Analítika

Estimadores de áreas pequeñas: cálculo de proporciones poblacionales para el caso ecuatoriano

Víctor Morales Oñate;
Carlos Jiménez Mosquera

BRINDAR TU CONOCIMIENTO AL PAÍS
Cuenta

Estimadores de áreas pequeñas: cálculo de proporciones poblacionales para el caso ecuatoriano[1]

Víctor Morales-Oñate; Carlos Jiménez-Mosquera

Instituto de Estadística, Universidad de Valparaíso, Valparaíso, Chile / Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito, Quito, Ecuador

victor.morales@uv.cl / cjimenez@usfq.edu.ec

Resumen

Este trabajo evalúa la posibilidad de mejorar la estimación de la proporción de indígenas y de adulto mayor en el Ecuador. Se aborda el problema haciendo uso de un conjunto de estimadores y explotando la similaridad espacial de los distritos analizados. El alcance de la aplicación abarca las provincias del Ecuador y los cantones de la provincia de Pichincha y el Guayas para el año 2010. Se usa la Encuesta de Empleo y Subempleo para la estimación directa, el Censo de Población y Vivienda (CPV) 2001 como información auxiliar y el CPV 2010 para el contraste de resultados.

Palabras clave: Estimador compuesto, similaridad espacial, ENEMDU.

Abstract

This paper evaluates the possibility of improving the estimate of the proportion of indigenous and elderly people in Ecuador. Dealing with this problem is by using a set of estimators and exploiting the spatial similarity of the districts surveyed addressed. The scope of this work covers the provinces of Ecuador and the cantons of the province of Pichincha and Guayas in 2010. The survey of Employment and Underemployment is used for direct estimation, the Census of Population and Housing 2001 as auxiliary information and CPV 2010 for contrasting results.

Keywords: Compound estimator, spatial similarity, ENEMDU.

Clasificador JEL: C13, C80, C81.

[1]Agradecimientos: a Nicholas Longford por su cordial ayuda en el proceso de estimación.

1. Introducción

Uno de los principales objetivos de las encuestas es determinar las características de una población tanto a nivel nacional como sub-nacional. En particular, suele ser de interés dominios geográficos como regiones, provincias, ciudades, entre otros. Los Institutos Nacionales de Estadísticas generalmente levantan información detallada de la población cada diez años a través de censos de población y vivienda. En Ecuador, además de los nuevos dominios geográficos de interés como las zonas de planificación, las dominios de interés tradicionales son: parroquias, cantones y provincias. Una de las herramientas de recolección de datos de mayor frecuencia temporal es la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). Hasta antes del 2014, esta encuesta carecía de representatividad estadística tanto a nivel de las provincias de la Amazonía así como en los cantones distintos de Quito, Guayaquil, Cuenca, Ambato y Machala, pero desde marzo del el 2014, la encuesta semestral cuenta con 24 dominos provinciales (Garcés *et al.*, 2014).

El cálculo de estimadores como el total o la media de variables de los dominios geográficos, para los cuales una encuesta es elaborada, se realiza mediante la estimación directa, esto es, un tipo de estimación que usa únicamente información del dominio, incluyendo su diseño muestral asociado (Sofronov, 2013) (bibliografía clásica en este tipo de estimación puede ser consultada en Cochran (1977), Särndal *et al.* (2003), entre otros). Sin embargo, cuando es de interés realizar estimaciones de variables para dominios no planificados, es necesario recurrir a otras técnicas de estimación: estimación de áreas pequeñas (*small area estimation* en inglés).

En términos muestrales, la ENEMDU tiene por objetivo *identificar a la población mayor de 10 años del país, que forma parte de la población económicamente activa, considerando diversos perfiles entre los que destacan: sexo, edad, grupo de ocupación, rama de actividad, categoría de ocupación, y nivel de instrucción* (Garcés *et al.*, 2014). Esto le permite obtener una determinada representatividad de la muestra en base a la cual se pueden conseguir indicadores de acceso o tenencia, plasmados en dimensiones de bienestar de la población como: acceso a servicios básicos, hacinamiento, educación, capacidad de generación de ingreso, entre otros.

Varias instituciones sociales y de planificación usan este instrumento -EMENDU- para otros fines como el cálculo de indicadores de pobreza, desigualdad, entre otros. Dichas mediciones ayudan a establecer líneas base, dar seguimiento y evaluar objetivos nacionales de desarrollo (SENPLADES, 2017). En diciembre del 2010, la ENEMDU contaba con información de 82774 personas en su muestra total. Este tamaño de muestra y su periodicidad hacen de la ENEMDU un insumo de información importante a ser explotado

en campos más allá del cálculo de indicadores de empleo.

Siguiendo a Longford (2010), el presente trabajo plantea abordar el problema de representatividad estadística para las provincias de la Amazonía así como los cantones de la provincia de Pichincha y Guayas para diciembre del año 2010 mediante el uso de una aplicación de estimadores de áreas pequeñas. En primer lugar, se plantea el uso de simulaciones de modo que se seleccione un estimador para el cálculo de la proporción. Luego, una vez elegido el estimador, se procede a aplicar la metodología en los dominios deseados.

2. Marco Teórico

Comparadas con los censos, las encuestas resultan ser una instrumento atractivo en términos de costo-beneficio para el levantamiento y posterior análisis de información. Se levantan teniendo presente una variable objetivo sobre la que se desea realizar inferencia del total de la población, un promedio o proporción. Pero también son de interés características de sub poblaciones (dominios), sean éstas de índole geográfica o de otro tipo. Un *estimador directo* es aquel que usa únicamente información del dominio donde se ha tomado la muestra, incluyendo la información de los *pesos* muestrales (distribución de probabilidad inducida por el diseño muestral), lo cual típicamente se realiza mediante el enfoque basado en el diseño. La estimación también puede ser *asistida* por modelos o con un enfoque netamente basada en el modelo (ver (Morales-Oñate y Morales-Oñate, 2017; Rao, 2015; Sterba, 2009) para más detalle en esta distinción).

Un dominio (área) será considerado *pequeño* si *la muestra específica del dominio no es lo suficientemente grande para soportar estimaciones directas de precisión adecuada* (Rao, 2015, p. 1). Esto generalmente ocurre cuando la muestra específica del dominio no fue planificada en el proyecto inicial. En este contexto, el esfuerzo por obtener información de áreas pequeñas ha sido principalmente realizado por demógrafos, quienes usan una gama de métodos indirectos para estimación de áreas pequeñas en años posteriores al levantamiento de datos censales. Generalmente estos métodos no incluyen muestreo y suelen enfocarse al cálculo población por área (urbano - rural), grupos de edad, sexo, migración, entre otros. Sin embargo, en la planificación de un país es necesario tener información de variables más específicas como pobreza, desnutrición, etnia, entre otros (Rao, 2015).

La idea central para abordar el problema de estimación de áreas pequeñas es *explotar* la *similaridad*. Por un lado, desde el enfoque basado en el diseño, la similaridad es aprovechada mediante la construcción de estimadores com-

puestos. Es decir, mediante la combinación convexa de estimadores directos y sintéticos. Un estimador sintético se basa en la idea de que las sub áreas que componen el área total tienen las mismas características de la del total (Gonzalez, 1973), trabajos como Longford (2012) aplican este enfoque. Por otro lado, otra forma de aprovechar la similaridad es a través del uso de modelos (generalmente jerárquicos) que se apoyan en la información dentro y entre dominios (Longford, 2006, p. 144). El modelo Fay-Herriot, y sus extensiones que consideran correlación espacial y espacio-temporal son ejemplos de este enfoque (Molina y Marhuenda, 2015a).

