Informe Metodológico

Estimaciones Comunales de Pobreza por ingresos en Chile Mediante Métodos de Estimación en Áreas Pequeñas

División Observatorio Social MDSF – CEPAL

Diciembre 2021

Este documento reemplaza al documento publicado en noviembre de 2020, correspondiendo a una versión más completa y actualizada del mismo.

Esta es una publicación conjunta entre el Ministerio de Desarrollo Social y de Familia de Chile y la Comisión Económica para América Latina y el Caribe (CEPAL). Por parte de CEPAL, este documento fue preparado por Andrés Gutiérrez, Asesor Regional en Estadísticas Sociales; Felipe Molina y Diego Lemus, consultores de la División de Estadísticas, bajo la supervisión del director de la División de Estadísticas de la CEPAL, Rolando Ocampo.

Las opiniones expresadas en este documento, que no ha sido sometido a revisión editorial, son de exclusiva responsabilidad de los/as autores/as y pueden no coincidir con las de la Organización de la Naciones Unidas o las de los países que representa.

Esta publicación debe citarse como: MDSF-CEPAL, “Estimaciones Comunales de Pobreza por ingresos en Chile Mediante Métodos de Estimación en Áreas Pequeñas - Informe Metodológico”,Santiago de Chile, 2022.

Tabla de contenidos

1. Antecedentes

El Ministerio de Desarrollo Social y Familia (en adelante, el Ministerio) tiene como parte de su misión proveer información acerca de la realidad social y económica del país. Para cumplir con este objetivo el Ministerio realiza la Encuesta de Caracterización Socioeconómica Nacional (Casen), que constituye el principal instrumento de medición de la realidad socioeconómica de los hogares del país, y es utilizada para el diseño y evaluación de la política social existente.

La Encuesta Casen se aplica desde el año 1987 con una periodicidad que ha estado entre dos y tres años. Mediante el uso de métodos estándar para el análisis de datos provenientes de encuestas complejas, esta encuesta permite conocer la situación de pobreza por ingresos y multidimensional de las personas y los hogares, y publicar estadísticas oficiales de la tasa de pobreza a distintos niveles de agregación territorial y por grupos de población prioritarios para la política social.

Si bien el tipo de estimador utilizado para estimar la tasa de pobreza a partir de la Encuesta Casen presenta propiedades deseables para la producción de estimadores insesgados y consistentes para los dominios de estudio que se han predefinido, es decir, a nivel nacional y regional; para subgrupos poblacionales que la Encuesta no abordó en su diseño, como las comunas, estos estimadores - basados únicamente en las unidades de muestreo observadas en la Encuesta – no resultan ser la elección más apropiada, puesto que con muestras pequeñas o nulas no solo reducen su precisión, sino que pierden sus propiedades de consistencia e insesgamiento.

Durante la última década, ha habido una creciente demanda por contar con estimaciones precisas y confiables de indicadores asociados al progreso y bienestar para subgrupos poblacionales no considerados en los diseños de las encuestas, con el fin de tomar decisiones informadas relativas a la implementación de políticas y asignación de recursos. Estos subgrupos poblacionales pueden ser grupos específicos de la población o áreas geográficas más reducidas. En el caso de la Encuesta Casen, esta demanda de información se centra en estimaciones de la tasa de pobreza por ingresos y multidimensional a nivel comunal.

En ese contexto, una posible solución es utilizar procesos de estimación que combinen la información obtenida a través de encuestas con información auxiliar proveniente de registros administrativos o censos. En las últimas décadas, se han producido importantes avances en el desarrollo de estas metodologías, que se conocen, por su sigla en inglés, como estimaciones SAE (*Small Area Estimation*). En Chile, el Ministerio decidió implementar desde el año 2009 una metodología de estimación de áreas pequeñas que ha permitido mejorar la precisión de la estimación de las tasas de pobreza por ingresos a nivel comunal. Para los años 2015 y 2017 se desarrolló e implementó la metodología SAE para estimar la pobreza multidimensional a nivel comunal.

Para las estimaciones 2015 y 2017 se realizó un diagnóstico de los aspectos a actualizar y mejorar de la metodología SAE a partir del desarrollo de la teoría de estimación de áreas pequeñas. Por lo anterior, y a partir del año 2019, se inició una revisión de la metodología implementada en la estimación SAE, con el objetivo de fortalecer los ejercicios venideros. Para este proceso de revisión, el Observatorio Social del Ministerio de Desarrollo Social y Familia contó con la asesoría de la División de Estadísticas de CEPAL.

El objetivo de este documento es presentar los temas tratados en este proceso, considerando diagnósticos, propuestas y conclusiones, que serán el insumo para evaluar la incorporación de algunas actualizaciones a la metodología utilizada hasta ahora, según las mejores prácticas disponibles en la literatura reciente.

1.1. Medición de pobreza a nivel comunal en Chile

Hasta la publicación de los resultados de la Encuesta Casen 2006, diseñados para ser representativos a nivel nacional, regional, y para un conjunto particular de comunas, para las cuales se consideraba que existían niveles aceptables de precisión, las estimaciones comunales se realizaban únicamente a partir de los datos provistos por la Encuesta Casen, utilizando los factores de expansión comunal correspondientes[1]. Desde la Encuesta Casen 2009 en adelante y a raíz de recomendaciones de la Comisión de Expertos convocada por el Ministerio con el objeto de hacer una revisión exhaustiva de la Encuesta Casen, se optó por desarrollar y aplicar métodos más confiables para la producción de estadísticas a nivel comunal, implementándose la metodología de estimación de áreas pequeñas que ha permitido mejorar la precisión de la estimación de las tasas de pobreza por ingresos a nivel comunal.

De esta forma, para las estimaciones comunales de la tasa de pobreza, debe considerarse como fuente oficial la serie de datos que ha sido construida por el Ministerio, utilizando la metodología SAE, la cual produce estimaciones para todas las comunas del país. Esta serie está disponible desde el año 2009 para la tasa de pobreza por ingresos y desde el año 2015 para la tasa de pobreza multidimensional.

La metodología de estimación SAE aplicada por el Ministerio ha evolucionado desde su implementación con el fin de adaptarse a las características de la encuesta y a la disponibilidad de información auxiliar en cada uno de los periodos. Lo anterior, junto con el avance en la aplicación de metodologías SAE en la literatura reciente, motiva una revisión comprehensiva de la metodología implementada en la estimación SAE en Chile con el objetivo de fortalecer los ejercicios venideros.

1.2. Usos de la estimación de la pobreza comunal en Chile

Las estimaciones de la tasa de pobreza a nivel comunal elaboradas por el Ministerio son utilizadas para la asignación de recursos públicos, por lo que es fundamental asegurar su calidad y precisión. Cabe destacar que su uso recomendado es en términos del nivel de pobreza en cada año de medición, y no respecto a la evolución de la tasa entre años.

En esta sección se describen brevemente tres mecanismos de asignación de recursos públicos que utilizan las tasas de pobreza estimadas por el Ministerio a partir de la metodología SAE: (i) Fondo Común Municipal, (ii) Subvención Escolar Preferencial y (iii) Definición de Zonas Rezagadas.

1. Ver http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/pobreza/DOCUMENTO_METODOLOGICO_SAE_2017.pdf

Fondo Común Municipal (FCM)

El Fondo Común Municipal (FCM) es un "mecanismo de redistribución solidaria de los ingresos propios entre las municipalidades del país"[2]. Si bien las municipalidades cuentan con ingresos propios permanentes, solo un pequeño grupo de comunas logra financiarse por medio de sus propios ingresos, por lo que el FCM es un mecanismo de redistribución en beneficio de las comunas más pobres del país.

Los ingresos del FCM se componen por el impuesto territorial, permisos de circulación, patentes comerciales, transferencias de vehículos, recaudación de multas de tránsito, impuesto territorial de inmuebles y aportes fiscales permanentes. Por su parte, la distribución del FCM entre los distintos municipios se calcula durante diciembre del año anterior y se realiza anualmente según los siguientes indicadores y ponderaciones:

- 25% por partes iguales entre las comunas del país.
- 10% con relación al número de pobres de la comuna, ponderado en relación con la población pobre por ingresos del país.
- 30% con relación al número de predios exentos de impuesto territorial de cada comuna.
- 35% en proporción directa a los menores ingresos propios permanentes per cápita.

El FCM establece que "el porcentaje de pobreza comunal equivale a la proporción de personas pobres estimadas de cada comuna, informada por el Ministerio sobre la base de la Encuesta Casen o por el instrumento que la reemplace, o la estimación efectuada por el referido Ministerio"[3].

Subvención Escolar Preferencial (SEP)

La Ley 20.248 del 2008 establece la forma como la Subvención Escolar Preferencial (SEP) entrega recursos adicionales para mejorar la calidad y equidad de la educación chilena en aquellos establecimientos educacionales que atienden a estudiantes con mayores necesidades socioeconómicas. Para ello, se paga una Subvención Escolar Preferencial a los sostenedores de los establecimientos educacionales incorporados al régimen de la Ley SEP, por cada alumno prioritario. Además, se paga una Subvención por Concentración para otorgar más recursos a los establecimientos que tienen un mayor porcentaje de alumnos prioritarios.

Los criterios para determinar la calidad de alumno prioritario para aquellos alumnos cuyos hogares no cuenten con la caracterización socioeconómica de su hogar[4], corresponden a los ingresos familiares del hogar, la escolaridad de la madre (padre o apoderado), la condición de ruralidad del hogar y el grado de pobreza por ingresos de la comuna.

2 https://obtienearchivo.bcn.cl/obtienearchivo?id=repositorio/10221/28638/1/BCN_FCM_ingresos_y_distribucion_GD_def.pdf
3 Decreto 1293.
4 Según el Sistema de Protección de Chile Solidario, el Registro Social de Hogares o Fonasa.

El criterio relativo a la pobreza comunal consiste en que la tasa de pobreza comunal donde tiene domicilio el estudiante sea mayor o igual al doble de la tasa de pobreza nacional, según la última Encuesta Casen o el instrumento que la reemplace[5].

Zonas Rezagadas

Por último, la estimación de la tasa de pobreza comunal se utiliza como insumo para definir las zonas rezagadas en materia de desarrollo social. En el marco del Decreto N°975 publicado el 14 de febrero de 2019 se aprueba el reglamento que fija la Política Nacional de Zonas Rezagadas en Material Social.

Esta política tiene como objeto propiciar el igual acceso de oportunidades entre las personas, independiente del lugar donde habiten, focalizando recursos en aquellos territorios que presentan brechas de mayor magnitud en su desarrollo social, a fin de que dichos territorios alcancen niveles de desarrollo no inferiores a su propia región[6]. Esta política surge como respuesta al diagnóstico de que en Chile existen territorios que presentan importantes rezagos en materia social respecto a la media regional y nacional. Situación acompañada de un considerable grado de aislamiento en relación con los principales centros urbanos y productivos de las regiones donde se ubican[7].

De acuerdo con lo dispuesto por la Política Nacional, la Subsecretaría de Desarrollo Regional y Administrativo (Subdere) debe publicar un listado con las comunas susceptibles de ser propuestas como zonas rezagadas por los gobiernos regionales[8]. Este listado incluye las comunas que cumplen con:

(i) Criterio de aislamiento: entendido como dificultad de accesibilidad y conectividad física, baja densidad poblacional, dispersión en la distribución territorial de sus habitantes, y baja presencia y cobertura de servicios básicos (según el Estudio de Localidades en Condición de Aislamiento elaborado por la Subdere).
(ii) Criterio de brechas sociales: definido como la distancia entre la pobreza comunal y la pobreza regional, lo que corresponde a las diferencias entre el promedio de la tasa de pobreza por ingreso y la tasa de pobreza multidimensional de cada comuna, respecto al promedio regional respectivo de ambas tasas, según la información entregada por el Ministerio a partir de la Encuesta Casen.

1.3. Asesoría de CEPAL

El énfasis de la Agenda 2030 para lograr los Objetivos de Desarrollo Sostenible (ODS), no dejar a nadie atrás (*leave no one behind)* y enfocarse en los más vulnerables primero, representa grandes desafíos y oportunidades para la comunidad estadística. Considerando que la desagregación de datos es fundamental para comprender si el desarrollo de los países está beneficiando a todo el espectro de la

5 https://registroycertificacionate.mineduc.cl/wp-content/uploads/sites/94/2018/03/Decreto_235-1.pdf
6 https://www.bcn.cl/leychile/navegar?idNorma=1128902
7http://territoriosdeconvergencia.subdere.gov.cl/files/doc_zonas_rezagadas/Informe%20Comunas%20Susceptibles%20Zonas%20Rezagadas.pdf
8 Solo podrán ser consideradas como territorios susceptibles de ser propuestos como zonas rezagadas, las comunas cuya brecha social sea mayor o igual a 0.

sociedad, es un desafío para los sistemas estadísticos nacionales generar datos precisos según distintas características de la población y por diferentes niveles geográficos.

En esta línea, el Ministerio acorde a su labor de analizar periódicamente la realidad social nacional a fin de detectar las necesidades sociales de la población y de las familias; y estudiar y proponer las metodologías que utilizará en la recolección y procesamiento de información para la entrega de información de encuestas sociales y otros indicadores, ha considerado necesario contar con la asesoría técnica de la Comisión Económica para América Latina y el Caribe (CEPAL) para la revisión y posterior actualización de metodologías que permitan producir información desagregada y precisa a nivel comunal, utilizando la Encuesta Casen. La División de Estadísticas de la CEPAL, por su parte, entrega asesoría técnica para el fortalecimiento de las capacidades nacionales relativas a la implementación de metodologías que permitan la generación de información desagregada para distintos grupos poblacionales y áreas geográficas de menor tamaño, en particular en lo referente a las metodologías de Estimación para Áreas Pequeñas.

Es por ello que el 19 de noviembre de 2019 se firmó un convenio de colaboración entre ambas instituciones que establece que CEPAL entregará asesoría técnica al Ministerio para la estimación de pobreza en áreas pequeñas con base en la Encuesta Casen. Para el proceso de revisión de la metodología de estimación de áreas pequeñas, se definió un plan de trabajo conjunto basado en el diagnóstico hecho por el Ministerio durante las estimaciones SAE para los años 2015 y 2017. Este plan de trabajo conjunto estuvo principalmente centrado en los siguientes temas:

- Precisión de los datos en áreas pequeñas y diseño muestral de la Encuesta Casen.
- Optimización de los factores de expansión para la obtención de estimaciones comunales.
- Tratamiento de varianzas directas y definición de un modelo suavizado.
- Definición de un modelo sintético predictivo y trasformación de variables.
- Intervalos de confianza para estimaciones sin muestra y viabilidad de la aplicación de bandas.
- Comparabilidad de estimaciones de pobreza entre series.
- Revisión del diseño muestral de la Encuesta Casen, e incorporación de las estimaciones comunales basadas en SAE como objetivo transversal de la operación estadística.
- Evaluación de otro tipo de modelos (espaciales y bayesianos) y validación de los resultados.

Como resultado de este proceso de asistencia técnica, se generaron importantes insumos metodológicos, teóricos y computacionales que permiten evidenciar una evolución del proceso de estimación de la pobreza comunal, incorporando procedimientos robustos que complementan los esfuerzos hechos por el Ministerio desde el 2009.

En este documento se presenta una compilación de esa asesoría, recorriendo de forma lógica y sistemática los procesos involucrados en la generación de las estimaciones de pobreza comunal en Chile. Después de una corta introducción, el segundo capítulo presenta algunas metodologías para la obtención de estimaciones desagregadas, y en particular la metodología implementada en Chile, el tercer capítulo describe el tratamiento de los factores de expansión, el cuarto capítulo presenta los criterios de inclusión que se definieron en esta consultoría, el quinto capítulo muestra la estimación de la varianza del estimador directo, el capítulo 6 presenta el procedimiento final de la estimación de pobreza comunal en Chile incluyendo el modelo sintético y su estimación de varianza, el capítulo 7 muestra la estimación del

error en el modelo Fay-Herriot, el capítulo 8 presenta la validación del modelo y finalmente, el noveno capítulo presenta las conclusiones más importantes en el desarrollo de esta asesoría.

2. Algunas metodologías para la obtención de estimaciones desagregadas

De forma general, el concepto de área pequeña se refiere a un subgrupo poblacional para el cual no es posible obtener resultados confiables con la información proveniente de las encuestas y por ende las estimaciones directas pierden precisión (Rao y Molina, 2015). La concepción de área pequeña no necesariamente está ligada al ámbito geográfico (provincias o comunas), sino que abarca de manera exhaustiva a cualquier subgrupo de una población que puede estar disperso en todo el territorio nacional, como por ejemplo etnias, nivel educativo, o cualquier otro subgrupo no planificado en el diseño de la encuesta, cuyo tamaño de muestra esperado es aleatorio. Nótese que el concepto de área pequeña puede ser válido incluso para grandes regiones que no están bien representadas en la encuesta. En estos casos se puede aplicar modelos SAE para obtener estimaciones más precisas en los dominios considerados como áreas pequeñas.

A continuación, se describen brevemente algunos modelos de estimación de áreas pequeñas utilizados para la construcción de indicadores de pobreza. Finalmente se resume la metodología que tradicionalmente es aplicada en Chile por el Ministerio de Desarrollo Social y Familia.

2.1 Algunos modelos de estimación en la construcción de mapas de pobreza

Los modelos de estimación en áreas pequeñas pueden ser clasificados según el nivel de agregación de las variables auxiliares. Las categorías más conocidas en la literatura relacionada son los modelos de área (o nivel agregado) y los modelos a nivel de unidad. En el primer grupo se tienen los modelos que enlazan un conjunto de variables agregadas dentro de las áreas pequeñas con sus respectivos estimadores directos. Dichos modelos se hacen necesarios cuando existen restricciones de confidencialidad en los datos a nivel de unidad o cuando estos no están disponibles. Dentro del segundo grupo se encuentran los modelos que permiten enlazar el valor de la variable de respuesta con el valor específico de un conjunto de covariables a nivel de unidad. A continuación, se presenta la descripción de los modelos SAE más reconocidos en la literatura académica relacionada, y que a su vez han sido considerados por la CEPAL para la construcción de mapas de pobreza en los países de América Latina y el Caribe.

2.1.1 Modelo Fay-Herriot (FH)

El modelo de área más reconocido en la literatura es el propuesto por Fay y Herriot (1979). El modelo FH está compuesto por dos niveles, considerando que el enlace entre un estimador directo $\hat{\theta}_d^{Dir}$ a nivel de área, insesgado bajo el diseño de muestreo, y un conjunto de covariables $\boldsymbol{x}_d$ disponibles varía de forma constante en los dominios (d) a través del siguiente modelo de regresión lineal:

$$\hat{\theta}_d^{Dir} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d, \qquad d = 1, \dots, D$$

por definición, $\boldsymbol{\beta}$ corresponde al vector de coeficientes de regresión común (constante) a todas las áreas y u_d efecto aleatorio o término de error del modelo específico para el dominio d. En este modelo u_d representa la heterogeneidad en el indicador de pobreza en los dominios de interés que no son capaces

de cuantificar las variables auxiliares consideradas y e_d representa el error de muestreo en el d-ésimo dominio, independiente de su respectivo efecto aleatorio. Bajo el modelo FH más simple, se supone que los efectos aleatorios son independientes entre las áreas y siguen una misma distribución con media constante e igual a cero y varianza común σ_u^2, es decir, $u_d \sim^{iid}(0, \sigma_u^2)$. Se asume también que los errores de muestreo son independientes entre sí, tienen media cero y varianzas conocidas ψ_d; es decir, $e_d \sim^{ind}(0, \psi_d)$ – Ver Molina, I. (2019) y Rao, J.N.K. y Molina, I. (2015) -.

En la práctica, σ_u^2 y ψ_d son desconocidas y son reemplazadas por estimadores consistentes $\widehat{\sigma_u^2}$ y $\hat{\psi}_d$ en el proceso de estimación del modelo. Para la primera se emplean técnicas tradicionales de estimación, como, por ejemplo, el método no paramétrico basado en momentos -Fay-Herriot (1979)-, el enfoque por máxima verosimilitud y el de máxima verosimilitud restringida o residual, mientras que para segunda se emplea, usualmente, la expresión de la varianza del estimador directo bajo el diseño de muestreo a partir de los datos de la encuesta de hogares, es decir, $\hat{\psi}_d = \widehat{\text{var}_\pi}(\hat{\theta}_d^{Dir}|F_{\alpha,d})$ para los dominios bajo estudio.

2.1.2 Modelo con errores anidados (BHF)

Battese, Harter y Fuller (1977) propusieron un modelo con errores anidados que relaciona linealmente una variable de interés (Y_{di}) con los valores de p covariables observados para el i-ésimo elemento del d-ésimo dominio d. El modelo a nivel de individuo propuesto sigue la siguiente ecuación:

$$Y_{di} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D$$

$\boldsymbol{\beta}$ corresponde al vector de coeficientes de regresión común (constante) a todas las áreas, u_d es el efecto aleatorio o término de error del modelo específico para el dominio d (suponiendo que, $u_d \sim^{iid}(0, \sigma_u^2)$) y e_{di} define el error a nivel de individuo. El modelo asume que $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, siendo k_{di} un conjunto de constantes conocidas asociadas a la posible heteroscedasticidad en los individuos dentro de cada área. Al igual que el modelo FH, se asume que los efectos aleatorios u_d son independientes al error a nivel de individuo e_{di}. –Ver Molina, I. (2019) y Rao, J.N.K. y Molina, I. (2015) -.

Este procedimiento divide la media poblacional respecto a los elementos seleccionados en la muestra de este dominio s_d (valores observados) y otra a los individuos no muestreados (r_d), así $\bar{Y}_d = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} Y_{di}\right)$. Molina, I. (2019) afirma que el EBLUP de $\bar{Y}_d$ bajo el modelo BHF se obtiene al ajustar el modelo a los datos de la muestra $(Y_{di} \in s_d)$ y pronosticar los valores de la variable de interés en los elementos no muestreados en cada dominio $(Y_{di} \in r_d)$, tal como se presenta en la siguiente ecuación,

$$\hat{\bar{Y}}_d^{BHF} = N_d^{-1}\left(\sum_{i\in s_d} Y_{di} + \sum_{i\in r_d} \hat{Y}_{di}^{BHF}\right)$$

2.1.3 Método de Elbers, Lanjouw y Lanjouw (ELL)

Elbers, C., Lanjouw, J.O. y Lanjouw, P. (2003), propusieron un modelo a nivel de unidad que puede modelar de forma precisa una gran variedad de indicadores si estos pueden ser expresados como una función del poder adquisitivo de los individuos de una población como, por ejemplo, el ingreso o gasto per cápita.

Corral, P., Molina, I. y Nguyen, M. (2020) afirman que este enfoque es el tradicionalmente implementado por el Banco Mundial con el fin de construir mapas de pobreza para los países a nivel mundial. El método ELL asume que la transformación logarítmica del poder adquisitivo (E_{di}) del i-ésimo hogar dentro de cada ubicación d está relacionado linealmente con un conjunto de características de este, es decir:

$$Y_{di}^{*} = \boldsymbol{x}_{di}'\boldsymbol{\beta} + u_d + e_{di}, \quad i = 1, \dots, N_d,\ d = 1, \dots, D$$

siendo $Y_{di}^{*} = \log(E_{di} + c)$ la transformación Log-Shift de la variable de bienestar considerada; $c > 0$ es una constante a ser estimada por algún método de optimización que garantice que genere una variable con distribución aproximadamente normal.

Bajo este modelo, u_d es el efecto aleatorio de la ubicación y e_{di} el termino de error asociado a cada hogar de la población. Al igual que el modelo BHF se asume que $u_d \sim^{iid}(0, \sigma_u^2)$ y $e_{di} \sim^{ind}(0, \sigma_e^2 k_{di}^2)$, siendo u_d y e_{di} independientes, y k_{di} constantes conocidas que representan la posible heteroscedasticidad. Finalmente, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión.

2.1.4 Mejor predictor empírico (EBP) bajo el modelo con errores anidados

Molina, I. y Rao, J.N.K. (2010) proponen un modelo con errores anidados que permite estimar indicadores no lineales como la incidencia y la brecha de pobreza. Según Molina, I. (2019), esta propuesta permite estimar un indicador FGT de orden α mediante un modelo que asume que la transformación Log-Shift del ingreso (E_{di}) del i-ésimo hogar dentro del área d está relacionado linealmente con un conjunto de características de este

$$Y_{di}^{*} = \mathrm{x}_{di}^{T}\boldsymbol{\beta} + u_d + e_{di}.\ \ i = 1, \dots, N_d,\ d = 1, \dots, D,$$

siendo $Y_{di}^{*} = \log(E_{di} + c)$, $\boldsymbol{\beta}$ es el vector de coeficientes de las covariables, u_d es el efecto aleatorio del área tal que $u_d \overset{i.i.d}{\sim} N(0, \sigma_u^2)$ y $e_{di} \overset{iid}{\sim} N(0, \sigma_e^2)$ son los errores a nivel individuo independientes de los efectos aleatorios. Al igual que el método ELL, $c > 0$ es una constante a ser estimada por algún método de optimización que garantice que genere una variable con distribución aproximadamente normal.

