Estimación de Áreas Pequeñas: aplicación en la Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC)

Subdepartamento de Investigación Estadística

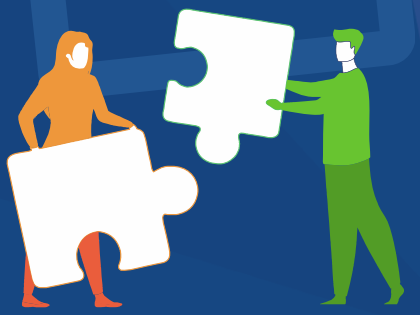
Departamento de Metodologías e Innovación Estadística

Instituto Nacional de Estadísticas

noviembre de 2022

Contenido

1. Motivación
2. Proyecto Estimación de áreas pequeñas
3. Aplicación ENUSC 2018



1.
Motivación

Motivación: la necesidad de las desagregaciones

- La generación de indicadores sociodemográficos y económicos confiables de la población **es esencial para la formulación de políticas públicas** de un país.

- Para ser informativos y eficaces, estos indicadores deben elegirse en el **nivel apropiado de desagregación** buscando equilibrar aspectos de interés del análisis con los recursos disponibles.

- Para ser eficaces, es menester construir un **sistema inferencial preciso y exacto** que trate de reducir la brecha entre las estadísticas oficiales (obtenidas a partir de encuestas por muestreo, planeadas solo para grandes dominios) y la solicitud local de datos o con mayor desagregación.

- Una forma de construir dicho sistema inferencial es a través de **metodologías de estimación de área pequeña** (**Small Area Estimation o SAE**, por sus siglas en inglés).

- En ese contexto, el Instituto Nacional de Estadísticas de Chile (INE) tiene como meta fomentar la aplicación de la metodología SAE en diferentes productos claves de la institución, incrementando la oferta estadística.

Motivación: estimación de áreas pequeñas

SAE en tres pasos:

1. Combinar datos de encuestas con registros y censos para estimar modelos que vinculen una variable y (encuesta) con variables x (encuesta, registros, censos).

2. Combinar los parámetros estimados del modelo con x, para unidades fuera de muestra, para formar predicciones.

3. Estimar los parámetros objetivo de la población con su correspondiente error de predicción.

El objetivo es construir un modelo predictivo que permita obtener estimaciones precisas de los parámetros objetivo del área o dominio de interés.

2.

Proyecto Estimación de áreas pequeñas

Proyecto Estimación de áreas pequeñas: objetivos y equipo de trabajo

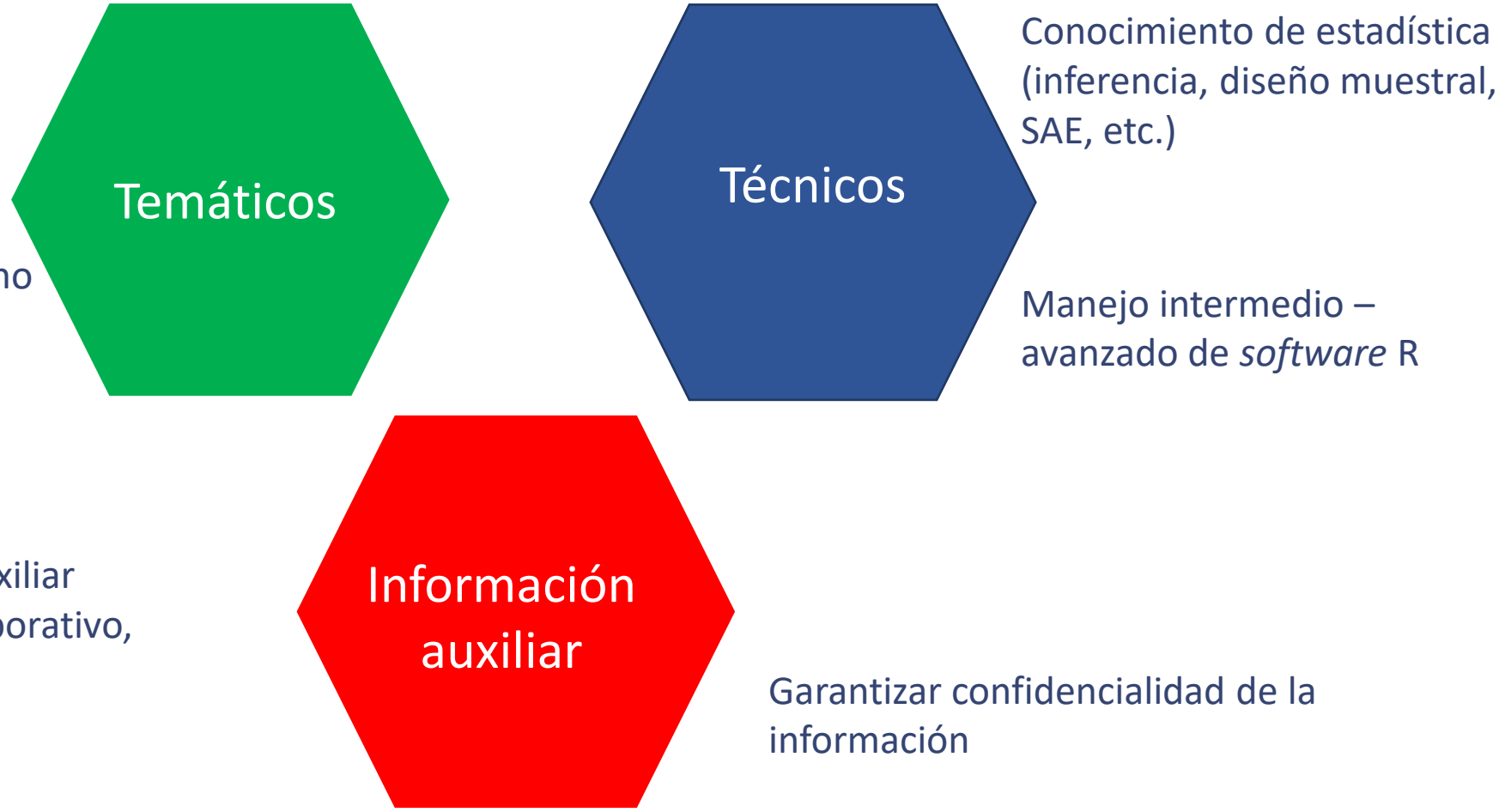
Objetivos de trabajo:

1. Diseñar una guía metodológica sobre el proceso de **estimación de áreas pequeñas** para la generación de estadísticas oficiales del Instituto Nacional de Estadísticas (INE), mediante el vínculo técnico – estadístico entre encuestas y fuentes de información externas.
2. Fomentar la aplicación de la metodología SAE en diferentes productos claves del INE, incrementando la oferta estadística.

Equipo de trabajo:

El equipo de trabajo está conformado por tres profesionales del **Subdepartamento de Investigación Estadística del INE**.