Existe software para ambos enfoques. Teniendo a Elbers *et al.* (2002) como referencia base, el Banco Mundial mantiene su propio software para el *mapeo* de pobreza basado en el modelo: *povmap*. Actualmente se encuentra en su versión 1.2 (lanzamiento 4), tiene su propio manual de usuario y está disponible en Windows (Bank, 2015; Zhao y Lanjouw, 2009). Asimismo, el paquete `sae` de `R` desarrollado por Molina y Marhuenda (2015b), contiene rutinas basadas en el diseño y el modelo. De hecho, se puede trabajar con este paquete como software de acompañamiento del libro seminal de Rao (2015), *Small Area Estimation*[2]. Por lo tanto, `sae` ilustra las metodologías presentadas en este trabajo, entre estas tenemos: mejor predicción lineal empírica (EBLUP) y bayes empírico (EB).

En general, este trabajo consiste en una aplicación del enfoque basado en el diseño debido a que el enfoque basado en el modelo es *bastante potente, pero su debilidad es que los resultados que arroja dependen de la validez del modelo* (Longford, 2006, p. 144). Claramente esto no descarta la posibilidad de que existan modelos que puedan mejorar las estimaciones aquí presentadas, pero muy probablemente los modelos necesitarían de una especificación particular en cada problema de estimación. No obstante, al prescindir de un supuesto distribucional de las variables, la aplicación aquí presentada puede ser usada de manera general y como un punto de partida en el problema de estimación de áreas pequeñas. En particular, se trata de una aplicación de la clase de estimadores compuestos propuestos en Longford (2010), donde se aprovecha la información de similaridad espacial, una métrica de distancia entre el domino objetivo y sus vecinos.

[2]Rao (2015) es una re edición de su primera aparición en 2013, la cual aún no tenía el desarrollo del paquete `sae`.

2.1. Notación y nociones de estimadores de áreas pequeñas

Siguiendo a Rao (2015) y a Longford (2006), a continuación se listan de manera más formal las ideas de áreas pequeñas:

- Se tiene una población P conformada por D áreas pequeñas y estas forman una partición P_d, $d = 1, \ldots, D$ y la variable Y está definida para todas las subpoblaciones (en adelante *distritos* para diferenciarlos de los dominios auto representados en la estimación directa) de P. El objetivo es calcular el estadístico θ_d de Y para cada distrito d
- Se asume que θ_d se puede calcular por una función Θ que puede ser usada para evaluar cualquier distrito de P, $\theta_d = \Theta(P_d)$.
- Una muestra S de P tiene una partición conforme con P_d en las subpoblaciones tal que $S_d = S \cap P_d$. El estimador directo $\hat{\theta}_d$ de θ_d depende solamente de los valores de Y y del diseño muestral. Se asume que $\hat{\theta}_d$, $d = 1, \ldots, D$ se relacionan de modo que $\hat{\theta}_d = \hat{\Theta}(S_d)$. Un ejemplo de Θ y $\hat{\Theta}$ es la media poblacional y la media muestral respectivamente.
- Los tamaños de las poblaciones de los distritos se denotan por N_d y n_d correspondientemente (siendo N y n sus totales). Las fracciones dentro de las subpoblaciones $f_d = \frac{n_d}{N_d}$ no necesariamente son idénticas. Se asume que $\hat{\Theta}_d$ es un estimador insesgado, entonces $v_d = var(\hat{\theta}_d)$; se asume que v_d es conocida.

Se define la media (1) y la varianza (2) para poblaciones finitas de un conjunto $(\theta_1, \ldots, \theta_D)$:

$$\theta = \frac{1}{D}(\theta_1 + \ldots + \theta_D) \tag{1}$$

$$\sigma_0^2 = \frac{1}{D}\sum_{d=1}^{D}(\theta_d - \theta)^2 \tag{2}$$

Se denota con subíndices $\mathcal{S}$ de $\mathcal{D}$ a los resultados muestrales y por distritos respectivamente, $v_d = var_{\mathcal{S}}(\hat{\theta}_d)$ y $\sigma_0 = var_{\mathcal{D}}(\theta_d)$. Se considera un conjunto de estimadores $\hat{\theta}_d^{(h)}$ para θ_d y su combinación convexa:

$$\tilde{\theta}_d = \sum_{h=0}^{H} b_d^{(h)} \hat{\theta}_d^{(h)} \tag{3}$$

En (3) se estiman los coeficientes $(b_d^{(1)}, \ldots, b_d^{(H)})$ y $b_d^0 = 1 - b_d^{(1)} - \ldots - b_d^{(H)}$ para los cuales el error medio cuadrático $EMC(\hat{\theta}_d, \theta_d) = E_{\mathcal{S}}\{(\hat{\theta}_d - \theta_d)^2 | \theta_d\}$ se minimiza. El estimador $\hat{\theta}_d = \hat{\theta}_d^0$ se asume insesgado.

Si $H = 1$ en (3) se tiene el estimador compuesto básico (4):

$$\tilde{\theta}_d = (1 - \hat{b}_d)\hat{\theta}_d + \hat{b}_d\hat{\theta} \tag{4}$$

donde se supone que los distritos son el resultado de un proceso de muestreo aleatorio simple estratificado $SSRS_d$, $\hat{b}_d = \frac{1}{(1+n_d\hat{\omega})}$ y ω es el ratio de las varianzas dentro y fuera de los distritos.

2.2. Similaridad Espacial

El supuesto de similaridad entre regiones geográficas pequeñas cercanas y/o contiguas es bastante intuitivo y ha sido abarcado en aplicaciones de varias índoles como estadística espacial, geoinformación y estimadores de áreas pequeñas (Cressie, 1993; Rahman *et al.*, 2010; Schmid y Münnich, 2014).

En esta aplicación se sigue a Longford (2010) quien define un conjunto de estimadores de θ_d libres de supuestos en cuanto a distribución y se asume una estructura natural de correlación relacionanda la distancia. Se asume que una función de distancia $\xi(d, d')$ está definida para cualquier pareja de distritos d y d'. Esta función es simétrica, no negativa e igual a cero cuando $d = d'$. La tabla 1 presenta el algoritmo para generar la matriz de similaridad espacial.

Basados en este algoritmo, para cada distrito se genera un $anillo - h$ como la población circundante a d,

$$\mathcal{P}_d^{(h)} = \bigcup_{\{d';\xi(d,d')\}} \mathcal{P}_{d'} \tag{5}$$

que en el caso muestral es $\mathcal{S}_d^{(h)}$.

2.3. Estimadores

Como se ha mencioando, Longford (2010) propone un conjunto de estimadores que toman en cuenta la matriz de similaridad espacial. Es decir, el estimador compuesto para cada distrito toma en cuenta la información de los $h - anillos$ alrededor del dominio d. El estimador de $\theta_d^{(h)}$ se define como,

$$\hat{\theta}_d^{(h)} = \frac{1}{n_d} \sum_{d' \in \mathrm{d}_d^{(h)}} n_{d'}\hat{\theta}_{d'} \tag{6}$$

Tabla 1: Algoritmo de similaridad espacial. Entrada: Matriz indicatriz simétrica ($M_{(d\times d)}$) de vecinos (1 si es vecino de d, 0 caso contrario) Salida: Matriz de similaridad espacial.

