Índice

Viviendas repetidas en el censo de 2010 de la Argentina: una exploración empírica[1]

Pablo De Grande[2]

Recibido: 25/07/2023
Aceptado: 30/08/2023

Resumen

El artículo muestra los resultados de un ejercicio de análisis tendiente a cuantificar y describir grupos de viviendas duplicadas en el Censo Nacional de Población, Hogares y Viviendas del año 2010 de la Argentina. Este análisis se realizó a partir de las bases REDATAM de datos de nivel de radio para el cuestionario básico. Se tomó como premisa que los atributos simples de grupos de viviendas sucesivas (con sus hogares y personas) deben permitir detectar la repetición artificial de datos, fuera de toda duda razonable de si se trata o no de viviendas genuinamente idénticas en el referente empírico. Como resultado, se confirma la existencia del fenómeno señalado (viviendas duplicadas) en las bases del cuestionario básico y se describe la distribución por jurisdicción de los casos duplicados que se han detectado, así como varias características que pueden inferirse respecto del proceso por el que habrían sido instrumentados.

Palabras clave: vivienda, censos de vivienda, censos de población, estadísticas de vivienda, calidad de los datos, evaluación, metodología estadística, Argentina.

[1] Esta investigación fue realizada con el apoyo del Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET). El autor agradece a Gonzalo Rodríguez y a Nicolás Sacco por los comentarios e intercambios de ideas durante esta investigación.

[2] Doctor en Ciencias Sociales y Humanidades por la Universidad de Quilmes. Licenciado en Sociología por la Universidad de Buenos Aires. Investigador en el Instituto de Estudios Histórico-Sociales (IEHS-IGEHCS) de la Universidad del Centro de la Provincia de Buenos Aires. Investigador del Consejo Nacional de Investigaciones Científicas y Técnicas de la Argentina (CONICET). Correo electrónico: pablodg@gmail.com.

Abstract

The article outlines the results of an analysis to quantify and describe groups of duplicated dwellings in the 2010 National Population, Household and Housing Census of Argentina. This analysis was performed using the REDATAM databases, at the level of the *radio* census geographic unit of the data, for the basic questionnaire. It was hypothesized that the simple attributes of successive groups of dwellings (including the households and persons) should enable artificial repetition of data to be detected, with no reasonable doubt as to whether they are truly identical dwellings according to the empirical referent. The findings confirmed the existence of duplicate dwellings in the basic questionnaire databases. The article then describes the distribution of detected duplicates by administrative areas, as well as some characteristics that can be inferred regarding the process that resulted in their duplication.

Keywords: housing, housing censuses, population censuses, housing statistics, data quality, evaluation, statistical methodology, Argentina.

Résumé

L'article présente les résultats d'un exercice d'analyse visant à quantifier et à décrire les groupes de logements dupliqués dans le recensement national de la population, des foyers et des logements de 2010 en Argentine. Cette analyse a été effectuée à partir des bases de données du niveau ratio de REDATAM pour le questionnaire de base. La prémisse était que des attributs simples de groupes de logements successifs (avec leurs ménages et leurs personnes) devraient permettre de détecter une répétition artificielle des données, au-delà de tout doute raisonnable quant à la question de savoir s'il s'agit ou non de logements véritablement identiques dans le cadre de l'étude de référence empirique. Le résultat confirme l'existence du phénomène indiqué (logements en double) dans les bases du questionnaire de base et décrit la distribution par juridiction des cas de duplication qui ont été détectés, ainsi que plusieurs caractéristiques qui peuvent être inférées concernant le processus par lequel ces cas auraient été créés.

Mots clés : logement, recensements du logement, recensements de la population, statistiques du logement, qualité des données, évaluation, méthodologie statistique, Argentine.

Introducción[3]

El Censo Nacional de Población, Hogares y Viviendas del año 2010 de la Argentina ha enfrentado una diversidad de obstáculos y controversias. Entre los señalamientos que han sido más recurrentes y problemáticos se encuentra la afirmación de que el censo tendría viviendas duplicadas. Ello supondría que, sin indicarse en la documentación técnica precisiones respecto a la duplicación o ponderación de casos en las bases de microdatos, algunas de las personas u hogares que se reflejan en dichas bases serían copias de otras, y no respuestas emergentes del relevamiento de campo.

De ser esto así, cabe preguntarse: ¿qué características tendrían esas personas u hogares? ¿Cuál sería el alcance geográfico del fenómeno? ¿Cuántos serían en total?

