Islas Canarias
Del 15 al 19 de noviembre de 2021

Escuela Andaluza de Salud Pública
Consejería de Salud y Familias

UNIVERSIDAD DE GRANADA

Estadística y cartografía

XXI Jornadas de Estadística de las Comunidades Autónomas

Uso de machine learning para ajustar pesos en encuestas panel. Aplicación a la Encuesta Sanitaria y Social.

Luis Castro Martín
María del Mar Rueda García
Andrés Cabrera León
Carmen Sánchez-Cantalejo Garrido
Ramón Ferri García
Jorge Hidalgo Calderón
Iria Enrique Regueira

https://orcid.org/0000-0002-0934-4219

Introducción

- La rápida expansión de la COVID-19 ha hecho necesaria una evaluación rápida y eficiente de la situación.
- La Encuesta Sanitaria y SOCial (ESSOC) (Sánchez-Cantalejo et al., 2021) es un proyecto de investigación que proporciona información a lo largo del tiempo sobre el impacto del COVID-19 que pueda ser de utilidad para tomar medidas.
- La ESSOC tiene un diseño de encuesta panel por superposición para garantizar estimaciones tanto transversales como longitudinales.

Introducción

- Las encuestas panel son ampliamente usadas para observar la evolución de ciertas características a lo largo del tiempo.
- Sin embargo, implican problemas de no-respuesta (Arcos et al., 2020) causados por la fatiga de la población al ser encuestada reiteradamente.
- Presentamos un nuevo método de reponderación que produce estimadores válidos a partir de muestras afectadas por falta de respuesta, usando información auxiliar disponible a distintos niveles.

Contexto

- U una población finita de tamaño N, $U = \{1, \dots, i, \dots, N\}$.
- y la variable de interés.
- $s_{(1)}$ una muestra de tamaño $n_{(1)}$ extraída en la 1ª medición $(t = 1)$ mediante muestreo aleatorio estratificado.
- h el estrato al que pertenece un individuo i $(h = 1, \dots, L)$ y $s_{(1)h}$ la muestra correspondiente al estrato h en $t = 1$.
- $s_{(1)rh} = \{i \in s_{(1)}/\text{responde en el estrato h}\}$
- $s_{(1)fh} = \{i \in s_{(1)}/\text{faltante en el estrato h}\}$

Contexto

- $m_{(1)h}$ el tamaño de $s_{(1)rh}$. $\sum_h m_{(1)h}$ es el tamaño de $s_{(1)r}$.
- $s_{(t)rh}$ la muestra efectiva para $t = 2, 3, \ldots, k$ en el estrato *h* de $s_{(1)}$, de tamaño $m_{(t)h}$.
- $s_{(t)new}$ una nueva muestra extraída mediante un muestreo aleatorio estratificado independiente. $n_{(t)hnew}$ su tamaño en el estrato *h*. $s_{(t)rhnew}$ su muestra efectiva, de tamaño $m_{(t)hnew}$.

Estimación del total en la 1ª medición

Dados $y_{ih}^{(t)}$ el valor de la variable de interés para el individuo i y d_{ih} su peso de diseño, el estimador Horvitz-Thompson viene dado como:

$$\hat{Y}_{\mathrm{HT}}^{(1)} = \sum_{i \in s_{(1)r}} d_{ih}\, y_{ih}^{(1)}$$

Estimación del total en la 1ª medición

Considerando la tasa de respuesta en cada estrato h, $r_h = \frac{m_{(1)h}}{n_{(1)h}}$, los pesos iniciales son reemplazados por $d_{ih}^{(1)} = \frac{d_{ih}}{r_h}$. Por tanto, el estimador viene dado como:

$$\hat{Y}^{(1)} = \sum_h \sum_{i \in s_{(1)rh}} \frac{N_h}{n_{(1)h}} \frac{n_{(1)h}}{m_{(1)h}} y_{ih}^{(1)} = \sum_h \sum_{i \in s_{(1)rh}} d_{ih}^{(1)} y_{ih}^{(1)}$$

Estimación del total en la medición *t*

En el caso de las siguientes mediciones, los pesos de la muestra $s_{(t)new}$ pueden ser ajustados siguiendo la idea anterior:

$$\hat{Y}_n^{(t)} = \sum_h \sum_{s_{(t)rhnew}} \frac{N_h}{n_{(t)hnew}} \frac{n_{(t)hnew}}{m_{(t)hnew}} y_{ih}^{(t)} = \sum_h \sum_{s_{(t)rhnew}} d_{ihn}^{(t)} y_{ih}^{(t)}$$

