ESTRUCTURAS LATENTES DE LA ECONOMÍA DEL CUIDADO EN COLOMBIA: UNA APROXIMACIÓN CON MODELOS DE MEZCLA, MODELOS ECONOMÉTRICOS Y APRENDIZAJE ESTADÍSTICO

LATENT STRUCTURES OF THE CARE ECONOMY IN COLOMBIA: AN APPROACH WITH MIXTURE MODELS, ECONOMETRIC MODELS AND STATISTICAL LEARNING

LAURA ESTEFANY PARDO VARGAS
lpardov@unal.edu.co
Universidad Nacional de Colombia
JOSE MIGUEL LEON PUENTES
joleonp@unal.edu.co
Universidad Nacional de Colombia

Resumen

La economía del cuidado se define como todas las actividades no remuneradas que se realizan en el hogar, relacionadas con el mantenimiento de la vivienda, los cuidados a otras personas del hogar o comunidad y el mantenimiento de la fuerza de trabajo remunerado [1], por lo que resulta fundamental en el sostenimiento del bienestar social e individual. Además, el valor económico derivado de estas actividades equivale al 20 % del PIB, que, de ser remuneradas, posicionarían al Trabajo de Cuidado No Remunerado (TCNR) como el sector más importante de la economía colombiana. Es por ello que, en aras de potenciar la productividad y aliviar la carga de los agentes que asumen el cuidado, se formulan modelos que permitan comprender en su complejidad la realidad del TCNR en Colombia, llevando a cabo un análisis exploratorio de segmentación multivariada del tiempo dedicado a este oficio, junto con variables económicas y sociodemográficas, por medio de técnicas de aprendizaje estadístico no supervisado. Posteriormente, se describe la distribución del tiempo empleado en estas actividades mediante modelos de mezcla finita y modelo GAM, con el fin de identificar grupos de individuos con patrones similares de cuidado.
El análisis se fundamenta en los datos recopilados por la Encuesta Nacional de Uso del Tiempo (ENUT) del Departamento Administrativo Nacional de Estadística - DANE Colombia [1] en su edición del 2020-2021. La exploración de los resultados se realiza mediante una segmentación que integra técnicas de agrupamiento no jerárquico (K-means). Este enfoque permite identificar tanto el número de grupos como el tiempo promedio dedicado al cuidado en cada uno de ellos. Asimismo, se realizan segmentaciones con enfoque territorial, social y económico, con el propósito de esclarecer las características de quienes asumen el cuidado en Colombia y aportar elementos que orienten el diseño de políticas públicas más precisas y equitativas.

Palabras claves : Economía del Cuidado, Trabajo no Remunerado, Modelos de Mezcla, Aprendizaje Estadístico, Microeconometría

Abstract

The care economy is defined as all unpaid activities carried out within the household, including domestic maintenance, care for other household or community members, and the upkeep of the paid labor force [1]. It plays a fundamental role in sustaining both social and individual well-being. Moreover, the economic value derived from these activities amounts to approximately 20% of Colombia's GDP; if monetized, Unpaid Care Work (UCW) would rank as the most significant sector of the Colombian economy. In this context, and with the aim of enhancing productivity and alleviating the burden borne by caregivers, we develop models that account for the complexity of UCW in Colombia. We conduct an exploratory multivariate segmentation analysis of the time devoted to care activities, alongside economic and sociodemographic variables, using unsupervised statistical learning techniques. Subsequently, the distribution of time allocated to these activities is modeled through finite mixture models and generalized additive models (GAM), in order to identify groups of individuals who share similar care patterns.

The analysis is based on data from the 2020–2021 edition of the Encuesta Nacional de Uso del Tiempo (ENUT) conducted by the Colombian National Administrative Department of Statistics (DANE) [1]. Results are explored through segmentation techniques that integrate non-hierarchical clustering (K-means), which allows for the identification of both the number of groups and the average time devoted to care within each cluster. In addition, territorial, social, and economic segmentations are carried out with the purpose of clarifying the characteristics of those who assume care responsibilities in Colombia and providing insights that may inform the design of more precise and equitable public policies.

Keywords : Care Economy, Unpaid Work, Mixture Models, Statistical Learning, Microeconometrics

1. Introducción

Incluya máximo 3 objetivos; el objetivo general y un par de objetivos específicos.

- Ejecutar la programación dinámica y simulación en el contexto de esta investigación.
- Determinar la distribución del recurso para cada uno de los canales de venta
- Modelar el comportamiento del funcionamiento de un hotel básico.

2. Metodología

Tratamiento y depuración de datos

La encuesta ENUT consta de 9 capítulos, de los cuales se destaca la importancia para este análisis del capítulo 8 (Uso del Tiempo), el cual pretende establecer la distribución en términos de tiempo, de las actividades de trabajo no remunerado así como las actividades personales realizadas por los miembros del hogar. Otros capítulos de la encuesta como lo son Condiciones de la Vivienda, Composición del Hogar, Educación, Salud y Fuerza del Trabajo, también fueron tenidos en cuenta.

De igual manera se filtra la población a personas con edad mayor a 13 años de nacionalidad colombiana.

Se realizo un proceso de imputación de datos para las siguientes 3 variables: nivel educativo, régimen de salud y estrato socioeconómico,, que respectivamente tenían 14 %, 6,7 % y 0,76 % de datos faltantes. Este proceso se llevo a cabo a través de imputación multivariante mediante ecuaciones encadenadas (MICE) [2]. Se generan múltiples imputaciones para cada valor faltante, creando varios conjuntos de datos completos. Los resultados se combinan para obtener una estimación final. Los métodos usados fueron: para las variable nominal régimen de salud, Multinomial logit model, y para las variables ordinales, nivel educativo y estrato socioeconómico, Ordered logit model. El número máximo de iteraciones se fijó en `maxit = 5`, se crearon `m = 10` bases de datos imputadas, y se uso la semilla `seed = 230125`.

Selección de variables

En una primera pre-selección se seleccionan alrededor de 397 variables a lo largo de los capítulos de la encuesta., viéndonos en la necesidad de reducir notablemente la dimensionalidad con la que se abordaría el análisis. Tras la eliminación de variables con gran cantidad de datos faltantes y agrupar variables desagregadas para conformar el tiempo total en minutos de TCNR y de tiempo personal, definido como actividades de cuidado personal, vida social, practicas de actividades físicas, culturales, recreativas o religiosas. Se llega a un numero aún considerable de variables.

En una última etapa del proceso de limpieza y selección de variables, se implementó el algoritmo de selección de características Boruta [3], este algoritmo esta diseñado como una envoltura alrededor de un algoritmo de clasificación Random Forest. Eliminando iterativamente las características que, según una prueba estadística, son menos relevantes que las sondas aleatorias. Una vez ordenadas las variables dependiendo su importancia media procedemos a comparar este ranking con los resultados obtenidos de la selección de variables realizada con Lasso y Elastic Net. Se realiza un análisis de colinealidad concluyendo que la mayor correlación de variables cuantitativas no supera el valor de 0,5. Para las variables categóricas, se usó la medida de asociación Cramér's V que permitió ayudar a descartar.
Además, se realiza un análisis a partir del GVIF (Generalized Variance Inflation Factor) con el cual se corroboró la ausencia de problemas serios de multicolinealidad registrando valores máximos de la métrica $\text{GVIF}^{(1/(2*\text{Df}))}$ de 1.6 lo cual es considerablemente menor a 2, valor de referencia usual.

Análisis exploratorio de los datos

Se analizaron un total de 120,653 individuos, los cuales corresponden a la muestra de interés con aquellas personas mayores a 13 años de edad, de las cuales, el 24 % manifestó no participar en actividades del TCNR. El restante 76 % reporta tiempos, en minutos, mayores a 0 dedicados a actividades de TCNR. Sin embargo la distribución de tiempos allí reportada es muy dispersa, como se evidencia en el Cuadro 1

Min	Q_1	Mediana	Media	Q_3	Max
1	70	170	212.4	300	2760

Cuadro 1: Distribución de la intensidad, en minutos, dedicada a actividades de TCNR

En la Figura 1 se puede evidenciar el comportamiento de los datos en su escala original, donde se observa que la distribución del tiempo de cuidado no remunerado presenta una alta asimetría positiva, con una gran concentración de observaciones en valores bajos (cercanos a cero) y una cola larga hacia la derecha. Esta distribución sesgada sugiere que una transformación logarítmica podría ser adecuada para estabilizar la varianza y aproximar una distribución más simétrica, lo que facilitaría un mejor ajuste del modelo y la interpretación de los efectos.

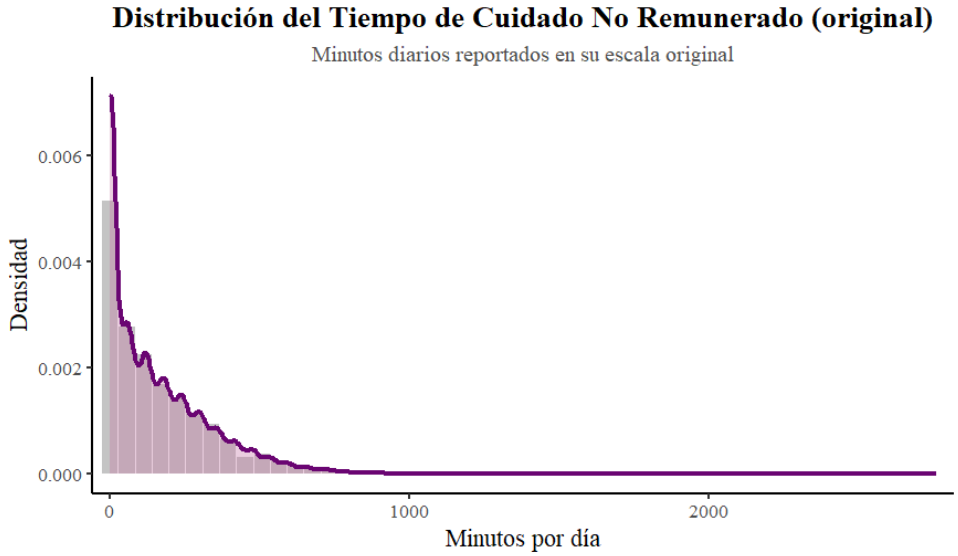

Figura 1: Histograma en escala original

Se decide hacer una transformación logarítmica para estabilizar la varianza de los datos, por lo que en la Figura 2 se presenta el histograma del tiempo de cuidado no remunerado en escala logarítmica. Esta transformación permite reducir la asimetría positiva observada en la distribución original y mitigar la influencia de valores extremos.

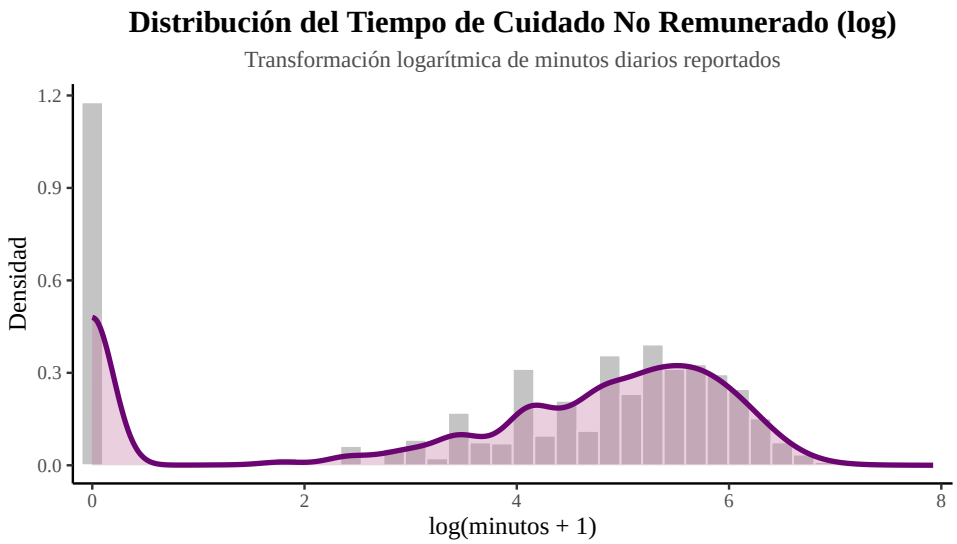

Figura 2: Distribución Tiempo TDCNR

La anterior gráfica corresponde a una transformación logarítmica de la distribución del tiempo dedicado a trabajo de cuidado no remunerado (minutos) dedicado semanalmente por los colombianos en el año 2021.

Como se puede observar, la distribución resultante tiene una alta concentración de valores en cero, lo que podría llegar a ser pensado como un problema de ceros inflados, o como se conoce en inglés, *zero-inflated distribution*. Además, se logra evidenciar ciertos picos de mayor concentración del tiempo, lo que tentaría a pensar en un distribución multinomial.

Segmentación y modelización

HDBSCAN

Grupo	Número de observaciones	Observaciones
5	69,502	Grupo principal
-1	47,928	Ruido o outliers
3	1,432	Subgrupo significativo
1	732	Subgrupo pequeño
2	379	Subgrupo pequeño
0	350	Subgrupo pequeño
4	330	Subgrupo pequeño

Cuadro 2: Distribución de observaciones por grupo según HDBSCAN

Se implementó el algoritmo HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) con el objetivo de realizar un agrupamiento sin necesidad de especificar a priori el número de clústeres. Adicionalmente, este enfoque permite identificar observaciones atípicas que no presentan pertenencia clara a ningún grupo. El procedimiento reveló, en concordancia con la naturaleza de los datos sociales, la existencia de un número considerable de individuos que no contribuyen de manera significativa a la formación de patrones de agrupamiento. Una propiedad fundamental de este método es que la conformación de los clústeres no se basa en una única variable de interés, sino en la estructura distribucional multivariada y en la densidad local de los datos, lo que permite una segmentación más fiel a la complejidad de la información analizada.

Kmeans

Se realizó una segmentación a partir del algoritmo de K-means para la cuál se fijó 4 clústers debido a estabilización de la suma de cuadrados intra-clúster. Con 4 clústers se observó que existe un grupo de individuos que no realiza TDCNR y otras dos agrupaciones cuya dedicación es mínima semanal comparado con el cúster número 1 el cual reporta en promedio 6 horas semanales dedicadas.

Clúster	Minutos promedio	Horas promedio
1	341.6	5.69
2	106.5	1.78
3	0.005	0
4	23.6	0.39

Cuadro 3: Resumen del tiempo en cuidado por clúster

Se emplearon 3 modelos para describir el comportamiento de las horas semanales dedicadas a TDCNR; Hurdle, Zero Inflated Negative Binomial, y GAM, los cuales se describen a continuación.

Hurdle:

Objetivo: Modelar una variable y que toma valores positivos solo para algunos individuos (participantes).

Supuestos clave:

- $d = 1$ si el individuo participa ($y > 0$), $d = 0$ si no ($y = 0$).

- Se estima $\Pr(d=1|x)$ con un modelo logit/probit.
- Para participantes, $f(y|d=1,x)$ modela los valores positivos.

El *modelo de dos partes* [4] para y se expresa entonces como

$$f(y|x) = \begin{cases} \Pr[d=0|x] & \text{si } y=0, \\ \Pr[d=1|x] f(y|d=1,x) & \text{si } y>0. \end{cases}$$

La estructura general de un modelo de dos partes [5] es:

$$E[Y|X] = Pr(Y>0|X) \times E[Y|Y>0,X]$$

1. Elección del modelo para participación:

- Logit o Probit, con formulación latente: $d=1$ si $x'\beta+\varepsilon>0$.

2. Distribución de y para los que participan:

- Distribuciones positivas: Log-normal, Gamma, NB truncada, etc.
- Se garantiza $y>0$ (normal truncada desde cero).

3. Estimación por máxima verosimilitud:

- Separada por partes:
  - Parte 1: Todas las observaciones (modelo binario).
  - Parte 2: Solo observaciones con $y>0$.

Zero Inflated Negative Binomial

Sea Y una variable aleatoria discreta (cantidad de minutos semales realizando trabajo de cuidado) y un proceso estructural (ser cuidador), entonces el modelo NBZI se especifica como

$$\begin{cases} Y_i^{\text{ind}} \sim \text{ZINB}(\mu_i, \phi, \tau, \pi_i), \\ g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \\ h(\pi_i) = \gamma_0 + \gamma_1 z_{i1} + \cdots + \gamma_q z_{iq} \end{cases}$$

Nota: Adecuado en presencia de sobredispersión ($\approx 1{,}4$)

Donde

$$Y \mid \nu = \begin{cases} 0 & \text{if} \quad \nu = 1 \\ \text{Neg. Binomial}(\mu, \phi, \tau) & \text{if} \quad \nu = 0 \end{cases} \quad \text{y } \nu \sim \text{Bernoulli}(\pi)$$

La función de masa de probabilidad de Y es

$$f_Y(y;\mu,\phi,\pi) = \begin{cases} \pi + (1-\pi)\left(\frac{\frac{\mu}{\phi}}{\mu+\frac{\mu}{\phi}}\right)^{\frac{\mu}{\phi}} & \text{if} \quad y=0 \\ (1-\pi)\frac{\Gamma\left(y+\frac{\mu}{\phi}\right)}{\Gamma\left(\frac{\mu}{\phi}\right)\Gamma(y+1)}\left(\frac{\mu}{\mu+\frac{\mu}{\phi}}\right)^{y}\left(\frac{\frac{\mu}{\phi}}{\mu+\frac{\mu}{\phi}}\right)^{\frac{\mu}{\phi}} & \text{if} \quad y=1,2,\ldots \end{cases}$$

- μ: Número esperado del conteo dado que no es 0.
- ϕ: Controla la sobredispersión del conteo.
- Funciones de enlace: $g(\mu_i) = log(\mu_i)$ y $h(\pi_i) = logit(\pi_i)$

GAM

El GAM es una extensión flexible del modelo lineal que permite capturar relaciones no lineales entre las variables independientes y la variable dependiente. En lugar de usar coeficientes lineales simples, emplea funciones suaves estimadas no paramétricamente (por ejemplo, splines) para cada predictor. Además, utiliza una función de enlace $g(\cdot)$ que conecta la media esperada de la respuesta con la suma de estas funciones suaves:

$$g(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p).$$

Se ajustaron dos modelos completos (Hurdle y ZINB) involucrando las siguientes variables.

Edad	Sexo	Parentesco	Etnia
Régimen de Salud	Actividad Semana Pasada	Percepción del Cuidado	Percepción del Tiempo
Clase	Región	Estrato	Vivienda
Servicio Doméstico	Total del Hogar	Subsidio	Tiempo de Ayuda Recibida
Tiempo Trabajado	Tiempo de Viaje al Trabajo	Nivel Educativo	Tiempo Personal

Cuadro 4: Variables incluidas en los modelos completos

De los modelos anteriores se decantaron las siguientes variables para ajustar sobre estas los modelos con interacciones Hurdle, ZINB, y GAM.

Edad	Sexo $\times$ Parentesco	Sexo $\times$ Percepción del Cuidado
Percepción del Tiempo	Clase $\times$ Sexo	Estrato
Total del Hogar	Tiempo de Ayuda Recibida	Nivel Educativo
Sexo $\times$ Tiempo Personal		

Cuadro 5: Variables e interacciones incluidas en los modelos reducidos

3. Resultados y análisis

Las gráficas de la figura 3 representan una comparación del desempeño de los modelos completos y con interacciones según 5 métricas.

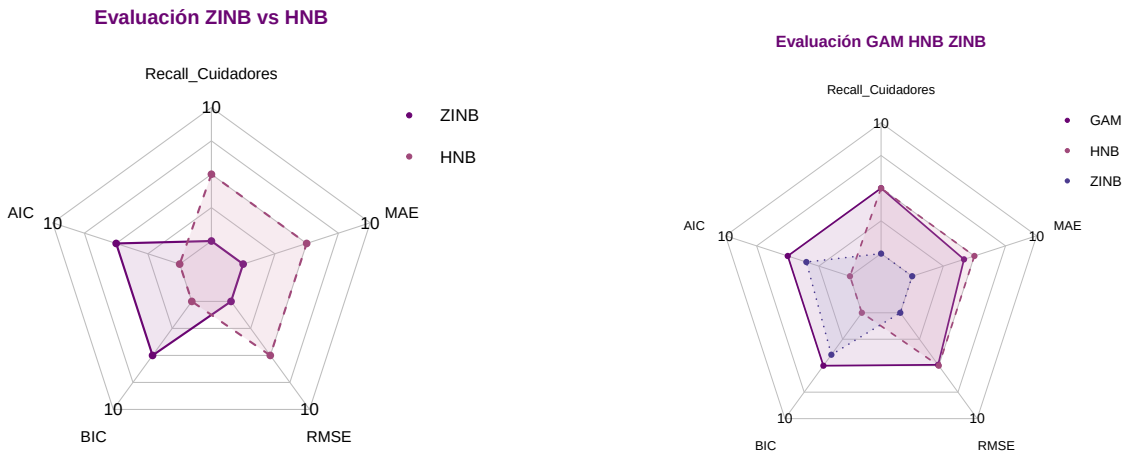

Figura 3: Desempeño modelos completos y con interacción

Como se observa, usando todas variables, el modelo Hurdle tiene una mejor capacidad predictiva que el modelo ZINB, sin embargo es el anterior el que tiene una mejor bondad de ajuste. Por otro lado, al considerar el segundo conjunto de variables con interacción, es posible resaltar ahora, que el mejor modelo es el GAM, siendo únicamente superado por el Hurdle en la métrica del error absoluto medio.
Los siguientes resultados fueron los obtenidos a partir del modelo ZINB.

Proceso estructural: no realizar cuidado

- Las mujeres tienen 96 % más chances de realizar tareas de cuidado que los hombres.
- Por cada año adicional en la edad, el chance de no realizar cuidado aumenta en un 8,9 %.
- A mayor número de personas en el hogar, el chance de no realizar cuidado aumenta en un 10 %.
- El chance de realizar cuidado siendo mujer y pareja del jefe de hogar aumenta en 58 %.
- Ser hija/hijo o nieta/nieto del jefe del hogar aumenta el chance de no realizar cuidado en 44 % y 66 % respectivamente.

Intensidad de tiempo realizando cuidado

- El tiempo esperado dedicado a labores de cuidado por las mujeres es aproximadamente 2.22 veces mayor que los hombres.
- Por cada año adicional de edad, el tiempo esperado realizando cuidado disminuye en un factor de aproximadamente 0.13 % por año.
- A mayor estrato el tiempo esperado realizando actividades de cuidado disminuye en un factor de 3,9 %.
- El alcanzar un mayor nivel educativo reduce la media esperada de tiempo de cuidado en un factor de 29.8 %

4. Conclusiones

- La feminización persistente revela cómo el cuidado se naturaliza como extensión del rol femenino.
- Variables demográficas como la edad y la composición del hogar (número de integrantes) aumentan las chances de no realizar cuidado, lo que sugiere que la participación en estas tareas puede desplazarse hacia otros miembros en hogares más grandes y a medida que las personas envejecen.
- Una vez inmersas en tareas de cuidado, las mujeres dedican más del doble de tiempo que los hombres, reflejando no solo un mayor involucramiento, sino también una mayor carga de trabajo.
- El envejecimiento tiene un efecto de reducción en la intensidad del cuidado, aunque marginal (0.13 % menos por cada año), indicando que las responsabilidades tienden a disminuir con la edad.
- La condición socioeconómica (estrato) actúa como un factor protector: personas de estratos más altos dedican menos tiempo al cuidado, lo que puede estar vinculado a la posibilidad de tercerizar estos servicios.
- Un mayor nivel educativo está fuertemente asociado con una disminución en la intensidad del tiempo de cuidado, lo que sugiere que las trayectorias educativas podrían reducir las barreras de género y redistribuir las cargas de cuidado.

5. Recomendaciones

Aún hay mucho camino por recorrer, algunas lineas de investigación adicionales para aquellos interesados en involucrarse a trabajar con estos datos son:

- Análisis de Clases Latentes
- Análisis Conjunto de las Ediciones de la ENUT
  - Datos Panel
- Modelos Mixtos
  - Efectos Aleatorios
- Aplicaciones de Muestreo
  - Estimadores, CV, IC
  - Calibración
  - Uso de variables auxiliares
- Análisis con la desagregación de las actividades de cuidado

Referencias

[1] Congreso de la República de Colombia. *Ley 1413 de 2010: Por la cual se regula la inclusión de la economía del cuidado en el sistema de cuentas nacionales*. Diario Oficial No. 47.890. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=40282. 2010.

[2] S. van Buuren y K. Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R". En: *Journal of Statistical Software* 45.3 (2011), págs. 1-67. DOI: 10.18637/jss.v045.i03.

[3] M. B. Kursa y W. R. Rudnicki. "Feature Selection with the Boruta Package". En: *Journal of Statistical Software* 36.11 (2010), págs. 1-13. DOI: 10.18637/jss.v036.i11.

[4] A. C. Cameron y P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge, UK: Cambridge University Press, 2005.

[5] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2010.