Bajo el modelo de errores anidados, la distribución de $\boldsymbol{y_{dr}}\,|\boldsymbol{y_{ds}}$, necesaria para calcular el mejor predictor (EBP), se obtiene al descomponer los vectores $\boldsymbol{y_d}$ y las matrices $\mathrm{X}_{\boldsymbol{d}}$ y $\mathrm{V}_{\boldsymbol{d}}$ en la parte asociada a los elementos observados en la encuesta de hogares y aquella fuera de la muestra, de la siguiente forma:

Siguiendo la propuesta de Molina, I. y Rao, J.N.K. (2010, 2015), un estimador del indicador de pobreza FGT se obtiene al desagregar el total en dos partes, una asociada a las observaciones dentro de la muestra y otra a las que están por fuera de esta, tal como se presenta a continuación:

$$\tilde{\theta}_d^{EBP}(\theta) = \frac{1}{N_d}\left(\sum_{i \in S_d} F_{\alpha,di} + \sum_{i \in r_d} \tilde{F}_{\alpha,di}^{EBP}\right)$$

En donde $F_{\alpha,di} = \left(\frac{z^* - Y^*_{di}}{z^*}\right) I(Y^*_{di} < z^*)$ y $\tilde{F}^{EBP}_{\alpha,di}$ es la esperanza (mejor predictor lineal insesgado) de las variables para los individuos que no perteneces a la muestra. Además, $z^* = log\,(z + c)$, en donde z corresponde a la línea de pobreza.

2.2 Metodología de estimación comunal de pobreza en Chile

Si bien existe una amplia variedad de métodos de estimación de áreas pequeñas, la metodología aplicada por el Ministerio se basa en el modelo áreas de Fay-Herriot, que plantea el uso de un estimador compuesto definido como una combinación lineal entre un estimador sintético y el estimador directo, mediante una ponderación obtenida como función de la estimación de las varianzas de cada estimador. Mientras más pequeña sea la estimación de la varianza asociada a la estimación directa, en comparación con la heterogeneidad no explicada por esta, mayor será la ponderación que se le otorgará a dicha estimación y viceversa.

El modelo de área de Fay-Herriot para las estimaciones de tasa de pobreza a nivel comunal fue elegido por sobre los modelos de unidad (por ejemplo, el método ELL) por considerarse que el Ministerio tiene a su disposición una gran riqueza de información administrativa que proviene de todos los Ministerios y Servicios del Estado; esto además se complementa con información proveniente del último Censo de Población y Vivienda disponible. De esta forma, el Ministerio está en una posición privilegiada para incorporar información auxiliar proveniente de distintas fuentes en las estimaciones SAE.

El modelo de Fay-Herriot utilizado en Chile, supone dos niveles de estimación: en el primer nivel se asume que la pobreza a nivel de la comuna se puede explicar por covariables observables $\boldsymbol{x}_d$ y un error comunal, denominado efecto aleatorio, u_d; mientras que en el segundo nivel se asume una interacción con la estimación directa, proveniente del modelo de muestreo a través de Casen, en el cual se aproxima la tasa de pobreza mediante una estimación que tiene asociado un error de muestreo e_d, que se supone conocido:

- Nivel 1: Modelo de vínculo (estimación sintética)
  $\theta_d = \boldsymbol{x}_d{}'\beta + u_d\,,\ con\ \ u_d \overset{iid}{\sim}(0, \sigma_u^2)$

- Nivel 2: Modelo de muestreo (estimación directa)
  $\hat{\theta}_d^{Dir} = Y_d + e_d,\quad con\ \ e_d \sim^{ind}(0, \psi_d^2)$

Estos dos niveles pueden ser especificados en una sola expresión algebraica:

$$\hat{\theta}_d^{FH} = \boldsymbol{x}_d{}'\beta + u_d + e_d,\quad con\quad u_d \overset{iid}{\sim}(0, \sigma_u^2)\,,\ e_d \overset{ind}{\sim}(0, \psi_d^2)$$

La cual, a su vez, puede expresarse como una ponderación de las estimaciones directas y las estimaciones sintéticas:

$$\hat{\theta}_d^{FH} = \hat{\gamma}_d \hat{\theta}_d^{Dir} + (1 - \hat{\gamma}_d)\hat{\theta}_d^{syn}\,,\ con\ d = 1, \dots, D$$

En donde $\hat{\theta}_d^{Dir}$ denota la estimación directa, $\hat{\theta}_d^{syn} = \boldsymbol{x}_d{}'\hat{\beta}$ corresponde a la estimación sintética usando variables auxiliares, y $\hat{\gamma}_d = \widehat{\sigma_u^2}/\left(\widehat{\sigma_u^2} + \hat{\psi}_d^2\right)$ corresponde al ponderador.

Según lo descrito en el modelo, para que las estimaciones SAE definan una buena aproximación a la pobreza comunal, se debe contar con estimaciones confiables, insesgadas y consistentes de las varianzas de los estimadores directos de la tasa de pobreza, además de asumir un modelo predictivo correcto para la estimación sintética. Siguiendo las recomendaciones internacionales (Eurostat, 2019), para realizar una estimación SAE apropiada es deseable contar con:

- Información administrativa disponible, actualizada y de calidad, para los niveles de desagregación de interés. Se debe explorar la capacidad explicativa de estas variables con el fenómeno que se busca estimar.
- Tamaños muestrales apropiados en la estimación directa, que entreguen estimaciones con cierto nivel de precisión.

De lo anterior se desprende la importancia de tener información de registros administrativos de calidad y de contar con estimaciones directas con ciertos niveles mínimos de precisión que permitan utilizar insumos de calidad en la producción de estimaciones en áreas pequeñas.

3. Tratamiento de los factores de expansión

Un diseño muestral apropiado debe garantizar que en los niveles de estimación definidos (nacional, urbano/rural y regional, para el caso de Casen), se obtendrán estimaciones insesgadas, precisas y consistentes que representan el verdadero fenómeno de la pobreza en el país y en las regiones. Con base en el diseño muestral, se definen de forma teórica los factores de expansión que corresponden al inverso de la probabilidad de selección y los procedimientos apropiados para la estimación de la varianza.

Como resultado del trabajo de campo, se deben realizar ajustes adicionales a los pesos de muestreo, entre los cuales se cuentan ajustes por cobertura (omisión de conglomerados), por elegibilidad (no elegibles y elegibilidad desconocida) y por ausencia de respuesta. Tradicionalmente, el último ajuste realizado a los factores de expansión corresponde a la calibración regional, provincial y comunal que busca cuadrar las estimaciones poblacionales obtenidas a partir de la Encuesta Casen con las proyecciones obtenidas del Censo.

Es posible que los factores de expansión teóricos inducidos por el diseño de muestreo original, se hayan debido ajustar de tal forma que se obtengan pesos más extremos e influyentes en algunas comunas. Un conjunto de pesos más dispersos o con valores extremos, tendrá implicaciones negativas en la estimación de la varianza, lo que a su vez impactará en la calidad de la ponderación que el modelo SAE le otorga a la estimación directa. Por lo anterior, para las estimaciones de área pequeña es necesario realizar un recorte o suavización de los factores de expansión, lo que se describe en la sección 3.2.

En última instancia, las encuestas multipropósito pueden realizar otro ajuste a los factores de expansión que tenga en consideración la distribución original de los factores teóricos. Los ajustes se hacen precisamente para evitar los pesos extremos y recuperar la distribución original de los factores de expansión que garantiza, de acuerdo con el diseño muestral, estimaciones insesgadas, precisas y consistentes.

3.1 Ajustes a los pesos de muestreo

En condiciones ideales el marco de muestreo debería coincidir plenamente con la población que se quiere estudiar; pero en general no siempre es posible contar con una lista de todos los elementos de la población. En el contexto de las encuestas de ingresos y gastos, no existe una lista que enumere a todos los hogares de un país de manera actualizada y, por ende, la práctica estándar es construir el marco de muestreo en varias etapas, seleccionando una muestra de áreas geográficas, luego realizando una actualización sobre las viviendas ocupadas por hogares particulares (empadronamiento exhaustivo de todos los hogares en las áreas seleccionadas), para finalmente seleccionar hogares.

Este esquema de selección permite establecer que al final todos los hogares (seleccionados o no) tengan una probabilidad conocida y distinta de cero de pertenecer a la muestra final. Sin embargo, como el marco de muestreo de las encuestas a hogares presenta imperfecciones, debido a la desactualización natural que conlleva este tipo de instrumentos, es necesario corregirlos para eliminar, o al menos reducir significativamente, el sesgo causado por estos inconvenientes no muestrales.

Valliant y Dever (2017) consideran tratar la ausencia de respuesta de manera diferenciada, mediante la clasificación de cada categoría del resultado final de la encuesta en alguno de los siguientes grupos de unidades:

- ER (unidades elegibles que fueron respondientes efectivos): casos elegibles para los cuales se ha recolectado una cantidad suficiente de información.
- ENR (unidades elegibles no respondientes): casos elegibles para los cuales no se recolectó ningún dato o la información recolectada es muy precaria.
- IN (unidades no elegibles): casos de miembros no elegibles que no hacen parte de la población de interés.
- UNK (unidades con elegibilidad desconocida): casos en donde no se puede conocer si la unidad es elegible o no.

Tras construir esta nueva clasificación, es posible hacer más expedito el manejo de los factores de expansión que, en general, puede seguir los siguientes procesos recomendados por CEPAL en la creación de los insumos para el análisis de las encuestas de hogares.

1. **Creación de los pesos básicos:** asociado a cada esquema particular de muestreo existe una única función que vincula a cada hogar con una probabilidad de inclusión en la muestra. De esta forma: $\pi_k = Pr(k \in s)$. Por lo tanto, el primer paso del anterior esquema induce la creación de los pesos básicos d_k que se definen como el inverso multiplicativo de la probabilidad de inclusión del hogar $d_{1k} = \frac{1}{\pi_k}$.

   Estos pesos son creados incluso para aquellas unidades que serán excluidas de la muestra porque no son elegibles o porque no entregaron ninguna información y luego serán modificados convenientemente.

2. **Ajuste por elegibilidad desconocida:** el siguiente paso consiste en redistribuir el peso de las estructuras cuyo estado de elegibilidad es desconocido. Esta situación se presenta a nivel del hogar cuando no se puede contactar porque nunca se atendió el llamado del encuestador y no se logró comunicar con nadie en la estructura (nadie en casa). En este caso existen dos posibilidades: que en la estructura sí haya hogares particulares o que en la estructura no habite ninguna persona. Por este motivo, se acostumbra a repartir los pesos de las unidades con elegibilidad desconocida entre las unidades que sí disponen de un estatus de elegibilidad (unidades elegibles encuestadas, unidades elegibles no encuestadas, unidades no elegibles).

   Si no es posible determinar la elegibilidad de algunas unidades que aparecen en el marco de muestreo, se tendrá una muestra s que contendrá el conjunto de las unidades elegibles en la muestra s_e, el conjunto de las unidades no elegibles en la muestra s_n y el conjunto de las unidades con elegibilidad desconocida s_u. En este último caso, la elegibilidad es desconocida a no ser que de manera arbitraria se clasifiquen como unidades elegibles no encuestadas o que se tenga información auxiliar en el marco de muestreo que permita imputar su estado de elegibilidad.

   Se recomienda formar B $(b = 1, \dots, B)$ categorías basadas en la información del marco de muestreo. Estas categorías pueden ser estratos o cruces de subpoblaciones. Siendo s_b la muestra

de unidades en la categoría b (que incluye unidades elegibles encuestadas, unidades elegibles no encuestadas y unidades con elegibilidad desconocida), se define el factor de ajuste por elegibilidad como $a_b = \frac{\sum_{s_b} d_{1k}}{\sum_{s_b \cap s_e} d_{1k}}$. Para la categoría b, los pesos ajustados por elegibilidad desconocida para aquellas unidades cuya elegibilidad sí pudo ser establecida (independientemente de su estado de respuesta) estarán dados por $d_{2k} = a_b * d_{1k}$.

3. **Descarte de las unidades no elegibles:** si hay estructuras seleccionadas desde el marco de muestreo que han cambiado su estado de ocupación y ahora no contienen ningún hogar, entonces el siguiente paso consiste en ajustar su peso básico de la siguiente manera:

$$d_{3k} = \begin{cases} 0, & \text{si la unidad } k \text{ no pertenece a la población objetivo} \\ d_{2k}, & \text{en otro caso} \end{cases}$$

   De esta forma, las unidades que se identificaron como no elegibles pierden su peso de muestreo y no expanden a la población final, con lo que se da cuenta del proceso de desactualización de los marcos de muestreo.

4. **Ajuste por ausencia de respuesta:** en este paso los pesos básicos de los ER se ajustan para tener en cuenta a los ENR. Al final del proceso, los pesos de los ER se incrementan para compensar el hecho de que algunas unidades elegibles no entregaron información. Al suponer que la distribución de las respuestas puede ser estimada, entonces la probabilidad de respuesta, conocida como *propensity score* (Särndal y Lundström 2006), está dada por ϕ_k.

   Si fuera posible tener acceso a las covariables $\mathbf{x}$, que son determinantes de que el hogar no realice la entrevista, entonces sería posible estimar el patrón de ausencia de respuesta mediante la siguiente expresión $\hat{\phi}_k = f(\mathbf{x}_k, \widehat{\boldsymbol{\beta}})$. Bajo este escenario, es posible definir el siguiente estimador insesgado $\hat{t}_y = \sum_{k \in s_r} d_{4k}\, y_k$. En donde $d_{4k} = \frac{d_{3k}}{\hat{\phi}_k}$ representa el ajuste al factor de expansión debido a la no respuesta.

5. **Calibración:** es un ajuste que se realiza a los pesos de muestreo con el propósito de que las estimaciones de algunas variables de control reproduzcan los totales poblacionales de estas variables (Silva 2004). Este es usualmente el último paso en el ajuste de los ponderadores y hace uso de información auxiliar que reduce la varianza para corregir los problemas de cobertura que no pudieron ser rectificados en los pasos previos[9].

   El objetivo de la calibración es obtener un nuevo sistema de ponderadores w_k que se encuentren cerca de los ponderadores básicos d_k, de tal forma que cuando los ponderadores sean usados para estimar los totales de las variables auxiliares, dichos totales sean reproducidos con exactitud.

6. **Recorte de los pesos:** un inconveniente que se genera debido a la multitud de ajustes en los factores de expansión es que, si bien el estimador resultante tendrá un sesgo cercano a cero, la

9 Ver Alvarado, M., y Pizarro, M. (2019a) para ver detalles de la nueva metodología de calibración de factores de expansión implementada por el INE en la Encuesta Nacional de Empleo.

distribución de los pesos calibrados puede mostrar datos extremos, tanto a la derecha (valores muy grandes) como a la izquierda (valores menores que uno), que hacen que la varianza del estimador crezca y que, por ende, la precisión de la inferencia decrezca. Para hacerle frente a este problema, es posible considerar un procedimiento de *trimming* o recorte de pesos (Valliant, Dever, y Kreuter, 2013, sec. 14.4), que puede ser resumido de la siguiente manera:

- Recortar cualquier peso mayor a un umbral establecido U.
- Cualquier peso con magnitud superior a U se trunca de la siguiente manera
  - $w_k^* = \begin{cases} U, \text{si } w_k \geq U \\ wk, \text{en otro caso} \end{cases}$
- Determinar la cantidad neta perdida debido al recorte de pesos extremos
  - $K = \sum_s (w_k - w_k^*)$
- Distribuir K equitativamente entre las unidades que no fueron recortadas.
- Iterar hasta que todos los nuevos pesos calibrados estén por debajo de U.

Al final del proceso se debe asegurar que los datos extremos en los factores de expansión han sido correctamente manejados y que la distribución general de los pesos no sufrió cambios estructurales en los subgrupos poblacionales de interés.

3.2 Método de Potter para estimaciones SAE

Uno de los primeros pasos en el ajuste de los modelos de estimación en áreas pequeñas es la revisión de los pesos de muestreo provenientes del diseño de muestreo de la Encuesta Casen, así como de los procesos de cobertura, elegibilidad y respuesta que la encuesta tuvo en el periodo de la recolección de información. Es posible que los pesos básicos, inducidos por el diseño de muestreo original, se hayan tenido que ajustar y que este proceso haya dado como resultado pesos más extremos e influyentes.

Por supuesto, un conjunto de pesos más disperso tendrá algunas repercusiones en la varianza de los estimadores, especialmente cuando la encuesta resulta ser auto-ponderada (factores de expansión idénticos a nivel de estrato). Para realizar el proceso de estimación en áreas pequeñas se requiere realizar una suavización de los factores de expansión que evite los efectos que pueden tener valores extremos sobre la estimación de varianza de la tasa de pobreza.

En particular, se ha elegido la metodología de Potter (1993), y el algoritmo NAEP, que consiste en recortar o suavizar los factores de expansión, utilizando el error cuadrático medio como medida de evaluación de la estrategia. Esta metodología puede ser considerada superior a otras metodologías de suavización (Valliant, Dever, y Kreuter 2018, página 411) puesto que no solo recorta los factores de expansión que pueden llegar a ser influyentes al momento de estimar las varianzas, sino que además asegura una minimización del sesgo y de la varianza en esta suavización.

La metodología de Potter consiste en elegir los factores de expansión recortados óptimos para que minimicen el Error Cuadrático Medio (ECM) de la variable de interés, en este caso la tasa de pobreza. Para

esta metodología se requiere en primer lugar, definir los grupos en los cuales se hará el procedimiento de suavización. En el caso de las estimaciones de pobreza realizadas por el Ministerio, se eligieron grupos conformados por el cruce de región y zona (urbano, rural), totalizando 30 grupos para los ejercicios SAE 2009 a 2015 y 32 grupos para el año 2017[10]. Para cada uno de estos grupos se define un umbral de recorte óptimo K_g:

$$K_g = \sqrt{C_g \frac{\sum w_{jg}^2}{n_g}}$$

Donde:
w_{jg}: factor de expansión comunal original asociado a cada individuo *j* en su grupo *g*.
n_g : tamaño de la muestra asociado a cada grupo *g*.
C_g : constante de optimización en cada grupo *g*.

El valor de C_g se determina de manera tal que los valores K_g minimicen el error cuadrático medio de la estimación de pobreza de su grupo. En la práctica se implementó un algoritmo de optimización no lineal para encontrar el valor de cada C_g; g = 1, 2, …, 40. Para cada dominio g y para cada umbral de truncamiento fueron calculados los errores cuadráticos medios de las estimaciones considerando estos nuevos ponderadores. Por consiguiente, definiendo a $\hat{\theta}_d^i$ como el estimador con los nuevos pesos y $\hat{\theta}_d^p$, como el estimador original, estas medidas de error cuadrático medio están dadas por la siguiente expresión

$$ECM(\hat{\theta}_d^i) = Var_p(\hat{\theta}_d^i) + (\hat{\theta}_d^i - \hat{\theta}_d^p)^2$$

A continuación, para cada grupo y para cada uno de los 40 valores posibles de C, se calculan los factores óptimos K_g y el ECM asociado, permitiendo elegir para cada grupo el valor de C_g y K_g que minimiza el ECM de la tasa de pobreza. El algoritmo utilizado para encontrar el valor óptimo es una derivación de la subrutina BOBYQA[11], el cual realiza una optimización sin derivadas, utilizando una aproximación cuadrática construida iterativamente para la función objetivo. La interfaz de R NLopt-BOBYQA admite tamaños de pasos iniciales desiguales en los diferentes parámetros (mediante el simple recurso de re-escalar internamente los parámetros proporcionalmente a los pasos iniciales), lo cual es importante cuando diferentes parámetros tienen escalas muy diferentes. Finalmente, una vez determinado K_g para cada grupo, los factores de expansión son truncados siguiendo la siguiente expresión:

$$w^T{}_{jg} = \begin{cases} w_{jg} & si \quad w_{jg} \le K_g \\ K_g & si \quad w_{jg} > K_g \end{cases}$$

10 En el año 2017 se creó la XVI región de Ñuble, totalizando 16 regiones en el país.
11 M. J. D. Powell, "The BOBYQA algorithm for bound constrained optimization without derivatives," Department of Applied Mathematics and Theoretical Physics, Cambridge England, technical report NA2009/06 (2009).

Donde $w^T{}_{jg}$: corresponde a los factores de expansión óptimos recortados para cada individuo *j* en cada grupo *g*. Dado que el recorte de los factores de expansión perjudica la calibración original del diseño de Casen, un último procedimiento que se realiza a estos factores es la calibración al total de la población que reside en viviendas particulares en las comunas en la muestra Casen. Efectivamente, el truncamiento de los valores máximos de los factores de expansión originales implica que la suma de los factores de expansión recortados (w_{jg}^T) es menor al total de la población de inferencia original, por lo que es necesario realizar un ajuste de manera de poder reproducir los totales de la población de interés. El factor de expansión recortado y calibrado w_{jg}^s viene dado por la expresión:

$$w_{jg}^s = w_{jg}^T * ajuste_g$$

$$ajuste_g = \frac{\sum w_{jg}}{\sum w_{jg}^T}$$

Donde:
$\sum w_{jg}$: sumatoria de los factores de expansión originales en cada grupo *g*.
$\sum w_{jg}^T$: sumatoria de los factores de expansión recortados en cada grupo *g*.

3.3 Exploración de los factores consolidados

En general, el comportamiento del algoritmo de Potter provee un buen punto de partida para la modelación de las estimaciones utilizando la metodología de Fay-Herriot. La tabla 1 muestra la estadística descriptiva y la distribución de los factores de expansión comunales, observándose que efectivamente existen pesos extremos todos los años, con una mayor dispersión en los años 2011 y 2013.

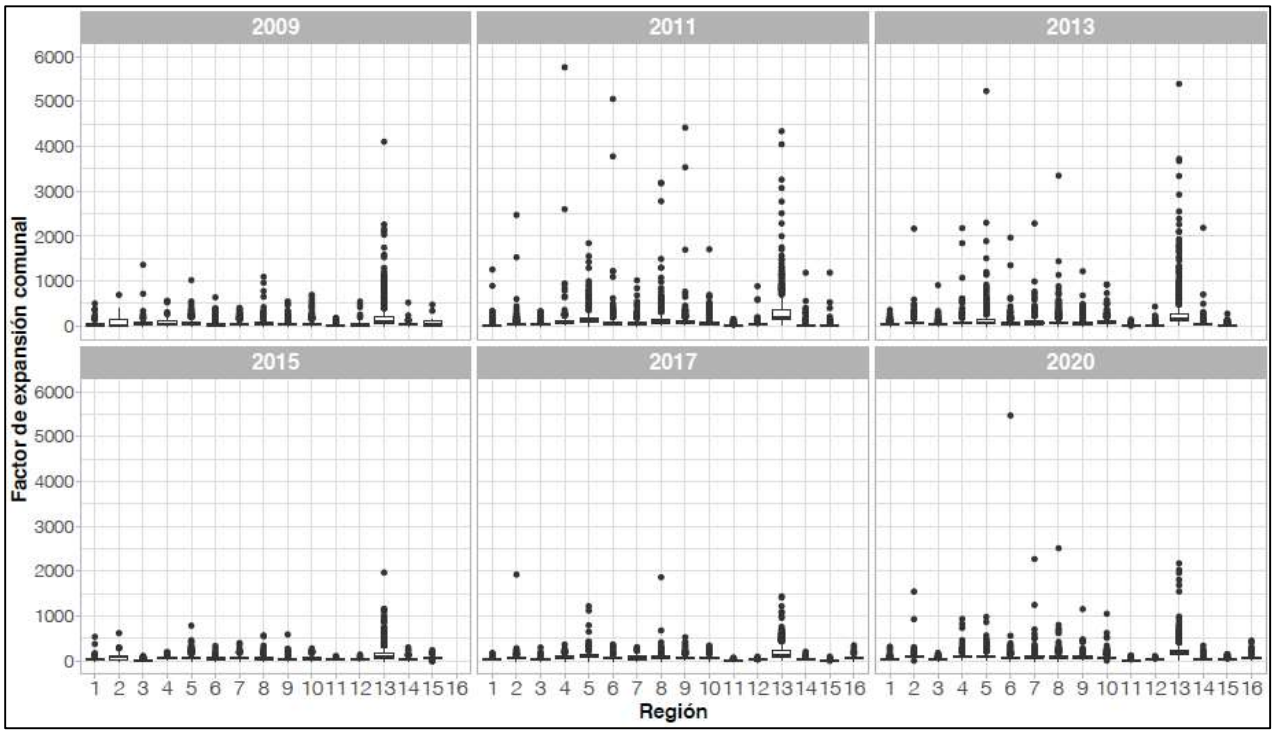
Tabla 1: Estadística descriptiva de los factores de expansión comunal originales. Casen 2009 – Casen en Pandemia 2020

Año	Promedio	Desv. estand	Mínimo	Máximo	Mediana
2009	85,80	161,09	1	4103	36
2011	106,40	296,89	2	15765	46
2013	102,88	203,87	2	6122	58
2015	71,03	78,47	1	1969	50
2017	85,88	91,42	2	1924	63
2020	102,93	184,89	3	15887	78

Fuente: Ministerio de Desarrollo Social y Familia - CEPAL. Casen años correspondientes.
Nota: Se toma una observación por estrato.

