Islas Canarias
Del 15 al 19 de noviembre de 2021

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

Luis Castro Martín
Universidad de Granada
luiscastro193@ugr.es

María del Mar Rueda García
Universidad de Granada
mrueda@ugr.es

Andrés Cabrera León
Escuela Andaluza de Salud Pública
andres.cabrera.easp@juntadeandalucia.es

Carmen Sánchez-Cantalejo Garrido
Escuela Andaluza de Salud Pública
carmen.sanchezcantalejo.easp@juntadeandalucia.es

Ramón Ferri García
Universidad de Granada
rferri@ugr.es

Jorge Hidalgo Calderón
Universidad de Granada
jorgehcal@ugr.es

Iria Enrique Regueira
Instituto de Estadística y Cartografía de Andalucía
iria.enrique@juntadeandalucia.es

Introducción

La rápida expansión de la COVID-19 ha hecho necesaria una evaluación rápida y eficiente de la situación. Los servicios sanitarios de todo el mundo han recurrido a encuestas para suplir esta necesidad de información sobre el impacto social, económico y sanitario de la enfermedad. Estas encuestas han facilitado la toma de decisiones en cuanto a medidas de prevención y protección de la población, especialmente a aquella con mayor riesgo de vulnerabilidad.

La Encuesta Sanitaria y SOCial (ESSOC) (Sánchez-Cantalejo et al., 2021) es un proyecto de investigación que surge ante esta necesidad. Su objetivo es proporcionar información específica y fiable sobre el impacto del COVID-19 a lo largo del tiempo en variables que puedan ser de utilidad a la hora de tomar medidas para controlar las consecuencias de la pandemia. Abarca la evolución de características del estado de salud, socioeconómicas, psicosociales, conductuales, laborales, medioambientales y clínicas; tanto de la población general como de colectivos vulnerables. El estudio integra datos de distintas fuentes basadas en encuestas y registros clínicos, epidemiológicos, poblacionales y ambientales. En el caso de las encuestas, éstas se han realizado mediante un diseño de encuesta panel por superposición.

Este tipo de diseños son ampliamente usados en estudios para observar la evolución de ciertas características a lo largo del tiempo. Sin embargo, implican problemas de no-respuesta causados, entre otros factores, por la fatiga de la población al ser encuestada reiteradamente. Este hecho hace necesario llevar a cabo nuevas encuestas hasta completar las muestras transversales de cada medición, reemplazando así a las personas que dejan de contestar de unas mediciones a otras y garantizando de esta manera obtener estimaciones igual de precisas en cada medición. Realizar estimaciones a partir de datos obtenidos con esta estructura no es tarea fácil, ya que implican importantes sesgos producidos por la falta de respuesta, así como de cobertura que afectan a la representatividad de la muestra. Si bien suelen emplearse métodos de reponderación para compensar dicho sesgo, estos métodos han sido propuestos generalmente para el caso de encuestas transversales y hay pocos estudios que proporcionen una metodología formal para su tratamiento en este tipo de paneles (Arcos et al., 2020).

Objetivos

Desarrollar métodos de ponderación para estimar totales, proporciones y diferencias de variables poblacionales a partir de datos obtenidos mediante encuestas panel combinando varias técnicas (Propensity Score Adjustment, machine learning y calibración) que hacen uso de información auxiliar a varios niveles.

Aplicar los métodos expuestos para obtener resultados sobre la ESSOC.

Contexto

Sea U una población finita de tamaño N, $U = \{1, \dots, i, \dots, N\}$. Queremos estimar un parámetro poblacional de una variable de interés y.

En la primera medición $(t = 1)$, una muestra $s_{(1)}$ de tamaño $n_{(1)}$ es extraída de la población mediante muestreo aleatorio estratificado. Siendo h el estrato al que pertenece un individuo i $(h = 1, \dots, L)$, notamos $s_{(1)h}$ a la muestra correspondiente al estrato h en la medición 1, de tamaño $n_{(t)h}$ (en adelante, muestra teórica).

