QUID 16 Revista del Área de Estudios Urbanos del Instituto de Investigaciones Gino Germani de la Facultad de Ciencias Sociales (UBA) - ISSN 2250-4060

Dossier

La integración de fuentes diversas de información en los estudios territoriales. Reflexiones sobre dos investigaciones

Germán Federico Rosati

Doctor en Ciencias Sociales, Universidad de Buenos Aires. Investigador Asistente en el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) con sede en el Instituto de Altos Estudios Sociales de la Universidad Nacional de San Martín (IDAES-UNSAM). Argentina.

E-mail: german.rosati@gmail.com

Fecha de recepción: 23/10/2021

Aceptación final: 16/02/2022

Los últimos años han estado caracterizados por un incremento notable en el volumen de información disponible para la investigación social. Se han publicado fuentes que podrían llamarse "tradicionales" (microdatos de censos y encuestas, series económicas, registros administrativos, etc.) y han aparecido nuevas fuentes de datos (redes sociales, plataformas, etc.). El presente trabajo busca reflexionar sobre el uso y la integración de fuentes heterogéneas en las ciencias sociales en general y en la problemática del territorio en particular. Para ello, reseña y analiza dos investigaciones que se basaron en este tipo de operaciones: una que busca detectar las diferentes modalidades de la expansión sojera en Argentina a nivel departamental y otra que desarrolló un índice de vulnerabilidad sanitaria a nivel radio censal para todo el país. Los problemas que surgen en dicha integración, la granularidad territorial de la información, la necesidad de avanzar en la utilización de cierto tipo de modelado que ayude a esta integración de fuentes y la posibilidad de replicabilidad y apertura de dicha información son algunos de los temas centrales del artículo.

Palabras clave: Estudios territoriales, integración de fuentes, modelos no supervisados, datos censales

Integration of diverse information sources in territorial studies. Reflections on two investigations

Abstract

The last few years have been characterized by a notable increase in the volume of information available for social research. Sources that could be called "traditional" have been published (microdata from censuses and surveys, economic series, administrative records, etc.) and new data sources have appeared (social networks, platforms, etc.). This work seeks to reflect on the use and integration of heterogeneous sources in the social sciences, in general, and in the problems of the

territory in particular. To do this, he reviews and analyzes two investigations that were based on this type of operations: one that seeks to detect the different modalities of soybean expansion in Argentina at the departmental level and another that developed a health vulnerability index at the radio census level for the entire country. The problems that arise in said integration, the territorial granularity of the information, the need to advance in the use of a certain type of modeling that helps this integration of sources and the possibility of replicability and openness of said information are some of the central issues. from the article.

Keywords: Territorial studies, source integration, unsupervised models, census data

Introducción

Los últimos años han estado caracterizados por un incremento notable en el volumen de información disponible para la investigación social. Su manifestación más notable se vincula con las nuevas fuentes de información proveniente de la llamada "revolución digital". Las redes sociales, las tecnologías mobile, el "internet de las cosas" y todas aquellas fuentes que pueden englobarse en el impreciso término de *big data* se muestran como un terreno fértil para el desarrollo de la investigación social (Salganik, 2018). En paralelo, se incrementan con el mismo ritmo las capacidades de cómputo para lidiar con esta información (Hilbert y López, 2011).

Suelen enumerarse las características del big data mediante las llamadas 3 "V": volumen, variedad y velocidad. La idea general es que son "muchos datos", que se producen en una gran variedad de formatos y estructuras (texto, imagen, sonido, video, grafos, etc.) y son creados de forma constante. Luego, fueron 5 "V", luego 7 "V" y se ha llegado a hablar de 42 "V"[1], casi parodiando la idea original. Lo cierto es que la información, que es producto de la era digital, tiene ciertas características que la diferencian de los datos cuantitativos tradicionales con los que se suele trabajar en las ciencias sociales. Salganik (2018) sistematizó estas características e identificó diez atributos importantes en el big data. Uno de los más importantes se vincula con la heterogeneidad de fuentes. Los datos de la era digital son **datos no estructurados** en varios aspectos (Sosa Escudero, 2019).

En primer lugar, el proceso de producción de estos datos es "no controlado". Al pensar en una encuesta como la Encuesta Permanente de Hogares (EPH) puede verse que existe un diseño altamente supervisado del instrumento de recolección: el cuestionario fue discutido largamente entre expertos nacionales e internacionales y se aplica de forma estandarizada a todas las unidades. En contraste, un *scraper* que trae información de sitios web recolecta la información tal y como es producida sin ninguna (o poca) mediación conceptual o metodológica (esta se produce luego de la recolección). Esta "espontaneidad" es un rasgo central en estas nuevas fuentes de datos.

[1] https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

En segundo lugar, mientras que en encuestas por muestreo probabilístico existe un proceso de selección cuidadoso de las unidades que forman parte del relevamiento (lo cual permite que con muy pocas unidades se puedan hacer inferencias del conjunto de la población objetiva), tal proceso de selección no existe en fuentes como Twitter o Facebook. Se obtiene información de los sujetos o de las unidades que estén adheridas a la plataforma (con todos los sesgos de selección derivados).

Por último, el resultado difiere considerablemente: mientras que la EPH produce un conjunto de tablas de datos ordenados que respetan la "estructura tripartita del dato" (Galtung, 1966), esta característica no suele hallarse en el big data: se trata de texto crudo, imágenes, audio, etc. Se requiere de otro proceso (que dependerá de la naturaleza de la información en cuestión) para producir esa estructura.

En este sentido, el trabajo con este tipo de información pone de manifiesto un problema: la integración/armonización de fuentes diversas. En efecto, lo que podría denominarse "triangulación de fuentes de datos" tiene un rol fundamental en el trabajo con la información producida en la "revolución digital". Dado que cada fuente es parcial, espontánea, con sus propios sesgos de selección y solamente ilumina un aspecto particular del objeto de estudio, el trabajo de integración de las mismas es sumamente importante. Este problema también puede trasladarse, sin demasiadas diferencias, al trabajo con fuentes tradicionales. Son muchas las situaciones en las que el uso de fuentes secundarias (encuestas, censos, etc.) plantea problemas similares que pueden servir como ejemplo para extraer algunas lecciones: en este trabajo se expondrán dos casos ilustrativos al respecto.

En paralelo a la aparición de estas nuevas fuentes, se ha ido desarrollando un proceso de apertura de información que podríamos llamar "tradicional", generalmente a partir de la decisión de los sistemas estadísticos nacionales y de organismos internacionales, de publicar una gran cantidad de información: microdatos de encuestas y censos de población[2], series de tiempo económicas[3], estandarización de registros administrativos[4] y otros son apenas algunos ejemplos de este proceso. Particularmente en Argentina, se ha generalizado la apertura de información censal en dos direcciones: 1) la publicación de información a diversos

[2] Es necesario mencionar el trabajo que el CELADE/CEPAL ha llevado adelante mediante la producción, difusión y apoyo a los sistemas estadísticos nacionales mediante su software REDATAM (https://redatam.org/es). Se generó una plataforma que hizo posible la publicación de microdatos correspondientes a diversos operativos censales en los países de Latinoamérica. También el Integrated Public Use Microdata Series (IPUMS), dependiente de la Universidad de Minesotta (https://www.ipums.org/), contribuyó de forma decisiva a la publicación de microdatos de relevamientos censales a nivel mundial, mediante un programa sistemático de extracción de muestras de los formularios censales.

[3] Puede mencionarse CEPALSTAT, portal de datos dependiente de la CEPAL con una gran cantidad de indicadores económicos, demográficos y laborales compilados y armonizados para los países de Latinoamérica, o la base de datos del Banco Mundial. Párrafo aparte merece la aplicación Povcalnet del Banco Mundial (http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx) que permite calcular diferentes líneas de pobreza para una gran cantidad de países del mundo y diferentes períodos de tiempo.

[4] Un ejemplo en esta dirección es el Observatorio de Empleo y Dinámica Empresarial dependiente del Ministerio de Trabajo (https://www.trabajo.gob.ar/estadisticas/oede/index.asp) que compila y publica información sobre los cambios en el empleo registrado y sobre demografía de empresas con base en datos del Sistema Integrado de Jubilaciones y Pensiones (SIJP).

niveles de desagregación (departamento y unidades menores), y 2) la publicación de bases usuarias de microdatos de los relevamientos censales.

Tanto la revolución digital como la apertura de datos oficiales han permitido el acceso a un gran caudal de información (en muchos casos, casi la única información existente para ciertos problemas de investigación) utilizable para el estudio de las transformaciones, los cambios y los movimientos en los territorios a diversas escalas analíticas. Al mismo tiempo, plantean una serie de problemas vinculados con la posibilidad de integrar ambos tipos de fuentes, las fuentes oficiales y las vinculadas a big data.

