Islas Canarias
Del 15 al 19 de noviembre de 2021

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA

XXI Jornadas de Estadística de las Comunidades Autónomas

MODELIZACION DE LOS DATOS DE LA ENCUESTA ESIE

Jorge Aramendi Rique
Marina Ayestaran Arregi

Introducción

Las encuestas están diseñadas para producir datos a un nivel de agregación determinado, para dominios/áreas, bien geográficas, bien socio-económicas, de la población.

Estimaciones de calidad a un nivel de agregación determinado

Demanda a un nivel más desagregado.

Métodos indirectos de estimación que no consideran solamente los datos muestrales relativos al dominio o área de interés, sino que además usan los datos muestrales provenientes de otras áreas o dominios. Utilización de información auxiliar relacionada con la variable de interés.

Estimaciones basadas en modelos a nivel de unidad, regresiones logísticas, que tienen en cuenta el diseño muestral de cada una de las encuestas.

Metodología empleada

Se han estimado variables discretas

ESIE
A16 (tiene internet sí/no)
C1 (tiene página web sí/no)
C21 (tiene página web en euskera sí/no)
B4 (hace trámites online sí/no)

- Metodología que tiene en cuenta la estructura de los datos, teniendo en cuenta el diseño y los estratos.
- Se ha definido un modelo para cada variable con las variables auxiliares disponibles.
- Para cada variable respuesta los análisis se han realizado período a período, 7 años.
- Se ha analizado la bondad de ajuste y capacidad predictiva de cada uno de los modelos y se ha optado por una estabilidad temporal de los modelos analizados.

Metodologia Estadística

- **Estimación**
- **Ajuste del modelo**
- **Capacidad Predictiva**
- **Validación del ajuste**
- **Extensión de la probabilidad al conjunto de la población**

Estimación

El diseño muestral complejo, las unidades tienen diferente probabilidad de selección y por lo tanto diferente peso/elevador y esto se ha tenido en cuenta en la fase de ajuste. En concreto en la función de máxima verosimilitud:

- **Para ello se define una función de pseudo-verosimilitud**

$$l_D(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i{}^{y_i w_i} (1 - p_i)^{(1-y_i)w_i}$$

donde w_i es el peso de cada unidad. Los pesos se tienen en cuenta a la hora de ajustar los modelos y obtener los parámetros. Un modelo de regresión logística

El ajuste del modelo se ha realizado utilizando el paquete `survey` de R. Que tiene en cuenta la pseudo-verosimilutud. ***"Lumley T (2011). Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R. John Wiley and Sons"***

Ajuste del modelo

- Para cada año/trimestre y variable se ha ajustado un modelo de regresión logística
- Se han validado todas las variables explicativas disponibles y se ha optado por una estabilidad temporal
- Se han seleccionado los mejores modelos en función del AIC eta MSE mas bajos y la capacidad predictiva (AUC mas alto)

- Podemos definir MSE como:

$$MSE = \frac{\sum_{r=1}^{k}\left[\left(\sum_{i=1_r}^{n_r} y_i\right) - n_r \hat{p}_r\right]^2}{\sum_{r=1}^{k} n_r}$$

Con, $\hat{p}_r,\ r = 1, \dots, k$ las probabilidades estimadas por el modelo, e $y_i,\ i = 1_r, \dots, n_r$ los valores observados.

- Y el parametro AIC , que tiene en cuenta el muestreo y el diseño (Lumley & Scott, 2015):

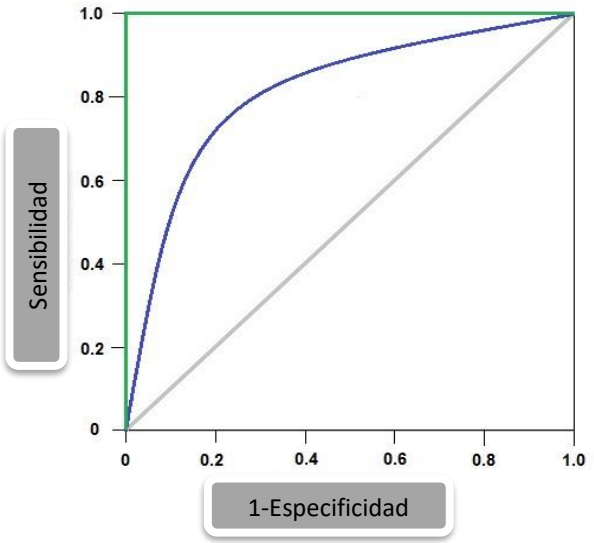
$$AIC = -2nl_D(\boldsymbol{\beta}) + 2q\hat{\tilde{\delta}}$$

donde $\hat{\tilde{\delta}}$ es la traza de la matriz del diseño y q el número de variables explicativas del modelo.