Paso	Descripción
1	Seleccionar un vector fila $m_{(i\times d)}$ de la matriz de entrada y generar un vector (vei) con las posiciones de los vecinos de $m_{(i\times d)}$.
2	Seleccionar una sub-matriz con las posiciones de las filas calculadas en uno con todas las columnas del paso 1
3	Si la longitud del vector calculado en el paso 1 es mayor a uno, generar un vector indicatriz de la sub- matriz calculada en 2) ($vei2$: vecinos de los vecinos)
4	Calcular: $m_{(i\times d)} + (max(M_{(d\times d)}) + vei2_{(1\times d)}) * (n_{(i\times d)} * vei2_{(1\times d)})$, donde $n_{(i\times d)}$ es el vector indicatriz de los NO vecinos de $m_{(i\times d)}$
5	Repetir el proceso para todos las filas de $M_{(d\times d)}$ para obtener $M'_{(d\times d)}$.
6	Mient*ras $min(M'_{(d\times d)}) = 0$, repetir los pasos del 1 al 5.
7	Llenar con ceros la diagonal de la matriz generada en el paso 6

Fuente: Longford (2010). Elaboración propia.

donde $\mathbb{d}_d^{(h)}$ en (6) es el conjunto de distritos que forman $\mathcal{P}_d^{(h)}$. De esta manera (3) queda definida completamente. Reescribiendo (3) se tiene:

$$\tilde{\theta}_d = \sum_{h=0}^{H} b_d^{(h)} \hat{\theta}_d^{(h)} = (\mathbf{1} - \mathbf{b}_d^T \mathbf{1})\hat{\theta}_d + \mathbf{b}_d^T \hat{\psi}_d$$

donde $\mathbf{b} = (b_d^{(1)}, \ldots, b_d^{(H)})^T$, $\hat{\psi}_d = (\hat{\theta}_d^{(1)}, \ldots, \hat{\theta}_d^{(H)})^T$, $\hat{\theta}_d^{(1)} = \hat{\theta}_d$ y los coeficientes $(b_d^{(0)} + \ldots + b_d^{(H)}) = 1$. Tanto $\mathbf{b}_d$ como $\hat{\psi}_d$ tienen longitud H.

En cuanto a la estimación de σ_0^2,σ_h^2 y la covarianza con los $h - anillos$ (γ_h), se tienen los siguientes estimadores:

$$\hat{\sigma}_0^2 = \frac{S_0}{D} - \frac{1}{D}\sum_{d=1}^{D}(1 - 2\frac{q_d}{q_+})\hat{v}_d - \hat{v}$$

donde $S_0 = \sum_{d=1}^{D}(\theta_d - \theta)^2$, q_d son los coeficientes de $\hat{\theta} = \frac{(q_1\hat{\theta}_1 + \ldots + q_D\hat{\theta}_D)}{q+}$ y q_+ es su total.

Sea $m_d^{(h)}$ el número de distritos en el $h - anillo$ del distrito d y $m_+^{(h)} = (m_1^{(h)} + \ldots + m_D^{(h)})$, entonces:

$$\hat{\sigma}_h^2 = \frac{1}{m_+^h}\{\sum_{d=1}^{D}\sum_{d'\in \mathbb{d}_d^{(h)}}(\hat{\theta}_d - \hat{\theta}_{d'})^2 - 2\sum_{d=1}^{D} m_d^{(h)} v_d\}$$

de modo que el estimador insesgado de la varianza es $\hat{\sigma}_h^2 = 2(\sigma_0^2 - \gamma_h)$, entonces,

$$\hat{\gamma}_h = \hat{\sigma}_0^2 - \frac{1}{2}\hat{\sigma}_h^2$$

De este modo se incorpora el criterio de los $h - anillos$ a la estimación de áreas pequeñas.

3. Metodología

Con el objetivo de mostrar los beneficios y limitaciones de la metodología, se han seleccionado tres aplicaciones específicas: proporción de indígenas a nivel provincial, a nivel cantonal dentro de la provincia de Pichincha y la estimación de adultos mayores a nivel cantonal dentro de la provincia del Guayas. La selección de estos casos específicos responde a varias consideraciones. Por un lado, antes del año 2014 las provincias de la Amazonía no eran autorepresentadas, es decir que la estimación directa no era posible para el cálculo de indicadores derivados de la ENEMDU y, por lo tanto, esta metodología permitiría la construcción de series históricas de información. Por otro lado, la aplicación a nivel cantonal es deseable y desafiante porque, i) Pichincha y Guayas muestran gran diferencia en el número de cantones que los componen (23 y 8 respectivamente) y ii) se evidencie cierta transversalidad en cuanto a la temática en la que se puede aplicar la estimación de áreas pequeñas (autoidentificación étnica y gerontología).

3.1. Población Indígena en el Ecuador

Según el Censo de Población y Vivienda (CPV) 2010, el 7 % de la población se auto identifica como indígena en el Ecuador. En términos generales, se puede apreciar en la figura 1 que la presencia de la población indígena en la mayoría de las provincias de la Región Costa es baja, media-alta en la Sierra y alta en la Amazonía. El rango de la proporción de población indígena se encuentra entre 0,18 % y 56,74 %. Esto implica que la media nacional *esconde* gran parte de los valores a nivel provincial. El mapa también sugiere que una aproximación a través de similaridad espacial podría ser un camino

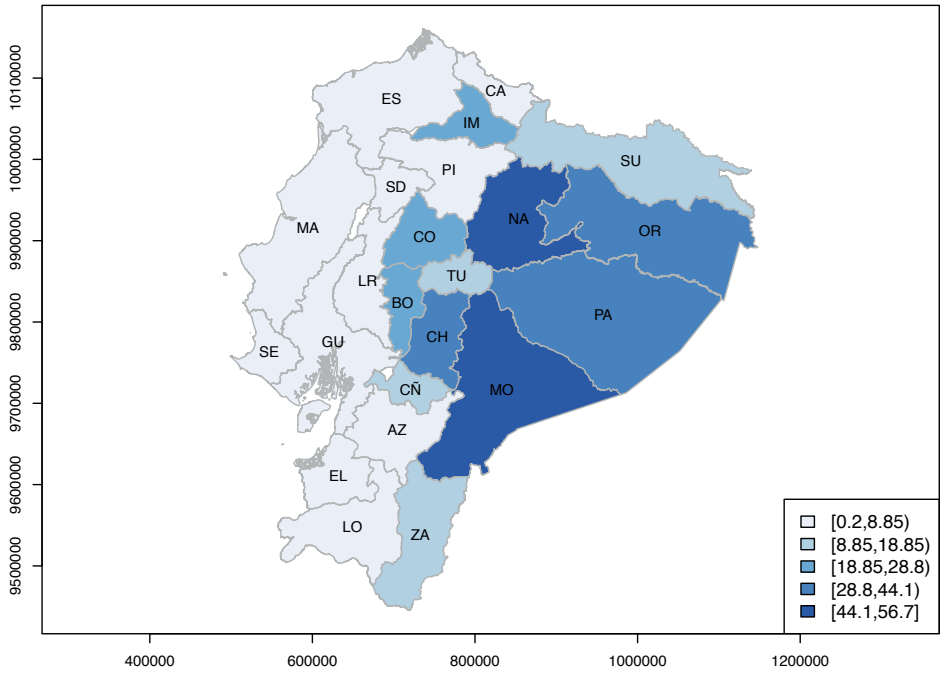

Figura 1: Proporción de indígenas a nivel provincial. Fuente: CPV 2010. Elaboración propia.

