Instituto Nacional de Estadísticas
Departamento de Metodologías e Innovación Estadística
Subdepartamento de Investigación Estadística

21 de enero de 2026

Síntesis metodológica:

Estimación de áreas pequeñas para la Encuesta Nacional Urbana de Seguridad Ciudadana 2024

1. Antecedentes

La "Encuesta Nacional Urbana de Seguridad Ciudadana" (ENUSC), tiene como objetivo 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 (ver resultados: ENUSC 2024). La planificación 2024 permite producir estadísticas oficiales con un nivel de confiabilidad aceptable para niveles nacional-urbano y regional-urbano. En consecuencia, el nivel comunal-urbano (136 comunas) no constituye un dominio planificado y puede ser considerado como un área pequeña. En este escenario, cobra relevancia el uso de estimadores indirectos basados en modelos, que permiten extender el sistema inferencial a desagregaciones de interés no planificadas en el diseño muestral. Dichos estimadores se construyen combinando la información de la encuesta con fuentes auxiliares externas —por ejemplo, censos, registros administrativos e imágenes satelitales— lo que contribuye a obtener estimaciones más precisas y exactas en estas áreas pequeñas. Este conjunto de técnicas que proporcionan cifras estadísticas a un nivel desagregado tiene lugar en una metodología estadística conocida como estimación en áreas pequeñas (en inglés, small area estimation o SAE**). Bajo ese contexto, a continuación, se presenta la implementación de la metodología SAE para la obtención de estimaciones a nivel comunal en el indicador de Victimización a Hogares por Delitos Violentos (VHDV) de la ENUSC 2024.**

2. Metodología e implementación

La estimación en áreas pequeñas se utiliza cuando, mediante las estimaciones directas, no es posible obtener resultados confiables para los dominios de interés, usualmente no planificados y que tienen un tamaño de muestra pequeño o nulo. Generalmente, los dominios están definidos como áreas geográficas (provincias, comunas, barrios) o grupos poblacionales (sexo, edad, grupos étnicos, LGBTIQ+, víctimas, etc.). Para su aplicación se utilizan estimadores indirectos los cuales consideran la información de la encuesta, variables auxiliares de otras fuentes de información y la relación existente con otras áreas o dominios. Al usar estimadores indirectos se está "sacrificando" el insesgamiento de los estimadores basados en el diseño y es necesario escoger un modelo adecuado que se ajuste correctamente a los datos, buscando mejorar la precisión de las estimaciones. Para la implementación de la metodología SAE, se utiliza un modelo de áreas, específicamente EBLUP (Empirical Best Linear Unbiased Predictor) basado en el modelo Fay-Herriot. A continuación, de forma general, se describe:

Sea θ_d el parámetro poblacional (en nuestro caso la proporción de hogares victimizados por delitos violentos VHDV) que se asume que puede ser explicado por un vector de variables auxiliares x_d, siguiendo una relación lineal:

$$\theta_d = x_d^T\beta + u_d, \qquad (\mathbf{1})$$

donde el error de regresión o efecto aleatorio es $u_d \overset{iid}{\sim} N(0, A)$, para $d = 1, \dots, D$.

Dado que θ_d es no observable, el modelo $(\mathbf{1})$ no se puede ajustar y se considera los estimadores directos $\hat{\theta}_d^{DIR}$ (basados en el diseño). En este caso, podemos representar el error debido al muestreo de este estimador mediante el modelo:

$$\hat{\theta}_d^{DIR} = \theta_d + e_d, \qquad (\mathbf{2})$$

donde $e_d \overset{ind}{\sim} N(0, \psi_d), d = 1, \dots, D$. La varianza del error de muestreo ψ_d se asume conocida. Combinando el nivel $(\mathbf{1})$ con $(\mathbf{2})$, el modelo puede expresarse como un modelo de regresión lineal mixto:

$$\hat{\theta}_d^{DIR} = x_d^T\beta + u_d + e_d. \qquad (\mathbf{3})$$

Se asume que los errores de muestreo e_d son independiente de los efectos aleatorios u_d para todo d. Luego, el estimador para θ_d que es lineal para las observaciones $\hat{\theta}_d^{DIR}$, que es insesgado bajo el modelo $(\mathbf{3})$ y que minimiza el error cuadrático medio bajo el modelo BLUP (en inglés, best linear unbiased predictor)