Proyecto Estimación de áreas pequeñas: requerimientos



Proyecto Estimación de áreas pequeñas: etapas

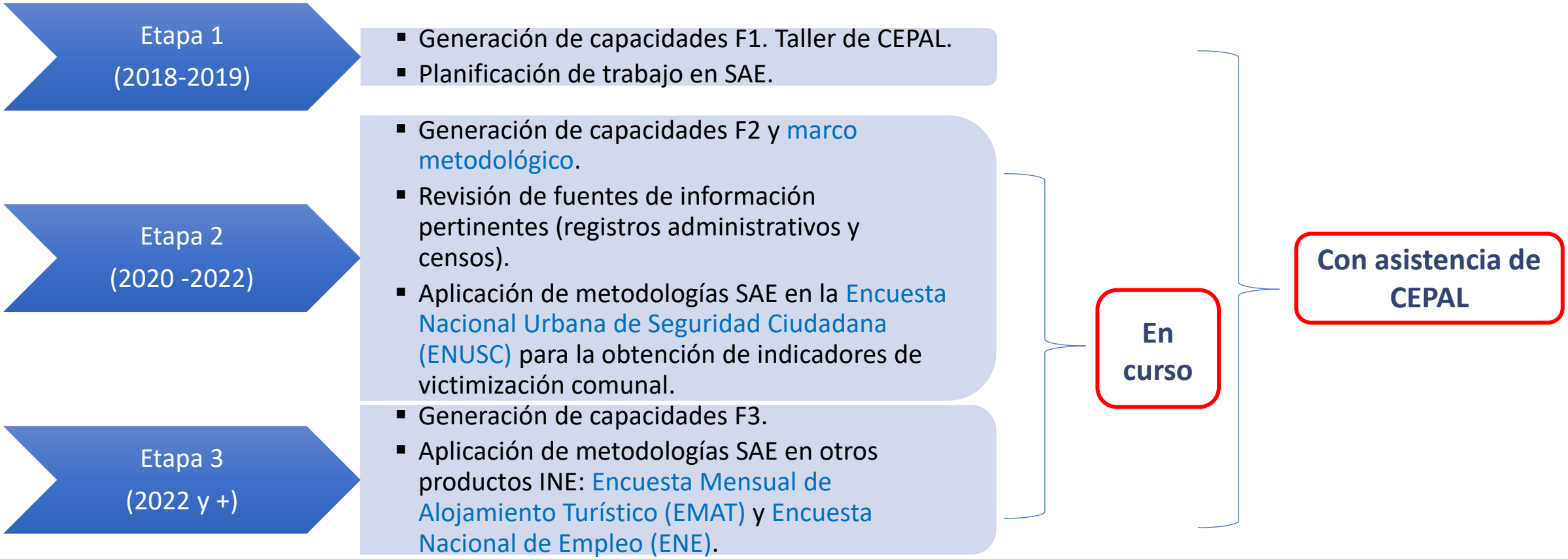


Fig. 1. Etapas del proyecto de estimación de áreas pequeñas. Fuente: Elaboración propia.

3.

Aplicación ENUSC 2018

Aplicación ENUSC 2018: necesidades de información

1

El objetivo de la Encuesta Nacional Urbana de Seguridad Ciudadana (ENUSC) es obtener información sobre la percepción de inseguridad, la reacción frente al delito y la victimización de personas y hogares a partir de una muestra de viviendas particulares ocupadas.

- **Permite producir estadísticas oficiales con un nivel aceptable para los niveles nacional y regional en la zona urbana del país.**

2

Estas estadísticas oficiales son utilizadas por las oficinas gubernamentales para la articulación, implementación y evaluación de programas

- **En materia de seguridad ciudadana y distribución de los recursos de las policías en el territorio nacional.**
- **Por este motivo, surge la necesidad desde el Ministerio del Interior y Seguridad Pública de contar con estadísticas oficiales en niveles de desagregación más bajos que los de la región, digamos el nivel comunal.**

Aplicación ENUSC 2018

- **Parámetro objetivo:** Tasa de victimización comunal.
- **Versión de la encuesta:** ENUSC 2018^{12}.
- **Indicadores:**
  - Victimización Agregada Delitos Consumados (VA_DC)
  - Robo por sorpresa (RPS)
  - Asalto con violencia o amenaza (AVI)
  - Hurtos (HUR)
  - Lesiones (LES)
  - Robo o hurto de vehículo (RHV)
  - Robo o hurto desde vehículo (RDV).

1 Se decide usar esta versión ya que dado el contexto de levantamiento de la versión 2019 (estallido social), fueron necesarios estudios exhaustivos para garantizar la calidad de las estimaciones muestrales.

2 Para más detalles sobre ENUSC 2018, ver documento disponible en https://www.ine.cl/docs/default-source/seguridad-ciudadana/metodolog%C3%ADa/2018/documento-metodol%C3%B3gico-de-dise%C3%B1o-muestral-xv-enusc-2018.pdf?sfvrsn=ea63c94e_4.

Aplicación SAE-ENUSC: resumen

Cuadro 1: Objetivo y visión del proyecto

Objetivo	Estimación de tasa de victimización comunal para los indicadores: victimización agregada, asalto con violencia o intimidación y robo por sorpresa)
Periodo(s)	2018
Cobertura encuesta	102 comunas urbanas de las 16 regiones incluidas en la muestra.
Datos	Microdatos encuesta (ENUSC 2018), microdatos censales (Censo 2017), información agregada a nivel de comunas (Censo 2017, RR.AA. de Carabineros provistos por la Subsecretaría de Prevención del Delito, SPD).
Covariables	Sociodemográficas (Censo 2017), tasas denuncias de delitos (SPD).
Metodología	Estimación de áreas pequeñas (SAE): Modelos de áreas (basados en Fay – Herriot, FH). Otras opciones estudiadas: ○ Modelos de unidad ○ Modelos multinivel
Resultado	Estadísticas comunales como caso de estudio.

Fuente: Elaboración propia.