El presente trabajo se propone reflexionar sobre el uso de diversas fuentes de datos (tanto oficiales como big data). Como se intentará mostrar, el uso flexible y combinado de fuentes de datos heterogéneas (datos censales, big data, derivados de imágenes satelitales, etc.), sin estar exento de una serie de problemas referidos a su heterogeneidad y a las dificultades para su integración y armonización, abre una serie de posibilidades para la investigación social. Para ello, se propone el análisis de dos investigaciones que utilizaron este tipo de fuentes. Las dos trabajan con niveles de desagregación diferentes (a nivel departamental y de radios censales). En ambos casos, se integran fuentes heterogéneas (derivadas de imágenes satelitales, información censal, cartografías a diferentes niveles, ruteadores y grafos de calle, etc.), las cuales fueron necesarias armonizar y homogeneizar. Se utilizaron, además, técnicas de análisis multivariadas no supervisadas como parte del proceso analítico.

Se intentará mostrar cómo el trabajo con este tipo de fuentes permite articular una alta resolución espacial con una mirada agregada a nivel total del país. De esta forma, puede combinarse un análisis de información a un nivel desagregado en términos territoriales sin perder por ello la referencia a procesos que operan en una escala mayor. Así, es posible establecer vinculaciones entre los procesos que ocurren a escala pequeña (locales) y procesos que operan en escalas regionales, provinciales o nacionales.

Primer caso: identificación de las modalidades de articulación territorial de la expansión sojera

El problema

El primer caso está enmarcado en una investigación en curso que busca comprender, analizar e identificar las principales relaciones entre los cambios que se producen en la frontera agrícola y las transformaciones en el uso del suelo a nivel de la formación social argentina. Dentro de este marco, se diseñó un dispositivo metodológico basado en datos secundarios que permitiera una primera aproximación a la articulación regional de las diferentes modalidades de expansión de la frontera agrícola, con especial énfasis en la expansión sojera (el principal vector de los movimientos de la frontera agrícola). De esta forma, se buscaba construir información que hiciera posible visibilizar la co-ocurrencia territorial de las distintas formas de la expansión del cultivo de soja en el país. Se seleccionó como primer ejercicio el período 1988-2002 por dos motivos. El primero tiene que ver con la disponibilidad de fuentes: el Censo Agropecuario más reciente es del 2018 y, al

momento de redactar el trabajo, se estaban empezando a publicar los primeros datos a nivel departamental. El segundo tiene que ver con el hecho (que se menciona más abajo) de que existen pocos análisis que combinen una mirada general con el procesamiento de información desagregada. A su vez, el área del estudio busca abarcar al "área sojera", es decir, todos los departamentos que hayan sembrado al menos un 4% de la superficie total implantada en 1ra. ocupación con soja en 2001/02.

Buena parte de la literatura existente muestra que la expansión sojera se da de cuatro formas básicas: 1) procesos de desmonte y deforestación; 2) desplazamiento de la actividad ganadera; 3) introducción de soja de segunda (también conceptualizada como "adición de producciones" que conlleva la posibilidad de realizar dos cosechas, por ejemplo, trigo-soja); y 4) sustitución del área sembrada con otras actividades (Aizen, Garibaldi y Dondo, 2009; Páez, 2016; Viglizzo y Jobbágy, 2010).

Esto desató una serie de debates que tenían que ver con cuál era la forma predominante de esta expansión: ¿la soja se expandió a costa de superficies de bosques nativos y/o de superficie con otros cultivos? ¿O más bien su movimiento se concentraba mediante la forma de "soja de segunda", es decir, sin desplazar a otros cultivos? En estas discusiones[5] queda planteado un problema metodológico: dado que habitualmente se cuenta con conteos crudos y superficies agregadas, es difícil cuantificar el peso de los procesos de reemplazo/sustitución en tanto no se conocen los valores desagregados por hectárea[6].

Ahora bien, más allá de llegar a un valor que cuantificara cada modalidad de expansión, en la investigación que se presenta se buscó realizar una primera aproximación a la articulación entre estas diversas modalidades: ¿cuál es la distribución espacial y la co-ocurrencia de estos procesos a nivel departamental?

La cuestión de la unidad de análisis (departamental) no resulta trivial: en efecto, los estudios que abordan la problemática de la expansión sojera tienden a hacerlo regionalmente. Estas miradas regionales pueden estar basadas tanto en estudios de caso, es decir, en el análisis de estos procesos en determinados departamentos y/o provincias (León, Prudkin y Reboratti, 1985; Rodríguez, 2008; Steimbreger, Radnonich y Bendini, 2003; Valenzuela, 2014) como en miradas más generales a partir de la consideración de estadísticas agregadas provincial o regionalmente (Ackhar et al, 2011; Azcuy Ameghino y León, 2005; Azcuy Ameghino y Ortega, 2010). La primera estrategia intenta aportar mediante un análisis más focalizado, ganando en una descripción detallada del fenómeno. La desventaja es que puede llevar a perder de vista procesos estructurales de mayor escala. La segunda, si bien logra brindar un panorama general del fenómeno, puede, a su vez, perder en dicha agregación la especificidad de la información analizada. Ambas estrategias pueden perder de vista el hecho de que las dinámicas del movimiento de las fronteras no

[5] Trigo y otros autores (2002) consideran que la expansión sojera se daba predominantemente sobre esta segunda forma. Páez (2016) sostiene el mismo planteo para el período 2004-2014. Sin embargo, la información presentada por Rodríguez (2008: 87) para el período 1991-2006 parece refutar este hecho, al menos en lo que se refiere a la provincia de Buenos Aires.

[6] Una alternativa consiste en el análisis de imágenes satelitales: puede verse un ejemplo de aplicación en Volante et al (2015).

tienen un comportamiento lineal en el tiempo ni en el espacio: como mencionan Kröger y Nygren (2020) la expansión forma patchworks (mosaicos) territoriales.

Resulta necesario, entonces, avanzar en el análisis conjunto de las diferentes modalidades de expansión y que logre integrar diferentes escalas: que pueda recuperar información a un alto nivel de desagregación, pero que, al mismo tiempo, pueda reconstruir una mirada sobre estos procesos a nivel más agregado.

Aproximación a los procesos de deforestación

Para poder dar cuenta de la articulación territorial entre las diferentes formas de la expansión sojera en el período mencionado (1988-2002), fue necesario vincular dos grandes tipos de fuentes: datos censales y derivados de imágenes satelitales. Cada una permitía una aproximación a dimensiones y modalidades diversas y fue, a su vez, generada mediante diferentes procedimientos, presentando diferentes niveles de agregación. Es por ello que un primer paso necesario consistió en la homogeneización de ambas y su agregación al nivel departamental: una unidad de nivel de agregación intermedia entre las diferentes fuentes. Este procedimiento trae aparejado necesariamente la imposibilidad de diferenciar los procesos al interior de los departamentos, pero, al mismo tiempo, permite tener una mirada conjunta sobre el total de la estructura agraria argentina, pudiendo identificar diferentes articulaciones (es decir, co-ocurrencias de las diferentes modalidades en tales departamentos) para todas las estructuras agrarias argentinas involucradas en el proceso de expansión sojera. A su vez, la etapa analítica utiliza diferentes técnicas de estimación (inferencia ecológica) y de análisis (clustering) para detectar la articulación entre dichos procesos.

Se utilizaron datos provistos por el Monitoreo de Deforestación en el Chaco Seco (http://monitoreodesmonte.com.ar/, LART-FAUBA-INTA-REDAF). Los mismos sirvieron como una aproximación a la primera modalidad de avance sojero[7]: el avance sobre bosques nativos mediante la deforestación. Se trata de datos en formato vectorial (polígonos) que representan el área deforestada desde 1976 hasta la actualidad. Los polígonos fueron digitalizados a través de la fotointerpretación de imágenes satelitales Landsat (Vallejos et al 2015).