$$\tilde{\theta}_d^{BLUP} = x_d^T\tilde{\beta}(A) + \tilde{u}_d(A), \qquad (\mathbf{4})$$

donde $\tilde{u}_d(A) = \gamma_d(A)\left(\hat{\theta}_d^{DIR} - x_d^T\tilde{\beta}(A)\right)$ es el efecto aleatorio predicho, $\gamma_d(A) = A/(A + \psi_d) \in (0,1)$ y $\tilde{\beta}(A) = [\sum_{d=1}^{D}(A + \psi_d)^{-1}x_d x_d^T]^{-1}\sum_{d=1}^{D}(A + \psi_d)^{-1}x_d\hat{\theta}_d^{DIR}$ es el estimador de β por mínimos cuadrados ponderados.

El BLUP se formula bajo el supuesto de que la varianza del efecto aleatorio A es conocida. En la práctica, A es desconocida, por lo que se reemplaza por un estimador consistente $\hat{A}$, obteniendo el EBLUP. Así, $\hat{\theta}_d^{EBLUP}$ puede interpretarse como una combinación convexa entre el estimador directo y el estimador sintético de regresión:

$$\hat{\theta}_d^{EBLUP} = \begin{cases} \hat{\gamma}_d\hat{\theta}_d^{DIR} + (1 - \hat{\gamma}_d)x_d^T\hat{\beta}, si\ d \in s, \\ x_d^T\hat{\beta}, si\ d \notin s, \end{cases} \qquad (\mathbf{5})$$

donde $\hat{\gamma}_d = \gamma_d(\hat{A}) = \hat{A}/(\hat{A} + \psi_d) \in (0,1)$ y $\hat{\beta} = \tilde{\beta}(\hat{A})$ y s es el conjunto de comunas seleccionadas para modelar. Note que, cuando el estimador directo es confiable, es decir, ψ_d es pequeño comparado con $\hat{A}$, entonces el EBLUP se acerca al estimador directo. En contraste, cuando el estimador directo no es confiable, es decir, ψ_d es grande comparado con $\hat{A}$, entonces el EBLUP se acerca al estimador sintético de regresión.

Una vez definido el estimador, el proceso secuencial utilizado para la estimación SAE de VHDV es el siguiente:

I. **Construcción de base de datos con fuentes auxiliares para la estimación sintética:** Se identifican dimensiones temáticas asociadas a la victimización a hogares por delitos violentos:
    - Características sociodemográficas: Variables asociadas a la exposición al riesgo según el perfil de las personas. Incluye edad, sexo/género, estado civil, nacionalidad y características del hogar.
    - Características socioeconómicas: Factores vinculados al nivel de recursos y oportunidades, que influyen en la

vulnerabilidad y capacidad de protección.

- Infraestructura urbana y entorno: Describe el contexto físico e institucional del territorio y su relación con el control social y la seguridad. Incluye calidad de vivienda, iluminación, densidad urbana, dotación policial, tránsito y acceso a servicios.
- Factores asociados al crimen y la delincuencia: Variables directamente relacionadas con la ocurrencia delictiva y el riesgo de victimización, principalmente provenientes de registros administrativos. Incluye denuncias policiales, tipos de delitos, percepción de inseguridad, consumo de alcohol o drogas y tenencia de armas.
- Características culturales y cohesión social: Factores que reflejan el grado de integración, confianza y control social de la comunidad. Considera participación social, confianza interpersonal, práctica religiosa e ideología política.

Luego, sobre las fuentes de información identificadas, son evaluados criterios como cobertura, consistencia, precisión y actualización, con el objetivo de asegurar la calidad de estos insumos para su uso en SAE. Esta fase, implica que algunos registros o variables deben pasar por un proceso de imputación y en otros casos, se descarta su uso en la implementación SAE.

II. Suavizamiento de la varianza directa mediante Función Generalizada de la Varianza (FGV): Se ajusta un modelo log-lineal a las varianzas estimadas por diseño a nivel de comunas, con el objetivo de obtener estimaciones más estables. Para el modelamiento, se consideran únicamente las comunas con $deff > 1$. Se corroboran los supuestos de normalidad y homocedasticidad de los residuos ($valor\ p > 0.05$), no se detectan observaciones influyentes (distancia de Cook < 1) y el modelo presenta un buen ajuste, con un R^2 ajustado de 94.2%.