Capacidad predictiva del modelo

El parámetro AUC (*area under the ROC curve*) mide la capacidad predictora del modelo

AUC tiene la capacidad de medir y diferenciar los individuos que han sido bien clasificados y no



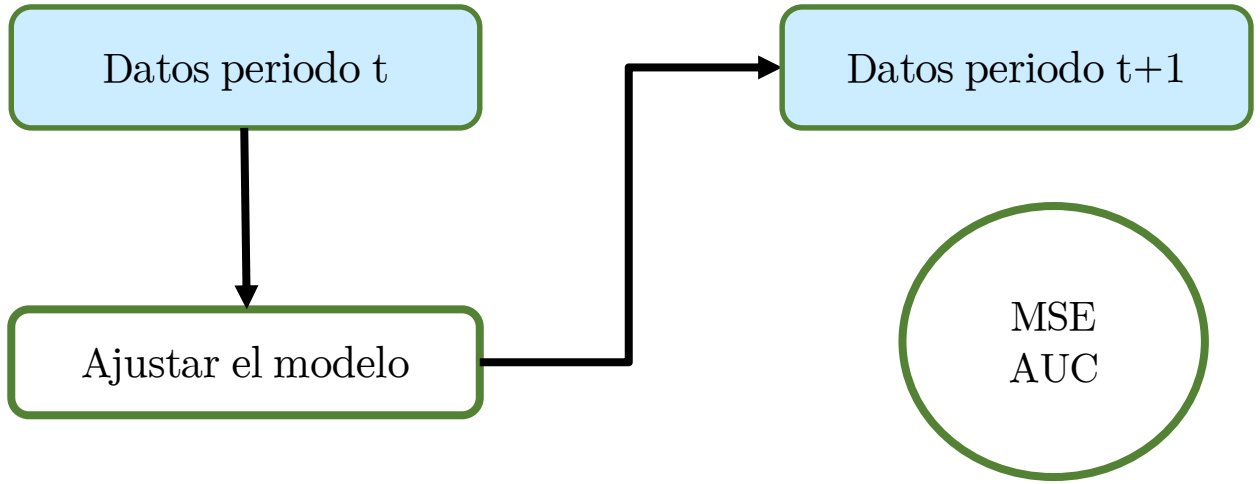
	Clasificación	
Respuesta	$\hat{p}(x_i) < c$	$\hat{p}(x_i) \geq c$
$Y = 0$	Negativo Verdadero (NV)	Falso Positivo (FP)
$Y = 1$	Falso Negativo (FN)	Positivo Verdadero (PV)

$\mathbf{Especificidad} = \mathbf{P(p(X) < c | Y = 0) = NV/(NV + FP)}$ -> **Detecta realmente los casos negativos**

$\mathbf{Sensibilidad} = \mathbf{P(p(X) \geq c | Y = 1) = PV/(PV + FN)}$ -> **Detecta realmente los casos positivos**

Validación del modelo

Aplicar el modelo a los datos del periodo siguiente



Y volver a calcular el MSE y el AUC, para validarlo.

Aplicación a la Encuesta ESI

- Objetivo: Proporcionar a los responsables políticos, agentes económicos y sociales, universidad, investigadores privados y ciudadanía en general, información periódica sobre la penetración de las nuevas tecnologías de la información y de la comunicación -TIC- en las empresas del País Vasco.

- Una encuesta anual. Panel con renovación anual

- Estratificación: Territorio Histórico x Actividad (2dig de la CNAE) x estrato de empleo (6 categorías)

- Elevación: Se reducen los estratos de empleo a 3 y se hace una corrección de no respuesta.

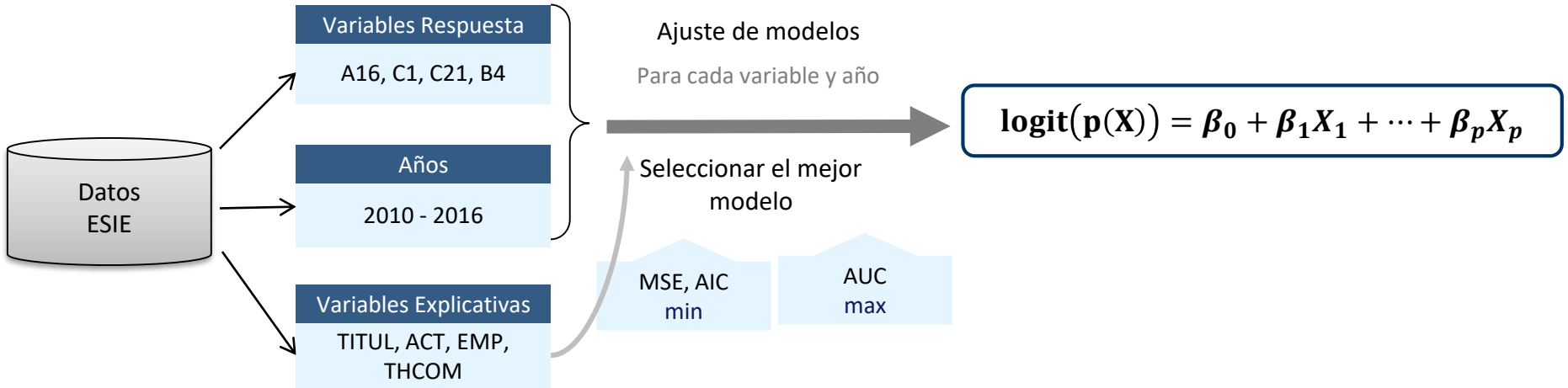
- Esta encuesta esta preparada para obtener resultados de calidad para los TH-s y a nivel de actividad para la CAE
- Problema: obtener estimaciones para variables principales a nivel de Comarca geográfica (20 + 3 capitales)
- Información auxiliar: DIRAE, Directorio de Actividades Económicas