Lo anterior se corrobora en la Figura 1 que muestra la distribución de los factores de expansión comunal, desagregados a nivel regional para los años 2009 a 2020.

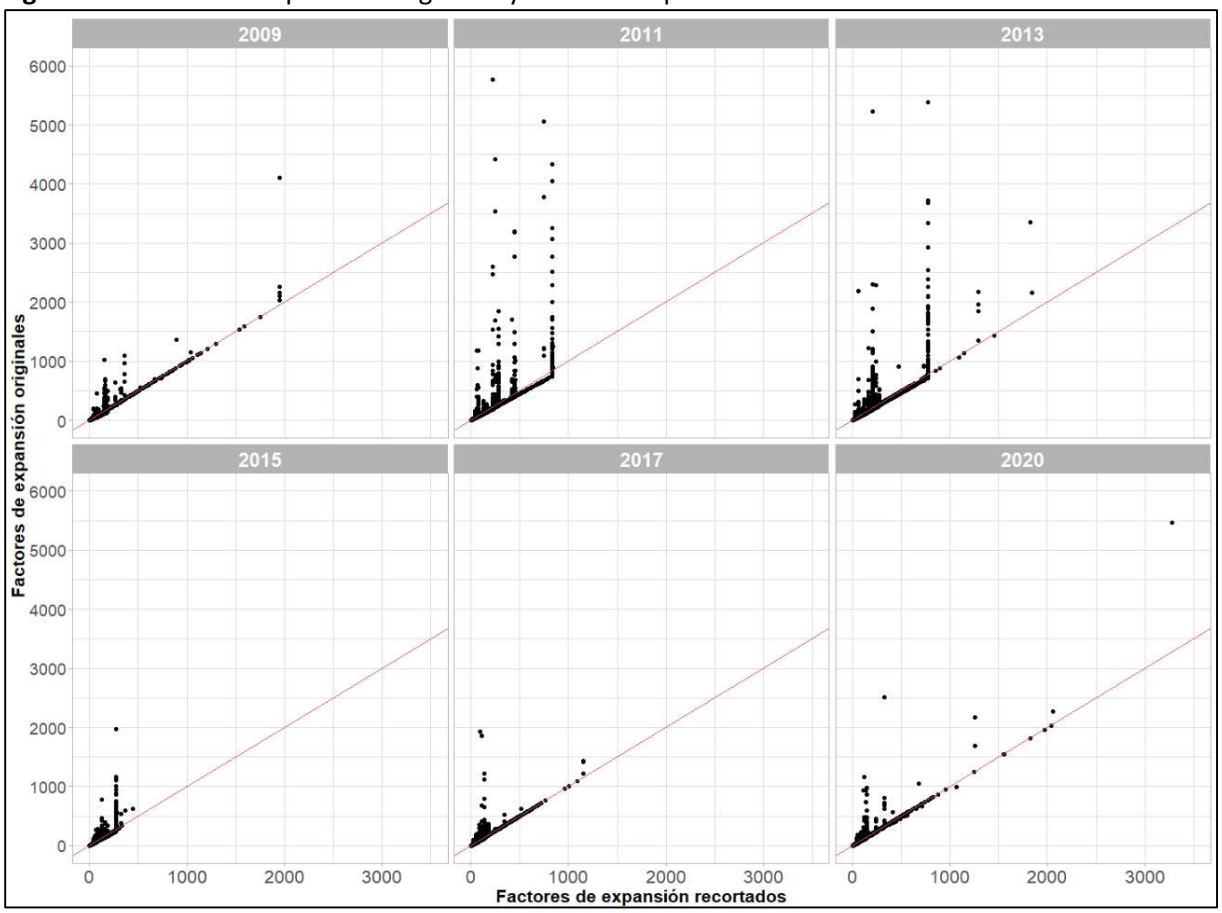
Figura 1: Distribución de factores de expansión comunal a nivel regional. Casen 2009 – Casen en Pandemia 2020



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL.
Nota: Se eliminan factores por sobre 10 mil, para mejor visualización del gráfico.

La Figura 2 contrasta los factores de expansión comunal originales y recortados, luego de aplicar la metodología de Potter. Es posible observar una disminución en los pesos extremos, manteniendo una posición en torno a la línea de 45 grados.

Figura 2: Factores de expansión originales y recortados por año. Casen 2009 – Casen en Pandemia 2020



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL.

La tabla 2 muestra en el panel a) las estadísticas descriptivas de los factores de expansión comunal originales y los pesos recortados, y en el panel b), la estimación de la tasa de pobreza y el efecto diseño. Luego de aplicar la metodología de Potter se puede apreciar que existe una menor dispersión de los pesos, conservando las medidas de posición. A nivel nacional se conserva la estimación puntual y se nota un descenso en la variabilidad. A partir de esta información se concluye que el recorte de los factores de expansión no introduce cambios en las estimaciones puntuales (que son insesgadas por diseño), y que la estimación de su varianza es menor en magnitud al proceso tradicional.

Tabla 2: Estadística descriptiva. Metodología de Potter. Casen 2009 – Casen en Pandemia 2020

a) Factores de expansión originales y recortados por año

	Factores originales					Factores recortados				
Año	**Promedio**	**DS**	**Mínimo**	**Máximo**	**Mediana**	**Promedio**	**DS**	**Mínimo**	**Máximo**	**Mediana**
2009	85,80	161,09	1	4103	36	83,52	140,30	1,0	1946,22	39,20
2011	106,40	296,89	2	15765	46	95,12	124,43	2,3	845,35	49,90
2013	102,88	203,87	2	6122	58	95,22	117,15	2,0	1840,28	58,55
2015	71,03	78,47	1	1969	50	69,81	59,72	1,0	445,92	53,33
2017	85,88	91,42	2	1924	63	84,49	81,79	2,0	1153,28	66,57
2020	102,93	184,89	3	15887	78	102,61	125,60	3,0	8883,75	82,92

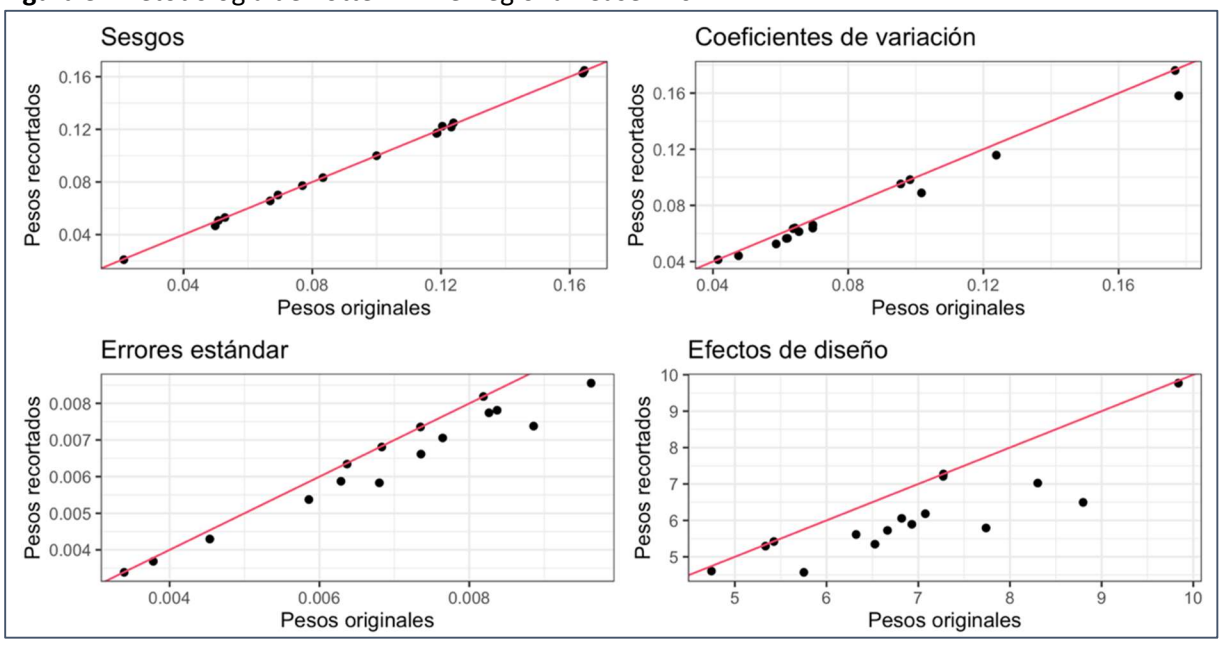
b) Tasas de pobreza y efecto diseño

	Factores originales			Factores recortados		
Año	**Pobreza**	**DS**	**DEFF**	**Pobreza**	**DS**	**DEFF**
2009	15,12%	0,32%	20,08	15,16%	0,30%	17,60
2011	22,32%	0,43%	21,49	22,35%	0,37%	16,04
2013	14,36%	0,30%	16,18	14,52%	0,25%	11,54
2015	11,57%	0,26%	17,28	11,60%	0,19%	9,88
2017	8,49%	0,19%	9,75	8,49%	0,18%	9,13
2020	10,91%	0,21%	8,60	10,94%	0,20%	7,46

Fuente: Ministerio de Desarrollo Social y Familia - CEPAL. Casen años correspondientes.
Nota: Se toma una observación por estrato. Estimaciones se realizan utilizando factor de expansión comunal. Estas no son exactamente iguales a las estimaciones oficiales a nivel nacional porque estas consideran el factor de expansión regional.

Esta misma verificación se realizó a nivel regional para todas las rondas de Casen, obteniendo igualmente resultados satisfactorios. La Figura 3 muestra el caso particular de la ronda 2017. Se puede verificar que las estimaciones puntuales utilizando los pesos originales y los recortados siguen mayormente el patrón de la línea de 45 grados (arriba-izquierda), indicando que en general las propiedades de insesgamiento se mantienen para las regiones. Además, los coeficientes de variación regionales (arriba-derecha), los errores estándar regionales (abajo-izquierda) y sus correspondientes efectos de diseño (abajo-derecha) muestran una disminución notable cuando se usan los pesos recortados, pues los puntos descansan debajo de la línea de 45 grados.

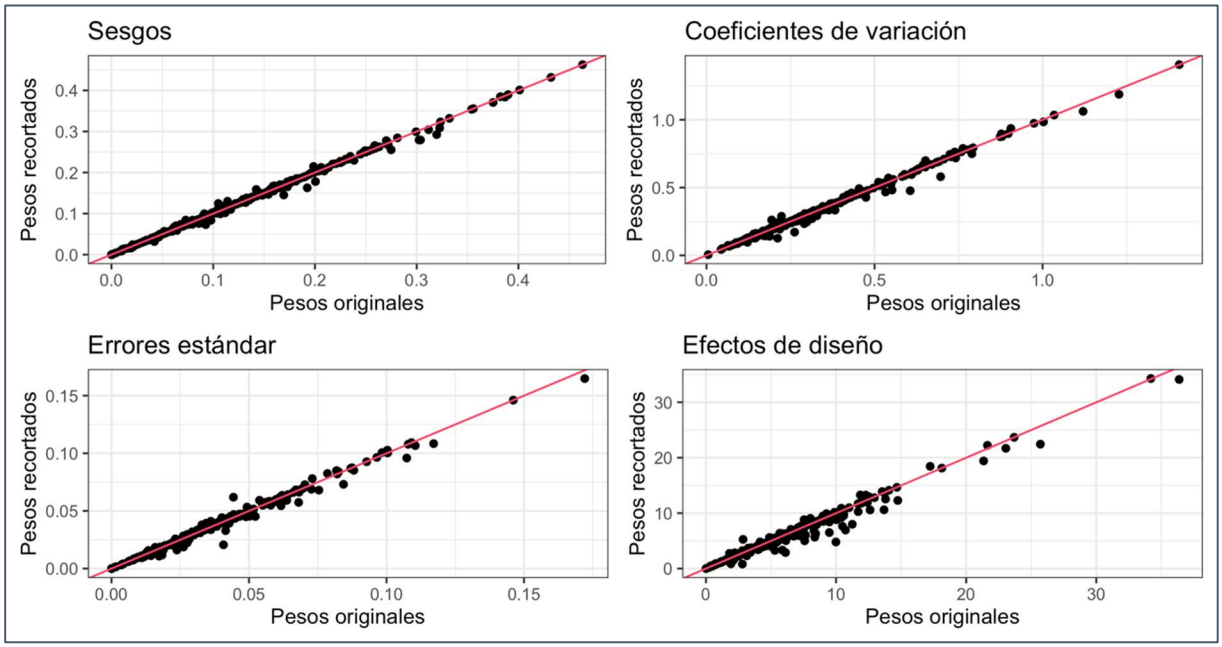
Figura 3: Metodología de Potter – nivel regional. Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL. Casen 2017.

A nivel de comuna, se observa el mismo comportamiento en todas las rondas de Casen. La siguiente figura muestra el caso particular para la ronda 2017, con las mismas conclusiones que las encontradas a nivel regional.

Figura 4: Metodología de Potter – nivel comunal Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL. Casen años correspondientes.

En general, el análisis gráfico muestra una distribución equivalente para los factores de expansión recortados con el algoritmo de Potter. Al observar el comportamiento de las estimaciones de pobreza para ambos métodos, es posible concluir que la parte del sesgo en el error cuadrático medio no aporta significativamente comparada con la varianza. La razón de esta afirmación es evidente al contemplar las ligeras desviaciones con respecto a la línea de 45 grados de las estimaciones óptimas con respecto a las estimaciones originales.

4. Criterios de Inclusión de comunas al modelo SAE

Según lo descrito en el capítulo 2, el estimador de Fay-Herriot calculado por el Ministerio se puede expresar como un promedio ponderado entre el estimador directo, proveniente de Casen, y un estimador sintético, que utiliza la relación de la pobreza estimada y algunas covariables auxiliares provenientes de registros administrativos y del Censo de Población y Vivienda. Además, el ponderador en cada comuna variará con respecto a la precisión de las estimaciones directas. Así, entre menor sea la varianza de la estimación directa, mayor será su ponderación en la estimación final, y viceversa. En la mayoría de los casos, a medida que el tamaño muestral aumenta, la estimación SAE tiende al estimador directo (Molina 2009).

En efecto, y considerando la forma particular del diseño de muestreo de Casen, se observa que las comunas con mayor tamaño de muestra tienden a presentar una mayor precisión en los estimadores directos (menores estimaciones de varianza), y por ende la estimación directa predomina en la ponderación del estimador SAE. Por la naturaleza de los estimadores directos, a partir de las encuestas no se puede conocer la varianza de la estimación, y por tanto este parámetro debe ser estimado. No obstante, cuando el tamaño de muestra de la comuna es muy pequeño, esta estimación resulta ser inadecuada, puesto que pierde sus propiedades estadísticas habituales como el insesgamiento, la precisión y la consistencia. Esto plantea un desafío en la ejecución de las estimaciones SAE, puesto que la ponderación sobre la estimación directa recae justamente en su precisión estimada. Es por esto que en el planteamiento de todas las aplicaciones de SAE realizadas hasta la fecha para Chile, la estimación sintética se ha realizado únicamente con las comunas que según las proyecciones del INE tienen más de 10 mil habitantes, como una aproximación de un mayor tamaño de muestra.

Puesto que el tamaño de muestra no es el único parámetro que influye en la calidad de las estimaciones directas, este capítulo tiene como objetivo profundizar en criterios adicionales que permitan tener una evaluación de la estimación directa de las comunas, con el objetivo de definir estándares necesarios para contar con un nivel de calidad aceptable en el muestreo y en la recolección de datos de la comuna, para una correcta aplicación de la metodología de estimación de áreas pequeñas. Los criterios aquí propuestos se relacionan con aspectos asociados a la calidad del muestreo y a la confiabilidad de la estimación de la varianza de la estimación puntual de la tasa de pobreza en cada comuna.

A partir de los criterios de inclusión especificados en este capítulo, se asignará la estimación de Fay-Herriot, que pondera las estimaciones directa y sintética para todas las comunas que cumplen con los criterios de inclusión. Mientras que a las comunas que no cumplen con los criterios de inclusión y para las comunas sin presencia en Casen, se asignará la estimación sintética[12].

12 Hasta el año 2015 las comunas sin presencia en Casen llevaban una estimación de tasa de pobreza que correspondía a una imputación de medias por conglomerado. En Casen 2017 se les asignó la estimación sintética.

4.1 Evaluación de la estimación directa

Como se menciona anteriormente, hasta el año 2017 se utilizó el criterio del tamaño poblacional para incluir a las comunas en la encuesta Casen en la derivación de los parámetros de interés asociados a la estimación sintética. La tabla 3 muestra que para el año 2017 esto implicó incluir 240 comunas en las estimaciones de los parámetros $\hat{\beta}$ y $\widehat{\sigma_u^2}$. No obstante, lo anterior, una vez derivados los coeficientes $\hat{\beta}$, estos eran utilizados para obtener estimaciones sintéticas para todas las comunas en la muestra Casen.

Tabla 3: Comunas incluidas y excluidas en la estimación sintética

	2009	2011	2013	2015	2017	2020
Comunas incluidas	235	241	241	185	240	258
Comunas excluidas	83	75	73	139	78	66

Fuente: Ministerio de Desarrollo Social y Familia - CEPAL. Casen años correspondientes.
Nota: El año 2015 se excluyeron de la estimación sintética 139 comunas para las cuales se aumentó el tamaño de muestra con el objetivo de mejorar la precisión comunal y, por lo tanto, se consideró que las estimaciones puntuales de estas comunas eran precisas, no pasando ellas a una estimación de SAE. El modelo sintético incluyó las 185 comunas restantes para las cuales no se aumentó el tamaño muestral por sobre lo estimado con un diseño de precisión regional y nacional.

De la misma forma, hasta el año 2017 se utilizaron todas las comunas con muestra Casen para la derivación del parámetro $\widehat{\psi}_d$, el cual, en conjunto con $\widehat{\sigma_u^2}$, son utilizados para el cálculo de $\hat{\gamma}_d$ que influye directamente en el peso otorgado a las estimaciones directa y sintética. Esta inclusión implica utilizar estimaciones de varianza a nivel comunal que, por diseño, no son precisas y que es el argumento central para el desarrollo de estimaciones SAE a ese nivel de desagregación.

La inclusión de comunas cuya precisión no alcanza niveles mínimos aceptables, puede tener repercusiones no deseadas en la metodología de estimación de áreas pequeñas, afectando la calidad de los insumos que entran en el modelo, y, por ende, la estimación de los parámetros relevantes. Por otro lado, la exclusión de las comunas según el tamaño poblacional refleja solo en cierto grado cómo el tamaño muestral de las comunas determina la precisión de la estimación, siendo más informativo evaluar indicadores asociados a los resultados del trabajo de campo en cada comuna, observando el tamaño muestral alcanzado, los grados de libertad asociados a la estimación, así como criterios que permitan inferir la calidad de la estimación de la varianza del estimador directo a nivel comunal (UN 2021)[13]. En la siguiente sección se describen los indicadores propuestos para evaluar la calidad de las estimaciones directas, mientras que en la sección 4.3 se define la regla de decisión finalmente aplicada.

4.2 Indicadores de calidad propuestos

Los indicadores que se van a revisar provienen de la literatura existente enfocada en medir la precisión de las estimaciones directas con el objetivo de publicación. El documento "Criterios de calidad en la estimación de indicadores usando encuestas de hogares. Una aplicación a la migración internacional"[14] (Gutierrez *et al.* 2020) resume algunas consideraciones en esta materia que han sido utilizados ampliamente en diversas encuestas y estudios. A continuación, se describen los indicadores analizados

13 Toolkit for using Small Area Estimation for the SDGs, United Nations, 2021.
14 Serie de Estudios Estadísticos N 101, junio 2020. https://repositorio.CEPAL.org/handle/11362/45681

para su eventual incorporación a la evaluación de la calidad estadística de las estimaciones que serán insumo del modelo SAE. Cabe destacar que los umbrales propuestos en la literatura para la publicación de cifras oficiales son mucho más exigentes que los que requiere un modelo de estimación de áreas pequeñas, por lo que deben ser ajustados a estas necesidades de información.

- **Grados de libertad (gl):** en general, los grados de libertad corresponden a la resta entre el número de unidades primarias de muestreo (UPM) y los estratos[15] en determinada subpoblación o área que se especifique (comunas, sexo, etc.). Es una medida de cuántas unidades de información independientes se tienen en la inferencia y, por lo tanto, impactan directamente en la amplitud del intervalo de confianza (IC), que a su vez depende del percentil asociado a una distribución t-student con determinados grados de libertad. Por ejemplo, el valor t con *gl*=1 es 12,7; con *gl*=5 es 2,5; con *gl*=40 es 2,02. En el caso de indicadores estimados a nivel nacional o regional, y por ende con grados de libertad muy altos, esta distribución converge a la distribución normal estándar, obteniéndose z=1,96.

- **Coeficiente de variación (CV):** el coeficiente de variación corresponde a la razón del error estándar y la estimación puntual. De esta forma, la magnitud de la dispersión se relaciona con la estimación puntual y permite una aproximación al error relativo de muestreo:

$$CV\left(\hat{\theta}_d^{Dir}\right) = \frac{\sqrt{\widehat{Var}\left(\hat{\theta}_d^{Dir}\right)}}{\hat{\theta}_d^{Dir}}$$

- **Coeficiente de variación logarítmico (CVL)[16]:** en el caso de proporciones el coeficiente de variación puede no funcionar de forma adecuada, ya que no es simétrico. Por ejemplo: si se está estimando una proporción P y esta estimación está cercana a cero, el CV será muy grande solo por efecto de la estimación puntual. Por el contrario, el CV del complemento $(1 - P)$ será muy pequeño. De esta forma, a pesar de que se está midiendo el mismo fenómeno, los CV son muy distintos. En general, CV y CVL entregan valores similares en rangos entre 0,2 y 0,8.

- **Efecto de diseño (DEFF):** el efecto diseño consiste en la razón entre la varianza de la estimación bajo el diseño complejo y la varianza bajo muestreo aleatorio simple de esa misma estimación. Bajo un diseño complejo no hay independencia entre las observaciones en determinadas áreas geográficas. El DEFF viene a dar cuenta de este efecto aglomeración que indica que entre más parecidas son las unidades entrevistadas dentro de la unidad primaria de muestreo, menos información aportan y por lo tanto menos precisa será la estimación. En general, el efecto de diseño será mayor cuando la distribución de la variable de interés esté más correlacionada con la distribución de las UPM en el marco de muestreo (correlación intraclase, ρ_y). Por ejemplo, en el

15 Para evaluar las estimaciones comunales se utilizan los conglomerados (UPM) y los estratos originalmente considerados en el diseño muestral de Casen y que se encuentra en su marco de selección. Mientras que las estimaciones para los dominios de representación de Casen utilizan pseudo-conglomerados y pseudo-estratos que resultan de la conglomeración de estas unidades en el marco. En el caso de las estimaciones SAE, se busca utilizar la información asociada directamente al proceso de selección de UPM dentro de los estratos originales del diseño para conocer la dispersión de la muestra en las áreas pequeñas (comunas).

16 $CV\left(\hat{L}\right) = \frac{se(\hat{L})}{\hat{L}} = \frac{CV(\hat{P})}{\hat{L}}$, $con\ \ \hat{L} = -\log\left(\hat{P}\right)$, $si\ \hat{P} \leq 0.5$ y $Var(\hat{L}) \approx \frac{Var(\hat{P})}{P^2}$

caso de la pobreza, esto ocurre si los hogares pobres están geográficamente aglomerados, segregados y separados de los hogares más acaudalados. Dado que esta relación entre la variable de interés y la UPM es más bien de carácter estructural y su variación usualmente solo se da en el largo plazo, en general el DEFF estimado se ve afectado por el número promedio de viviendas entrevistadas en cada UPM ($\overline{m}$), donde a mayor ($\overline{m}$) mayor será también el DEFF.

$$DEFF = \frac{Var\left(\hat{\theta}_d^{Dir}\right)}{Var_{mas}\left(\hat{\theta}_d^{Dir}\right)} = 1 + (\overline{m} - 1)\rho_y$$

- **Tamaño de muestra (n):** el tamaño de muestra afecta el error estándar y, por lo tanto, la amplitud del intervalo de confianza. El error estándar generalmente decrece a medida que el tamaño de muestra se hace más grande.

