Actualización de tabulados censales via SAE

Angela Luna Hernández
A.Luna.Hernandez@soton.ac.uk

Department of Social Statistics and Demography
University of Southampton

Noviembre 2018

Agradecimientos

Especiales agradecimientos a Li-Chun Zhang y Nikos Tzavidis, de la Universidad de Southampton, por su contribución al desarrollo de MSPREE y a ONS, en particular Alison Whitworth, Kirsten Piller, Solange Correa y Philip Clarke, por llevar a cabo buena parte de los análisis aquí presentados.

Este trabajo ha recibido financiamiento de la oficina nacional de estadísticas británica - ONS y el consejo de investigación económico y social británico ESRC. Nueva investigación es parcialmente financiada por el programa de investigación e innovación de la Unión Europea, Horizon 2020, a través de el proyecto INGRID-2 (grant No. 730998).

Outline

Motivación

- Interés en obtener información de fenómenos sociales y demográficos a altos niveles de desagregación.
- Encuestas por muestreo no cuentan con el tamaño muestral suficiente para dar resultados a estos niveles. Alta incertidumbre y áreas ausentes.
- Censos de población no son suficientemente frecuentes. Brechas de información.
- Otras fuentes de información disponibles (registros administrativos) usualmente no contienen las variables de interés.

Motivación

Tabulado censal (gold standard) de una o más variables categóricas de acuerdo a una geografía predeterminada:

Area	1	...	J	Total
1				$Y_{1\cdot}$
2				$Y_{2\cdot}$
3				$Y_{3\cdot}$
		Y_{aj}		
...				...
A				$Y_{A\cdot}$
Total	$Y_{\cdot 1}$	...	$Y_{\cdot J}$	$Y_{\cdot\cdot}$

Motivación

Tabulado censal (gold standard) de una o más variables categóricas de acuerdo a una geografía predeterminada:

Area	1	...	J	Total
1				$Y_{1\cdot}$
2				$Y_{2\cdot}$
3				$Y_{3\cdot}$
		Y_{aj}		
...				...
A				$Y_{A\cdot}$
Total	$Y_{\cdot 1}$	...	$Y_{\cdot J}$	$Y_{\cdot\cdot}$

Objetivo: actualización de tabulados censales en el período inter-censal.
SAE: optimizar la relación sesgo-varianza de diferentes fuentes de datos.

Fuentes de datos disponibles

Censos de población:

- Tabulado de interés, cruces con otras variables asociadas.
- Pros: cobertura total. No error muestral.
- Contras: desactualización $\rightarrow$ sesgo.

Registros administrativos:

- Registros de vivienda, condiciones de alojamiento, educación, empleo, ingresos (vía impuestos), ...
- Pros: no error muestral. Actualización periódica.
- Contras:
  - diferencias en población objetivo, definiciones, geografía, método de recolección, informante, etc $\rightarrow$ sesgo.
  - falta de autonomía para la oficina de estadísticas.

Fuentes de datos disponibles

Encuestas por muestreo:

- Tabulados parciales o agregados provenientes de encuestas a hogares (mercado laboral, ingresos y gastos,...), conjuntos de márgenes.
- Pros: insesgamiento. Disponibilidad de variables asociadas.
- Contras: baja precisión para estimativas desagregadas.

Otras fuentes:

- Big data, información geoespacial, mobile phone data, etc...
- Pros: bajo costo, actualización 'en tiempo real', granularidad.
- Dudas: capacidad predictiva, facilidad de acceso, errores de medición, complejidad de manejo.

Consideraciones

Area-level vs unit-level models

- Parámetros de interés son frecuencias (celda) o proporciones (|area).
- Posibles predictores son categóricos (datos agrupados) → Modelamiento de tablas de contingencia → 'area'-level models.

Jerarquía de tipos de datos

1. Tabulado previo (censo) y márgenes actualizadas (encuesta) (Proxy association) + (Asignación) → SPREE.
2. Estimativas muestrales: actualización de la estructura de asociación → GSPREE, MSPREE.
3. Tabulados (parciales) de fuentes administrativas.
4. Otras fuentes.

Métodos

Distancia:

Chi-cuadrado (Dostál et al., 2016), CSPREE (Molina et al., 2008).

Regresión:

Modelos multinomiales con efectos aleatorios. Molina et al. (2007), Scealy (2010), Saei & Taylor (2012), López-Vizcaíno et al. (2013), López-Vizcaíno et al. (2015).