Aplicación ENUSC 2018: metodología

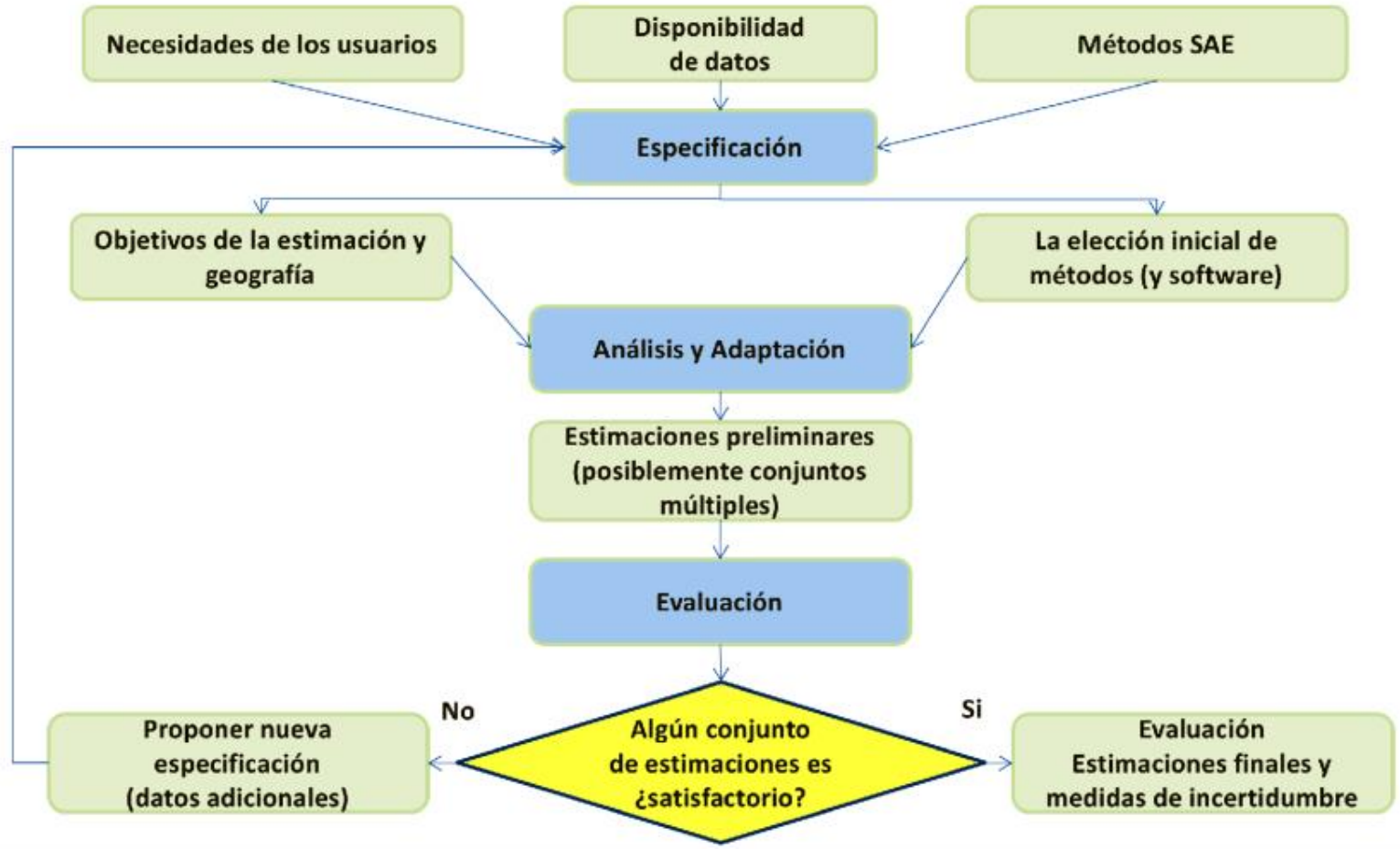


Fig. 3: Proceso de estimación. Fuente: Adaptación de Tzavidis et al. (2018).

Aplicación ENUSC 2018: (I) Especificación, información auxiliar

Cuadro 2: Covariables relacionadas con la victimización

Grupo	Covariables	Fuente
Criminológicas	— Delitos de connotación social (violaciones sexuales y los robos con intimidación) — Delitos a la propiedad, como los robos, hurtos, robos de vehículos motorizados, robos con violencia y robos con violencia en el hogar.	(Fay, Planty y Diallo, 2013)
Socioeconómicas	Educación, ingresos, índices de desarrollo, etc.	(ISUC, 2013); (Armas y Herrera, 2018); (Buil Gil, 2019); (Olavarría, 2006)
Sector más victimizado	Posición política y/o ideológica, dimensión valórica, etc.	(Alvira y Rubio, 1982)
Demográficos	Rango etario, sexo, condición de actividad, densidad poblacional, etc.	(INEGI, 2017); (Armas y Herrera, 2018); (Buil Gil, 2019); (Buil Gil, Medina, y Schlomo, 2020); (Olavarría, 2006)
Medidas de seguridad	Medidas comunitarias de seguridad y de política pública desde el enfoque de seguridad y también desde el enfoque social.	(Armas y Herrera, 2018)

Fuente: Elaboración propia basada en la revisión realizada en la aplicación SAE – ENUSC.

Aplicación ENUSC 2018: (I) Especificación, información auxiliar (Cont.)

Cuadro 3a: Covariables sociodemográficas

ID	Indicador / Título del recurso
p_pob_menores15	Proporción de la población menor de 15 años
p_pob_15_24	Proporción de la población entre 15 y 24 años
p_pob_25_49	Proporción de la población entre 25y 49 años
p_pob_50_64	Proporción de la población entre 50 y 64 años
p_pob_65ymas	Proporción de la población de 65 y más
p_pob_14_35	Proporción de la población entre 14 y 35 años
p_hombres	Proporción de la población de hombres
p_mujeres	Proporción de la población de mujeres
p_ocupados	Proporción de la población ocupada
p_desocupados	Proporción de la población desocupada
p_inactivos	Proporción de la población inactiva
p_pob_menores18	Proporción de la población menor de 18 años
p_educ_0	Proporción de la población sin nivel de escolaridad
p_educ_primaria	Proporción de la población con nivel de escolaridad primaria
p_educ_secundaria	Proporción de la población con nivel de escolaridad secundaria
p_educ_terciaria	Proporción de la población con nivel de escolaridad terciaria
p_pob_25a65_educ_terciaria	Proporción de la población entre 25 y 64 años con nivel de educación terciaria
p_educ_sec_ter	Proporción de la población adulta con al menos educación secundaria completa
p_pob_educ_sup	Proporción de la población adulta que alcanzo la educación superior (terciaria)
p_pob_educ_hprim	Proporción de la población adulta que tiene a lo sumo educación primaria incompleta
p_N_agua	Proporción de viviendas con acceso a la red pública de alcantarillado
p_N_viv_inade	Proporción de viviendas en condiciones inadecuada
p_hacina	Proporción de viviendas en condiciones de hacinamiento. Se considera hacinamiento cuando el cociente entre el número de piezas y el número de personas es superior a tres

Fuente: Elaboración propia basado en la información provista por el Censo 2017.

Cuadro 3b: Covariables sobre casos policiales

ID	Indicador / Título del recurso
ts_viole_asal_sex	Tasa de violaciones/asaltos sexuales
ts_robo	Tasa de robos
ts_hurtos	Tasa de hurtos
ts_robo_vehi	Tasa de robo de vehículos
ts_robo_viole_intim	Tasa de robo con violencia e intimidación
ts_robo_viole_hog	Tasa de robos con violencia en el hogar
ts_asal_simple	Tasa de asaltos simples
ts_homi	Tasa de homicidios
ts_robo_resi_priv	Tasa de robos en locales residenciales privados
ts_dist_prob_drog	Tasa de denuncias por distribución de drogas
p_crim	Tasa de criminales

Nota 1: Tasas de casos por cada 100.000 habitantes.

Nota 2: Casos policiales en la comuna donde ocurrió el delito. No necesariamente coinciden con la comuna de residencia de la víctima.

Fuente: Elaboración propia basado en la información provista por la SPD.

Nota : Las covariables en los cuadros 3a y 3b fueron construidas bajo los siguientes criterios:

- Considerando el total de la población de la comuna
- Aplicando filtro que considera solo la fracción urbana de la comuna

Aplicación ENUSC 2018: (II) Análisis y adaptación, y (III) Evaluación

Las etapas de análisis en esta investigación son las siguientes:

1. Estimación de tasas de victimización comunal para los indicadores objetivo por método de estimación directa;

2. Exploración de datos, es decir, verificar la distribución y correlación de los datos, con respecto a los indicadores objetivo;

3. Estimación de la varianza del estimador directo, mediante función de varianza generalizada (FVG), para cada uno de los indicadores: estimaciones y diagnósticos;

4. Selección de variables para modelo de áreas pequeñas:

    - Árboles de decisión;
    - Random Forests. Importancia del poder predictivo de las covariables;
    - Regresión lineal. Criterio de información de Akaike (AIC). Evaluación de modelos considerando diferentes métricas (RMSE, PRESS, R2 Adj., etc.);
    - Criterios temáticos y de control.

Aplicación ENUSC 2018: (II) Análisis y adaptación, FVG

Una forma de obtener un estimador eficiente es modelando ψ_d a través de un modelo log-lineal como sigue

$$\log(\psi_d) = \boldsymbol{x}_d^T\alpha + \varepsilon_d, \qquad d = 1, \dots, D, \tag{1}$$

donde $\varepsilon_d \sim N(0, \sigma_\varepsilon^2)$ independientes, $d = 1, \dots, D$; $\boldsymbol{x}_d^T$ es un vector de variables explicativas, α es un vector de parámetros del modelo que deben ser estimados.