Tabla 2: Población por etnia - miles.

Acr	Provincia	Indígena	Afro	Negro	Mulato	Montubio	Mestizo	Blanco	Otro	Total
AZ	AZUAY	17,6	10,8	0,9	3,9	2,9	637,9	36,7	1,3	712,1
BO	BOLIVAR	46,7	1,2	0,2	0,6	2,1	127,8	4,9	0,2	183,6
CÑ	CAÑAR	34,2	4,2	0,5	1,3	2,4	172,6	9,6	0,4	225,2
CA	CARCHI	5,6	6,8	1,7	2,1	0,4	142,9	4,7	0,2	164,5
CO	COTOPAXI	90,4	4,8	0,4	1,6	7,3	294,8	9,3	0,5	409,2
CH	CHIMBORAZO	174,2	3,6	0,2	1,2	1,2	267,9	10	0,4	458,6
EL	EL ORO	4,1	24,2	4,7	12,6	16,9	489,8	46,8	1,7	600,7
ES	ESMERALDAS	15	123,1	56,6	54,9	13	238,6	31,3	1,6	534,1
GU	GUAYAS	46,2	204,3	36,4	111,4	411	2461,74	355,3	19,1	3645,5
IM	IMBABURA	102,6	12,2	4,1	5,2	1,2	261,7	10,8	0,5	398,2
LO	LOJA	16,5	8,3	0,6	1,8	3,2	404,9	13,2	0,5	449
LR	LOS RIOS	5	30,3	7,1	10,7	272,7	411,9	38,5	2	778,1
MA	MANABI	2,5	62,2	8,6	11,4	262,7	954,2	64,3	3,9	1369,8
MO	MORONA SANTIAGO	71,5	1,1	0,2	0,6	0,3	68,9	4,6	0,8	147,9
NA	NAPO	58,8	0,8	0,2	0,7	0,6	39,5	2,8	0,2	103,7
PA	PASTAZA	33,4	0,6	0,2	0,5	0,3	46,4	2,4	0,1	83,9
PI	PICHINCHA	137,6	65,4	12,8	38,4	34,6	2115	163,2	9,4	2576,3
TU	TUNGURAHUA	62,6	4,7	0,3	2,2	2,3	414,5	17,4	0,7	504,6
ZA	ZAMORA CHINCHIPE	14,2	0,8	0,2	0,3	0,2	73,4	1,9	0,3	91,4
SU	SUCUMBIOS	23,7	4,3	2,3	3,8	1,7	132,4	8	0,4	176,5
OR	ORELLANA	43,3	2,6	1,7	2,4	1,6	78,4	6	0,3	136,4
SD	SANTO DOMINGO	6,3	16,4	3,8	8,2	9	298,2	25,1	1	368
SE	SANTA ELENA	4,2	20,6	1,5	4,2	15,2	244,3	11,4	7,4	308,7

Fuente: CVP 2010. Elaboración propia

apropiado para la estimación de este porcentaje a nivel provincial ya que provincias cercanas muestran tonos similares.

Cabe notarse la diferencia que se establece en la tabla 2 al analizarse los datos en valores absolutos. Por ejemplo, la provincia de Pichincha, con aproximadamente 138 mil habitantes que se autoidentifican como indígenas en el año de análisis, ubicándose segunda a nivel nacional en este indicador. Ahora, si se analiza en términos relativos, este valor representa un 5,3 % de su composición por etnias, siendo Napo la provincia que tiene mayor proporción de indígenas a nivel nacional (56,7 %).

3.2. Población Indígena en Pichincha

Según el CPV 2010, el 5 % de la población se auto identifica como indígena en Pichincha. En el caso de la proporción de indígenas en los cantones de la provincia de Pichincha (figura 2), este indicador tiene un rango de 0,5 % a 34 %. Su rango es menor que el caso analizado a través de provincias pero sigue habiendo diferencias marcadas.

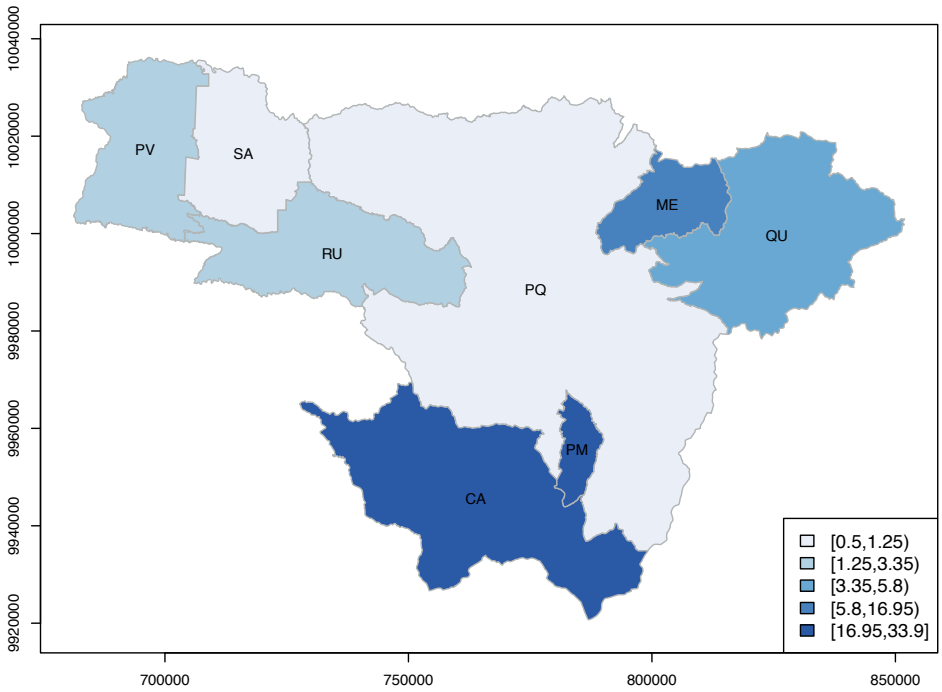

Figura 2: Proporción de Indígenas a nivel Cantonal - Pichincha. Fuente: CPV 2010. Elaboración propia.

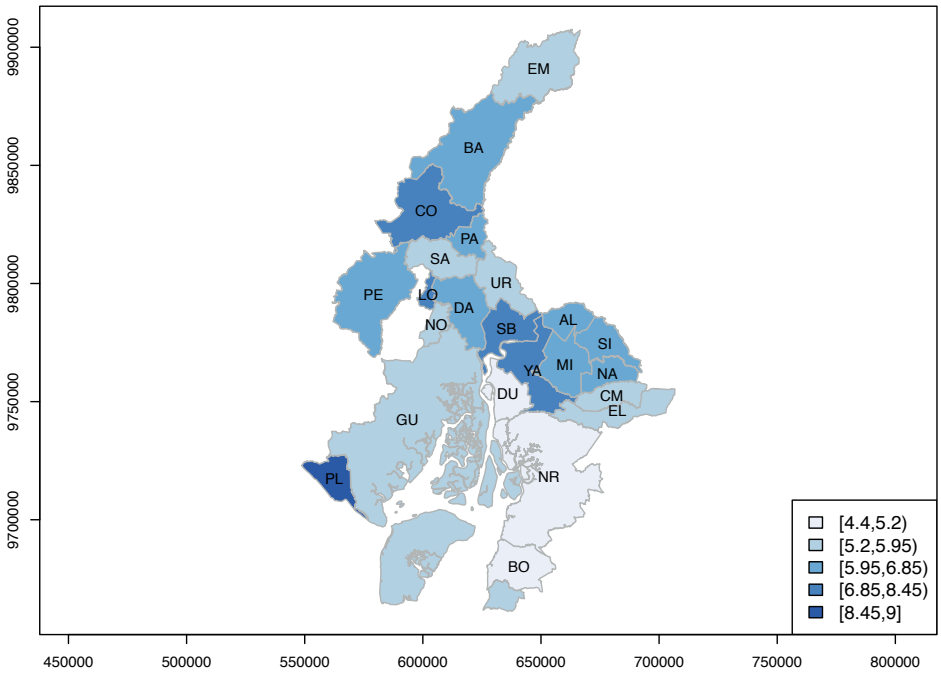

Figura 3: Proporción de Adultos mayores a nivel Cantonal - Guayas. Fuente: CPV 2010. Elaboración propia.