Aproximación a los cambios en el uso del suelo

Ahora bien, para lograr un acercamiento al resto de los procesos (los movimientos de sustitución de producciones y adición de soja de segunda) se utilizó información de los Censos Nacionales Agropecuarios de 1988 y 2002. Se tomó para cada departamento la cantidad de hectáreas sembradas de soja y la superficie sembrada total. Sobre esta matriz se estimaron diferentes modelos de inferencia ecológica (King, 1997) para estimar la cantidad de hectáreas que la soja sustituyó en cada departamento, lo que llamamos el "efecto sustitución". Estos modelos buscan realizar una estimación de datos a nivel individual disponiendo de datos agregados,

7

problema que suele surgir en estudios electorales. En esta investigación se utilizó el modelo de Gary King (1997), uno de los más recientes y que constituye una síntesis de varios modelos anteriores. El objetivo es tratar de estimar las celdas interiores de una tabla como la siguiente:

Tabla 1. Relación entre los parámetros a estimar

	Soja 2001 2002	Resto 2001 2002	Total
Soja 1987 1988	β_s^i	$1-\beta_s^i$	$S_{87,88}^i$
Resto 1987 1988	β_r^i	$1-\beta_r^i$	$1-S_{87,88}^i$
Total	$S_{01,02}^i$	$1-S_{01,02}^i$	1

Fuente: Elaboración propia basada en King (1997)

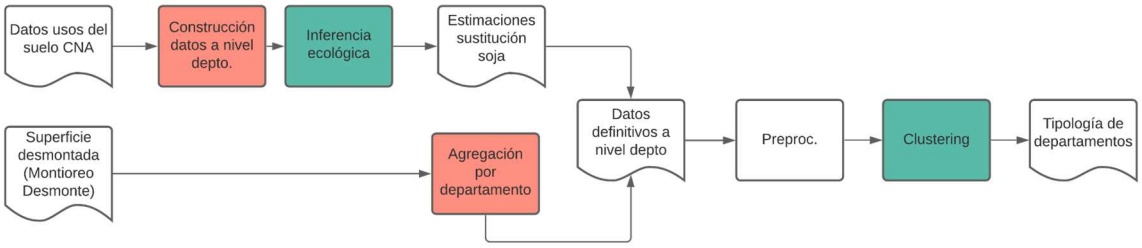
Se dispone de la información censal de los marginales de la tabla. Particularmente, el parámetro β_r^i es la cantidad de hectáreas que fueron sembradas con soja en 2001/02 pero estaban ocupadas con otros cultivos en 1987/88: este será el parámetro de sustitución.[8]

Por último, utilizando las fuentes censales cuantificamos la segunda forma de expansión sojera mencionada: el doble cultivo. Si bien el foco central estará en los procesos de expansión de la soja de primera, debido a que la doble ocupación del suelo introduce un problema numérico de estimación (en muchos departamentos la doble ocupación es nula) para los modelos de inferencia ecológica utilizados, se

[8] La propuesta de King (1997) utiliza tanto la información proveniente del método de límites (Duncan and Davis, 1953) como de la regresión de Goodman (1959) para fortalecer las estimaciones de los parámetros a nivel desagregado. Luego, a partir de ambas estimaciones desagregadas, estima los parámetros agregados. En primer lugar, se determina el rango de valores posibles de los parámetros en las unidades de interés. Luego, se modela la distribución de probabilidad conjunta de ambos parámetros. Así, en lugar de asumir (como en el modelo de Goodman) que los parámetros son "constantes" a lo largo de las unidades territoriales, el modelo de King asume que existe cierto nivel de correlación. Al mismo tiempo, el modelo produce valores dentro del rango [0,1]. Ambos requisitos se logran asumiendo que son extraídos de una distribución bivariada normal truncada (TBN). Luego se estiman los cinco parámetros de la TBN mediante el uso de regresiones lineales.
Finalmente, se producen estimaciones de las cantidades de interés (en nuestro caso, las celdas interiores de la Tabla 1) usando la información determinística sobre los límites de los parámetros en base a variantes de la ecuación 3. Esto se logra mediante procesos de simulaciones iterativas, dado que la derivación analítica del modelo es compleja.
Los resultados de la aplicación del modelo son los siguientes: una estimación de las cantidades agregadas de interés, n estimaciones (donde n es la cantidad de unidades en el dataset) de los parámetros a nivel distrito y otros productos como los límites de los parámetros para cada unidad, desvíos estándar de cada parámetro estimado, etcétera.

cuantificó el peso de la soja de segunda a nivel departamental en el 2002. Puede verse a continuación un esquema del flujo de trabajo.

Figura 1. Flujo de trabajo para la construcción de la tipología de departamentos



Fuente: Elaboración propia.

Detección de la articulación territorial de las diferentes formas de expansión sojera

El paso siguiente fue unificar los diferentes datos parciales: la información de deforestación, las estimaciones de los modelos de inferencia ecológica y la proporción de soja de segunda. La información construida generó una matriz en la que cada unidad constituía un departamento y cada columna una variable que cuantificaba algunos aspectos de la forma de expansión sojera (proporción de hectáreas deforestadas, variación de la superficie sembrada total, proporción de soja de 2da). A su vez, se calculó, para cada departamento a partir de los parámetros estimados mediante el modelo de inferencia ecológica, la cantidad (y la proporción) de hectáreas que se siembra bajo cada una de las cuatro modalidades (sustitución de soja, conservación de soja, sustitución de otros cultivos y conservación de otros cultivos).

Ahora bien, dado que lo que buscaba era poder encontrar combinaciones (co-ocurrencias) de las diferentes formas de expansión sojera a nivel departamental, se realizó un análisis de clustering jerárquico. El análisis de clústeres busca construir grupos de objetos (en este caso, departamentos) intentando que cada grupo tenga la máxima homogeneidad al interior y la máxima diferencia entre los grupos en función de las variables consideradas. (James, Witten, Hastie y Tibshirani, 2017). Luego de estandarizar mediante puntaje z todas las variables, se utilizó como métrica de distancia la de Manhattan y el método de aglomeración jerárquico de Ward. La cantidad de clústeres fue determinada mediante la inspección visual del dendrograma resultante.

Así, se detectaron seis formas de expansión sojera a nivel departamental en Argentina:

- Tipo 1. Zonas en las que la soja se encuentra en expansión. Esta expansión se da tanto bajo la forma de sustitución como del doble cultivo sojero.

- Tipo 2. Zonas en las que la soja se encuentra en expansión y en las que la misma se da predominantemente mediante sustitución de otros cultivos.

- Tipo 3. Zonas cuyos cambios en el uso del suelo son relativamente estables y en las que la soja se da en buena medida predominantemente en forma de doble cultivo.

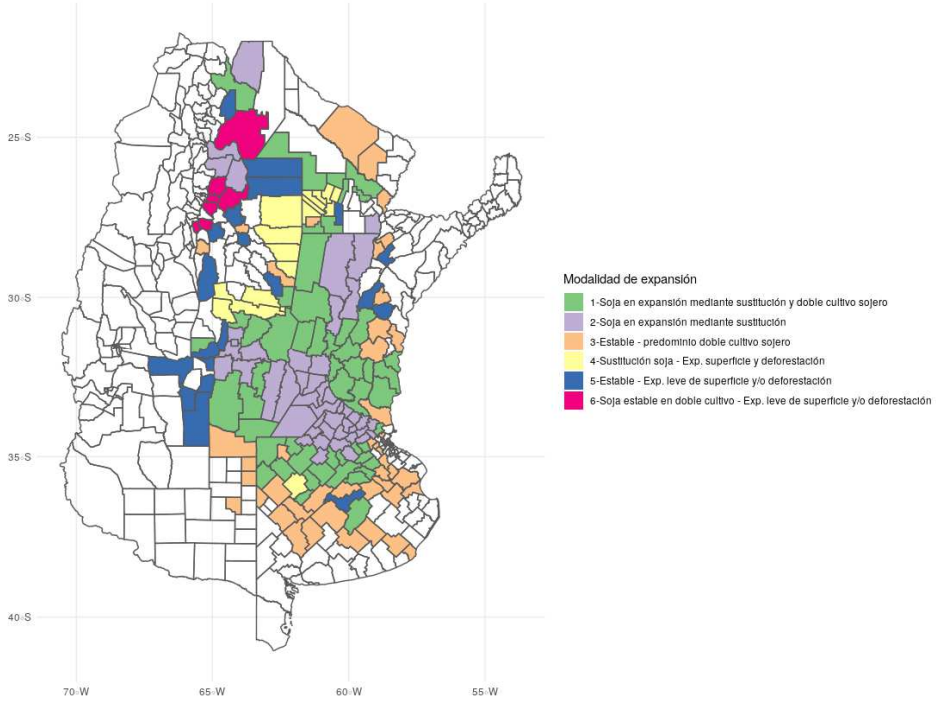
- Tipo 4. Zonas en las que el avance sojero se da mediante la sustitución de otros cultivos. A su vez, se encuentra articulado con expansiones fuertes del área sembrada y/o de avance de la deforestación.

- Tipo 5. Zonas cuyos cambios en el uso del suelo podemos caracterizar como estable. En algunos de estos departamentos se puede observar algún peso moderado de expansión de la superficie sembrada y/o de deforestación.

- Tipo 6. Zonas en las que la superficie con soja se encuentra relativamente estable, en las que predomina el doble cultivo y en las que se observa una leve expansión de la superficie total sembrada y cierto peso de los procesos de deforestación.

Al mismo tiempo fue posible mapear la distribución geográfica de estos tipos de expansión.

Mapa 1. Modalidades de expansión sojera, Argentina (zonas sojeras) 1988-2002



Fuente: elaboración propia basada en CNA y Monitor de Desmonte

Los resultados son coincidentes con lo que encuentran otros trabajos. Por ejemplo, Viglizzo y Jobbágy (2010) detectan frentes de expansión, retrocesos y "estacionarios" en zonas que tienden a coincidir con los diferentes clústeres identificados.