- **Tamaño de muestra efectivo (n_{eff}):** como ya se mencionó, en un muestreo complejo las observaciones no son independientes, lo que implica que la correlación que existe entre las UPM y la variable de interés afecta la varianza; por lo que para obtener determinados niveles de precisión es necesario aumentar la muestra de personas. Como contraparte, el tamaño efectivo corresponde al cociente entre el tamaño definido en el diseño complejo y el DEFF, y es una aproximación a la información efectiva que aporta la muestra lograda[17].

$$n_{eff} = \frac{n}{DEFF}$$

- **Casos no ponderados (y):** corresponde a los casos en la muestra en donde se observa la variable de interés; en el caso de Casen se refiere al número de personas en situación de pobreza. Un número de casos bajo se reflejará en CV muy altos y en IC muy amplios. Notar que este criterio se aplica en conjunto con otros criterios de precisión para diferenciar las situaciones en que una baja observación de la variable de interés se debe a una baja prevalencia en la comuna de aquellos donde la baja prevalencia de la variable de interés se explica por problemas asociados a la precisión de la estimación.

La tabla 4 a continuación resume los umbrales que señala la literatura para cada uno de estos indicadores con el objetivo de evaluar la publicación de cifras de una encuesta.

17 Si la variable de interés es muy homogénea en una misma UPM, entonces la información que aportan los hogares en esa UPM no será muy distinta. En ese sentido, la cantidad de individuos que están aportando a la inferencia del indicador no es el número de personas, ni el número de hogares en la muestra; sino que el tamaño de muestra efectivo que corresponde al tamaño muestral deflactado por los efectos de aglomeración.

Tabla 4: Valores de referencia para indicadores de calidad de estimación

Indicador	Valor	Referencia
Intervalos de confianza (IC)	-	-
Coeficiente de Variación CV	20%	Gutiérrez et.al. (2018)
Coeficiente de Variación Logarítmico (CVL)	17,5%	Barnett-Walker et al. (2003)
Efecto diseño (DEFF)	-	-
Tamaño muestra (n)	100	Barnett-Walker et al. (2003)
Tamaño muestra efectiva (neff)	68	Barnett-Walker et al. (2003)
Grados de libertad (gl)	8	Parker, Talih, y Malec (2017)
Casos no ponderados	50	National Research Council (2015)

Fuente: Criterios de calidad en la estimación de indicadores usando encuestas de hogares. Una aplicación a la migración internacional (Serie estudios estadísticos N101. CEPAL, junio 2020).

4.3 Reglas de decisión para la inclusión de comunas

En el trabajo conjunto con CEPAL, y siguiendo las recomendaciones de Eurostat (2019)[18], se decidió partir con los criterios de calidad estándar que se usan para evaluar la publicación de resultados. Dado que el ejercicio de inclusión de comunas no tiene un objetivo de publicación, los umbrales fueron flexibilizados para ser utilizados en los ejercicios SAE de Chile. Es importante destacar que para el ejercicio SAE, el objetivo no es incluir comunas precisas, ya que el mismo modelo Fay-Herriot incluye una ponderación por precisión, sino más bien, el objetivo es incluir comunas con un muestreo de calidad que indique que los insumos que entren al modelo (precisos o no precisos) son confiables.

Una vez analizados todos los criterios que se utilizan en la literatura se decidió priorizar los siguientes:

- Indicadores asociados al muestreo y logro de la encuesta, ya que evalúan la calidad de la muestra mediante la cual se hará inferencia.
- Indicadores asociados a la precisión de la estimación de varianza que entreguen alertas respecto a la validez de esta estimación.

Criterio 1: Grados de libertad

Asumiendo una alta homogeneidad dentro de cada UPM, la información efectiva que contribuye a las estimaciones puntuales viene dada por el número de UPM en cada comuna más que por la cantidad de viviendas entrevistadas. Los grados de libertad corresponden a la diferencia entre el número de UPM y los estratos, representando la información independiente que se tiene en la inferencia. Este indicador impacta de manera directa en la amplitud del intervalo de confianza.

Los datos presentados en la tabla 5, muestran una media de 10 grados de libertad en 2009, cuando el diseño Casen se basaba en un marco de secciones solamente; mientras que para los años 2011 a 2020 los grados de libertad fluctúan entre 19 y 31 con una gran variación entre mínimos y máximos cada año, lo que refleja el cambio de marco desde uno de secciones a uno de manzanas.

Para este indicador se definen 2 umbrales. En primer lugar, aquellas comunas con 14 o más grados de libertad se incluyen automáticamente en las estimaciones SAE, porque se considera que este valor indica

18 "Guidelines on small area estimation for city statistics and other functional geographies".

que hay suficientes unidades que aportan información independiente sobre el fenómeno de la pobreza. Para la aplicación SAE, se ha fijado el valor en 14 grados de libertad para la inclusión de la comuna de forma automática, lo que implica que el resto de las comunas pasarán por los siguientes criterios que evalúan mediante un conjunto de otros indicadores la calidad del muestreo en la comuna[19].

En segundo lugar, se utilizará como criterio de exclusión automática de las comunas con 2 o menos grados de libertad, ya que esto impactaría en la información independiente de la cual se dispone para realizar las estimaciones de varianza (ver figura 1).

Tabla 5: Grados de libertad

GRADOS DE LIBERTAD (*)						
Año	Min	Media	Mediana	Max	comunas con gl>=14	comunas con gl<=2
2009	0	10,5	9	22	77	8
2011	0	19,8	9	377	107	70
2013	1	24,6	11	372	134	11
2015	0	31,1	11	497	153	59
2017	1	27,3	11	390	127	14
2020	2	31,7	11	525	141	14

Fuente: Ministerio de Desarrollo Social y Familia- CEPAL. Casen años correspondientes.
(*) Se utilizan las UPM y estratos provenientes del marco de selección de la encuesta y no las variables agrupadas varstrat y varunit utilizadas para las estimaciones en los dominios de representación de la encuesta.

Criterio 2: Muestra lograda

Uno de los criterios de exclusión se refiere al tamaño de muestra logrado en la comuna. A medida que el tamaño muestral aumenta, se espera que los estimadores converjan a su verdadero valor y que el error estándar disminuya junto con los intervalos de confianza, haciendo las estimaciones más precisas. Para este criterio se fija un umbral de 50 observaciones en la comuna. Cabe destacar que las observaciones corresponden a personas en la muestra, porque los indicadores de interés que se busca estimar, pobreza multidimensional y pobreza por ingresos, se encuentran a este nivel.

Este criterio de exclusión se evalúa en conjunto con los otros indicadores de exclusión, sobre las comunas que no pasaron el criterio de inclusión por grados de libertad (ver figura 5).

Criterio 3: Efecto de Diseño

Este indicador se puede interpretar como el número de unidades muestrales independientes (como si fuera muestreo aleatorio simple) que hay en las viviendas seleccionadas. Por ejemplo, una muestra que

19 A modo de referencia, el INE plantea una serie de criterios y umbrales para evaluar la precisión de las estimaciones provenientes de encuestas de hogares, dentro de los cuales sugiere 9 grados de libertad, sin embargo, este límite también se analiza en conjunto con otros indicadores de precisión tales como el CV o el error estándar. Documento de Trabajo: Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares. Instituto Nacional de Estadísticas. 28 de febrero de 2020.

bajo un diseño complejo contemple la selección de 200 viviendas y que tuvo un efecto diseño de 20 implicaría que esas 200 viviendas en realidad están aportando información efectiva de 10 viviendas.

El criterio de exclusión en base al efecto de diseño fija el umbral en valores menores a 1. Debido a que al tener Casen un diseño complejo, efectos de diseño menores que uno implicarían que la varianza del diseño complejo es más precisa que aquella obtenida bajo un diseño aleatorio, lo cual es contraintuitivo en el caso de Casen, debido a que el efecto de la conglomeración sobre la estimación de la varianza es mayor al efecto de la estratificación. La fórmula del DEFF en la sección 3.1 muestra que el DEFF nunca podría ser menor que uno dado que el coeficiente de correlación intra-clase es, en general, positivo y m es mayor que uno. Esta evidencia, en conjunto con el hecho de que las comunas tienen menos de 14 grados de libertad, indica que la estimación del DEFF es poco confiable, y por ende, inadecuadas para ser incluidas en la estimación SAE.

Tabla 6: Efecto diseño

EFECTO DISEÑO				
Año	Min	Media	Max	comunas con deff<1
2009	0,15	7,76	36,8	5
2011	0,00	7,51	91,73	58
2013	0,02	6,02	50,0	35
2015	0,00	5,54	62,17	59
2017	0,00	4,69	43,34	55
2020	0,03	4,48	28,54	35

Fuente: Ministerio de Desarrollo Social y Familia - CEPAL. Casen años correspondientes.
(*) Se excluyen del cálculo los efectos de diseño indefinidos.

Criterio 4: Número observado en la muestra de personas en situación de pobreza

Otro criterio de exclusión se refiere a un bajo número de casos en la variable de interés, en este caso, personas en situación de pobreza. Este indicador incide directamente en mayores coeficientes de variación, mayor amplitud del IC y mayor imprecisión de la estimación. Adicionalmente, valores muy bajos en áreas pequeñas dificultan saber si se debe a que en realidad hay pocas personas en situación de pobreza o se debe a un efecto puntual del diseño de ese año en particular, en donde, por azar en las UPM seleccionadas se encontraron pocos hogares pobres.

Se define como umbral de exclusión comunas con menos de 15 personas en situación de pobreza por ingresos. Este criterio se evalúa de forma secuencial después del criterio de inclusión por grados de libertad y en conjunto con los otros criterios de exclusión.

Tabla 7: Número de personas en pobreza

NUMERO PERSONAS EN POBREZA (y)				
Año	Min	Media	Max	comunas con y<15
2009*	0	123,5	402	9
2011	0	151,13	1709	11
2013	2	113,8	1228	16
2015	0	113,1	1080	21
2017	0	64,4	651	41
2020	0	65,8	790	46

Fuente: Ministerio de Desarrollo Social y Familia - CEPAL. Casen años correspondientes.
(*): Los datos del año 2009 corresponden a la metodología antigua de pobreza por ingresos.

Los indicadores seleccionados, sus umbrales y las reglas de decisión se resumen en la figura 4 a continuación.

Figura 5: Criterios de inclusión de comunas a modelo FH

Fuente: Ministerio de Desarrollo Social y Familia- CEPAL.

La tabla 8 a continuación muestra los resultados de aplicar la regla secuencial de inclusión de comunas y las comunas incluidas/excluidas en cada paso.

Tabla 8: Criterios de inclusión por año

Año	Inclusión grados de libertad	Exclusión tamaño muestral	Exclusión efecto de diseño	Exclusión número de pobres	Exclusión grados de libertad	**Total Incluir**	**Total Excluir**
	gl>=14	n<50	deff<1	y<15	gl<=2		
2009	77	0	5	5	8	322	12
2011	107	1	56	7	70	232	92
2013	134	0	35	11	11	279	45
2015	153	0	51	17	59	234	90
2017	127	0	52	36	14	242	82
2020	141	1	34	43	14	256	68

Fuente: Ministerio de Desarrollo Social y Familia – CEPAL. Casen años correspondientes

Nota: Se debe notar que las condiciones deff < 1, y < 15 y gl <=2 se aplican en un mismo paso del proceso de inclusión. En este sentido, estimaciones que se excluyen por el criterio del efecto de diseño también pueden no cumplir el criterio de los grados de libertad o de número de pobres.

5. Estimación de la varianza del estimador directo

Como ya se mencionó anteriormente, el estimador directo no es el único insumo del modelo de áreas de Fay-Herriot; también lo es su varianza. El estimador puntual da un indicio de la localización del parámetro, y su varianza presenta el nivel de certeza o confianza sobre esta localización. De esta forma, una comuna con un estimador directo cuya varianza es muy grande, será una comuna para la cual los intervalos de confianza serán mucho más amplios que una comuna cuya estimación directa induzca una varianza más pequeña.

No hay que dejar de lado que, al tratar con cifras provenientes de procesamientos con encuestas de hogares, es indispensable siempre tener en cuenta que el sustento inferencial recae en la estrategia de muestreo, definida como la dupla compuesta por el diseño de muestreo y el estimador escogido. En particular, teniendo en cuenta los ajustes al factor de expansión mencionados en el capítulo 3. Denotando como y_k al resultado de la observación de la variable de interés (pobreza) en el individuo k, la estimación de la proporción de personas en condición de pobreza a nivel nacional está dada por la siguiente expresión:

$$\hat{\theta}^{Dir} = \frac{\sum_h \sum_i \sum_{k \in s} w_k \ y_k}{\sum_h \sum_i \sum_{k \in s} w_k} = \frac{\hat{t}_y}{\widehat{N}}$$

Nótese que las tres sumatorias corresponden a los estratos de muestreo ($h = 1 \dots, H$), las unidades primarias de muestreo ($i = 1, \dots, n_{Ih}$) y los individuos dentro de los hogares ($k = 1, \dots, n_i$). Este estimador corresponde a una razón entre dos conteos y, tanto el numerador como el denominador son aleatorios. Por un lado, el numerador estima la cantidad total de personas en condición de pobreza, mientras que el denominador estima la cantidad de personas dentro de la población objetivo de Casen. Nótese además que el ponderador w_k corresponde al peso de muestreo óptimo calibrado. En particular, la estimación para la proporción de personas en condición de pobreza dentro de un subgrupo de interés U_d es:

$$\hat{\theta}_d^{Dir} = \frac{\sum_h \sum_i \sum_{k \in s_d} w_k \ y_k}{\sum_h \sum_i \sum_{k \in s_d} w_k} = \frac{\hat{t}_{y_d}}{\widehat{N}_d}$$

En donde s_d hace referencia a la muestra observada en el dominio. Es importante detenerse un momento en esta expresión y mencionar que, en algunas ocasiones, la cardinalidad de s_d (es decir, el tamaño de muestra efectivo en el subgrupo de interés) es muy pequeña, o incluso nula. Este es un detalle no menor porque existirá una cota en el tamaño de muestra para la cual la distribución del estimador empiece a converger a la distribución teórica (t-student con *gl* grados de libertad). Por ende, si el subgrupo de interés está muy subrepresentado, la estimación directa, junto con su varianza perderán las propiedades estadísticas de la inferencia directa: insesgamiento, precisión y consistencia.

5.1 Aproximación de la varianza para la encuesta Casen

En general, para cualquier estrategia de muestreo, es posible definir expresiones genéricas para el estimador de la varianza de un estimador directo. Por ejemplo, la varianza para un estimador de total $\hat{t}_y$ está dada por la siguiente expresión:

$$Var(\hat{t}_y) = \sum_{k \in U} \sum_{j \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Esta expresión depende de las probabilidades de inclusión de los elementos π_k, de las covarianzas entre las variables de membresía a la muestra aleatoria Δ_{kl}, y de los valores de la característica de interés y_k en todos los individuos de la población; por esta razón el subscrito de las sumatorias $(k \in U, j \in U)$ depende de U. Por supuesto, es imposible conocer este último insumo, y se plantea una afirmación muy importante: no es posible calcular la varianza de un estimador directo en las encuestas de hogares.

Ante estas limitaciones, el camino tradicional indica que, para poder realizar un proceso inferencial robusto, es necesario estimar la varianza del estimador de directo. No existe una solución única, pero un acercamiento habitual es considerar el siguiente estimador insesgado:

$$\widehat{Var}(\hat{t}_y) = \sum_{k \in s} \sum_{j \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

El cual depende de las probabilidades de inclusión de segundo orden π_{kl}, y de los valores de la característica de interés en la muestra. Sin embargo, como cualquier estimador, sus bondades dependerán de que se cumplan algunas condiciones de regularidad, entre las cuales cabe resaltar dos: 1) que la población *U* sea grande, 2) que la muestra *s* sea grande. Sin el cumplimiento de estos requisitos, este estimador tendrá un desempeño muy pobre en cuanto a su precisión. Por supuesto, en términos de la inferencia nacional, regional, y de los dominios de representatividad de la encuesta Casen, estas condiciones se dan por cumplidas.

En particular, el estimador de la varianza para la proporción de personas en condición de pobreza, se torna un poco más complejo. Dado que la forma funcional del estimador $\hat{\theta}^{Dir}$ es no lineal por tratarse de un cociente de variables aleatorias, es necesario recurrir a los principios de linealización de Taylor para encontrar una aproximación lineal del estimador de varianza (Gutiérrez, 2016). En particular, un estimador aproximadamente insesgado para la varianza de $\hat{\theta}^{Dir}$ toma la siguiente forma:

$$\widehat{AVar}(\hat{\theta}^{Dir}) = \sum_{k \in s} \sum_{j \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}$$

En donde la variable linealizada e_k toma está definida por la siguiente expresión:

$$e_k = \frac{1}{\hat{N}}(y_k - \hat{\theta})$$

Debido a las complejidades algebraicas y computacionales, estimar la varianza de estos estimadores, utilizando las expresiones matemáticas exactas en encuestas complejas que contemplan esquemas de selección en varias etapas, estratificación y uso de pesos desiguales, puede tornarse bastante costoso e ineficiente. Por lo anterior, el acercamiento recomendado a la estimación de las varianzas es por medio del uso de aproximaciones. En general, las aproximaciones computacionales, que en su mayoría se basan en la técnica del último conglomerado (Wolter, 2007), generan una salida práctica al problema de la estimación de la varianza. Si bien, estas expresiones no brindan estimaciones de varianza estrictamente insesgadas, sí constituyen un acercamiento bastante preciso, aceptado y adoptado en la mayoría de los procesos de estimación de estadísticas oficiales en la mayoría de las oficinas nacionales de estadística en el mundo.

Con base en lo anteriormente expuesto, la aproximación de la varianza para la proporción de personas pobres a nivel nacional estaría dada por la siguiente expresión (Valliant, Dever y Kreuter 2018, sección 15)

$$\widehat{Var}(\hat{\theta}^{Dir}) = \sum_h \frac{n_h}{n_h - 1} \sum_{i \in s_h} \left(\hat{t}_{hi} - \bar{\hat{t}}_h\right)^2$$

En donde, $\hat{t}_{h_i} = \sum_{k \in s_{hi}} w_k \, e_k$ y $\bar{\hat{t}}_h = \frac{1}{n_h} \sum_{i \in s_h} \hat{t}_{hi}$.

5.2 Estimación de varianza para modelo SAE en Chile

La estimación de la varianza de la tasa de pobreza utilizadas como insumo en las estimaciones SAE reviste dos grandes desafíos que deben ser abordados: (i) en primer lugar, la estimación de la varianza depende de la proporción estimada, generando un problema de endogeneidad que viola el supuesto de varianza constante del modelo de Fay-Herriot. (ii) en segundo lugar, la estimación de varianza en áreas con un tamaño de muestra pequeño será muy imprecisa, incorporando ruido en la estimación del verdadero proceso de varianza.

Ambos temas han sido abordados en la literatura principalmente a través de dos métodos, el uso de la transformación arcoseno (Jiang *et al.* 2001, Hadam *et al.* 2020), y la estimación de una Función Generalizada de Varianza (Hidiroglou *et al.* 2019, Fuquene *et al.* 2019, Rivest and Belmonte 2000).

En las estimaciones de área pequeña realizadas por el Ministerio para los años 2009 a 2015, se aplicó una transformación arcoseno para resolver el problema de endogeneidad, permitiendo, además, contar con una aproximación simple y conocida para la estimación de la varianza a través de los efectos de diseño comunales.

Sin embargo, esta transformación no aborda la volatilidad que pueden tener las estimaciones a nivel de comuna, generando patrones de estimación muy variables que no permiten aproximar de forma confiable la precisión del estimador directo, lo que afecta las estimaciones de área pequeña.

Frente a esta volatilidad comunal observada, en los años 2009 a 2015 se tomó la decisión de aproximar la estimación de los efectos de diseño comunal por su equivalente regional, aportando una mayor estabilidad a esta estimación. Esto asumía que el efecto diseño regional era una buena aproximación a los efectos diseño de las comunas pertenecientes a cada región.

En el año 2017 se realizó el diagnóstico de que la utilización de efectos de diseño regional podría estar sobre o subvalorando la precisión de la estimación directa comunal, además de estar asignando a todas las comunas en una misma región el mismo efecto de diseño regional. Como resultado de este diagnóstico, se estimaron los efectos diseño comunales y regionales, y se utilizó cada uno dependiendo de la muestra lograda en cada comuna[20].

A continuación, se describen la Función Generalizada de Varianza y la transformación arcoseno, como una propuesta para abordar la endogeneidad y la variabilidad en áreas pequeñas. Se presentan los resultados de simulación obtenidos para las estimaciones comunales del año 2009.

5.3 Transformaciones FGV

Como se mencionó anteriormente, uno de los insumos más importantes en el modelo de áreas es la varianza del estimador directo, a nivel de comuna, la cual no puede calcularse de ningún modo. En correspondencia, este valor debe estimarse desde los datos recolectados en cada comuna. Sin embargo, en comunas en las que se cuenta con un tamaño de muestra muy pequeño, estas estimaciones no tendrán un buen comportamiento. Por ende, es muy útil utilizar un modelo de suavizamiento de las varianzas para eliminar el ruido y la volatilidad de estas estimaciones y extraer la verdadera señal del proceso.

Hidiroglou (2019) afirma que $E_{mp}(\hat{\theta}_d^{dir}) = \mathbf{x}_d{}'\boldsymbol{\beta}$ y $V_{mp}(\hat{\theta}_d^{dir}) = \sigma_u^2 + \tilde{\psi}_d^2$, en donde el suscrito mp hace referencia a la inferencia doble que se debe tener en cuenta en este tipo de ajustes y define la medida de probabilidad conjunta entre el modelo y el diseño de muestreo:

- m hace referencia a la medida de probabilidad inducida por el modelamiento y la inclusión de las covariables auxiliares ($\mathbf{x}_d$).
- p hacer referencia a la medida de probabilidad inducida por el diseño de muestreo complejo que induce las estimaciones directas.

