CEDLAS
Facultad de Ciencias Económicas
UNIVERSIDAD NACIONAL DE LA PLATA

DOCUMENTOS DE TRABAJO

Big Data y Algoritmos para la Medición de la Pobreza y el Desarrollo

Walter Sosa Escudero

Documento de Trabajo Nro. 319
Octubre, 2023
ISSN 1853-0168
www.cedlas.econo.unlp.edu.ar

Cita sugerida: Sosa Escudero, W. (2023). Big Data y Algoritmos para la Medición de la Pobreza y el Desarrollo. Documentos de Trabajo del CEDLAS Nº 319, Octubre, 2023, CEDLAS-Universidad Nacional de La Plata.

Big data y algoritmos para la medición de la pobreza y el desarrollo

Walter Sosa Escudero[1]

Eritrea es un pequeño país del noreste de África, al que casi todos los rankings ubican entre los tres más pobres del mundo. Eritrea hace 30 años que no tiene un censo (Jerven, 2013). Eritrea es como un hipertenso severo que no puede medirse la presión, que no puede consultar a un médico.

La medición estándar de la pobreza depende de la implementación periódica de un sistema de encuestas, que excede las posibilidades de un país paupérrimo como Eritrea. Y aún lejos de la situación extrema de países como los del Africa Subsahariana, los costos de la medición oficial de la pobreza en países de desarrollo intermedio, como la Argentina, hacen que las cifras disponibles no logren captar con precisión las áreas rurales, los barrios marginales o ciertos grupos de interés (como los pueblos originarios o los jóvenes), particularmente afectados por el azote de la privación.

La revolución del combo big data-machine learning-inteligencia artificial ha invadido todos los campos del conocimiento y, esperablemente, el de la medición del bienestar no es una excepción. Y, naturalmente, urge preguntar si los enormes problemas de cuantificación de la pobreza o la desigualdad no encontraran una solución rápida y efectiva que provenga de la combinación de datos masivos de big data y los poderosos algoritmos de machine learning y la inteligencia artificial.

Esta nota es una introducción técnicamente accesible a los logros y desafíos del uso big data y machine learning para la medición de la pobreza, el desarrollo, la desigualdad y otras dimensiones sociales. Se basa en Sosa Escudero, Anauati y Brau (2022), un artículo abarcativo y técnico, que estudia con detalle el estado de las artes en lo que se refiere al uso de machine learning para los estudios de desarrollo y bienestar, al cual remitiremos para mayores detalles y referencias específicas.

[1] Profesor plenario de UdeSA, investigador principal del CONICET, investigador CEDLAS/UNLP. wsosa@udesa.edu.ar Se agradecen los comentarios de Leopoldo Tornaroli a algunas inquietudes en relación con la medición de la pobreza en la Argentina.

¿Midiendo lo inmedible?

El problema de la medición de la pobreza es complejo, porque no existe ninguna forma inequívoca de definirla, y, aun habiendo logrado cierto acuerdo conceptual, no hay formas indiscutibles de medirla. Las estrategias comúnmente utilizadas, como el "enfoque de líneas" (que considera que un hogar es pobre si sus ingresos no alcanzan para comprar una canasta básica de bienes o "línea de pobreza"), son meros acuerdos técnicos y operativos, que negocian las dificultades que implica lidiar con un objeto tan complejo y multidimensional como la privación, con la necesidad de contar con mediciones relevantes para el diagnóstico y la implementación de políticas efectivas para paliar las necesidades de quienes menos tienen. Es decir, la medición de la pobreza pretende ser útil aun cuando no necesariamente buena.

En los países de desarrollo intermedio, como la Argentina, la solución a este dilema entre fineza conceptual y pragmatismo descansa en un complejo sistema de encuestas periódicas. A modo de ejemplo, la Encuesta Permanente de Hogares que implementa el INDEC argentino permite medir la pobreza en forma razonable, estable y con una frecuencia apropiada, y habilita comparaciones temporales o regionales de modo de saber no solo la magnitud del fenómeno de la pobreza (cuantos hogares pobres hay en una región en cierto momento) sino también si hay más pobres en una región que en otra, o si en una misma región la pobreza subió o bajó.

