Caracterización Social

Procedimiento de cálculo de la Tasa de Pobreza a nivel Comunal mediante la aplicación de Metodología de Estimación para Áreas Pequeñas (SAE).

OBSERVATORIO SOCIAL

Serie Documentos Metodológicos, Nº1
11 de Febrero de 2013

Ministerio de Desarrollo Social
Gobierno de Chile

Versión Preliminar

Contenido

I. Introducción

Uno de los objetivos del Ministerio de Desarrollo Social es proveer información acerca de la realidad social y económica del país. Para ello, el Ministerio levanta la Encuesta de Caracterización Socioeconómica Nacional (Casen) desde 1987 y ha publicado estadísticas oficiales de la tasa de pobreza nacional, regional y comunal, utilizando métodos estándar para el análisis de datos provenientes de encuestas complejas. El tipo de estimador utilizado presenta propiedades deseables para la producción de estimaciones insesgadas y consistentes a nivel nacional y regional[1][2]. A nivel comunal, sin embargo, la propiedad de consistencia del estimador se va perdiendo (Cochran, 1977; Lohr, 1999; Särndal *et al.*, 1992; Rao, 2003).

Durante el año 2010, el Ministerio de Desarrollo Social convocó a una Comisión de Expertos con el objeto de hacer una revisión exhaustiva de la Encuesta Casen[3]. Una de las recomendaciones de esta Comisión fue que se utilizaran métodos más confiables para la producción de estadísticas a nivel comunal. En las últimas décadas se han desarrollado un conjunto de metodologías que reciben el nombre genérico de Estimaciones para Áreas Pequeñas[4] que cuentan con mejores propiedades para la producción de estadísticas a niveles geográficos menores, a partir de la combinación de datos de encuestas con datos de otras fuentes, como registros administrativos o censales.

Dado el amplio uso y la creciente demanda por estadísticas a nivel comunal, el Ministerio acometió la tarea de producir estadísticas más precisas y exactas a este nivel territorial. Para ello, el año 2011, en conjunto con el Programa de las Naciones Unidas para el Desarrollo (PNUD) en Chile, dio inicio a un proyecto de investigación para el desarrollo de una metodología de estimación para áreas pequeñas a partir de los datos de la Encuesta Casen 2009 y que sería replicable en versiones siguientes. Para tal efecto se contó con la asesoría del experto internacional Partha Lahiri[5], quien ha trabajado en diversos países en el desarrollo e implementación de metodologías de estimación para áreas pequeñas.

Como se presentará más adelante, el principal resultado de este trabajo de investigación, son las importantes ganancias en términos de la precisión del estimador de área pequeña para la producción de estadísticas a nivel comunal. Esto es de particular relevancia, ya que permitirá a los tomadores de decisión contar con insumos de mejor calidad para distinguir territorios que presentan distintos niveles de pobreza.

Este documento fue desarrollado con el objetivo de describir y documentar el proceso de cálculo de las tasas de pobrezas comunales estimadas a partir de Casen 2009 y 2011, utilizando la metodología de estimación para áreas pequeñas desarrollada por el Ministerio[6].

---

[1] En el contexto de inferencia a poblaciones finitas, la propiedades de insesgamiento y consistencia de los estimadores se prueban con respecto al mecanismo probabilístico que genera los datos de la encuesta, es decir, el proceso de muestreo probabilístico asociado al diseño muestral respectivo.

[2] Es importante recordar que las propiedades de insesgamiento y consistencia están asociadas a los estimadores, no a las estimaciones derivadas a partir de los estimadores. Sin perjuicio de lo anterior, las estimaciones derivadas de un estimador más confiable que otro, por ejemplo, son a su vez más confiables. Por esta razón, para facilitar el uso del lenguaje, a lo largo del texto se utiliza indistintamente la terminología en relación a los estimadores o las estimaciones derivadas a partir de ellos.

[3] Ver documento "Informe Final" (Comisión de Técnicos Casen, 2010) en referencias.

[4] En inglés, conocidas por la sigla SAE - *Small Area Estimation*. Ver Rao (2003), Jiang y Lahiri (2006) y Pfeffermann (2002, 2013) para una revisión de las metodologías de estimación para áreas pequeñas.

[5] El Doctor Partha Lahiri es Profesor del Programa Conjunto de Metodología de Encuestas (*Joint Program in Survey Methodology*, JPSM) de la Universidad de Maryland y Profesor del Instituto de Investigaciones Sociales (*Institute for Social Research*, ISR) de la Universidad de Michigan.

[6] Este documento fue elaborado en conjunto por Jenny Encina y Carolina Casas-Cordero (Ministerio de Desarrollo Social) y Rodrigo Herrera (PNUD). Se agradecen los comentarios y revisiones a versiones de este documento

Para conocer las estimaciones de las tasas de pobreza comunal para los años 2009 y 2011, y sus respectivos intervalos de confianza, revisar el documento "Incidencia de la Pobreza a nivel Comunal, según Metodología de Estimación para Áreas Pequeñas. Chile 2009 y 2011"[7]. Para una revisión detallada de la literatura en estimación par áreas pequeñas recomendamos revisar Rao (2013) y Pfeffermann (2002, 2013) y los textos especializados sugeridos a lo largo de este documento[8].

En lo que sigue el documento se organiza de la siguiente manera: la sección II presenta brevemente los problemas de la metodología actual para las estimaciones de nivel comunal; la sección III revisa aspectos metodológicos y la literatura de estimación de áreas pequeñas; la sección IV describe paso a paso el proceso de estimación de la metodología implementada en Chile; la sección V presenta limitaciones de la metodología actual y áreas de investigación futura; y la sección VI presenta las referencias.

II. Diagnóstico: problemas del estimador actual para estadísticas de nivel comunal

En Chile, las estadísticas oficiales de pobreza son estimadas a partir de los datos recolectados en la Encuesta Casen. Esta encuesta 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[9].

Los resultados de toda encuesta están sujetos a errores de muestreo, ya que las estimaciones se basan en datos recolectados a partir de una muestra y no de un censo de la población objetivo. Una muestra permite seleccionar un subconjunto de observaciones que son una réplica aproximada, pero no exacta, de la población total. Estimaciones derivadas a partir de una muestra para un parámetro de interés como la tasa de pobreza son, por tanto, una aproximación del verdadero valor del parámetro, donde la precisión de la aproximación viene determinada por el tipo de estimador utilizado, el tamaño de la muestra, y las características del área de interés. La precisión se mide, generalmente, a través de estimaciones del error de muestreo como el error estándar, el intervalo de confianza o el coeficiente de variación de la estimación[10].

El error de muestreo depende de múltiples factores, sin embargo, bajo una estrategia de estimación dada, el error de muestreo es mayor cuando el tamaño de la muestra es más pequeño. Por ejemplo, cuando se desea producir estimaciones para subgrupos de la población (ej. discapacitados) o para áreas geográficas pequeñas (ej. comunas). A mayor error de muestreo es menor el grado de precisión que se tiene de la estimación de interés.

---

profesionales del Ministerio de Desarrollo Social (Isabel Millán, Alvaro Krause, Alvaro Herrera, Alfredo Martin) y PNUD (Osvaldo Larrañaga y Denisse Falk).

[7] Ver en referencias documento "Incidencia de la Pobreza a nivel Comunal, según Metodología de Estimación para Áreas Pequeñas. Chile 2009 y 2011" (Ministerio de Desarrollo Social, 2013).

[8] Para una mirada sucinta del problema, ver las presentaciones realizada por el Dr. Partha Lahiri en Chile el 18 de Mayo de 2011 (Lahiri 2011a, Lahiri 2011b) en las referencias.

[9] El levantamiento de la Encuesta se realiza de manera periódica y entrega información acerca de la incidencia, magnitud y características de la pobreza. Esta es una encuesta de hogares residentes en viviendas particulares, representativa a nivel nacional, regional y urbano/rural. Para mayores antecedentes sobre la Encuesta Casen y el cálculo de la tasa de pobreza con metodología actual ver "Manual del Investigador: Encuesta de Caracterización Socioeconómica Nacional 2011", Observatorio Social, Ministerio de Desarrollo Social.

[10] La teoría estadística estándar, conocida como *estimación para poblaciones finitas (*Särndal *et al.*, 1992*),* permite estimar estos errores para muestras probabilísticas como las de la encuesta Casen.

La metodología estándar de estimación, diseñada para estimación en áreas grandes, tiene dos importantes limitaciones para la producción de estadísticas en áreas pequeñas:

1. La precisión de las estimaciones se reduce a medida que disminuye el tamaño de la muestra.
2. La falta de precisión en las estimaciones no permite realizar comparaciones confiables entre unidades o entre años para una misma unidad de análisis.

Los tomadores de decisión, tanto públicos como privados, necesitan contar con información de mayor precisión que permitan, por una parte, discriminar territorios que presentan diferencias en sus tasas de pobreza y, por otra, detectar los cambios ocurridos en el tiempo para evaluar adecuadamente el impacto a nivel local de las políticas implementadas.

En las últimas décadas, se han producido importantes avances en el desarrollo de metodologías que permiten combinar datos provenientes de encuestas y datos de registros administrativos y censos para obtener estimaciones más confiables (robustas) a menores niveles territoriales. En la siguiente sección se presenta la metodología de estimación para áreas pequeñas que el Ministerio de Desarrollo Social ha desarrollado con el objetivo de contar con mayor precisión y exactitud en la producción de estimaciones de las tasas de pobreza a nivel comunal.

III. Aspectos metodológicos y literatura sobre estimación en áreas pequeñas

Existen distintos tipos de estimadores que pueden ser utilizados para obtener un parámetro de interés como la tasa de pobreza. En la literatura especializada, el estimador de un parámetro se denomina "directo" si está basado únicamente en los datos de la muestra asociada a un área pequeña determinada. Por otra parte, un estimador es "indirecto" si se basa en información que no está asociada a la muestra de dicha área pequeña. Existe una gran variedad de indicadores, denominados simplemente "sintéticos", que combinan de diversas formas datos de encuestas u otras fuentes (ver en Cochran, 1977; Lohr, 1999; Särndal et al., 1992; Rao, 2003). A continuación se discuten los más relevantes para el caso de la metodología implementada en Chile.

1. Áreas Pequeñas

Un área pequeña es una subpoblación para la cual las estimaciones en base a métodos estándar (estimaciones directas) son inadecuadas, debido a que si la muestra de la subpoblación es pequeña, entonces el estimador directo tendrá una alta variabilidad, lo que hace que éste sea muy impreciso[11]. En este contexto, es posible asimilar como un problema de estimación de área pequeña la situación en la que se encuentran los estimadores directos comunales provenientes de la encuesta Casen. Tal como se señaló en la sección anterior, en muchos casos el estimador directo de pobreza comunal se caracteriza por tener poca precisión.

Los tomadores de decisión necesitan contar con información de la mayor precisión posible, de tal forma que ésta permita, por una parte, discriminar territorios que presentan diferencias en sus indicadores de pobreza y, por otra, detectar los cambios ocurridos en el tiempo para evaluar adecuadamente el impacto a nivel local de las políticas implementadas.

[11] "*Un dominio (área) se considera grande (o mayor) si la muestra asociada a ese dominio es lo suficientemente grande como para producir "estimaciones directas" de precisión adecuada. Un dominio se considera "pequeño" si la muestra asociada a ese dominio no es lo suficientemente grande como para producir estimaciones directas de precisión adecuada".* Rao (2003, pág. 1).

En este contexto, existe en Chile una creciente demanda de diversos organismos gubernamentales y no gubernamentales, tanto a nivel central como a nivel comunal, para que se produzcan estimaciones de mejor calidad sobre una amplia gama de indicadores sociales que permitan evaluar de mejor manera el bienestar en los distintos niveles territoriales y coordinar las acciones requeridas. Por lo tanto, disponer de técnicas de estimación para áreas pequeñas de mayor exactitud y precisión es fundamental si se quiere maximizar la cantidad de información disponible para estas áreas.

La historia de los métodos de estimación para áreas pequeñas se remontan a Inglaterra en el siglo XI y a Canadá en el siglo XVII (Brackstone, 1987). Estos primeros métodos utilizaban en su mayoría datos provenientes de diversos registros administrativos y censos. Con el advenimiento de las nuevas técnicas metodológicas en el desarrollo de encuestas, diferentes organismos gubernamentales han explorado la posibilidad de utilizar datos provenientes de encuestas en la producción de estadísticas asociadas a áreas pequeñas[12]. La principal ventaja de utilizar datos de encuestas, es que éstas permiten producir estadísticas sobre una amplia gama de variables de manera regular y con mayor eficacia.

Los estimadores basados en la teoría estadística estándar generalmente son confiables cuando se dispone de tamaños muestrales grandes. Sin embargo, y como se ha dicho, pueden ser muy poco fiables cuando los datos disponibles son escasos (como puede ser el caso para gran parte de las comunas consideradas en el diseño muestral de la encuesta Casen).

La literatura ha sugerido diversas estrategias de diseño muestral para incorporar factores que influyen en la calidad de los datos en presencia de áreas pequeñas (véase Rao 2003, pp 21-24 y su referencia). Si bien estos cambios en los diseños muestrales pueden mejorar el rendimiento de los estimadores directos provenientes de encuestas en áreas pequeñas, siguen existiendo problemas debido a que algunos de ellos pueden ser muy costosos de implementar.

En un sentido amplio, esto exige una estrategia global para la generación de estimadores de alta calidad en un contexto de áreas pequeñas, que implica utilizar información relevante tanto de encuestas como de registros administrativos y censales, y la aplicación de diferentes técnicas de estimación disponibles para ello. Lo anterior, supone un ahorro de costos ya que la combinación de estos datos sirve como alternativa a la recolección de nueva información a partir de encuestas (que pueden ser costosos) y una reducción en los errores de estimación, ya que se está trabajando a partir de la vinculación de distintas bases de datos existentes. Este es un método indirecto de obtener estimadores asociados a un área pequeña en particular.

2. Estimación en áreas pequeñas

En este contexto, una decisión clave a considerar es la selección del método a utilizar para combinar la información proveniente de las distintas fuentes disponibles. Dependiendo del tipo de problema que se esté analizando, se pueden obtener estimaciones indirectas de áreas pequeñas usando diversas herramientas y técnicas estadísticas[13].

[12] En Estados Unidos, por ejemplo, programas federales que utilizan métodos de estimación de áreas pequeñas incluyen: *Infant and maternal health for states (NCHS); Personal income for states and counties (BEA); Post-census population for counties (USCB); Employment and unemployment for states (BLS); Livestock, crop production for counties (NASS); Disabilities, hospital utilization, physician visits for states (NCHS); Median income for 4-person families for states (USCB); and Prevalence of alcohol abusers and illicit drug users*.