III. Aplicación de criterios de calidad para estimaciones directas: Para asegurar que las estimaciones directas de VHDV cumplan las propiedades estadísticas esperadas, se aplica un flujo de criterios de calidad, que considera: Grados de libertad (gl); Tamaño de muestra logrado (n); Efecto de diseño ($Deff$); Casos no ponderados (Y_{np}). La **Tabla 1** muestra el detalle de los criterios de inclusión y exclusión al modelo SAE. Esto implica que, para las 21 comunas excluidas, la estimación SAE solo tendrá el componente sintético asociado al modelo derivado de las fuentes de información utilizadas.

Tabla 1. Resumen de criterios de inclusión y exclusión

	Inclusión	**Exclusión**					
	grados de libertad	tamaño muestral	efecto de diseño	conteo casos no ponderados	grados de libertad	**Total incluir**	**Total excluir**
	$\boldsymbol{gl \geq 14}$	$\boldsymbol{n < 50}$	$\boldsymbol{deff < 1}$	$\boldsymbol{Y_{np} \leq 2}$	$\boldsymbol{gl < 2}$		
N° Comunas	**46**	**6**	**0**	**18**	**0**	**115**	**21**

Fuente: Instituto Nacional de Estadísticas (INE).

IV. Especificación de modelo sintético:

Para el modelamiento del indicador VHDV se utiliza la transformación arcoseno de la raíz cuadrada. En la especificación

del modelo sintético se prioriza la inclusión de variables que otorguen cobertura a cada una de las dimensiones temáticas descritas en la primera parte del proceso (**I.**), asegurando además su significancia estadística. Asimismo, se privilegia la selección de un modelo con adecuada bondad de ajuste ($R^2_{Hidiroglou} = 84\%$). El modelo finalmente seleccionado corresponde a:

Tabla 2. Covariables en modelos SAE para VHDV

Variable	Coeficientes	Error estándar	Valor t	Valor p
(Intercepto)	0.4982	0.0779	6.3929	0.0000
Prop.RoboConIntimidación	0.1498	0.0482	3.1049	0.0019
MedIngAsaDep	-0.1134	0.0455	-2.4897	0.0128
PropVivInadecuada	-0.1916	0.0636	-3.0119	0.0026
PropCampamentos	0.0114	0.0054	2.1032	0.0354
dummy Tarapacá	-0.0204	0.0294	-0.6928	0.4884
dummy Antofagasta	-0.1238	0.0339	-3.6556	0.0003
dummy Atacama	-0.1066	0.0344	-3.1008	0.0019
dummy Coquimbo	-0.1178	0.0424	-2.7751	0.0055
dummy Valparaíso	-0.1371	0.0457	-2.9991	0.0027
dummy Metropolitana	-0.1180	0.0457	-2.5844	0.0098
dummy O'Higgins	-0.1367	0.0454	-3.0092	0.0026
dummy Maule	-0.1646	0.0482	-3.4123	0.0006
dummy Ñuble	-0.1655	0.0502	-3.2973	0.0010
dummy Biobío	-0.1580	0.0474	-3.3318	0.0009
dummy La Araucanía	-0.1180	0.0485	-2.4314	0.0150
dummy Los Ríos	-0.1716	0.0475	-3.6143	0.0003
dummy Los Lago	-0.1565	0.0433	-3.6176	0.0003
dummy Aysén	-0.1634	0.0479	-3.4116	0.0006
dummy Magallanes	-0.1816	0.0485	-3.7415	0.0002
AIC	-359.96			
BIC	-302.31			
$R^2_{Hidiroglou}$	**84%**			

Fuente: Instituto Nacional de Estadísticas (INE).

V. **Cálculo del error cuadrático medio (ECM):** El ECM fue estimado mediante método de bootstrap (B=200) y para la transformación inversa se utiliza Naive.

VI. **Evaluación del modelo:** La **Tabla 3** resume los test de diagnóstico para el modelo bajo un nivel de significancia $\alpha = 0.05$, no evidenciando incumplimiento de los supuestos ($Valor\ p > 0.05$).