Y aun cuando resulte útil para ciertos propósitos, la medición por líneas basada en encuestas periódicas tiene severas limitaciones de cobertura. Las áreas rurales son de difícil alcance por este tipo de mediciones, igual que los barrios marginales o grupos de interés como los jóvenes, o los inmigrantes, lo que demanda una "granularidad" que se contrapone con el alcance mayoritariamente urbano y limitado a grandes aglomerados de las encuestas tradicionales.

Los problemas de medición se agigantan cuando se reconoce la naturaleza *multidimensional* del bienestar, influida por los trabajos seminales de Amartya Sen (Sen, 1985). El enfoque de líneas focaliza en los aspectos de la pobreza que pueden ser captados (directa o indirectamente) por el ingreso. Dimensiones relevantes como la disponibilidad de activos, la calidad del capital humano o el acceso a redes familiares de contención son difíciles, cuando no imposibles, de captar con los sistemas tradicionales de encuestas. Gasparini, Cicowiez y Sosa Escudero (2013) describen con detalle los múltiples problemas que conlleva la medición del bienestar, y una introducción informal es el capítulo 6 en Sosa Escudero (2022).

¿Big o new data?

La visión más inocente de big data lleva a creer si todavía no se dispone de todos los datos, es solo cuestión de esperar un poco. Y así, en no muchos años y como en la Biblioteca de Babel de Borges, aparecerán en el océano de big data los datos de las encuestas hoy inexistentes en Eritrea o los que faltan en la Encuesta Permanente de Hogares (EPH) argentina para medir la pobreza rural.

Y el argumento tiene algo de cierto y también de engañoso. Engañoso porque los datos de big data son de una naturaleza distinta a los de una encuesta científicamente diseñada. Las encuestas obedecen a un estricto diseño muestral, que garantiza que con unos pocos datos (2.640 hogares en la EPH del Gran Buenos Aires) es posible medir el bienestar de una población mucho más grande (aproximadamente 5.320.000 hogares). 2.640 es un número ínfimo para los estándares de big data, en donde cualquier *celebrity* menor tiene seguidores en twitter que se cuentan de a millones. La enorme diferencia es que el paradigma de *small data* que rige a las encuestas o a los experimentos científicos es que, por su diseño, muy pocos datos contienen muchísima información, como una cucharadita de la olla de una salsa bien revuelta. Dicho de otra forma, big data no es más de lo mismo, es un monstruo grande pero de una naturaleza completamente distinta: son datos espontáneos, anárquicos, observacionales, libres de estructura.

La parte optimista de la promesa de big data no se relaciona con la masividad sino con su naturaleza innovadora. Los datos faltantes para medir la pobreza en Erítrea o en Santiago del Estero no aparecerán por arte de magia en un olvidado "baul virtual", sino que cierta inteligencia encontrará en la masividad de datos una forma de aproximarlos razonablemente con *otro* tipo de datos. Y ahí está la verdadera promesa de big data: que el océano de datos anárquicos contenga alguna laguna de datos cristalinos que permita aproximar el bienestar en forma simple, sin necesariamente depender de la costosa institucionalidad de los sistemas de encuestas periódicas. En lo que refiere a la cuestión de la medición del bienestar, la pobreza o la desigualdad, más que de *big*, la revolución es de *new* data.

Medir, extrapolar, dimensionar y visualizar la pobreza con big data y algoritmos

El trabajo seminal de Blumenstock, Cadamuro y Ok (2005) es indicativo del potencial y las limitaciones del uso de big data y algoritmos para medir el bienestar y la pobreza. Ruanda es un país similar a Eritrea en lo que respecta a la urgencia de la pobreza y a la inviabilidad de apelar