La presencia de la población indígena cantonal en orden descendente es: Cayambe (CA), Pedro Moncayo (PM), Mejía (ME), Quito (QU), Pedro Vicente Maldonado (PV), Rumiñahui (RU), San Miguel de los Bancos (SA) y Puerto Quito (PQ). Similar al caso anterior, la figura 2 sugiere que la estimación del indicador analizado puede ser mejorada al tomar en cuenta la similaridad espacial de los cantones objetivo.

3.3. Adulto mayor en los cantones del Guayas

La provincia del Guayas cuenta con 8435 individuos en su muestra para diciembre de 2010. Es la provincia que tiene la mayor muestra de hogares a nivel Nacional y está dividida en 23 cantones. Según el CPV 2010, las personas de 65 años y más representan el 5,8 % (205128 personas) del total provincial. La figura 3 muestra este valor distribuido a nivel cantonal.

El rango de la proporción de adultos mayores se encuentra entre 4,4 y 9,0. Es mucho menor al rango de las secciones 3 y 3.2. La menor volatilidad y el número de cantones de esta Provincia puede ser de ayuda para mejorar

la estimación.

3.4. Selección del Estimador de área pequeña

Se han generado 500 muestras con muestreo aleatorio simple estratificado donde los estratos son los distritos d ($SSRS_d$, por sus siglas en inglés). Estas muestras son generadas a partir de las medias poblacionales de las provincias del Ecuador y de los cantones de Pichincha y Guayas según el caso. Finalmente se calculan las proporciones y tamaños muestrales de cada distrito y cada muestra para ser evaluadas por un conjunto de estimadores.

De la familia de estimadores posibles, el presente trabajo se orienta a la estimación directa, compuesta univariada y compuesta bivariada. Para estos dos últimos tipos de estimadores se tiene los casos en los que se puede tomar en cuenta la matriz de similaridad espacial. Las etiquetas de las estimaciones compuestas vienen dadas por $K - Comp - H$, donde K indica el uso de información auxiliar ($K = 1$ si no usa y $K = 2$ caso contrario) y H el valor de la matriz de similaridad espacial ($H = 1$ si no la usa y $H > 1$ con el valor de la matriz de similaridad espacial truncada en H).

4. Resultados

Los resultados conseguidos permiten obtener la mejor variante $K-Comp-H$ para cada aplicación mediante el análisis del error de cada opción. La tabla 3 resume estos resultados para el caso de las provincias del Ecuador.

Tabla 3: Conjunto de estimadores: Proporción de indígenas en las provincias del Ecuador.

$K - Comp - H$	Directo	1-Comp-1	1-Comp-2	1-Comp-3	2-Comp-1	2-Comp-2
Directo		0 (0)	0 (0)	0 (0,001)	0,001 (0,001)	0,001 (0,001)
1-Comp-1	0,003 [16]		0 (0)	0 (0,001)	0,001 (0,001)	0,001 (0,001)
1-Comp-2	0,005 [17]	0,003 [15]		0 (0,001)	0,001 (0,001)	0,001 (0,001)
1-Comp-3	0,008 [15]	0,005 [16]	0,004 [21]		0,001 (0,001)	0,001 (0,001)
2-Comp-1	0,016 [23]	0,014 [22]	0,014 [22]	0,016 [22]		0 (0)
2-Comp-2	0,019 [21]	0,016 [23]	0,017 [21]	0,019 [22]	0,006 [17]	

Fuente: CPV 2010. Elaboración propia

Se aprecia que la tabla está dividida en dos paneles. En la parte inferior de la diagonal se muestra la suma del valor absoluto del error de estimación. En corchetes se expresa el número de provincias en que el estimador de la fila es superior al estimador de la columna. En la parte superior de la diagonal se muestra la media del valor absoluto del error de estimación y su desviación estándar encerrada en paréntesis. El rango de la suma del valor absoluto de

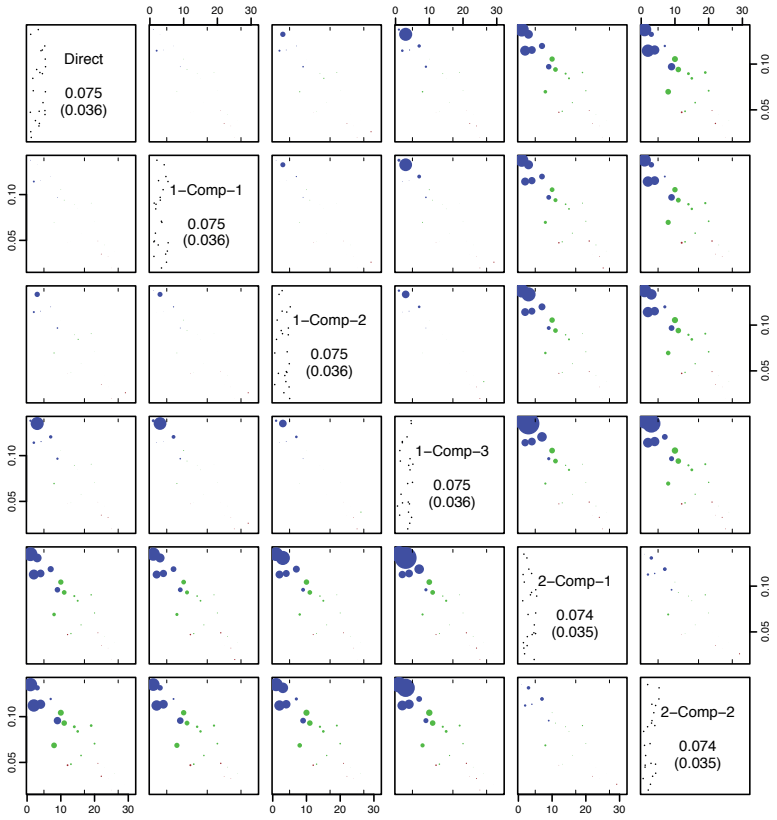

Figura 4: Suma de errores absolutos por provincias del Ecuador. Fuente: CPV 2010. Elaboración propia.

los errores está entre 0,003 y 0,019 para todas las comparaciones, en general es un valor pequeño. Si se compara el conjunto de estimadores propuesto con la estimación directa, la estimación $1-Comp-1$ (estimación compuesta que no usa información auxiliar ni la matriz de similaridad espacial) es la peor y la estimación $2-Comp-1$ (estimación compuesta que usa información auxiliar y no usa la matriz de similaridad espacial) es la mejor. Al tomar en cuenta ambos criterios, la suma del valor absoluto del error y el número de provincias en que mejora la estimación, el estimador $2-Comp-1$ sigue siendo la mejor opción. Pese a que existen otros estimadores con valores inferiores en el primer criterio, el número de provincias en que se mejora la estimación es considerablemente mayor (23 a 16, 23 a 17 y 23 a 15). En resumen, la tabla 3 sugiere usar la estimación $2-Comp-1$ para el cálculo de las proporciones de indígenas a nivel provincial.