Además, $\tilde{\psi}_d^2 = E_m(\hat{\psi}_d^2)$ es la varianza suavizada del estimador directo $\hat{\theta}_d^{dir}$. Un aspecto importante en este tipo de modelos es que, en general, no es posible tratar a ψ_d^2 como un valor fijo puesto que no es estrictamente una función de las covariables auxiliares. De hecho, es endógena con respecto al mismo parámetro que se pretende estimar. Partiendo del hecho de que se tiene acceso a un estimador insesgado de ψ_d^2, denotado por $\hat{\psi}_d^2$ se tiene que:

20 En el caso de comunas con una varianza regional estimada mayor a la estimación de la varianza comunal se asignó la varianza comunal en el caso de tener más de 20 UPM y más de 100 viviendas, en el caso de tener 20 o menos UPM se utilizó la aproximación regional. Para el resto de las comunas en que la estimación de la varianza comunal era mayor que la región, se optó por mantener la estimación comunal. Comunas con estimaciones de cero fueron excluidas del modelo Fay-Herriot, asignándoseles la estimación sintética como estimación SAE.

$$E_{mp}(\hat{\psi}_d^2) = E_m(E_p(\hat{\psi}_d^2)) = E_m(\psi_d^2) = \tilde{\psi}_d^2$$

La anterior igualdad puede interpretarse como que un estimador insesgado y simple de $\tilde{\psi}_d^2$ puede ser $\hat{\psi}_d^2$. Sin embargo, este estimador de muestreo es inestable cuando el tamaño de muestra es pequeño, que es justo el paradigma dominante en la estimación de áreas pequeñas. Rivest and Belmonte (2000) consideran modelos de suavizamiento para la estimación de las varianzas directas definidos de la siguiente manera:

$$\log(\hat{\psi}_d^2) = \mathbf{z}_d{}'\boldsymbol{\alpha} + \varepsilon_d$$

En donde $\mathbf{z}_d$ es un vector de covariables explicativas que son funciones de $\mathbf{x}_d$, $\boldsymbol{\alpha}$ es un vector de parámetros que deben ser estimados, ε_d son errores aleatorios con media cero y varianza constante, que se asumen idénticamente distribuidos condicionalmente sobre $\mathbf{z}_d$. Del anterior modelo, la estimación suavizada de la varianza de muestreo está dada por:

$$\tilde{\psi}_d^2 = E_{mp}(\hat{\psi}_d^2) = \exp(\mathbf{z}_d{}'\boldsymbol{\alpha}) \cdot \Delta$$

En donde, $E_{mp}(\varepsilon_d) = \Delta$. No hay necesidad de especificar una distribución paramétrica para los errores de este modelo. Al utilizar el método de los momentos, se tiene el siguiente estimador insesgado para Δ:

$$\hat{\Delta} = \frac{\sum_{d=1}^{D} \hat{\psi}_d^2}{\sum_{d=1}^{D} \exp\left(\mathbf{z}_d{}'\boldsymbol{\alpha}\right)}$$

De la misma forma, al utilizar mínimos cuadrados ordinarios, la estimación del coeficiente de parámetros de regresión está dada por la siguiente expresión:

$$\hat{\boldsymbol{\alpha}} = \left(\sum_{d=1}^{D} \mathbf{z}_d\, \mathbf{z}_d{}'\right)^{-1} \sum_{d=1}^{D} \mathbf{z}_d \log(\hat{\psi}_d^2)$$

Por último, el estimador suavizado de la varianza muestral está definido por:

$$\hat{\tilde{\psi}}_d^2 = \exp(\mathbf{z}_d{}'\hat{\boldsymbol{\alpha}})\hat{\Delta}$$

Rivest and Belmonte (2000) además concluyeron que este estimador no sobrestima ni subestima la varianza suavizada de la inferencia doble, puesto que el promedio de las estimaciones suavizadas $\hat{\tilde{\psi}}_d^2$ coincide con el promedio de las varianzas directas $\hat{\psi}_d^2$. Por tanto:

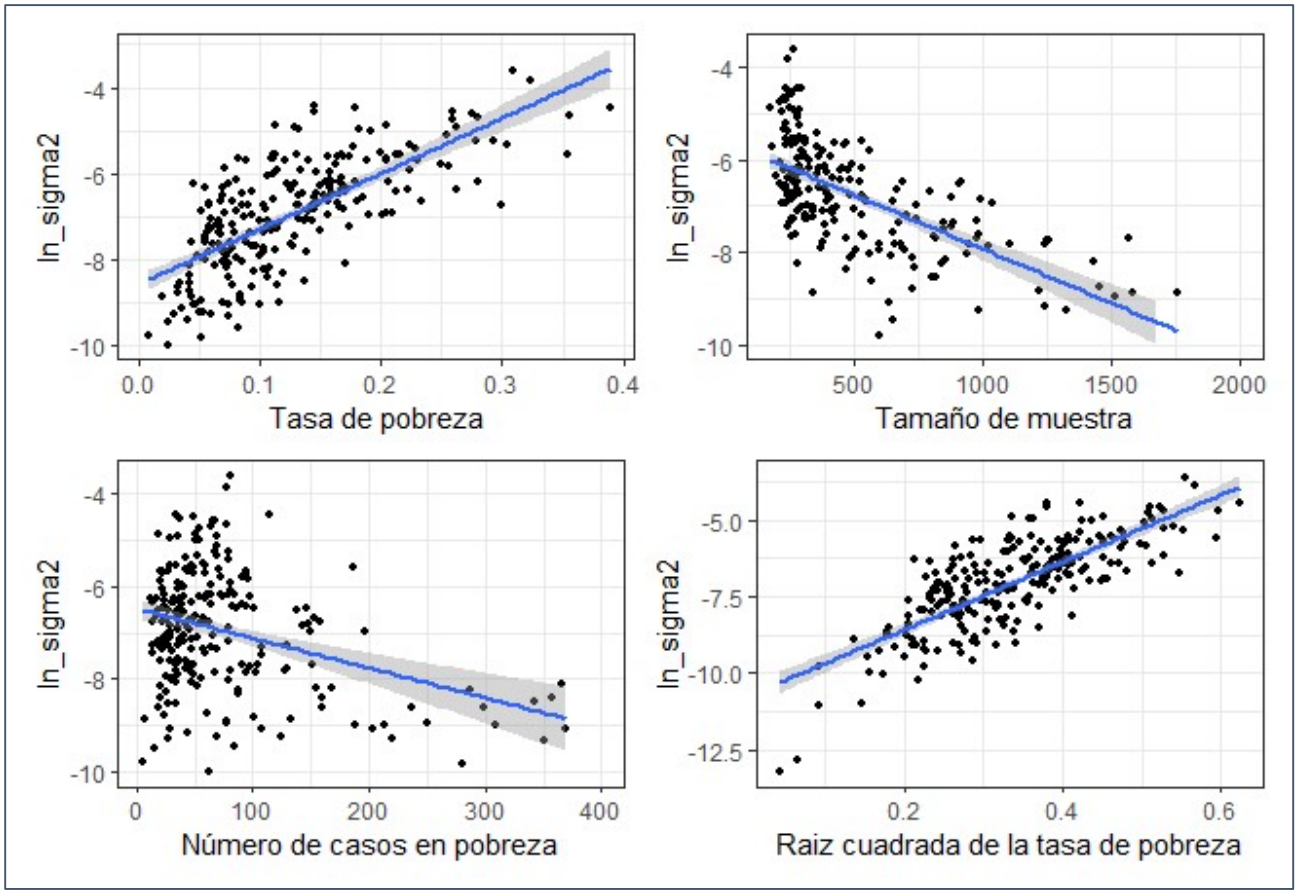
$$\frac{\sum_{d=1}^{D} \hat{\tilde{\psi}}_d^2}{D} = \frac{\sum_{d=1}^{D} \hat{\psi}_d^2}{D}$$

Para el ejercicio de estimación del Ministerio, el modelo considerado definió como variable dependiente el logaritmo natural de la varianza directa y como covariables se incluyó al intercepto, a la estimación directa de la tasa de pobreza, al tamaño de muestra comunal, a la interacción entre la tasa de pobreza y

el tamaño de muestra, a la raíz cuadrada de la tasa de pobreza, a la raíz cuadrada del tamaño de muestra y, por último, a la raíz cuadrada de la interacción entre la tasa de pobreza y el tamaño de muestra. Las comunas incluidas en la modelación que, pasando los criterios de calidad (capítulo 4), tienen una tasa nula de pobreza, y por consiguiente una estimación nula de la varianza del estimador directo no fueron incluidas en el ajuste del modelo[21], pero sí se obtuvieron las predicciones de sus varianzas.

La figura 6 muestran cuatro esquemas descriptivos que justifican las covariables y las relaciones establecidas en el modelo para el año 2017. Estas relaciones se mantienen para todos los años. Además, el factor de ajuste $\hat{\Delta}$ estuvo cercano a 1.2 en todas las series estudiadas.

Figura 6: Variable dependiente y covariables utilizadas en modelo FGV Casen 2017



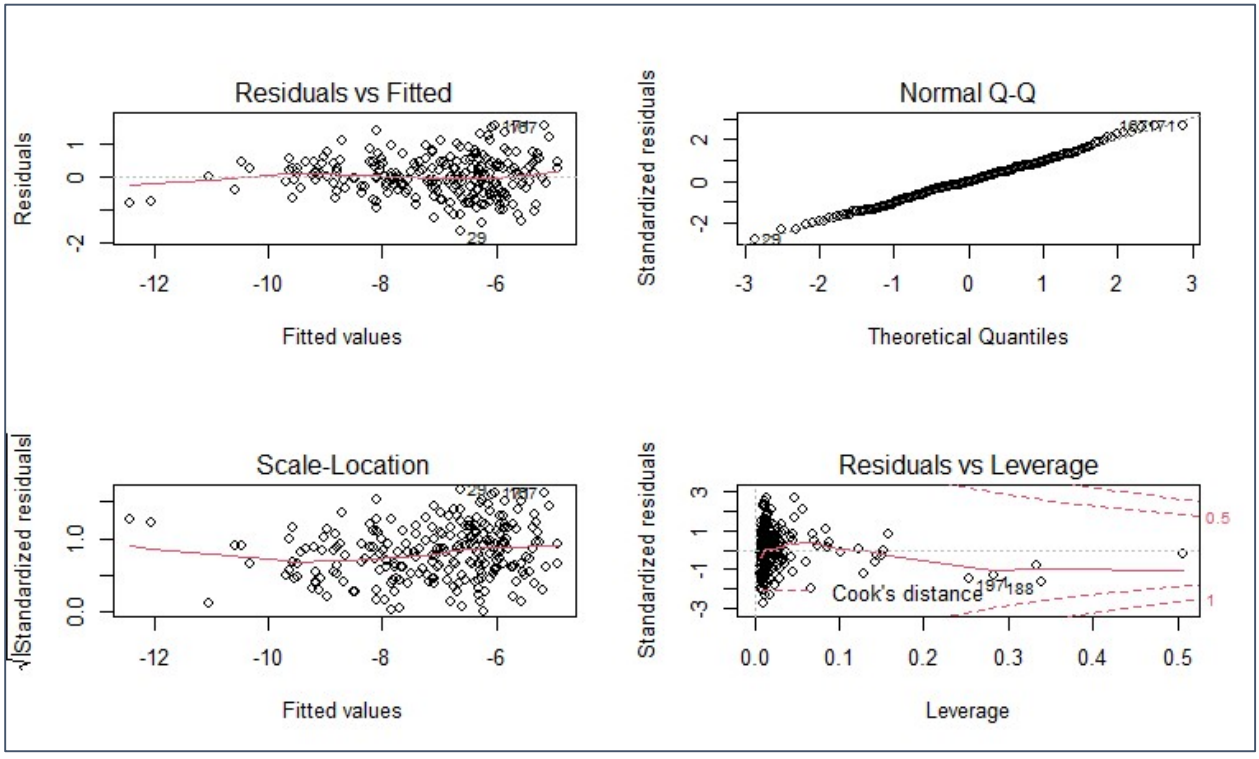
Fuente: Ministerio de Desarrollo Social y Familia – CEPAL

21 Notar que el modelo utiliza como variable dependiente el logaritmo natural de la varianza directa, el cual se indefine con varianzas nulas.

A continuación, se muestra el resultado de la modelación de la varianza para el año 2017 (figura 7). En particular, se nota que:

- De la esquina superior izquierda, los residuales no muestran ningún patrón evidente, lo que sugiere que las covariables identificadas sí alcanzan a describir el fenómeno con bastante precisión.
- Por otra parte, en la esquina superior derecha se puede observar que el gráfico de normalidad QQ muestra una coherencia adecuada entre los percentiles de una distribución normal estándar y la distribución empírica de los residuales estandarizados, por ende, existe evidencia de que los errores del modelo sí tienen una distribución normal.
- De la misma forma en la esquina inferior izquierda se muestra el gráfico de la raíz cuadrada de los residuales estandarizados contra la predicción de los valores del modelo: en general no se evidencia ningún patrón anormal que pueda indicar presencia de heteroscedasticidad que deba ser tenida en cuenta en el modelo.
- Finalmente, el gráfico de la distancia de Cook en la esquina inferior derecha no muestra comunas que se encuentren por fuera de los márgenes de la distancia de Cook y por ende se concluye que en general, no existen comunas influyentes en el modelo FGV.

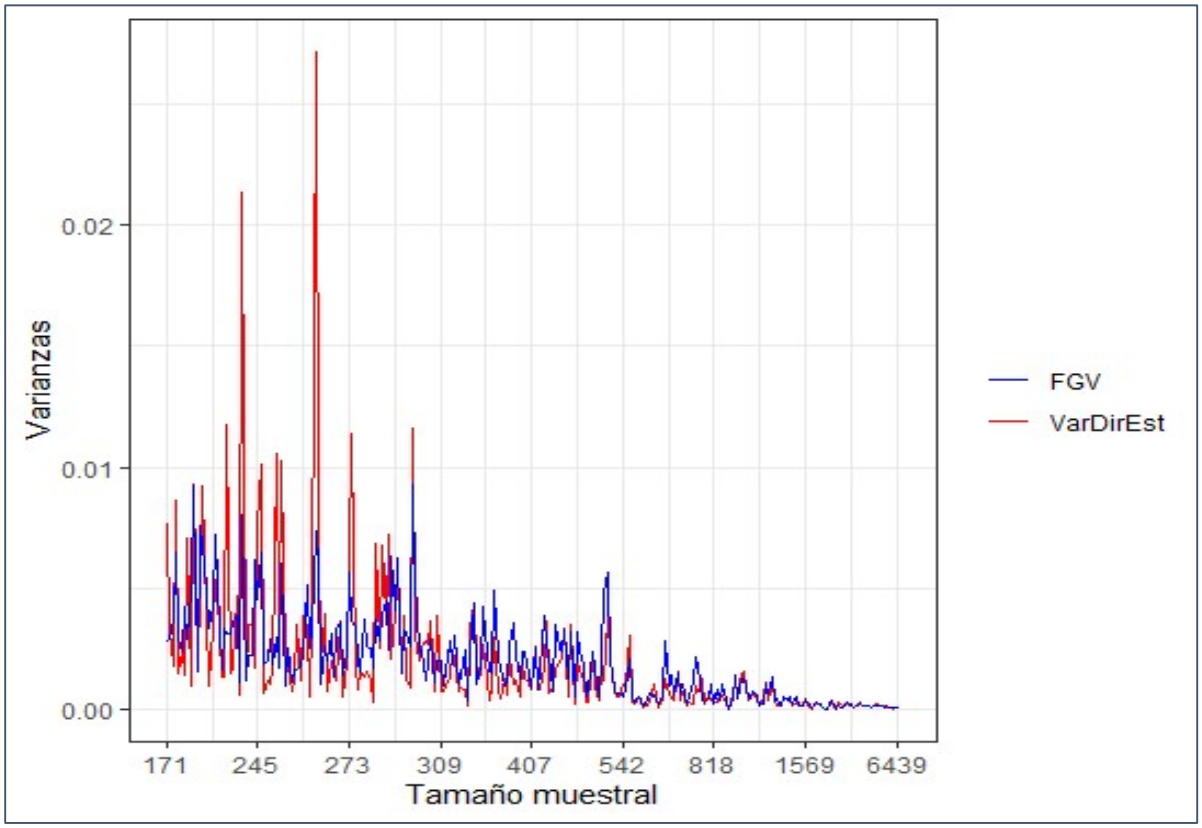
Figura 7: Análisis de calidad de ajuste del modelo FGV Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

A partir de la anterior estructura, se realiza la predicción de las varianzas directas comunales que, por definición, están suavizadas. En particular, obsérvese la figura 8 en donde se muestra la comparación entre la estimación directa de la varianza del estimador (línea roja) y la estimación suavizada del estimador (línea azul). Como las abscisas del diagrama están ordenadas ascendentemente de acuerdo con el tamaño de muestra en las comunas incluidas en la muestra, se puede verificar fácilmente que, como era de esperarse, la magnitud de la varianza disminuye a medida que el tamaño de muestra aumenta. Además, es indiscutible que el modelo suavizado induce una mayor estabilidad a través de las comunas.

Figura 8: Varianza directa y suavizada Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

5.4 Transformación arcoseno

Como se indicó en la introducción de este documento, en su concepción más básica, el modelo de FH es una combinación lineal de covariables. Sin embargo, el resultado de esta combinación (las predicciones) pueden tomar valores que se salen del rango aceptable en el que puede estar una proporción; es decir, en general el estimador de Fay-Herriot $\hat{\theta}_d^{FH} \in \mathbb{R}$, mientras que el estimador directo $\theta_d \in (0,1)$. Esta característica de los modelos lineales que intentan predecir proporciones hace que se deban utilizar alternativas metodológicas con base en transformaciones de la estimación directa y que permitirán garantiza que el estimador de Fay-Herriot $\hat{\theta}_d^{FH} \in (0,1)$.

Además, el estimador de la varianza de la estimación directa para una proporción en un muestreo complejo depende de la proporción estimada y por ende se crean problemas de endogeneidad en los parámetros. Jiang *et al.* (2001) propusieron una transformación que permite evadir las dificultades de trabajar con la estimación directa. En particular, se consideró que las proporciones directas estimadas son transformadas mediante la función arcoseno de la raíz cuadrada, $g^{-1}(\cdot) = \text{arcsin}(\sqrt{(\cdot)})$. Esta transformación asegura que los valores predichos por el modelo siempre estén en el rango de valores aceptables para una proporción.

Por otro lado, el problema de endogeneidad también es abordado por esta misma transformación. En primer lugar, nótese que, asumiendo independencia de las selecciones en cada área y suponiendo que la fracción de muestreo es pequeña, entonces, bajo un muestreo complejo, el estimador de la varianza de un estimador directo de una proporción $\hat{\theta}_d^{dir}$ está dado por la siguiente expresión:

$$\widehat{Var}_p(\hat{\theta}_d^{Dir}) \approxeq \widehat{DEFF}_d \times \frac{\hat{\theta}_d^{Dir}(1-\hat{\theta}_d^{Dir})}{n_d}$$

Y, por supuesto, esta expresión es endógena con respecto a la estimación directa $\hat{\theta}_d$. Ahora, al realizar la transformación, ésta queda formulada de la siguiente forma (no lineal):

$$\hat{z}_d = \text{arcsin}\left(\sqrt{\hat{\theta}_d^{Dir}}\right)$$

Es necesario utilizar una aproximación por series de Taylor para linealizar esta expresión y obtener su esperanza y su varianza. Por ende, la parte lineal del desarrollo de Taylor evaluada en $\hat{\theta}_d = \theta_d$ está dada por:

$$\hat{z}_d \approx g^{-1}(\theta_d) + (\hat{\theta}_d^{Dir} - \theta_d)\frac{d\, g^{-1}(\hat{\theta}_d^{Dir})}{d\, \hat{\theta}_d}|_{\hat{\theta}_d^{Dir}=\theta_d}$$

Nótese que

$$\frac{d\ \text{arcsin}\left(\sqrt{\hat{\theta}_d^{Dir}}\right)}{d\ \hat{\theta}_d^{Dir}}|_{\hat{\theta}_d=\theta_d} = \frac{1}{\sqrt{1-\theta_d}}\frac{1}{2\sqrt{\theta_d}}$$

Lo cual nos lleva, una vez más, a establecer que la parte lineal del desarrollo de Taylor de la transformación es:

$$\hat{z}_d \approx \text{arcsin}\left(\sqrt{\theta_d}\right) + (\hat{\theta}_d^{Dir} - \theta_d)\frac{1}{\sqrt{1-\theta_d}}\frac{1}{2\sqrt{\theta_d}}$$

Ahora, es posible analizar esta variable en términos de su esperanza y varianza, de la siguiente manera:

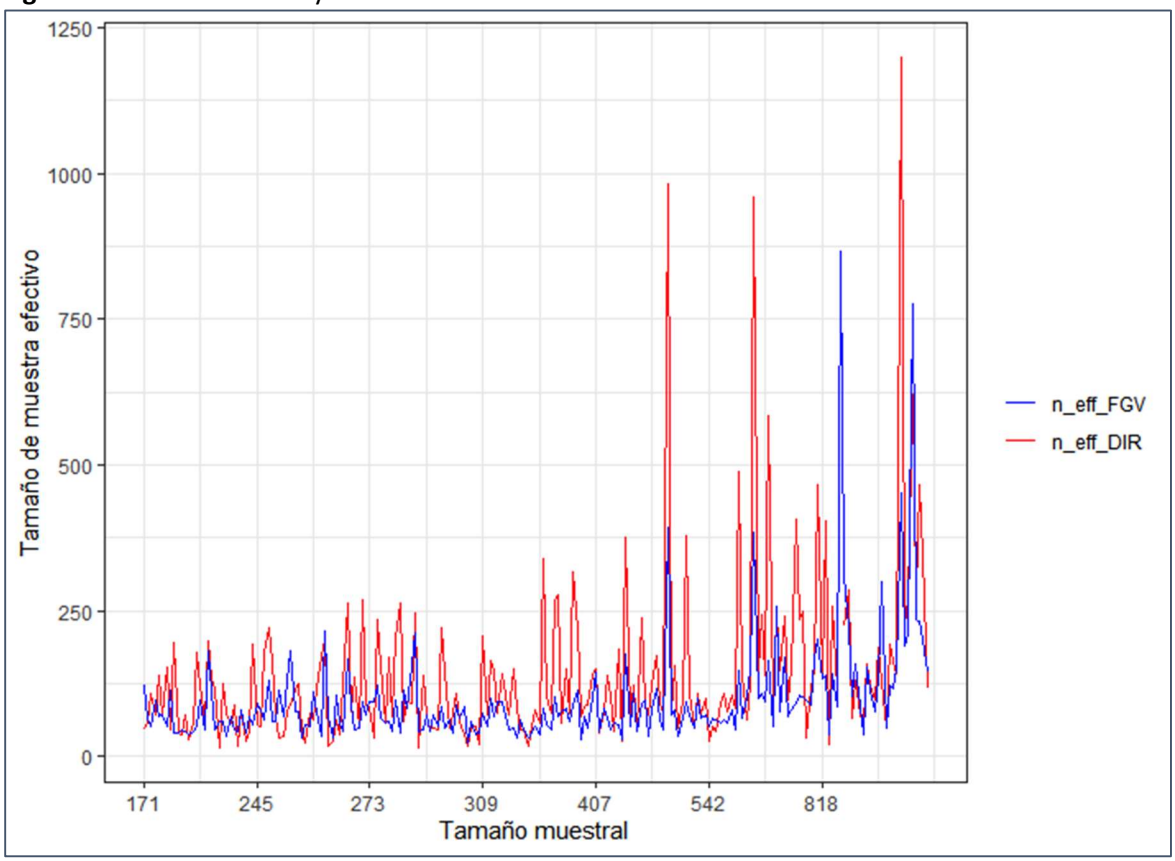
$$E_p(\hat{z}_d) \approx E_p\left[\text{arcsin}\left(\sqrt{\theta_d}\right) + (\hat{\theta}_d - \theta_d)\frac{1}{\sqrt{1-\theta_d}}\frac{1}{2\sqrt{\theta_d}}\right] = \text{arcsin}\left(\sqrt{\theta_d}\right)$$

Teniendo en cuenta que el tamaño de muestra efectivo está definido como la división entre el tamaño de muestra del conjunto de respondientes y el efecto de diseño en el área d, entonces esta varianza aproximada es igual a:

$$Var_p(\hat{z}_d) = \frac{\widehat{DEFF}_d}{4 \times n_d} = \frac{1}{4 \times n_{d,efectivo}}$$

Es decir, que la varianza no depende de los parámetros que se quieren estimar. Al utilizar esta transformación es necesario contemplar una evaluación empírica de lo que sucedería si se utilizara el suavizamiento de la varianza directa en la expresión anterior. Al respecto, la figura 9 muestra la relación entre el tamaño de muestra clásico (número de hogares en la muestra) y el tamaño de muestra efectivo (corregido por el efecto de diseño). En particular, si no se utilizara la función generalizada de varianza, se estaría introduciendo una dinámica bastante volátil en el modelo de áreas; mientras que, al considerar la modelación de la varianza directa, se está extrayendo una información bastante más estable con respecto a la variación del estimador directo. Lo anterior es evidente al observar la figura 8 en donde la línea roja representa el comportamiento del tamaño de muestra efectivo directo y la línea azul el tamaño de muestra efectivo suavizado.

Figura 9: Varianza directa y suavizada Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL.
Nota: Para una mejor visualización, figura se restringe a valores de muestra efectivo menores a 1250.

5.5 Estimación de los efectos de diseño

Realizar una estimación apropiada del efecto de diseño es fundamental para poder estimar apropiadamente la varianza del estimador directo transformado, la cual está en función del tamaño de muestra efectivo, que a su vez depende del efecto de diseño comunal. El proceso de estimación de Fay-Herriot con la transformación sugerida hará necesario estimar los efectos del diseño de muestreo complejo de Casen. Esta cantidad, se estima de la siguiente manera:

$$DEFF = \frac{Var(\hat{\theta}_d^{Dir})}{Var_{MAS}(\hat{\theta}_d^{Dir})}$$

Kish (1965, página 258) afirma que el efecto de diseño es la relación entre la varianza real de una muestra y la varianza real de una muestra aleatoria simple del mismo número de elementos. Esa frase final es la razón por la cual los diferentes softwares usan la varianza ponderada (ver más abajo), en vez de la varianza muestral. Además de lo anterior, la idea del efecto de diseño trata de evaluar el mismo estimador bajo diferentes escenarios de muestreo. Como el estimador que se está estudiando $\hat{\theta}_d$ viene ponderado por los factores de expansión de la encuesta, entonces lo más conveniente es utilizar el mismo rasero para evaluar ambas estrategias de muestreo. Es posible encontrar una discusión más profunda sobre el efecto de diseño en Gambino (2009, sección 4.), Sarndal, Swensson y Wretman (2003, página 188) y Gutierrez, Zhang y Montano (2016, página 101).

En resumen, $Var(\hat{\theta})$ es la varianza del diseño de muestreo complejo (estratificado, bietápico) para el estimador $\hat{\theta}$, y $Var_{MAS}(\hat{\theta}_d)$ es el estimador de la varianza bajo un muestreo aleatorio simple sin reemplazo para el estimador $\hat{\theta}$. En particular, bajo muestreo aleatorio simple, se tiene que

$$Var_{MAS}(\hat{\theta}_d^{Dir}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S_U^2$$

En donde N es el número de elementos en la población, n es el tamaño de muestra efectivo y S_U^2 es la varianza poblacional de la variable de interés, dada por

$$S_U^2 = \frac{\sum_U (y_k - \theta)^2}{N - 1}$$

Un estimador insesgado de esta varianza S_U^2 es la varianza muestral ponderada, la cual está dada por la siguiente expresión:

$$\hat{S}_U^2 = \left(\frac{n}{n-1}\right) \frac{\sum_s w_k (y_k - \hat{\theta}_d^{Dir})^2}{\sum_s w_k - 1}$$

De esta forma, una estimación de la varianza $Var_{MAS}(\hat{\theta}_d^{Dir})$ bajo muestreo aleatorio simple está dada por la siguiente expresión:

$$\widehat{Var}_{MAS}(\hat{\theta}_d^{Dir}) = \frac{1}{n}\left(1 - \frac{n}{\widehat{N}}\right)\hat{S}_U^2$$

En donde $\widehat{N} = \sum_s w_k$. Por lo tanto, la estimación del efecto de diseño DEFF está dada por

$$\widehat{DEFF} = \frac{\widehat{Var}(\hat{\theta}_d^{Dir})}{\widehat{Var}_{MAS}(\hat{\theta}_d^{Dir})}$$

6. Procedimiento final de estimación de la pobreza comunal para Casen

Como complemento a las estimaciones directas de pobreza y su estimación de varianza, el modelo de Fay-Herriot requiere la estimación de un modelo sintético de pobreza, así como la estimación de su varianza o errores aleatorios. Estas estimaciones buscan modelar la variable de interés como función de un conjunto de variables auxiliares relacionadas, buscando mejorar así la precisión de las estimaciones directas obtenidas por medio de la encuesta Casen.

