-rep(nomUC_R3,vivxr500$viviendas)#Se pega UC_R3 en las viviendas
viv.datos<-rbind(viv.datosR500,viv.datosr500) #Se juntan ambas bases de viviendas
```

Análisis de Correspondencias Múltiple de Sección 5.1.1

```
library(FactoMineR)
mca1 <- MCA(datos, graph = FALSE)) #sólo las columnas
```

Algoritmo de Elliott: Elliott_R3 (se toman las variables 15,17 y19 para unir los radios) de Sección 5.1.1

```
tract<-datos.radios[,1]; lat<-datos.radios[,2]; long<-datos.radios[,3];pop<-
datos.radios[,4]
factor<-as.matrix(datos.radios[,c(15,17,19)])
N<-sum(pop); K<-111; n<-length(tract); nm1<-n-1
STOP<-trunc(N/K-N/(2*n)); nK<-trunc(n/K)

ddist2<-matrix(rep(0,n*n),n,n) ###Create distance matrix###
for(j in 2:n){ddist2[1,(j-1)]<-sqrt(((lat[1]-lat[j])^2)+((long[1]-long[j])^2))}
for(i in 2:nm1){
  for(j in 1:(i-1)){ddist2[i,j]<-sqrt(((lat[i]-lat[j])^2)+((long[i]-long[j])^2))}
  for(j in (i+1):n){ddist2[i,j]<-sqrt(((lat[i]-lat[j])^2)+((long[i]-long[j])^2))}}
for(j in 1:n){ddist2[n,j]<-sqrt(((lat[n]-lat[j])^2)+((long[n]-long[j])^2))}
holdvar<-rep(0,n); holddist<-rep(0,n); holdcv<-rep(0,n)

for(try in 1:n){
  ctr<-try
#ctr<-hold ###Una vez que elige el radio se vuelve a correr a partir de acá*
  pick<-rep(0,n); temp<-order(ddist2[ctr,]); j<-1; tp<-0; add<-NULL
  repeat{
    pick[temp[j]]<-1
    tp<-tp+pop[temp[j]]
    if (tp>STOP) break
    j<-j+1 }
  for(k in 2:(K-1)){
    nk<-n-sum(pick>0)
    kemp<-order(ddist2[ctr,]*(pick==0)) ###Se saca del for step y converge
    teststep<-rep(0,nk)
    for(step in 1:nk){
      tpick<-pick
      tctr<-kemp[sum(pick>0)+step]
      temp<-order(ddist2[tctr,])
      j<-1; tp<-0; add<-NULL
      repeat{
        if(tpick[temp[j]]==0){add<-c(add,temp[j]); tp<-tp+pop[temp[j]]}
        if(tp>STOP) break
        j<-j+1
```

```
        if(j==n+1) break} #Agregado para que no siga
      tpick[add]<-rep(k,length(add))
      teststep[step]<-summary.lm(lm(factor~as.factor(tpick)))$sigma }
  #step<-order(teststep)[1] #Estrato
  step<-order(teststep,decreasing = T)[1] #Para cluster se usa decreasing
    ctr<-order(ddist2[ctr,]*(pick==0))[sum(pick>0)+step]
    temp<-order(ddist2[ctr,])
    j<-1; tp<-0; add<-NULL
    repeat{
      if(pick[temp[j]]==0){ add<-c(add,temp[j]) ; tp<-tp+pop[temp[j]]}
      if(tp>STOP) break
      j<-j+1
      if(j==n+1) break #agregado para que no siga}
    pick[add]<-rep(k,length(add))}
  pick[(pick==0)]<-rep(K,sum(pick==0))

  holdvar[try]<-summary.lm(lm(factor~as.factor(pick)))$sigma
  vivxgrupo<-by(pop,pick,sum)  #Suma cantidad de viviendas por grupo
  holdcv[try]<-sd(vivxgrupo)/mean(vivxgrupo) #Calcula el cv de vivxgrupo
  tsum<-0; denom<-0
  for(k in 1:K){
   tpick<-c(1:n)[(pick==k)]
    for(i in 1:(length(tpick)-1)){ ###Suma distanciasxgrupo, se usa también Sec.5.1.2
       tsum<-tsum+sum((ddist2[tpick[i],c(tpick[(i+1):length(tpick)])])^2, na.rm =
T)#Se agrega el na.rm=T
      denom<-denom+(length(tpick)-1)}}
  holddist[try]<-tsum/denom}

  ###Pick optimal clustering under equal size criterion###
tholdvar<-c(holdvar); tholddist<-c(holddist); holdcv<-c(holdcv)
sholdvar<-(tholdvar-mean(tholdvar))/sqrt(var(tholdvar))
sholddist<-(tholddist-mean(tholddist))/sqrt(var(tholddist))
sholdcv<-(holdcv-mean(holdcv))/sqrt(var(holdcv))
#zholdvar<-sholdvar-min(sholdvar) #minimo error->forma estratos
zholdvar<-max(sholdvar)-sholdvar #maximo error->forma cluster
zholddist<-sholddist-min(sholddist) #mínima distancia promedio entre grupos
zholdcv<-sholdcv-min(sholdcv)
hold<-order((zholdvar^2)+(zholddist^2)+(zholdcv^2))[1] #Se agrega cv(viv x grupo)
ctr<-hold ###Se vuelve a correr *desde la segunda línea del for try hasta su final

datos.radios$Elliott_R3<-pick
```

Cálculo del efecto de diseño: para estimador PEA agrupado por UC_R3 en Sección 5.1.2