[13] Por ejemplo Empirical Best Linear Unbiased Prediction (E-BLUP), Empirical Bayes (EB) y método Hierarchical Bayes (HB).

Una de las aplicaciones más conocidas de la técnica de estimación en áreas pequeñas es la construcción de mapas de pobreza, que son descripciones espaciales de la distribución de la pobreza generados para distintas unidades geográficas de un país o territorio determinado. Estos mapas son una herramienta útil de visualización para los responsables de la política pública pero también para el público no experto. A los responsables de generar políticas públicas, les permite considerar la dimensión geográfica de la pobreza, sobre todo en unidades geográficas pequeñas como ciudades y comunas.

Siendo el mapeo de la pobreza esencialmente una herramienta asociada a un problema de estimación en áreas pequeñas, diversas aplicaciones de estimación basadas en modelos han sido desarrolladas y aplicadas en diferentes países. Por ejemplo, a través del programa SAIPE[14] (Small Area Income and Poverty Estimates) Estados Unidos utiliza desde 1993 un modelo de estimación para la producción de estadísticas oficiales de pobreza a distintos niveles de agregación geográfica (ej. estados, condados y distritos escolares)[15]. El modelo de estimación utilizado por el programa SAIPE es una variación del modelo propuesto por Fay-Herriot en 1979 (véase Fay-Herriot, 1979) y es actualmente utilizado para estimar las cifras oficiales de pobreza y para asignar fondos públicos entre localidades[16]. En términos generales, el modelo de Fay-Herriot es una técnica estadística que permite calcular estimadores asociados a áreas pequeñas como un promedio ponderado de un estimador sintético y de un estimador directo. Mientras el estimador directo utiliza datos recolectados en encuestas, el estimador sintético se basa en datos provenientes de registros administrativos y censales.

Otro de los métodos conocidos para realizar mapeos de pobreza es el desarrollado por el Banco Mundial (conocido también como método de ELL) desarrollada por Elbers, Lanjouw y Lanjouw (2003). Este método combina información proveniente de encuestas de hogares y censos. El objetivo es contar con estimaciones de pobreza en función de una medida de bienestar como el ingreso o el consumo, pero a mayores niveles de desagregación. El método consiste en estimar modelos de ingreso (consumo) con información proveniente de las encuestas de hogares, considerando como variables explicativas solamente a aquellas que se encuentren también en el censo[17]. El resultado son estimaciones sintéticas de la variable de bienestar de interés (provenientes de la aplicación de un modelo), con diferentes desagregaciones geográficas, aún en zonas donde no se dispone de datos provenientes de encuestas.

Recientemente, Molina y Rao (2010) propusieron un método (Empirical Bayes (best) Method) para generar mapas de pobreza que utiliza datos provenientes de encuestas y datos censales. Este método es bastante similar al que propone ELL, a excepción de la inclusión de un efecto aleatorio específico para cada área pequeña.

[14] Véase, Citro *et al.* (1997), Maples y Bell, 2005; Bell *et al.*, 2007 para algunos detalles sobre la metodología.
[15] El U.S. Census Bureau desarrolla, desde 1997, estimaciones de áreas pequeñas para la mediana del ingreso, el total de pobres, el total de niños pobres menores de 5 años, el total de niños de 5-17 años en familias pobres y el total de pobres menores de 18 años a través de su programa SAIPE (Small Area Income and Poverty Estimates). Ver National Research Council (2000, pág. 1).
[16] Las estadísticas de las tasas de pobreza infantil, producidas por el programa SAIPE del El U.S. Census Bureau, sirven como base para la distribución de fondos de alimentación bajo la legislación *"Improving America´s School Act"*. Esta legislación, conocida como "Título I", regula la asignación de fondos públicos al programa de educación primaria y secundaria más grande en Estados Unidos, responsable de suplementar fondos locales y estatales para localidades con niños de bajo desempeño, especialmente en escuelas de bajos ingresos.
[17] Para una descripción completa de la metodología del Banco Mundial y las aplicaciones del método en diferentes países, consulte Elbers *et al.* (2008), Neri *et al.* (2005), y otras muchas aplicaciones en diferentes países, consulte http://go.worldbank.org/9CYUFEUQ30

El tema central en la generación de mapas de pobreza, al igual que en cualquier otro problema de estimación de áreas pequeñas, es la elección del modelo a utilizar. Es posible que los diferentes métodos de mapeo de la pobreza puedan diferir dependiendo de las necesidades específicas del país. Por ejemplo, si la mayoría de las áreas pequeñas no tienen muestra o la muestra es muy pequeña, y además no existen errores no muestrales atribuidos a la vigencia de los datos del censo o a la comparabilidad de las variables explicativas entre la encuesta y el censo (por ejemplo, Tarozzi y Deaton, 2009), ELL podría ser un método sensato de aplicar. Sin embargo, en presencia de errores muestrales atribuidos a la obsolescencia de los datos muestrales o a discrepancias evidentes en variables del censo y la encuesta, el método de Molina-Rao podría ser una opción, siempre que la mayoría de las áreas pequeñas dispongan de algunos datos muestrales.

A continuación se describe el modelo de Fay-Herriot, modelo que usa actualmente como base el programa SAIPE en Estados Unidos y que sirve también de base para el método desarrollado para Chile.

3. El modelo de Fay-Herriot

En su trabajo de 1979, Fay y Herriot adaptaron el estimador de James-Stein[18] en el contexto de un modelo de dos niveles para estimar el ingreso per cápita para un conjunto de áreas pequeñas, con población inferior a 1.000 habitantes, en Estados Unidos. En términos simples, Fay y Herriot realizan dicha estimación como un promedio ponderado entre un estimador directo y un estimador sintético.

Un estimador sintético "puro" hace fuerte uso de la información de datos administrativos y censales, pudiendo generar poca variación local, lo que puede contribuir a un importante incremento del sesgo de no existir cierta homogeneidad entre las áreas bajo estudio. Por lo tanto, para evitar el potencial sesgo del estimador sintético y la inestabilidad del estimador directo "puro" (discutido en sección II), el método de Fay y Herriot considera como estimador final una combinación lineal entre un estimador directo y uno sintético. La ponderación que toman dichos estimadores viene dada por la variabilidad asociada a cada estimador. De esta manera, si el estimador directo tiene poca varianza en relación al estimador sintético, será mayor su ponderación en la estimación final, y viceversa.

Formalmente, los dos niveles de estimación propuestos en Fay y Herriot (1979) son:

Nivel 1: $$y_i \mid \Theta_i \rightarrow N(\Theta_i, D_i), \quad i = 1, \ldots, m \qquad (1)$$

Nivel 2: $$\Theta_i \rightarrow N(X_i \beta, A), \quad i = 1, \ldots, m \qquad (2)$$

Con m siendo el número de áreas pequeñas.

[18] El estimador de James-Stein es el resultado de "contraer" (*shrink*) todos los promedios individuales de una población bajo estudio hacia un gran promedio (el promedio de los promedios). De esta manera, si el promedio de una unidad en particular es superior al gran promedio, se debe disminuir dicho valor, y si no alcanza el gran promedio, se deberá aumentar. Ver más detalles de la comparación de los estimadores James-Stain y Fay-Herriot en Fay y Herriot (1979).

Notar que este modelo de dos niveles también puede ser escrito como el siguiente modelo lineal mixto:

$$Y_i = \Theta_i + e_i = X_i\beta + v_i + e_i \quad , \quad i = 1,\ldots,m \qquad (3)$$

Donde: $v_i \to N(0, A)$ y. $e_i \to N(0, D_i)$

La ecuación del Nivel 1 señala que para cada área i, Y_i es un estimador de Θ_i (el verdadero valor de la variable de interés) y que tiene una varianza dada por D_i, la cual se asume conocida. La ecuación del Nivel 2 señala que Θ_i tiene una distribución previa, normal con varianza A. En este contexto, el estimador bayesiano de Θ_i es:

$$\Theta_i^{EB} = Y_i^* + (1 - (D_i/(A + D_i))) * (Y_i - Y_i^*) \qquad (4)$$

Como se muestra en la formula anterior[19], se expresa el estimador Θ_i^{EB} como un promedio ponderado de Y_i^* e Y_i, siendo Y_i un estimador directo e Y_i^* un estimador sintético de Y_i que utiliza información auxiliar sobre Θ_i contenida en el vector X_i, donde:

$$Y_i^* = X_i * \left((X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y\right) \qquad (5)$$

Por su parte, Σ es una matriz diagonal con $\Sigma_{ii} = D_i + A$ y entrega la mínima varianza insesgada estimada de $x'_i\beta$, la media previa de Θ_i.

El factor de constricción[20] $(1 - B_i) = (1 - (D_i/(A + D_i)))$, presente en el segundo término de la ecuación (4) y luego también en la ecuación (14), indica que si D_i es más grande con respecto a A, entonces el estimador Θ_i^{EB} se contraerá hacia el estimador sintético Y_i^*, de lo contrario se contraerá hacia el estimador directo Y_i. En concreto, se pondera más al estimador con menos varianza.

En este caso, el vector X_i y la varianza muestral D_i son considerados como conocidos, pero β y A deberán ser estimados a partir de los datos.

[19] Notar que esta fórmula también puede ser expresada como en la ecuación (14):
$\hat{\Theta}_i^{EB} = (A/(A + D_i)) * Y_i + (D_i/(A + D_i)) * Y_i^*$

[20] En inglés "*shrinkage factor".*

Si bien el método original propuesto por Fay y Herriot utiliza una solución iterativa para encontrar el valor del estimador de la varianza A, la literatura reciente ha considerado distintas formas de obtener dicho estimador (véase Li y Lahiri (2009)). Uno de estos métodos corresponde al de máxima verosimilitud, utilizando un ajuste propuesto por Li y Lahiri con el fin de obtener estimaciones estrictamente positivas del valor de A. Las estimaciones de máxima verosimilitud son los valores estimados de los parámetros que maximizan una función de verosimilitud determinada. En otras palabras, estiman los valores de los parámetros bajo los cuales los datos observados habrían tenido la más alta probabilidad de ocurrencia.

En este contexto, Li y Lahiri proponen la estimación del parámetro A mediante la especificación de una función de verosimilitud que es ajustada de la siguiente forma:

$$L_{adj}(A) = L(A)\, A \qquad (6)$$

Donde $L(A)$ es una función de verosimilitud determinada. En este caso, la elección de $L(A)$ corresponde a (*profile likelihood function*):

$$L(A) = C\left|\sum\ \right|^{-1/2} \exp\left\{\frac{1}{2} y' P\ y\right\} \qquad (7)$$

Por lo tanto, reemplazando en (6):

$$L_{adj}(A) = C\left|\sum\ \right|^{-1/2} \exp\left\{\frac{1}{2} y' P\ y\right\} A \qquad (8)$$

Donde C es una constante libre de A, y $P = \sum^{-1} - \sum^{-1} X\left(X'\sum^{-1} X\right)^{-1} X'\sum^{-1}$

El estimador de máxima verosimilitud ajustado de A es obtenido maximizando la función de verosimilitud definida $L_{adj}(A)$ o, equivalentemente, la correspondiente función de verosimilitud en logaritmo natural de A (*log likelihood*), denotada por $l_{adj}(A)$.

Aplicando logaritmo natural a (8) se tiene que:

$$l_{adj}(A) = c - (1/2)\left(\log\left|\sum\ \right| - y' P\ y\right) + \log(A) \qquad (9)$$

Sea $l_{adj}(A)^{(j)}$ la j-ésima derivada de $l_{adj}(A)$ respecto de A, y luego de derivar (9) respecto de A, se obtiene:

$$l_{adj}(A)^{(1)} = \frac{1}{2}\left[y' P^2 y - tr\left(\sum^{-1}\right)\right] + \frac{1}{A} \qquad (10)$$

A partir de lo cual es posible obtener el valor de $\hat{A}$.

Una vez encontrado el valor único de $\hat{A} \geq 0$, es posible estimar el set de betas (β) mediante una estimación realizada por mínimos cuadrados ponderados:

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1}Y \qquad (11)$$

Donde, $\Sigma_{ii} = D_i + A$

Con los coeficientes (β) se obtiene el estimador sintético:

$$Y_i^* = \hat{\beta}X \qquad (12)$$

Finalmente, conociendo el valor de $\hat{A} \geq 0$ y $\hat{\beta}$ es posible calcular el estimador Bayesiano de pobreza $\hat{\Theta}_i^{EB}$ como:

$$\hat{\Theta}_i^{EB} = Y_i^* + (1 - \hat{B}_i) * (Y_i - Y_i^*) \qquad (13)$$

Donde $\hat{B}_i = D_i / (\hat{A} + D_i)$ entrega más peso a la estimación directa si la varianza asociada a la estimación directa (D_i) es relativamente menor que la varianza asociada al modelo ($\hat{A}$).

Notar que la ecuación (13) también puede ser definida como:

$$\hat{\Theta}_i^{EB} = (1 - \hat{B}_i) Y_i + \hat{B}_i Y_i^* \qquad (14)$$

O puesto de otra forma:

$$\hat{\Theta}_i^{EB} = (A/(A + D_i)) * Y_i + (D_i/(A + D_i)) * Y_i^* \qquad (15)$$

Para el caso de Chile, la estimación del parámetro de la tasa de pobreza a nivel comunal, ($\hat{\Theta}_i^{EB}$ estimador Bayesiano de la tasa de pobreza) tiene como punto de partida el resultado de la combinación lineal entre la tasa de pobreza directa (Y_i) -proveniente de la encuesta Casen- y una tasa de pobreza sintética (Y_i^*) -estimada a partir de un modelo lineal que

toma ventaja del uso de información auxiliar proveniente de registros administrativos y datos censales de cada comuna. En la sección IV, se entregan más detalles acerca del proceso completo de estimación utilizado.

4. Caso de Chile

En el contexto chileno, la principal fuente de información relacionada con la pobreza es la Encuesta Casen, cuyo diseño muestral permite tener datos para casi todas las comunas del país. Como se ha señalado, el método de estimación estándar tiene poca precisión para la producción de estadísticas de nivel comunal y se requiere contar con una metodología de estimación que produzca estimaciones más confiables para los niveles de desagregación requeridos. Las metodologías de estimación para áreas pequeñas, basadas en modelos, pueden servir este propósito. Considerando la diversidad de técnicas disponibles, los siguientes requisitos parecen razonables de exigir a los estimadores de áreas pequeñas basados en modelos:

a. El estimador de pobreza basado en un modelo debe maximizar el uso de los datos provenientes de la encuesta;
b. En áreas pequeñas con un tamaño de muestra grande, este estimador de pobreza debiera ser cercano al estimador directo de pobreza proveniente de la encuesta;
c. Cuando se agreguen las estimaciones de pobreza para áreas pequeñas (e.g. comunas) a una zona geográfica mayor (e.g. región), el resultado debiera coincidir con la estimación directa asociada a esta área geográfica mayor, ya que esta última estimación se considera insesgada y más confiable (desde un punto de vista del diseño muestral).