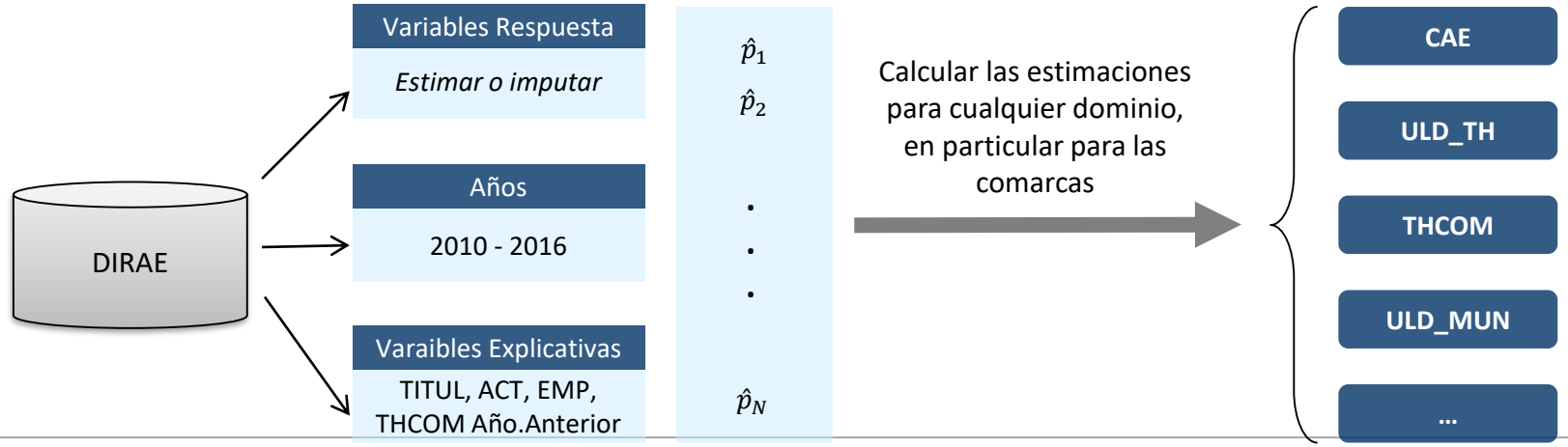
Aplicación a la Encuesta ESI

- Agregaciones de la CNAE
- Empleo estratificado: 1) 0-2, 2) 3-9, 3) 10-19, 4) 20-49, 5) 50-99, 6) 100-
- Titularidad
- Presencia en el estado
- Estimación sintética de la variable en el año anterior.

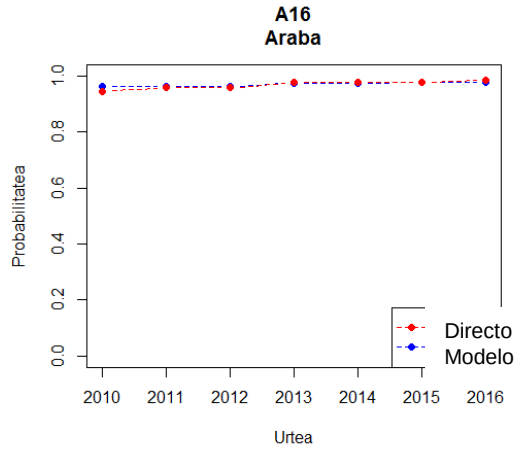
	2010	2011	2012	2013	2014	2015	2016
THCOM23	0.002	0.072	0.775	0.266	0.112	0.029	0.640
A21	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.003
EMP	<0.001	<0.001	<0.001	0.009	<0.001	0.023	<0.001
TITUL	<0.001	0.001	0.005	0.061∗	0.002	0.058∗	0.009
A16.2		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
AIC	7152	4064	4139	3844	3483	3420	3316
MSE	0.117	0.050	0.057	0.049	0.043	0.038	0.039
	(0.118)	(0.067)	(0.053)	(0.048)	(0.047)	(0.049)	
AUC	0.864	0.956	0.946	0.948	0.947	0.942	0.925
	(0.866)	(0.940)	(0.943)	(0.944)	(0.930)	(0.904)	