Con la intención de indagar sobre estas cuestiones, en este artículo se realiza un análisis sistemático sobre la base de datos correspondiente al cuestionario básico del Censo Nacional de Poblaciones, Hogares y Viviendas de 2010. Si bien existen antecedentes de propuestas orientadas a detectar casos repetidos en bases de datos censales (Marshall, 2008; Abbott y Large, 2009), partían del supuesto de que cada caso fuese bien conocido (por ejemplo, por su nombre o apellido) y que la oficina estadística no los hubiese duplicado voluntariamente, sino que se trataría de personas que por diferentes razones se habían censado más de una vez. En el análisis que aquí se presenta, en cambio, se debía definir y aplicar una metodología que permitiera aproximarse a estimar las viviendas que podían considerarse réplicas de otras, incluso sin disponer de la información típicamente identificatoria (como el nombre o el domicilio de las personas).

Para lograr esto, en el procedimiento se tomaron como grupos de interés todas las secuencias de cinco viviendas o más que tuvieran iguales valores en todos los indicadores seleccionados (a nivel de vivienda, hogar e individual). La idea de usar viviendas subsecuentes tuvo por objetivo reducir el número de falsos positivos que podían presentarse si se tomaban viviendas en forma individual. Ello se debe a que dos viviendas pueden estar representadas por valores idénticos en el registro censal por tratarse simplemente de casos similares (sin ser réplicas). Sin embargo, esta situación es menos probable si se hallan dos viviendas sucesivas e idénticas en diferentes lugares del país, aún menos si son tres o más. El punto de corte para determinar qué se puede considerar una réplica de baja probabilidad de ocurrencia aleatoria (cinco hogares) se estableció utilizando el censo de 2001 como grupo de control, es decir, seleccionando un nivel que mostrara en dicho censo una escasez de casos coincidentes para ese volumen de atributos idénticos subsecuentes.

[3] Los datos producidos en relación con este artículo pueden consultarse y descargarse desde la cartografía publicada en el sitio Poblaciones (véase [en línea] https://mapa.poblaciones.org/map/87101). El código fuente de la aplicación elaborada para el cálculo de los identificadores, pares y grupos de viviendas se encuentra disponible mediante la licencia GNU-GPL3 en el repositorio GitHub (véase [en línea] https://github.com/discontinuos/duplicates-finder). Allí también se encuentra la aplicación en forma ejecutable para Windows, que ofrece una interfaz visual para la selección de los datos a procesar y la ejecución de las comparaciones y cálculos.

A partir de dicho análisis, se realiza una primera caracterización de viviendas y personas que habrían sido representadas más de una vez en los microdatos del censo de población argentino de 2010. Con este fin se aplica un procedimiento definido en forma *ad hoc* para el análisis. El mismo procedimiento puede ser de utilidad para la validación de otros resultados censales, por lo que en el apartado metodológico se detalla de manera exhaustiva.

Para avanzar hacia el objetivo general, en la siguiente sección se da cuenta de dos antecedentes que abonaron la hipótesis de la presencia de casos duplicados en el mencionado censo. Luego se presenta la estrategia metodológica con que se trabajó para evaluar la problemática descrita. A continuación, se presentan los resultados de la aplicación de dicha estrategia. Finalmente, en la sección de conclusiones, se retoman las preguntas iniciales.

A. Antecedentes

Este artículo se inspira en una presentación realizada en 2017 por Florencia Molinatti en las XIV Jornadas Argentinas de Estudios de Población, titulada "Las migraciones internas en Argentina: posibilidades, alcances y desafíos para su captación mediante el Censo de 2010" (Molinatti, 2017). En ella, la autora mostraba resultados de investigación que daban cuenta de datos anómalos en los registros de la base de datos ampliada del censo Nacional de Población, Hogares y Viviendas de 2010. En concreto, relataba las dificultades que había enfrentado al utilizar parte de esa información para analizar el comportamiento migratorio de los habitantes. Había encontrado indicios de que muchas de las personas que se encontraban en pequeñas localidades de la provincia de Córdoba no habrían sido censadas en ellas. Se trataba, según la autora, de personas relocalizadas, en forma *ex post*, de otros centros urbanos a esas zonas censales.