La suma de los valores absolutos de los errores de cada provincia se compara en la figura 4. Cada panel de la figura 4 fuera de la diagonal tiene la misma escala. Cada conjunto de estimadores están representados en el eje vertical y en el eje horizontal están las provincias según el orden ascendente de sus tamaños poblacionales. El diámetro de los puntos en cada sub-gráfico muestra la magnitud de la diferencia entre la estimación del panel horizontal y vertical; sus colores indican la Región Natural a la que pertenece (Sierra: verde, Costa: café, Amazonía: azul). En la diagonal principal están las medias y las desviaciones estándar de la suma del valor absoluto de los errores así como los valores de los estimadores del eje horizontal distribuidos verticalmente en forma aleatoria.

Comparando con la estimación directa, a medida que se agrega información en los estimadores $K - Comp - H$ se puede observar que las diferencias de las sumas de los valores absolutos crecen. Al agregar información de la matriz de similaridad espacial (H), las diferencias en la Amazonía son mucho más notorias mientras que para las demás regiones es constante. Si se aumenta información auxiliar a la estimación, las diferencias de las regiones de la Amazonía y de la Sierra crecen más que en el caso anterior y la Costa permanece constante. Esto se debe a la proporción indígena por regiones, Sierra 11 %, Costa 1 % y Amazonía 33 %. Se encuentra un patrón consecuente entre el aumento de información y la mejora de la estimación dada por el conjunto de estimadores $K - Comp - H$. Para los fines de este trabajo, se procede a usar la estimación $2 - Comp - 1$ para el cálculo de proporciones de indígenas a nivel provincial. La tabla 9 muestra la estimación de cada provincia. Se puede notar que las estimaciones en provincias representativas como Azuay es muy parecida para el caso $2 - Comp - 1$ (2.957 %), la estimación directa (2.457 %) y la información del indicador del CPV 2001 usada como auxiliar (2.477 %).

4.1. Estimador para los cantones de Pichincha

La figura 4 resume las comparaciones de los estimadores en el caso de los cantones de la provincia de Pichincha.

Tabla 4: Comparaciones de conjuntos de estimadores para provincias.

$K - Comp - H$	Directo	1-Comp-1	1-Comp-2	1-Comp-3	2-Comp-1	2-Comp-2
Directo		0,002 (0,003)	0,002 (0,002)	0,002 (0,003)	−0,004 (0,01)	−0,003 (0,012)
1-Comp-1	0,016 [6]		0 (0,001)	0 (0,002)	−0,006 (0,011)	−0,004 (0,014)
1-Comp-2	0,018 [7]	0,008 [5]		0 (0,001)	−0,007 (0,012)	−0,005 (0,013)
1-Comp-3	0,02 [7]	0,009 [5]	0,004 [3]		−0,007 (0,012)	−0,005 (0,013)
2-Comp-1	0,048 [4]	0,063 [3]	0,062 [3]	0,065 [3]		0,002 (0,017)
2-Comp-2	0,065 [5]	0,074 [5]	0,067 [5]	0,067 [5]	0,072 [5]	

Fuente: CPV 2010. Elaboración propia

El rango de la suma del valor absoluto de los errores está entre 0,004 y 0,074 para todas las comparaciones, las variaciones cantones con 4 veces mayores a las provinciales. La figura 4 sugiere que sea la estimación $1 - Comp - 1$ es la mejor en términos de la suma de la diferencia de los valores absolutos. Asimismo, serán las estimaciones $1 - Comp - 2$ y $1 - Comp - 3$ las candidatas a ser seleccionadas al considerar el criterio del número de cantones que presentan mejora. Complementariamente a las sugerencias que muestra la figura 4, se presenta una medición individual de los cantones en la figura 5.

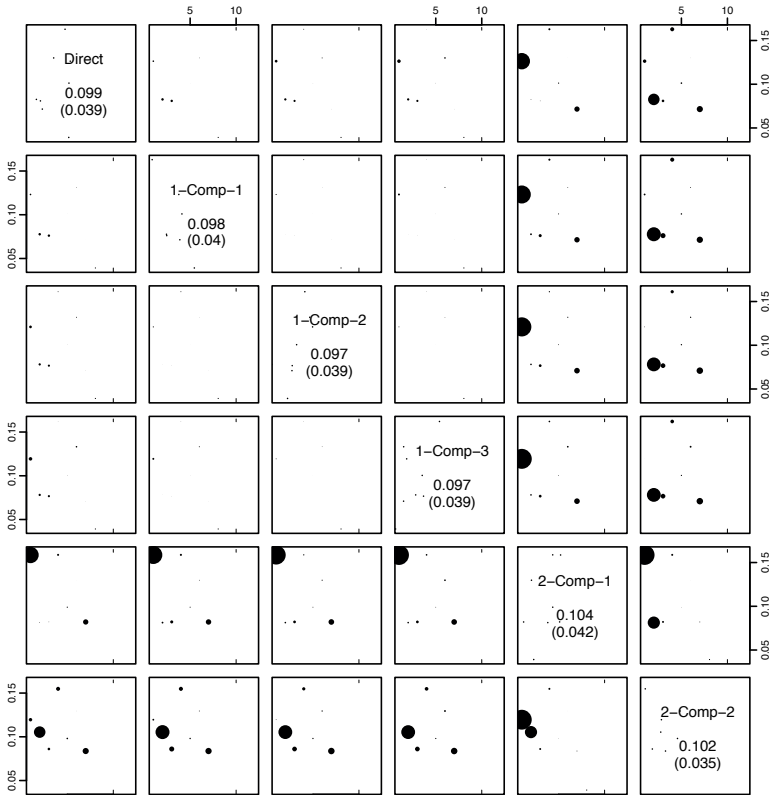

Figura 5: Suma de errores absolutos por cantones de la provincia de Pichincha. Fuente: CPV 2010. Elaboración propia.

Esta figura no sugiere un patrón de disminución del error a medida que aumenta la población. Cuatro de los ocho cantones de la provincia de Pichincha tienen valores de la proporción de indígenas muy cercanos a cero. Esto corrobora la falta de evidencia de un patrón poblacional. Pedro Moncayo muestra un comportamiento aberrante, es el punto atípico en todos los casos. Esto se debe a que entre el 2001 y el 2010 la proporción de la población que se considera indígena para este cantón es la que más creció (7,8 %).

Es claro que los estimadores compuestos mejoran los resultados en todos los casos. Sin embargo, no es posible determinar uno en particular para la estimación a partir de la evidencia mostrada en la figura 5. Para el caso de los cantones de Pichincha, se procede a realizar el cálculo de las proporciones con el supuesto de que a mayor información, mejor resultado ($2-Comp-2$). La tabla 10 muestra la estimación de cada cantón de Pichincha. Se puede notar que las estimaciones en cantones representativas como Quito es muy parecida para el caso $2-Comp-2$ (3.036 %), la estimación directa (2.37 %) y la información del indicador del CPV 2001 usada como auxiliar (4.085 %).

4.2. Estimador para los cantones de Guayas: Adulto Mayor

La tabla 5 resume las comparaciones de los estimadores en el caso de los cantones de la provincia de Guayas para la estimación de la proporción de adultos mayores.