La falta de respuesta en la muestra $s_{(1)}$ la divide en:

$$s_{(1)rh} = \{i \in s_{(1)}/\text{responde en el estrato h}\}$$

$$s_{(1)fh} = \{i \in s_{(1)}/\text{faltante en el estrato h}\}$$

Sea $m_{(1)h}$ el número de respuestas de $s_{(1)rh}$. Es decir, $\sum_h m_{(1)h}$ es el tamaño de $s_{(1)r}$.

Para cada una de las siguientes mediciones ($t = 2, 3, \dots, k$), notamos $s_{(t)rh}$ a la muestra de individuos del estrato *h* que han respondido en la medición *t* perteneciendo a la muestra original $s_{(1)}$, de tamaño $m_{(t)h}$ (en adelante, muestra efectiva). Para compensar por la falta de respuesta, una nueva muestra $s_{(t)new}$ es extraída de la población *U* mediante un muestreo aleatorio estratificado independiente de la muestra $s_{(1)}$. Asumiremos que las muestras $s_{(t)new}$ y la muestra $s_{(1)}$ tienen una intersección vacía. Finalmente, notamos $n_{(t)hnew}$ al tamaño de $s_{(t)new}$ en el estrato h y $s_{(t)rhnew}$ a su muestra efectiva en dicho estrato, de tamaño $m_{(t)hnew}$.

A continuación se describe la metodología utilizada para obtener estimadores del total para cada medición (muestras transversales), así como del cambio absoluto y relativo de una medición con respecto a otra, considerando que parte de las muestras son con solapamiento (muestras longitudinales).

Metodología: Estimación del total

Dados $y_{ih}^{(t)}$ el valor de la variable de interés para el individuo *i* y d_{ih} su peso de diseño, definido como la inversa de su probabilidad de inclusión, el estimador Horvitz-Thompson para el total *Y* en la primera medición viene dado como:

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

En el caso de un muestreo aleatorio estratificado simple, $d_{ih} = \frac{N_h}{n_{(1)h}}$. Sin embargo, estos pesos de diseño deben ser ajustados para considerar la no respuesta de algunos individuos. Este ajuste reducirá el sesgo producido cuando hay diferentes propensidades de respuesta para diferentes grupos.

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

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

En el caso de las siguientes mediciones, habrá disponibles dos muestras (la original y la nueva) de las que obtener estimaciones. Siguiendo la idea anterior, los pesos de la muestra $s_{(t)new}$ pueden ser ajustados obteniendo el estimador:

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

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

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

Ambos estimadores pueden ser combinados considerando

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

donde α_1 y α_2 son constantes no negativas tales que $\alpha_1 + \alpha_2 = 1$.

Los valores óptimos para los coeficientes α_1 y α_2 serían:

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

Sin embargo, desconocemos los valores de $\mathrm{V}\left(\hat{Y}^{(t)}\right)$ y $\mathrm{V}\left(\hat{Y}_n^{(t)}\right)$. Podrían sustituirse por una estimación obtenida a partir de las muestras pero eso no garantiza que sea óptimo. Una solución sencilla consiste en considerar los tamaños de cada muestra:

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

Por tanto, el estimador resultante sería:

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

También se puede aplicar una reponderación por calibración para tener en cuenta información auxiliar. Incluir este tipo de información aumenta la consistencia entre las estimaciones de diferentes muestras y puede mejorar la precisión de los estimadores (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, $X_h^{*(t)} = \sum_{\mathcal{U}_h} \boldsymbol{x}_{ih}^{*(t)}$.

Nuevos pesos w_{ih}^t son calculados de manera que sean lo más cercanos posibles, dada una 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

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

Por tanto, el estimador calibrado se obtiene como:

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

Metodología: Estimación del cambio absoluto

Este tipo de encuestas realizan varias mediciones con el objetivo principal de estimar los cambios a lo largo del tiempo. Un parámetro de interés sería entonces el cambio absoluto de una medición a otra para nuestra variable de interés, al que notaremos:

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

Tener muestras con solapamiento nos permite estimar estas variaciones con más precisión. Sin embargo, la no respuesta debido al desgaste de los encuestados introduce un sesgo que debe ser corregido reponderando los pesos.

