Proyecto de tesis de maestría

UNC Universidad Nacional de Córdoba

Identificación de potencial sesgo por no respuesta en encuestas de hogares del nordeste argentino sobre el ingreso familiar y el nivel de pobreza

---

Tesista	Lic. Celine Iliana Cabás
Directora	Dra. Patricia Caro
Co-Director	Dr. Carlos Matías Hisgen

Diciembre 2024

Maestría en Estadística Aplicada
Universidad Nacional de Córdoba

Contenidos

Introducción

Introducción

"Si la decisión de responder depende estadísticamente de las variables bajo investigación, entonces la submuestra de encuestados no reflejará con precisión la distribución real de las variables de interés en la población y esto, a su vez, dará como resultado inferencias basadas en muestras sistemáticamente sesgadas". [1]

- No respuesta en encuestas de hogares.
- Caso particular: Encuesta Permanente de Hogares (EPH).
- Producto estadístico utilizado para medir distribuciones de ingresos y nivel de pobreza por aglomerado urbano.
- Cobertura geográfica: Nordeste Argentino.

¿Qué ocurre en el NEA?

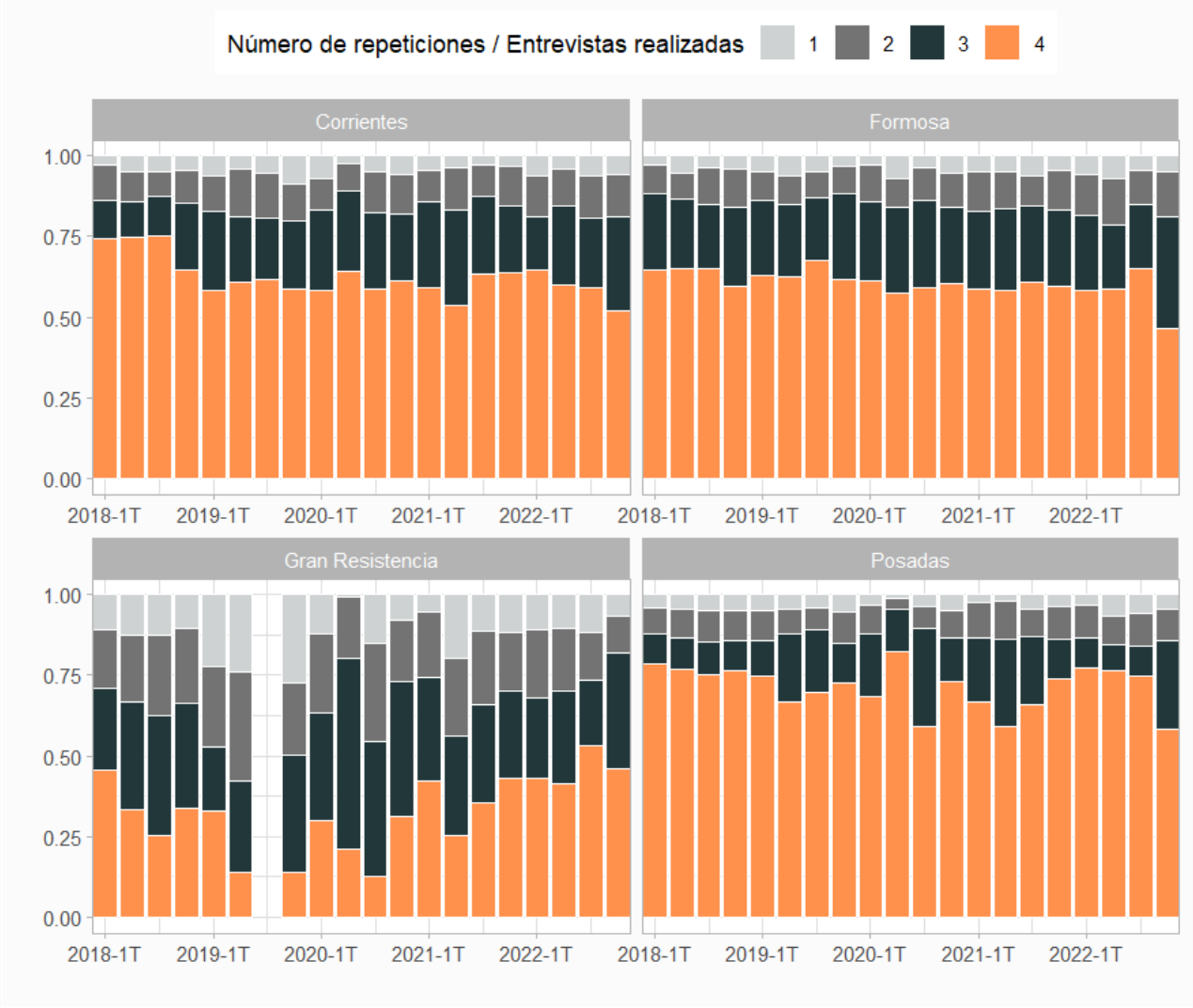

Figura 1: Estructura de respuesta por aglomerado urbano, 2018-2022.

Introducción

- ¿Qué ocurriría si la estructura de respuesta depende del ingreso?
- ¿De qué manera puede equilibrarse la representatividad de los hogares en la muestra? ¿Cómo mejorar la representatividad de quienes tienden a no responder?
- La selección del NEA viene justificada por el interés de estudiar si las diferencias entre el nivel de pobreza monetaria de Gran Resistencia y los demás aglomerados urbanos de la región puede verse justificada por diferencias observables en sus estructuras de no respuesta.

Antecedentes

Marco teórico I

→ **¿Qué entendemos por no respuesta?**

No respuesta al ítem o *no respuesta unitaria* [1]

→ **¿Cuándo existe sesgo por no respuesta?**

Cuando las personas que sí respondieron a la encuesta difieren significativamente de aquellas que no lo hicieron, es decir, la no respuesta *no se comporta de manera aleatoria* [2].

→ **¿Cómo detectar el sesgo por no respuesta?**

- Test de Little [3];
- Test Chi-cuadrado de contraste de independencia [4];
- Modelos logísticos de respuesta binaria o politómica [4]

Marco teórico II

→ **¿Cómo lidiar con el sesgo? ¿De qué manera corregirlo?**

- Técnicas de imputación para sustituir individuos no encuestados. Métodos de calibración basados en variables auxiliares que deben conocerse en la población [4] [5];
- **Propensity Score Adjustment (PSA)**. Modelos de estimación de la probabilidad de respuesta (propensión a responder):
  - Modelos logísticos de respuesta binaria;
  - Modelos multinomiales de respuesta politómica ordinal;
  - Modelos de aprendizaje automático (CART, RF, XGBoost) [6].
- Recalibrar los pesos de los hogares mejorando la representatividad en la muestra de aquellos con baja propensión a responder respecto a aquellos con alta propensión a responder.

Antecedentes empíricos

Algunos trabajos aplicados:

- Korinek [1] trabaja con la encuesta Current Population Survey (CPS) de U.S. Census Bureau para analizar la sensibilidad de la distribución acumulada del ingreso frente a ajustes en los factores de expansión basados en la no respuesta unitaria.
- *En Argentina*, el trabajo de Comari y Hoszowski [7] estudia el efecto de la no respuesta en la EPH (2005-2011) sobre la estimación de la *tasa de desempleo*. Encuentran que los hogares con mayor número de entrevistas realizadas tienen menores tasas de desempleo.

Actualizaciones metodológicas

En Argentina, el INDEC ha incorporado actualizaciones metodológicas referidas a:

- *Ajustes por probabilidad de respuesta* en el factor de expansión basados en las variables: nivel educativo, edad, cantidad de ocupados y desocupados, régimen de tenencia de la vivienda, entre otras [8].
- *Ajustes sobre el nivel de ingreso*. La no respuesta fue abordada desde el enfoque de "no respuesta al ítem" con correcciones mediante ajustes a los pesos de diseño asignando a los no respondentes el comportamiento de los respondentes por estratos [9].

Formulación del problema y objetivos

Formulación del problema

Pregunta de investigación

¿Qué efectos tiene la presencia de sesgo por no respuesta en encuestas de hogares del nordeste argentino sobre la estimación del ingreso familiar y el nivel de pobreza?

Objetivo general

Identificar la potencial presencia de sesgo por no respuesta en encuestas de hogares del nordeste argentino y sus efectos sobre el ingreso familiar y el nivel de pobreza durante el período 2018-2022.

Objetivos específicos

- Comparar las estructuras de respuesta de la encuesta de hogares entre los aglomerados urbanos del nordeste argentino para el período 2018-2022.
- Comprobar si la predisposición a responder por parte de los hogares depende significativamente del ingreso familiar u otras variables en los distintos aglomerados urbanos del nordeste argentino.
- Comparar modelos para predecir la probabilidad de los hogares de responder de manera completa el esquema de entrevistas de la encuesta.
- Plantear una corrección del sesgo por no respuesta basada en la probabilidad predicha de responder mediante técnicas de reponderación de los datos.
- Contrastar las distribuciones de ingresos y el nivel de pobreza estimado antes y después de la corrección por no respuesta.

Fuentes de información

Fuentes de información

Este proyecto se plantea mediante el uso de las siguientes **fuentes secundarias** de información relevadas por el Instituto Nacional de Estadísticas y Censos (INDEC) de la República Argentina:

- Encuesta Permanente de Hogares (EPH), individual y hogar.
- Índice de Precios al Consumidor (IPC) para deflactar ingresos.
- Valorización de la Canasta Básica Total (CBT) para la determinación de la condición de pobreza de los hogares.

Variables preliminares

Cuadro 1: Variables preliminares a utilizar en EPH individual y hogar.

Variable	Descripción
Identificación	
CODUSU	Código de identificación de la vivienda
NRO_HOGAR	Código de identificación del hogar
REGION	Código de región geográfica
AGLOMERADO	Código de aglomerado urbano
ANO4	Año de relevamiento
TRIMESTRE	Trimestre de relevamiento
Base individual	
CH03	Relación de parentesco (Jefe de hogar=1)
CH04	Sexo
CH06	Edad
NIVEL_ED	Nivel educativo
ESTADO	Condición de actividad
CAT_OCUP	Categoría ocupacional
Base hogar	
IV1	Tipo de vivienda
IX_TOT	Cantidad de miembros del hogar
ITF	Ingreso total familiar
IPCF	Ingreso per cápita familiar
Variable de respuesta	
NRO_REP	Número de entrevistas realizadas en el período 2018-2022
Ponderador	
PONDIH	Ponderador del ITF y del IPCF

¿Cómo medimos la propensión a responder?

Esquema de rotación trimestral
Idealmente, una misma vivienda debe ser encuestada dos trimestres consecutivos, descansar los dos trimestres subsiguientes y volver a ser encuestada dos trimestres consecutivos más para garantizar la estructura de panel de datos que caracteriza a la encuesta.

↓

No respuesta
No todos los hogares completan el esquema de la encuesta.

↓

Entrevistas realizadas (Categórica ordinal)
Número de entrevistas realizadas (del 1 al 4) como variable proxy para medir la menor o mayor tendencia a responder de los hogares.

Metodología

Metodología i

Primera etapa: Análisis descriptivo exploratorio

Análisis descriptivo de la **estructura de respuesta de la muestra** para los cuatro aglomerados urbanos que representan al nordeste argentino en la encuesta (Gran Resistencia, Corrientes, Formosa y Posadas).

Segunda etapa: Modelos para detectar sesgo por no respuesta

Ajuste de un **modelo lineal generalizado multinomial de respuesta politómica** ordinal que explique el número de entrevistas realizadas por hogar (del 1 al 4) para testear la aleatoriedad teórica que esta variable debería tener respecto al ingreso.

Metodología ii

Tercera etapa: Predecir la probabilidad de respuesta

Comparación de modelos alternativos para PSA de aprendizaje automático para la predicción de la propensión de los hogares a completar el esquema de la encuesta. Mejorar la representatividad de los casos con baja propensión a responder.

Cuarta etapa: Analizar distribuciones de ingresos

Comparar distribuciones acumuladas de ingresos pre y post ajuste del factor de expansión. Analizar la sensibilidad de las colas de la distribución.

Resultados esperados

Resultados esperados

- La decisión de responder más o menos veces depende sistemáticamente del ingreso del hogar. **La no respuesta no se comporta de manera aleatoria**.
- **Buen desempeño en un modelo predictivo** de la probabilidad de que el hogar complete el esquema de la encuesta.
- **Mejorar la representatividad de los casos con baja propensión a responder** mediante reponderación de los datos.
- Dado que la estructura de respuesta en Gran Resistencia se encuentra notablemente desbalanceada, se espera que la implementación del método de reponderación modifique en mayor medida la distribución acumulada del ingreso per cápita de los hogares de este aglomerado y en menor medida las distribuciones de Corrientes, Formosa y Posadas.

Referencias i

[1] A. Korinek, J. A. Mistiaen y M. Ravallion. **"An econometric method of correcting for unit nonresponse bias in surveys".** En: *Journal of Econometrics* 136 (2007), págs. 213-235.

[2] F. Butar Butar y C. Chang. **"Weighting Methods in Survey Sampling".** En: *Survey Research Methods* (2012), págs. 4768-4782.

[3] L. González Allendes. **"Propuesta de tratamiento de la no respuesta parcial para la medición de la Pobreza Multidimensional en Chile".** Tesis de mtría. Universidad de Chile, 2019.

[4] J. Bethlehem, F. Cobben y B. Schouten. **Handbook of nonresponse in household surveys.** Wiley, 2011, pág. 474.

[5] R. Ferri-García y M. Del Mar Rueda. **"Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys".** En: *SORT* 42 (2018), págs. 159-182.

Referencias ii

[6] R. Ferri-García et al. **"Estimating response propensities in nonprobability surveys using machine learning weighted models".** En: *Mathematics and Computers in Simulation* 225 (2024), págs. 779-793.

[7] C. Comari y A. E. Hoszowski. **"Non response in rotating panel surveys: analysis on Argentina's labor force survey".** En: *Joint Statistical Meetings*. 2014.

[8] Instituto Nacional de Estadística y Censos. **Encuesta Permanente de Hogares: Consideraciones metodológicas sobre el tratamiento de la información del segundo trimestre de 2020.** 2020.

[9] Instituto Nacional de Estadística y Censos. **Encuesta Permanente de Hogares: Diseño de registro y estructura para las bases preliminares Hogar y Personas.** 2024.

¡Muchas gracias!