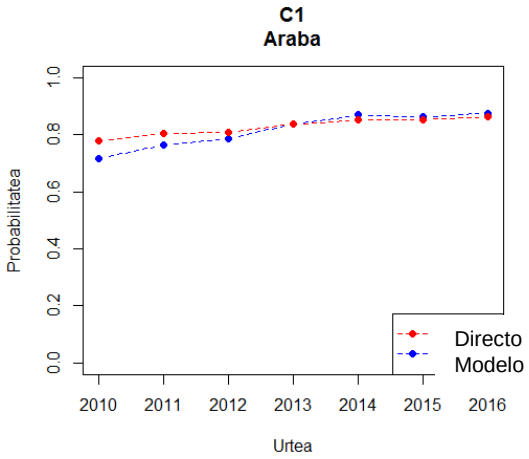
ESIE

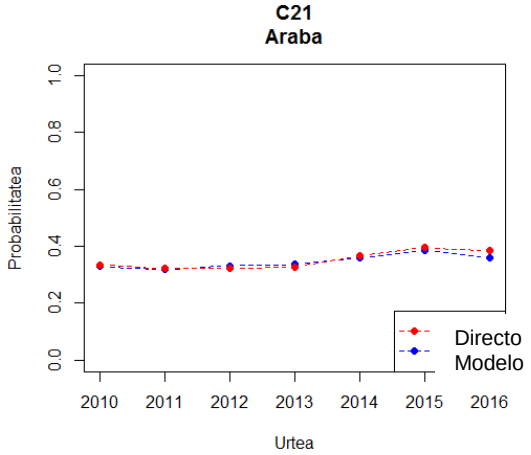

Variables Respuesta
A16, C1, C21, B4
Años
2010 - 2016
Variables Explicativas
TITUL, ACT, EMP, THCOM
Datos ESIE
Ajuste de modelos
Para cada variable y año
Seleccionar el mejor modelo
MSE, AIC min
AUC max
$\mathbf{logit}(\mathbf{p}(\mathbf{X})) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \cdots + \boldsymbol{\beta_p X_p}$


Variables Respuesta
Estimar o imputar
Años
2010 - 2016
Varaibles Explicativas
TITUL, ACT, EMP, THCOM Año.Anterior
DIRAE
$\hat{p}_1$
$\hat{p}_2$
$\hat{p}_N$
Calcular las estimaciones para cualquier dominio, en particular para las comarcas
CAE
ULD_TH
THCOM
ULD_MUN
...

Resultados para Álava

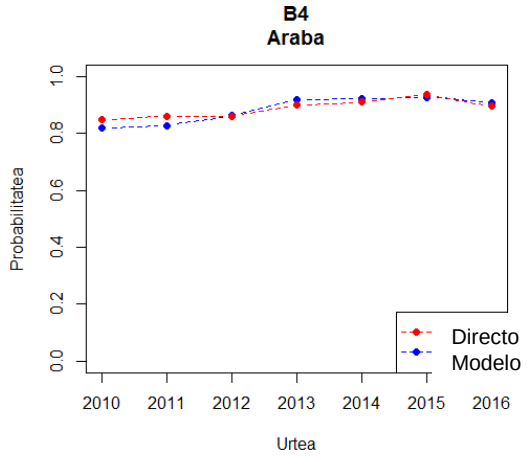




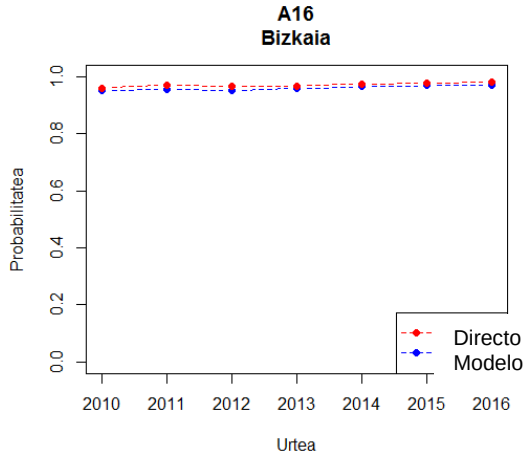


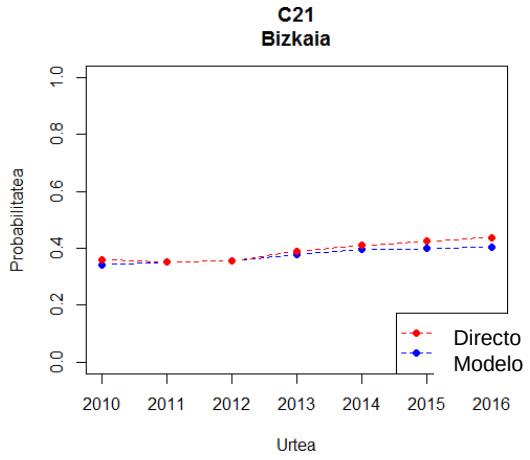
Diferencias entre la probabilidad estimada y la estimación directa

	2010	2011	2012	2013	2014	2015	2016
A16	0.015	0.006	0.005	-0.004	-0.003	0.000	-0.008
C1	-0.062	-0.039	-0.023	0.000	0.017	0.010	0.013
C21	-0.005	-0.003	0.008	0.008	-0.007	-0.011	-0.026
B4	-0.030	-0.033	0.002	0.021	0.012	-0.008	0.011