Estos resultados permiten mostrar los diferentes matices que la expansión sojera tuvo a nivel territorial. Si bien al analizar los números agregados se observa un crecimiento vertiginoso, las manifestaciones locales de este proceso tienen una

amplia gama de variabilidad: incluso parecen existir zonas en las que el resto de los cultivos mantuvieron su peso, articulándose con soja de segunda y con otros procesos de expansión.

Otro punto importante tiene que ver con la relación entre deforestación y sojización: lejos de constituir un proceso unívoco, en algunos de los departamentos en los que la soja se expandió a costa de otros cultivos podemos observar una articulación con procesos de deforestación; en otros, en cambio, esta combinación no es tan evidente.

Estos diferentes tipos identificados permiten matizar fuertemente la imagen de una expansión sojera homogénea a lo largo de las diferentes zonas del país. A su vez, este tipo de análisis permitió hacer observable algunos rasgos que los procesos de expansión de las fronteras de commodities tendrían. En efecto Kröger y Nygren (2020) sobre la base de algunos indicadores sumamente agregados postulaban la existencia de no linealidad espacial y temporal en las expansiones de frontera. Este rasgo de *patchwork* se hace visible a nivel departamental en la expansión sojera argentina y no sería visible mediante el estudio de estadísticas a nivel agregado (nacional o provincial) ni tampoco a partir del estudio en profundidad de algunos departamentos. Particularmente evidentes resultan estos mosaicos en las zonas de expansión de frontera propiamente dicha (el oeste de Chaco y Tucumán y este/norte de Santiago del Estero, por ejemplo), en donde se observa la articulación de cuatro formas de expansión sojera bien delimitables.

Segundo caso: generación de un mapa de vulnerabilidad sanitaria

Problema general

La segunda investigación[9] que se utilizará como ejemplo buscó construir información con alta resolución espacial sobre la distribución geográfica de la Vulnerabilidad sanitaria en Argentina. En términos generales, el concepto de "vulnerabilidad sanitaria" se relaciona con los llamados determinantes de la salud: existen ciertas variables que se encuentran fuertemente relacionadas con la salud de un individuo o de una población. Estos determinantes pueden ser de diferentes tipos (sociales, ambientales, individuales, biológicos, etc.) y se relacionan con el acceso desigual a los servicios de salud (tanto en términos cualitativos como cuantitativos). No todos los sectores de la población tienen las mismas oportunidades para acceder a un tratamiento médico general. En términos generales, aquellas poblaciones más vulnerables en términos socioeconómicos tienden a residir en zonas sumamente desventajosas en términos del acceso a los servicios de salud.

Un primer obstáculo que surge al momento de intentar cuantificar la vulnerabilidad surge de la necesidad de incorporar diferentes factores que pueden explicar este acceso (o falta de acceso) al sistema de salud. Los principales factores asociados tienen que ver con altos niveles de pobreza, pertenencia a minorías, presencia de enfermedades preexistentes, ausencia de cobertura de salud, etcétera. A su vez,

[9] Pueden encontrarse todos los detalles técnicos y metodológicos, así como los principales resultados, en Rosati, Olego y Vázquez Brust (2020).

existen otros factores que operan a nivel ambiental y no individual. Es por ello que se construyó un indicador que tratara de identificar zonas en las que se observaba un alto nivel de vulnerabilidad sanitaria. En este sentido, se buscó detectar "zonas calientes" caracterizadas por situaciones extremas de falta de acceso al sistema de salud.

Y en este punto se presentaba el segundo problema: la variable óptima para incluir en el índice hubiese sido la cobertura de salud de la población; no obstante, la información censal disponible solo presentaba este dato a nivel departamental[10]. Y, dado que el objetivo general era poder detectar zonas con déficit en el acceso a servicios de salud de forma lo más granular posible, se adoptó una definición reducida del concepto: un componente central del índice sería la distancia (a pie) que existe entre la residencia de una persona y el centro de salud más próximo. Esto permitiría producir información a nivel de radios censales. A su vez, se incorporó información proveniente del Censo de Población del año 2010 para incluir información correspondiente a las condiciones de vida en cada uno de los radios censales.

Fuentes de datos y métodos

Para la construcción del Índice de Vulnerabilidad Sanitaria (IVS) se utilizaron las siguientes fuentes de datos:

1. Microdatos del Censo Nacional de Población y Viviendas 2010

2. Cartografía de radios censales (2010)

3. Nombre, nivel y dirección geocodificada de establecimientos de salud de dependencia pública (hospitales, centros de salud y postas sanitarias)

4. Cartografía de calles

Como se mencionó previamente, la vulnerabilidad sanitaria puede ser subdividida en dos dimensiones: el acceso al sistema de salud y una serie de factores que podrían englobarse bajo el rótulo de "condiciones de vida". Es por ello que el IVS es el resultado de combinar otros dos índices. La construcción de la tabla de datos de establecimientos de salud implicó la necesidad de combinar y armonizar diferentes fuentes de datos:

1. Base nacional de Hospitales y Centros de Atención Primaria: la misma fue compilada por el Sistema de Información Sanitaria Argentina (SISA), obtenido a través del SEDRONAR en el sitio de IDERA (http://catalogo.idera.gob.ar).

2. Efectores de salud del programa SUMAR: se aplicaron técnicas de *web scraping* sobre el sitio para la obtención de los listados de efectores con la dirección de cada centro de salud.

[10] Esto se debe a que la pregunta del Censo sobre cobertura de salud se incluyó en el llamado "cuestionario ampliado", el cual fue aplicado a una muestra de la población que solo permite una desagregación a nivel departamental. En cambio, las variables utilizadas en la investigación fueron incluidas en el "cuestionario ampliado" que sí fue aplicado a la totalidad de la población, por lo cual se dispone de información desagregada a nivel de radio censal. Para mayores precisiones pueden consultarse los cuestionarios censales en el sitio del INDEC.

3. Listados de hospitales y centros de atención de salud del Programa Nacional de Salud Sexual y Procreación Responsable (Ministerio de Salud).

4. Otras fuentes a nivel provincial, generalmente, ministerios o secretarías de salud.

Las fuentes 2, 3 y 4 solamente tenían disponible las direcciones, por lo cual fue necesario geocodificarlas mediante la aplicación de Google Geocoding API. Luego de la integración de todos los datasets, el número de efectores de salud pública georeferenciados aumentó de los 4.419 (en el dataset de SISA) a 16.564 en el dataset final.

Accesibilidad a centros de salud

A partir de la información geolocalizada de establecimientos de salud y de la cartografía censal, el primer paso implicó el cálculo de las distancias a pie a los centros de salud más cercanos. Para cada uno de los 52.000 radios censales de Argentina se calculó una métrica de acceso: el tiempo que se tarda en llegar a pie desde el radio censal al centro de salud más cercano. Ahora bien, un primer problema se deriva de la forma irregular de los radios censales. En efecto, al tratarse de una subdivisión del territorio con fines logísticos (se trata de la mínima unidad territorial que dos censistas pueden relevar en el operativo censal) se observa un contraste muy fuerte en las formas y tamaños de los radios censales en áreas de alta y baja densidad de población. Esto implicaría que tomar los centroides de los polígonos de radios podría introducir algunas distorsiones.

Para solucionar este problema, las distancias a nivel radio censal fueron calculadas muestreando cinco puntos al interior de cada radio. El procedimiento fue el siguiente:

1. Se seleccionaron 5 puntos al azar en cada radio.

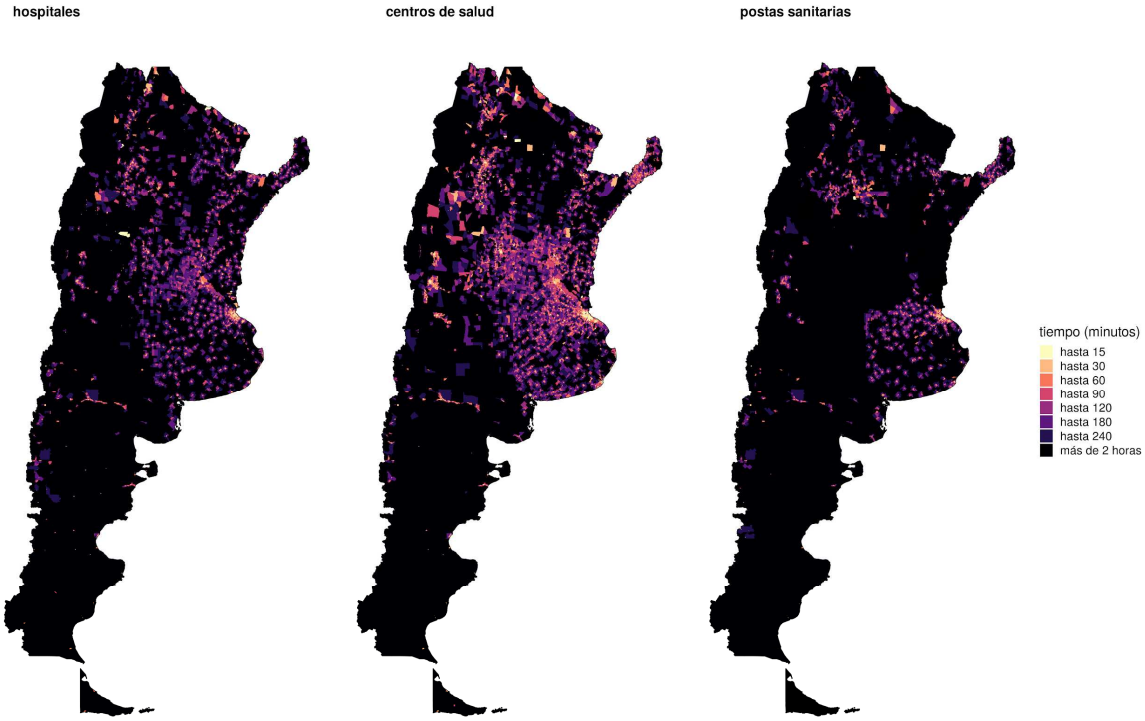
2. Para cada punto, se identificó (mediante un algoritmo de vecinos más cercanos) el establecimiento más cercano.