Tabla 5: Comparaciones de conjuntos de estimadores para cantones del Guayas.

$K-Comp-H$	Directo	1-Comp-1	1-Comp-2	1-Comp-3	2-Comp-1	2-Comp-2
Directo		−0,001 (0,004)	−0,002 (0,005)	−0,002 (0,005)	−0,001 (0,004)	−0,001 (0,005)
1-Comp-1	0,072 [11]		−0,001 (0,001)	−0,001 (0,001)	0 (0,001)	−0,001 (0,001)
1-Comp-2	0,086 [10]	0,027 [2]		0 (0,001)	0,001 (0,001)	0 (0,001)
1-Comp-3	0,089 [10]	0,029 [2]	0,001 [13]		0,001 (0,002)	0 (0,001)
2-Comp-1	0,068 [11]	0,021 [12]	0,036 [19]	0,04 [19]		−0,001 (0,001)
2-Comp-2	0,079 [10]	0,025 [8]	0,019 [19]	0,023 [16]	0,026 [2]	

Fuente: CPV 2010. Elaboración propia

Al comparar los valores de la suma del valor absoluto de la primera columna de la tabla 5, se aprecia que el de menor valor es la estimación $2-Comp-1$ y mejora a la estimación directa en 11 cantones. Sin embargo, la estimación $2-Comp-2$ supera a la anterior en dos cantones a cambio de un ligero aumento en el error. La tabla 5 sugiere entonces el uso de la estimación $2-Comp-2$ para la proporción de adultos mayores.

Todas las estimaciones diferentes de la estimación directa muestran mejoras en cuanto a precisión. Al comparar la estimación directa con la $2-Comp-2$ se tiene que la estimación directa muestra un error promedio de 0,935 y la segunda de 0,937, pero la varianza de la primera es mayor. De esta forma, en función de la tabla 5 y la figura 6 se puede concluir que la estimación $2-Comp-2$ es una buena opción para la estimación de la proporción de adultos mayores.

En esta sección se resume los resultados obtenidos a partir de los modelos $2-Comp-1$ y $2-Comp-2$ para provincias del Ecuador y cantones de Pichincha y Guayas respectivamente.

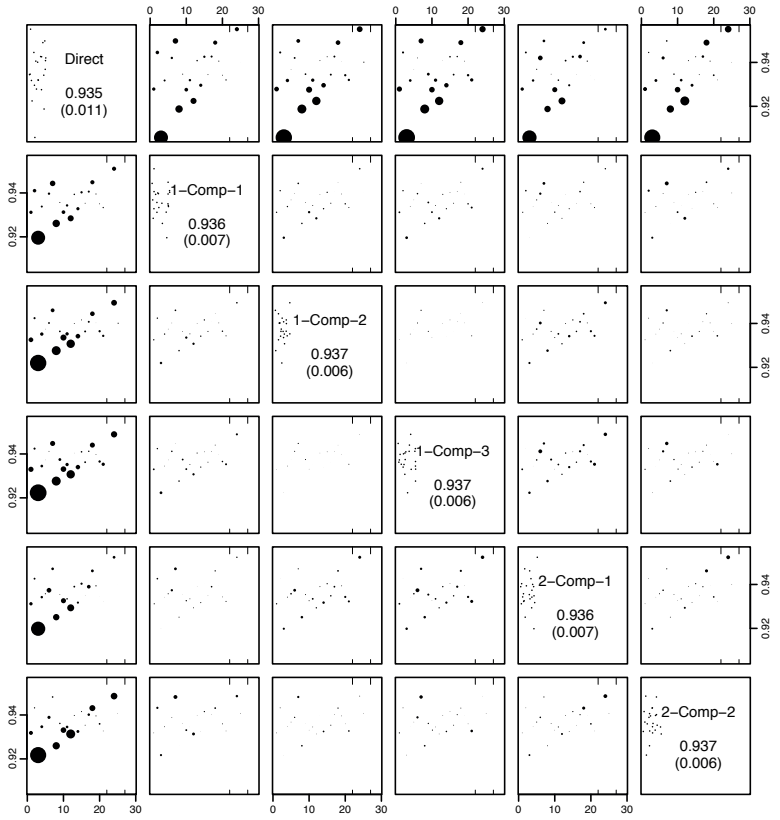

Figura 6: Suma de errores absolutos por cantones de la provincia de Guayas. Fuente: CPV 2010. Elaboración propia.

Tabla 6: Matriz de distancias de los estimadores de áreas pequeñas - Provincial.

	2-Comp-1	ENEMDU	CPV-2010
2-Comp-1	0	0,004	24,339
ENEMDU	0,004	0	24,343
CPV-2010	24,339	24,343	0

Fuente: CPV 2010. Elaboración propia

La matriz de distancia entre los estimadores $2 - Comp - 1$, ENEMDU y CPV 2010 se presenta en la tabla 6. La distancia entre la estimación proveniente de la ENEMDU es mayor que la distancia que tiene $2 - Comp - 1$ respecto de las proporciones del Censo de Población y Vivienda. De esta manera se logra el objetivo de mejorar la estimación de las proporciones de indígenas a nivel provincial. En particular la estimación mejora en 4 de las provincias de la región amazónica.

Tabla 7: Matriz de distancias de los estimadores de áreas pequeñas - Cantonal Pichincha.

	2-Comp-2	ENEMDU	CPV-2010
2-Comp-2	0	0,005	48,191
ENEMDU	0,005	0	48,194
CPV-2010	48,191	48,194	0

Fuente: CPV 2010. Elaboración propia

La tabla 7 muestra la misma información que la tabla 6, en este caso para los cantones de la Provincia de Pichincha, se puede observar que existe una distancia menor entre el CPV 2010 y el estimador $2 - Comp - 2$ que la distancia entre la proporción obtenida a partir de la ENEMDU y el CPV 2010. Al igual que el caso anterior, mejora la estimación de la proporción de indígenas para los cantones de Pichincha. La tabla 8 resume los resultados para la estimación de los adultos mayores en los cantones de la provincia de Guayas.

Tabla 8: Matriz de distancias de los estimadores de áreas pequeñas - Cantonal Guayas.

	2-Comp-2	ENEMDU	CPV-2010
2-Comp-2	0	0,155	21,617
ENEMDU	0,155	0	21,705
CPV-2010	21,617	21,705	0

Fuente: CPV 2010. Elaboración propia

La estimación compuesta usando información del 2001 y la matriz de similaridad espacial es más precisa que la estimación directa. La distancia que existe entre la estimación de este trabajo y los resultados del CPV 2010 es la menor de todas en la tabla 8. Esto puede deberse a la menor heterogeneidad del indicador estimado. La tabla 11 muestra la estimación de cada cantón de Guayas. Se puede notar que las estimaciones en cantones representativos como Guayaquil es muy parecida para el caso $2 - Comp - 2$ (6.37 %), la estimación directa (6.37 %) y la información del indicador del CPV 2001 usada como auxiliar (5.81 %).

5. Conclusiones

A través de las tablas 3 y 4 se ha puesto en evidencia que las estimaciones a través de composición superan a las estimaciones directas en todos los casos tanto para la estimación de la proporción de indígenas en las provincias del Ecuador como en los cantones de la provincia de Pichincha.

Cuando se cuenta con un número considerable de distritos, se puede observar patrones de comportamiento error-población. A medida que aumenta la población disminuye el error de estimación en todos los casos de la aplicación provincial.