Siguiendo la revisión de métodos de estimación, podemos notar que la estimación realizada por el método de ELL, que se basa enteramente en los datos censales, no cumple el requisito a). El método de Molina-Rao, por el contrario, utiliza datos reales de la encuesta, sin embargo, al igual que el método ELL, la estimación final de la variable de interés requiere la vinculación de los datos de la encuesta a los datos del censo anterior a nivel del hogar, lo que es una tarea difícil en el contexto chileno debido a que no se cuenta con datos censales cercanos a la Encuesta Casen (el último censo disponible corresponde al año 2002). Además, ninguno de estos métodos ha sido probado para analizar si cumplen con b) y c). Por estas razones, metodologías ELL y Molina-Rao resultan no ser apropiadas para aplicarlas al contexto chileno. Es importante notar que las cuestiones de errores ajenos al muestreo planteadas anteriormente pueden ser problemáticas.

Por estas razones, la metodología aplicada por el Ministerio de Desarrollo Social toma como modelo base la desarrollada por Fay y Herriot, pero integra a la vez las mejores prácticas disponibles en la literatura reciente. Es importante mencionar algunas de las principales características de la metodología desarrollada por el Ministerio para estimación de tasas de pobreza a nivel comunal, las que serán detalladas a lo largo del presente documento:

a. La estimación directa proveniente de la encuesta Casen contribuye a la estimación final de tasas de pobreza. Esto representa una clara ventaja en comparación, por ejemplo, con otros métodos de estimación para áreas pequeñas que se basan 100% en estimaciones sintéticas[21].

[21] Ver método de Elbers, Lanjouw y Lanjouw (2003).

b. La contribución de la estimación proveniente de Casen está relacionada con su nivel de precisión. En comunas con alta precisión la estimación directa proveniente de Casen pondera más que la estimación sintética. En comunas con baja precisión, la estimación sintética pondera más que la estimación directa.

c. El método considera protecciones contra fallas del modelo. Se implementan bandas, específicas para cada comuna, que ponen un tope a las predicciones del modelo. El tope es de +/- 1 error estándar (de la estimación directa de pobreza) y permite poner un límite a predicciones fuera de rango para las estimaciones sintéticas[22].

d. Se suavizan los factores de expansión de la encuesta. Esta medida se implementa para evitar que valores extremos[23] en el factor de expansión influencien en forma negativa la contribución de la estimación directa a la tasa de pobreza de áreas pequeñas. Esto, ya que valores extremos del factor de expansión pueden hacer más variables (menos precisas) las estimaciones directas.

e. El método considera las estimaciones regionales de tasas de pobreza como marco de referencia para las estimaciones comunales. El procedimiento utilizado, conocido como benchmarking, sirve dos objetivos. Primero, asegura que las nuevas estimaciones comunales de pobreza sean consistentes con la correspondiente tasa de pobreza regional estimada en forma directa[24]. Segundo, la estimación de los factores de ajuste para que la agregación de las comunas de una región sea consistente con la estimación de pobreza en la región, permiten evaluar la calidad del modelo para cada región – si el modelo es bueno, el factor de ajuste en cada región estará en torno a 1.

[22] Ver más detalles del método en Efron y Morris (1975).
[23] Los valores extremos (*outliers*) corresponden a valores anormalmente extremos o fuera de rango.
[24] La consistencia entre las estimaciones comunales y regionales se refiere, en este punto, a que la suma del total de personas pobres estimados a partir de las tasas de pobreza comunal en una región, utilizando la metodología de áreas pequeñas *es exactamente igual* a total de personas pobres estimado para esa región utilizando el método estándar de estimación para áreas grandes. Más detalles acerca de cómo se implementa este ajuste en la sección 4.11.

IV. Implementación de la metodología de estimación en áreas pequeñas para el caso de las comunas de Chile

El objetivo de esta investigación es desarrollar una metodología que permita producir estadísticas más confiables de la tasa de pobreza comunal en las comunas en la muestra Casen. Con este objetivo, el Ministerio de Desarrollo Social desarrolló una metodología que sigue de cerca los métodos empleados actualmente por el programa SAIPE, llevado adelante por el Bureau del Censo de Estados Unidos, para la producción de estadísticas de ingreso y pobreza que son utilizados, entre otros, para la distribución de fondos federales entre localidades a lo largo de Estados Unidos.

Los pasos seguidos, y elementos considerados, para realizar la aplicación del modelo de estimación de tasas de pobreza comunales para el caso de Chile fueron los siguientes:

- Suavización de factores de expansión;
- Estimación de tasas de pobreza directa (Y_i) [25];
- Construcción de una base de datos a nivel comunal, para el vector de datos administrativos (X_i);
- Transformación de las tasas de pobreza para estabilizar la varianza muestral;
- Derivación de los parámetros $\hat{A}, D_i, B_i$;
- Selección del modelo de área y estimación de los parámetros $\widehat{\beta}$;
- Cálculo de las estimaciones sintéticas de pobreza (Y_i^*);
- Cálculo de las estimaciones Bayesianas de la tasa de pobreza (Θ_i);
- Truncamiento de la estimación Bayesiana;
- Transformación de las estimaciones Bayesianas de la tasa de pobreza a su escala original;
- Cálculo de la "tasa de pobreza SAE" (P_i^{SAE}), mediante la calibración de las estimaciones del nivel comunal al nivel regional;
- Cálculo de los intervalos de confianza de la tasa de pobreza SAE $(IC(P_i^{SAE}))$.

A continuación se describe en detalle cada paso de este proceso. Los gráficos y tablas que se presentan muestran los resultados del proceso de estimación con datos Casen 2009. Los mismos procedimientos se aplicaron para la producción de las estimaciones a partir de los datos de la Encuesta Casen 2011.

[25] Este cálculo se realiza utilizando los factores suavizados calculados en el primer paso.

1. Suavización de factores de expansión

Históricamente, se han producido dos factores de expansión para el análisis de los datos de la Encuesta Casen[26]:

- Los factores comunales se usan para producir estimaciones a nivel de comuna o grupos de comunas, ya que expanden al total de la población que reside en hogares particulares del país en las comunas en la muestra Casen. En la muestra 2009 hay 11 comunas excluidas de la muestra y la población de ellas no está cubierta por los factores comunales. En la muestra 2011 hay 22 comunas excluidas[27].

- Los factores regionales se usan para producir estimaciones a nivel nacional, área y regiones, ya que expanden al total de la población que reside en hogares particulares del país, tanto de las comunas en la muestra como las fuera de la muestra. Es decir, toda la población del país está cubierta por los factores regionales.

Tanto los factores de expansión comunales (*expc*) como los regionales (*expr*) están disponibles en las bases de datos públicas de la Encuesta Casen. Para la producción de las estimaciones directas a nivel comunal se utilizaron los factores de expansión comunal. Antes de proceder con la estimación del modelo, se realiza un análisis estadístico con el fin de suavizar los factores de expansión. Esta medida se implementa para evitar que *valores extremos* (outliers) en el factor de expansión influencien en forma negativa la contribución de la estimación directa a la tasa de pobreza de áreas pequeñas. Para realizar esta suavización se utilizó uno de los métodos en Potter (1993). En síntesis, el método utilizado busca encontrar una distribución de factores de expansión que minimice el error cuadrático medio (MSE) de la variable de interés.

$$\hat{MSE}(\hat{P}_g^T) = \hat{\text{var}}(\hat{P}_g^T) + (\hat{P}_g^T - \hat{P}_g)^2 \qquad (16)$$

Donde *g* corresponde a un grupo o dominio de estimación, donde $\hat{P}_g^T$ y $\hat{\text{var}}(\hat{P}_g^T)$ corresponden a la estimación de la tasa de pobreza y varianza de la tasa de pobreza en el grupo *g* utilizando los factores de expansión truncados y $\hat{P}_g$ corresponde a la estimación de la tasa de pobreza utilizando los factores de expansión originales de la encuesta Casen.

En la práctica, este truncamiento se realizó agrupando a las comunas en g=30 grupos definidos por la variable región y zona (urbano, rural). En cada grupo *g* se definió un factor óptimo K_g :

$$K_g = \sqrt{C_g \frac{\sum w_{jg}^{\ 2}}{n_g}} \qquad (17)$$

[26] Para detalles sobre el proceso de desarrollo de los factores de expansión de las encuestas Casen 2009 y 2011 ver documentos metodológicos en las referencias.
[27] Las comunas no incluidas en la muestra Casen 2011 corresponden a las "áreas de difícil acceso" que no están incluidas en el marco muestral del Instituto Nacional de Estadísticas (INE). Ver detalles en documento de diseño muestral Casen 2011 (Ministerio de Desarrollo Social 2011, págs. 18-20).

Donde w_{jg} corresponde al factor de expansión comunal original asociado a cada individuo *j* en su grupo *g*, n_g es el tamaño de la muestra asociado a cada grupo *g* y C_g es un valor arbitrario 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 probó C=1, 2, … , 40 en cada grupo. Una vez determinado K_g para cada grupo, los factores de expansión son truncados siguiendo la siguiente expresión:

$$w^T{}_{jg} = \begin{cases} w_{jg} & si \quad w_{jg} \le K_g \\ K_g & si \quad w_{jg} > K_g \end{cases} \qquad (18)$$

Donde $w^T{}_{jg}$ corresponde a los factores de expansión óptimos truncados para cada individuo *j* en cada grupo *g*.

La suma de los factores de expansión comunales originales (w_{jg}) corresponde a una estimación del total de la población que reside en viviendas particulares, a Noviembre de 2009, en las comunas en la muestra Casen. El truncamiento de los valores máximos de los factores de expansión originales implica, por lo tanto, que la suma de los factores de expansión truncados (w_{jg}^T) es menor al total de la población de inferencia original. Antes de continuar con el procedimiento, se deben corregir los factores truncados en (18) de manera de poder reproducir los totales poblaciones de interés. El factor de expansión suavizado y corregido w_{jg}^S viene dado por la expresión:

$$w_{jg}^S = w_{jg}^T * ajuste_g \qquad (19)$$

$$ajuste_g = \frac{\sum w_{jg}}{\sum w_{jg}^T} \qquad (20)$$

Donde $\sum w_{jg}$ corresponde a la sumatoria de los factores de expansión originales en cada grupo g y $\sum w_{jg}^T$ corresponde a la sumatoria de los factores de expansión truncados en cada grupo g.

La Tabla 1 presenta información sobre los límites (K_g) utilizados en el truncamiento de factores por grupo, así como el número de observaciones truncadas (frecuencia muestral). En total, 17.815 de los factores de expansión originales fueron truncados, lo que equivale a un 7,2% del total de observaciones en la muestra Casen 2009.

Tabla 1.

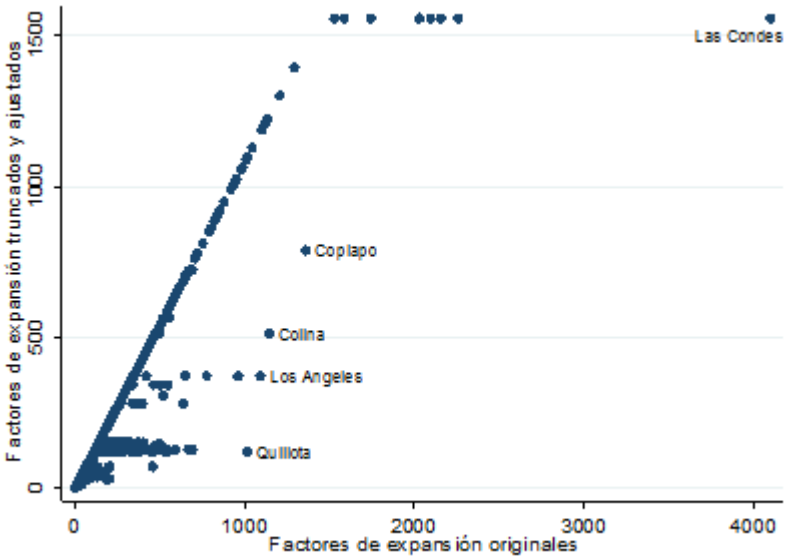
Estadística descriptiva del factor de expansión comunal original, K óptimo y total de factores truncados, según grupo g. Casen 2009.

Grupo g	Media (w_{jg})	Mínimo (w_{jg})	Máximo (w_{jg})	Óptimo K_g	Total de factores truncados (freq.)
1	87,4902	5	501	137,6	748
2	10,1405	2	34	63,0	0
3	94,3949	3	692	672,7	32
4	4,5907	1	27	32,5	0
5	59,8353	6	1.363	731,7	12
6	14,3449	4	47	83,4	0
7	95,7634	7	558	524,6	82
8	27,3082	4	134	127,6	58
9	74,0826	5	1.020	112,6	4.117
10	23,0577	3	100	75,7	105
11	53,7106	2	637	262,0	229
12	22,6640	2	110	87,7	171
13	69,0955	3	405	141,5	1.471
14	26,0215	4	160	49,8	1.760
15	66,7037	4	1.095	346,3	105
16	20,2520	4	203	30,6	2.929
17	60,0223	4	548	318,7	225
18	28,2342	6	183	37,2	2.152
19	70,5936	2	693	118,5	1.547
20	22,8171	3	461	67,9	673
21	37,9711	5	183	52,8	497
22	9,9334	3	34	11,8	328
23	85,7458	8	544	777,8	0
24	9,8642	2	41	14,1	67
25	147,5910	5	4.103	1.445,2	81
26	38,0404	2	1.147	476,6	18
27	53,8487	6	520	287,1	86
28	31,1182	6	237	135,4	54
29	106,1280	1	475	133,2	268
30	14,3313	1	57	103,1	0
Total					17.815

Fuente: Estimaciones propias, Encuesta Casen 2009.

El Gráfico 4 presenta una comparación entre los factores de expansión originales (w_{jg}) y los factores de expansión suavizados (w_{jg}^{S}). Si bien no existen grandes diferencias entre los factores de expansión originales y suavizados, el gráfico permite distinguir aquellas unidades muestrales que se ven afectadas por el truncamiento de su factor de expansión original. A modo de ejemplo, notar que factores de expansión originalmente superiores a 2.000 fueron truncados a un valor cercano a los 1.500.