3. Para cada punto, se calculó la distancia a pie a través de la grilla de calles (es decir, no en línea recta) al establecimiento más cercano.

El procedimiento fue repetido para cada tipo de establecimiento de salud (hospitales, centros y postas). El índice final a nivel radio censal, se constituyó en la mediana de las 15 distancias calculadas (5 puntos para cada uno de los tres tipos de establecimiento de salud).

Para determinar la ruta óptima y el tiempo a pie que sería necesario recorrer, se utilizó el Open Source Routine Machine (OSRM), un sistema de ruteo que busca el camino más corto entre dos puntos a través de la red de calles.

Mapa 2. Tiempo de viaje a pie a centros de salud según tipo de establecimiento de salud (hospitales, centros y postas)



Fuente: elaboración propia basada en Rosati, Olego y Vázquez Brust (2020)

Puede notarse en el Mapa 2 la desigual distribución de la accesibilidad a los servicios de salud en dos dimensiones. Por un lado, en términos territoriales: se observa claramente cómo zonas con mayor densidad poblacional aparecen con mejores valores de accesibilidad. Por otro lado, los diferentes centros de salud presentan una distribución diferenciada: las postas sanitarias muestran distancias a pie notablemente más elevadas. Es más, se observa una relación inversa entre la distancia a hospitales/centros de salud y la distancia a postas sanitarias. De hecho, puede verse en el mapa cómo extensas áreas del centro del país presentan los valores más altos. Esta diferencia puede deberse a dos factores: 1) la alocación de postas sanitarias es una potestad de los gobiernos provinciales; y 2) deficiencias o subregistro en los datos de origen.

Índice de Condiciones de Vida / Nivel socioeconómico

El segundo componente del IVS está constituido por una métrica que intenta resumir diferentes características vinculadas a las condiciones de vida y al nivel socioeconómico de la población. Se utilizó información proveniente del Censo Nacional de Población y Vivienda 2010 y se calculó el Índice de Nivel Socioeconómico (INSE) a nivel de hogar. Se utilizaron las siguientes variables: condición de propiedad de la vivienda, calidad de los materiales de construcción, calidad de conexión a servicios básicos, calidad de construcción, hacinamiento, presencia de algún indicador NBI, nivel educativo del jefe de hogar, cantidad de

personas desocupadas en el hogar, presencia de servicio doméstico y condición de actividad. Estas variables remiten al acceso a ciertos servicios, a la calidad de la vivienda y a otros aspectos vinculados a la composición y situación ocupacional del hogar. Si bien estos bienes se encuentran asociados a trayectorias individuales y sociales, también mantienen una relación con el lugar de residencia.

El INSE fue construido mediante la combinación de estos 11 indicadores. El objetivo era reducir la información presente en los indicadores a un solo valor numérico que lograra captar la mayor correlación posible entre los mismos. Para realizar dicha combinación se utilizó una técnica de reducción de dimensionalidad llamada "autoencoder"[11].

En este sentido, el INSE busca resolver algunas de las limitaciones que tiene el índice clásicamente utilizado: Necesidades Básicas Insatisfechas (NBI). En efecto, el NBI solamente provee una medida dicotómica: cada hogar es clasificado como pobre o no, según presente valores positivos en alguno de los indicadores que forman parte del mismo. A su vez, cada uno de esos indicadores son variables binarias. Si alguno de estos valores es positivo, el hogar será clasificado como "pobre". El INSE[12], en contraste, fue diseñado como una medida de intensidad o "gradiente". Esto se observa en dos aspectos. Por un lado, se mantiene el carácter ordinal de los indicadores que lo componen, atributo que se pierde en el NBI. Por ejemplo, NBI considera el material de construcción de la vivienda de forma dicotómica (es deficitario o no). En cambio, el INSE rescata el hecho de que la variable fue codificada originalmente en cuatro estados. Por otro lado, el índice resultado del INSE es una variable cuantitativa que varía entre -1 y +1. Este enfoque permite captar con mayores matices los extremos de la distribución: los hogares en peores condiciones y los hogares en mejores condiciones.

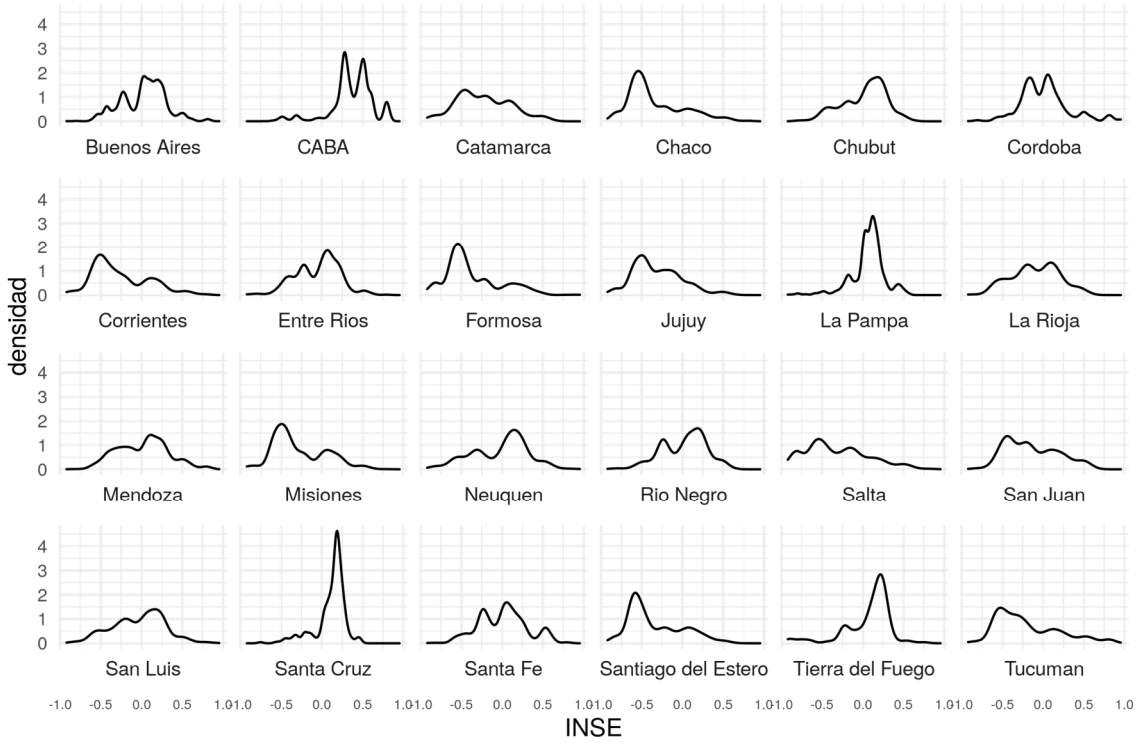
Dado que el INSE fue estimado a nivel hogar, para obtener el mapa final, los valores de los hogares de cada radio censal fueron agregados a través de una métrica conocida como la "Trimedia de Tukey", la cual consiste simplemente en el promedio ponderado de los cuartiles de la distribución.

En el Gráfico 1 (en el eje X se ubican los valores del índice y en el Y los valores de densidad) puede verse que la distribución del INSE es consistente a nivel provincial: se encuentra sesgado a la izquierda (es decir, hacia los valores más bajos) en provincias clásicamente caracterizadas por altos niveles de pobreza y carencias diversas como Chaco, Santiago del Estero y Catamarca. A su vez, se encuentra sesgada hacia los valores más altos en jurisdicciones como CABA y varias provincias patagónicas.