A fin de obtener estimaciones para ψ_d , los valores predichos, ψ_d , del modelo en (2) se obtienen empleando la siguiente expresión:

$$\hat{\psi}_d = \exp\left(\boldsymbol{x}_d^T\hat{\alpha}\right) \times \hat{\Delta}(\hat{\alpha}), \qquad d = 1, \dots, D, \tag{2}$$

donde $\hat{\Delta}(\hat{\alpha}) = \frac{\sum_{d=1}^{D} \psi_d}{\sum_{d=1}^{D} \exp(\boldsymbol{x}_d^T\alpha)}$ es un estimador insesgado de un término de corrección de sesgo propuesto por Hidiroglou et al. (2019) $\left[\Delta = \mathrm{E}\left(\exp(\varepsilon_d)\right)\right]$ y $\hat{\alpha}$ es el estimador de α obtenido mediante el método de mínimos cuadrados, con $\hat{\alpha} = \left(\sum_{d=1}^{D} x_d x_d^{-1}\right)^{-1} \sum_{d=1}^{D} x_d \ \log(\psi_d)$.

Aplicación ENUSC 2018: (II) Análisis y adaptación, FVG

Una propiedad deseable de $\hat{\psi}_d$ es que el promedio del estimador de varianza suavizada, $\hat{\psi}_d$, es igual al promedio de la varianza del estimador directo, ψ_d , es decir

$$\frac{\sum_{d=1}^{D} \hat{\psi}_d}{D} = \frac{\sum_{d=1}^{D} \psi_d}{D}$$

Esto asegura que $\hat{\psi}_d$ sistemáticamente no sobreestima o subestima $\psi_d = E(\hat{\psi}_d)$.

Aplicación ENUSC 2018: (II) Análisis y adaptación, y (III) Evaluación

5. Estimación de tasas de victimización comunal para los indicadores objetivo y RMSE con EBLUP, utilizando modelo FH

    - Uso de transformación arcoseno: estimación de parámetros y estimaciones sintéticas;
    - Verificación que $X^T \boldsymbol{\beta} \in [0,1]$;
    - Verificación de los signos de los $\boldsymbol{\beta}$**s**, es decir, que estén en la dirección temáticamente correcta o esperada;
    - Verificación de supuestos: rango de estimaciones, normalidad, homocedasticidad de los residuales, etc.
    - Verificación de presencia de puntos atípicos y puntos influyentes (distancias de Cook);
    - Cálculo de medida de bondad de ajuste $(R^2_{ideal})^3$. Relación entre estimaciones directas y estimaciones FH;
    - Ajustes para cumplir con propiedad de *benchmarking* con respecto a total regional.

6. Selección del mejor modelo comparando el valor RMSE en cada modelo de estimación.

3 Hidiroglou, M., Beaumont, J-F. & Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, 45(1), 122-123. Statistics Canada, Catalogue No. 12-001-X.

Aplicación ENUSC 2018: (II) Análisis y adaptación, Fay – Herriot (FH)

- Fay y Herriot (1979) analizaron los ingresos per cápita para áreas pequeñas con menos de 1.000 habitantes en Estados Unidos.

- **Modelo de muestreo:**

$$\hat{\delta}_d^{DIR} = \delta_d + e_d, \text{ con } e_d \overset{iid}{\sim} N(0, \psi_d) \tag{3}$$

- **Modelo de vínculo:**

$$\delta_d = \boldsymbol{x}_d'\boldsymbol{\beta} + u_d, \text{ con } u_d \overset{iid}{\sim} N(0, \sigma_u^2) \tag{4}$$

- **Combinando (3) y (4):**

$$\hat{\delta}_d^{DIR} = \boldsymbol{x}_d'\boldsymbol{\beta} + u_d + e_d, \text{ con } u_d \overset{iid}{\sim} N(0, \sigma_u^2) \text{ y } e_d \overset{iid}{\sim} N(0, \psi_d) \tag{5}$$

Aplicación ENUSC 2018: (II) Análisis y adaptación, Fay – Herriot (FH) (Cont.)

- El estimador de Fay – Herriot, $\hat{\delta}_d^{FH}$, puede ser expresado como una combinación convexa entre el estimador directo y el estimador sintético de regresión como sigue:

$$\hat{\delta}_d^{FH} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1 - \hat{\gamma}_d)\, \boldsymbol{x}'_d \boldsymbol{\beta} \qquad (6)$$

Donde:

$\hat{\delta}_d^{DIR}$: Estimador directo del parámetro para el área (o dominio) d.

$\hat{\gamma}_d$: Peso para el estimador directo dado por $\hat{\gamma}_d = \sigma_u^2/(\sigma_u^2 + \psi_d) \in (0,1)$, con σ_u^2 bondad de ajuste del modelo sintético.

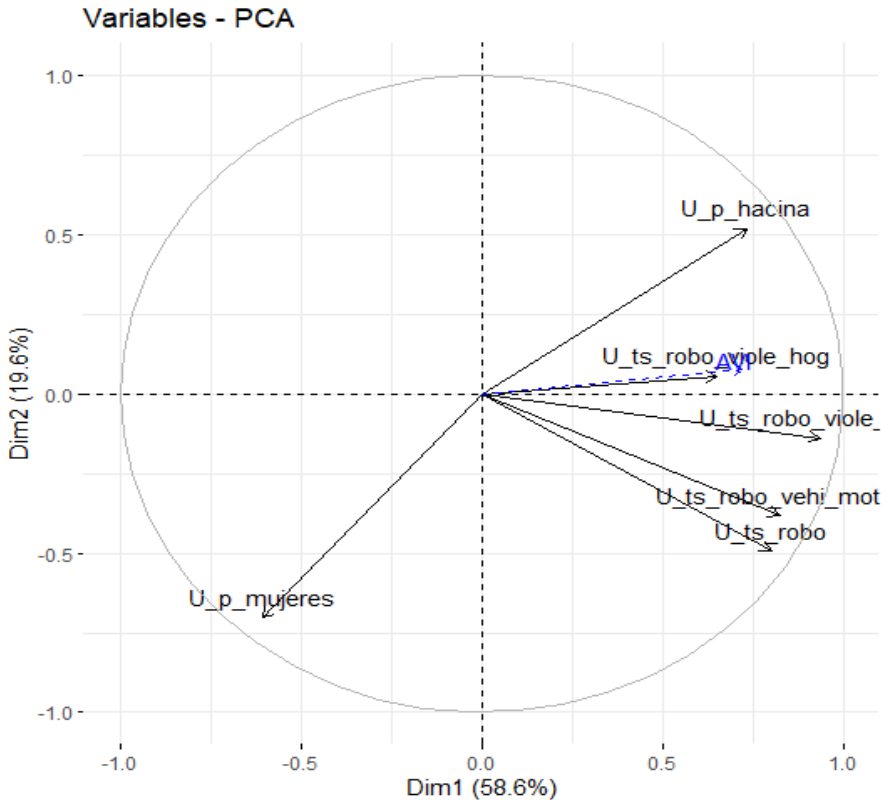
ψ_d : Varianza del estimador directo. Al ser una estimación, puede ser imprecisa.

$\boldsymbol{x}'_d$: Vector de covariables para el área (o dominio) d.

$\boldsymbol{\beta}$: Vector de parámetros obtenido por mínimos cuadrados ponderados.