Debido a la incongruencia en las variables de residencia anterior, Molinatti sugería que ciertos casos habían sido imputados a partir de casos de otras localidades y que se había omitido la rectificación de ciertas variables, como la de residencia anterior. En estas circunstancias, la población completa de ciertas localidades declaraba no haber vivido allí cinco años antes, cuando el conocimiento del terreno permitía afirmar que eso no era cierto.

A partir de estas evidencias, se abrían muchas preguntas sobre lo que había sido anunciado desde el Instituto Nacional de Estadística y Censos (INDEC) como el mejor censo de la historia argentina (Infobae, 2017) ¿Eran estos casos irregularidades aisladas en la información censal? ¿Se circunscribían estos fenómenos al cuestionario ampliado, que investigó Molinatti, o la muestra básica tenía problemas similares? La muestra básica carecía de las preguntas de residencia anterior, por lo que el descubrimiento de Molinatti no podía aplicarse en ella.

Los resultados e informes censales, publicados entre 2012 y 2014, no fueron acompañados por estimaciones de cobertura (como sí sucedió en el censo de 2001),

ni se precisaron detalles sobre los procedimientos aplicados en datos faltantes o en la relocalización de casos. No había explicaciones metodológicas que justificaran la existencia de casos repetidos o desplazados.

En un comunicado del Instituto de Estadísticas y Censos (INDEC) con fecha de julio de 2016 —un año antes de la presentación de Molinatti— las entonces nuevas autoridades del organismo declaraban haber radicado una denuncia penal con relación a "la detección de irregularidades en la base de datos definitiva del Censo Nacional de Población, Hogares y Viviendas 2010" (INDEC, 2016). En el texto difundido daban cuenta de hechos similares a los descritos por Molinatti en su presentación. El comunicado indicaba lo siguiente:

> En particular, se ha detectado la traspolación de los datos de una población a otra. Por ejemplo, la población de Humahuaca estaba replicada sobre otra ubicación geográfica, distinta a la de referencia.
>
> Hasta el momento se han detectado réplicas de registros de personas en aproximadamente 400.000 casos; por ejemplo, réplicas o clonación de registros de individuos en un rango que va desde duplicaciones de dos (2) hasta ciento treinta (130) veces; de estos últimos, en 94 casos los mismos registros se replicaron 130 veces (INDEC, 2016).

Luego de esto, sin embargo, el INDEC no publicó informes técnicos o bases de datos que rectificaran la información ya publicada por el organismo, o que describieran el alcance o las características de los problemas encontrados.

Esta combinación de hechos dejaba, cuando menos, dudas sobre la información censal disponible y los usos que se pudiera hacer de ella. Si esos eran los casos detectados hasta el momento del comunicado, ¿se detectaron más casos luego? ¿Esa cifra aproximada de 400.000 casos representaba la cantidad de personas que aparecían duplicadas, o el total de réplicas?

El procedimiento que se presenta aquí estuvo dirigido a intentar responder, al menos parcialmente, algunas de estas preguntas.

B. Metodología

1. Fuentes de información y herramientas

Para realizar las comparaciones entre los atributos de viviendas, hogares y personas se utilizó la base de datos en formato REDATAM del censo nacional de 2010 (cuestionario básico). Se tomó como muestra de control la base de datos del censo nacional de 2001.

En ambos casos, se convirtieron en listados de viviendas, hogares y personas, utilizando el paquete de código abierto *Conversor REDATAM* (De Grande, 2016). Este *software* permite realizar exportaciones de datos desde bases de REDATAM sin pérdida de información debido al uso de formatos de microdatos en archivos CSV o SPSS.

Las comparaciones de viviendas y grupos que se indican en la siguiente sección se realizaron con una aplicación en lenguaje C♯ sobre dichos datos y se construyeron bases de datos intermedias en formato SQLite3. Los análisis cuantitativos posteriores se hicieron con el paquete estadístico SPSS.

2. Selección de indicadores

Como se comentó anteriormente, el propósito de esta investigación fue buscar si había personas repetidas en la base de datos del censo de población de 2010. Diversas fuentes señalaban la existencia de estos casos, pero la falta de identificadores únicos en las bases públicas censales dificultaba su caracterización. Los atributos simples de una persona podían no ser suficientes para distinguirla de cualquier otra persona con iguales características demográficas. ¿Eran el mismo registro censal, insertado dos veces en la base de datos, o simplemente eran dos personas con iguales características?