Las variables auxiliares se basan en la disponibilidad de registros administrativos, censales o de otras fuentes. Dos elementos esenciales referidos a las variables auxiliares son: (i) tener información disponible en los niveles geográficos o poblacionales requeridos, y (ii) que esta información sea medida de manera consistente a través de todas las áreas (Eurostat, 2019; Gutiérrez, 2018). Además, esta información debe ser de calidad tanto en sus registros como en su cobertura.

En esta sección se describe la estimación del modelo sintético, incluyendo las variables analizadas, la selección del modelo, estimación de los coeficientes de la regresión y de la varianza.

6.1 Covariables incluidas en el análisis

En el caso de Chile, las estimaciones de área pequeña son responsabilidad del Ministerio de Desarrollo Social y Familia, el cual cuenta con acceso a la información de registros administrativos que se levanta en el sector público, y que se complementa con información del Censo de Población. La fuente de los registros administrativos corresponde a un set de información que proviene directamente de los Ministerios, tales como Ministerio de Salud, de Vivienda, de Educación y del Trabajo; de sistemas de información tales como el Registro Social de Hogares (RSH) y del Sistema Nacional de Información Municipal (SINIM); de otras instituciones públicas, tales como el INE, la Superintendencia de Salud, Superintendencia de Pensiones y la Agencia de Calidad de la Educación.

A continuación, se describen brevemente los tipos de variables que se han considerado en la selección del modelo sintético.

- **Variables del Censo:** Se consideran variables relativas a la proporción de población urbana y rural, proporción perteneciente a un pueblo originario, proporción de viviendas con hacinamiento, origen del agua según fuente de origen, materialidad de la vivienda (aceptable, recuperable, irrecuperable), déficit cuantitativo de vivienda, años de escolaridad, asistencia, analfabetismo, tasa de inmigración, tasa de dependencia, índice de envejecimiento, entre otros.

- **Variables administrativas de salarios y ocupación:** provienen de los registros de la Superintendencia de Pensiones y corresponde a información de los trabajadores asalariados formales, de estos registros se obtienen el ingreso promedio imponible, porcentaje de trabajadores con ingreso imponible menor al 50% de la mediana, porcentaje de trabajadores con ingreso imponible igual al ingreso mínimo y la tasa de ocupación formal dependiente.

- **Variables administrativas de salud:** provienen de la Superintendencia de Salud, FONASA. Ministerio de Salud y del INE, obteniéndose indicadores sobre la proporción de afiliados al sector público de salud (FONASA), afiliados al sector privado (ISAPRE), tasa de mortalidad infantil, esperanza de vida, tasa de años de vida potencialmente perdidos, proporción de adultos mayores y niños según estado nutricional, entre otros.

- **Variables administrativas de educación:** provienen del Ministerio de Educación y de la Agencia de Calidad, las variables corresponden a tasas de matrícula, escolaridad promedio, puntajes promedio del Simce, proporción de estudiantes según índice de vulnerabilidad escolar.

La producción de estadísticas a partir de registros administrativos requiere un análisis detallado de la calidad de estos datos. Para la metodología SAE, resulta importante evaluar la cobertura poblacional de estos indicadores, en tanto se espera que estos sean representativos del fenómeno de interés que se busca modelar al nivel de desagregación de interés (comunas). En particular, varios de estos indicadores proveen información parcial de la población, como por ejemplo los afiliados al sector público de salud, las personas que trabajan en el sector formal del mercado del trabajo, etc. La excepción son los indicadores del Censo, ya que proveen información para toda la población[22].

Por lo anterior, es necesario incluir correcciones por cobertura en los indicadores cuando sea necesario. Los indicadores administrativos se han trabajado utilizando variables en tasas o porcentajes que controlan por la población incluida en el indicador, cuando es necesario. En el caso de los indicadores de salud, se trabaja con trienios o quinquenios, ya que esta información suele ser muy volátil de año a año, sobre todo para indicadores con baja prevalencia, como por ejemplo la mortalidad infantil. De la misma forma, en algunos casos, hay comunas que no tienen información para todos los indicadores, por lo que se realiza una imputación según el valor de la provincia. Aunque estos son casos aislados, este hecho se toma también en consideración al momento de evaluar las variables que se van a incluir en el modelo, tratando de privilegiar comunas que no tengan valores imputados.

La disponibilidad del Censo 2017 tuvo repercusiones positivas en cuanto a la mejoría en la calidad de los datos de fuentes externas, ya que proporcionó una alternativa para la corrección de variables por cobertura, dando una solución al denominador de esta. Por ejemplo, para el año 2017 fue posible tener el número total de ocupados, lo que permitió corregir variables de ingresos formales por el porcentaje de participación de este nivel de formalidad en cada comuna. Desde el año 2018, el Ministerio de Desarrollo Social y Familia, está trabajando de manera continua en la generación de datos administrativos para distintas desagregaciones en base a homologaciones y estandarizaciones que aseguren tener información de calidad.

6.2 Selección de variables y estimación de coeficientes de regresión

Como se describió en la sección 6.1, las estimaciones SAE en Chile, han incluido un amplio set de variables auxiliares, las que requieren un procedimiento de selección para incluir en el modelo final. El objetivo es modelar la tasa de pobreza comunal como función de un conjunto de variables auxiliares $\boldsymbol{x}_d$ y un error aleatorio, de acuerdo con:

22 Aunque en zonas pequeñas o aisladas también podría haber problemas de cobertura.

$$\theta_d = \boldsymbol{x}_d{}'\beta + u_d\,,\ \ con\quad u_d \overset{iid}{\sim} (0, \sigma_u^2)$$

Entre 2009 y 2017, esta selección se realizó mediante un procedimiento de *stepwise* que parte de un set amplio de covariables y realiza una selección sistemática con base en la significancia de los regresores. Aunque el procedimiento de *stepwise* es realizado automáticamente por el software estadístico, de todas formas, es necesario verificar que el modelo final tenga propiedades deseables, para esto, se realizan algunas revisiones que incluyen:

- Que el modelo tenga consistencia teórica, es decir, que las variables finales tengan una relación con la tasa de pobreza y que sus coeficientes tengan el signo esperado.
- Revisión del ajuste mediante el estadístico R^2 y t-test de significancia estadística para los parámetros estimados.
- Análisis de los residuos para verificar que el error es ruido blanco y comprobar supuestos de normalidad y homocedasticidad.
- Revisión de que el modelo sea parsimonioso mediante los criterios de información (AIC y BIC)

La tabla 9 muestra, para cada año, las variables finales seleccionadas, al igual que el R^2 y las comunas incluidas en el modelo. Cada año se han incluido 5 a 7 variables que han resultado significativas, además de las variables dicotómicas regionales. Estas variables han entrado al modelo transformadas mediante logaritmo (variables continuas) o mediante proporciones[23]. En los Anexos se puede encontrar la lista de variables auxiliares consideradas en el análisis para el año 2020.

Nótese que la estimación de los parámetros del modelo sintético se realiza mediante un procedimiento de mínimos cuadrados generalizados (MCG), donde la variable dependiente corresponde a la tasa de pobreza comunal obtenida directamente de Casen, luego de aplicar el método de Potter a los factores de expansión, y las variables independientes provienen del set de variables auxiliares seleccionadas $\boldsymbol{x}_d$. La ponderación utilizada por MCG corresponde a la suma de las varianzas de la estimación directa suavizada $\left(\hat{\tilde{\psi}}_d^2\right)$ y a la varianza del modelo sintético o error aleatorio ($\hat{\sigma}_u^2$). De esta forma los coeficientes $\hat{\beta}$, se obtienen de la siguiente fórmula:

$$\hat{\beta} = \left[\sum_{d\epsilon S} \boldsymbol{x}_d \boldsymbol{x}_d{}' / \left(\hat{\sigma}_u^2 + \hat{\tilde{\psi}}_d^2\right)\right]^{-1} \left[\sum_{d\epsilon S} \boldsymbol{x}_d \hat{\theta}_d^{Dir} / \left(\hat{\sigma}_u^2 + \hat{\tilde{\psi}}_d^2\right)\right]$$

23 Hasta el año 2017 las variables de proporciones se transformaron con el arcoseno para mantener la consistencia con la transformación de la variable dependiente y para suavizar posibles valores extremos. Sin embargo, desde el año 2020 estas variables entran en su escala original.

Tabla 9: Variables auxiliares utilizadas en estimación sintética

	2009	2011	2013	2015	2017	2020
R cuadrado SAE[24]	0,80	0,82	0,82	0,84	0,82	0,86
N comunal	322	232	279	234	242	256
Remuneración promedio de los afiliados al AFC y AFP por cobertura (Mintrab)	x				x	
Proporción de población en el 40% más vulnerable según RSH						x
Cobertura del RSH						x
Porcentaje de pobreza histórica (Casen)	x				x	
Porcentaje población rural (Censo 2002)	x					
Tasa asistencia escolar (Sinim)	x					
Cobertura de afiliados a AFC y AFP con remuneración inferior al salario mínimo					x	
Cobertura de afiliados a AFC y AFP respecto del total de ocupados a nivel comunal					x	
Tasa de ocupación formal dependiente						x
Promedio de años de escolaridad (Censo 2017)					x	
Puntaje Simce historia 8vo básico						x
Porcentaje de población perteneciente a pueblo indígena (Censo)		x	x	x	x	x
Porcentaje afiliados Seguro de Cesantía con remuneración imponible menor al salario mínimo (AFC/AFP)		x	x	x		
Porcentaje de población afiliada a Fonasa A o B (Fonasa)		x	x	x		
Porcentaje de población afiliada a Isapre (Isapre)		x	x	x	x	
Diagnóstico nutricional de niños (obesos y sobrepeso)						x
Tasa de hacinamiento medio						x
Tasa de analfabetismo (Censo 2002)	x	x	x	x		

Fuente: Ministerio Desarrollo Social y Familia, metodología de estimación SAE (distintos años).
Nota: Valores del R cuadrado corresponden al modelo que se usó en cada año (mismas covariables), pero con la metodología presentada en este documento. La fórmula utilizada se encuentra en la sección 8.1 de este documento.

6.3 Estimación de la varianza del modelo sintético

[24] Este indicador no debe confundirse con el estimado para una regresión de mínimos cuadrados ordinarios. El detalle sobre su estimación se encuentra desarrollado en la sección 8.1 de este documento.

La estimación de la varianza de los efectos aleatorios, $\hat{\sigma}_u^2$, puede ser estimada mediante varios métodos que asumen una estimación conocida o suavizada de $\hat{\psi}_d^2$ (Hidiroglou 2019). Los métodos más conocidos son *Restricted Máximum Likelihood* (REML) de Rao y Molina (2015) y *Adjusted Density Maximization* (ADM) de Li y Lahiri (2010).

Por una parte, el método REML aproxima el valor de $\hat{\sigma}_u^2$mediante la log-verosimilitud del modelo Fay-Herriot, utilizando el algoritmo recursivo de *Fisher Scoring* mediante la siguiente ecuación:

$$\sigma_u^{2(a+1)} = \sigma_u^{2(a)} + \left[I\left(\sigma_u^{2(a)}\right)\right]^{-1} s\left(\sigma_u^{2(a)}\right)$$

Donde:

$$I(\sigma_u^2) = \frac{1}{2} tr(P^2)$$
$$s(\sigma_u^2) = -\frac{1}{2} tr(P) + \frac{1}{2} \hat{\theta}_d^{Dir\prime} P' P \hat{\theta}_d^{Dir}$$

Este método puede arrojar valores de $\hat{\sigma}_u^2$, menores que cero, lo que generalmente se corrige aproximando el valor a cero, implicando que la estimación de Fay-Herriot corresponderá a la estimación sintética. En 2010 Li y Lahiri generaron el método ADM que soluciona el problema de valores estimados menores que cero mediante la maximización del producto de la función de verosimilitud por la varianza del modelo. Si bien, este método evita valores negativos, Hidiroglou (2019) recomienda utilizarla con precaución ya que podría derivar en una sobreestimación del parámetro σ_u^2.

Por otro lado, el método ADM planteado por Li y Lahiri define la siguiente función de verosimilitud:

$$L_{adj}(\sigma_u^2) = \sigma_u^2 L_{reml}(\sigma_u^2)$$

De esta forma la log-verosimilitud se expresa como:

$$l_{adj}(\sigma_u^2) = log(\sigma_u^2) + l_{reml}(\sigma_u^2)$$

La condición de primer orden para encontrar el máximo es:

$$\frac{\partial l_{adj}(\sigma_u^2)}{\partial \sigma_u^2} = \frac{1}{\sigma_u^2} + s(\sigma_u^2) = \frac{1}{\sigma_u^2} - \frac{1}{2} tr(P) + \frac{1}{2} \hat{\theta}_d^{Dir\prime} P' P \hat{\theta}_d^{Dir} = 0$$

Como resultado del trabajo conjunto entre CEPAL y el Ministerio de Desarrollo Social y Familia, se encontró que la mejor forma de aproximar la estimación del error aleatorio del modelo sintético es primeramente utilizando el método REML, y en caso de que se obtengan estimaciones con valores negativos, utilizar ADM. Es importante mencionar que en ninguno de los años considerados en Casen se obtuvieron estimaciones negativas o nulas.

6.4 La transformación inversa y consistencia con las cifras nacionales

Es importante recordar que la estimación obtenida está transformada con la función arcoseno, por lo que es necesario devolver el estimador a su escala original. Esto se puede realizar mediante la aplicación de la función $sin^2\left(\hat{\theta}_d^{FH}\right)$. Sin embargo, y tal como lo señala Hadam, Wurz y Kreutzmann (2020), esto tiene asociado un sesgo producido por la no linealidad de la transformación.

Para solucionar lo anterior se puede utilizar una transformación que corrige por este sesgo incorporando el hecho de que el estimador de Fay-Herriot tiene una distribución asociada. De esta forma el estimador transformado a su escala original corresponde a:

$$\hat{\theta}_d^{FH,trans} = E\left(sin^2\left(\hat{\theta}_d^{FH}\right)\right) = \int_{-\infty}^{\infty} sin^2(t) f_{\hat{\theta}_d^{FH}}(t)dt$$

$$= \int_{-\infty}^{\infty} sin^2(t) \frac{1}{\sqrt{2\pi}\frac{\tilde{\sigma}_u\,\tilde{\psi}_d}{\tilde{\sigma}_u + \tilde{\sigma}_u\tilde{\psi}_d}} exp\left(-\frac{\left(t-\hat{\theta}_d^{FH}\right)^2}{2\frac{\tilde{\sigma}_u\tilde{\psi}_d}{\tilde{\sigma}_u + \tilde{\sigma}_u\tilde{\psi}_d}}\right) dt$$

Finalmente, con el objetivo de que las estimaciones comunales sean consistentes con las estimaciones regionales y nacionales que corresponden a los niveles de representatividad en la encuesta Casen (en donde se tiene garantizado el insesgamiento, la precisión y la consistencia de los estimadores) se realiza un proceso de *benchmarking* sobre las estimaciones de pobreza comunal. El ajuste aplicado corresponde a la razón entre el número de pobres en la región r, obtenido de Casen (estimación directa) y aquél que se obtiene de agregar las estimaciones comunales para la misma región r. De esta forma el ajuste viene dado por la siguiente expresión:

$$R_r = \frac{\hat{\theta}_r^{Dir} P_r}{\sum_{d \in R_r} \hat{\theta}_d^{FH,trans} P_c}$$

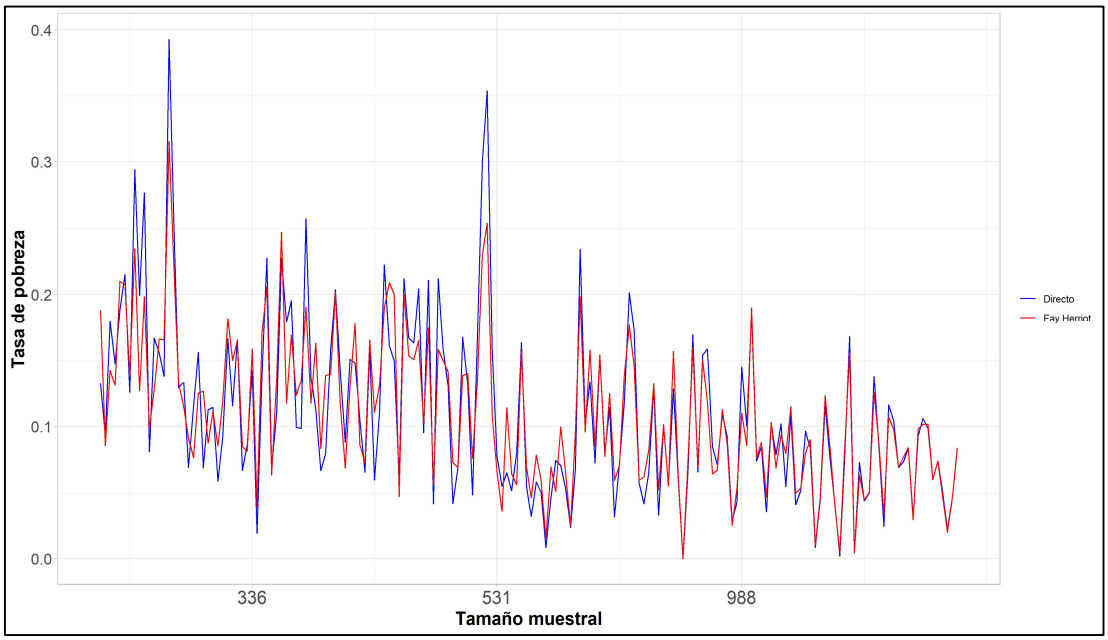
En donde, $\hat{\theta}_r^{Dir}$ corresponde a la estimación directa regional, $\hat{\theta}_d^{FH,trans}$ es la estimación de Fay-Herriot comunal en escala original, mientras que P_r y P_c corresponden a las poblaciones regionales y comunales respectivamente. De esta forma la estimación final de Fay-Herriot a nivel comunal corresponde a:

$$\hat{\theta}_d^{FH} = \hat{\theta}_d^{FH,trans} \times R_r$$

Este ajuste arroja resultados consistentes con las estimaciones regionales. En general, y como se observa en la figura 11, las estimaciones se mantienen en torno a la línea de los 45 grados. En la figura 10 se pueden observar las comunas ordenadas según el tamaño muestral. Los datos muestran que, a mayor tamaño de la muestra, las estimaciones Fay-Herriot son más cercanas a las estimaciones directas. Esto ocurre porque al aumentar el tamaño muestral, las estimaciones directas son más precisas y por consiguiente el modelo de Fay-Herriot otorga más peso a estas estimaciones.

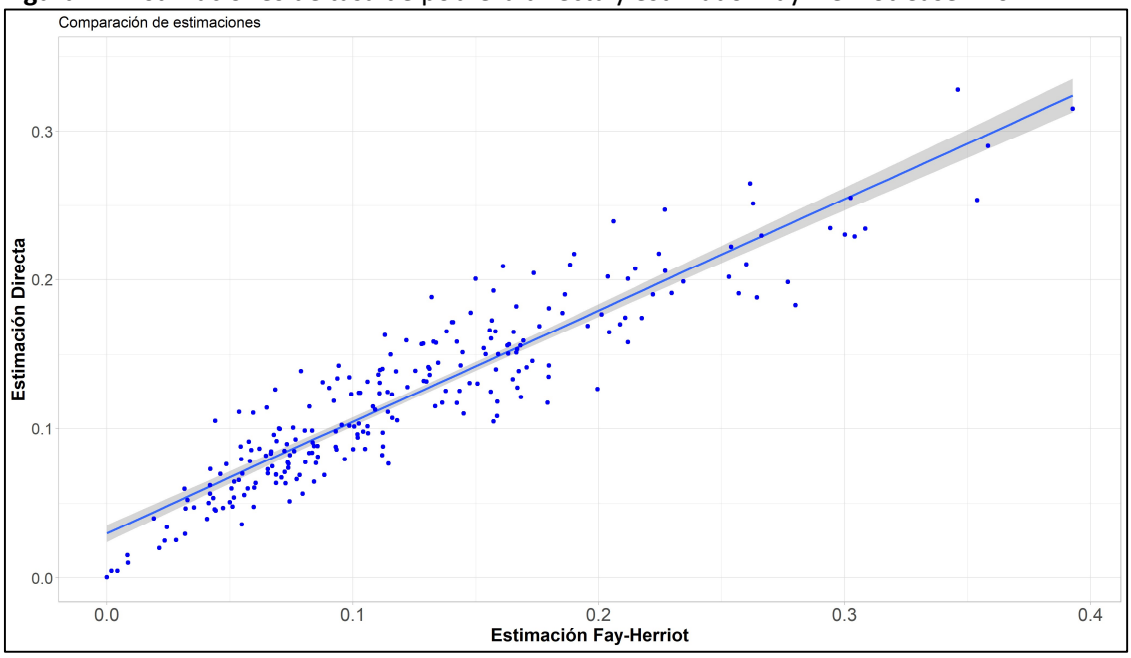
La figura 10, muestra una relación lineal entre la estimación de Fay-Herriot y la estimación directa. Lo que grafica que el modelo de Fay-Herriot, permite obtener estimaciones más precisas, sin alterar significativamente las estimaciones puntuales. Esto permite mantener la consistencia con las estimaciones puntuales a los niveles de representatividad de la encuesta (regional y nacional).

Figura 10: Estimaciones de tasa de pobreza directa y estimador Fay-Herriot, según tamaño de muestra Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

Figura 11: Estimaciones de tasa de pobreza directa y estimador Fay-Herriot Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

7. Estimación del error en el modelo

Por definición, el error cuadrático medio (ECM) es una medida de desempeño de un estimador que mide el promedio de los errores al cuadrado, en términos de la diferencia entre el parámetro y la estimación. Esta medida es ampliamente utilizada como indicador para validar la calidad estadística de las estimaciones obtenidas por cada dominio de interés con el fin de identificar aquellas áreas sobre las cuales se podrá realizar una inferencia apropiada. En general, si una comuna tiene un ECM muy alto, el modelo debe ser revisitado hasta obtener una ECM apropiado. En caso contrario, la comuna debería ser excluidas de cualquier conclusión, hipótesis o inferencia, y no deberían ser incluidas en el mapa de pobreza con el fin de no reportar estimaciones que no corresponden a la realidad de las condiciones de estos agregados geográficos en los países bajo análisis.

7.1 Estimación básica del ECM

En el capítulo anterior se presentó la fundamentación teórica de la metodología de Fay-Herriot con la cual se generaron las estimaciones de la tasa de pobreza comunal. Bajo el supuesto de normalidad sobre el efecto aleatorio, $u_d \overset{i.i.d}{\sim} N(0, \sigma_u^2)$ y el término de error a nivel de individuo $e_{di} \overset{iid}{\sim} N(0, \sigma_e^2)$, Prasad y Rao (1990) proveen la siguiente aproximación del error cuadrático medio del estimador Fay-Herriot:

$$\widehat{ECM}(\hat{\theta}_d^{FH}) \approx g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2)$$

donde σ_u^2 corresponde a la varianza del efecto aleatorio de la comuna. Los términos de la ecuación anterior son, respectivamente:

$$g_{1d}(\hat{\sigma}_u^2) = \hat{\gamma}_d \, \hat{\psi}_d$$

El cual es un término que, al aumentar el número total de dominios D, es de orden $O(1)$ y está asociado al error debido a la predicción del efecto aleatorio del área u_d. El segundo término $g_{2d}(\sigma_u^2)$ es el error debido a la estimación del vector de coeficientes de regresión $\boldsymbol{\beta}$.

$$g_{2d}(\hat{\sigma}_u^2) = \left(1 - \hat{\gamma}_d\right)^2 \boldsymbol{x}'_d \left(\sum_{d=1}^{D} (\hat{\sigma}_u^2 + \hat{\psi}_d)^{-1} \boldsymbol{x_d} \boldsymbol{x}'_{\boldsymbol{d}} \right)^{-1} \boldsymbol{x_d}$$

Mientras que $g_{3d}(\sigma_u^2)$ es el error debido a la estimación de la varianza σ_u^2

$$g_{3d}(\hat{\sigma}_u^2) = \left(1 - \hat{\gamma}_d\right)^2 (\hat{\sigma}_u^2 + \hat{\psi}_d)^{-1} \bar{V}(\hat{\sigma}_u^2)$$

Donde $\bar{V}(\hat{\sigma}_u^2)$ es la varianza asíntótica del estimador de σ_u^2. Según Molina, I. (2019), los dos últimos términos de la ecuación -$g_{2d}(\sigma_u^2)$ y $g_{3d}(\sigma_u^2)$- son de orden $O(D^{-1})$; es decir, que para un gran número de áreas D estos términos tienden a ser nulos. Por el contrario, para un número de áreas moderado es necesario tener en cuenta los tres términos para evitar la subestimación del ECM.