Estimación del total en la medición *t*

- De igual manera se ajustan los pesos de $s_{(t)r}$:

$$\hat{Y}^{(t)} = \sum_h \sum_{i \in s_{(t)rh}} \frac{N_h}{n_{(t)h}} \frac{n_{(t)h}}{m_{(t)h}} y_{ih}^{(t)} = \sum_h \sum_{i \in s_{(t)rh}} d_{ih}^{(t)} y_{ih}^{(t)}$$

- Ambos estimadores se combinan como:

$$\hat{Y}_{\alpha}^{(t)} = \alpha_1 \hat{Y}^{(t)} + \alpha_2 \hat{Y}_n^{(t)}$$

Estimación del total en la medición *t*

- Los valores óptimos para los coeficientes α_1 y α_2 serían:

$$\alpha_1 = 1 - \alpha_2 = \frac{V\left(\hat{Y}_n^{(t)}\right)}{V\left(\hat{Y}^{(t)}\right) + V\left(\hat{Y}_n^{(t)}\right)}$$

- Como desconocemos los valores de $\mathrm{V}\left(\hat{Y}^{(t)}\right)$ y $\mathrm{V}\left(\hat{Y}_n^{(t)}\right)$, una solución sencilla consiste en considerar los tamaños de cada muestra:

$$\alpha_1 = 1 - \alpha_2 = \frac{m_{(t)h}}{m_{(t)h} + m_{(t)hnew}}$$

Estimación del total en la medición *t*

El estimador resultante es:

$$\hat{Y}_{c1}^{(t)} = \sum_h \frac{N_h}{m_{(t)h} + m_{(t)hnew}} \left(\sum_{i \in s_{(t)rh}} y_{ih}^{(t)} + \sum_{i \in s_{(t)rhnew}} y_{ih}^{(t)} \right)$$

$$= \sum_h \sum_{s_{(t)rh} \cup s_{(t)rhnew}} d_{ihc}^{(t)} \, y_{ih}^{(t)}$$

Estimación del total en la medición *t*

- También se puede aplicar una reponderación por calibración para tener en cuenta información auxiliar. Incluir este tipo de información aumenta la consistencia y puede mejorar la precisión (Deville y Särndal, 1992).

- Sea $x^{*(t)}$ un conjunto de variables auxiliares relacionadas con *y* tales que sus totales poblaciones para cada estrato son conocidos, $X_h^{*(t)} = \sum_{\mathcal{U}_h} \boldsymbol{x}_{ih}^{*(t)}$.

Estimación del total en la medición *t*

Nuevos pesos w_{ih}^{t} minimizan la distancia *G* a los pesos $d_{ihc}^{(t)}$:

$$\min \sum_{s_{(t)rh} \in s_{(t)rhnew}} G\left(w_{ih}^{(t)}, d_{ihc}^{(t)}\right)$$

cumpliendo la condición de calibración

$$\sum_{s_{(t)rh} \cup s_{(t)rhnew}} w_{ih}^{(t)} x_{ih}^{*(t)} = \sum_{\mathcal{U}_h} x_{ih}^{*(t)}$$

Estimación del total en la medición *t*

Por tanto, el estimador calibrado se obtiene como:

$$\hat{Y}_{\mathrm{CAL}}^{(t)} = \sum_{h} \sum_{s_{(t)rh} \cup s_{(t)rhnew}} w_{ih}^{(t)} y_{ih}^{(t)}$$

Estimación del cambio absoluto

- Un parámetro de interés es el cambio absoluto de una medición a otra:

$$\theta_t = Y^{(t)} - Y^{(1)}$$

- Para estimarlo usamos muestras con solapamiento pero hay que considerar la no respuesta debido al desgaste de los encuestados.
- Para ello, usamos el popular método Propensity Score Adjustment (PSA)(Lee, 2006; Lee y Valliant, 2009).
- Consiste en modelizar la probabilidad $\pi_{(t)k}$ que un individuo de la muestra $s_{(1)r}$ tiene de volver a responder en la ocasión t.