a un sistema de encuestas periódicas para medirla. Estos autores parten de notar que el uso de teléfonos celulares en Ruanda se encuentra lo suficientemente extendido como para que exista alguna relación entre la intensidad de su uso y el bienestar. De modo que procedieron a *entrenar* un modelo simple para predecir la pobreza sobre la base de a la intensidad de uso de celulares. Los datos de bienestar vienen de una pequeña pero cuidadosa encuesta, y los de celulares, de una empresa privada. Luego de entrenado, el modelo es utilizado para predecir la pobreza en todo el territorio de Rwanda, con una "granularidad" de 1 kilómetro cuadrado y, de acuerdo a los autores, con un costo 500 veces inferior y una disponibilidad 20 veces más rápida que la de una encuesta tradicional. Lo relevante de este enfoque es que los datos más importantes para el estudio (la intensidad del uso de teléfonos) no vienen de ninguna encuesta ni un experimento formal sino de la "huella digital" que dejan los usuarios al usar sus celulares.

Los últimos años han sido testigos de un auténtico aluvión de métodos similares. Así, al uso de celulares se suman estrategias basadas en la intensidad de las luces captadas por imágenes satelitales, los datos de redes (sociales, de comercio, postales, migratorias), los de empresas privadas como Ebay o Linkedin, el análisis geográfico del tipo de material usado para las construcciones de viviendas y la textura de sus techos (captados por imágenes satelitales), entre muchas alternativas, descriptas con detalle en Sosa Escudero et al. (2022). Todas estas estrategias sustituyen los datos de bienestar obtenidos por encuestas por otros, "de big data", que permiten aproximar el fenómeno de la medición de la pobreza.

Un problema en donde big data ha provisto soluciones valiosas es el de la interpolación o extrapolación de información relevante. El temprano trabajo de Elbers, Lanjouw y Lanjouw (2003) es pionero en la "microestimacion" de la pobreza a partir de datos más desagregados. Esto permitió obtener avances considerables en la medición de la pobreza en áreas de difícil acceso, como los barrios de emergencia o las zonas rurales, o la interpolación de datos censales, disponibles, en el mejor de los casos, cada diez años, como en Argentina.

La visualización de la información del bienestar dista de ser un punto menor, todo lo contrario: es una herramienta comunicacional crucial para la focalización de las políticas públicas y para la creación de consensos que las apoyen. El proyecto Atlas del Capital Social liderado por Raj Chetty (Chetty et al. 2022), es una ambiciosa herramienta visual para explorar el alcance del capital social (las relaciones comunitarias y personales) basada en 21 mil millones de relaciones de amistad medidas en Facebook. Se trata de un enfoque innovador, que marca el futuro del tipo de investigación social moderna en los años por venir.

La cuestión de la dimensionalidad es un punto importante en la caracterización del bienestar y también un tema central a machine learning. Los trabajos que sucedieron al de Sen (1985) muestran que las mediciones basadas solo en el ingreso no logran captar adecuadamente el bienestar, y que hacen falta otras variables o dimensiones. Por otro lado, una contribución central del bagaje de machine learning es un conjunto de técnicas que permiten estudiar la dimensión subyacente a un conjunto complejo de información. Más concretamente, tres preguntas obvias son: 1) ¿es realmente multidimensional el bienestar? 2) Si lo es ¿cuántas dimensiones hacen falta para captarlo apropiadamente? 3) Y si el bienestar es esencialmente multidimensional, ¿cuán erradas son las medidas basadas en ver nada más que el ingreso?

Algunos artículos en los que estuvo involucrado el autor de esta nota apuntan a echar luz sobre esta dimensión. Gasparini et al (2013) usan análisis de clúster y factores, y muestran que, efectivamente, los algoritmos sostienen la hipótesis de multidimensionalidad y que, sorprendentemente, el ingreso hace un papel razonable en captar el bienestar. Edo et al (2022) desarrollan un método moderno de selección de variables para reducir la dimensión del bienestar, y lo utilizan para detectar y medir la clase media argentina.