Para poder realizar una clasificación de casos duplicados, se decidió sortear la dificultad de no poder reconocer a las personas en la base de datos públicos debido a la escasez de atributos. Con ese fin, el ejercicio de comparación de personas idénticas se convirtió en un ejercicio de comparación de grupos idénticos de personas sucesivas idénticas.

Esto implicaba adoptar el supuesto de que la probabilidad de que dos grupos de personas independientes compartan el total de sus atributos (ser idénticas en orden y contenido de todas sus características) decrece a medida que aumenta el tamaño de los grupos comparados. Si se incorporan cantidades suficientes de atributos, y de individuos, esta probabilidad sería lo suficientemente pequeña como para no afectar una estimación de duplicaciones producidas de manera artificial.

De ser así, si se tomaran en conjunto atributos de todas las personas de un grupo, como la edad, el sexo, la condición ocupacional, los materiales de la vivienda, los años de educación o la condición de asistencia a un establecimiento educativo, sería muy improbable encontrar otro grupo con idénticas características, con igual número de personas e iguales valores de los indicadores.

Un elemento a tener en cuenta para evaluar la viabilidad de este procedimiento es que el número efectivo de grados de libertad de las variables censales es difícil de establecer, en especial cuando aparecen en forma combinada (Villa Diharce, 2004). Las variables no solamente presentan distribuciones heterogéneas (el hecho de poseer una heladera es un valor fijo "en sí" en casi todos los hogares, mientras que disponer de microondas es un factor de mayor variación), sino que muchas de ellas covarían. Por ejemplo, quienes no tienen heladera difícilmente tendrán microondas, y es posible que las variables de calidad constructiva de su vivienda guarden relación con la citada falta de heladera.

Todo lo anterior hace que la libertad de variación de los valores (en definitiva, aquello que gobierna la probabilidad de encontrar dos hogares o dos grupos de hogares idénticos) diste mucho de poder calcularse como el mero producto entre cada conjunto de categorías

posibles. Su aproximación debe ser empírica, y por ese motivo se utilizó la información de los últimos dos censos nacionales.

El conjunto de variables que existe en ambos censos (afinidad necesaria para poder utilizar uno como control del otro) es relativamente extenso, e incluye variables de vivienda, de hogar y de personas. Además, algunas de estas variables, como la edad y el sexo, no son constantes ni a nivel de barrio ni a nivel de viviendas y se complementan con características del hogar (como la disponibilidad de bienes) que, tomadas en conjunto, ofrecen una importante cantidad de atributos a contrastar.

3. Procedimiento

El objetivo del procedimiento no fue detectar diferencias en estimadores poblacionales (por ejemplo, comparar la varianza que tienen ciertos atributos en el censo con relación a la varianza esperable), sino identificar a las personas duplicadas.

Las variaciones en los estimadores podían no ser significativas (dado que la duplicación de casos fue parcial, y hecha en grupos; no se copiaron o reponderaron hogares, sino grupo de hogares). Incluso en caso de manifestarse, habrían confirmado lo que el mismo INDEC ha señalado con anterioridad, a saber, la existencia de réplicas de carácter no específico en la base de datos censal.

Sin embargo, dicha estrategia no habría tenido la flexibilidad necesaria para permitir un abordaje más específico, centrado en tareas tales como la de estimar la distribución de estos casos por provincia o su incidencia y mapear la distribución geográfica de las copias y los casos que las originaron, entre otras.

a) Comparación individual

Para proceder caso por caso, era menester poder comparar entre sí las viviendas (y los grupos de viviendas). Antes de hacer referencia a la comparación de grupos, cabe detallar cómo se realizó la comparación entre cada par de viviendas en el censo de población (para luego poder comparar todas las viviendas del censo entre sí, y después, en grupos).

Una vivienda en una base de datos censal contiene características que la describen, principalmente, como construcción edilicia (el tipo de techo, el material de los pisos, entre otros factores). Al mismo tiempo, se asocia a dos niveles más de información: i) los atributos de los hogares, que describen, entre otras cosas, la disponibilidad de bienes de cada hogar (por ejemplo, heladera, microondas), o su estructura sociodemográfica, y ii) los atributos de cada uno de los integrantes de esos hogares (por ejemplo, edad, ocupación, nivel educativo alcanzado).

Ante la necesidad de tener un mecanismo para comparar viviendas sin reducir su variabilidad original, se definió que se debían tomar en cuenta atributos en los tres niveles, es decir, encontrar viviendas con atributos iguales como viviendas (por ejemplo, el tipo de techo o de paredes), con hogares iguales en su interior (misma cantidad y composición) y con personas iguales en cada uno de ellos.