Preservan la estructura:

SPREE (Purcell & Kish, 1980), GSPREE (Zhang & Chambers, 2004), MSPREE (Luna-Hernández, 2016).

Para contraste de algunos métodos ver: Little & Wu (1991), Suesse et al. (2017).

Métodos

Chi-cuadrado: (Dostál et al., 2016)

$M = (m_{aj})$ tabla de frecuencias censales (adj.), $a = 1, ..., A$; $j = 1, ..., J$.

La idea es obtener $N = (n_{aj})$ que minimice:

$$\sum_{a,j} \frac{\left(\frac{n_{aj}}{n_{\bullet\bullet}} - \frac{m_{aj}}{n_{\bullet\bullet}} \right)^2}{m_{aj}}$$

sujeto a restricciones $n_{a\bullet} = \sum_j n_{aj}$ y $n_{\bullet j} = \sum_a n_{aj}$ dados.

Métodos

CSPREE: (Molina et al., 2008, cited Rao & Molina, 2015)

$N = (n_{aj})$ tabla de frecuencias censales, $a = 1, ..., A$; $j = 1, ..., J$,
$\hat{M} = (\hat{m}_{aj})$ tabla de estimativas muestrales.

Para cada area a y un conjunto de constantes α_{aj}, obtener $\tilde{M}_a = (\tilde{m}_{aj})$ que minimice:

$$\sum_j \left[\alpha_{aj} \frac{(n_{aj} - \tilde{m}_{aj})^2}{n_{aj}} + (1 - \alpha_{aj}) \frac{(\hat{m}_{aj} - \tilde{m}_{aj})^2}{\hat{m}_{aj}} \right]$$

sujeto a restricciones $\sum_j \tilde{m}_{aj} = M_{a\bullet}$ dados.

α_{aj} puede ser escogido para garantizar que las estimativas sean cercanas a $\hat{m}_{aj}$ cuando el tamaño de muestra es grande y a n_{aj} en caso contrario.

SPREE

Y_{aj} variable aleatoria. Frecuencia de la categoría j en el area a.
$\mathbb{E}(Y_{aj}) = \mu_{aj}$.

Es posible representar el tabulado de interés como:

$$\log(\mu_{aj}) = \alpha_{\bullet\bullet} + \alpha_{a\bullet} + \alpha_{\bullet j} + \alpha_{aj}$$

Modelo log-linear saturado. $\hat{\mu}_{aj} = y_{aj}$ para $a = 1, \ldots, A$; $j = 1, \ldots, J$. Por ejemplo, definiendo $z_{aj} = \log(y_{aj})$:

$\alpha_{\bullet\bullet} = \bar{z}_{\bullet\bullet}$; $\alpha_{a\bullet} = \bar{z}_{a\bullet} - \bar{z}_{\bullet\bullet}$; $\alpha_{\bullet j} = \bar{z}_{\bullet j} - \bar{z}_{\bullet\bullet}$;
$\alpha_{aj} = z_{aj} - \bar{z}_{a\bullet} - \bar{z}_{\bullet j} - \bar{z}_{\bullet\bullet}$.

Restricciones $\sum_a \alpha_{a\bullet} = 0$; $\sum_j \alpha_{j\bullet} = 0$; $\sum_j \alpha_{aj} = 0$; $\sum_a \alpha_{aj} = 0$.

SPREE

$$\log(\mu_{aj}) = \alpha_{\bullet\bullet} + \alpha_{a\bullet} + \alpha_{\bullet j} + \alpha_{aj}$$

asignación + asociación

- Las márgenes del tabulado determinan completamente la estructura de asignación. Pueden ser estimadas con precisión vía proyecciones poblacionales y agregados muestrales.
- Conocidas las márgenes, solo resta modelar la estructura de asociación (interacciones).

SPREE

Purcell & Kish (1980)

Fuentes: 'Proxy' X + márgenes actualizadas $Y_{a\bullet}$, $Y_{\bullet j}$.
Supuesto: $\alpha^{Y}_{aj} = \alpha^{X}_{aj}$
Método:

- IPF $\rightarrow$ Raking multiplicativo. Por ejemplo:

$$Y^{(1)}_{aj} = X_{aj} \times \frac{Y_{\bullet j}}{X_{\bullet j}} \quad ; \quad Y^{(2)}_{aj} = Y^{(1)}_{aj} \times \frac{Y_{a\bullet}}{Y^{(1)}_{a\bullet}}$$