Gráfico 4
Comparación de factores de expansión comunal
originales (w_{jg}) y truncados/ajustados (w_{jg}^{S})



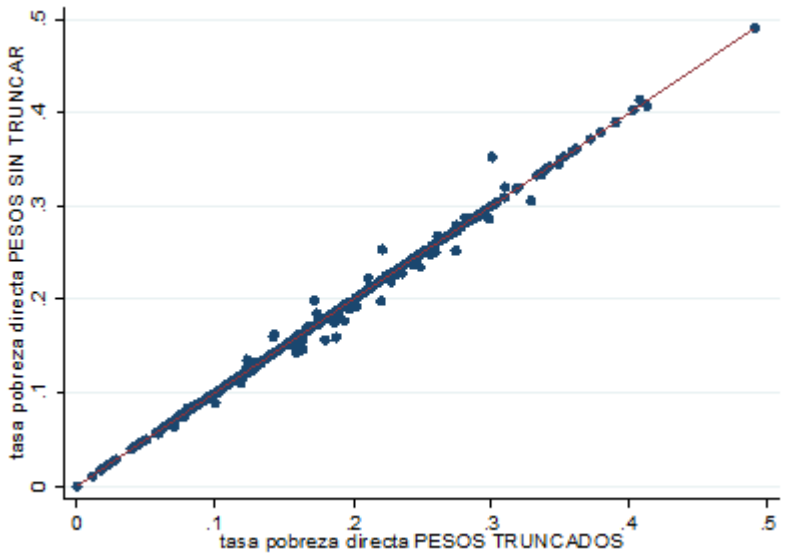
Fuente: Estimaciones propias, Encuesta Casen 2009.

2. Estimación de tasas de pobreza directa (Y_i)

Una vez obtenida la nueva distribución de factores de expansión, truncados y ajustados, se procede a re-estimar las tasas de pobreza y sus respectivas varianzas, tanto a nivel comunal como a nivel regional. En adelante, todos los cálculos y estimaciones son realizados considerado las tasas de pobreza basadas en los factores de expansión comunal truncados y ajustados (w_{jg}^{S}), en vez de aquellas estimadas con los factores de expansión comunal originales (w_{jg}).

El Gráfico 5 muestra la relación que existe entre la tasa de pobreza estimada mediante el uso de los factores de expansión comunal originales y la tasa de pobreza estimada mediante el uso de factores de expansión comunal truncados y ajustados. En general, se observa que para la mayoría de las comunas el cambio en el valor en las tasas de pobreza es marginal.

Gráfico 5
Comparación de Tasas de Pobreza estimada utilizando
factores de expansión comunal originales y truncados/ajustados. Casen 2009.



Fuente: Estimaciones propias, Encuesta Casen 2009.

3. Construcción de base de datos a nivel comunal

Uno de los requisitos fundamentales para realizar la estimación de la tasa de pobreza mediante el método de áreas pequeñas, es disponer de información auxiliar que permita relacionar determinadas características de la comuna con el nivel de pobreza en la comuna. En Chile, esta información pueden provenir de: i) registros administrativos generados por el Estado y ii) datos censales de carácter estructural[28]. En otras palabras, lo que se desea es contar con un conjunto amplio de variables auxiliares que se correlacionen conceptualmente a la situación de pobreza de las comunas, que permita luego realizar un proceso de selección estadístico de aquellas variables que resulten ser más importantes para explicar la variabilidad en las tasas de pobreza entre comunas.

Considerando el conjunto de indicadores que forman parte de este marco, y sobre la base de la información administrativa y censal disponible, se dispuso de un listado inicial de variables asociadas a la pobreza a nivel comunal. Se consideraron como parte de la base de datos exclusivamente variables[29]:

- Asociadas a estado o condición social[30], que reflejaran la situación social de las comunas;
- Provenientes de fuentes administrativas confiables;
- Elaboradas en forma periódica, para contar con actualización en ejercicios futuros.

[28] Dado que el último Censo de población y Vivienda disponible para Chile corresponde al del año 2002, es deseable el uso de variables que no cambien de manera significativa en el corto plazo.
[29] El proceso de identificación y selección de variables a evaluar fue desarrollado en conjunto por los equipos técnicos del ministerio y PNUD. También se realizó consulta a académicos del área.
[30] Describen los resultados sociales que las políticas intentan influenciar, es decir, describen las condiciones generales de la población a nivel de ingresos, salud, educación, empleo, seguridad ciudadana, entorno, etc.

Adicionalmente a las variables y fuentes de datos antes señaladas, la base de datos también incorporó una variable extraída de la misma serie de encuestas Casen - el promedio de las estimaciones directas de la tasa de pobreza para los últimos tres años[31]. El porcentaje de personas que vivieron bajo condiciones de pobreza en una comuna en años anteriores, es probablemente un buen predictor del porcentaje de personas que viven bajo pobreza en periodos siguientes. La estimaciones Casen de un año en particular, sin embargo, son imprecisas para muchas comunas. Para compensar en parte este problema, se optó por utilizar el promedio de los tres últimos años para cada comuna, con la finalidad de reducir en parte el error de medición asociado al uso de las estimaciones puntuales de cada año. En el Anexo 1 de este documento se encuentra la lista de variables consideradas para la elección del modelo (fuente, periodicidad y año de disponibilidad de los datos).

4. Transformación de las tasas de pobreza para estabilizar la varianza muestral

La estimación del modelo de áreas definido en (1)-(4), para el caso de una variable dependiente como la tasa de pobreza, requiere de la aplicación de una transformación que permita estabilizar la varianza de las estimaciones provenientes de la muestra (Carter y Rolph, 1974).

La tasa de pobreza comunal (P_i) se puede estimar, a partir de los datos muestrales, como la proporción de personas pobres en una comuna sobre el total de la población comunal. La varianza de este indicador viene dada por la expresión:

$$\text{var}(P_i) = (P_i)(1-P_i)/(n_i - 1) \qquad (21)$$

Como se puede observar, la varianza de la tasa de pobreza (expresada como D_i en la ecuación (1)) depende de P_i , lo que implica que no es constante sino que varía en función del valor de P_i . Esto viola el supuesto de varianza constante del modelo empírico bayesiano propuesto por James-Stein[32]. Para corregir este problema, diversos autores han propuesto la aplicación de procedimientos que permitan estabilizar la varianza del estimador bajo estudio. Para el caso de una proporción, se propone el uso de la transformación arcoseno sobre la raíz cuadrada de P_i (Carter y Rolph 1974, pág. 882).

Al aplicar la transformación arcoseno sobre la raíz de la tasa de pobreza P_i tenemos que, bajo *muestreo aleatorio simple*, la varianza de la variable transformada ya no depende del valor de la tasa de pobreza. Bajo estas condiciones, el modelo definido en (1)-(4) es aproximadamente válido.

$$Y_i = \sin^{-1}\sqrt{P_i} \qquad ; \qquad Var(Y_i) \cong \frac{1}{4n_i} \qquad (22)$$

[31] Para el modelo 2009 se utilizó el promedio de las estimaciones directas asociadas a los años 2000, 2003 y 2006. Para el modelo 2011 se utilizó el promedio de años 2003, 2006 y 2009. Notar que para el ejercicio con los datos 2011 se utilizó la pobreza directa del 2009, no la pobreza SAE 2009.

[32] El estimador de James-Stein asume $Y_i \rightarrow_{ind} N(\theta_i, D)$, es decir, que las observaciones Y_i se asumen independientes e idénticamente distribuidas, de acuerdo a una distribución normal con media θ_i y varianza D, donde D se asume como un escalar constante conocido. Ver en Fay y Herriot (1979, pág 270).

Para el caso de Chile, las estimaciones de la tasa de pobreza comunal P_i no provienen de un diseño muestral aleatorio simple, sino de un *diseño muestral complejo* que involucra diversos grados de estratificación, conglomeración y probabilidades de selección desigual[33]. La expresión para la varianza en (22), por lo tanto, no aplica directamente al caso de las estimaciones provenientes de la muestra Casen. Una forma relativamente sencilla, y ampliamente utilizada para incorporar el efecto del diseño muestral complejo a una estimación de varianza, es el reemplazo del tamaño de la muestra (n_i) por el *tamaño efectivo de muestra* (Kish, 1965)[34][35].

La expresión que sigue representa una aproximación de la varianza de la transformación propuesta por Carter y Rolph (1974) bajo muestreo complejo, la cual se obtiene reemplazando en la ecuación (22) el *tamaño efectivo de muestra* (m_i) por el tamaño de la muestra (n_i) en la comuna *i*.

$$D_i = Var(Y_i) \cong \frac{1}{4m_i} = \frac{deff_i}{4n_i} \qquad (23)$$

La expresión en (23) se utilizará, por lo tanto, para estimar el parámetro D_i bajo el diseño muestral complejo de Casen.

Para estimar el efecto diseño en cada comuna ($deff_i$), se requiere contar con estimaciones confiables de la varianza de la tasa de pobreza comunal bajo el diseño muestral complejo de la Encuesta Casen. Sin embargo, al igual que para la tasa de pobreza comunal, las estimaciones de la varianza de la tasa de pobreza comunal son poco confiables en áreas pequeñas.

Existen diversas alternativas en la literatura para tratar este problema (ver en Bell, 2008), tales como el uso de aproximaciones del error cuadrático medio (Wang y Fuller, 2003; Rivest y Vandal, 2003), modelamiento de la varianza muestral (Arora y Lahiri, 1997; You y Chapman, 2006; Liu, Lahiry y Kalton, 2007) y desarrollo de funciones generalizadas de varianza (Gershunskaya y Lahiri, 2005; Huff, Eltinge y Gershunskaya, 2002; Cho et al., 2002; Eltinge, Cho y Hinrichs, 2002; Otto y Bell, 2008). Para esta aplicación, se optó por una aproximación sintética simple que consiste en reemplazar la estimación del efecto diseño comunal $(deff_i)$ por la estimación del efecto diseño del nivel regional $(deff_r)$. Esta solución asume que la estimación del efecto diseño regional es una aproximación razonable a la estimación del efecto diseño de cada una de las comunas en la región respectiva[36].

---

[33] Para detalles sobre el diseño muestral de las encuestas Casen 2009 y 2011 ver documentos metodológicos respectivos en las referencias.

[34] El *tamaño efectivo de muestra* (m) se define como la razón entre el tamaño muestral (n) y el efecto diseño asociado a la variable de interés ($deff$). Esta expresión, propuesta en Kish (1965), trata de ilustrar el efecto del diseño muestral en la precisión de las estimaciones relacionando incremento en el efecto diseño a reducción en el tamaño de muestra.

[35] El efecto diseño es un parámetro poblacional que depende del diseño muestral (s) y del estimador de interés (z) y se define como la razón entre la varianza de un estimador bajo un determinado diseño muestral $\mathrm{var}(z)_s$ y la varianza de ese mismo estimador bajo muestreo aleatorio simple $\mathrm{var}(z)_{mas}$ (Kish, 1965).

[36] El uso de esta aproximación sigue un razonamiento estadístico. En áreas grandes (*e.g.* región) los errores de muestreo de los estimadores son suficientemente pequeños para ser escogidos como los mejores estimadores, pero en áreas pequeñas (*e.g.* comunas), es preferible sustituir estimadores sesgados con error de muestreo despreciable

La fórmula para el cálculo del efecto diseño asociado a cada comuna *i* corresponde a:

$$deff_i = deff_{r\in i} = \frac{V_r^*}{\frac{P_r(1-P_r)}{n_r^v}} \qquad (24)$$

Donde la expresión en el numerador corresponde a la varianza de la tasa de pobreza de la región *r* bajo el diseño muestral complejo de Casen, y la expresión en el denominador corresponde a la varianza de la tasa de pobreza bajo muestreo aleatorio simple. En esta última expresión, P_r corresponde a la estimación de la tasa de pobreza directa[37] proveniente de la encuesta Casen para la región *r* y n_r^v corresponde al tamaño muestral, en términos de viviendas (v) encuestadas, en la región *r*. La Tabla 2 muestra el efecto diseño estimado para cada región:

Tabla 2.
Efectos diseño de la tasa de pobreza a nivel regional, utilizando factores de expansión comunal suavizados. Casen 2009.

	Región	Deff
1	Tarapacá	3,280
2	Antofagasta	5,750
3	Atacama	6,477
4	Coquimbo	4,665
5	Valparaíso	3,390
6	O´Higgins	4,307
7	Maule	4,870
8	Biobío	5,506
9	Araucanía	5,618
10	Los Lagos	6,095
11	Aysén	2,843
12	Magallanes	2,323
13	Metropolitana	3,290
14	Los Ríos	8,681
15	Arica y Parinacota	2,864

Fuente: Estimaciones propias, Encuesta Casen 2009.

(estimación regional) por estimadores con errores de muestreo grande (estimación comunal). Ver breve comentario en relación a este tipo de práctica en Fay y Harriot (1979, pág. 269).

[37] Notar que tanto la pobreza regional, así como su varianza, han sido estimadas utilizando los factores de expansión comunal previamente suavizados (ver sección 4.2).

5. Derivación de los parámetros relevantes

Dado que las comunas del país son heterogéneas en cuanto a sus características, observables y no observables, es importante para los efectos de la estimación de tasas de pobreza comunal, obtener estimadores y parámetros robustos, que sean representativos de la verdadera relación que existe entre pobreza y las características asociadas a cada comuna.

Dado que la variable dependiente, la tasa de pobreza, proviene de datos muestrales, la precisión de la estimación disminuye con el tamaño de muestra, lo que puede afectar la robustez de las estimaciones de los parámetros del modelo definido en (1)-(4). Por esta razón, los parámetros $\hat{A}$ y $\widehat{\beta}$ son derivados a partir de un modelo que se estima sólo en base al subconjunto de comunas con más de 10.000 habitantes. Los parámetros resultantes son utilizados, posteriormente, en la predicción de las estimaciones sintéticas de la tasa de pobreza para cada comuna (Y_i^*).

El factor de constricción $(1-B_i)$ se estima para cada comuna *i*, mediante la fórmula $\hat{B}_i = \frac{D_i}{D_i + \hat{A}}$. Mientras que D_i, que se asume conocido, se calcula a partir de la ecuación (23). A continuación viene una breve descripción de la obtención de los parámetros $\hat{A}, D_i, B_i$.

5.1 *Estimación de* $\hat{A}$

La estimación de $\hat{A}$, o varianza asociada al modelo, se realiza mediante el método de máxima verosimilitud ajustado descrito en la sección III. Es importante destacar que se estima un solo valor de $\hat{A} \geq 0$ para toda la población, el que resultó ser 0,00234134 el año 2009.