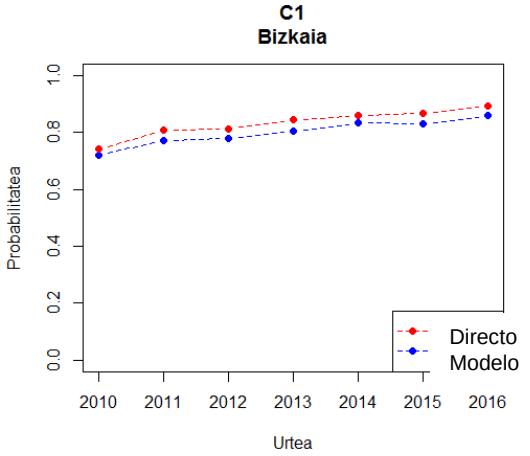
Resultados para Bizkaia

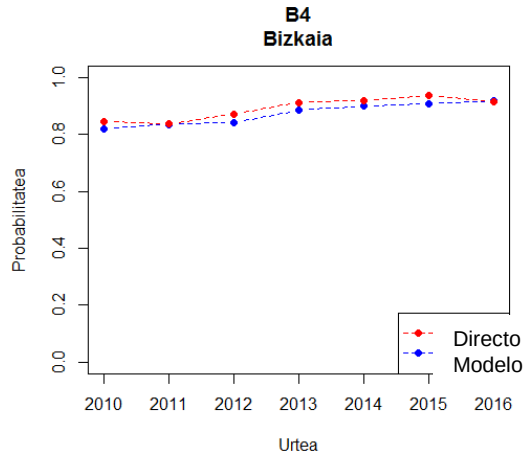




Diferencias entre la probabilidad estimada y la estimación directa

	2010	2011	2012	2013	2014	2015	2016
A16	-0.009	-0.015	-0.014	-0.007	-0.009	-0.008	-0.012
C1	-0.021	-0.038	-0.034	-0.039	-0.026	-0.035	-0.034
C21	-0.017	0.000	0.002	-0.012	-0.014	-0.027	-0.033
B4	-0.025	-0.001	-0.030	-0.026	-0.019	-0.028	0.004

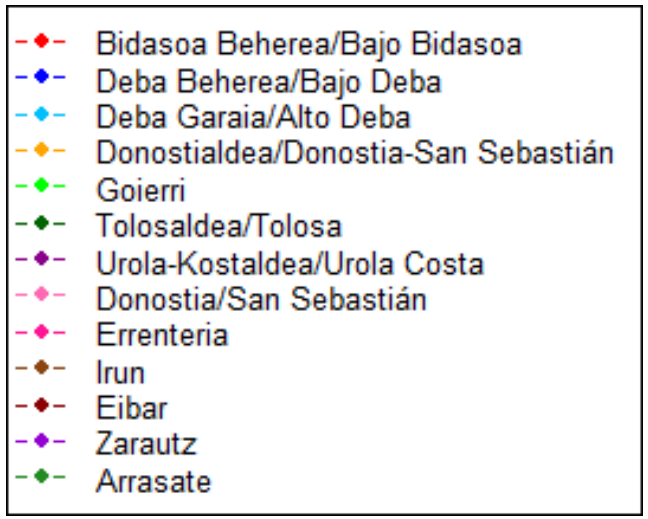


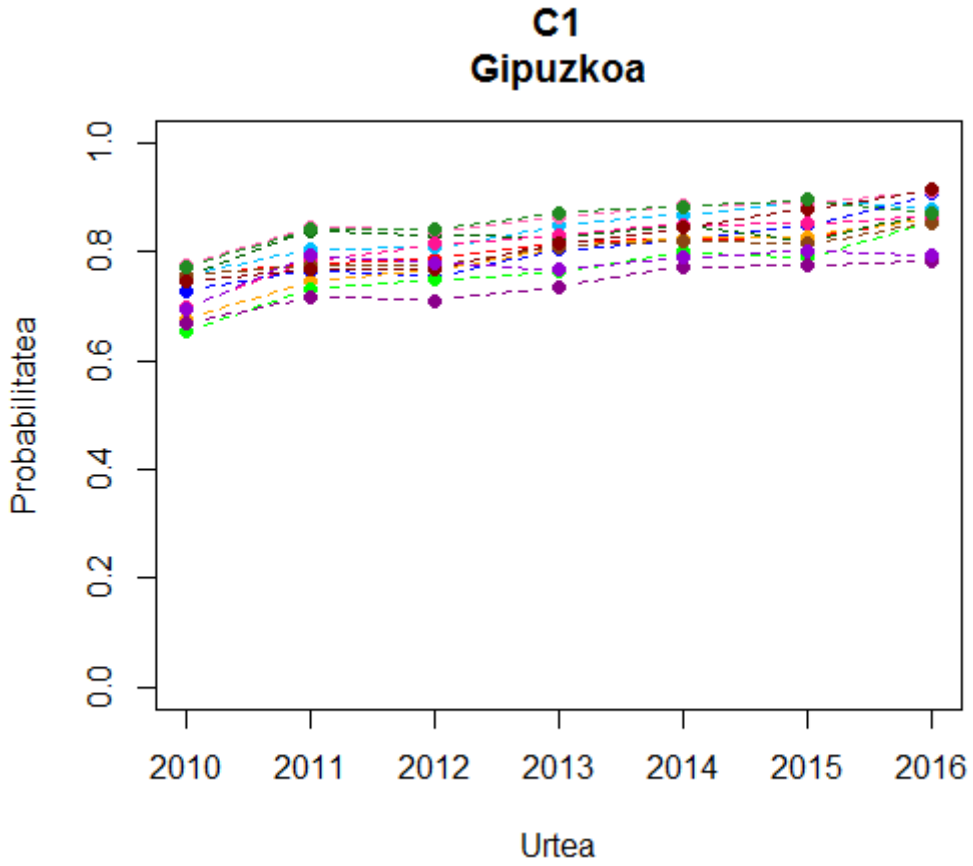


Resultados por comarcas Gipuzkoa

Variable respuesta: C1
Variables explicativas:

- THCOM23
- A21
- EMP (5 categorías)
- TITUL
- C1.2





PUNTOS FUERTES

- Con esta metodología se pueden obtener resultados para cualquier dominio de la población.
- Se ha tenido en cuenta el diseño del muestreo en la estimación del modelo.
- La capacidad predictiva del modelo es muy alta.