Estimación del cambio absoluto

- Para cada individuo de $s_{(1)r}$ definimos $\delta_{(t)k} = 1$ si $k \in s_{(t)r}$ y $\delta_{(t)k} = 0$ si $k \in s_{(1)r} - s_{(t)r}$.
- Asumimos que el mecanismo de selección de respuesta es ignorable:

$$\pi_{(t)k} = P\left(\delta_{(t)k} = 1 \middle| y_k, x_k\right) = P(\delta_k = 1 | x_k); k \in s_{(1)r}$$

- También asumimos que el mecanismo sigue un modelo paramétrico:

$$P\left(\delta_{(t)k} = 1 \middle| y_k, x_k\right) = f_t(x_k)$$

Estimación del cambio absoluto

- $\pi_{(t)k}$ puede ser estimado mediante un método del estado del arte del Machine Learning como es XGBoost (Chen y Guestrin, 2016).
- XGBoost construye árboles de decisión ensamblados que optimizan una función objetivo usando Gradient Tree Boosting (Friedman, 2001).
- El modelo se entrena usando $s_{(1)r}$ e incluyendo todas las variables.
- Se minimiza la pérdida logística para $\delta_{(t)k}; k \in s_{(1)r}$:

$$l(\hat{\pi}_{(t)}) = \sum_{k \in s_{(1)r}} -\delta_{(t)k} \log(\hat{\pi}_{(t)k}) - (1 - \delta_{(t)k}) \log(1 - \hat{\pi}_{(t)k})$$

Estimación del cambio absoluto

Definimos el siguiente estimador de θ para la medición t:

$$\hat{\theta}_t = \sum_h \sum_{i \in s_{(t)rh}} d_{ih}^{(1)} \frac{1}{\hat{\pi}_{(t)ih}} \left(y_{ih}^{(t)} - y_{ih}^{(1)} \right) = \sum_h \sum_{i \in s_{(t)rh}} d_{ihPSA}^{(t)} \left(y_{ih}^{(t)} - y_{ih}^{(1)} \right)$$

Estimación del cambio absoluto

Estos pesos $d_{ihPSA}^{(t)}$ también pueden ser calibrados con totales poblacionales conocidos. Por tanto, en una siguiente fase, se aplica calibración para reponderarlos, obteniendo unos pesos $v_{ih}^{(t)}$. El estimador final viene dado por:

$$\hat{\theta}_{ct} = \sum_{h} \sum_{i \in s_{(t)rh}} v_{ih}^{(t)} \left(y_{ih}^{(t)} - y_{ih}^{(1)} \right)$$

Estimación del cambio relativo

- Otro parámetro de interés en las encuestas panel es el cambio relativo:

$$\theta_{rt} = \frac{Y^{(t)} - Y^{(1)}}{Y^{(1)}}$$

- El ratio de las estimaciones puede usarse como estimador:

$$\hat{\theta}_{rt} = \frac{\hat{\theta}_{ct}}{\hat{Y}^{(1)}_{\text{CAL}}}$$

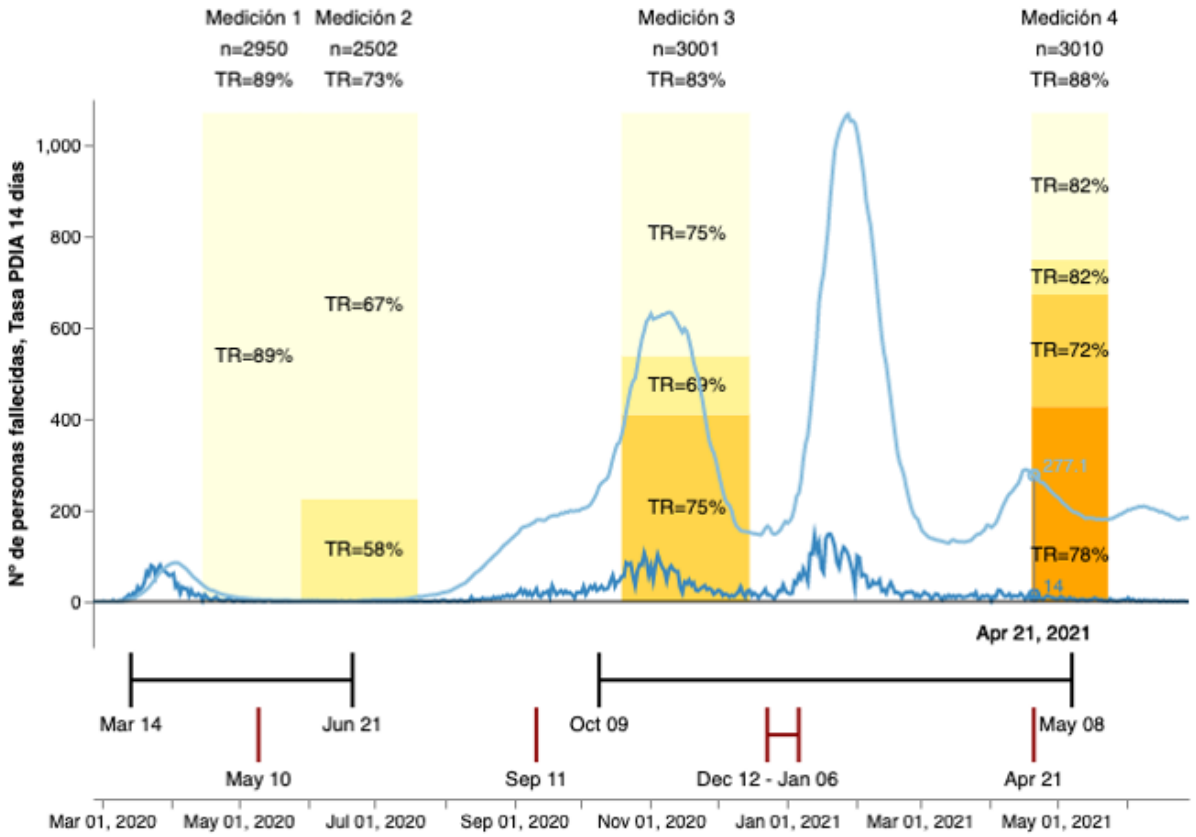
ESSOC

- Los métodos propuestos se aplican a la Encuesta Sanitaria y SOCial.
- Incluye 4 mediciones sobre la población andaluza mayor de 16 años.
- $s_{(1)}$, coincide con el comienzo del confinamiento en abril de 2020. $s_{(2)}$, $s_{(3)}$ y $s_{(4)}$ corresponden a 1, 6 y 12 meses después, respectivamente.
- La estratificación se hizo en función de la provincia y el nivel de urbanización a partir de la Base de Datos Longitudinal de Población de Andalucía (IECA, 2020).
- Los pesos son calibrados con los totales por provincia, sexo, edad, nacionalidad y tamaño total de la población.

Tamaños muestrales y no-respuesta

Encuesta Sanitaria y Social: trabajo de campo y evolución de la pandemia por SARS-COV-2 en Andalucía

n: Muestra efectiva, TR: Tasa de respuesta, PDIA: Pruebas Diagnósticas de Infección Activa
Fuente: IECA (Instituto de Estadística y Cartografía de Andalucía)



Evolución de:
- PDIA_14días
- Personas fallecidas

Muestras longitudinales:
- Muestra longitudinal 1
- Muestra longitudinal 2
- Muestra longitudinal 3
- Muestra longitudinal 4

Fechas relevantes:
- Mar 14: Inicio del estado de alarma
- May 10: Desescalada: movilidad con restricciones
- Jun 21: Fin del estado de alarma
- Sep 11: Restricciones de movilidad
- Oct 09: Inicio del estado de alarma
- Dic 23: Permiso de reagrupación familiar
- Jan 06: Fin de permiso
- Apr 29: Apertura movilidad entre provincias
- May 08: Fin del estado de alarma