5.2 *Estimación de* D_i

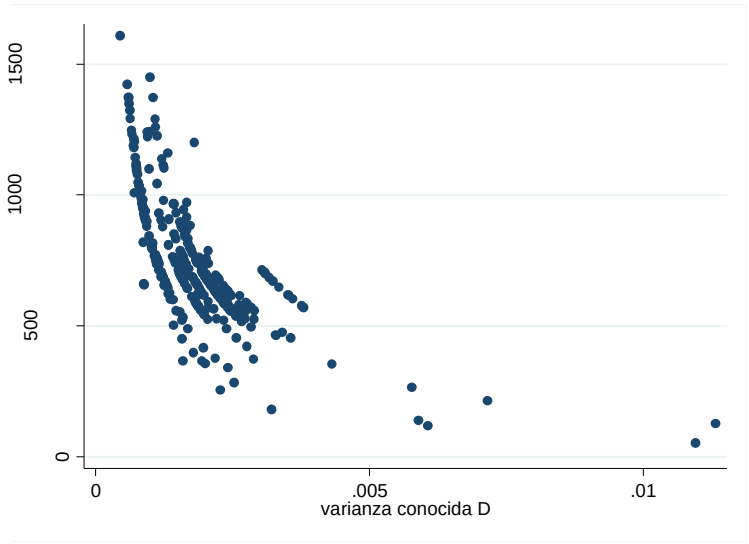
La Tabla 4 presenta estadísticos descriptivas de la varianza asociada a la estimación directa de la tasa de pobreza (D_i). El tramo 1 agrupa a las comunas con menor varianza de la estimación directa, mientras que el tramo 4 agrupa a las comunas que presentaron mayor variabilidad. En promedio, en las comunas de mayor varianza directa, ésta es cercana a tres veces la existente en comunas de menor varianza directa. Es decir, hay una fuerte diferencia en la precisión de los estimadores de la tasa de pobreza entre las distintas comunas del país. El Gráfico 6 muestra que, en la mayoría de los casos, las comunas que tienen un mayor tamaño muestral (en la encuesta Casen) tienen asociado un menor nivel de varianza y viceversa. También se observa que hay comunas que teniendo un mismo tamaño muestral presentan distintos niveles de varianza en su estimador de la tasa de pobreza.

Tabla 4.
Estadísticas descriptivas de la varianza de la tasa de pobreza directa (D_i), según cuartiles de la distribución de D_i.

Cuartiles de la distribución de D_i	Media	mediana	Min	Max
1 = menor D_i	0,0009047	0,0008931	0,0004453	0,0012006
2	0,0014632	0,0015082	0,0012023	0,0016854
3	0,0019475	0,0019501	0,001686	0,0021987
4 = mayor D_i	0,0030616	0,0025722	0,0022082	0,0113181
Total	0,0018417	0,0016857	0,0004453	0,0113181

Fuente: Estimaciones propias, Encuesta Casen 2009.

Gráfico 6.
Relación entre tamaño muestral de la comuna y la varianza del estimador directo de pobreza



Fuente: Estimaciones propias, Encuesta Casen 2009.

5.3 *Estimación de* B_i

La Tabla 5 presenta las estadísticas descriptivas del factor B_i, el cual refleja el peso relativo de las varianzas del estimador directo y el sintético. Este factor se utiliza para estimar la ponderación de los estimadores directo y sintético de cada comuna, en la determinación del estimador Bayesiano de la tasa de pobreza. Cabe recordar que en cada comuna este ponderador se determina a partir del valor de la varianza asociada a la estimación directa de la tasa de pobreza (D_i) y la varianza asociada al modelo (A, que es igual para todas las comunas), y dependiendo de esta relación, la estimación directa de la tasa de pobreza tiene más o menos peso en la tasa de pobreza SAE.

A partir de los resultados presentados, es posible señalar que, el año 2009, cerca de un cuarto de las comunas tiene un ponderador inferior a 50% para la estimación directa de

pobreza. Esto se debe a que en esas comunas el valor de la varianza asociada a la estimación directa D_i es mayor que el de la varianza asociada a la estimación sintética proveniente del modelo $(\hat{A})$. En otras palabras, dado que las estimaciones directas (Y_i) tienen muy poca precisión, se le cree en mayor medida a la estimación proveniente del modelo (Y_i^*).

En promedio, el año 2009, la tasa de pobreza directa recibe una ponderación de 58% y la tasa de pobreza sintética se pondera en 42% para obtener la tasa de pobreza comunal Bayesiana. Estas ponderaciones, sin embargo, varían entre comunas. En la Tabla 5 se observa la relación entre la varianza del estimador directo y la ponderación del estimador sintético. Comunas que tienen menor varianza en su estimación directa de pobreza (tramo 1) tienen asociado un menor ponderador para estimación sintética (B_i) y por lo tanto un mayor ponderador para estimación directa $(1-B_i)$.

Tabla 5.

Estadísticas descriptivas del factor B_i, según cuartiles de la distribución de la varianza de la tasa de pobreza directa (D_i).

Cuartiles de la distribución de D_i	Media	Min	Max
1 = menor D_i	0,2765	0,1598	0,3390
2	0,3836	0,3393	0,4186
3	0,4535	0,4186	0,4843
4 = mayor D_i	0,5474	0,4854	0,8286
Total	0,4150	0,1598	0,8286

Fuente: Estimaciones propias, Encuesta Casen 2009.

6. Selección del modelo y estimación parámetros $\hat{\beta}$

Previo a la selección de las variables a incluir en el modelo de áreas, las variables que fueron evaluadas fueron transformadas respecto de su escala original. Tasas y proporciones fueron transformadas a una escala arcoseno, mientras que variables continuas y de cuenta fueron transformadas a logaritmo (ver sección 4.3).

El proceso de selección del modelo tuvo como objetivo buscar un buen ajuste (medido por el R^2 ajustado) y parsimonia (buscar un mínimo set de predictores)[38]. Con este fin, se utilizó el procedimiento *stepwise* en Stata 11, de manera de reducir el número de variables independientes en atención a su capacidad explicativa del modelo. El modelo de selección de los predictores se estimó con Mínimos Cuadrados Ordinarios. El modelo para la estimación de los parámetros $\hat{\beta}$, sin embargo, se estimó utilizando Mínimos Cuadrados Ponderados,

[38] A modo de referencia, el modelo de área para las estimaciones a nivel de condados que utilizó el programa SAIPE (*Small Area Income and Poverty Estimates*) para la estimación del número de pobres en el 2011 incluyó sólo 5 variables: número de devoluciones de impuestos para aquéllos cuyo ingreso bruto es menor al umbral de pobreza; beneficiarios de "food stamps"; población menor de 18 años; número de exenciones tributarias por tenencia de niños y datos del Censo 2000 de número de niños entre 5 y 17 años en situación de pobreza. Ver detalles en National Research Council (2000).

siguiendo la expresión en la ecuación (11), donde el vector X_i corresponde a las variables finales del modelo.

Las variables finalmente incluidas en el modelo fueron[39]: Remuneraciones promedio de los trabajadores dependientes; tasa de pobreza comunal promedio de 3 últimos años; porcentaje de población rural; porcentaje de población analfabeta y porcentaje de asistencia escolar. Además de esto se incluyeron tres variables mudas (*dummys*) para las regiones séptima, octava y novena.

El modelo estimado por Mínimos Cuadrados Ordinarios entregó un R^2-ajustado de 0,67. El supuesto de normalidad de los residuos se evaluó mediante las pruebas de Shapiro-Wilks, así como la evaluación mediante histogramas y gráficas de *qnorm*. Por otra parte, la presencia de heterocedasticidad se evaluó mediante el análisis de un diagrama de dispersión entre los valores residuales, y los valores ajustados y el coeficiente de correlación de Spearman.

En el Anexo 2 se presenta el modelo de área que sirvió de base para la selección de variables a incluir en la estimación sintética. El Gráfico 1 presentado en el Anexo 3 indica que los residuos tiene un aspecto normal. Por su parte el diagramas de cuantiles (Q-Q plots) presentado en el Gráfico 2, indica que no se observan desviaciones significativas de la hipótesis de normalidad. Lo anterior se ratifica mediante la prueba de Shapiro-Wilks (W=0.99485), que señala que no es posible rechazar la hipótesis nula de normalidad.

El Gráfico 3 del Anexo 3 permite inferir la no presencia de heterocedasticidad, pues la varianza de sus residuos parece ser constante (los residuos no reflejan patrón determinado). Lo anterior se confirma al analizar los coeficientes de correlación de Spearman presentados en la Tabla 1 del citado Anexo. Correlaciones estadísticamente significativas entre el cuadrado de los errores estandarizados y las covariables consideradas en la regresión, podrían ser tomadas como evidencia de heterocedasticidad. Sin embargo, los datos presentados en la Tabla confirman lo que se observa gráficamente - no es posible rechazar el supuesto de homogeneidad en la varianza en cualquiera de las covarianzas analizadas.

Las variables seleccionadas en el modelo se correlacionan conceptualmente a la situación de pobreza:

- La teoría muestra evidencia de que la escolaridad se relaciona con el nivel de ingresos (Mincer, 1958) y por lo tanto con la pobreza, en este sentido la asistencia escolar y analfabetismo nos permiten capturar este efecto y nos permite ver si diferencias en asistencia entre comunas se relacionas a diferencias en pobreza. La variable asistencia también es utilizada en el cálculo del Índice de Desarrollo Humano (IDH) elaborado por el Programa de las Naciones Unidas para el Desarrollo (PNUD) en Chile.

- La tasa de pobreza en Chile se calcula relacionando los ingresos de un hogar con una línea de pobreza, en este sentido, los ingresos están directamente relacionados a la condición de pobreza. Por esta razón, la inclusión de la remuneración de los trabajadores dependientes es una variable relevante para incluir en el modelo de áreas. La variable que promedia las tasas de pobreza de los últimos tres años para cada comuna busca capturar parte de este fenómeno.

[39] Ver en Anexo 1 mayor información de las variables seleccionadas, fuente y fecha de la información.

- La probabilidad de ser pobre puede estar influenciada por el hecho de haber sido pobre en el pasado, esto generalmente se ve reflejado en el término conocido como "pobreza dura". Este tipo de pobreza ha generado el desarrollo de programas tales como "Chile Solidario", "Chile Barrios" y otros. Aunque la evidencia para Chile (Contreras *et al.*, 2004) muestra alta movilidad de las personas en torno a la línea de pobreza (caen y salen de situación de pobreza), también muestra un porcentaje de personas que permanece en situación de pobreza.

- La ruralidad puede ser otro factor determinante en las tasas de pobreza, para el caso de Chile, el ingreso en zonas rurales es inferior al ingreso en zonas urbanas[40] lo que tiene un impacto directo en la pobreza. Para Latinoamérica, aunque el número de pobres se concentra en zonas urbanas, la tasa de pobreza sigue siendo un fenómeno rural (Echeverría, 2000)

7. Cálculo de las estimaciones sintéticas de pobreza $\left(Y_i^*\right)$

Una vez obtenido el conjunto de Betas asociados a las variables administrativas se procede a estimar una tasa de pobreza sintética (Y_i^*) para cada comuna. El estimador sintético se obtiene a partir de la expresión en la ecuación (12).

8. Cálculo de las estimaciones Bayesianas de la tasa de pobreza $(\hat{\Theta}_i^{EB})$

El estimador bayesiano de la tasa de pobreza resulta del promedio ponderado de la tasa de pobreza directa $\left(Y_i\right)$ y la tasa de pobreza sintética $\left(Y_i^*\right)$. Las ponderaciones de los estimadores directo y sintético para cada comuna, utilizadas en la determinación de $\hat{\Theta}_i^{EB}$, se realizan utilizando el factor de constricción $(1-B_i)$, como se ilustra en la ecuación (14).

9. Truncamiento de la estimación Bayesiana de la tasa de pobreza

Cuando el modelo considerado en la determinación del estimador bayesiano es razonable, se espera que este estimador $(\hat{\Theta}_i^{EB})$ tenga un mejor desempeño que el estimador directo (Y_i). Sin embargo, como se hace mención en Fay-Herriot (1979), los estimadores Bayesianos pueden presentar un desempeño adecuado *en general*, pero un desempeño pobre para algunos de los componentes *en particular*. Aplicado al contexto chileno, esto quiere decir que el modelamiento que beneficia a *la mayoría* de las comunas puede ser inapropiado para *algunas* comunas. Esta situación se puede dar, por ejemplo, por mala especificación del modelo, datos administrativos con error de medición distinto entre comunas, o por la presencia de valores extremos (outliers).

[40] Según datos de Encuesta casen 2011, el ingreso autónomo per cápita es 60% mayor en zonas urbanas. (150.980 en zona rural versus 242.184 en zona urbana)

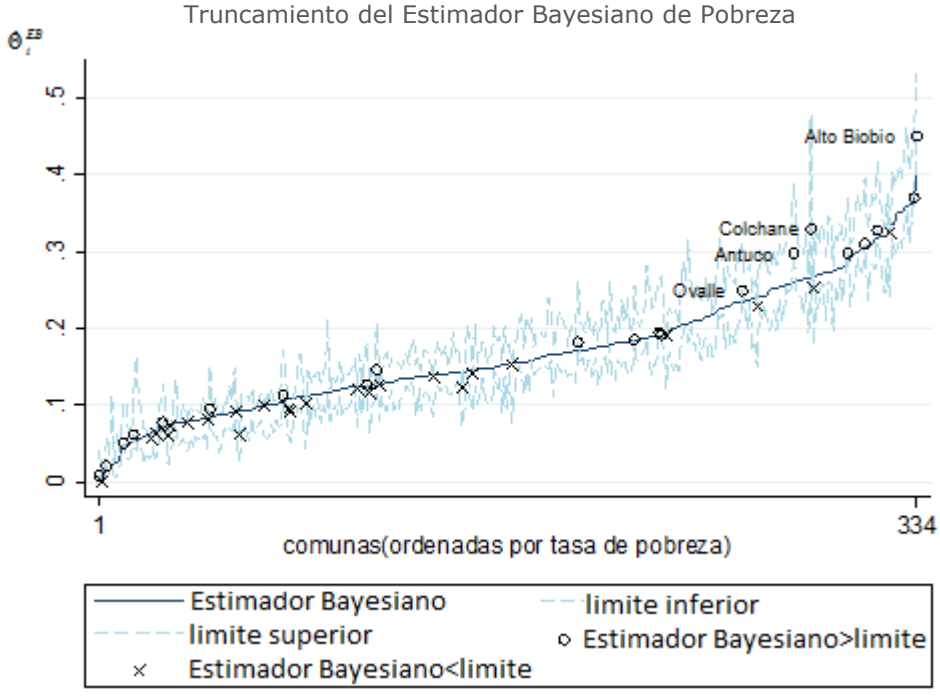
Para corregir por las potenciales fallas del estimador Bayesiano $(\hat{\Theta}_i^{EB})$, se sugiere utilizar aplicar la siguiente corrección (Efron y Morris, 1972; Fay y Herriot, 1979):

$$\hat{\Theta}_i^{EB} = \begin{cases} \hat{\Theta}_i^{EB} & si \quad Y_i - c\sqrt{D_i} \le \hat{\Theta}_i^{EB} \le Y_i + c\sqrt{D_i} \\ Y_i - c\sqrt{D_i} & si \quad \hat{\Theta}_i^{EB} \le Y_i - c\sqrt{D_i} \\ Y_i + c\sqrt{D_i} & si \quad \hat{\Theta}_i^{EB} \le Y_i + c\sqrt{D_i} \end{cases} \qquad (25)$$

Donde c es una constante, definido como c=1 para las aplicaciones del programa SAIPE (EE.UU.) y también para la aplicación en Chile. La idea general de este truncamiento es limitar la desviación del estimador bayesiano $\hat{\Theta}_i^{EB}$ del estimador directo Y_i. Esto reduce la perdida de eficiencia en áreas o comunas donde el modelo falla, sin perder los beneficios que provoca en aquellas áreas o comuna donde éste funciona bien.