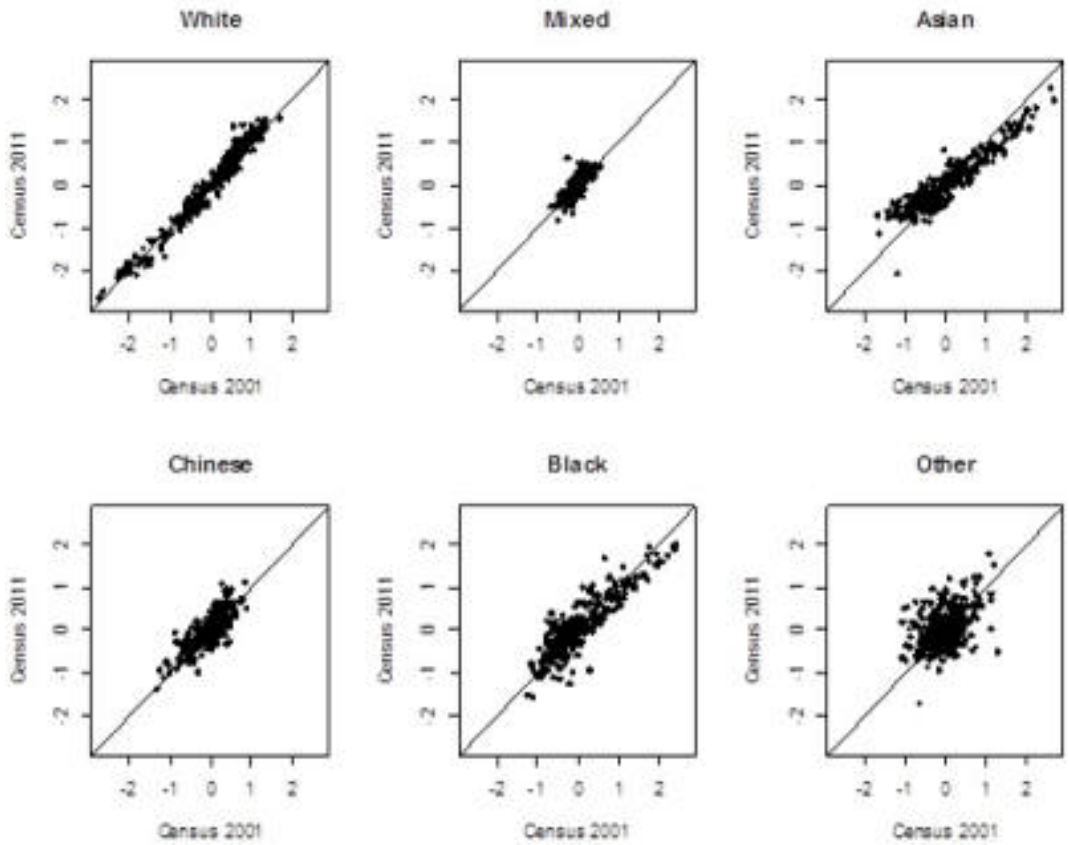
- Modelo log-lineal: Genere $\tilde{X}$ usando $Y_{a\bullet}$, $Y_{\bullet j}$ y asumiendo independencia y ajuste un modelo de efectos principales con offset α^{X}_{aj}.

Variables asociadas sirven para definir tablas de contingencia de más de dos vías. Uso de márgenes parciales.

GSPREE

Zhang & Chambers (2004)

Fuentes: 'Proxy' X + estimativa muestral $\hat{Y}$ + márgenes $Y_{a\bullet}$, $Y_{\bullet j}$.
Supuesto: $\alpha_{aj}^{Y} = \beta\alpha_{aj}^{X}$



GSPREE

Zhang & Chambers (2004)

Método:

- Generalized Structural Linear Model (Zhang & Chambers, 2004)
- Modelo Multinomial: $\pi_{aj} = \mathrm{P}(\text{estar en categoría } j | \text{ area } = a)$.

$$\log\left(\frac{\pi_{aj}}{\pi_{aJ}}\right) = \phi_j + \beta(\alpha^X_{aj} - \alpha^X_{aJ}); \quad j = 1, \ldots, J-1$$

- Modelo Poisson: $\mu^Y_{aj} = \mathbb{E}(Y_{aj})$

$$\log(\mu^Y_{aj}) = \gamma_a + \lambda_j + \beta\alpha^X_{aj}$$

Si las estimativas provienen de una muestra compleja, el modelo puede ajustarse maximizando una Quasi-Verosimilitud usando IWLS.
Estimador de efectos mixtos con efectos aleatorios a nivel de celda.