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. Para cada individuo de $s_{(1)r}$ definimos la variable indicadora $\delta_{(t)k} = 1$ 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, es decir, que solo depende de un conjunto de variables auxiliares x_k:

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

También asumimos que el mecanismo sigue un modelo paramétrico:

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

Bajo esos supuestos, $\pi_{(t)k}$ puede ser estimado usando 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 mediante el algoritmo Gradient Tree Boosting (Friedman, 2001). El modelo se entrena usando toda la muestra $s_{(1)r}$ e incluyendo todas las variables disponibles como x_k. La función objetivo a minimizar debe ser definida como la pérdida logística para $\delta_{(t)k}; k \in s_{(1)r}$:

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

Un factor a tener en cuenta es que los valores que nos interesan, $\hat{\pi}_{(t)k}$ para $k \in s_{(1)r} \cap s_{(t)r}$, vienen de un subconjunto de los valores usandos para entrenar el modelo, $\delta_{(t)k}$ para $k \in s_{(1)r}$. Esto implica que es probable que se produzca un sobreajuste y que por tanto se obtengan valores muy cercanos a 1 en lugar de las verdaderas propensidades. Aplicar una optimización de hiperparámetros será esencial para evitar este problema.

Las propensidades estimadas $\hat{\pi}_{(t)ih}$ para cada individuo *i* del estrato *h* de la muestra $s_{(t)rh}$ son usadas para reponderar los pesos, obteniendo el siguiente estimador de θ para la medición *t*:

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

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

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

Metodología: Estimación del cambio relativo

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

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

Dado que no es más que el ratio de dos valores que ya hemos estimado, puede usarse el ratio de las estimaciones como estimador:

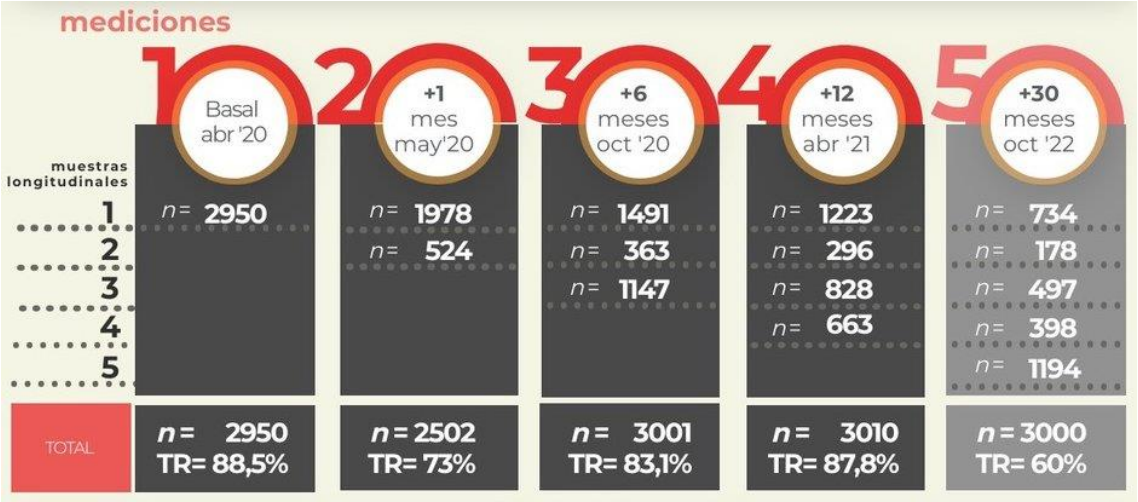
$$\hat{\theta}_{rt} = \frac{\hat{\theta}_{ct}}{\hat{Y}_{\text{CAL}}^{(1)}}$$

Resultados: Aplicación a la ESSOC

Hemos aplicado los métodos propuestos a la Encuesta Sanitaria y SOCial (ESSOC) descrita en la introducción. Los datos incluyen 4 mediciones sobre la población andaluza mayor de 16 años. La primera, $s_{(1)}$, coincide con el comienzo del confinamiento en abril de 2020. Las tres siguientes, $s_{(2)}$, $s_{(3)}$ y $s_{(4)}$ fueron llevadas a cabo 1, 6 y 12 meses después, respectivamente.