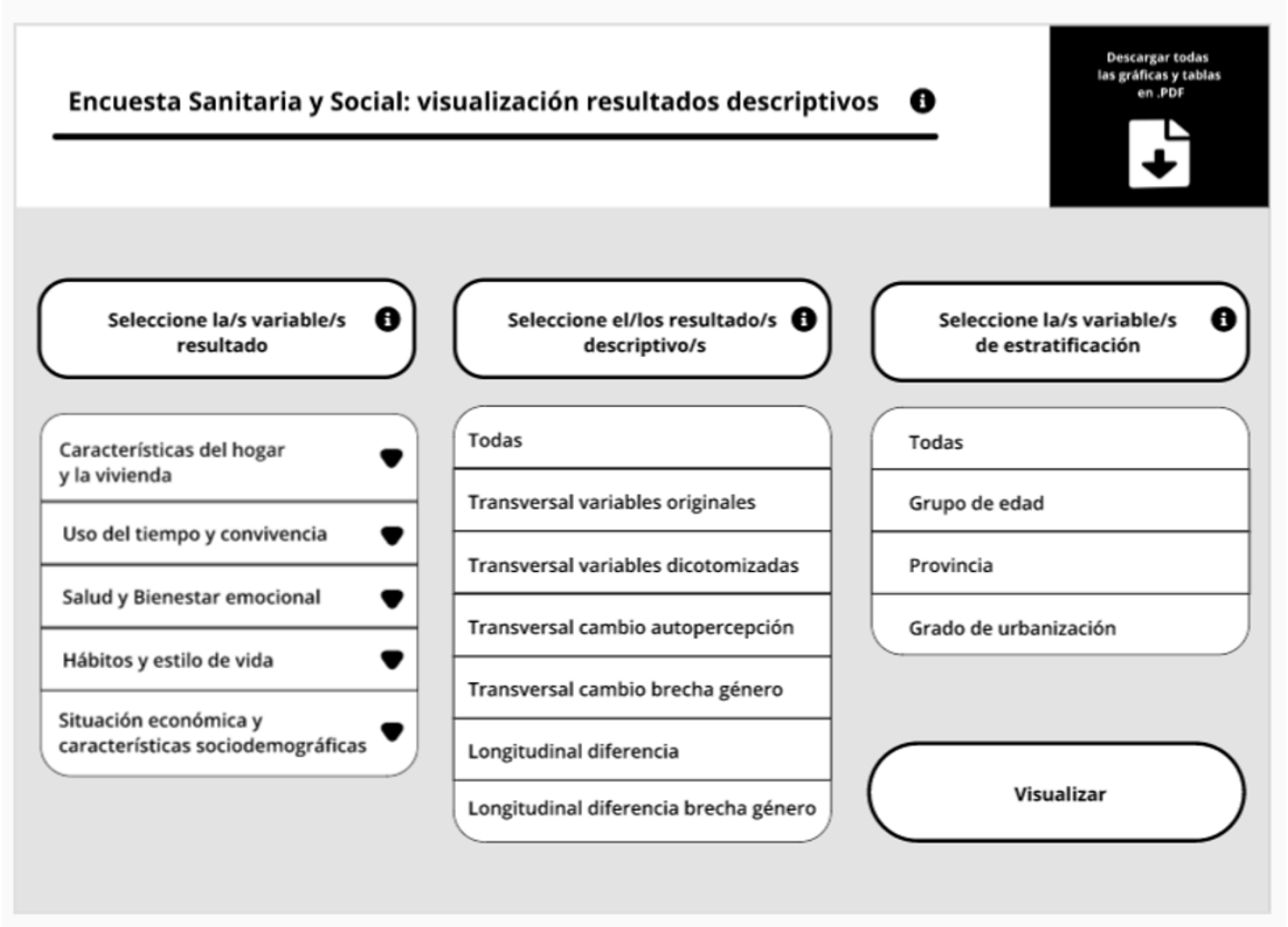
Optimización de hiperparámetros

- Número de estimadores $\in [10, 400]$: El número de árboles que forman el ensamblado.
- Tasa de aprendizaje $\in [0.01, 1]$: La reducción aplicada da los pesos tras cada iteración de boosting.
- Profundidad máxima $\in [1,60]$: El número máximo de particiones en cada árbol.
- Peso mínimo $\in [1,6]$: El mínimo del total de pesos que debe contener una rama para considerar una nueva partición.