Con este fin, no se recurrió a la construcción de un índice sintético (por adición de los valores, o cálculos de componentes principales, entre otras cosas) para representar a cada vivienda, sino que se utilizó la totalidad de la información disponible en los indicadores seleccionados. Se incluyeron todos los que se hubieran publicado para ambos censos y se excluyeron de la lista las variables que pudieran tener un anclaje geográfico, como el municipio, la localidad, el departamento o la provincia. Esto se debió a que uno de los objetivos era determinar los casos que tal vez se hubieran duplicado hacia jurisdicciones diferentes a las de origen. Si se hubiera incluido la localización geográfica entre las características del hogar, habría sido imposible detectar la semejanza completa entre hogares de diferentes áreas.

En el cuadro 1 puede verse la lista de variables seleccionadas para la comparación de los grupos de viviendas de ambos censos. Estas variables se utilizaron para realizar el análisis del censo de 2010. Se utilizó el mismo procedimiento que se había aplicado en el censo de 2001 para tener un marco de referencia respecto de la plausibilidad de que determinadas viviendas o grupos de viviendas presentaran valores idénticos en ese mismo conjunto de variables para una población censal.

Cuadro 1

Argentina: variables utilizadas para la caracterización de los hogares, disponibles en ambos relevamientos censales, 2001 y 2010

Variables		Descripción	Categorías	
2001	2010		2001	2010
CH13	H06 y H07	Material de la cubierta exterior de los techos	15	9 y 3
H5	H05	Material predominante de los pisos	5	5
CH9	H10 y H11	Servicio sanitario	5	3 y 3
H10	H08	Tenencia de agua	4	4
H11	H09	Procedencia del agua	10	7
CH6	PROP	Régimen de tenencia	7	7
CH23	H19A	Tenencia de heladera o *freezer*	4	3
CH26	H19B	Tenencia de computadora y conexión a Internet	4	3
CH25	H19C y H19D	Tenencia de teléfono	5	3 y 3
H21	H16	Cantidad de habitaciones o piezas en total		
H19	H14	Combustible usado principalmente para cocinar	7	8
H16	H13	Baño o letrina de uso exclusivo	3	3
CC4	H12	Desagüe del inodoro	4	5
CP63[a]	CONDACT	Condición de actividad	4	4
P1	P01	Relación de parentesco	11	10
P2	P02	Sexo	3	3
P3	P03	Edad	112	112
P9A	P05	En qué país nació	3	3
P9APAIS	P06	País de nacimiento		
P4	P07	Sabe leer y escribir	3	3
CP3	P08	Condición de asistencia escolar	5	4
CP12		Años de escolaridad aprobados	19	

Variables		Descripción	Categorías	
2001	2010		2001	2010
	P09	Nivel educativo que cursa o cursó		9
	P11A	Último grado o año que aprobó en ese nivel		8
	P10	Completó el nivel		3
Casos en 2001		36 260 130		
Casos en 2010		40 117 096		

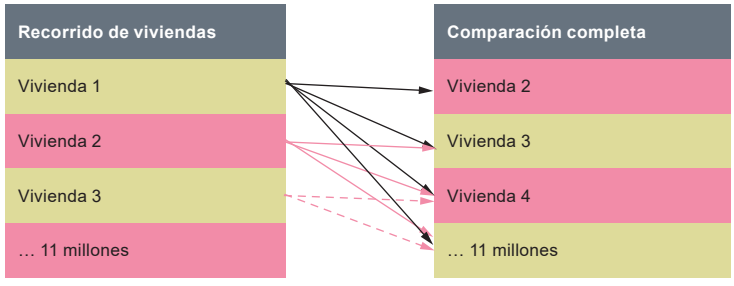
Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.
[a] Recodificada en cuatro categorías iguales a las de 2010.

Como se comentó anteriormente, se consideró que cualquier operación de agregación simple de los datos (promedios, conteos o sumas) de las viviendas en indicadores sintéticos habría reducido la variación posible entre las viviendas. Esta reducción habría producido un aumento de la posibilidad de dar con falsos positivos evitables, por ejemplo, viviendas que resultaran iguales en un índice sintético, pero que fueran diferentes en alguno de los atributos al considerarlos individualmente.