El desafío de la evaluación de las políticas

Una parte central de la agenda de la política social estuvo dominada por la evaluación de efectos causales, como los que surgen de un experimento en las disciplinas clásicas como la biología. Varias cuestiones éticas y operativas hacen que la ruta experimental se vea restringida en las disciplinas sociales. Asi es que la llamada “revolución de credibilidad” en econometría dedico mucha energía al diseño de métodos estadísticos que permiten realizar inferencias causales aun con datos observacionales, que no vienen de un experimento concreto. Si bien todavía incipiente, la combinación de big data y machine learning para el análisis causal es un área de investigación fértil. Recientemente, estos métodos, que mezclan el análisis causal clásico y machine learning, han resultado útiles para el mismo diseño de los experimentos, para decidir quiénes deben recibir una política social (como un subsidio), para medir efectos heterogéneos de las políticas (si cierto subsidio beneficia a todos por igual o lo hace en forma distinta para ciertos grupos) o para combinar fuentes de datos tradicionales (experimentos o encuestas) con información “de big data”. Se trata de una temática de frontera, que se espera que sintetice el mundo de la estadística inferencial clásica con la visión moderna de aprendizaje.

Comentarios finales

Las estadísticas sociales son un fenómeno tan técnico como social. La medición de la pobreza es un *acuerdo* que negocia la imposibilidad de hacerlo en forma indiscutible con la necesidad pragmática de disponer de alguna cifra que asista al diagnóstico y a la implementación de la política social. Este acuerdo es de una naturaleza científica ("los métodos funcionan y son útiles") y también política y comunicacional. Posiblemente, el principal desafío de la medición moderna de la pobreza en base a big data sea la construcción de un consenso que convenza a la sociedad (científicos, políticos, comunicadores, ciudadanos) de la confiabilidad de estas nuevas medidas.

Las estadísticas sociales clásicas no serán reemplazadas por machine learning, tal vez todo lo contrario: su estructura puntillosa hace que las mismas funcionen como "piedra de Roseta" para el entrenamiento y evaluación de métodos alternativos.

La estabilidad de las mediciones es un requisito crucial de la estadística social, lo que pone un freno natural al impulso de machine learning en la cosa social. Un método novedoso (en términos de eficiencia o alcance) tiene que lidiar con el inevitable problema de "comparar peras y manzanas": cambiar el método conduce a discutir si la pobreza bajo (o subió) porque efectivamente lo hizo en la realidad o porque cambio el método. Una vez más, es la interacción entre el sistema científico y el político lo que garantiza que los beneficios de la innovación más que compensen a los conflictos comunicacionales de cambiar las mediciones.

A la larga, y paradójicamente, las contribuciones de machine learning y big data para la pobreza deberían conducir a que se hable poco de medirla y mucho diseñar y evaluar políticas que asistan a los que menos tienen.

Referencias y bibliografia

Blumenstock, J., Cadamuro, G. y On, R., 2015, Predicting poverty and wealth from mobile phone metadata, *Science*, 350(6264), 1073-1076.

Edo, M., Sosa Escudero, W., y Svarc, M., A multidimensional approach to measuring the middle class, 2021, *Journal of Economic Inequality*, 19, pp. 139–162.

Gasparini, L., Cicowiez, M., y Sosa Escudero, W., 2013, *Pobreza y Desigualdad en América Latina. Conceptos y Herramientas Analíticas*, 2013, Editorial Temas, Buenos Aires.

Gasparini, L., Marchionni, M., Olivieri, S. y Sosa Escudero, W., Multidimensional Poverty in Latin America and the Caribbean: New Evidence from the Gallup World Poll, 2013, *Journal of Economic Inequality*, 11, 195–214.

Jerven, M., 2013, *Poor Numbers*, Cornell University Press, Ithaca, NY.

Sen, A., 1985, *Commodities and capabilities*, Oxford University Press.

Sosa Escudero, W., Anauati, V. y Brau, W., 2022, Poverty, inequality and development studies with machine learning, capitulo 9 en Chan, F. y Matyas, L. (eds) *Econometrics with Machine Learning*, Springer, New York

Sosa Escudero, W., 2019, *Big data*, 8ª edición, Siglo XXI Editores, Buenos Aires.

Sosa Escudero, W., 2022, *Que es (y que no es) la estadística*, 2da edición revisada, Siglo XXI Editores, Buenos Aires.