Optimización de hiperparámetros

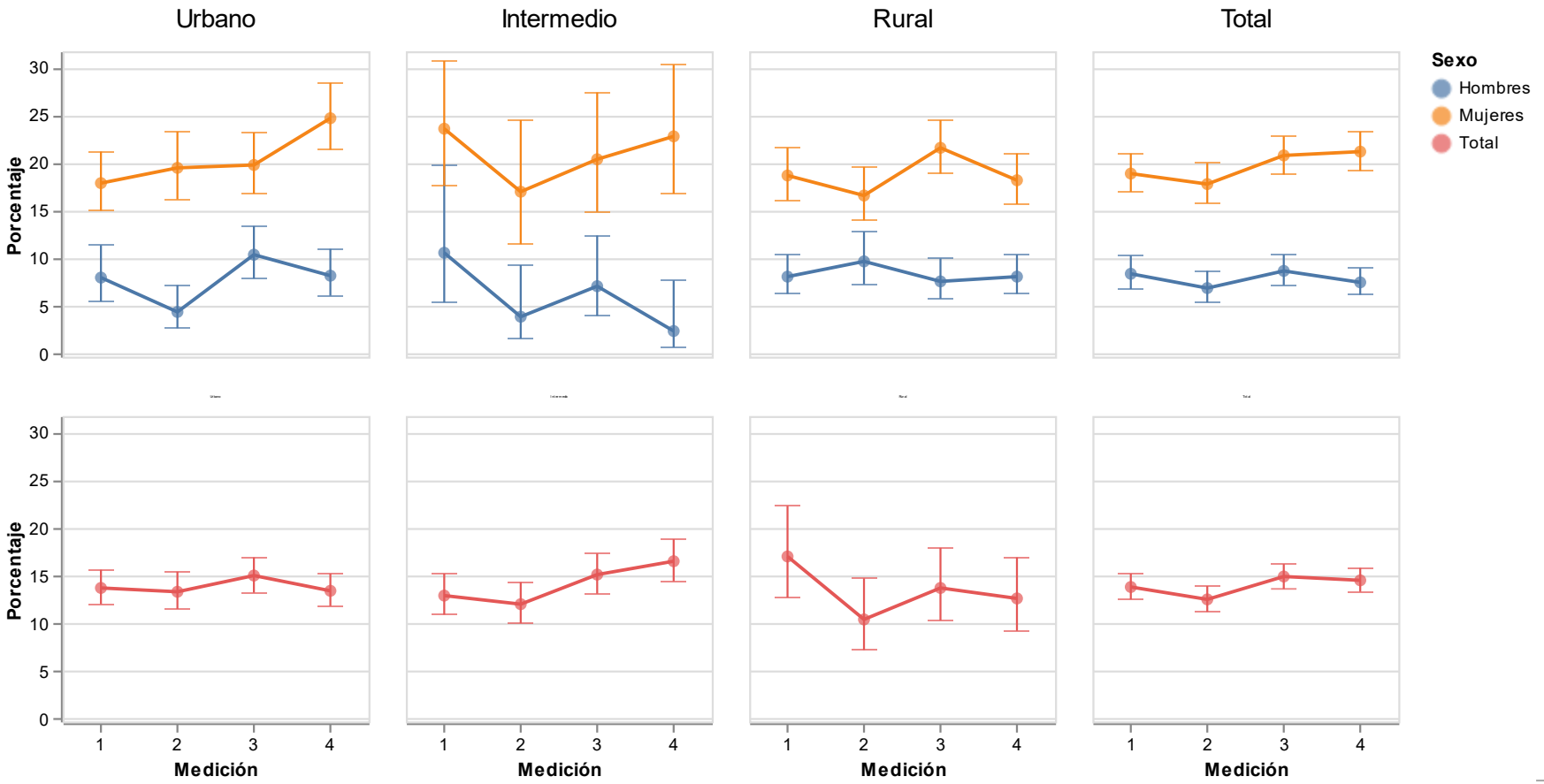
- La precisión del algoritmo se comprueba mediante validación cruzada.
- Los valores de los hiperparámetros que minimizan el error se obtienen mediante el algoritmo Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011).
- Este proceso evita que se produzca un sobreajuste (propensidades estimadas extremadamente cercanas a 0 o a 1).

Resultados


Encuesta Sanitaria y Social: visualización resultados descriptivos
Descargar todas las gráficas y tablas en .PDF
Seleccione la/s variable/s resultado
Seleccione el/los resultado/s descriptivo/s
Seleccione la/s variable/s de estratificación
Características del hogar y la vivienda
Uso del tiempo y convivencia
Salud y Bienestar emocional
Hábitos y estilo de vida
Situación económica y características sociodemográficas
Todas
Transversal variables originales
Transversal variables dicotomizadas
Transversal cambio autopercepción
Transversal cambio brecha género
Longitudinal diferencia
Longitudinal diferencia brecha género
Todas
Grupo de edad
Provincia
Grado de urbanización
Visualizar