[11] En líneas generales, un "autoencoder" tiene como objetivo encontrar una representación de los datos de input generalmente con el objetivo de realizar una reducción de la dimensionalidad. En general, funcionan simplemente aprendiendo a replicar los inputs en los outputs. Suelen ser utilizados para la compresión de archivos de audio, imágenes o video. Puede consultarse al respecto en Geron (2017) y Goodfellow, Bengio y Courville (2016).

[12] Es importante notar que el INSE no se presenta como una métrica de "pobreza" o de "carencia" como el NBI, sino más bien de "condiciones de vida" o "nivel socioeconómico". En este sentido, excede la noción de "pobreza" (más precisamente de peores condiciones de vida) en tanto la misma se ubicaría en cierta región del INSE (sus valores más bajos).

Gráfico 1. Distribución del Índice de Nivel Socioeconómico según provincia

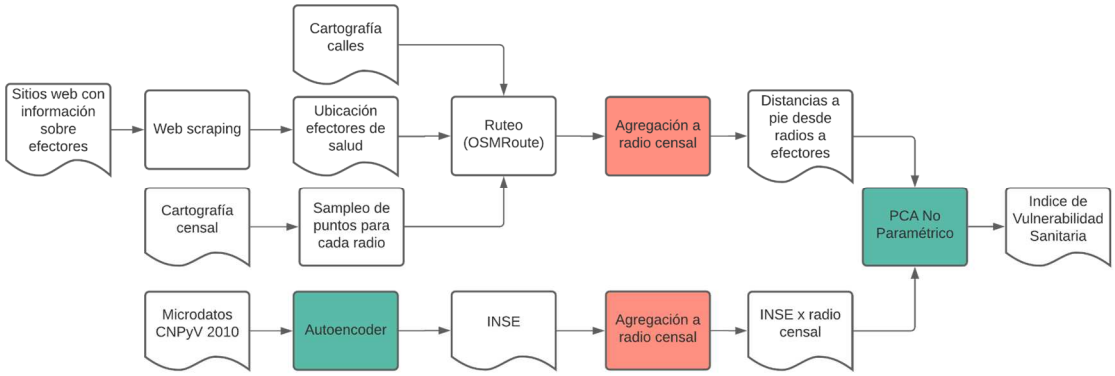


Fuente: elaboración propia basada en Rosati, Olego y Vázquez Brust (2020)

Índice de Vulnerabilidad Sanitaria

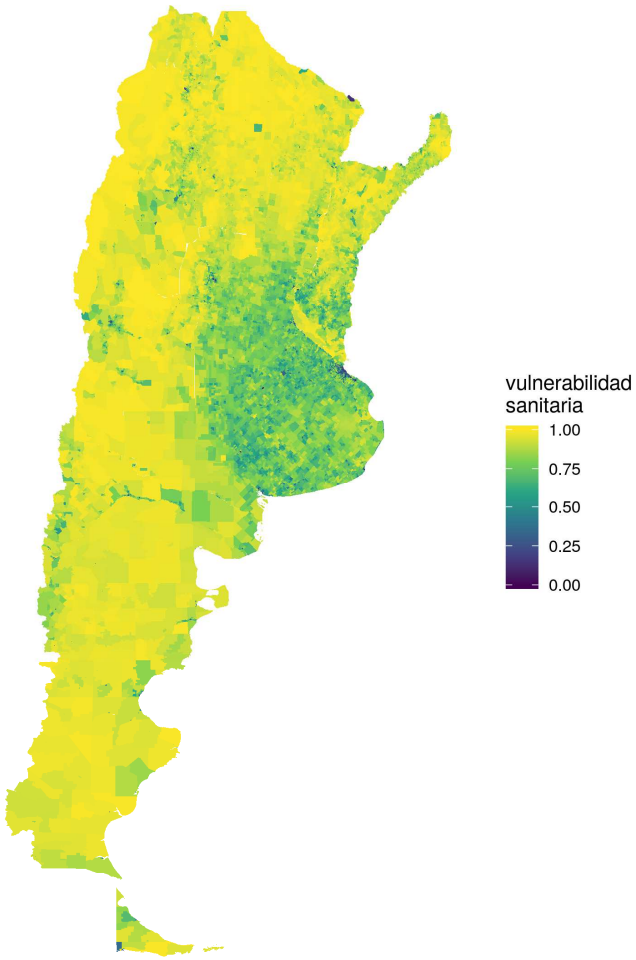
El último paso para construir el Índice de Vulnerabilidad Sanitaria final consistió en combinar ambas variables a nivel de radio censal. Para ello, luego de transformar de forma adecuada los valores de cada subíndice, los mismos fueron combinados mediante otra técnica de reducción de la dimensionalidad llamada "Análisis de Componentes Principales No Paramétrico" (Jolliffe, 2006). Se resume el flujo completo de trabajo en la siguiente figura:

Figura 2. Flujo de trabajo completo para la construcción del IVS



El resultado a nivel radio censal puede verse en el Mapa 3:

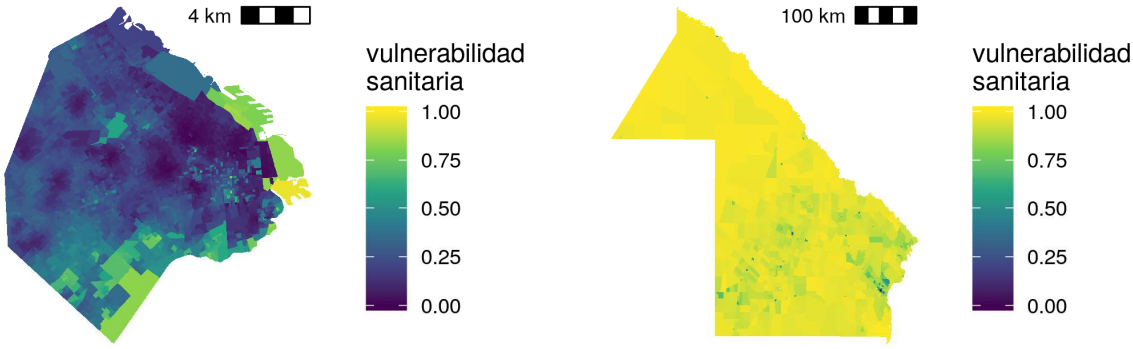
Mapa 3. Distribución del Índice de Vulnerabilidad Sanitaria a nivel radio censal. Argentina, 2010-2018



Fuente: elaboración propia basada en Rosati, Olego y Vázquez Brust (2020)

Allí se observa la pauta esperada: las zonas de mayor densidad demográfica presentan menores índices de vulnerabilidad sanitaria. Se observan dos zonas claramente delimitadas: una zona con bajos valores del índice (Buenos Aires, la Ciudad Autónoma de Buenos Aires y los grandes aglomerados urbanos); otra zona con valores altamente deficitarios, que agrupa al resto del país y a las zonas menos pobladas. No obstante, esta escala de agregación debe tomarse con recaudos: si se observa con mayor detalle una mirada más focalizada en algunas regiones (Mapa 4), se advierten patrones que no son detectables en el Mapa 3.

Mapa 4. Distribución del Índice de Vulnerabilidad Sanitaria a nivel radio censal. CABA y Chaco, 2010-2018



Fuente: elaboración propia basada en Rosati, Olego y Vázquez Brust (2020)

Aparecen en CABA zonas en situaciones críticas del índice (zona Sur en barrios como Lugano, por ejemplo). En Chaco, (una provincia caracterizada por una mala situación sanitaria general) se observan zonas con valores bajos de vulnerabilidad (la zona alrededor de Resistencia o Sáenz Peña).

Aprendizajes de la integración de fuentes diversas: escalas y modelado de datos

Por supuesto que el uso de estas fuentes no está exento de problemas: la periodicidad de los relevamientos censales dista de ser óptima para ciertos objetivos de investigación, las definiciones conceptuales no siempre se adaptan a los problemas investigados, etcétera. A su vez, tanto en las fuentes clásicas de información como en aquellas vinculadas a las nuevas tecnologías, pueden existir serios problemas de cobertura y sesgos de selección. Estas limitaciones son más conocidas y han sido mencionadas frecuentemente. No obstante, no siempre se ha hecho el mismo énfasis en las potencialidades que las mismas tienen (utilizadas de forma correcta) para las ciencias sociales y los estudios territoriales específicamente.

La reseña de ambas investigaciones pretende funcionar como un ejemplo de algunas de esas potencialidades mediante la integración de fuentes secundarias. Mientras que el análisis de las formas de expansión sojera se basó en información que podríamos denominar "tradicional" (derivada de imágenes satelitales, fuentes y cartografías censales), la construcción del IVS supuso la combinación de fuentes censales, junto con fuentes y técnicas de recolección y construcción de información más "modernas" (técnicas de *web scraping*, utilización de APIS de geocodificación y ruteo, etcétera).