Aplicación ENUSC 2018: análisis PCA para AVI y RPS

AVI



RPS

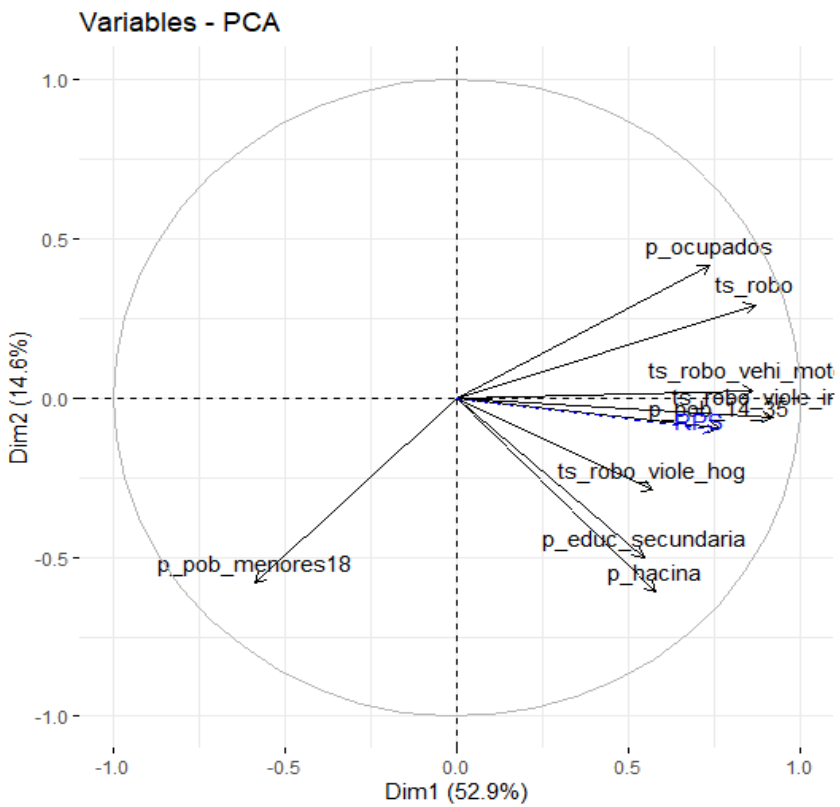
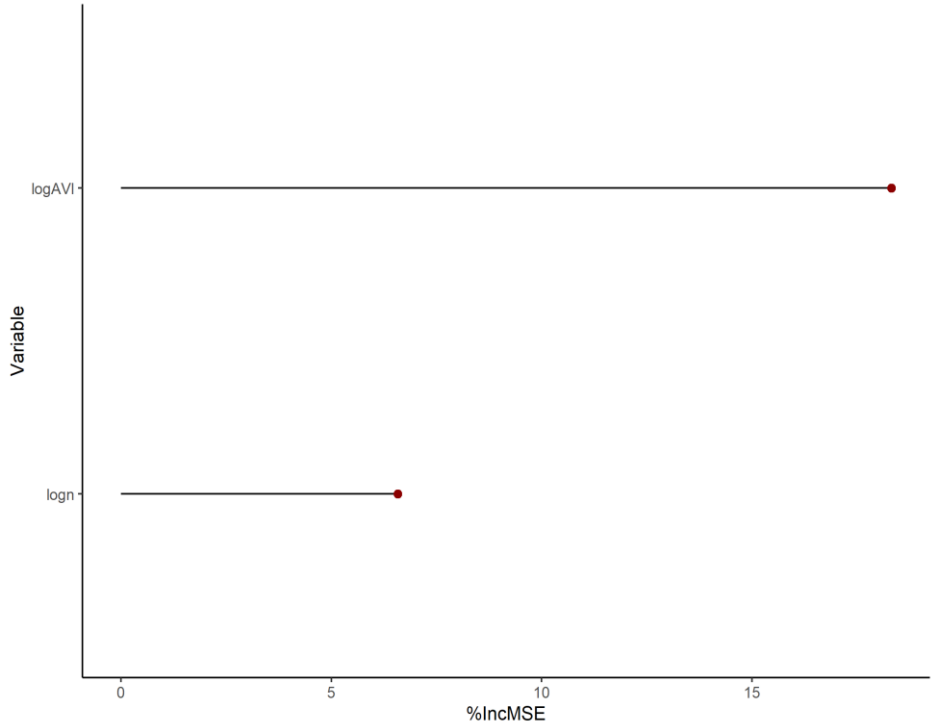


Fig. 4: Análisis exploratorio, PCA: AVI (izquierda), RPS (derecha). Fuente: Elaboración propia.

Aplicación ENUSC 2018: importancia de variables para FVG, AVI y RPS

log(VarAVI)~ logAVI + logn + logAVI*logn, R2 Adj.: 98,65%

log(VarRPS)~ logRPS + logn + logRPS*logn, R2 Adj.: 98,73%



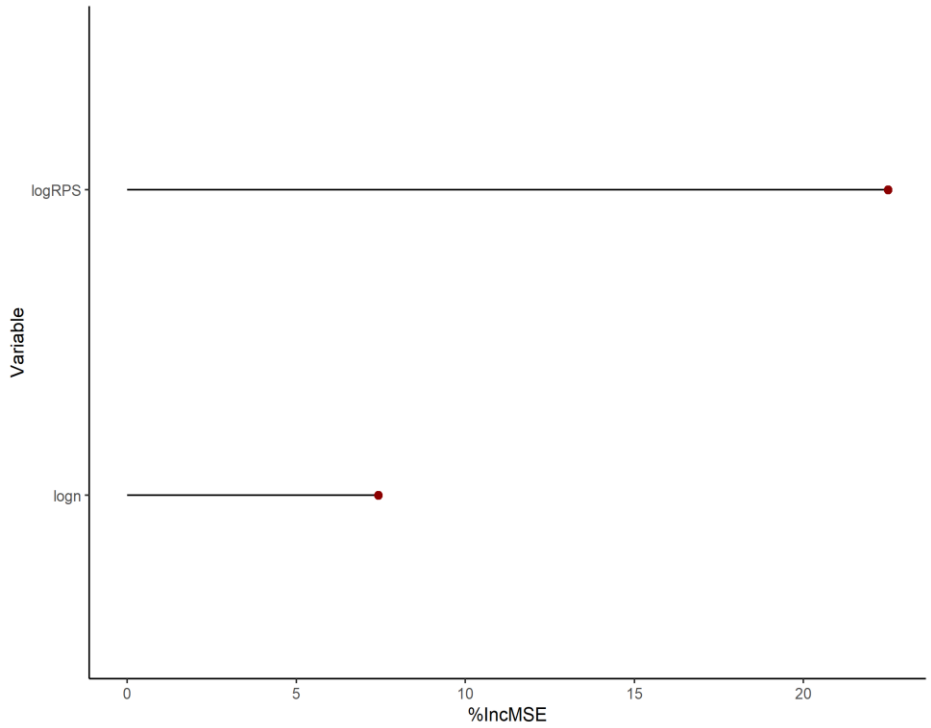


Fig. 5: Importancia de variables (poder predictivo) según *Random Forests* para FVG: AVI (izquierda), RPS (derecha). Fuente: Elaboración propia.