```
library(PracTools)
BW<-BW2stageSRS(datos$pea,psuID=datos$UC_R3) #Para MSA

vivxR3=aggregate(list("viviendas"=datos$vivienda),by=list("R3"=datos$UC_R3),sum)
pp.R3<-vivxR3$viviendas/sum(vivxR3$viviendas)
BW<-BW2stagePPT(datos$pea, pp=pp.R3, psuID=datos$UC_R3) #Para PPT
deff<-BW[5]*(1+(BW[6]*9)) #k*(1+(delta*(n-1)))
```

Formación de CP por FRB: para indicadores de carencias en Sección 5.2.1.3

```
library(FRB)
pca_boot<-FRBpcaMM.default(datos) #Probar con FRBpcaMM(datos) si FRB se descarga sola
valorar<-as.data.frame(t(t(pca_boot$eigvec[,1])%*%t(as.matrix(datos))))
datos$CP_BRF<-valorar$V1
```

Formación de estratos por L-H: para CP_FRB en 5 estratos en Sección 5.2.1.4

```
library(stratification)
datos$cp_5<-strata.LH(datos$CP_FRB,CV=0.1,Ls=5,algo = "Sethi")$stratumID
```

Cálculo de los CV con asignación proporcional estratos en Sección 5.2.2

```
#*****************Funcion: calcula CV para una tasa=Y/Z, según estrato *****
cv_razon=function(Y,Z,estrato,n,UPM)
{
  r=sum(Y)/sum(Z)
  tz=sum(Z)
  razon<-(Y-(r*Z))/tz  #Estadística linealizada por Series de Taylor
  base_lin<-data.frame(cbind(razon,estrato,UPM))
  lista.UPM<-unique(base_lin[,2:3])
  N<-length(lista.UPM$UPM)
  Nh<- by(lista.UPM$UPM,INDICES=lista.UPM$estrato,FUN=length)
  a_h<-(Nh/N)*n
  est<-length(table(estrato))
  var.prop=var.p.total<-NULL
  for (i in 1:est) {
   base_estra<-subset(base_lin, base_lin$estrato==i)
   pp.UPM <- table(base_estra$UPM) / nrow(base_estra)
   razon<-as.numeric(base_estra$razon)
   Ni<-table(base_estra$UPM)
   tots<-by(razon, INDICES=base_estra$UPM, FUN=sum)
   vars<-by(razon, INDICES=base_estra$UPM, FUN=var)
   b2<-sum(pp.UPM * (tots/pp.UPM - sum(tots))^2)
   w2<- sum(Ni^2 * vars/pp.UPM)
   m<-a_h[i]
    ifelse(n>40, var.prop[i]<-(b2/m)+(w2/(m*10)),var.prop[i]<-(b2/m)+(w2/(m*15)))
  }#Establecido en 5.2.1.1
  var.p.total<-sum(var.prop)
  cv<-sqrt(var.p.total)*100/r
  return(cv)
}
CV<- cv_razon(datos$desocup,datos$pea,datos$cp_5,n=34,datos$UPM)


#*****************Funcion: calcula CV para el total Y, según estrato *****
cv_total=function(Y,pop,estrato,n,UPM)
{
  base_lin<-data.frame(cbind(Y,estrato,UPM))
  lista.UPM<-unique(base_lin[,2:3])
  N<-length(lista.UPM$UPM)
  Nh<- by(lista.UPM$UPM,INDICES=lista.UPM$estrato,FUN=length)
  a_h<-(Nh/N)*n
  est<-length(table(estrato))
  var.prop=var.p.total<-NULL
  for (i in 1:est) {
   base_estra<-subset(base_lin, base_lin$estrato==i)
   pp.UPM <- table(base_estra$UPM) / nrow(base_estra)
   Ni<-table(base_estra$UPM)
   tots<-by(Y, INDICES=base_estra$UPM, FUN=sum)
   vars<-by(Y, INDICES=base_estra$UPM, FUN=var)
   b2<-sum(pp.UPM * (tots/pp.UPM - sum(tots))^2)
   w2<- sum(Ni^2 * vars/pp.UPM)
   m<-a_h[i]
    ifelse(n>40, var.prop[i]<-(b2/m)+(w2/(m*10)),var.prop[i]<-(b2/m)+(w2/(m*15)))
  }#Establecido en 5.2.1.1
  var.p.total<-sum(var.prop)/(sum(pop)^2)
  t=sum(Y)/sum(pop)
  cv<-sqrt(var.p.total)*100/t
  return(cv)
}

CV <-cv_total(datos$pea,datos$pop,datos$cp_5,n=34,datos$UPM))
```