Tabla 3. Resumen de criterios de inclusión y exclusión

	Test Shapiro-Wilk		Test Breusch-Pagan	Test Durbin-Watson
	Efectos Aleatorios	Estandarizados	Estandarizados	Estandarizados
VHDV				
Estadístico	0.982	0.982	0.449	1.826
Valor p	*0.134*	0.133	0.503	0.173

Fuente: Instituto Nacional de Estadísticas (INE).

Además, se realizan análisis complementarios para el modelo, que incluyen la revisión de multicolinealidad, la identificación de datos influyentes mediante la distancia de Cook y la evaluación de medidas de precisión como el CVL y el RMSE. Finalmente, se efectúa un pre-benchmarking, que es la agregación de estimaciones EBLUP a nivel regional, corroborando que estos estén contenidos en los intervalos de confianza de las estimaciones regionales directas.

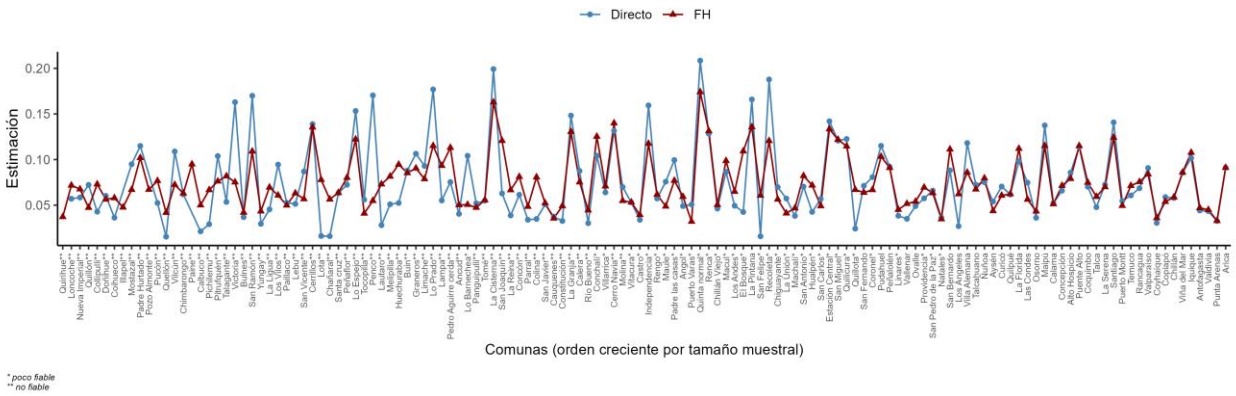
VII. Benchmarking para consistencia con estimaciones regionales:

Cada estimación EBLUP para VHDV se ajusta con un respectivo factor de consistencia regional, esto con el fin de asegurar consistencia entre las estimaciones obtenidas mediante un modelo y las estimaciones oficiales obtenidas bajo diseño muestral de la encuesta. Los 16 factores de corrección son cercanos a uno, dado que el modelo es consistente (pre-benchmarking).

3. Resultados

La ***Figura 1***, muestra las estimaciones para VHDV, con las comunas ordenadas según tamaño muestral de menor a mayor en el eje de las abscisas. Se observa que, en general, las estimaciones EBLUP (en rojo) son más conservadoras que las obtenidas utilizando los estimadores directos (azul) en las áreas con menor tamaño de muestra. Además, como se esperaba, para las comunas con los tamaños de muestra más grandes, por ejemplo, Punta Arenas y Arica, ambas estimaciones son muy similares.