PUNTOS DÉBILES

- La información relativa a las variables predictoras que se quieran introducir en el modelo tiene que estar disponible en DIRAE para poder obtener estimaciones en cualquier dominio de la población.

CALIDAD DE LOS RESULTADOS

- **Mediante un estudio de simulación se ha estudiado la calidad de las estimaciones obtenidas a partir de los modelos desarrollados**
- **No existe una población de referencia por lo que se ha simulado una población para cada una de las variables y se han realizado 500 muestreos siguiendo el diseño de la encuesta.**
- **Para cada una de estas muestras se ha ajustado el modelo y se han obtenido las estimaciones.**

CREACION: Pseudo población

Se ha creado la Pseudo-población estratificando y aplicando aleatoriamente los resultados de la encuesta. No se ha utilizado el mismo modelo que para estimar , para evitar sobre ajustes de los datos

CALIDAD DE LOS RESULTADOS

- **Se han estimado las probabilidades para cada una de las variables explicativas a partir de las r=1,…,500 muestras y se han extendido a la población de referencia, para cada uno de los subdominios estudiados.**
- **Parámetros para medir la calidad**
  - **Sesgo medio**

$$Sesgo_D = \frac{\sum_{r=1}^{500} \hat{p}_r^D}{500} - p_{true}^D$$

  - **Sesgo absoluto medio**

$$Sesgo.Abs_D = \frac{\sum_{r=1}^{500} \left| \hat{p}_r^D - p_{\boldsymbol{true}}^{\boldsymbol{D}} \right|}{500}$$

  - **Varianza**

$$Varianza_D = \frac{\sum_{r=1}^{500} (\hat{p}_r^D)^2}{500} - \left[\frac{\sum_{r=1}^{500} \hat{p}_r^D}{500} \right]^2$$

  - **Error cuadrático medio (mse)**

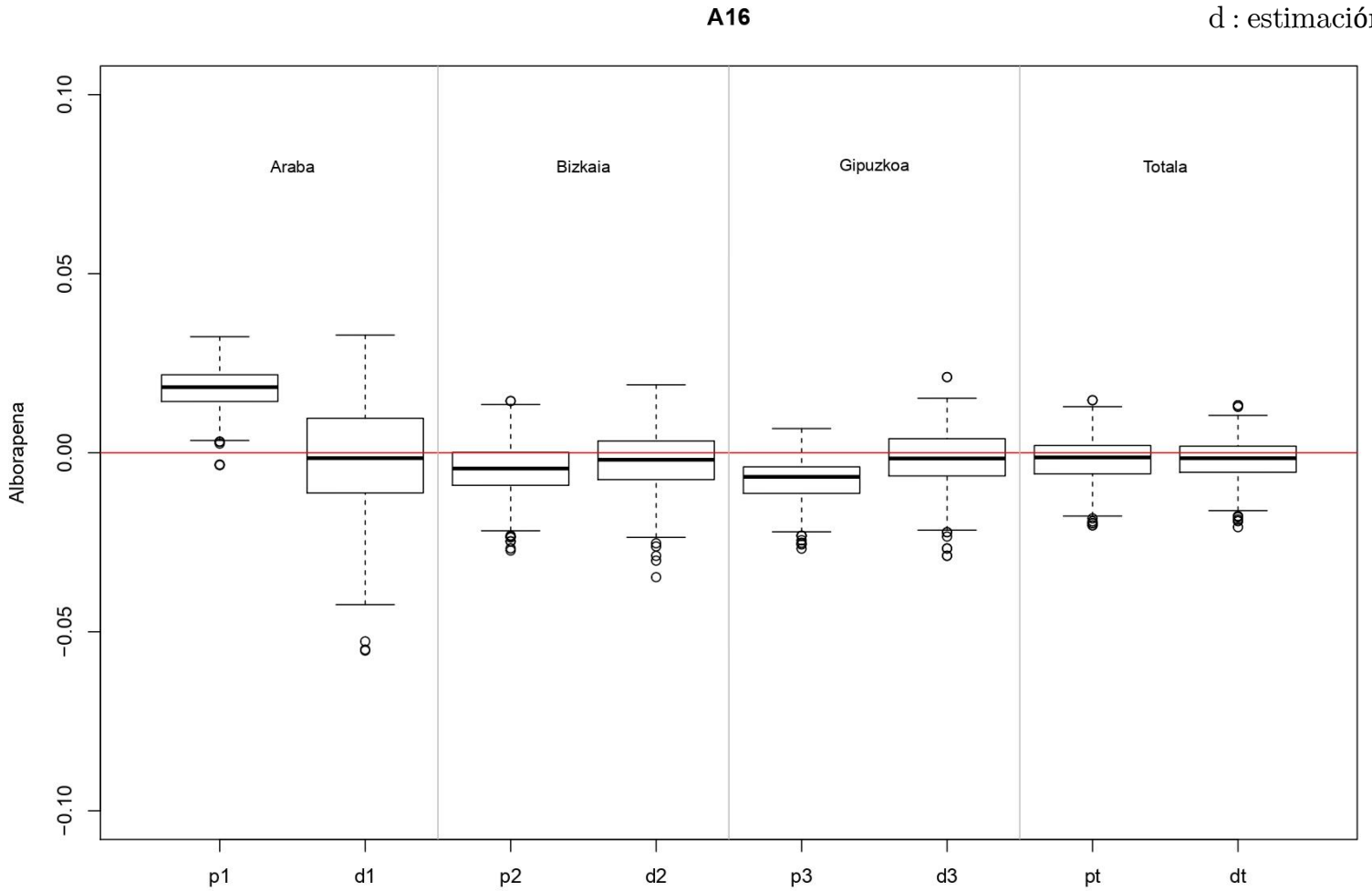
$$Mse_D = \frac{\sum_{r=1}^{500} (\hat{p}_r^D - p_{true}^D)^2}{500}$$