En consecuencia, para reconocer y extraer de las bases de datos censales los grupos de viviendas que tuvieran valores iguales en el total de los indicadores seleccionados (siendo algunos del nivel de la vivienda, otros del hogar y otros, de las personas), se eligió una estrategia en que cada vivienda se viera representada por la concatenación de todos sus atributos. De este modo se evitaría la pérdida de información, lo que permitiría hacer una mejor captación de los casos repetidos.

Cabe aclarar que se utilizaron solamente variables presentes en ambos censos (cuestionario básico de 2010 y cuestionario de 2001, véase el diagrama 1), para así reforzar su comparabilidad y evitar imprecisiones en la estimación del tamaño mínimo aceptable de los grupos de viviendas sucesivas que pudieran surgir de diferencias intercensales de la longitud en los identificadores.

Diagrama 1
Argentina: recorrido de viviendas, enumerando las que tienen iguales valores en sus atributos y los de sus habitantes



Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

b) Rastreo de grupos idénticos

Para poner esta modalidad de comparación individual al servicio de la detección de grupos de viviendas repetidas, se definieron los pasos descritos a continuación.

i. Detección de pares de viviendas iguales

En primer lugar, se procedió a calcular un identificador único para cada vivienda a partir de sus valores. Con este procedimiento se abarcaron los atributos de todos los miembros de cada hogar, en cada vivienda, sumándolos en forma de texto. De esta manera, si las primeras tres variables de la vivienda (material de techos, material de pisos y servicio sanitario) tienen valores 1, 2, y 1, se almacena el texto "1|2|1", para luego seguir agregando las demás variables de la vivienda. Después, en la misma unión de valores, se agrega cada atributo de los hogares declarados en ella (por ejemplo, tenencia de heladera, tenencia de teléfono) y se procede de igual modo con los valores de las personas en cada hogar.

Esto produce un identificador en forma de texto extenso de valores concatenados "1|2|1|3|4|1|3|2|1|2|3|1|7|...". Su longitud varía en función del número de hogares y de personas que residan en la vivienda, y debe ser construido para cada una de ellas.

En segundo lugar, para almacenar estos valores y poder compararlos de manera más eficiente, se implementó un mecanismo de *hashing*, es decir, una conversión a un identificador numérico de tamaño fijo, que asegura una pérdida casi nula del grado de singularidad del dato. Se realiza un cálculo de *hashes* con el algoritmo SHA-512[4]. De esta manera, cada vivienda queda representada por un identificador de 64 *bytes,* es decir, por un valor entre 0 y 2^512, y se obtiene un listado donde figura en cada fila el identificador numérico secuencial de la vivienda (1, 2, 3, 4) junto al *hash* de su contenido (código de 64 *bytes*).

Una vez calculado el *hash* de cada vivienda (el identificador basado en sus atributos unidos a los de sus habitantes) se procede a determinar las viviendas que tengan iguales atributos. Para esto, se recorre la lista de viviendas, comparando el *hash* de cada vivienda con los de todas las viviendas posteriores a ella en la lista (véase el diagrama 1)[5].

Se establecen así los pares de viviendas que tienen el mismo valor en su *hash*, es decir, que presentan valores idénticos en todos sus atributos censales.

ii. Identificación de grupos

Se había tomado como premisa que la existencia de dos viviendas con atributos idénticos en un censo podía responder a una coincidencia fortuita de viviendas iguales en la realidad, por lo que se decidió aceptar como posibles copias (artificialmente generadas) solamente los grupos de viviendas sucesivas que fueran idénticos entre sí.

---

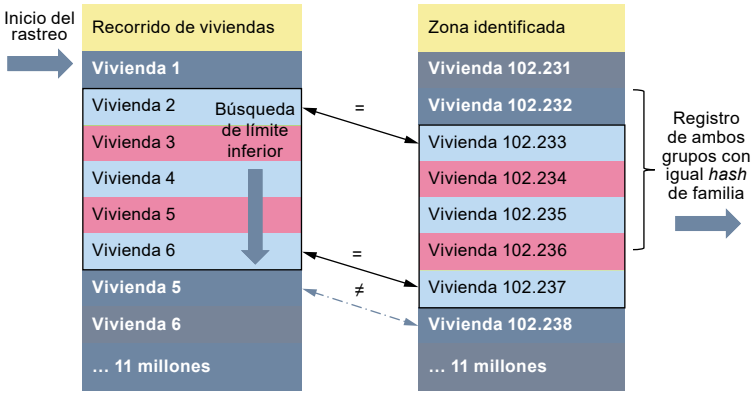
4 Los *hash* son valores producidos por algoritmos de criptografía que permiten obtener un valor transformado (no reversible) y usualmente más breve de un conjunto de datos. Se busca una máxima variabilidad entre los valores del *hash* a partir de los valores ingresados (Blain Escalona y Vázquez Inclán, 2011).