Mayor homogeneidad en la estimación del indicador de adultos mayores refleja mayor precisión al momento de incorporar estimaciones compuestas en todos los casos.

La elección del modelo estimado depende de varios criterios, incluyendo en el número de distritos en los que la estimación mejora y al menor valor de del error causado.

El contar con información auxiliar suele ser uno de los grandes inconvenientes para la estimación de áreas pequeñas. En este sentido, el uso de la matriz de similaridad espacial ayuda mejorar las estimaciones directas del indicador. Por ejemplo, las estimaciones 1-Comp-1 y 1-Comp-3 mejoran la estimación en 17 y 15 provincias respectivamente.

El conjunto de estimadores $K-Comp-H$ son libres de uso de un modelo. Esto puede ser una gran ventaja al dejar de lado la subjetividad que puede generarse en la especificación del modelo usado para estimadores de áreas pequeñas.

Dado que los estimadores socioeconómicos son insumos para la planificación nacional y local, de su adecuada estimación depende el hecho de que ciertos recursos sean asignados a una determinada localidad, que una persona sea considerada o no en la participación de algún programa social, entre otros. En este contexto, cualquier mejora en la estimación puede ser determinante en la toma de decisiones.

Se ha utilizado como información auxiliar a los datos del CPV 2001 debido a que era la última información censal disponible para los puntos de tiempos analizados. Pero, al ser información de casi diez años atrás, es probable que esto haya afectado la precisión de la estimación compuesta donde la información censal actúa como auxiliar. Se recomienda el uso de información auxiliar más reciente, misma que puede provenir de registros administrativos o encuestas especializadas. En particular, para el caso de adultos mayores se recomienda el uso de información del Registro Social.

En el presente estudio la determinación de la muestra a través de $SSRS_d$ se realiza asumiendo una fracción de la población constante de $\frac{1}{200}$. Este valor

podrá ser considerado diferente para cada distrito y evaluar el rendimiento del conjunto de estimadores en este caso.

Referencias

Bank, W. (2015). Software for poverty mapping.

Cochran, W. (1977). Survey sampling. *J. Wiley. New York*.

Cressie, N. A. (1993). *Statistics for spatial data*. Wiley Online Library.

Elbers, C., Lanjouw, J. O., y Lanjouw, P. (2002). *Micro-level estimation of welfare*, volumen 2911. World Bank Publications.

Garcés, C., Albán, A., y Troya, P. (2014). *Metodología del diseño muestral de la encuesta de empleo y desempleo ENEMDU*. Instituto Nacional de Estadísticas y Censos.

Gonzalez, M. E. (1973). Use and evaluation of synthetic estimates. En *Proceedings of the social statistics section*, pp. 33–36. American Statistical Association.

Longford, N. T. (2006). *Missing data and small-area estimation: Modern analytical equipment for the survey statistician*. Springer Science & Business Media.

Longford, N. T. (2010). Small area estimation with spatial similarity. *Computational Statistics & Data Analysis*, 54(4):1151–1166.

Longford, N. T. (2012). Allocating a limited budget to small areas. *Journal of the Indian Society of Agricultural Statistics*, 66(1):31–41.

Molina, I. y Marhuenda, Y. (2015a). sae: An r package for small area estimation. *The R Journal*, 7(1):81–98.

Molina, I. y Marhuenda, Y. (2015b). sae: An R package for small area estimation. *The R Journal*, 7(1):81–98.

Morales-Oñate, V. y Morales-Oñate, B. (2017). Regresión lineal bajo diseños complejos: un enfoque aplicado. *Analítik*, 14(2):103–124.

Rahman, Azizur and Harding, Ann and Tanton, Robert and Liu, Shuangzhe and others (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation*, 3(2):3–22.

Rao, J. N. (2015). *Small-Area Estimation*. Wiley Online Library.

Särndal, C.-E., Swensson, B., y Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schmid, T. y Münnich, R. T. (2014). Spatial robust small area estimation. *Statistical Papers*, 55(3):653–670.

SENPLADES (2017). *Plan Nacional de Desarrollo*. Secretaría Nacional de Planificación y Desarrollo.

Sofronov, G. (2013). A hybrid algorithm for spatial small area estimation under models with complex contiguity. En *Differential Evolution (SDE), 2013 IEEE Symposium on*, pp. 25–30. IEEE.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate behavioral research*, 44(6):711–740.

Zhao, Q. y Lanjouw, P. (2009). Using povmap2-a users guide. *The World Bank*.

ANEXO

A. Anexo 1

	2-Comp-1	Directo	Auxiliar
AZ	2.957	2.457	2.477
BO	23.531	23.39	25.44
CÑ	14.565	12.812	15.193
CA	5.407	8.028	3.434
CO	22.811	21.557	22.101
CH	37.546	37.096	37.989
EL	0.971	0.896	0.676
ES	2.537	2.336	2.813
GU	1.016	0.728	1.268
IM	25.296	25.309	25.773
LO	3.51	3.964	3.67
LR	0.642	0.435	0.638
MA	0.384	0.225	0.179
MO	41.526	41.901	48.356
NA	60.111	65.316	56.747
PA	31.439	25.895	39.792
PI	4.866	5.469	5.339
TU	13.137	11.602	12.403
ZA	11.003	9.802	15.561
SU	6.344	2.238	13.421
OR	26.274	22.197	31.767
SD	1.516	1.017	1.717
SE	0.842	0.709	1.349

Tabla 9: Estimación de proporción de indígenas a nivel provincial. La columna *2-Comp-1* es la estimación sae, *Directo* es la estimación directa de la ENEMDU y *Auxiliar* es la información del indicador correspondiente al censo 2001. Fuente: CPV 2010. Elaboración propia

	2-Comp-2	Directo	Objetivo
QU	3.036	2.37	4.085
CA	41.3	74.94	33.868
ME	4.317	4.05	7.48
PM	13.465	1.55	26.42
RU	7.929	3.04	1.893
SA	0.92	1.38	0.58
PV	7.933	3.22	2.608
PQ	0.703	1.04	0.523

Tabla 10: Estimación de proporción de indígenas a nivel cantonal en la provincia de Pichincha. La columna *2-Comp-2* es la estimación sae, *Directo* es la estimación directa de la ENEMDU y *Auxiliar* es la información del indicador correspondiente al censo 2001. Fuente: CPV 2010. Elaboración propia

	2-Comp-2	Directo	Objetivo
GU	6.37	6.37	5.81
AL	13.82	13.83	6.3
BO	6.26	6.24	4.89
BA	12.62	12.64	6.37
CO	6.28	6.26	7.91
DA	8.31	8.31	6.4
DU	7.01	7.01	4.37
EM	8.75	8.75	5.75
EL	8.26	8.26	5.57
MI	12.45	12.45	6.6
NR	9.73	9.73	4.96
NA	7.91	7.91	6.51
PA	15.13	15.22	6.62
PE	11.09	11.11	6.71
SB	13.34	13.37	7.42
SA	7.86	7.86	5.57
UR	14.19	14.2	5.72
YA	6.53	6.52	7.02
PL	5.48	5.46	8.98
SI	5.63	5.6	6.13
CM	2.97	2.86	5.64
LO	12.73	12.75	7.12
NO	10.49	10.53	5.38

Tabla 11: Estimación de proporción de adultos mayores a nivel cantonal en la provincia del Guayas. La columna *2-Comp-2* es la estimación sae, *Directo* es la estimación directa de la ENEMDU y *Auxiliar* es la información del indicador correspondiente al censo 2001. Fuente: CPV 2010. Elaboración propia