MSPREE

Luna-Hernandez (2016)

Fuentes: 'Proxy' X + estimativa muestral $\hat{Y}$ + márgenes $Y_{a\bullet}$, $Y_{\bullet j}$.
Supuesto:

$$\begin{bmatrix} \alpha^{Y}_{a1} \\ \vdots \\ \alpha^{Y}_{aJ} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \dots & \beta_{1J} \\ \vdots & \ddots & \vdots \\ \beta_{J1} & \dots & \beta_{JJ} \end{bmatrix} \begin{bmatrix} \alpha^{X}_{a1} \\ \vdots \\ \alpha^{X}_{aJ} \end{bmatrix}$$

Método: Modelos Multinomial o Poisson, Quasi-verosimilitud.
Estimador de efectos mixtos con efectos aleatorios a nivel de celda.

Habitantes por autoridad local según grupo étnico

- Grupo étnico es una de las variables identificadas como clave en la consulta de tópicos para el Censo 2021 (ONS-UK)
- Estimaciones requeridas para asignación de recursos y monitoreo de políticas antidiscriminación, entre otros.
- ONS produjo proyecciones experimentales por grupo étnico (componentes) 2006-2011. Descontinuadas por mostrar grandes diferencias con otras fuentes.
- Trabajo conjunto para evaluar el potencial de uso de SPREE para producir estas estimativas.

Fuentes de datos

Tabla 'proxy'

- Censo de población. Tabulado desagregado por grupo étnico, autoridad local, edad (3 grupos) y sexo. Cobertura de 93 % (2011).
- Censo escolar de Inglaterra. Periodicidad anual. Cubre la población 2-19 años, con prácticamente 100 % de cobertura para 5-15 años.

Encuestas por muestreo

- Annual Population Survey (APS). Combina LFS con boost samples. Mayor tamaño de muestra disponible (alrededor de 250.000 personas/año). La muestra incluye todas las autoridades locales.

Márgenes

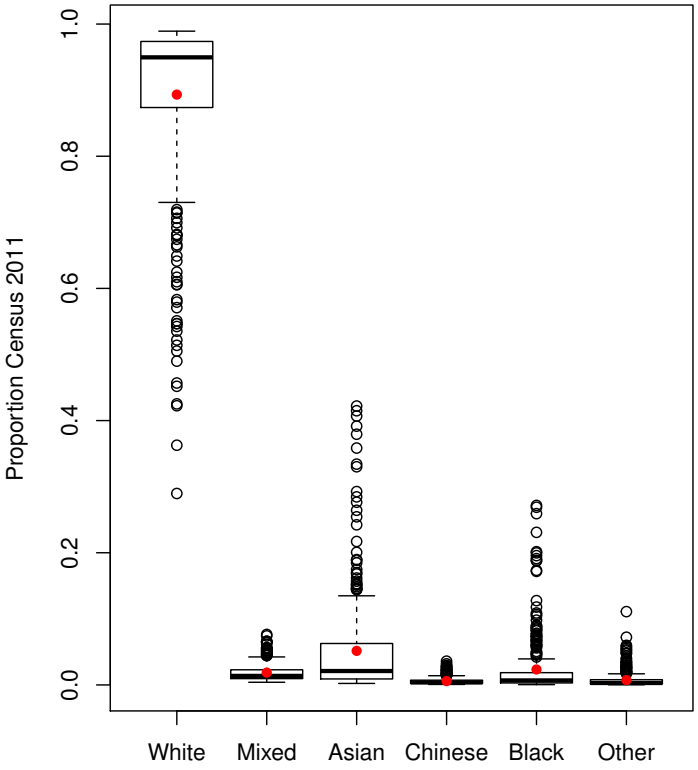
- Filas: Proyecciones de población.
- Columnas: Estimativas de la APS a nivel nacional.

Grupo étnico

- En APS, 15 subcategorías agrupadas en 7 grandes categorías.
- En Censo 2011, 18 subcategorías agrupadas en 5 grandes categorías.

Fuente	Blanco	Mixto	Negro	Asiático	Chino	Otro
Censo 2011	X	X	X	—Chino—		X
Censo escolar	X	X	X	X	—Chino—	
APS	X	X	X	X	X	X

Grupo étnico



Categoría	Censo11	APS10-11	Dif.rel.
Blanco	85.42	86.44	+1
Mixto	2.25	1.44	-36
Asiático	7.10	6.55	-8
Chino	0.72	0.51	-29
Negro	3.48	3.33	-4
Otro	1.03	1.74	+69

Márgenes:
Fila: Proyecciones poblacionales (componentes).
Columna: APS Nacional.