La estratificación se hizo en función de la provincia y el nivel de urbanización. Para la calibración usamos los totales poblacionales de las siguientes variables: provincia, sexo, edad, nacionalidad y tamaño total de la población; obtenidos de la *Encuesta Continua de Hogares* (INE, 2019) para la primera medición ($t = 1$) y el Padrón Municipal de Habitantes (IECA, 2020) para el resto de mediciones ($t = 2, 3, 4$).

El tamaño teórico de las muestras es de 5000 individuos. Los detalles de no-respuesta, así como del tamaño efectivo de las muestras correspondientes a cada medición pueden consultarse en el siguiente gráfico:

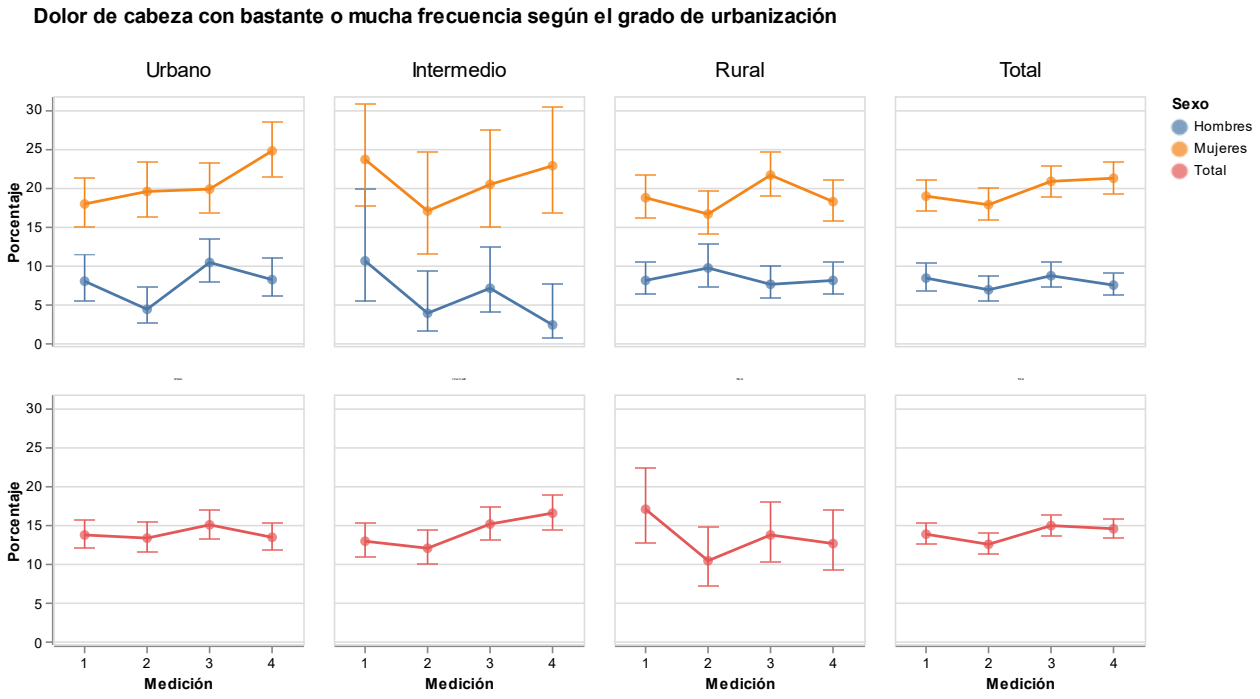


A la hora de aplicar el modelo de XGBoost que modela la no-respuesta $\pi_{(t)k}$ para $k \in s_{(t)r}$ con $t > 1$, hemos considerado los siguientes hiperparámetros a optimizar:

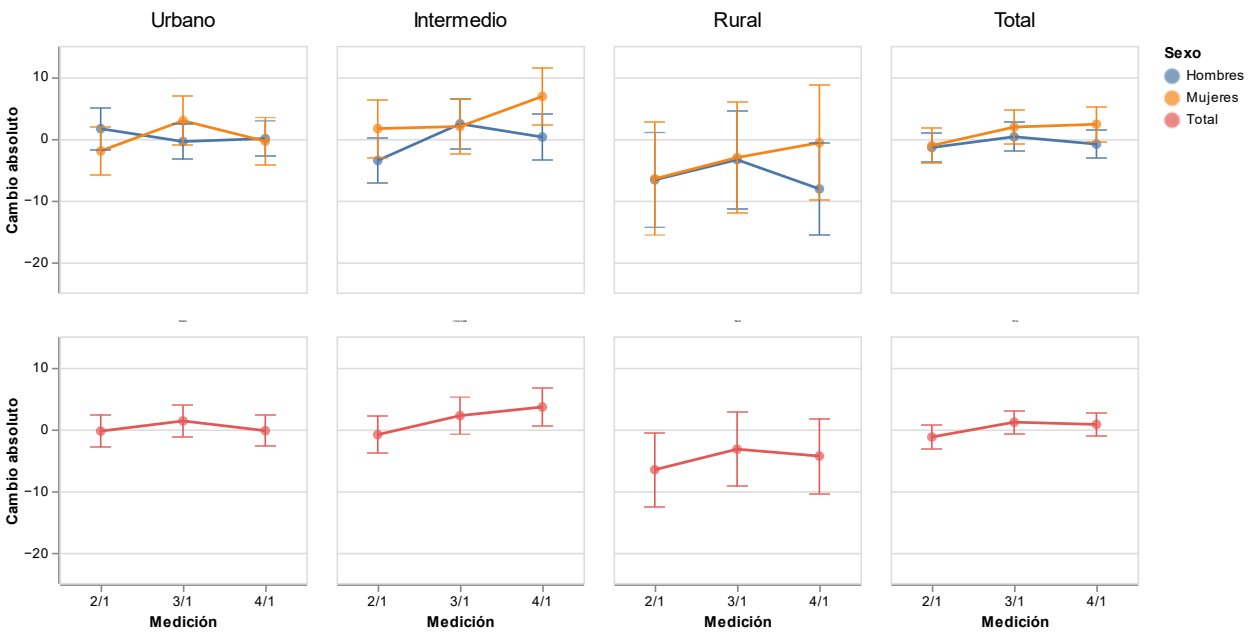
- Número de estimadores $\in [10, 400]$: El número de árboles que forman el ensablado.
- Tasa de aprendizaje $\in [0.01, 1]$: La reducción aplicada da los pesos tras cada iteración de boosting.
- Profundidad máxima $\in [1,60]$: El número máximo de particiones en cada árbol.
- Peso mínimo $\in [1,6]$: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

La precisión del algoritmo se comprueba mediante validación cruzada. Usando esa métrica, aplicamos el algoritmo Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011) para obtener los valores de los hiperparámetros que minimizan el error.