La aplicación de este criterio para el caso chileno afecta, en la estimación 2009, a 45 comunas: en 21 comunas la estimación bayesiana supera en un error estándar o más la estimación directa Y_i, y en 24 comunas es inferior en un error estándar o más de tal estimación directa. El Gráfico 7 presenta a las comunas ordenadas según el valor que tiene su estimador bayesiano de tasa de pobreza posterior a la aplicación de este criterio. Para cada comuna, el Gráfico también ilustra el valor que tienen los límites presentados en la ecuación (25). Se marca a través de círculos aquellas comunas donde la estimación bayesiana es mayor que la estimación directa en un error estándar ó más, y a través de cruces donde es menor en un error estándar o más (ambos símbolos marcan el valor de truncamiento).

Gráfico 7
Truncamiento del Estimador Bayesiano de Pobreza



Fuente: Estimaciones propias, Encuesta Casen 2009.

A excepción de algunos pocos casos, el truncamiento de las estimaciones fue relativamente marginal, ya que como se observa en el gráfico, la mayoría de las cruces y círculos están cerca de la línea continua (color negro) y lejos de las bandas (color gris) que marcan las desviaciones por sobre y por debajo de 1 desviación estándar.

Sin perjuicio de lo anterior, en algunas comunas el truncamiento adquiere importancia. Para las tres comunas con mayor truncamiento, la diferencia entre tasa de pobreza directa y bayesiana (antes de truncar) fue de aproximadamente 10 puntos porcentuales. Después del truncamiento, para estas mismas tres comunas, la diferencia se redujo a aproximadamente 5 puntos porcentuales. Estos datos sugieren que la estimación sintética de pobreza en estas comunas difiere de manera importante de la estimación directa, y que además tiene una ponderación mayor que esta última. Sin embargo, la estimación Bayesiana de pobreza se encuentra fuera de rango respecto de la estimación directa por lo que se prefiere acotar dicho valor previniendo cualquier posible error en las variables administrativas que determinan la tasa de pobreza sintética.

10. Transformación de las estimaciones Bayesiana de la tasa de pobreza a su escala original

Una vez obtenido el estimador bayesiano de las tasas de pobreza comunales, ellos deben ser transformados a su escala original. Lo anterior se realiza mediante la siguiente fórmula:

$$P_i^{EB} = \sin(\hat{\Theta}_i^{EB})^2 \qquad (26)$$

Donde P_i^{EB} corresponde al estimador bayesiano de la tasa de pobreza en su escala original.

11. Cálculo de la tasa de pobreza SAE (P_i^{SAE})

Una propiedad deseable para las estimaciones de un fenómeno de interés es que ellas sean consistentes a distintos niveles de agregación. Esto quiere decir, para el caso de la tasa de pobreza por ejemplo, que las estimaciones de porcentaje de personas pobres en una región X coincidan – independientemente de si la estimación se realiza en forma directa a nivel regional o bien a partir de los estimadores bayesianos derivados para las comunas en la región. Dicho de otra forma, dado que las estimaciones directas a nivel regional son confiables, se espera que las estimaciones bayesianas del nivel comunal sean consistentes con la estimación regional correspondiente.

Con esta finalidad, se realiza un ajuste final a las estimaciones bayesianas de pobreza comunal con el fin de imponer una consistencia lógica a los resultados. El ajuste, propuesto en Fay y Herriot (1979), consiste en hacer coincidir los niveles de pobreza regional obtenidos mediante la estimación Bayesiana $\left(P_i^{EB}\right)$ con la estimación regional directa desde la encuesta $\left(Y_r\right)$.

El ajuste aplicado corresponde a la razón entre el número de pobres en la región obtenido de Casen (estimación directa, Y_r) y aquél que se obtiene para las comunas en la región usando la metodología de áreas pequeñas (estimación Bayesiana, P_i^{EB}). De esta forma, el ajuste viene dado por:

$$R_r = \frac{Y_r \quad N_r}{\sum_{i=1}^{com} P_i^{EB} \quad N_i} \qquad (27)$$

Donde Y_r corresponde a la estimación directa de tasas de pobreza regional (estimada utilizando el factor de expansión regional original, *expr*), N_r corresponde al total de la población en la región *r*[41], y $\sum_{i=1}^{com} P_i^{EB} \; N_i$ corresponde a la sumatoria del número de pobres en las comuna en la región *r* (P_i^{EB} multiplicado por el tamaño poblacional de la comuna)[42].

Realizando el ajuste especificado, se tiene que la tasa de pobreza final (en adelante denominada tasa de pobreza SAE) para cada comuna es:

$$P_i^{SAE} = P_i^{EB} \times R_r \qquad (28)$$

La Tabla 7 presenta los factores de ajustes para cada región del país estimados, para el año 2009, mediante la ecuación (27). En promedio estos factores están en torno a uno, lo que implica que las estimaciones comunales de pobreza SAE son consistentes con la correspondiente tasa de pobreza regional estimada en forma directa, considerando a este último un dato confiable para usar como marco de referencia. Esto permite señalar que el modelo implementado en la estimación de las tasas de pobreza SAE funciona de buena manera en la mayoría de las regiones[43].

[41] Cifras de la población en la región provienen de estimaciones del INE.

[42] Es importante destacar que el ajuste (calibración) regional, requiere que todas las comunas del país posean una estadística de pobreza, pues de otra forma, la pobreza de comunas no incluidas en la muestra sería asignado al resto de las comunas de la región, que sí son parte de Casen. El procedimiento de áreas pequeñas se realiza sólo a las comunas con muestra Casen (334 comunas para 2009). Para obtener cifras de pobreza para el resto de comunas sin representación en Casen, el Ministerio utiliza un modelo de componentes principales, mediante el cual identifica grupos o conglomerados de comunas con similares características (sobre la base de datos del Censo de Población y Vivienda 2002). Realizada tal agrupación, asigna a comunas sin representación en Casen, el promedio de la tasa de pobreza comunal de las comunas en el conglomerado al cual pertenecen.

[43] Un coeficiente cercano a 1 indica que el agregado de tasas de pobreza bayesiana a nivel regional es coincidente con el dato regional extraído de la encuesta Casen, por lo que el método implementado permite obtener estimaciones consistentes con lo observado para niveles geográficos confiables (estimadores más precisos).

Tabla 7.

Factores de ajuste por región. Casen 2009.

	Región	R_r
1	Tarapacá	1,12172
2	Antofagasta	0,97455
3	Atacama	1,06685
4	Coquimbo	1,04309
5	Valparaíso	1,00387
6	O´Higgins	1,00430
7	Maule	1,05292
8	Biobío	0,99010
9	Araucanía	1,01628
10	Los Lagos	1,04088
11	Aysén	1,06255
12	Magallanes	0,97368
13	Metropolitana	0,97765
14	Los Ríos	1,08572
15	Arica y Parinacota	0,99486
Total		1,014982

Fuente: Estimaciones propias, Encuesta Casen 2009.

12. Derivación de los intervalos de confianza

En este apartado se describe la creación de intervalos de confianza para el estimador de pobreza SAE $\left(P^{SAE}\right)$. Estos intervalos son obtenidos a partir del estimador Bayesiano de pobreza antes de transformar a la escala original, esto es Θ_i^{EB} . Es importante recalcar la necesidad de construir estos intervalos de confianza debido a que, si bien las estimaciones SAE son más precisas que las estimaciones directas de tasas de pobreza comunal, éstas siguen teniendo asociado un nivel de imprecisión.

12.1. Problemas para estimar la precisión del estimador Bayesiano Θ_i^{EB}

El estimador Bayesiano de la tasa de pobreza Θ_i^{EB} es una combinación lineal de dos estimadores que tienen asociado cierto nivel de imprecisión, por lo tanto, un desafío adicional consiste en producir una medida que permita reflejar el grado de imprecisión que también existe en la estimación Bayesiana[44].

Existen varios estimadores que permiten reflejar el grado de imprecisión de un estimador de interés (z), entre ellos el error estándar (SE(z)), el coeficiente de variación (CV(z)) y los intervalos de confianza (IC(z)). Para el caso de Chile se optó por desarrollar estimaciones

[44] El método original propuesto por Fay y Herriot no consideró la construcción de intervalos de confianza para el estimador $\hat{\Theta}_i^{EB}$.

del intervalo de confianza, ya que es un estimador de variabilidad relativamente más conocido y de más fácil comunicación.

En la literatura de áreas pequeñas, la variabilidad de estimadores derivados de modelos lineales mixtos (*empirical best linear unbiased predictor*, EBLUP) - como el estimador Bayesiano desarrollado para Chile - se mide tradicionalmente a partir del valor medio del cuadrado del error de predicción (*mean squared prediction error*, MSPE). En concreto, las estimaciones de intervalos de confianza típicamente siguen la forma $EBLUP \pm z_{\alpha/2}\sqrt{MSPE}$, donde MSPE es un estimador del verdadero MSPE del EBLUP y $z_{\alpha/2}$ es el valor crítico superior 100(1-α/2) de la distribución normal estándar.

Chatterjee, Lahiri y Li (2006) discuten los problemas asociados a este tipo de métodos para la estimación de los intervalos de confianza y proponen un enfoque de bootstrap paramétrico para estimar la distribución completa de un estimador centrado y escalado en forma apropiada[45]. El método propuesto produce intervalos de predicción altamente precisos[46]. En consecuencia, se utiliza este método para la producción de intervalos de confianza para el estimador Bayesiano de las tasas de pobreza comunal.

12.2. El método Chatterjee, Lahiri y Li para estimar los intervalos de confianza del estimador Bayesiano Θ_i^{EB}

A continuación se resumen en seis pasos para la aplicación del cálculo de los intervalos de confianza del estimador Bayesiano de la tasa de pobreza para las comunas de Chile. Para detalles de la implementación del método y la revisión de las propiedades del método en contraste con los otros métodos disponibles ver Chatterjee, Lahiri y Li (2006).

Paso 1: Generar, para cada comuna i, cinco mil muestras bootstrap (*k*=1, …,5000) de los estimadores Θ_i e y_i, utilizando los parámetros $\widehat{A}$, D_i, y $\hat{\beta}$ derivados en las secciones IV.5 y IV.6. Las muestras se generan siguiendo estas expresiones:

$$\tilde{\Theta}_i \rightarrow N(x'_i\,\hat{\beta}, \widehat{A}), \quad i=1,\ldots,m \tag{31}$$

$$\tilde{y}_i \mid \tilde{\Theta}_i \rightarrow N(\tilde{\Theta}_i, D_i), \quad i=1,\ldots,m \tag{32}$$

Paso 2: Re-calcular los parámetros $\hat{\hat{A}}$, $\hat{\hat{B}}_i$ y $\hat{\hat{\beta}}$ para cada muestra de $\tilde{y}_i$ generada en el Paso 1. Con estos parámetros calcular para cada muestra la estimación Bayesiana de pobreza $\hat{\tilde{\Theta}}_i^{EB} = (1-\hat{\tilde{B}}_i)\tilde{y}_i + \hat{\tilde{B}}_i\,(x'_i\,\hat{\tilde{\beta}}\,)$, donde $\hat{\tilde{B}}_i = D_i/(\hat{\tilde{A}}+D_i)$ y D_i corresponde al calculado en la

[45] *"Las probabilidades de cobertura de este tipo de intervalos pueden converger al valor nominal de (1-α), pero los intervalos no son eficientes, en el sentido de que tienen problemas de subcobertura o sobrecobertura, dependiendo de la elección ´particular del estimador MSPE. Más precisamente, el error de cobertura de este tipo de intervalos es de orden $O(n^{-1})$ o mayor, lo cual no es suficientemente preciso para la mayoría de las aplicaciones de los estudios de áreas pequeñas, muchos de los cuales involucran tamaños de muestra pequeños"* (Chatterjee, Lahiri y Li 2006, pág. 1222).

[46] Resultados de simulaciones en Chatterjee, Lahiri y Li (2006) demuestran la superioridad del método de bootstrap paramétrico por sobre otras técnicas existentes utilizadas en la construcción de intervalos de confianza en el contexto de modelos lineales mixtos.

sección IV.5.2. Al término de este paso, se habrán generado cinco mil valores de $\tilde{\hat{\Theta}}_i^{EB}$ para cada comuna *i*.

Paso 3: Calcular t_i siguiendo la expresión (33) para cada muestra, donde $\tilde{\Theta}_i$ proviene de las muestras de bootstrap generadas en el Paso 1; $\tilde{\hat{B}}_i$ y $\tilde{\hat{\Theta}}_i^{EB}$ fueron calculadas en el Paso 2 y D_i corresponde al calculado en la sección IV.5.2. Al término de este paso, se habrá generado una distribución de 5.000 valores de t para cada comuna *i*.

$$t_i = (\tilde{\Theta}_i - \tilde{\hat{\Theta}}_i^{EB}) / \sqrt{D_i(1-\tilde{\hat{B}}_i)} \qquad (33)$$

Paso 4: Ordenar de menor a mayor los valores de t calculados en el Paso 3 para construir una distribución empírica de t para cada comuna *i*. Esta distribución se utiliza para encontrar los valores críticos de t para cada comuna *i*. Los valores críticos que se buscan corresponden al percentil 2,5 (t_{i1}) y 97,5 (t_{i2}) de la distribución de t para cada comuna *i*.