Cálculos propios.

Resultados

- Aún no constituyen estadísticas oficiales.
- Estimativas para 2013. Evaluación de fuentes de datos y modelos. (Luna et al., 2015).
- Ejercicio de validación Censo 2011 (ONS, 2017a).
- Estimativas para 2015. Evaluación de fuentes de datos y modelos. Intervalos de confianza. (ONS, 2017b).

Zero estimativas

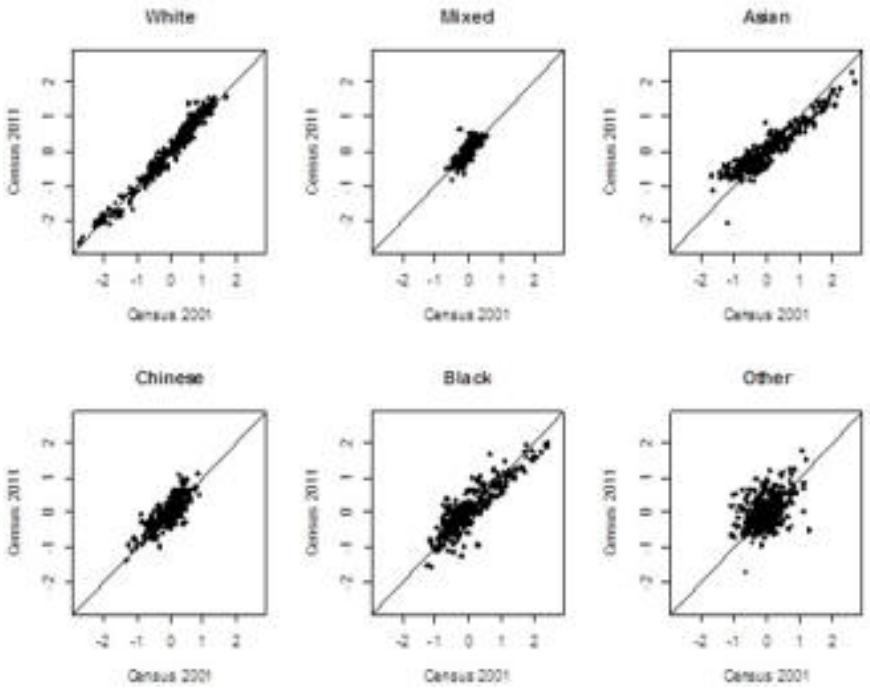
Numero de 0's por area	Frecuencia	%
Ninguna	155	44.80
Una	61	17.63
Dos	55	15.90
Tres	34	9.83
Cuatro	27	7.80
Cinco	14	4.04

Categoría	Estimativas 0 (areas)	%
Blanco	0	0.00
Mixto	58	15.93
Asiático	43	11.81
Chino	156	42.86
Negro	96	26.37
Otro	98	26.92

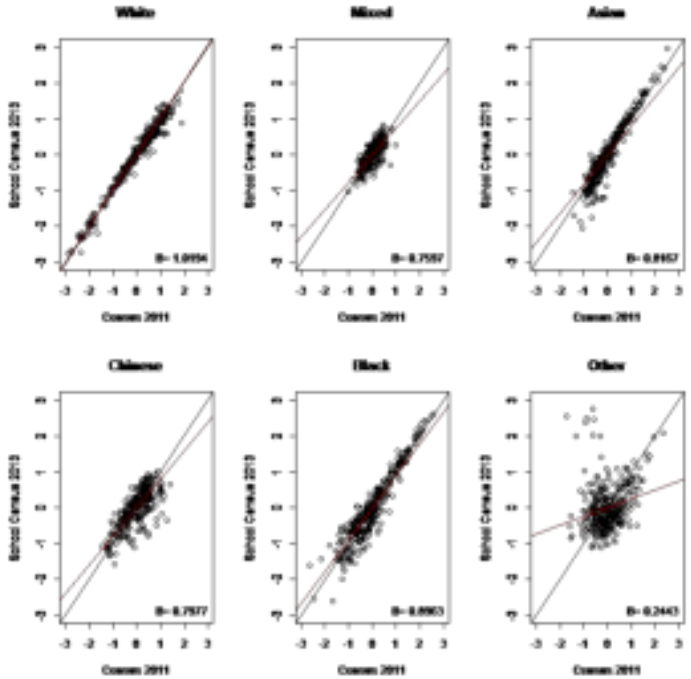
Indicador	Categoría					
	Blanco	Mixto	Asiático	Chino	Negro	Otro
cve(mediana)	0.0157	0.4702	0.4769	0.8228	0.6153	0.6797
cve (media)	0.0192	0.5675	0.6369	0.9219	0.6966	0.7850
prop (media)	0.8985	0.0131	0.0491	0.0044	0.0223	0.0126