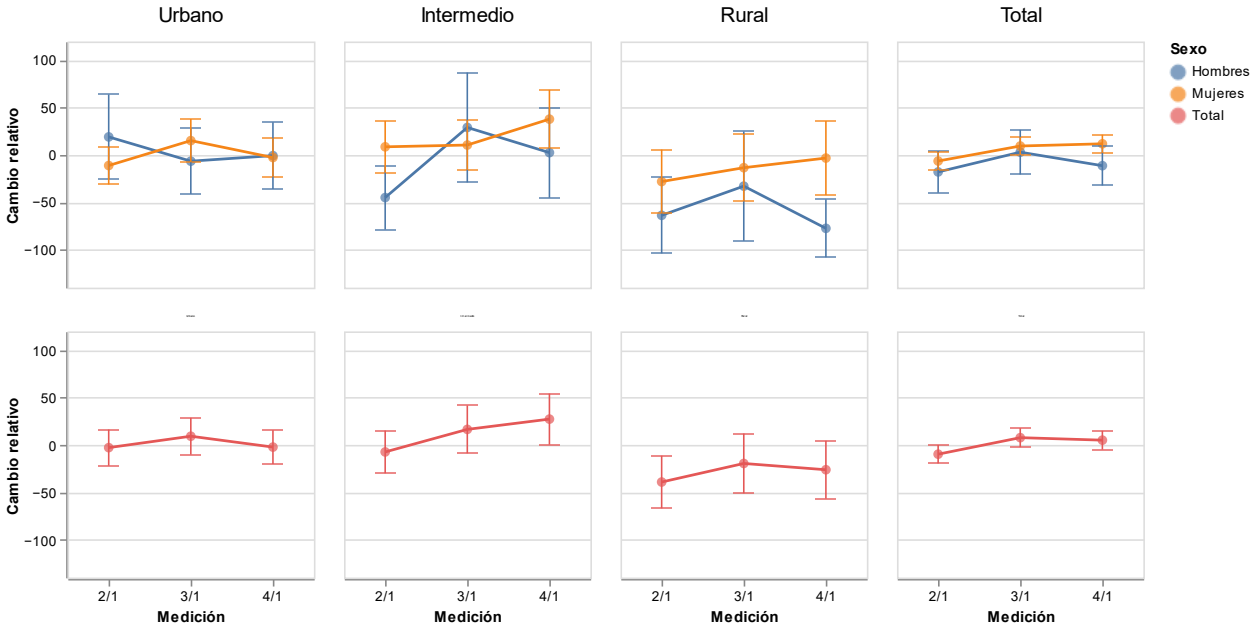
A continuación se muestran las gráficas resultantes tras aplicar todos los estimadores propuestos a la variable “Dolor de cabeza con bastante o mucha frecuencia”:



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



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



Conclusión

Las encuestas panel con superposición son muy útiles cuando es de interés analizar la evolución de ciertas variables a lo largo del tiempo y, al mismo tiempo, ofrecer estimaciones igual de precisas en cada momento o medición. Sin embargo, este tipo de encuestas también implican problemas de no-respuesta que introducen sesgo en las estimaciones obtenidas. En esta ponencia proponemos una serie de estimadores específicos para reducir dicho sesgo y poder obtener conclusiones más fiables.

También utilizamos un caso práctico de aplicación de los estimadores propuestos a la ESSOC, obteniendo información sobre el impacto del COVID-19 en la población a lo largo del tiempo y en momentos específicos de la pandemia. El uso de algoritmos y

técnicas del estado del arte del Machine Learning nos han permitido obtener resultados óptimos en variables clave a la hora de tomar decisiones sobre la pandemia.

Financiación

Se ha obtenido financiación de las convocatorias competitivas del Fondo SUPERA COVID-19 de Santander Universidades (SAUN), la Conferencia de Rectores de Universidades Españolas (CRUE), y el Consejo Superior de Investigaciones Científicas (CSIC), además del Programa de Ayudas Competitivas COVID-19 de Pfizer Global Medical Grants. Este trabajo también está apoyado en parte por la beca IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 y por el Ministerio de Ciencia e Innovación, España [número de beca PID2019-106861RB-I00/AEI/10.13039/501100011033].

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.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
- IECA (2020). Padrón Municipal de Habitantes.
- INE (2019). Encuesta continua de hogares.
- 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, M. M., Saez, M., Enrique, I., Ferri-García, R., De la Fuente, M., Castro-Martín, L., ... & Cabrera-León, A. (2021). 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.