Los paneles (a), (b) y (c) en el Gráfico 8 presenta la distribución de t para tres comunas en la muestra Casen 2009. Para la comuna Puchuncaví, se observa que la distribución de t es relativamente simétrica en torno a cero, por lo tanto los valores de los umbrales de corte asociados a los percentil 2,5 (t_{i1}) y 97,5 (t_{i2}) serán de magnitudes relativamente parecidas. Comunas como Providencia y Peumo, sin embargo, presentan distribuciones asimétricas en torno a cero, lo cual dará origen a valores de corte de magnitudes distintas para t_{i1} y t_{i2}.

Paso 5: Calcular los intervalos de confianza, utilizando la expresión (34), a partir de valores (t_{i1}, t_{i2}) obtenidos del Paso 4 para cada comuna *i*:

$$I(t) = \left\{ \hat{\Theta}_i^{EB} \pm t\sqrt{D_i(1-\hat{B}_i)} \right\} \qquad (34)$$

Donde $\hat{\Theta}_i^{EB} = (1-\hat{B}_i)Y_i + \hat{B}_i\, x'_i\, \hat{\beta}$ es el estimador empírico Bayesiano de la tasa de pobreza estimado a partir de la ecuación (25) en la sección IV.5.9 (calculado a partir de la muestra original); $\hat{B}_i = D_i/(\hat{A} + D_i)$ es el factor estimado en la sección IV.5.3 (calculado a partir de

la muestra original); y los umbrales de corte t_{i1} y t_{i2} provienen de la expresión (33) (calculados a partir de las 5.000 muestras bootstrap).

Paso 6: Transformar las estimaciones del límite inferior y el límite superior del intervalo de confianza, calculados a partir de la expresión (34), a la escala original de las tasas de pobreza, utilizando la transformación arcoseno:

$$IC(P_i^{SAE})_{\alpha=5\%} = \sin\left[I(t_{i1})\right]^2 = \sin\left[\hat{\Theta}_i^{EB} - t_{i1}\sqrt{D_i(1-\widehat{B}_i)}\right]^2 \quad (35)$$

$$IC(P_i^{SAE})_{\alpha=95\%} = \sin\left[I(t_{i2})\right]^2 = \sin\left[\hat{\Theta}_i^{EB} + t_{i2}\sqrt{D_i(1-\widehat{B}_i)}\right]^2 \quad (36)$$

Como cabe esperar de este procedimiento *bootstrap* paramétrico, las estimaciones de los intervalos de confianza resultantes tienen cierto grado de asimetría. En la Tabla 8 se ilustra este punto para las tres comunas en el Gráfico 8. La tabla presenta las estimaciones de la tasa de pobreza SAE (columna 2), los límites inferior y superior del intervalo de confianza asociado a la tasa de pobreza SAE (columnas 1 y 3), y el largo del intervalo de confianza[47] superior e inferior de la tasa de pobreza SAE (columnas 4 y 5).

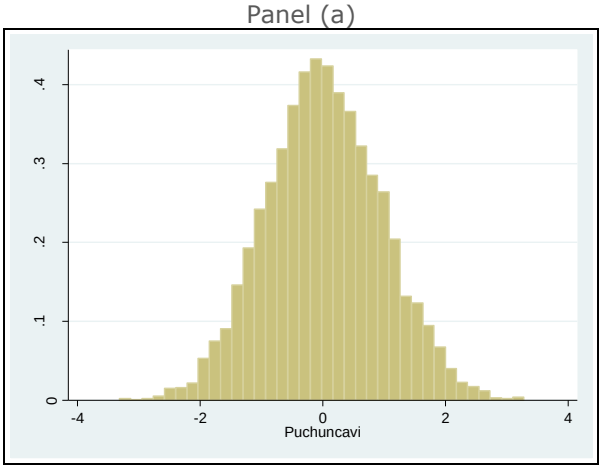
La comuna de Puchuncaví, que en el gráfico 8 presenta una distribución de t relativamente simétrica, el largo de los intervalos de confianza es de similar magnitud para el límite inferior (=-3,4pp) y el límite superior (=4,00pp). Las comunas de Providencia y Peumo, por otra parte, presentan largos del intervalo de confianza bastante asimétricos. El límite superior del intervalo de la comuna de Providencia tiene mayor precisión (=0,10pp) que el límite inferior (=-0,60pp). Lo contrario pasa para la comuna de Peumo, donde el límite inferior tiene mayor precisión (=-1,60pp) que el límite superior (=7,30pp). Estos resultados son consistentes con lo que se observa en el gráfico 8 para estas comunas, donde la comuna de Providencia presenta menor dispersión en la cola derecha de la distribución de t lo que luego se traduce en mayor precisión en la estimación del intervalo superior del intervalo de confianza respectivo. El caso contrario se observa para la comuna de Peumo.

Es importante recordar esta particularidad del procedimiento de estimación de los intervalos de confianza del estimador de *áreas pequeñas*, ya que es mucho más común encontrar intervalos de confianza simétricos en la inferencia estadística para *áreas grandes*.

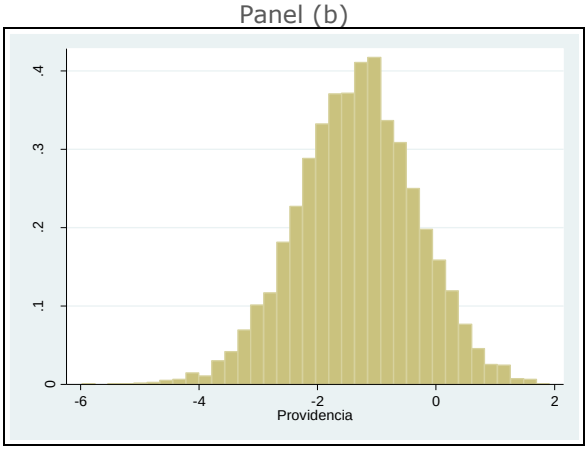
[47] El largo del intervalo de confianza corresponde simplemente a la diferencia entre el límite del intervalo y el estimador de la tasa de pobreza. En la tabla 8, para el largo inferior corresponde a Columna (1) – Columna (2), y para el largo superior corresponde a Columna (3) – Columna (2).

Grafico 8
Distribución de estadísticos t, bootstrap (k=5.000)

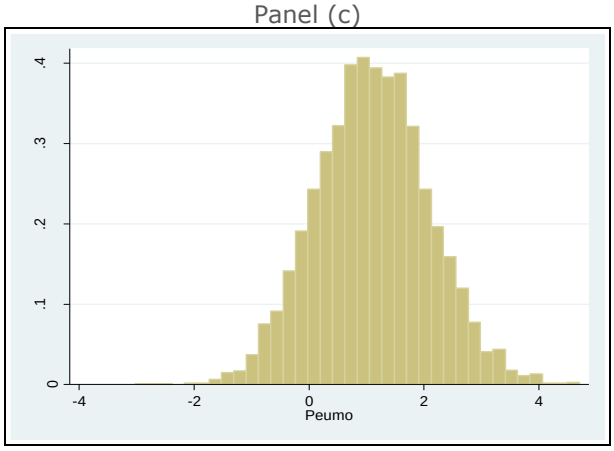
Panel (a)



Panel (b)



Panel (c)



Fuente: Estimaciones propias, Encuesta CASEN 2009.

Tabla 8.
Estimador Bayesiano, Intervalos de confianza y Largo del Intervalo de confianza de la Tasa de Pobreza SAE. Casen 2009.

Nombre Comuna	Límite inferior IC (1)	Tasa pobreza SAE (2)	Límite superior IC (3)	Largo del IC Inferior (4)	Largo del IC superior (5)
Puchuncaví	10,7%	14,1%	18,1%	-3,40pp	4,00pp
Providencia	**-0,4%**	0,2%	0,3%	-0,60pp	**0,10**pp
Peumo	13,1%	14,7%	**22,0%**	**-1,60**pp	7,30pp

Fuente: Estimaciones propias, Encuesta Casen 2009.
Nota: pp quiere decir puntos porcentuales.

V. Limitaciones e Investigación Futura

El método aquí propuesto utiliza el *estado del arte* en metodología de estimación empírica Bayesiana, la cual ha demostrado mejorar las estimaciones directas en numerosos estudios (*e.g*. Efron y Morris, 1975; Fay y Herriot, 1979; *Bell et al.*, 2007). Al igual que en dichos estudios, hemos notado mejoras similares en el contexto Chileno en términos de reducción del largo de los intervalos de confianza para la mayoría de las comunas en las muestras Casen 2009 y 2011. Estos resultados resultan prometedores, sin embargo hay aspectos de la metodología desarrollada que podrían beneficiarse de un programa continuo de investigación y evaluación: (1) mejorar la estrategia para estimar la varianza asociada al estimador directo comunal, (2) mejorar la especificación del modelo de la tasa de pobreza comunal, (3) identificar mejores variables auxiliares para el modelo comunal, (4) evaluar el desempeño de los resultados del modelo lineal mixto en comparación con otros modelos de estimación para áreas pequeñas (e.g. método ELL, métodos no pareados, y otros en Rao (2003)), y (5) desarrollar estimaciones de áreas pequeñas para otras variables de interés, por mencionar algunos.

Los estimadores empíricos Bayesianos, a diferencia de los estimadores directos, se basan en modelos mixtos y por lo tanto existe una necesidad de monitorear continuamente la metodología y desarrollar investigación que permita evaluar y actualizar los métodos aquí desarrollados. Este tipo de evaluación debiera ser productivo para el desarrollo de un sólido programa de producción de mapas de pobreza local en Chile. En esta línea, es importante recordar que mejoras en la producción de estadísticas comunales no solo depende del uso de metodologías de estimación adecuadas -- los resultados del proceso de análisis son condicionales al diseño de la muestra. Esto quiere decir, por lo tanto, que se requiere avanzar en el desarrollo de una *estrategia integral* para la estimación en áreas pequeñas en la cual se deben dedicar esfuerzos tanto a la etapa de *diseño muestral* como a la de *estimación*. Estrategias de este tipo se describen en Singh *et al.* (1994), Marker (2001) y Choudhury *et al.* (2012).

Para finalizar, cabe hacer una reflexión sobre las limitaciones y los usos de las estimaciones aquí desarrolladas. Las estimaciones derivadas a partir de metodología de estimación para áreas pequeñas, al igual que las estimaciones directas, están sujetas a errores muestrales y no muestrales. El objetivo de esta investigación ha sido mejorar la precisión de las estimaciones comunales. Los nuevos estimadores desarrollados, aunque más robustos que los anteriores, también están sujetos a cierto nivel de imprecisión. Es importante considerar este hecho al momento de hacer uso de estos resultados, especialmente si los mecanismos de asignación de recursos han sido diseñados para utilizar las estimaciones puntuales, y no sus errores muestrales asociados. En Chile, al igual que en otros países, las estimaciones de pobreza a nivel local son utilizadas en procedimientos de asignación de recursos entre localidades. Existe, por lo tanto, la necesidad de desarrollar reglas robustas para la distribución de recursos a nivel local, de manera que las diferencias dentro de los márgenes de error sean consideradas como tales, sea al comparar estimaciones entre comunas en un mismo período, o al contrastar tasas de pobreza estimadas para una misma comuna en distintos años. Para avanzar en posibles estrategias de asignación de presupuesto entre localidades, que usan como insumos estimaciones en base a muestras pequeñas, se sugiere revisar la discusión al respecto disponible en: "*Statistical issues in allocating funds by formulas, Panel on Formula Allocations*", Thomas *et al*. (Editors), National Research Council (2003).

VI. Referencias

1. Literatura sobre estimación para áreas pequeñas

Arora, Vipin and Lahiri, (1997), On the Superiority of the Bayesian Method Over the BLUP in Small Area Estimation Problems, Statistica Sinica.

Bell, W.R. (2008), "Examining Sensitivity of Small Area Inferences to Uncertainty about Sampling Error Variances," Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association, pp 327-334.

Bell (1997). Models for county and state poverty estimates. Preprint, Statistical Research Division, U. S. Census Bureau.

Bell, W., Basel, W., Cruse, C, Dalzell,L., Maples,J., OHara,B. and Powers,D. (2007), Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties, Census Report.

Brackstone, G.J. (1987), Small area data: policy issues and technical challenges, in Small Area Statistics, (R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh eds.) 3-20, Wiley, New York.

Carter, G. and Rolph, J. (1974), Empirical Bayes methods applied to estimating fire alarm probabilities, Journal of the American Statistical Association 69, 880-885.

Chatterjee, A., Lahiri, P. and Li, H. (2008), Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models, The Annals of Statistics 36, 1221-1245.

Cho, Eltinge, Gershunskaya and Huff, (2002), Evaluation of generalized Variance Function Estimators for the US Current Employment Survey. Proceedings of the American Statistical Association, Survey Research Methods Section.

Choudhry, G.H., Rao, J.N.K., and Hidiroglou, M.A. (2012), On sample allocation for efficient domain estimation, Survey Methodology, 38,

Cochran, W. (1977), Sampling Techniques. 3d edition. New York: Wiley.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999), Hierarchical Bayes estimation of unemployment rates for the U.S. states, Journal of the American Statistical Association, 94, 1074-1082.

Efron, B. and Morris, C. (1975), Data analysis using Stein's estimator and its generalizations, Journal of the American Statistical Association 70, 311-319.

Elbers, C., J. Lanjouw, and P. Lanjouw, "Micro-Level Estimation of Poverty and Inequality," Econometrica 71:1 (2003), 355–364.

Elbers, Chris, Peter Lanjouw, and Philippe George Leite, "Brazil within Brazil: Testing the Poverty Map Methodology in Minas Gerais," World Bank policy research working paper no. 4513 (2008).

Esteban, M.D., Morales, D., Pérez, A. and Santamaría, L. (2011), Small area estimation of poverty proportions under area-level time models, Computational Statistics and Data Analysis. DOI: 10.1016-j.csda.2011.10.015.

Eltinge, Cho and Hinrichs, (2002), Use of generalized Variance Functions in Multivariate Analysis. Proceedings of the American Statistical Association, N°74.

Fay, R.E., and Herriot, R.A. (1979), Estimates of income for small places: An application of James_Stein procedure to census data, Journal of the American Statistical Association 74, 269_277.

Gabler, S., Haeder, S., and Lahiri, P. (1999), A model-based justi_cation of Kish's formula for design e_ects for weighting and clustering, Survey Methodology, 25, 105-106.