Aplicación ENUSC 2018: importancia de variables para AVI y RPS

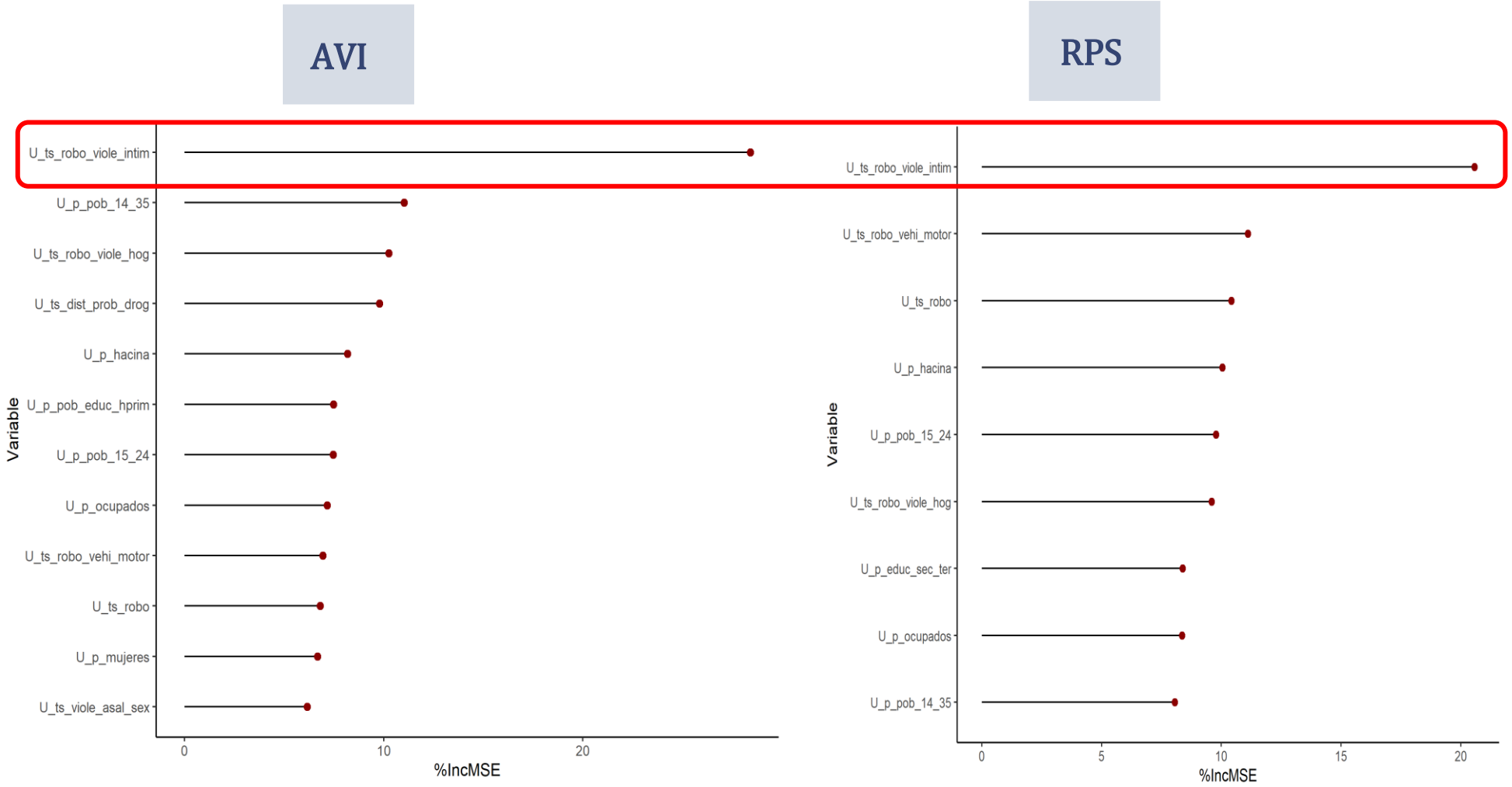
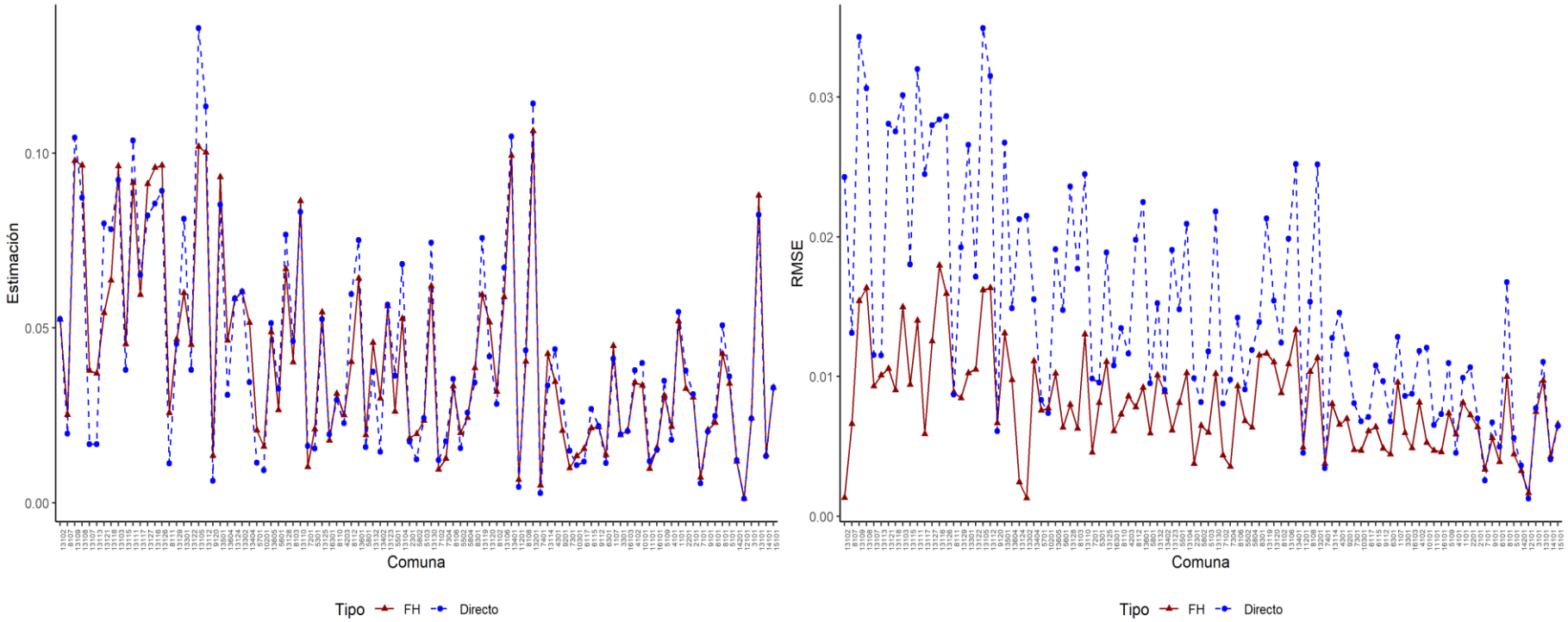


Fig. 6: Importancia de variables (poder predictivo) según *Random Forests*: AVI (izquierda), RPS (derecha). Fuente: Elaboración propia.