Es importante marcar que existen algunas operaciones de homogeneización que son análogas en ambos casos: en los ejemplos anteriores, la operación más importante fue la definición de una escala de observación o resolución espacial que permitiera

hacer comparables los diferentes registros. En el caso de la expansión sojera, la escala fue departamental. En cambio, la resolución espacial cambió para el caso del IVS: se procesó y construyó información a nivel de los radios censales. En los flujos de trabajo correspondientes a ambos proyectos (Figuras 1 y 2) se han destacado en rojo las etapas en las que se efectúan operaciones de agregación o armonización de escalas en los datos.

Otro aspecto importante es el rol que adquieren ciertas técnicas de modelado de datos en el proceso de integración de fuentes. Muchas veces, el resultado de estas operaciones de integración suele ser una tabla de alta dimensionalidad, es decir, una gran cantidad de columnas. En este sentido, la etapa analítica requiere (casi ineludiblemente) de algún proceso que sintetice dicha información, la cual en muchos casos (como en los ejemplos anteriores) suele ser altamente descriptiva y de la que no se dispone de modelos conceptuales o mecanismos causales claros.

Las llamadas técnicas de aprendizaje no supervisado son sumamente útiles en estos casos: se caracterizan por no disponer de una variable dependiente la cual se intenta modelar. A diferencia, por ejemplo, de una regresión en el que una variable de respuesta (Y) se modela como una combinación de un conjunto de predictores (X), posiblemente lineal ($Y_i = \beta_0 + \beta_1 X_1 + + \beta_p X_p$), en las técnicas no supervisadas se busca detectar patrones en un conjunto de variables independientes sin una variable dependiente. Dos suelen ser las tareas más comunes de este tipo de herramientas: clusterización y reducción de dimensionalidad[13].

En el primer caso, se busca agrupar de forma semiautomática un conjunto de datos en una cantidad finita de clases (no existentes de antemano). Un clúster satisface dos criterios básicos: 1) las observaciones son similares al interior de cada grupo y 2) las unidades son heterogéneas entre los diferentes grupos. Existe una gran cantidad de métodos de clustering (k-medias, jerárquicos, DBSCAN, etcétera), pero la gran mayoría coincide en el cálculo de una matriz de distancias entre los casos y en alguna forma de agrupamiento de los datos en función de la misma. Dicho en términos sociológicos: se busca automatizar la construcción de tipologías, es decir, la inferencia de clases a partir de la combinación de los diversos atributos X.

Este es el uso que se le dio en el trabajo sobre la expansión sojera. Cada una de las variables e indicadores utilizadas buscaron operacionalizar las diferentes formas identificada por la literatura en que la misma se producía a nivel departamental: las estimaciones de los modelos de inferencia ecológica buscaban captar los cambios y transiciones entre los diferentes cultivos (particularmente, los de sustitución); la proporción de soja de segunda trató de cuantificar la existencia de procesos de adición de producción; la variación en la superficie sembrada constituyó una aproximación a la incorporación de nuevas tierras cultivables; y la proporción de hectáreas deforestadas era un intento de ponderar el peso de los procesos de avance sobre bosques nativos. El objetivo era, entonces, llegar a una tipología de departamentos tratando de observar de qué formas se articulan territorialmente estos procesos.

[13] Una explicación más detallada al respecto puede encontrarse en James, Witten, Hastie y Tibshirani (2017) y en Sosa Escudero (2018a).

El segundo tipo de tareas no supervisadas mencionadas busca resumir la información contenida en las columnas de un conjunto de datos. Seguramente, el Análisis de Componentes Principales (PCA, por sus siglas en inglés) sea el más conocido: a partir de una matriz X de dimensión nxp se busca representar la información contenida en X en una matriz reducida de dimensión $nxp*$ donde $p* < p$ mediante la combinación lineal de los diferentes predictores. Variando la forma de combinar las X se obtienen diferentes métodos de reducción de la dimensionalidad. Una utilidad (no la única) de este tipo de técnicas radica en la posibilidad de generar índices "multidimensionales" compuestos por diversas combinaciones de variables de un dataset.

Así, el IVS hace un uso intensivo de este tipo de técnicas en dos instancias. En primer lugar, mediante la construcción del índice de condiciones de vida. En efecto, en este caso, se asume que las condiciones en que se desarrolla la vida de la población (el "índice de nivel socioeconómico") es una variable latente que surge de alguna forma de combinación entre los indicadores que se mencionaron más arriba. De esta forma, se utiliza un autoencoder como una forma de pasar de unos 12 indicadores (que se traducen en aproximadamente unas 40 columnas en la tabla de datos al realizar la codificación de cada una de ellas) a una sola variable que logra captar una proporción sustancial de la información contenida en ellas. Luego, se utilizó otra técnica de reducción (Análisis de Componentes Principales no paramétrico) para construir el índice final de vulnerabilidad sanitaria, combinando el índice de nivel socioeconómico y la distancia a pie a centros de salud[14]. Como pueden verse en las Figuras 1 y 2, se nota que las etapas de modelado de datos (marcadas en verde) ocupan un rol central en diferentes etapas de cada uno de los procesos.

En ambos casos, y dada la homogeneización e integración previa de las fuentes de información, las técnicas de clustering y de reducción de dimensionalidad cumplen un rol relevante en la etapa analítica y de síntesis de la información. Las mismas permiten resumir un conjunto relativamente grande de variables, ya sea en términos de una tipología de carácter discreto, y generalmente cualitativo, o en términos de la identificación de una variable latente cuantitativa y que expresa un gradiente continuo de situaciones. A su vez, la integración de fuentes diversas de un alto nivel de desagregación permitió en el caso de las modalidades de expansión sojera caracterizar los procesos al interior de una unidad territorial determinada (un departamento). De esta forma, puede potenciarse la utilidad de estudios de campo de mayor grado de focalización territorial, en tanto permite la inclusión del caso de estudio en diversas escalas. Pero también permite contar con información relativamente granular para decidir futuros estudios de casos. Así, al analizar la información del mapa sobre las formas de la expansión sojera, podrían seleccionarse zonas y departamentos que sean contrastantes en esta variable. De esta forma, podría decidirse seleccionar un departamento "típico" de cada modalidad de expansión:

[14] En Rosati y Chazarreta (2020) puede encontrarse un ejemplo de articulación de fuentes censales de población y agropecuarias para la identificación de diferentes estructuras agrarias para el total del país. Se procesó información a nivel departamento y se utilizó PCA para reducir la dimensionalidad de la tabla de datos y, luego, un procedimiento de clustering para generar la tipología.

- Almirante Brown (Chaco), representa muy bien el tipo 1 "soja en expansión mediante sustitución y doble cultivo sojero"

- Vera (Santa Fe) sería un caso del tipo 2 "soja en expansión mediante sustitución"

- Aguirre (Santiago del Estero) ilustra el tipo 3 "estable - predominio doble cultivo sojero"

- Independencia (Chaco) es un buen ejemplo del tipo 4 "sustitución soja - exp. superficie y deforestación"

- Alberdi (Santiago del Estero) es un espécimen del tipo 5 "estable - exp. leve de superficie y/o deforestación"

- Anta (Salta) es un ejemplo del tipo 6 "soja estable en doble cultivo - exp. leve de superficie y/o deforestación"

Se seleccionaron departamentos en las zonas del NEA y NOA, pero también podrían seleccionarse casos haciendo variar de forma controlada ciertos aspectos agroecológicos y haciendo jugar diferentes regiones.

Algo similar ocurre en el caso del IVS: la alta resolución espacial del mapa permite detectar tanto situaciones y regiones localizadas de alta vulnerabilidad sanitaria como poder generar métricas agregadas para caracterizar unidades mayores (departamentos, provincias) o incluso generar regionalizaciones ad-hoc basadas en la distribución a nivel radio censal del índice.

Comentarios finales

Este trabajo intentó plantear algunas reflexiones sobre el uso y la integración de fuentes heterogéneas en las ciencias sociales en general y en la problemática del territorio en particular. Se analizaron dos investigaciones en dos contextos diferenciados: un estudio sobre la expansión sojera entre 1988 y 2002 y otro centrado en el desarrollo de índice de vulnerabilidad sanitaria a nivel radio censal para todo el país.

Ambas permitieron identificar algunas de las características que presenta este tipo de integración. La granularidad territorial de la información, el problema de la definición de una escala de análisis común, la necesidad de avanzar en la utilización de cierto tipo de modelado que ayude a esta integración de fuentes y la posibilidad de replicabilidad y apertura de dicha información son algunos de los temas centrales del artículo.

Este tipo de fuentes de información hace posible, entonces, establecer relaciones conceptuales y empíricas tanto a nivel de diferentes unidades a un mismo nivel de agregación (radios, departamentos, etc.) como aportar información sobre una escala mayor.