5 Solo se consideraron las viviendas posteriores para evitar construir pares repetidos (A->B, B->A). La lista se recorrió en forma ordenada ascendente por identificador.

Para lograr la identificación de grupos de viviendas, cabía rastrear secuencias de pares iguales, sin conocerse *a priori* el tamaño que podían tener estos grupos (es decir, se trataba de grupos de tamaño potencialmente variable).

El procedimiento implementado para este fin supuso un recorrido por los pares detectados en el paso anterior, con el fin de verificar si las viviendas posteriores a cada par también eran iguales entre sí (si había un par vecino al primer par). En caso afirmativo, se repetía la misma verificación con la vivienda siguiente, para comprobar si también formaba par con la vivienda siguiente a los pares (copias) de las anteriores (véase el diagrama 2).

Diagrama 2
Argentina: rastreo de márgenes inferiores para la identificación de grupos de viviendas idénticas



Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

De esta forma, se establece el límite inferior de la semejanza entre viviendas para cada tramo del listado (las viviendas se recorren en orden, por lo que siempre se comienza desde el límite superior, o sea, la primera vivienda coincidente). En el diagrama 2 puede verse el ejemplo de la identificación de un grupo de cinco pares de viviendas sucesivas con *hashes* idénticos (valores iguales en todos sus atributos). Este rastreo permite encontrar grupos de pares de viviendas sucesivas con valores iguales (para la explicación que sigue, se considerarán "pares de grupos de viviendas", más que "grupos de pares de viviendas"). Con estos (grupos de viviendas de los que existe otro grupo igual en la base de datos), el procedimiento produce una lista donde se establece, respecto de cada uno de ellos, su punto de inicio (la primera vivienda en ambos conjuntos de hogares idénticos), el tamaño (la cantidad de hogares y de personas involucradas) y un identificador.

Para construir el identificador del grupo, se unen todos los *hashes* de las viviendas que lo componen y, a partir de esa información, se produce un nuevo *hash*. Este sirve de código (o identificador único de contenido) para el grupo y representa la suma de atributos de todas sus viviendas y miembros.

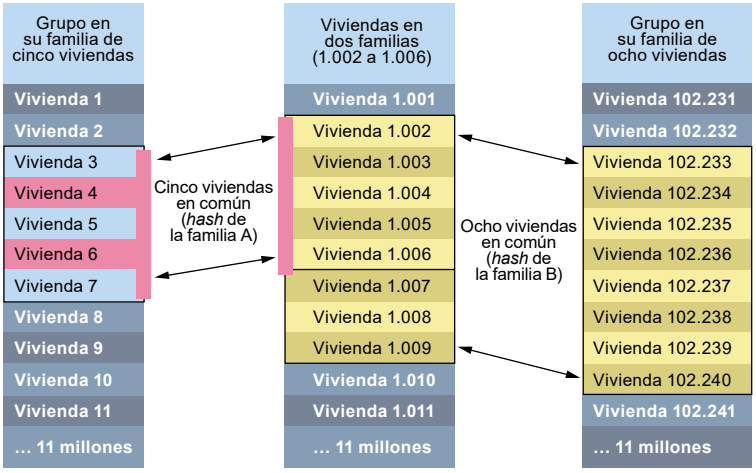
Al construirse mediante un *hash* de sus valores (y no como un número creciente asignado en forma exógena), este código es idéntico en todos los grupos de viviendas con iguales atributos internos, por lo que puede denominarse código (o *hash*) de "familia".

Una familia de grupos de viviendas está compuesta por todos los grupos repartidos en la base de datos censal cuyos atributos internos son iguales. El *hash* de familia permite identificar de manera coherente al grupo, sin importar cuántas veces aparece en la base de datos. Gracias a ello, es posible contabilizar cuántas copias iguales hay de un mismo grupo en los registros. El conjunto de grupos iguales entre sí se denominará "familia".