Fuente: Luna-Hernandez (2016), APS 2012-13

Asociación



(a) Census 2001 vs Census 2011

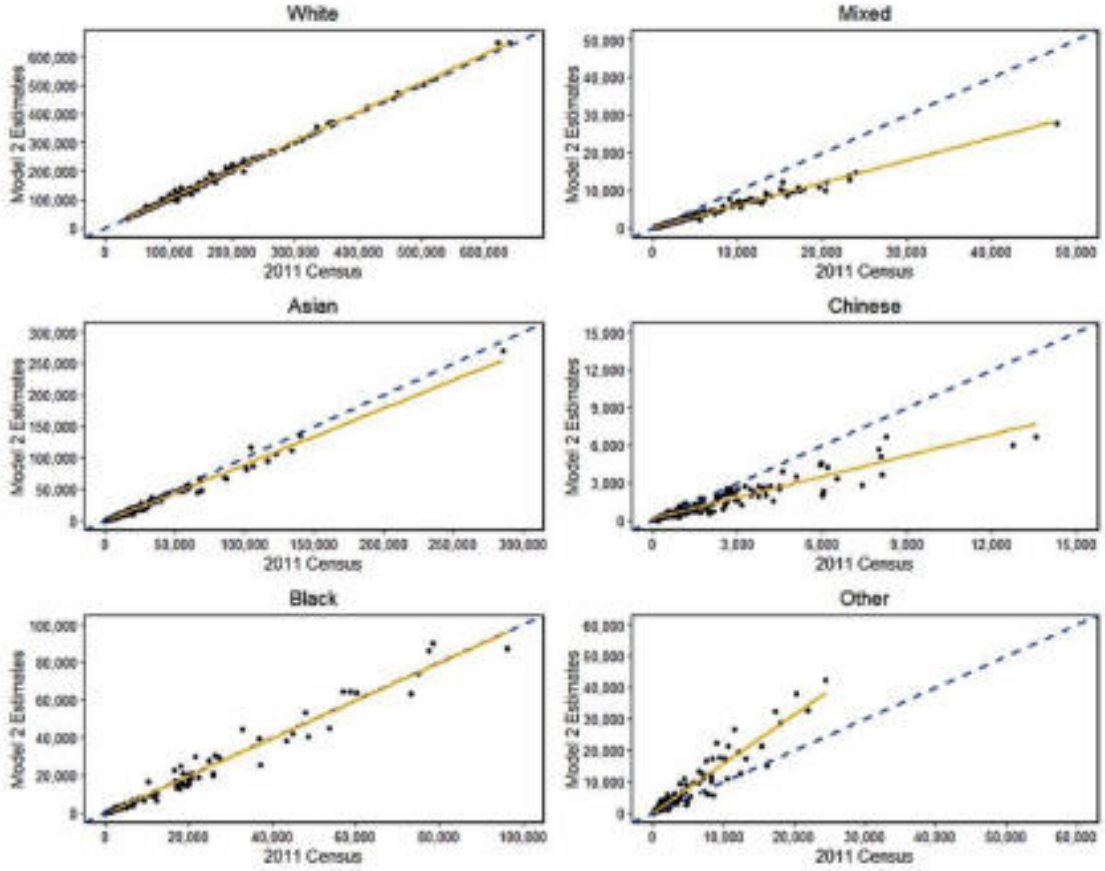


(b) Census 2011 vs School census 2013

Fuente: (a) Luna et al. (2015). (b) Cálculos ONS.