Por último, cabe mencionar que los datos que funcionaron como base para la construcción de las diversas métricas aquí presentadas son, en todos los casos, fuentes públicas y accesibles (datos censales, información de Open Street Map,

Monitoreo de Deforestación). También la información procesada se encuentra disponible para ser consultada:

- los mapas de las modalidades de expansión sojera a nivel departamento se encuentran en el siguiente link: https://gefero.github.io/CONICET_mapas_soja/

- la información a nivel radio censal sobre accesibilidad a centros de salud, índice de nivel socioeconómico e índice de vulnerabilidad sanitaria pueden descargarse del sitio https://poblaciones.org/

Este constituye un punto importante en términos de las buenas prácticas de la llamada "ciencia abierta" (Open Science) particularmente, aquellas que se centran en la búsqueda de la replicabilidad tanto de los datos crudos como de los principales resultados de la investigación[15].

Bibliografía

ACKHAR, Marcel; DOMÍNGUEZ, Ana; DÍAZ, Ismael y PESCE, Fernando (2011) "La intensificación del uso agrícola del suelo en el litoral oeste del Uruguay en la última década". *Revista Pampa*, No. 7, pp. 143-157.

AIZEN, Marcelo; GARIBALDI, Lucas y DONDO, Mariana (2009) "Expansión de la soja y diversidad de la agricultura argentina". *Ecología Austral*, Vol. 19, No. 1, pp. 45-54.

AZCUY AMEGHINO, Eduardo y LEÓN, Carlos A. (2005) "La sojización: contradicciones, intereses y debates". *Revista Interdisciplinaria de Estudios Agrarios*, No. 23, pp. 133-157.

AZCUY AMEGHINO, Eduardo y ORTEGA, Lucía (2010) "Sojización y expansión de la frontera agropecuaria en el NEA y NOA: transformaciones, problemas y debates" *Documentos del PIEA*, Vol. 5, pp. 141-159.

DUNCAN, Otis y DAVIS, Beverly (1953) "An Alternative to Ecological Correlation" *American Sociological Review*, Vol. 18, No. 6, pp. 665–666.

GALTUNG, Johann (1966) *Teoría y método de la investigación social*. Buenos Aires: EUDEBA.

GOODFELLOW Ian; BENGIO Yoshua y COURVILLE Aaron (2016) *Deep Learning*. Massachusetts: MIT Press.

[15] La OECD define a la Ciencia Abierta como el hecho de "hacer los productos primarios de la investigación con fondos públicos -publicaciones y datos de investigación- de acceso público en un formato digital con ninguna o mínimas restricciones" (OECD, 2015:7). Sin embargo, esta definición es parcial, en tanto la ciencia abierta se propone extender los principios de "apertura" (*openness*) a todo el proceso de investigación, fomentando la colaboración entre colegas y disciplinas e incluyendo la posibilidad de replicar todos los resultados (intermedios y finales) de la investigación. Tales objetivos no se basan solamente en un imperativo moral, sino que la escasez de investigación reproducible y replicable atenta contra los mismos resultados de la actividad científica. Fenómenos como el llamado p-hacking o p-harking (una mala estimación e interpretación de las pruebas de hipótesis conduciendo a resultados falsos o poco fiables) tienen como una de sus causas la dificultad en el control colaborativo de los resultados por parte de la comunidad científica. Un diagnóstico y una propuesta de solución a esos problemas puede encontrarse en Munafò, Nosek,Bishop et al. (2017).

GERÓN, Aurélien (2017) *Hands-on machine learning with Scikit-Learn and TensorFlow. Concepts, tools, and techniques to build intelligent systems*. New York: O'Reilly.

GOODMAN, Leo (1959) "Some Alternatives to Ecological Correlation". *American Journal of Sociology*, Vol. 64, No. 6, pp. 610-624.

HILBERT, Martin y LÓPEZ, Priscilla (2011) "The World's Technological Capacity to Store, Communicate, and Compute Information". *Science*, Vol. 332, pp. 60-65.

JOLLIFFE, Ian (2002) *Principal Component Analysis*. New York: Springer.

JAMES, Gareth; WITTEN, Danielle; HASTIE, Trevor y TIBSHIRANI, Robert (2017) *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

KING, Gary (1997) *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

KRÖGER, Markus y NYGREN, Anya (2020) "Shifting frontier dynamics in Latin America". *Journal of Agrarian Change*, Vol. 20, No. 3, pp. 364-386.

LART/FAUBA (2004) *Patrones espaciales y temporales de la expansión de soja en Argentina. Relación con factores socio-económicos y ambientales*. Buenos Aires: Laboratorio de Análisis Regional y Teledetección (LART)/FAUBA.

LEÓN, Carlos; PRUDKIN, Nora y REBORATTI, Carlos (1985) "El conflicto entre producción, sociedad y medio ambiente: la expansión agrícola en el sur de Salta". *Desarrollo Económico*, Vol. 25, No. 99, pp. 399-420.

MUNAFÒ, Marcus; NOSEK, Brian; BISHOP, Dorothy; BUTTON, Katherine; CHAMBERS, Christopher; PERCIE DU SERT, Nathalie; SIMONSOHN, Uri; WAGENMAKERS, Eric; WARE, Jennifer y IOANNIDIS , John (2017) "A manifesto for reproducible science". *Nature Human Behavior*, Vol. 1, No. 0021. https://doi.org/10.1038/s41562-016-0021

OECD (2015) "Making Open Science a Reality". *OECD Science, Technology and Industry Policy Papers*, No. 25.

PÁEZ, Sergio (2016) "Soja en Argentina a principios del siglo XXI: el sistema agropecuario y la competencia por el uso del suelo productivo". *Cuadernos De Economía Crítica*, Vol. 3, No. 5, pp. 135-169.

RODRÍGUEZ, Javier (2008) *Consecuencias económicas de la soja transgénica. Argentina 1996-2006*. Buenos Aires: CLACSO.

ROSATI, Germán y CHAZARRETA, Adriana (2020) "Tipos de estructuras sociales agrarias en la formación social argentina. Un análisis a nivel departamental: 2001-2002". *Mundo Agrario*, Vol. 21, No. 48. https://doi.org/10.24215/15155994e153

ROSATI, Germán; OLEGO, Tomás y VAZQUEZ BRUST, Antonio (2020) "Building a sanitary vulnerability map from open source data in Argentina (2010-2018)". *International Journal of Equity in Health*, Vol. 19, No. 204. https://doi.org/10.1186/s12939-020-01292-3

SALGANIK, Matthew (2018) *Bit by bit. Social research in the digital age*. Oxford: Princeton University Press.

SOSA ESCUDERO, Walter (2018a) "Big data y aprendizaje automático: ideas y desafíos para economistas" En: Ahumada, Hidegart; Gabrielli, María; Herrera, Marcos y Sosa Escudero, Walter, *Una nueva econometría. Automatización, big data, econometría espacial y estructural*. Bahía Blanca: Asociación Argentina de Economía Política, pp. 157-201.

SOSA ESCUDERO, Walter (2019) *Big Data. Breve manual para conocer la ciencia de datos que ya invadió nuestros días*. Buenos Aires: Siglo XXI Editores.

STEIMBREGER, Norma; RADONICH, Marta y BENDINI, Mónica (2003) "Expansiones de frontera agrícola y transformaciones territoriales: procesos sociales diferenciales" En: Bendini, Mónica y Steimbreger, Norma (coords.) *Territorios y organización social de la agricultura*. Buenos Aires: La Colmena, pp. 17-39.

VALENZUELA, Cristina (2014) "Implicancias del avance de la 'frontera' agropecuaria en el Nordeste Argentino en las últimas dos décadas". *Estudios Socioterritoriales*, Vol. 2, No. 16, pp. 95-109.

VIGLIZZO, Ernesto y JOBBÁGY, Esteban (eds.) (2010) *Expansión de la Frontera Agropecuaria en Argentina y su Impacto Ecológico-Ambiental*. Buenos Aires: Ediciones INTA.

VOLANTE, José; MOSCIARO, Jesús; MORALES POCLAVA, María; VALE, Laura; CASTRILLO, Silvana; SAWCHIK, Jorge; TISCORNIA, Guadalupe; FUENTE, Marcel; MALDONADO IBARRA, Isaac; VEGA TRUJILLO, Richard; CORTÉZ, L.; PARUELO, José (2015) "Expansión agrícola en Argentina, Bolivia, Paraguay, Uruguay y Chile entre 2000-2010: caracterización espacial mediante series temporales de índices de vegetación". *Revista de Investigaciones Agropecuarias*, Vol. 41, No. 2, pp. 179-191.

TRIGO, Eduardo; CHUDNOVSKY, Daniel; CAP, Eugenio y LÓPEZ, Andrés (2002) *Los transgénicos en la agricultura argentina: una historia con final abierto*. Buenos Aires: Libros del Zorzal.

VALLEJOS, María; VOLANTE, José; MOSCIARO, Jesús; VALE, Laura; BUSTAMANTE, Laura y PARUELO, José (2015) "Transformation dynamics of the natural cover in the Dry Chaco ecoregion: A plot level geo-database from 1976 to 2012". *Journal of Arid Environment*, Vol. 123, pp. 3-11.