donde,

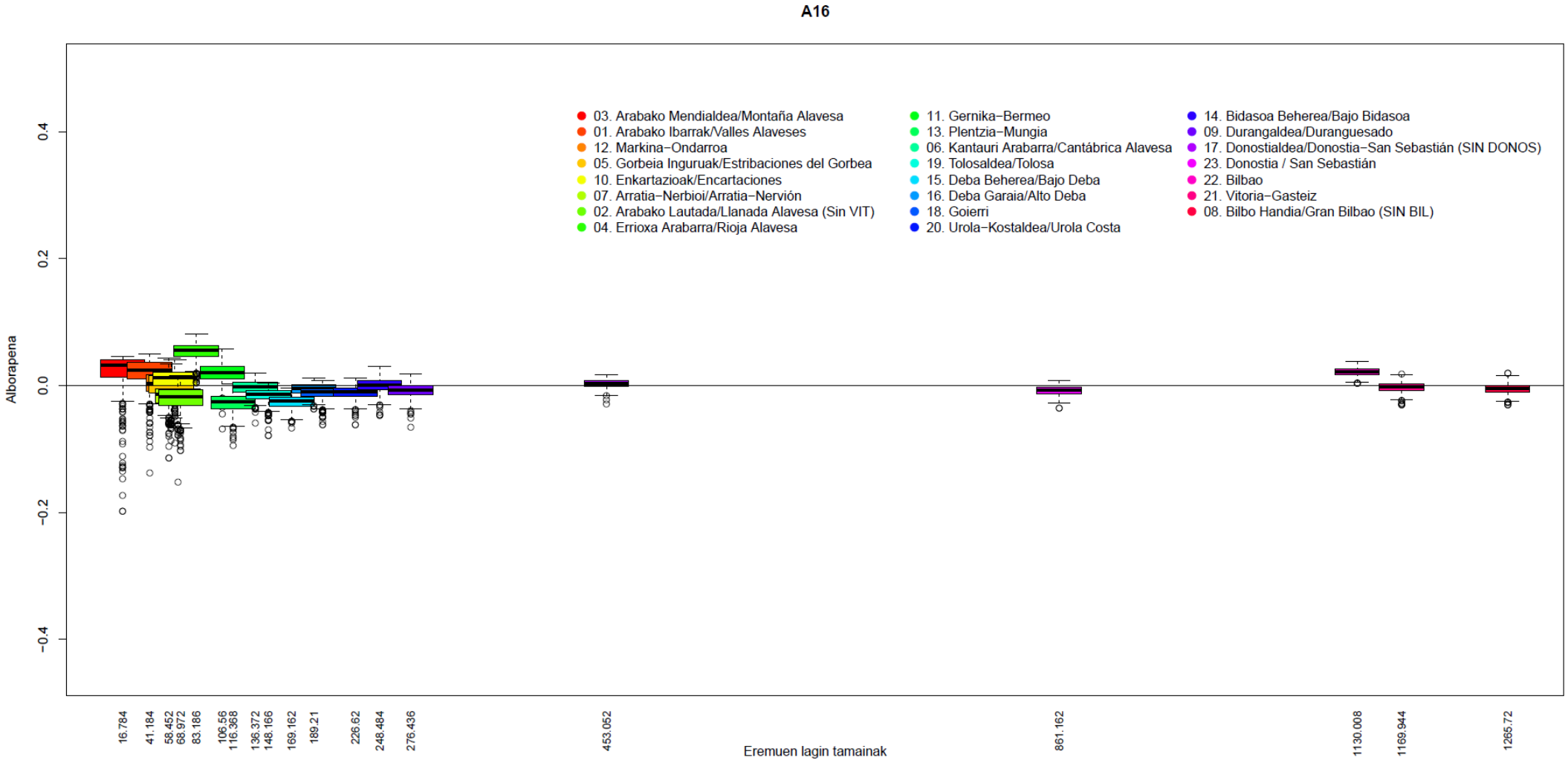
$$\hat{p}_r^D = \frac{\sum_{i=1}^{N_D} \hat{p}_{r,i}}{N_D} \qquad r = 1, \dots, 500 \qquad p_{true}^D = \frac{\sum_{i=1}^{N_D} y_i}{N_D} \qquad Mse_D = {Sesgo_D}^2 + Varianza_D$$