En el caso de las comunas que no hacen parte de la muestra de la encuesta Casen, el estimador Fay-Herriot toma la siguiente forma:

$$\hat{\theta}_d^{FH} = \hat{\theta}_d^{syn} = \boldsymbol{x}_d{}'\beta$$

Por lo tanto, para estas comunas específicas, el estimador del error cuadrado medio del estimador de Fay-Herriot será igual al estimador del error cuadrado medio del estimador sintético. Rao y Molina (2015), muestran que, en este caso, el MSE toma la siguiente forma:

$$\widehat{ECM}(\hat{\theta}_d^{FH}) = \hat{\sigma}_u^2 + \boldsymbol{x}_d'\left(\sum_{d=1}^{D}(\hat{\sigma}_u^2 + \hat{\psi}_d)^{-1}\boldsymbol{x_d}\boldsymbol{x}'_{\boldsymbol{d}}\right)^{-1}\boldsymbol{x_d}$$

7.2 Métodos no paramétricos en la estimación del ECM

Bajo las condiciones previamente mencionadas, el ECM puede ser estimado por medio de las soluciones analíticas presentadas por Prasad, N. N. y Rao, J.N.K. (1990); Datta, G.S. y Lahiri, P. (2000); Slud, E.V. y Maiti, T. (2006). En el caso de una transformación sinusoidal inversa, hasta donde sabemos, no se dispone de una solución analítica.

Según Jiang, J., Lahiri, P. y Wan, S. (2002), el proceso de estimación del ECM del mejor predictor empírico (EBP) para modelos mixtos no normales y no lineales, y en general, para los estimadores generales de los parámetros del modelo requiere de técnicas o metodologías estadísticas soportadas en las potentes capacidades computacionales actuales. Por tal motivo, los métodos *Jackknife* o *Bootstrap* son muy útiles para estimar el MSE, principalmente debido a la sencillez de su implementación.

7.2.1 Estimador *Jackknife* del ECM en el modelo de Fay-Herriot

En esta sección, explicamos el estimador *Jackknife* del ECM para el modelo tipo Fay-Herriot propuesto por Jiang, J., Lahiri, P. y Wan, S. (2002). Para el modelo de Fay-Herriot bajo transformación arcoseno, el estimador Jackknife del ECM de EBLUP viene dado por:

$$\widehat{ECM}_J(\widehat{\theta_i}) = g_{1i}(\hat{A}) - \frac{m-1}{m}\sum_{d=1}^{m}[g_{1i}(\hat{A}_{-d}) - g_{1i}(\hat{A})] + \frac{m-1}{m}\sum_{d=1}^{m}(\widehat{\theta}_{i,-d} - \widehat{\theta_i})$$

Donde $\hat{A}_{-d}(\hat{\beta}_{-d})$ es el estimador de $A(\beta)$ después de extraer la d-ésima área pequeña del proceso de estimación. En esta ecuación

$$\hat{\theta}_{i,-d} = \hat{\gamma}_{i,-d}\hat{\theta}_i^{DIR} + (1 - \hat{\gamma}_{i,-d})x_i'\hat{\beta}_{-d}$$

siendo

$$\hat{\gamma}_{i,-d} = \frac{\hat{A}_{-d}}{\hat{A}_{-d} + R_i}$$

$$g_{1i}(\hat{A}_{-d}) = \hat{A}_{-d}(1 - \hat{\gamma}_{i,-d})$$

Para caso especial del modelo de Fay-Herriot con $R_i = R$ y $x_i'\beta = \mu$ $(i = 1, \dots, m)$. Lahiri, P. y Rao, J.N.K. (1995) demostró que el estimador Jackknife del ECM está dado por

$$\widehat{ECM}(\hat{\theta}_i) = g_1(\hat{A}) + g_2(\hat{A}) + \frac{R^2}{m(\hat{A}+R)}(b_2 - 1) + \frac{R^2}{m(\hat{A}+\mathrm{R})^2}(b_2 - 1)(y_i - \bar{y})^2 - \frac{2R^2}{m(\hat{A}+R)^{\frac{3}{2}}}\sqrt{b_i}(y_i - \bar{y})$$

Donde

$$b_1 = \frac{m_3^2}{(\hat{A}+R)^3};\ b_2 = \frac{m_4}{(\hat{A}+R)^2};\ g_2(\hat{A}) = \frac{R_i^2}{(A+R_i)^2}x_i'\left(\sum_{d=1}^{m}\frac{x_d x_d'}{A+R_d}\right)^{-1}x_i$$

Por lo tanto, a diferencia de los estimadores del ECM basados en la normalidad, el estimador Jackknife del ECM involucra términos estimados de asimetría (m_3^2) y curtosis (m_4^2). Jiang, J., Lahiri, P. y Wan, S. (2002) demostraron que el sesgo del estimador *Jackknife* del ECM es de orden $o(m^{-1})$ y que este estimador posee buenas propiedades frecuentistas y bayesianas. Los resultados anteriores son para la escala transformada. Para proporcionar resultados en la escala original, los autores aproximan el ECM de $h^{-1}(\widehat{\theta_i})$ empleando

$$\widehat{ECM}\left[h^{-1}(\widehat{\theta_i})\right] = h^{-1'}(\widehat{\theta_i})\widehat{ECM}(\widehat{\theta_i}),$$

donde $h^{-1'}(x)$ denota la primera derivada de $h^{-1}(x)$ con respecto a x y $mse(\widehat{\theta_i})$ es una estimación del MSE obtenido por el método jackknife previamente descrito.

7.2.2 Estimador Bootstrap del ECM en el modelo con transformación

En la actualidad existen dos enfoques metodológicos que permiten obtener estimaciones para el error cuadrático medio (ECM) y generar intervalos de confianza bajo un modelo FH que considera la transformación arcoseno para la variable dependiente. Autores como Casas-Cordero et al. (2016), Burgard et al. (2016) y Schmid et al. (2017), utilizaron el método de *Bootstrap* paramétrico, en el que el ECM y los umbrales de los intervalos de confianza del bootstrap se construyen en la escala transformada con la posterior transformación inversa naive $h^{-1}(x) = \sin^2(x)$. En contraste con este enfoque, Hadam, S., Wurz, N. y Kreutzmann, A-K. (2020) transforman el estimador FH, al nivel original de la variable, teniendo en cuenta el sesgo de transformación inversa.

Bajo el segundo enfoque se considera la propuesta metodológica desarrollada por Burgard et al. (2016) y Sugasawa, S. y Kubokawa, T. (2017), en donde se presenta el modelo FH que considera de forma específica el sesgo de transformación inversa de la variable dependiente con transformación arcoseno. El paquete desarrollado por Kreutzmann et al. (2018) en el software estadístico R permite calcular los errores cuadráticos medios empleando las dos metodologías previamente mencionadas. A continuación, se presenta el algoritmo que desarrolla el Bootstrap paramétrico propuesto por Hadam, S., Wurz, N. y Kreutzmann, A-K. (2020):

1. Realizar la transformación arcoseno a las estimaciones directas $\hat{\theta}_d^{Dir}$ junto a las varianzas de la transformación arcoseno $\arcsin\left(\sqrt{\hat{\theta}_d^{D}}\right)$ definidas por Jiang, J., Lahiri, P. y Wan, S. (2002) como $\sigma_{e_d}^2 = 1/4\tilde{n}_d$ donde $\tilde{n}_d$ corresponde a los tamaños de muestra efectivos en las áreas pequeñas. A partir de estos cálculos, estimar $\hat{\sigma}_u^2$ y $\hat{\beta}$ usando las áreas seleccionadas.

2. Reemplazar los parámetros del modelo con sus estimaciones para obtener el estimador FH en el nivel transformado

$$\hat{\theta}_i^{FH*} = \hat{\gamma}_i \sin^{-1}\left(\sqrt{\hat{\theta}_i^{Dir}}\right) + (1-\hat{\gamma}_i)X_i^T\hat{\beta}$$

3. Mediante técnicas de integración numérica se obtiene el estimador FH en la escala original $\hat{\theta}_i^{FH,trans}$. En este paso se transforma el estimador $\hat{\theta}_i^{FH*}$ considerando los supuestos del modelo FH

$$\sin^{-1}\left(\sqrt{\hat{\theta}_i^{Dir}}\right) = X_i^T\hat{\beta} + u_i + e_i, \qquad u_i \sim N(0,\hat{\sigma}_u^2) \;\; y \;\; e_i \sim N\left(0,\hat{\bar{\psi}}_i^2\right)$$

Para tal fin, se emplea la siguiente fórmula

$$\hat{\theta}_i^{FH,trans} = E\left\{\sin^2\left(\hat{\theta}_i^{FH*}\right)\right\} = \int_{-\infty}^{\infty} \sin^2(t)\, f_{\hat{\theta}_i^{FH*}}(t)\, dt$$

$$= \int_{-\infty}^{\infty} \sin^2(t) \frac{1}{2\pi\frac{\hat{\sigma}_u^2\hat{\bar{\psi}}_i^2}{\hat{\sigma}_u^2+\hat{\bar{\psi}}_i^2}} exp\left(-\frac{\left(t-\hat{\theta}_i^{FH*}\right)^2}{2\frac{\hat{\sigma}_u^2\hat{\bar{\psi}}_i^2}{\hat{\sigma}_u^2+\hat{\bar{\psi}}_i^2}}\right) dt$$

para evitar el sesgo debido a la no linealidad de la transformación; lo anterior bajo la conocida distribución del estimador $\hat{\theta}_i^{FH*} \sim N\left(\hat{\theta}_i^{FH*}, \frac{\hat{\sigma}_u^2\hat{\bar{\psi}}_i^2}{\hat{\sigma}_u^{2+}\hat{\bar{\psi}}_i^2}\right)$

4. Generar efectos Bootstrap para los dominios de interés d a partir de la densidad normal dada por

$$u_d^{*(b)} \sim_{iid} N(0,\hat{\sigma}_u^2), \qquad d = 1,2,\dots,D$$

5. Generar los errores Bootstrap de forma similar pero independiente a $u_d^{*(b)}$, empleando la densidad normal dada por

$$e_{di}^{*(b)} \sim_{iid} N\left(0,\hat{\bar{\psi}}_i^2\right), i = 1,2,\dots,N_d; d = 1,2,\dots,D$$

6. Generar la población Bootstrap de la variable respuesta transformada empleando los hogares seleccionados en la encuesta Casen y la ecuación del modelo

$$\left(arcsin\left(\sqrt{\hat{\theta}_d^{Dir}}\right)\right) = \boldsymbol{x}_{di}^T\hat{\boldsymbol{\beta}} + u_d^{*(b)} + e_{di}^{*(b)}$$

donde, $u_d^{*(b)} \sim_{iid} N(0, \hat{\sigma}_u^2)$ fue generado en el paso 4 y $e_{di}^{*(b)} \sim_{iid} N(0, \hat{\tilde{\psi}}_i^2)$ en el paso 5.

7. Para cada iteración Bootstrap, obtener las estimaciones de FH en la escala original $\left(\hat{\theta}_{i,(b)}^{FH,trans}\right)$ utilizando la integral definida para la corrección del sesgo. En este paso se deben utilizar las estimaciones Bootstrap generadas de las áreas muestreadas para generar un modelo Bootstrap con las mismas covariables utilizadas. Este proceso genera nuevas estimaciones Fay - Herriot tanto para las áreas muestreadas como para las no muestreadas. Para las áreas no muestreadas, la estimación Fay - Herriot del Bootstrap corresponderá a la estimación sintética generada en este nuevo modelo a partir de los efectos aleatorios y errores de muestreos generados.

8. Para cada iteración Bootstrap calcular

$$\theta_{d,(b)}^{trans} = E(\sin^2(x_d'\hat{\beta} + u_d^*))$$

9. Estimar el MSE y los intervalos de confianza del 95%

$$MSE_B\left(\hat{\theta}_d^{FH,trans}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{d,(b)}^{FH,trans}, \theta_{d,(b)}^{trans}\right)^2$$

El procedimiento Bootstrap paramétrico bajo transformación inversa "naive" tiene una secuencia similar a la previamente ilustrada solo que en lugar de utilizar integración numérica hace uso de forma explícita la expresión $h^{-1}(x) = \sin^2(x)$. Resulta importante señalar que, Según la desigualdad de Jensen (Jensen et al., 1906), una transformación inversa ingenua $\left(\sin^2\left(\hat{\theta}_i^{FH*}\right)\right)$ conduce a resultados sesgados debido a la no linealidad de la transformación -ver Hadam, S., Wurz, N. y Kreutzmann, A-K. (2020).

7.3 Coeficiente de variación e intervalos de confianza

Con el error cuadrático medio estimado se puede calcular una medida de variación relativa, similar al coeficiente de variación, con el fin de tener un criterio objetivo que permita definir la calidad de las estimaciones. El coeficiente de variación bajo el estimador del ECM de Bootstrap paramétrico propuesto por Hadam, S., Wurz, N. y Kreutzmann, A-K. (2020) se define como sigue:

$$\widehat{CV} = \frac{\sqrt{MSE_B\left(\hat{\theta}_d^{FH,trans}\right)}}{\hat{\theta}_d^{FH,trans}} * 100$$

Se recomienda que las comunas con un $\widehat{CV}$ mayor a 30% sean excluidas de los mapas de pobreza al considerarse que no tienen la precisión requerida. Es posible considerar dos metodologías diferentes para la estimación por intervalo del indicador de interés en las comunas chilenas. El primer enfoque consiste en un intervalo de confianza tipo Wald establecido a partir del enfoque de Prasad, N. N. y Rao, J.N.K. (1990); Datta, G.S. y Lahiri, P. (2000) para la estimación del error cuadrático medio

$$\hat{\theta}_d^{FH} \pm Z_{\frac{\alpha}{2}} \sqrt{MSE_B\left(\hat{\theta}_d^{FH,trans}\right)}$$

Por el contrario, bajo el enfoque del Boostrap paramétrico propuesto por Hadam, S., Wurz, N. y Kreutzmann, A-K. (2020), el intervalo de confianza del 95% del estimador FH con transformación inversa corregida por sesgo se obtiene a partir de los cuantiles 2.5% ($q_{0.025}$) y 97.5% ($q_{0.975}$) de las réplicas Bootstrap del algoritmo previamente descrito, lo que corresponde a la siguiente expresión:

$$CI\left(\hat{\theta}_d^{FH,trans}\right) = [\hat{\theta}_{d,(b)}^{FH,trans} + q_{0.025}\left(\hat{\theta}_{d,(b)}^{FH,trans}, \theta_{d,(b)}^{trans}\right);$$
$$\hat{\theta}_{d,(b)}^{FH,trans} + q_{0.975}\left(\hat{\theta}_{d,(b)}^{FH,trans}, \theta_{d,(b)}^{trans}\right)]$$

Como la distribución teórica del estimador Fay - Herriot es normal (paso 3 del algoritmo) y considerando que la transformación realizada con la integral es insesgada. Podemos escribir los intervalos de confianza del 95 % como:

$$CI\left(\hat{\theta}_d^{FH,trans}\right) = (\hat{\theta}_d^{FH,trans} \pm 1.96\, RMSE\left(\hat{\theta}_d^{FH,trans}\right))$$

El Ministerio estimó los Intervalos de confianza de las estimaciones 2011 a 2017 mediante un procedimiento de bootstrap que consistió en replicar las estimaciones Fay-Herriot y con base en estas iteraciones construir la distribución de los estadísticos de prueba, para obtener los valores críticos $t_{0.025}$ y $t_{0.975}$ que intervendrían en los Intervalos de Confianza de acuerdo con la siguiente expresión:

$$I_d(t) = \left\{\hat{\theta}_d^{FH} - t_{0.025}\sqrt{\hat{\psi}_d^2(1-\hat{B})}\hat{\theta}_d^{FH}; \hat{\theta}_d^{FH} + t_{0.975}\sqrt{\hat{\psi}_d^2(1-\hat{B})}\hat{\theta}_d^{FH}\right\}$$

Este procedimiento permite calcular los Intervalos de Confianza solo para las comunas que tienen muestra en Casen, ya que dependen del parámetro $\hat{\sigma}_e^2$ (varianza directa de la estimación Casen). Desde 2020 se adopta la metodología de cálculo del Intervalo de Confianza mediante el estimador bootstrap del ECM (sección 7.2.2), lo que permite obtener Intervalos de Confianza para todas las comunas del país.

8. Validación del modelo Fay – Herriot

El modelo Fay-Herriot requiere que se cumplan algunos supuestos asociados a la homocedasticidad y a la normalidad de los residuos para asegurar su validez. Además, es recomendable verificar algunas propiedades deseables como la bondad de ajuste y la significancia estadística de los parámetros estimados. Este capítulo tiene por objetivo verificar que los supuestos se cumplen y analizar la calidad del modelo. Por simplicidad, los análisis expuestos en este capítulo se refieren al año 2017, pero las conclusiones se mantienen para las estimaciones de área pequeña realizadas desde 2009 a la fecha.

8.1 Bondad de ajuste y Test-t de los parámetros

El coeficiente de determinación, R^2, corresponde a la proporción de la variabilidad muestral total explicada por la regresión, además de servir como insumo para evaluar el poder predictivo de las variables auxiliares x_d. En el contexto específico de los modelos de área, dicho coeficiente se define como se muestra a continuación (Hidiroglou and Yung, 2019):

$$R^2 = 1 - \frac{\hat{\sigma}_u^2}{\frac{(D-p)}{(D-1)}\hat{\sigma}_u^2 + S^2(\hat{\beta})}$$

Donde p es el número de covariables utilizadas en el modelo, D es el número de comunas muestreadas y la varianza muestral de $x_d'\hat{\beta}$, denotada como $S^2(\hat{\beta})$ toma la siguiente forma:

$$S^2(\hat{\beta}) = \frac{\sum_{d=1}^{D}(x_d'\hat{\beta} - \bar{x}_d'\hat{\beta})^2}{D-1}$$

Tal como lo señalan Hidiroglou y Yung (2019), el investigador no tiene ningún interés en conocer el coeficiente de determinación asociado al modelo $\hat{\theta}_d^{Dir} = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d + e_d$, puesto que el objetivo de la metodología nunca será predecir el comportamiento del estimador directo, $\hat{\theta}_d^{Dir}$. Al contrario, como lo que se desea es predecir el parámetro θ_d, entonces el interés recae sobre el modelo $\theta_d = \boldsymbol{x}_d{}'\boldsymbol{\beta} + u_d$.

Para el caso particular de Chile en el año 2017, el R^2 es de 0,816 lo que significa que el modelo tiene un buen grado de ajuste. En la tabla 10, se muestra el coeficiente de determinación para todos los años disponibles:

Tabla 10: Ajustes del modelo para todos los años según R^2 (coeficiente de determinación)

Año	Coeficiente de determinación
2009	0,801
2011	0,817
2013	0,820
2015	0,842
2017	0,816
2020	0,856

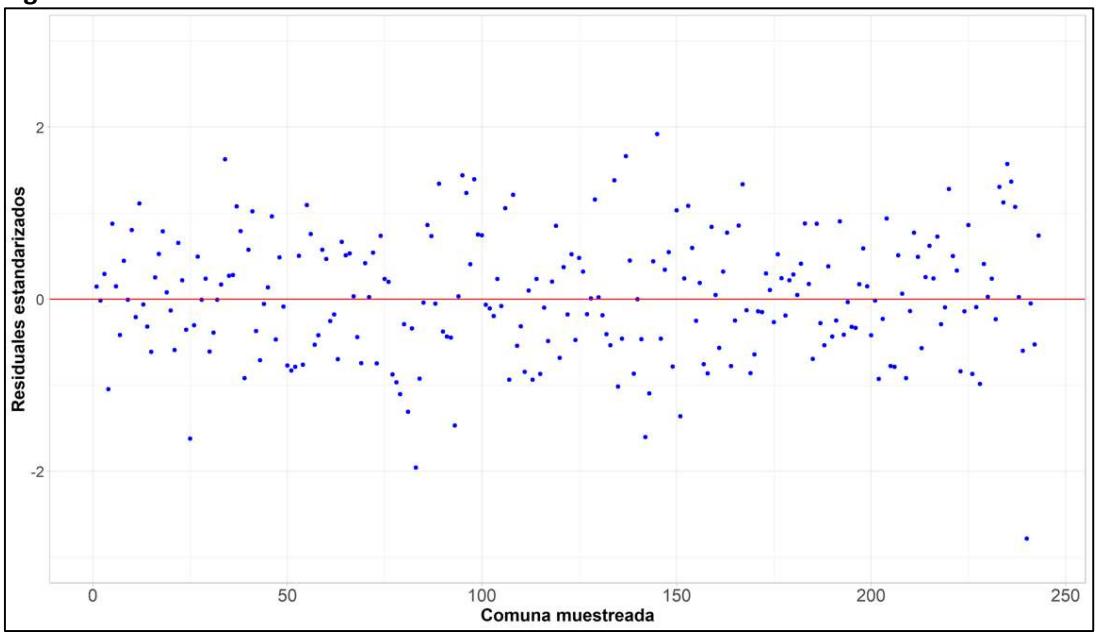
Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

8.2 Homocedasticidad de los residuos

El supuesto de homocedasticidad de los errores significa que la varianza es la misma para todos los valores de x e igual a σ^2. La figura 12 muestra los residuos estandarizados $(\hat{\theta}_d^{FH} - x_d'\hat{\beta})/\sqrt{\hat{\sigma}_u^2 + \hat{\tilde{\psi}}_d^2}$ versus las comunas muestreadas. Mientras que la figura 13 muestra los residuos estandarizados versus los valores predichos estandarizados $x_d'\hat{\beta}/\sqrt{\hat{\sigma}_u^2 + \hat{\tilde{\psi}}_d^2}$.

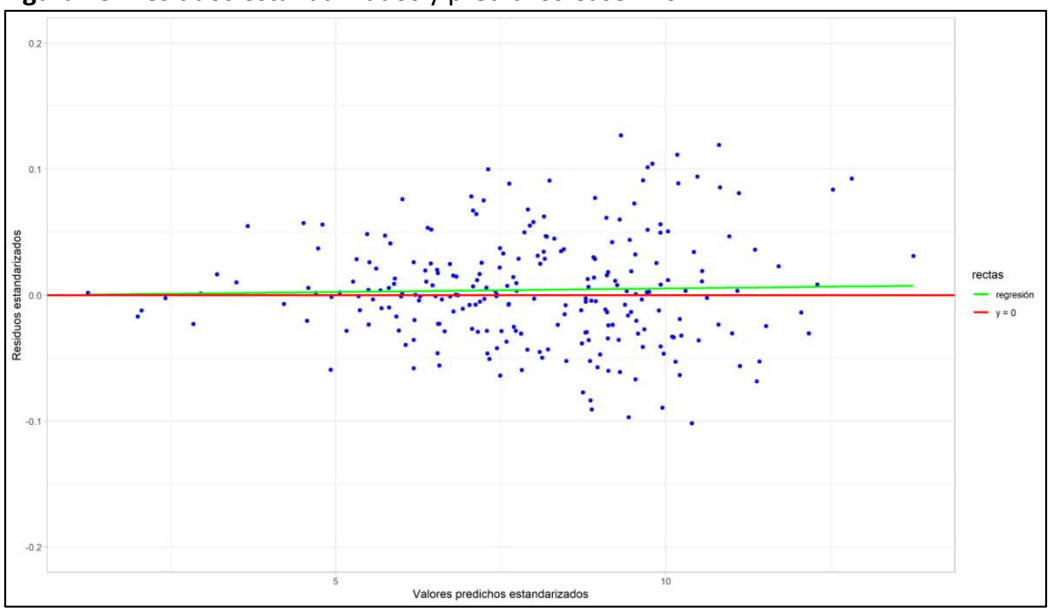
Las figuras 12 y 13 muestran que la dispersión de los residuos no tiene un patrón específico, concentrándose en torno a la recta que pasa por y=0, mostrando que la especificación del modelo es apropiada. Además, se observa que los residuos poseen una variabilidad similar a lo largo de las comunas muestreadas, lo que permite verificar el supuesto de homocedasticidad de los errores.

Figura 12: Residuos estandarizados Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

Figura 13: Residuos estandarizados y predichos Casen 2017

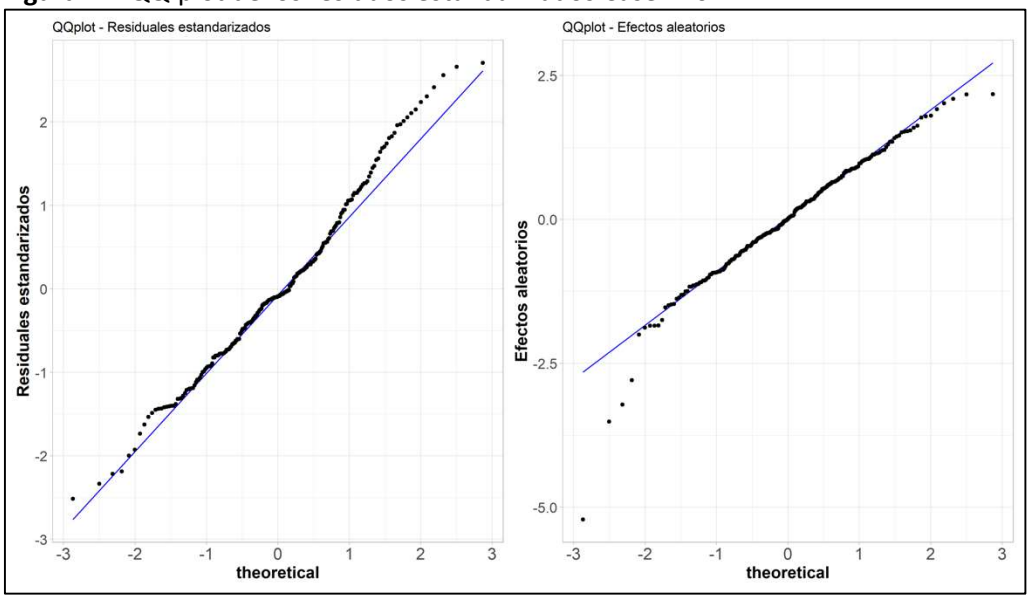


Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

8.3 Normalidad de los errores

Para evaluar la normalidad de los errores se utiliza el gráfico QQ-plot que permite comparar la distribución de los residuos estandarizados con los cuantiles teóricos de una distribución normal. La figura 14 muestra algunas desviaciones en ambas colas de la distribución, pero no se rechaza la hipótesis de normalidad. Además, el modelo de Fay-Herriot es robusto a desviaciones leves de la normalidad (Rao y Molina, 2015).

Figura 14: QQ-plot de los residuos estandarizados Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia- CEPAL

8.4 Distancia de Cook

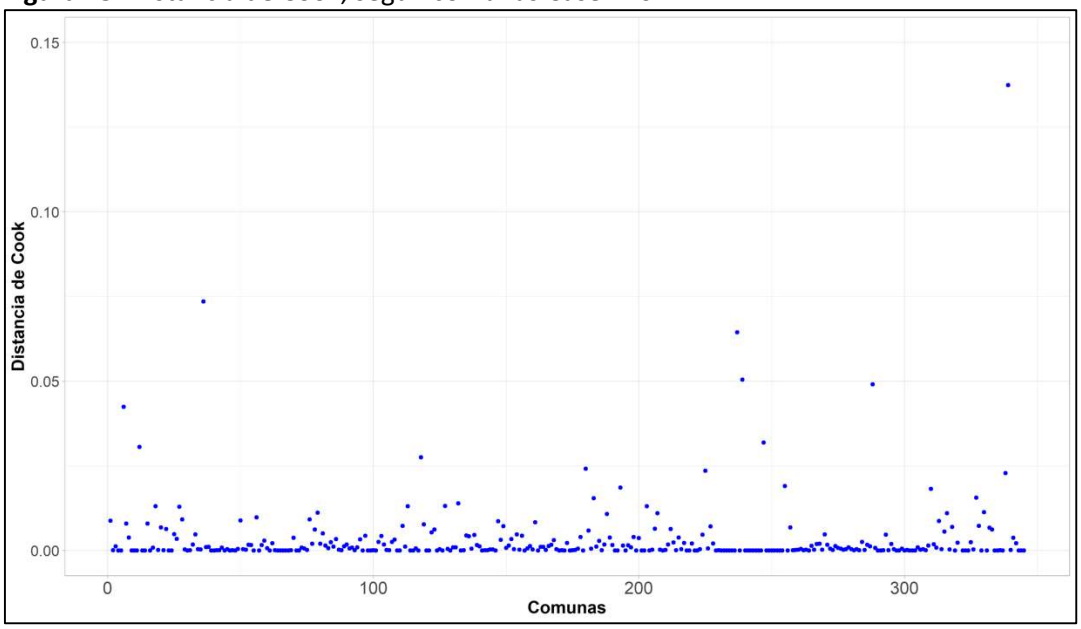
Un *outlier* es una observación que no se ajusta al modelo y que se manifiesta en un residuo excesivamente grande. Por otro lado, una observación con alto *leverage* corresponde a una observación con un valor extremo para alguno de los predictores que podría ser un punto influyente. La distancia de Cook combina en un único valor la magnitud del residuo y el nivel de *leverage*. Valores mayores a 1 se pueden considerar como observaciones influyentes, que son aquellas que influyen de forma importante en el modelo, es decir, su exclusión incide en el ajuste del modelo.

Se obtuvo la distancia de Cook para identificar a las comunas que podrían ser valores influyentes en la estimación de los coeficientes $\hat{\beta}$. La distancia de Cook para la comuna i está dada por:

$$D_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}^{(-i)})' \sum_{d \in S} \frac{x_d x_d'}{\hat{\sigma}_u^2 + \hat{\tilde{\psi}}_d^2} (\hat{\beta} - \hat{\beta}^{(-i)})$$

Donde $\hat{\beta}^{(-i)}$ corresponde a la estimación para β luego de quitar al área i del modelo. En la figura 15 se muestra la distancia de Cook para las comunas ordenadas por tamaño.

Figura 15: Distancia de Cook, según comunas Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

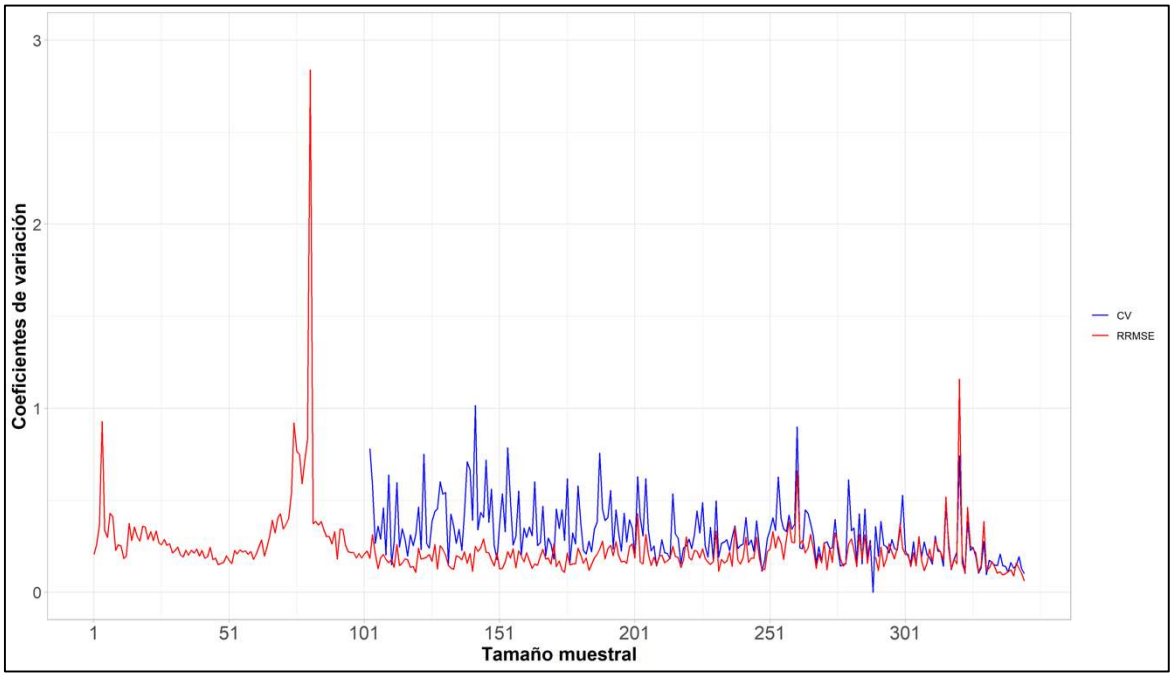
8.5 Coeficiente de variación y RRMSE

Luego de definir el modelo y obtener sus errores cuadráticos medios, se calculó el error cuadrático medio relativo de la raíz (RRMSE) para las estimaciones de Fay-Herriot. Adicionalmente, se obtuvo el coeficiente de variación (CV) para las estimaciones directas. La comparación de ambos estadísticos permite evaluar la ganancia en eficiencia del estimador de área pequeña en relación con el estimador directo.

El RRMSE se define como $\sqrt{\widehat{ECM}(\hat{\theta}_d^{FH})}/\hat{\theta}_d^{FH}$, donde el numerador corresponde a la raíz del error cuadrático medio del modelo para el área d y el denominador a la estimación del modelo para el área d. Por su parte, el CV está dado por $\sqrt{\hat{\psi}_d^2}/\hat{\theta}_d^{DIR}$.

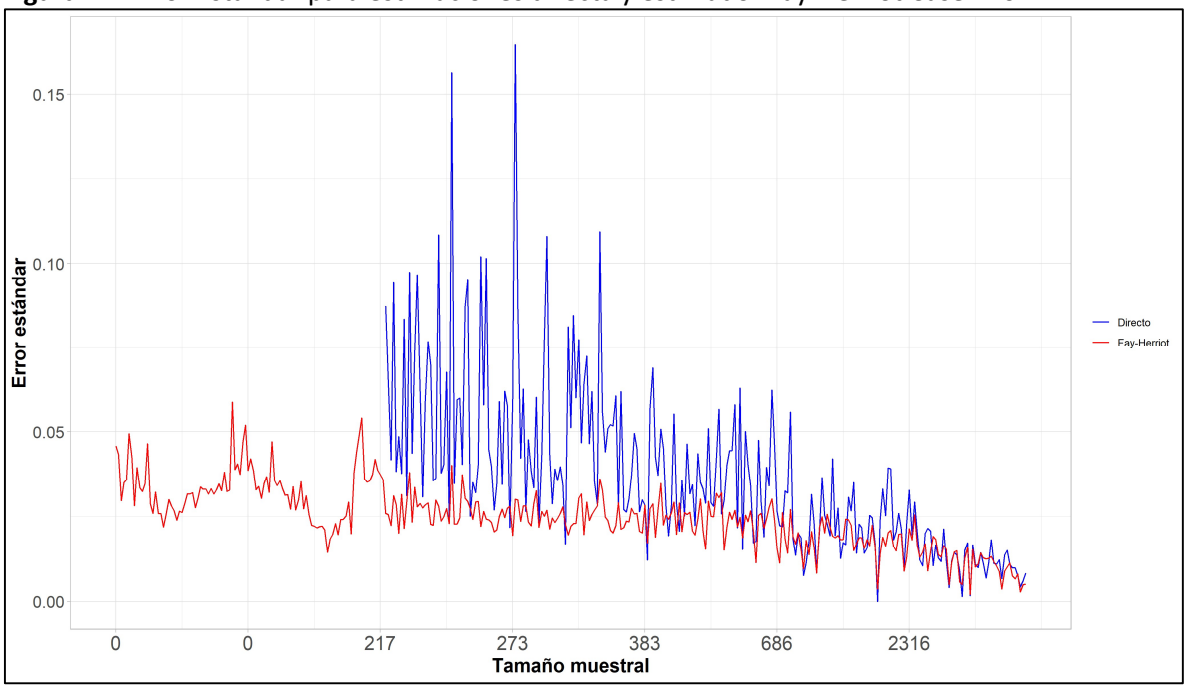
La figura 16 muestra que las estimaciones de Fay-Herriot logran una mejor precisión, lo que es más evidente en áreas con tamaños de muestra más reducidos. Adicionalmente, la figura 17 muestra que el error estándar del modelo Fay-Herriot es en general menor que el de la estimación directa. Esto implica que el modelo Fay-Herriot es más eficiente.

Figura 16: Coeficiente de Variación y RRMSE para estimaciones directa y estimador Fay-Herriot Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

Figura 17: Error Estándar para estimaciones directa y estimador Fay-Herriot Casen 2017



Fuente: Ministerio de Desarrollo Social y Familia - CEPAL

9. Discusión y recomendaciones generales

El convenio de colaboración suscrito entre el Ministerio de Desarrollo Social y Familia y CEPAL tuvo por objetivo hacer una revisión exhaustiva de la metodología de áreas pequeñas implementada por el Ministerio. La asesoría técnica realizada por CEPAL incorporó la revisión teórica de la metodología, al igual que la revisión de las programaciones implementadas. Para el proceso de revisión de la metodología de estimación de áreas pequeñas, se definió un plan de trabajo conjunto basado en el diagnóstico hecho por el Ministerio durante las estimaciones SAE para los años 2015 y 2017.

El resultado de esta consultoría mostró la correcta implementación del modelo Fay-Herriot en las estimaciones comunales de pobreza por ingreso en Chile. No obstante, también permitió realizar una actualización a los últimos estándares en la materia y generó importantes insumos metodológicos, teóricos y computacionales que permiten evidenciar la evolución del proceso de estimación de la pobreza comunal, incorporando procedimientos robustos que complementan los esfuerzos hechos por el Ministerio desde el 2009.

El principal avance realizado consiste en establecer de forma clara que la estimación de áreas pequeñas requiere información desagregada y de calidad, tanto de la estimación directa como de los registros administrativos. En ese sentido, es importante reconocer que no todas las comunas incluidas en Casen poseen un nivel de precisión adecuada para ser utilizadas como insumo en las estimaciones Fay-Herriot. De esta forma, el avance en criterios de inclusión que permitan identificar de manera objetiva la calidad de las estimaciones comunales permite mejorar la calidad de los insumos que entran en el modelo y, por ende, en la calidad final de las estimaciones Fay-Herriot.

Otro avance importante es relevar la selección de las variables auxiliares que entran en el modelo y la importancia que reviste para tener un insumo de calidad por el lado de las estimaciones sintéticas. En esta consultoría se establecieron estándares para evaluar la calidad de la estimación sintética que permiten evaluar el ajuste y la parsimoniosidad del modelo, así como otros aspectos asociados a evitar la multicolinealidad y comprobar que se cumplen el supuesto de ruido blanco en el error.

En años anteriores se había diagnosticado la inestabilidad del efecto diseño a nivel comunal, y a pesar de que había sido abordado mediante distintas estrategias (ver capítulo 5), en esta consultoría se determinó que el uso de una Función Generalizada de Varianza permitía reducir la inestabilidad de la estimación de la varianza directa a nivel comunal. Este método tiene la ventaja de estar muy estudiado en la literatura, ser utilizado en otras aplicaciones de estimación en áreas pequeñas y ser un modelo relativamente simple de estimar.

Las mejoras o actualizaciones numéricas realizadas, como, por ejemplo: la estimación de los errores aleatorios, el recorte de los factores de expansión y la corrección del sesgo al volver las estimaciones a su escala original permiten refinar los aspectos más técnicos de la estimación de áreas pequeñas y evitar posibles sesgos que, aunque se comprobó que eran pequeños en la serie 2009 a 2020, podrían no serlo en estimaciones futuras. De la misma forma la incorporación del paquete EMDI del software R, permite ahorrar tiempo en los pasos más estándares de las estimaciones de Fay-Herriot, permitiendo concentrar

los esfuerzos en la definición del modelo sintético, análisis de la calidad de los datos comunales y estimación de la varianza directa.

La actualización del cálculo de los intervalos de confianza que ahora se basan en la estimación del ECM mediante un procedimiento de bootstrap, permite obtener intervalos de confianza para todas las comunas del país, a la vez que entregan una herramienta para medir el desempeño del estimador de Fay-Herriot. Por último, se definieron las validaciones finales para verificar que se cumplen los supuestos teóricos que asume el modelo Fay-Herriot. La definición de estas validaciones y la estimación del ECM fueron muy valiosas en el sentido que permiten estandarizar la forma en que se evalúa el funcionamiento del modelo final. A futuro, es importante hacer revisiones periódicas a la metodología que permitan ir actualizando los procedimientos a las últimas prácticas que se realicen en la literatura.

Como trabajo futuro para la metodología de áreas pequeñas que implementa el Ministerio, se pueden destacar algunos desafíos relacionados a avanzar en la aplicación de esta metodología para otros indicadores relevantes en el diseño de la política pública, por ejemplo, indicadores relacionados a la infancia o discapacidad. Otro aspecto importante es el estudio de la comparabilidad de la serie de estimación de áreas pequeñas a través de los años.

Aspectos más específicos asociados a la metodología se refieren a la inclusión de nuevas variables administrativas que reflejen la realidad territorial, como por ejemplo información satelital, y analizar la inclusión de efectos espaciales y temporales en las estimaciones.

Referencias

- Burgard, J. P., Münnich, R. and Zimmermann, T. (2016), Impact of Sampling Designs in Small Area Estimation with Applications to Poverty Measurement. In Analysis of Poverty Data by Small Area Estimation (ed. M. Pratesi), 85–108. Hoboken: John Wiley & Sons.

- Casas-Cordero, C., Encina, J. and Lahiri, P. (2016), Poverty Mapping for the Chilean Comunas. In Analysis of Poverty Data by Small Area Estimation (ed. M. Pratesi), 379–403. Hoboken: John Wiley & Sons.

- Datta, G. S. and Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica, 613–627.

- Gambino, Jack G. 2009. "Design Effect Caveats." The American Statistician 63 (2): 141–46. https://doi.org/10.1198/tast.2009.0028.

- Gutierrez, Andres, Hanwen Zhang, and Cristian Montano. 2016. "Cálculo Del Tamano de Muestra Para La Estimacion de Una Varianza En Poblaciones Finitas Con Funciones En r." Comunicaciones En Estadistica 9 (1): 109.

- Hadam, Sandra, Würz, Nora, & Kreutzmann, Ann-Kristin (2020), Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany. Working paper. Freie Universität.

- Hidiroglou, Michel A. 2019. "Development of a Small Area Estimation System at Statistics Canada." Survey Methodology 45 (1): 101–26.

- Jiang, J., Lahiri, P. y Wan, S. (2002), A unified jackknife theory for empirical best prediction with M-estimation. The Annals of Statistics. 30 (6), 1782-1810, https://doi.org/10.1214/aos/1043351257

- Kish, Leslie. 1965. Survey Sampling. John Wiley; Sons.

- Kreutzmann, Ann-Kristin & Pannier, Sören & Rojas-Perilla, Natalia & Schmid, Timo & Templ, Matthias & Tzavidis, Nikos. (2018), The R package emdi for the estimation and mapping of regional disaggregated indicators. Journal of Statistical Software. 91. 10.18637/jss.v091.i07.

- Lahiri, P., y Rao, J.N.K. (1995), Robust estimation of mean squared error of small area estimators, Journal of the American Statistical Association, 90, 758-766.

- Molina, I. (2019), Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas, Series Estudios Estadísticos, No 97, (LC/TS.2018/82/Rev.1), Santiago, Comisión Económica para América Latina y el Caribe, (CEPAL).

- Prasad, N. N. y Rao, J. N. (1990), The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association, 85, 163–171.

- Rao, J.N.K. y Molina (2015), Small area estimation, Second Ed., Hoboken, NJ: Wiley.

- Rivest, Louis-Paul, and Eve Belmonte. 2000. "A Conditional Mean Squared Error of Small Area Estimators." Survey Methodology 26 (1): 67–78.

- Sarndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 2003. Model Assisted Survey Sampling. Springer Science & Business Media.

- Schmid, T., Bruckschen, F., Salvati, N. y Zbiranski, T. (2017), Constructing sociodemographic indicators for national statistical institutes using mobile phone data: Estimating literacy rates in Senegal. Journal of the Royal Statistical Society: Series A, 180, 1163–1190.

- Slud, E. V. y Maiti, T. (2006), Mean-squared error estimation in transformed Fay–Herriot models. Journal of the Royal Statistical Society: Series B, 68, 239–257.

- Sugasawa, S. y Kubokawa, T. (2017), Transforming response values in small area prediction. Computational Statistics & Data Analysis, 114, 47–60.

- Valliant, Richard, Jill A. Dever, and Frauke Kreuter. 2018. Practical Tools for Designing and Weighting Survey Samples. Statistics for Social and Behavioral Sciences. Springer International Publishing. https://doi.org/10.1007/978-3-319-93632-1.

- Wolter, Kirk M. 2007. Introduction to Variance Estimation. 2nd ed. Statistics for Social and Behavioral Sciences. Springer.

ANEXO 1: Variables auxiliares 2020

En este Anexo se presenta una lista de las variables consideradas en la selección del modelo. La mayoría de las variables tienen desagregaciones por edad, nivel educacional, etc. Por razones de extensión, la tabla a continuación muestra las variables sin desagregaciones.

Variable	Desagregación	Fuente
Tasa de matrícula neta	Nivel educacional: parvulario, básica, media, superior	Administrativo
Tasa de matrícula bruta	Nivel educacional: parvulario, básica, media, superior	Administrativo
Tasa rezago población		Administrativo
Tasa de rezago matrícula		Administrativo
Tasa ocupación formal dependiente	Tramos de edad: 15 a 29, 30 a 44, 45 a 59, 60 y más. Sexo: hombres, mujeres	Administrativo
Promedio ingreso imponible de asalariados formales dependientes	Tramos de edad: 15 a 29, 30 a 44, 45 a 59, 60 y más. Sexo: hombres, mujeres	Administrativo
Promedio ingreso imponible de la población en edad de trabajar	Tramos de edad: 15 a 29, 30 a 44, 45 a 59, 60 y más. Sexo: hombres, mujeres	Administrativo
Porcentaje de asalariados formales dependientes con ingreso imponible menor al 50% de la mediana	Tramos de edad: 15 a 29, 30 a 44, 45 a 59, 60 y más. Sexo: hombres, mujeres	Administrativo
Porcentaje de asalariados formales dependientes con ingreso imponible igual al ingreso mínimo	Tramos de edad: 15 a 29, 30 a 44, 45 a 59, 60 y más. Sexo: hombres, mujeres	Administrativo
Mediana del ingreso imponible de los asalariados dependientes	Tramos de edad: 15 a 29, 30 a 44, 45 a 59, 60 y más.Sexo: hombres, mujeres	Administrativo
Proporción afiliados a fonasa	Grupo: (A, B, C, D). Sexo: hombres, mujeres	Administrativo
Tasa de victimización		Administrativo
Mortalidad infantil (por cada mil nacidos vivos)		Administrativo
Años de vida potencialmente perdidos (por cada mil habitates)		Administrativo
Distribución ingresos municipales	Ingresos propios permanentes, Fondo Común Municipal	Administrativo

Variable	Desagregación	Fuente
Ingreso municipal: Ingresos propios permanentes per cápita		Administrativo
Tasa de estado nutricional adulto mayor	Bajo peso, normal, obeso, sobrepeso	Administrativo
Indice de Vulnerabilidad Escolar	Prioridades: primera, segunda, tercera, no vulnerables	Administrativo
Simce	Lenguage, matemáticas, historia. 4 básico, 8 básico	Administrativo
Tasa de estado nutricional niño	Desnutrido, normal, obeso, sobrepeso, riesgo desnutrición	Administrativo
Porcentaje de afiliados al sistema de isapre		Administrativo
Proporción población perteneciente a un pueblo originario (todos)		Censo 2017
Proporción población perteneciente a un pueblo originario (reconocidos por ley)		Censo 2017
Proporción de población nacida en otro país		Censo 2017
Porcentage poblacion urbana/rural		Censo 2017
Porcentage poblacion hombre/mujer		Censo 2017
Promedio anios escolaridad mayores de 15		Censo 2017
Promedio anios escolaridad mayores de 18		Censo 2017
Porcentage poblacion niños		Censo 2018
Porcentage poblacion adulto mayor		Censo 2019
Proporción de viviendas según hacinamiento	Sin hacinamiento, hacinamiento medio, hacinamiento crítico	Censo 2017
Proporción de hogares según hacinamiento	Sin hacinamiento, hacinamiento medio, hacinamiento crítico	Censo 2017
Proporción de viviendas según materialidad	Recuperable, irrecuperable	Censo 2017
Proporción de viviendas segun tipo origen del agua	Red pública, pozo o noria, camión aljibe, vertiente	Censo 2017
Proporción viviendas con déficit cuantitativo	Viviendas irrecuperables, hogares allegados, hogares hacinados	Censo 2017

Variable	Desagregación	Fuente
Tasa de dependencia familiar		Censo 2017
Índice de vejez		Censo 2017

Fuente: Ministerio de Desarrollo Social y Familia- CEPAL