Aplicación ENUSC 2018: estimaciones para asalto con violencia o intimidación

$$\arcsin(AVI) \sim ts\_robo\_viole\_intim\_d*** + p\_pob\_14\_35\_d* + factor\ (region)**,$$
$$R2\ (Hidiroglou): 91{,}99\%$$

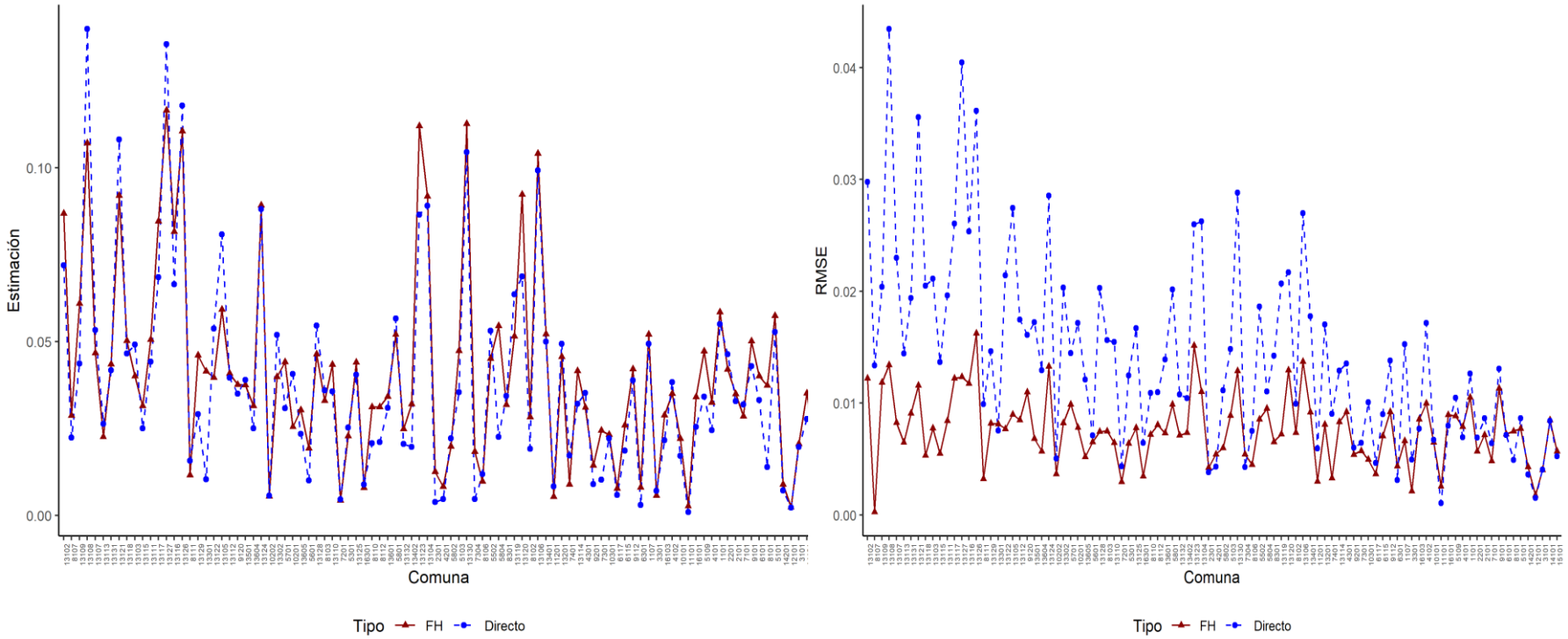


Nota: *: p-valor < 0.10, **: p-valor < 0.05, ***: p-valor<0.01.

Fig. 7: Resultados para asalto con violencia o intimidación: estimaciones (izquierda), RMSE (derecha). Fuente: Elaboración propia.

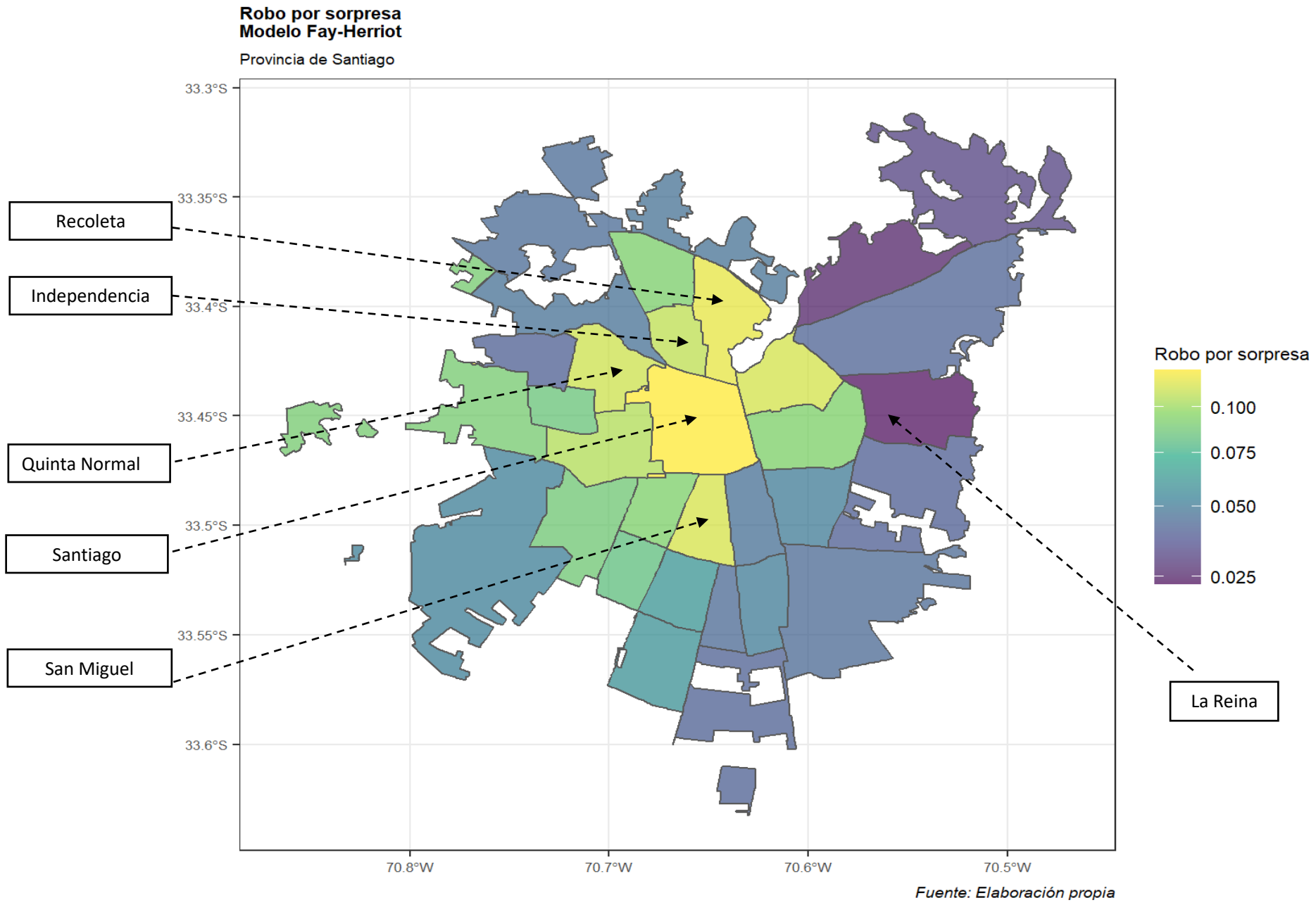
Aplicación ENUSC 2018: estimaciones para robo por sorpresa

arcsin(RPS) ~ ts_robo_viole_intim** ts_rps_d***+
p_pob_14_35_d***+p_educ_sec_ter**+ factor (region)**, R2 (Hidiroglou): 94,78%