Gershunskaya, J. and Lahiri, P. (2005), Variance Estimation for Domains in the US Current Employment Statistics program. Proceedings of the American Statistical Association, Survey Research Methods Section.

Haslett S. and Jones, G. (2006) Small area estimation of poverty, caloric intake and malnutrition in Nepal. Nepal Central Bureau of Statistics / World Food Programme, United Nations / World Bank, Kathmandu, September 2006, 184pp, ISBN 999337018-5.

Huff, Eltinge and Gershunskaya, (2002), Exploratory Analysis of generalized Variance Function models for the US Current Employment Survey. Proceedings of the American Statistical Association, Survey Research Methods Section.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. Test 15 1–96. MR2252522.

Kish, (1965), Survey Sampling. John Wiley & Sons, Inc., New York, London.

Liu, Lahiry and Kalton, (2007), Hierarchical Bayes modeling of Survey Weighted Small Area Proportions. Proceedings of the American Statistical Association, Survey Research Methods Section.

Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. J. Multivariate Anal. 101 882–892. MR2584906

Lohr, S (1999), Sampling: Design and Analysis (Advanced Series). Pacific Grove, CA:Brooks/Cole Publishing

Maples, J., and Bell, W.R. (2005). Evaluation of school district poverty estimates: Predictive models using IRS income tax data. Proceedings of the Survey Research Methods Section, American Statistical Association.

Marker, D.A. (2001), Producing small area estimates from national surveys: methods for minimizing the use of indirect estimators, Survey Methodology, 27, 183-188.

Molina, I. and Rao, J.N.K. (2010), Small area estimation of poverty indicators, Canadian Journal of Statistics, 38, 369-385.

National Research Council (2000). Small-area estimates of school-age children in poverty, in: Evaluation of Current Methodology. Citro, C. and Kalton, G. (Eds.), National Academy Press, Washington DC.

National Research Council (2003). Statistical Issues in Allocating Funds by Formula. Panel on Formula Allocations. Thomas A. Louis, Thomas B. Jabine, and Marisa A. Gerstein, Editors. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Otto and Bell, (2008), Bayesian Assessment of Uncertainty in Seasonal Adjustment with Sampling Error present. Research report 92/12, Statistical Research Division, US Census Bureau.

Pfeffermann, D. (2002). Small area estimation—new developments and directions. International Statistical Review 70 125–143.

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. Statist. Sci. Volume 28, Number 1 (2013), 40-68.

Potter F.J. (1993), The effect of weight trimming on nonlinear survey estimates; 1993, San Francisco, CA: American Statistical Association.

Rao, J. N. K. Small Area Estimation, Wiley, New York, 2003.

Rivest and Vandal, (2003), Mean Squared Error Estimation for Small Areas When the Small Area variances are estimated. Proceedings of the International Conference on recent Advances in Survey Sampling, ed. J.N.K. Rao.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994), Issues and strategies for small area data, Survey Methodology, 20, 3-22.

Särndal, Swensson and Wretman, (1992); Model Assisted Survey Sampling. New York: Springer-Verlag.

Tarozzi, A. and Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. Review of Economics and Statistics, 91, 773–792.

Wang, Fuller, and Wayme (2003). The Mean Squared Error of Small Area Predictors Constructed with Estimated Error variances. Journal of the American Statistical Association, N°98.

Ybarra, Y.M.L and Lohr, S. (2008), Small area estimation when auxiliary information is measured with error, Biometrika, 95, 919-931.

You and Chapman, (2006), Small Area Estimation Using Area Level Models and estimated Sampling Variances. Survey Methodology, N°32.

2. Referencias a documentos de trabajo y otros

Contreras, Cooper, Herman and Neilson. (2004), Dinámica de la Pobreza y Movilidad Social: Chile 1996-2001. Departamento de Economía, Universidad de Chile.

Echeverría, R. (2000), Options for rural poverty reduction in Latin America and the Caribbea, Cepal Review N° 70.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures, Econometrica, 52, 761–766.

Mincer, J. (58), "Investment in Human Capital and personal income distribution", Journal of Political Economy, August 1958.

Neri, L., Ballini, F. and Betti, G. (2005). Poverty and inequality in transition countries. Statistics in Transition, 7, 135–157.

Comisión de Técnicos Casen (2010), "Informe Final", Octubre 2010.

Lahiri, P. (2011a). "Estimación de Áreas Pequeñas: Parte I". Seminario Estimación de Áreas Pequeñas, Santiago, 18 mayo de 2011.

Lahiri, P. (2011b). "Estimación de Áreas Pequeñas: Parte II". Seminario Estimación de Áreas Pequeñas, Santiago, 18 mayo de 2011.

Ministerio de Planificación (2010). "Informe Metodológico Casen 2009".

Ministerio de Desarrollo Social (2012). "Metodología del Diseño Muestral y Factores de Expansión Encuesta de Caracterización Socioeconómica Nacional (Casen) 2011". Serie Documentos Metodológicos Casen 2011 Nº1.

Ministerio de Desarrollo Social (2013). "Incidencia de la Pobreza a nivel Comunal, según Metodología de Estimación para Áreas Pequeñas. Chile 2009 y 2011". Serie Documentos Metodológicos Caracterización Social Nº1.

VII. ANEXOS

Anexo 1
Variables contenidas en la base de datos comunal

Variable	Descripción	Periodicidad	Ultimo año disponible	Fuente
Remuneraciones promedio de los trabajadores dependientes	La renta imponible o remuneración imponible corresponde a la renta informada en la planilla de pago de cotizaciones. En el caso que existan dos cotizaciones en un mes para una misma relación laboral, se suman las rentas informadas.	Anual	2010	Subsecretaría del trabajo, Ministerio del Trabajo
Razón de analfabetos respecto a la población de 10 y más años en la comuna	El cálculo del indicador "razón de analfabetismo" consistió en dividir el número de población analfabeta de 10 años y más por la población total de 10 años y más.	Decenio	2002	Censo de Población y Vivienda, Instituto Nacional de Estadísticas (INE)
Promedio SIMCE 4º y 8ª básico lenguaje y matemáticas		Anual y año por medio	2008	Prueba SIMCE, Ministerio de Educación
Porcentaje de capital humano avanzado respecto a la población comunal	Considera profesionales universitarios y técnicos.	Decenio	2002	Censo de Población y Vivienda, INE
Porcentaje de vivienda social	Expresa el porcentaje de casas o departamentos cuyo avalúo fiscal es de hasta 520 UF, según señala la Ordenanza General de Urbanismo y Construcción.	Anual	2010	Observatorio Habitacional, Ministerio de Vivienda y Urbanismo (Minvu)
Esperanza de vida al nacer	Es una estadística que indica la media de la cantidad de años que vive una determinada población en un cierto periodo de tiempo.	Anual y quinquenal	2000-2005*	Depto. Epidemiología, Ministerio de Salud (Minsal)
Tasa de mortalidad infantil	Defunciones de menores de 1 año por mil nacidos vivos.	Anual	2008	DEIS, (Minsal)
Índice de SWARROP	También llamado tasa de mortalidad proporcional. Tasa empleada en demografía y epidemiología para comparar la mortalidad entre poblaciones con diferente estructura por edades. Proporción de fallecimientos entre personas de 50 años o más por cada 100 defunciones totales.	Anual	2008	DEIS, (Minsal)

Variable	Descripción	Periodicidad	Ultimo año disponible	Fuente
Años de vida potencialmente perdidos	Sería bueno incorporar explicación	Anual	2008	DEIS, (Minsal)
Porcentaje de sobre peso, obesidad y retraso en talla en niños/as menores de 6 años	Considera población que se atiende en el sector público de salud.	Anual	2010	DEIS, (Minsal)
Porcentaje de desnutrición, sobre peso, obesidad en adultos mayores		Anual	2010	DEIS, (Minsal)
Fecundidad declarada en menores de 19 años	Qué se entiende por fecundidad: embarazos o niños nacidos de madres adolescentes?	Decenio	2002	Depto. Demografía, INE
Presupuesto comunal per cápita		Anual	2009	SINIM, Subsecretaría de Desarrollo Regional
Índice de dependencia familiar	Estimación indirecta del número de personas mayores de edad que depende de sus hijos. Cuociente entre los mayores de 79 años y la población del grupo de edades de 40 a 44 años, por cien	Anual	2010	Depto. Demografía, INE
Índice de vejez	Número de adultos mayores de 60 años por cada 100 niños/as menores de 15 años	Anual	2010	Depto. Demografía, INE
Tasa de migración comunal	Movimientos de población, tanto de recepción de habitantes de una comuna distinta a la de residencia, como de expulsión, considerando el período 1997-2002.	Decenio	2002	Depto. Demografía, INE
Tasa de crecimiento natural de la población comunal	Nacimientos menos defunciones registradas en la comuna.	Decenio	1992-2002	Depto. Demografía, INE
Porcentaje población con puntaje FPS menor a 8.500 respecto a población comunal		Anual	2010	Ficha de Protección Social, Ministerio de Desarrollo Social
Cobertura aplicación FPS respecto población comunal		Anual	2010	Ficha de Protección Social, Ministerio de Desarrollo Social
Porcentaje de	Porcentaje de población rural	Decenio	2002	Censo de

Variable	Descripción	Periodicidad	Ultimo año disponible	Fuente
Población Rural	que posee la comuna respecto al total de población comunal. Se entiende por población rural a toda aquella población que vive en asentamientos de menos de 2000 habitantes, o en aquellos que poseen menos de 1000 (Población Rural / Población Comunal Estimada para el Año)			Población y Vivienda, Instituto Nacional de Estadísticas (INE)
Porcentaje de Asistencia Escolar Comunal	A mayor valor del indicador, mayor será el nivel de asistencia de los alumnos de colegios municipales de la comuna. Asistencia Promedio Mensual Establecimientos Municipales de Educación / Promedio Mensual de niños matriculados	Mensual	2012	SINIM, Subdere
Tasa de pobreza histórica	Promedio de pobreza 3 años anteriores disponibles.	Bi anual o trienal	2011	Encuesta Casen, Ministerio de Desarrollo Social

Anexo 2

Resultados de la regresión por Mínimos Cuadrados Ordinarios (MCO)

Variable dependiente ARCOSENO de la tasa de pobreza comunal (directa) Y_i

Comunas de más de 10.000 habitantes, Casen 2009

	Betas MCO ($\widehat{\beta}$)	Betas estandandarizados. MCO
Remuneraciones promedio de los trabajadores dependientes (log)	-0,09575646 3,52**	-0,21927953 3,52**
% pobreza histórica (arcsin)	0,49548266 7,92**	0,48474029 7,92**
% población rural (arcsin)	-0,13409847 4,96**	-0,39252745 4,96**
% población analfabeta (arcsin)	0,40349163 2,57*	0,25176513 2,57*
% asistencia escolar (arcsin)	-0,21883535 2,23*	-0,0938032 2,23*
región 7 (=1)	0,03442978 2,11*	0,08671043 2,11*
región 8 (=1)	0,03882056 2,67**	0,12474226 2,67**
región 9 (=1)	0,105632 6,04**	0,28328927 6,04**
Constante	1,61477028 4,24**	-0,00203088 0,06
Observaciones	235	235
R-cuadrado ajustado	0,67	0,67

Estimaciones realizadas usando Mínimos Cuadrados Ordinarios (MCO).
Se presentan valores de t-estadístico. * variable estadísticamente significativa con un 95% de confianza; ** variable estadísticamente significativa con un 99% de confianza.

Anexo 3

Gráfico de residuos del modelo de regresión

Gráfico 1
Distribución errores estandarizados (línea azul)

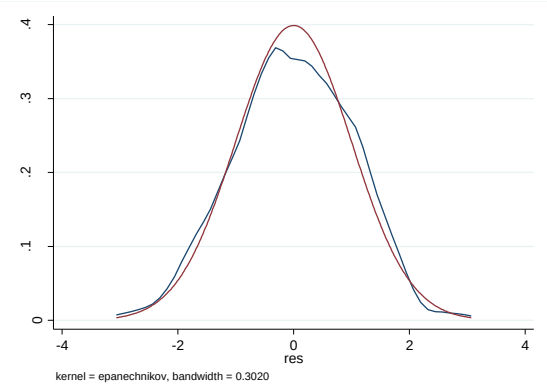
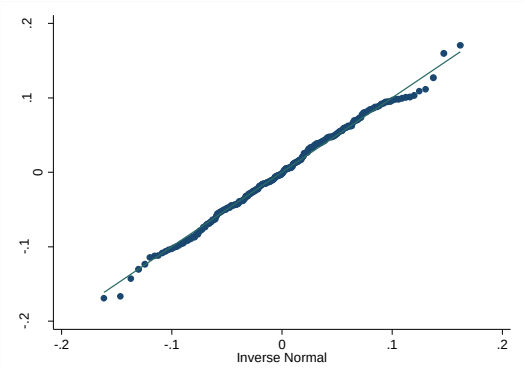
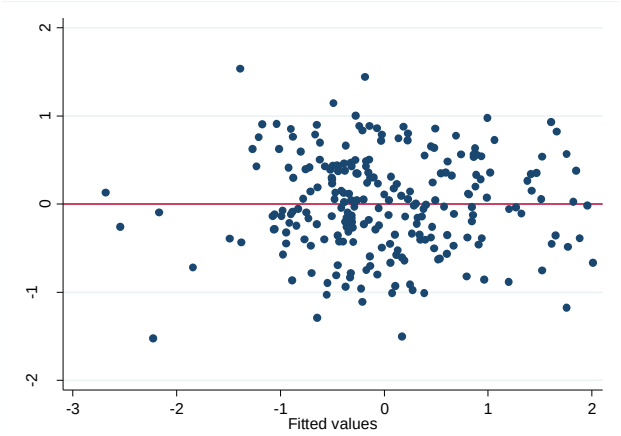


Gráfico 2
Cuartiles de residuos estandarizados vs cuartiles de una distribución normal



Fuente: Estimaciones propias, Encuesta Casen 2009.

Gráfico 3
Homocedasticidad de los residuos



Fuente: Estimaciones propias, Encuesta Casen 2009.

Tabla 1
Coeficientes de correlación de Spearman (ρ) para el cuadrado de los residuos estandarizados

Covariables	ρ de Sperarman	t-estadístico asintótico
Remuneraciones promedio de los trabajadores dependientes	-0.0144	0.8264
% pobreza histórica	-0.0148	0.8214
% población rural	-0.0065	0.9214
% población analfabeta	-0.0092	0.8882
% asistencia escolar	0.0072	0.9126
región 7	0.0337	0.607
región 8	-0.095	0.1467
región 9	-0.0061	0.9256

Fuente: Estimaciones propias, Encuesta Casen 2009.