Sesgo

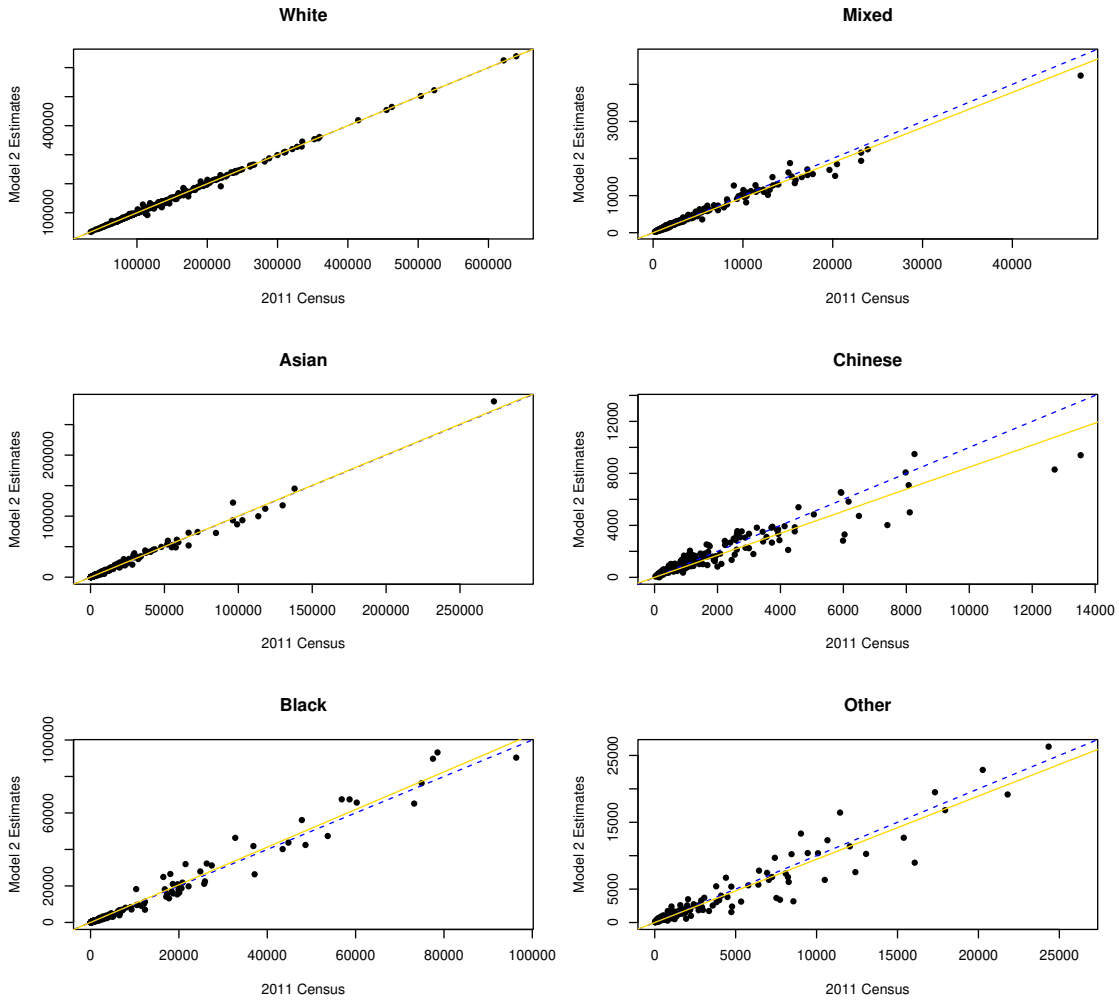
GSPREE 2011. Fuentes: Censo 2001, Censo escolar 2011, APS 2010-11.
Benchmark: APS 2010-11.



Fuente: ONS (2017a)

Sesgo

Márgenes? Benchmark a Censo 2011



Cálculos ONS.

Jerarquía de fuentes de datos

Peso del Censo Poblacional y el Censo escolar en la tabla 'proxy' final

$$\alpha_{aj}^{X} = \delta\alpha_{aj}^{CE} + (1-\delta)\alpha_{aj}^{CP}$$

Peso	2011-2013 (a)			2001-2011 (b)		
	0-4	5-15	16+	0-4	5-15	16+
Censo escolar	0.22	0.323	0.087	0.829	0.853	0.383
Censo poblacional	0.78	0.677	0.913	0.171	0.147	0.617

Fuente: (a) Luna et al. (2015) y (b) ONS (2017a)

ECM

Bootstrap paramétrico bajo un modelo con efectos aleatorios.