CALIDAD DE LOS RESULTADOS

p : estimación del modelo
d : estimación directa



CALIDAD DE LOS RESULTADOS



CALIDAD DE LOS RESULTADOS

	Dominio	Tamaño muestral medio	(min - max)	ptrue	Sesgo		Sesgo abs	
					Media	sd	Media	sd
3	Arabako Mendialdea/Montaña Alavesa	16,78	(9 - 26)	0,9545	0,0182	0,0508	0,036	0,0402
1	Arabako Ibarrak/Valles Alaveses	41,18	(27 - 54)	0,9492	0,0187	0,0248	0,027	0,0154
12	Markina-Ondarroa	58,45	(39 - 80)	0,9524	-0,0002	0,0229	0,0171	0,0152
5	Gorbeia Inguruak/Estribaciones del Gorbea	60,49	(43 - 76)	0,9621	-0,001	0,0214	0,0165	0,0137
10	Enkartazioak/Encartaciones	63,82	(42 - 88)	0,9535	0,0093	0,0169	0,0162	0,0104
7	Arratia-Nerbioi/Arratia-Nervión	66,70	(50 - 84)	0,985	-0,0179	0,0178	0,0189	0,0167
2	Arabako Lautada/Llanada Alavesa (Sin VIT)	68,97	(52 - 91)	0,9806	-0,0216	0,0199	0,0225	0,0188
4	Errioxa Arabarra/Rioja Alavesa	83,19	(63 - 105)	0,9153	0,0529	0,0134	0,0529	0,0134
11	Gernika-Bermeo	106,56	(83 - 132)	0,9353	0,0191	0,0163	0,0218	0,0124
13	Plentzia-Mungia	116,37	(92 - 141)	0,9845	-0,0277	0,0154	0,0277	0,0154
6	Kantauri Arabarra/Cantábrica Alavesa	136,37	(111 - 165)	0,9695	-0,0032	0,0118	0,0094	0,0079
19	Tolosaldea/Tolosa	148,17	(120 - 180)	0,9823	-0,016	0,0118	0,0161	0,0115
15	Deba Beherea/Bajo Deba	169,16	(139 - 202)	0,9925	-0,026	0,0111	0,026	0,0111
16	Deba Garaia/Alto Deba	189,21	(163 - 227)	0,9761	-0,0057	0,0094	0,0083	0,0073
18	Goierri	197,14	(165 - 234)	0,9772	-0,0124	0,0112	0,013	0,0105
20	Urola-Kostaldea/Urola Costa	226,62	(194 - 269)	0,9756	-0,0111	0,0107	0,0119	0,0098
14	Bidasoa Beherea/Bajo Bidasoa	248,48	(216 - 292)	0,9514	-0,0004	0,012	0,0094	0,0075
9	Durangaldea/Duranguesado	276,44	(241 - 321)	0,9638	-0,0082	0,0115	0,0107	0,0093
17	Donostialdea/Donostia-San Sebastián (SIN DONOS)	453,05	(405 - 498)	0,9628	0,0029	0,0071	0,0062	0,0045
23	Donostia / San Sebastián	861,16	(793 - 903)	0,9701	-0,0078	0,0073	0,0084	0,0066
22	Bilbao	1130,01	(1090 - 1178)	0,9436	0,0214	0,0062	0,0214	0,0062
21	Vitoria-Gasteiz	1169,94	(1096 - 1231)	0,9576	-0,0029	0,0079	0,0065	0,0054
8	Bilbo Handia/Gran Bilbao (SIN BIL)	1265,72	(1194 - 1332)	0,96	-0,0054	0,0078	0,0073	0,0061

CALIDAD DE LOS RESULTADOS

- **A16: el sesgo absoluto medio más alto obtenido es de 0,036 con una desviación estándar de 0,04.**

- **B4, C1 y C21: el sesgo absoluto medio crece, estando en todos los casos próximo a 0,13. En cualquier caso, este valor se da en el dominio más pequeño que se ha estudiado, Arabako Mendialdea / Montaña Alavesa (de media tiene un tamaño muestral de 15 individuos). Si dejamos a un lado el dominio Arabako Mendialdea/ Montaña Alavesa, el sesgo absoluto medio más alto se reduce a la mitad.**

- **B4 y C21: se han utilizado dos métodos diferentes para calcular el sesgo. En base a las simulaciones realizadas en los dominios analizados, no se han encontrado diferencias entre ambos métodos. Sin embargo, en el caso de C21, los resultados obtenidos con el método B son ligeramente mejores.**

CALIDAD DE LOS RESULTADOS

- **A16**
  - **Muestra entre 50≤n≤100, desviación media absoluta < 0,025**
  - **Muestra entre 100 < n≤500, desviación media absoluta < 0,02**
  - **Muestra entre n > 500, desviación media absoluta < 0,008**
- **B4**
  - **Muestra entre 50≤n≤100 , desviación media absoluta < 0,08**
  - **Muestra entre 100 < n≤500 , desviación media absoluta ≤ 0,05**
  - **Muestra entre n > 500 , desviación media absoluta < 0,02**
- **C1**
  - **Muestra entre 50≤n≤100 , desviación media absoluta < 0,08**
  - **Muestra entre 100 < n≤500 , desviación media absoluta ≤ 0,05**
  - **Muestra entre n > 500, desviación media absoluta < 0,03**
- **C21**
  - **Muestra entre 50≤n≤100 , desviación media absoluta ≤ 0,05**
  - **Muestra entre 100 < n≤500, desviación media absoluta ≤ 0,026**
  - **Muestra entre n > 500, desviación media absoluta < 0,02**

Islas Canarias
Del 15 al 19 de noviembre de 2021

Eskerrik asko!

¡Gracias!

Eustat
EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADÍSTICA