Aproximación a la distribución funcional entre trabajadores.
Un análisis a partir de técnicas multivariadas para el período 2006-2021 en Uruguay.

Keywords— Distribución funcional del ingreso, Análisis de Cluster, Análisis Discriminante

Abstract

En Uruguay si bien se dispone de numerosos estudios sobre la desigualdad personal del ingreso para los últimos 30 años, la distribución funcional del ingreso ha sido escasamente estudiada. Con este segundo enfoque podemos estudiar las diferencias en la participación de las fuentes de ingreso en el excedente, lo que permite considerar por separado grupos económicos con características muy diferentes. Sin embargo no permite considerar heterogeneidades entre trabajadores, siendo estas importantes determinantes de la desigualdad de ingresos. El principal objetivo de este trabajo consiste en analizar la participación de la masa de remuneraciones laborales en el ingreso, y analizar su apropiación por parte de distintos grupos de trabajadores. Ello requiere la identificación de una clasificación de trabajadores a partir de características personales y del puesto de trabajo, para estudiar la composición de la masa salarial y su evolución. Para esto se propone utilizar métodos multivariados. Se aplica el análisis de cluster sobre variables que reflejen características individuales de los trabajadores y su puesto de trabajo, con la finalidad de identificar grupos diferenciados de trabajadores. Los grupos son identificados en un año base y luego con el análisis discriminante se asigna a los individuos de los restantes años en la partición de grupos formada en dicho año base. Teniendo a los individuos de todos los años clasificados en grupos, se verá la evolución de la participación de estos dentro de la masa salarial en los años seleccionados, y su capacidad de apropiación sobre la misma. Se utilizan datos de Encuestas Continuas de Hogares (INE) para los años 2006, 2011, 2019 y 2021, entendiendo que representan momentos marcadamente distintos de la dinámica económica reciente.

1 Metodología

Para el **análisis de cluster** se contó con distintas variables categóricas, por lo que se eligió como métrica al coeficiente de similitud de Jaccard, definido como el tamaño de la intersección dividido por el tamaño de la unión de los conjuntos de muestras:

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{|A \bigcap B|}{|A| + |B| - |A \bigcup B|}$$

En su aplicación particular a este problema se requiere de él para considerar similares a quienes compartan categorías de las variables. De esta forma luego del análisis individuos que compartan un conjunto de categorías estarán asignados a un mismo grupo.

Se optó por métodos de clusterización no jerárquicos, la cantidad de grupos se definió por el criterio de la máxima silueta promedio. En particular el método a utilizar es K-Medoides o partición en torno a los medoides (PAM), donde cada cluster está representado por el individuo mediano candidato. Dado el tamaño de los datos se utilizará la implementación CLARA de K-Medoides (Kaufman y Rousseeuw, 1990), que se basa en el enfoque del muestreo.

Este análisis se realiza sobre un año base (2011) y como resultado se conocerá la pertenencia de los individuos a los grupos formados. Para lograr la misma clasificación de grupos para los cuatro años seleccionados se aplica el **análisis discriminante**, derivando una regla para asignar las observaciones externas (datos de los años 2006, 2019 y 2021) a los grupos haciendo mínima la probabilidad de clasificar incorrectamente.

Entre las distintas funciones discriminantes se eligió el discriminante logístico dado que todas las variables son cualitativas (Peña, 2002).

Adelantando algunos resultados, como se llega a más de dos grupos, suponemos que la variable que indica las subpoblaciones proviene de una distribución multinomial. Se presentan a continuación las probabilidades logarítmicas entre grupos (James et al., 2021):

$$log(\tfrac{Pr(Y=k|X=x)}{Pr(Y=K|X=x)}) = \beta_{k0} + \beta_{k1}x_1 + ... + \beta_{kp}x_p$$

Los individuos se asignan a un grupo o a otro comparando las probabilidad a priori, con esta regla de clasificación el grupo más probable es aquel donde su verosimilitud es más alta.

Para evaluar el modelo se utiliza una medida de precisión definida como la cantidad de datos bien clasificados sobre el total, valores altos de la misma implican un buen desempeño del modelo. Como es de interés clasificar datos externos, para evitar problemas de sobreidentificación se utiliza el método de validación cruzada $k - folds$ (James et. al., 2021).

El trabajo fue principalmente realizado en R, desde el procesamiento de datos con la librería tidytable, la generación de visualizaciones a partir de ggplot2 (v3.4.2; Wickham, 2016), y la implementación de los métodos con las librerías cluster (v2.1.4; Maechler, 2022) y nnet (v7.3-19; Venables y Ripley, 2002). Adicionalmente se crearon las funciones en R necesarias para la aplicación de la cross-validation, y se utilizó como herramienta Docker para fomentar su reproducibilidad.

2 Resultados

Se lograron identificar 4 grupos con diferencias muy marcadas en las características de los trabajadores y su puesto de trabajo. Uno de menores ingresos y con los peores indicadores, y los otros tres con niveles similares de ingresos pero diferenciados por variables de educación, tareas y sector. La clasificación formada permitió hacer distintos análisis sobre la distribución del ingreso laboral entre estos grupos. Los resultados son alentadores para la búsqueda de mayor complejidad en el esquema de la distribución funcional, considerando heterogeneidades entre trabajadores a partir de métodos multivariados.

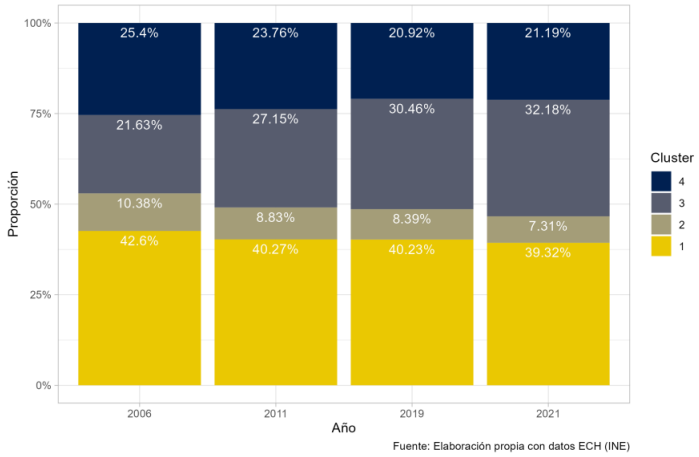

Figure 1: Composición de la masa salarial

Referencias:

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Kaufman, L., Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley Sons.
- Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2022). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source), https://CRAN.R-project.org/package=cluster.
- Peña, D. (2002). Análisis de datos multivariantes (Vol. 24). Madrid: McGraw-hill.
- Venables WN, Ripley BD (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/.
- Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4