Estadísticas para MSE(GSPREE)/MSE(Direct)

	Blanco	Mixto	Asiático	Chino	Negro	Otro
Q1	0.147	0.116	0.105	0.230	0.078	0.359
Median	0.309	0.216	0.235	0.395	0.214	0.637
Mean	0.437	0.288	0.404	0.462	0.402	0.843
Q3	0.582	0.357	0.562	0.600	0.596	1.138

1/2 ancho IC(95 %) para GSPREE 2013 en puntos porcentuales

	Blanco	Mixto	Asiático	Chino	Negro	Otro
1/2 IC (media)	1.49	0.33	1.10	0.29	0.61	0.97
Estimativa Nacional	86.43	1.53	6.68	0.55	3.15	1.66
Media estimativas	90.02	1.28	4.91	0.45	2.14	1.19

Cálculos propios.

Lecciones aprendidas

- Relevancia de contar con varias fuentes de información
- Impacto de diferencias en definiciones u operacionales. Cuál es el parámetro que puedo estimar?
- Ejercicios de validación. Utiles para 'calibrar' la metodología y comunicar confianza a los usuarios.
- Aproximación top to bottom? Benchmarking?

Siguientes pasos

- Actualmente en desarrollo de software (conjuntamente con ONS)
- Validez para geografías más granulares? Simulaciones preliminares sugieren bajo riesgo de sesgo en la estimación de β. Mayor varianza puede comprometer la calidad del estimador.
- Múltiples encuestas? Actualización periódica
- Otros modelos para datos composicionales?

Bibliografía

Dostál, L., Münnich, R., Gabler, S. & Ganninger, M., (2016). *Frame correction modelling with applications to the German register-assisted census 2011.* Scandinavian Journal of Statistics, 43(3): 904-920.

Little, R.J.A., & Wu, M.M., (1991). *Models for contingency tables with known margins when target and sampled populations differ.* Journal of the American Statistical Association 86(413): 87-95.

López-Vizcaíno, E., Lombardía, M. J., & Morales, D., (2013). *Multinomial-based small area estimation of labour force indicators.* Statistical modelling, 13(2): 153–178.

López-Vizcaíno, E., Lombardía, M. J., & Morales, D., (2015). *Small area estimation of labour force indicators under a multinomial model with correlated time and area effects.* Journal of the Royal Statistical Society: Series A, 178(3): 535–565.

Luna, A., Zhang, L-C., Whitworth, A., Piller, K., (2015). *Small area estimates of the population distribution by ethnic group in England: a proposal using structure preserving estimators.* Statistics in Transition new series 16(4): 585-602.

Bibliografía (2)

Luna-Hernandez, A., (2016). Multivariate Structure Preserving Estimation for population compositions. Unpublished PhD thesis. Department of Social Statistics and Demography, University of Southampton.

Molina, I., Rao, J.N.K., & Hidiroglou, M.A., (2008). *SPREE techniques for small area estimation of cross-classifications.* Unpublished manuscript.

Molina, I., Saei, A., & Lombardía, M. J., (2007). *Small area estimates of labour force participation under a multinomial logit mixed model.* Journal of the Royal Statistical Society: Series A, 170(4): 975–1000.

Office for National Statistics, (2017a). *Research Outputs: An approach for estimating ethnicity from survey and administrative data, 2011*. Available online at: `https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/populationcharacteristics/researchoutputsanapproachforestimatingethnicityfromsurveyandadministrativedata2011`.

Bibliografía (3)

Office for National Statistics, (2017b). *Ethnicity estimates from survey and administrative data, 2015*. Available online at: `https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/populationcharacteristics/researchoutputsethnicityestimatesfromsurveyandadministrativedata2015`.

Purcell, N. J. & Kish, L., (1980). *Postcensal estimates for local areas (or domains).* International Statistical Review, 48(1): 3–18.

Rao, J. N. K. & Molina, I., (2015). *Small area estimation.* John Wiley & Sons, 2nd edition.

Saei, A. & Taylor, A., (2012). *Labour force status estimates under a bivariate random components model.* Journal of the Indian Society of Agricultural Statistics, 66(1): 187–201.

Bibliografía (4)

Scealy, J., (2010). *Small area estimation using a multinomial logit mixed model with category specific random effects.* Research paper, Australian Bureau of Statistics. Retrieved from `http://www.abs.gov.au/ausstats/abs@.nsf/cat/1351.0.55.029`.

Suesse, T., Namazi-Rad, M. R., Mokhtarian, P., & Barthélemy, J., (2017). *Estimating cross-classified population counts of multidimensional tables: an application to regional Australia to obtain pseudo-census counts.* Journal of Official Statistics 33(4): 1021-1050.

Zhang, L.-C. & Chambers, R. L., (2004). *Small area estimates for cross-classifications.* Journal of the Royal Statistical Society: Series B, 66(2): 479–496.