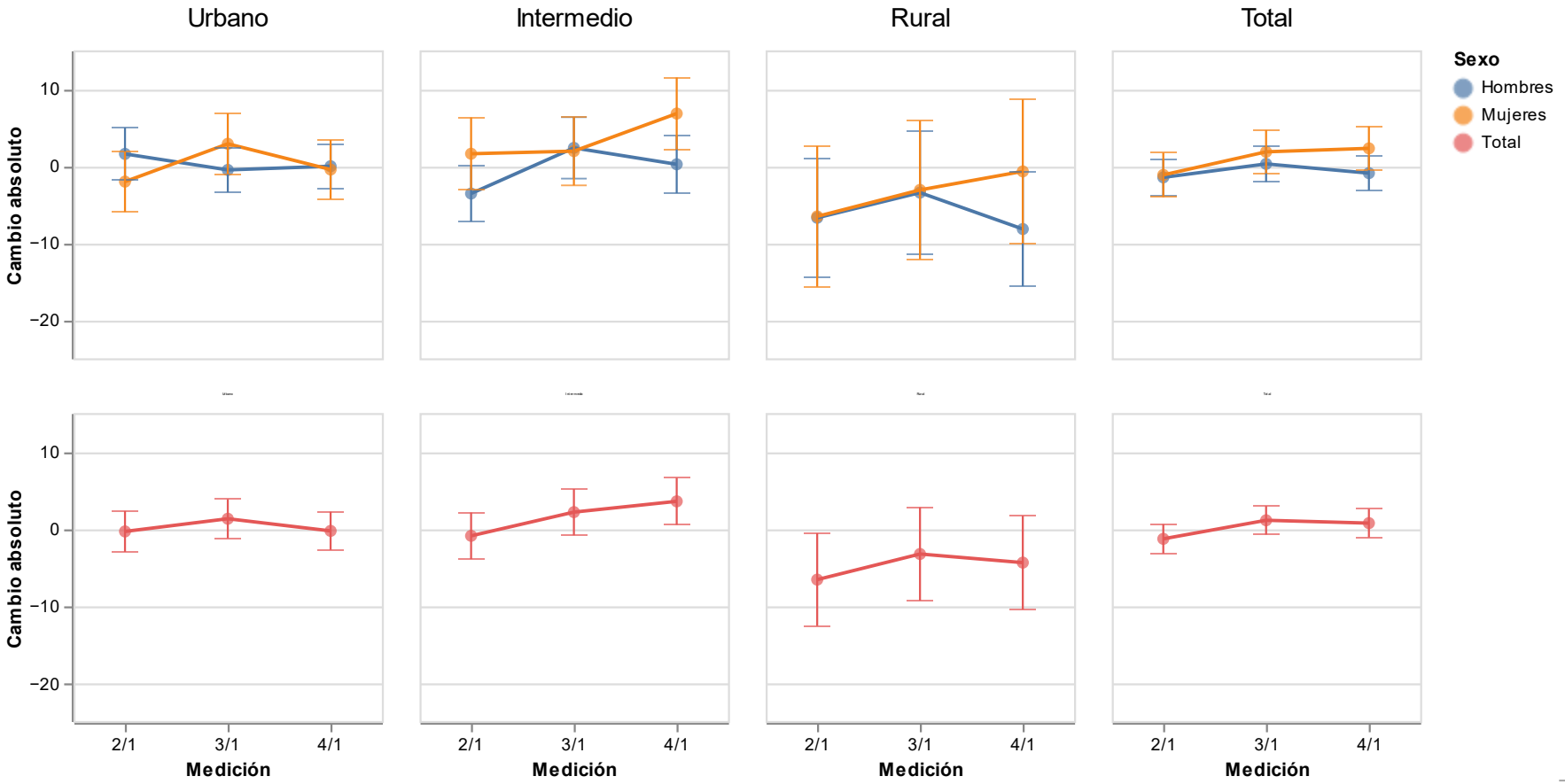
Resultados

Dolor de cabeza con bastante o mucha frecuencia según el grado de urbanización



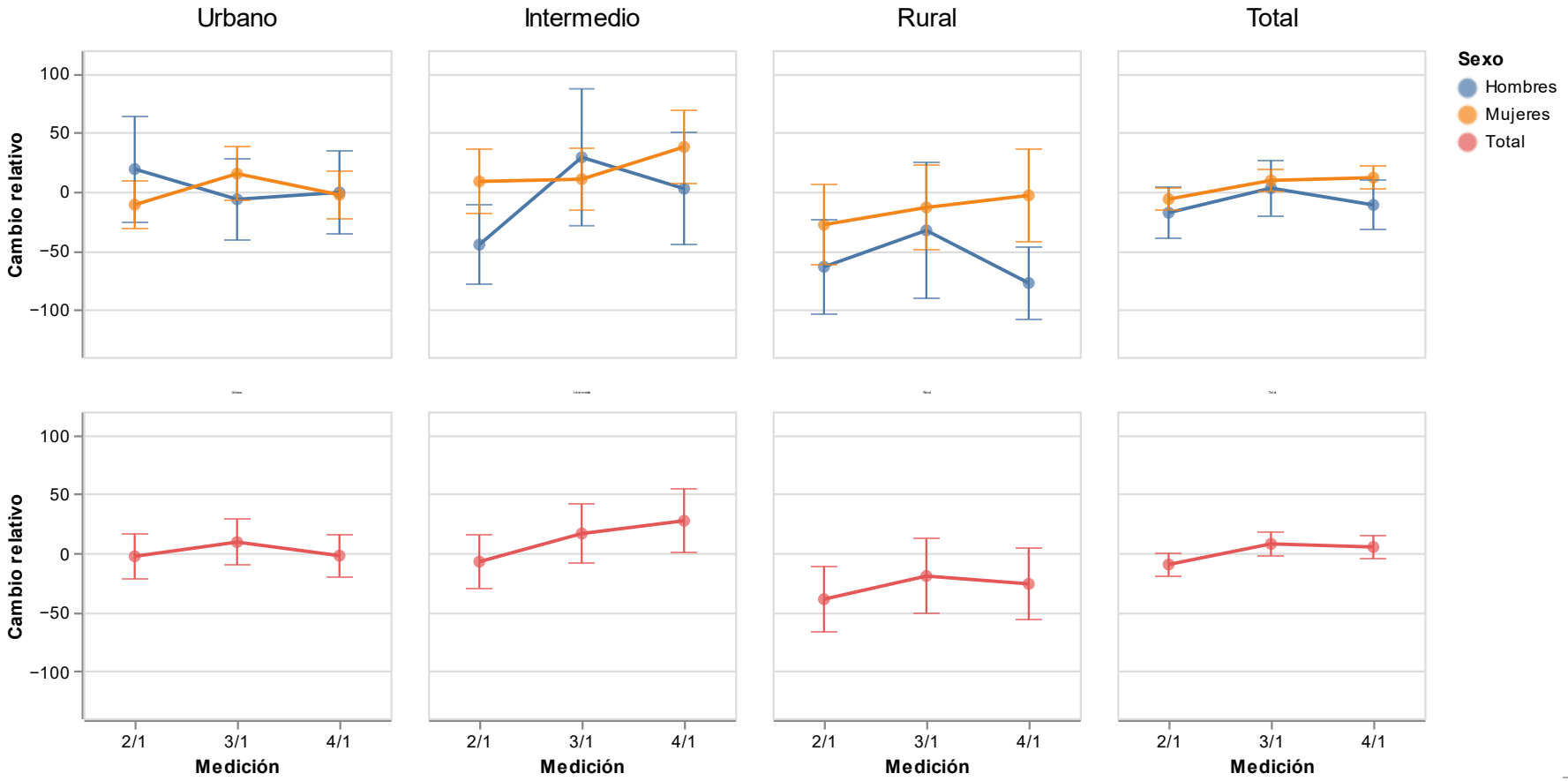
Resultados

Cambio absoluto en dolores de cabeza con bastante o mucha frecuencia según el grado de urbanización



Resultados

Cambio relativo en dolores de cabeza con bastante o mucha frecuencia según el grado de urbanización



Conclusiones

- Las encuestas panel con superposición son muy útiles cuando es de interés analizar la evolución de ciertas variables a lo largo del tiempo.
- Sin embargo, también implican problemas de no-respuesta que requieren estimadores específicos.
- Dichos estimadores pueden ser aplicados usando algoritmos del estado del arte del machine learning para obtener estimaciones óptimas.
- Estas técnicas nos han permitido obtener resultados más fiables en variables clave a la hora de tomar decisiones sobre la pandemia.

Referencias

- Arcos, A., Rueda, M. M., & Pasadas-del-Amo, S. (2020). Treating Nonresponse in Probability-Based Online Panels through Calibration: Empirical Evidence from a Survey of Political Decision-Making Procedures. Mathematics, 8(3), 423.
- Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011, December). Algorithms for hyper-parameter optimization. In 25th annual conference on neural information processing systems (NIPS 2011) (Vol. 24). Neural Information Processing Systems Foundation.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
- Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418), 376-382.

Referencias

- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- IECA (2020). Base de Datos Longitudinal de Población de Andalucía.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. Journal of official statistics, 22(2), 329.
- Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociological Methods & Research, 37(3), 319-343.
- Sánchez-Cantalejo C, Rueda MdM, Saez M, Enrique I, Ferri R, Fuente MdL, Villegas R, Castro L, Barceló MA, Daponte-Codina A, Lorusso N, Cabrera-León A. Impact of COVID-19 on the Health of the General and More Vulnerable Population and Its Determinants: Health Care and Social Survey–ESSOC, Study Protocol. International Journal of Environmental Research and Public Health. 2021; 18(15):8120. https://doi.org/10.3390/ijerph18158120

Islas Canarias
Del 15 al 19 de noviembre de 2021

¡Gracias!

Escuela Andaluza de Salud Pública
Consejería de Salud y Familias

UNIVERSIDAD DE GRANADA

Estadística y cartografía

https://orcid.org/0000-0002-0934-4219