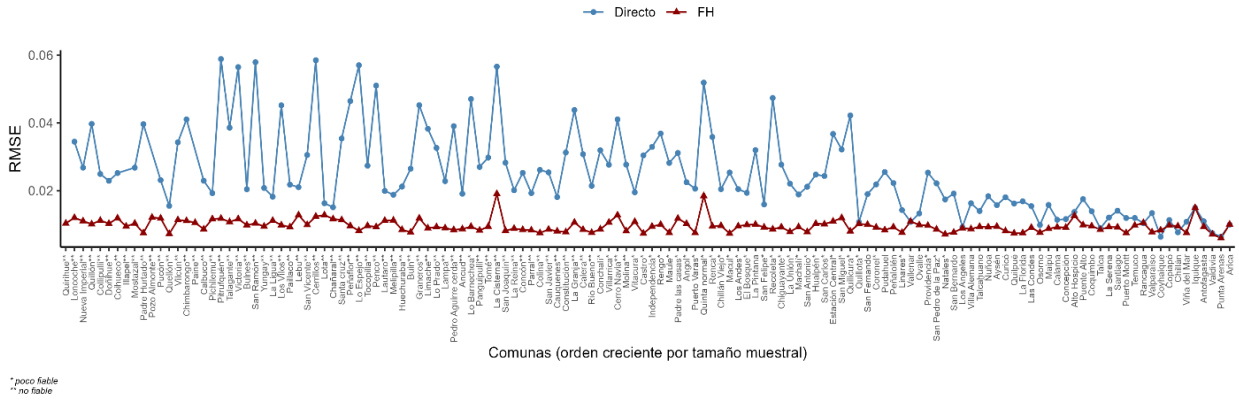
Figura 1. *Estimaciones comunales SAE vs. Directas para VHDV*



Fuente: Instituto Nacional de Estadísticas (INE).

La **Figura *2*** muestra el RMSE (error cuadrático medio de la raíz) estimados de las estimaciones directas y EBLUP. Se observa que, en general, las estimaciones de áreas pequeñas obtenidas a través de EBLUP (rojo) son significativamente más eficientes que las estimaciones directas (azul), especialmente en las áreas de menor tamaño muestral.

Figura 2. RMSE comunales SAE vs. Directas para VHDV



Fuente: Instituto Nacional de Estadísticas (INE).

4. Alcances y conclusiones

La aplicación de la metodología de estimación para áreas pequeñas, mediante el estimador EBLUP basado en el modelo de Fay–Herriot, permite mejorar la precisión de las estimaciones comunales del indicador de VHDV para la ENUSC 2024, especialmente en aquellas comunas donde las estimaciones directas no cumplen con los estándares de calidad definidos por el INE. Previo a la implementación del SAE, se revisa la calidad de las fuentes de información auxiliar, considerando criterios como la cobertura geográfica y temporal, la precisión en los registros y la consistencia en las definiciones operativas de las variables. Asimismo, se logra identificar dimensiones relevantes asociadas a la victimización, lo que permite seleccionar covariables pertinentes en la etapa de especificación del modelo. En la etapa de validación y evaluación, se aplicaron criterios de calidad para las estimaciones directas comunales y una revisión exhaustiva de los modelos, que incluye la evaluación de la consistencia teórica de las covariables, la bondad de ajuste y significancia estadística, la verificación de supuestos para los efectos aleatorios y los residuales estandarizados, y la identificación de observaciones atípicas o influyentes. Previo al benchmarking, se verifica la consistencia de las estimaciones SAE, constatando que su agregación regional se encuentra dentro de los intervalos de confianza de las estimaciones regionales directas. En términos de precisión, el coeficiente de variación logarítmico promedio obtenido mediante SAE en 2024 fue inferior al de las estimaciones directas.

En cuanto a los alcances, los resultados dependen tanto de la calidad de las fuentes de información auxiliar como de las características de las estimaciones directas, en particular de sus medidas de error. Algunos registros administrativos presentaron limitaciones de cobertura, datos faltantes o inconsistencias, lo que requirió imputación o, en ciertos casos, su exclusión del modelo. Adicionalmente, se identificaron desafíos asociados al levantamiento muestral y a la calidad de ciertas estimaciones directas comunales, los cuales fueron abordados mediante ajustes metodológicos específicos. Finalmente, si bien la metodología SAE permitió obtener estimaciones comunales confiables para el periodo 2024, no es posible aplicar las pruebas estadísticas habituales para comparar estos resultados con el periodo 2023, dado que en dicho año las estimaciones comunales se produjeron exclusivamente mediante métodos directos. Esta limitación evidencia la necesidad de avanzar en el desarrollo de estrategias metodológicas que permitan asegurar la comparabilidad interanual de las estimaciones para áreas pequeñas, mediante la definición de modelos consistentes en el tiempo y la continuidad de las fuentes de información y covariables utilizadas.