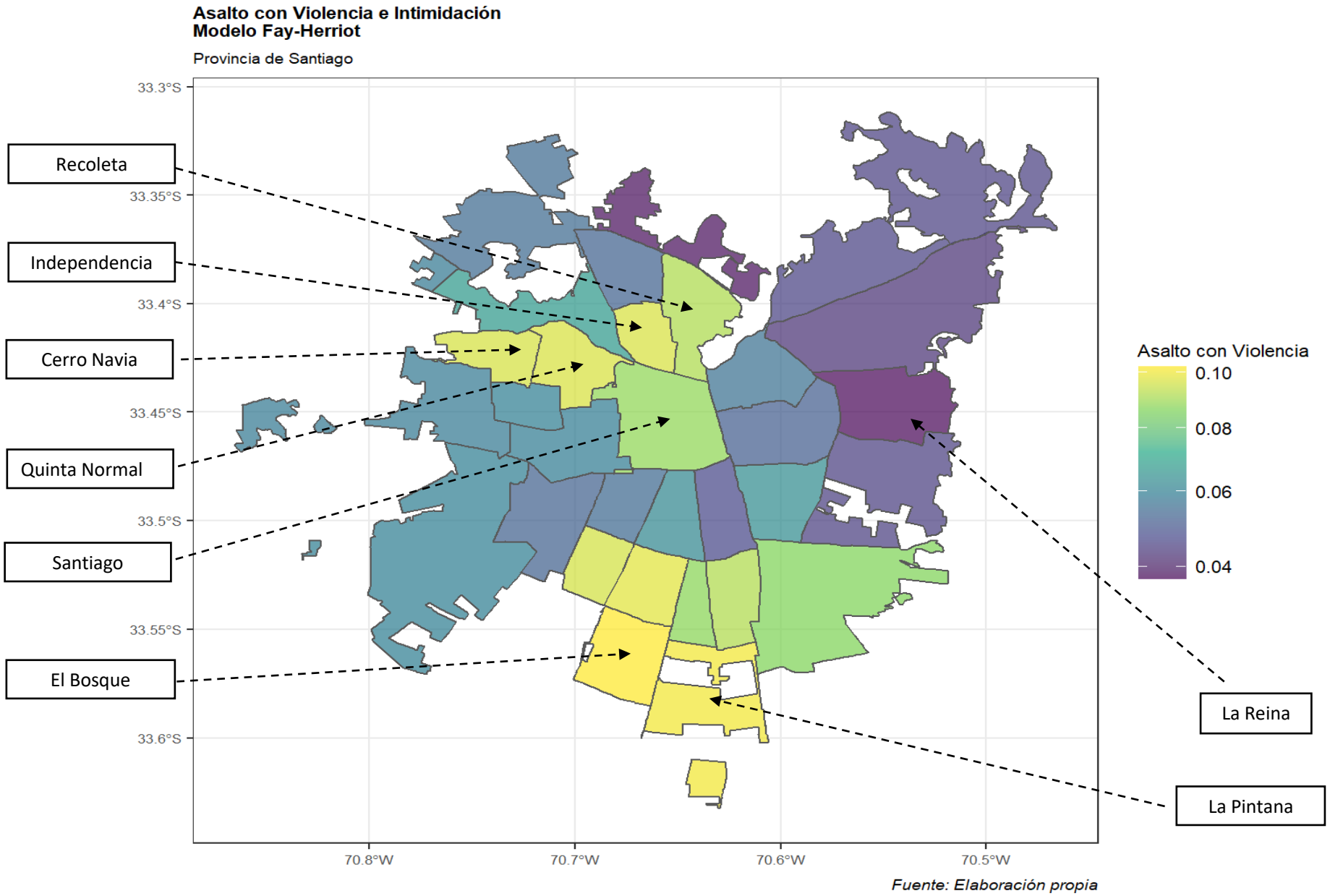


Nota: *: p-valor < 0.10, **: p-valor < 0.05, ***: p-valor<0.01.

Fig. 8: Resultados para robo por sorpresa: estimaciones (izquierda), RMSE (derecha). Fuente: Elaboración propia.



Fuente: Elaboración propia


Asalto con Violencia e Intimidación
Modelo Fay-Herriot
Provincia de Santiago
33.3°S
33.35°S
33.4°S
33.45°S
33.5°S
33.55°S
33.6°S
70.8°W
70.7°W
70.6°W
70.5°W
Asalto con Violencia
0.10
0.08
0.06
0.04
Recoleta
Independencia
Cerro Navia
Quinta Normal
Santiago
El Bosque
La Reina
La Pintana
Fuente: Elaboración propia

Discusión

- Las encuestas sobre delincuencia tienen sus propios problemas metodológicos y el error de medición puede deberse a que las víctimas no recuerdan, sobrestiman o subestiman las situaciones (Buil Gil et al., 2020).

- Notamos que las estimaciones de la ENUSC se producen para las tasas de victimización del área (comuna) donde residen las víctimas de diferentes delitos, mientras que el registro policial utilizado como insumo para los modelos da cuenta de la totalidad de casos policiales que ocurren en cada área.

- Esto puede complicar los esfuerzos para interpretar los resultados del modelo y comparar nuestras estimaciones con los registros policiales.

- Como medida de mitigación, hemos utilizado como registro policial los casos policiales ocurridos en el área de residencia de la víctima. Estos registros poseen una alta correlación con los registros utilizados en los modelos (>0,95).

Aplicación ENUSC 2018: conclusiones, reflexiones y próximos pasos

1. **Precisión**. Para ambos indicadores, el modelo FH entrega estimaciones del RMSE, en general, menores a las obtenidas de manera directa a partir del diseño muestral.

2. **Robustez** (áreas grandes que fueron muestreadas). Las estimaciones SAE son similares a las estimaciones obtenidas con el estimador directo, pero con un error de estimación considerablemente menor. **Los CV obtenidos son menores a 30%**.

3. **Benchmarking**. Las estimaciones comunales son consistentes con las estimaciones regionales, oficialmente publicadas.
4. **Áreas no muestreadas**. El modelo FH permite estimar en áreas no muestreadas.

5. **Inferencia estadística eficiente.** El modelo FH permite realizar un proceso de inferencia estadística eficiente sin la necesidad de elevar mayormente los costos en aras de obtener un mayor nivel de precisión.
6. **Cooperación productiva** entre la academia, instituciones y los profesionales.

7. **Próximos pasos:**
   - Producir estimaciones para los otros indicadores de la encuesta.
   - Estudiar el uso de modelos FH espacio temporal para producir serie de estimaciones por indicador, y modelos FH multivariados para modelar indicadores que estén muy bien correlacionados.
   - Estimación SAE Bayesiano e integración de datos de fuentes alternativas. Uso de RR.AA. con indicadores en contexto de pandemia, Censo de Población y Vivienda 2024, etc.
   - Recopilar la opinión de expertos temáticos y expertos locales en las áreas de interés (comunas).
   - Pasar de un caso de estudio a producción de estadística oficial sobre indicadores de victimización comunal.

4.
Referencias

Fay, R.E., and Herriot, R.A. (1979), Estimates of income for small places: An application of James Stein procedure to census data, *Journal of the American Statistical Association*, 74(366), 269 – 277.

Hidiroglou, M., Beaumont, J-F. & Yung, W. (2019). Development of a small estimation system at Statistics Canada. *Survey Methodology*, 45(1), 101-126. Statistics Canada, Catalogue No. 12-001-X.

Molina, I. (2019). Desagregación de datos en encuestas de hogares: Metodología de estimación en áreas pequeñas, Estudios Estadísticos, CEPAL, 97.

Rao, J. N. (2014). *Small-area estimation*. Wiley StatsRef: Statistics Reference Online.

Tzavidis, N., Zhang, L.-C., Luna Hernandez, A., Schmid, T., y Rojas-Perilla, N., (2018). From Start to Finish: A Framework for the Production of Small Area Official Statistics. *Journal of the Royal Statistical Society, Series A*.

Instituto Nacional de Estadísticas · Chile

GRACIAS

ine.gob.cl