Es importante considerar, para los conteos que se presentan en la sección de resultados, que una vivienda (o grupo de viviendas) puede pertenecer a varias familias (y de hecho ocurre) en los casos en que se encuentren duplicadas en diferentes zonas de la base de datos en series de diversa longitud. Esto hace necesario distinguir, por una parte, las coincidencias entre viviendas, y por otra, el conteo de viviendas involucradas.

De esa forma, una vivienda puede pertenecer a una primera "familia" (que puede constituirse de conjuntos de cinco viviendas repetidas), pero también puede considerarse perteneciente a una segunda familia de repeticiones cuyo tamaño sea de ocho miembros (o).

Diagrama 3
Argentina: relación entre viviendas y múltiples familias



Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

Además, una validación que se tuvo en cuenta para el análisis espacial de estas "familias" consistió en verificar si estos grupos de viviendas iguales respetaban la delimitación del radio censal, o si, por el contrario, un grupo familiar podía extenderse (en el listado de viviendas), trascendiendo los puntos de corte que los radios suponen.

Del análisis de la base de datos se desprende que no existe una circunscripción por radio del fenómeno de duplicación de grupos de viviendas. Es decir, una secuencia de hogares puede tener como origen una lista de viviendas que exceda los límites del radio. Esto fue así en 214 ocasiones a nivel nacional, lo que hace suponer que en la réplica de valores no se haya tomado el radio como criterio de selección de viviendas a duplicar.

Para comentar un caso en particular (de grupos partidos), la familia 1001 es un grupo de 11 viviendas del radio 068613210 que se encuentran duplicadas varias veces, y en orden, en los radios 068612005 y 068612006. De esas 11 viviendas, solo 10 están copiadas hacia el final del radio 068612005. Este proceso de copia continúa en el radio 068612006 con la vivienda número 11 de la serie visible en el radio 068613210. Luego, hacia el final del radio 068612006, se encuentran otras dos copias completas de la familia, que continúan también con un corte parcial sobre el radio 068612007.

iii. Preparación de las bases de datos para el análisis

Una vez definidos los grupos de viviendas equivalentes y las familias que conforman, resta construir las bases de datos que sirven para el análisis estadístico de los resultados. Estas bases se enumeran a continuación:

- Una base de datos con la información descriptiva de los grupos, que posee un registro por cada grupo, en que se indica su tamaño, la vivienda de inicio y su *hash* de familia.
- Una base de pares de viviendas, a partir de la expansión de los grupos de familias, que permita crear listas y agrupar por radio las viviendas que poseen "familiares" en otros radios por el parentesco de alguna de sus familias de pertenencia.
- Una base de datos de personas en "familias", que permita caracterizar a nivel nacional y provincial a las poblaciones cuyos datos aparecen más de una vez en el censo en comparación con los hogares que no pertenecen a ninguna "familia".

A continuación, se presenta el análisis elaborado a partir de las bases de datos resultantes para los microdatos del censo de 2010, con uso del censo de 2001 como línea base o grupo de control.

C. Resultados

1. Umbral de aceptabilidad (para evitar falsos positivos)

Un paso que pareció necesario en el procesamiento de esta información consistió en establecer los tamaños de grupo que debían considerarse suficientemente grandes para que fuera despreciable la probabilidad de tratarse de atributos iguales por causa del azar.

Si el umbral era demasiado bajo, se podían encontrar casos en que la similitud se debiera simplemente a la coincidencia espontánea de sus respuestas y no a problemas del procedimiento censal. La idea de comparar grupos de viviendas (no solamente una) tiene por objeto aclarar esta relación entre azar y redundancia de la información.

Al mismo tiempo, si se fijara un umbral demasiado exigente, como el de considerar solamente como válidos los grupos de al menos 50 viviendas sucesivas, se podrían perder casos que deberían ingresar en la clasificación.

La incorporación del censo de 2001 al análisis tiene por objetivo dar una base empírica a la elección de este punto de corte. Considérese a esos fines que el Censo de Población, Hogares y Viviendas de 2001 no posee viviendas duplicadas voluntariamente en su procesamiento. Los casos en que hubiera viviendas idénticas sucesivas serían explicables por errores de carga aislados o por azar genuino de la información (por mera casualidad).

El censo de 2001 facilita así el establecimiento de un punto de partida en una línea de base en que las repeticiones se correspondan con coincidencias fortuitas, o con errores de carga, que deberían ser relativamente escasos si se aplica un umbral apropiado de clasificación[6].

Al utilizar el listado de grupos (el resultado del último paso del cálculo), es posible establecer la cantidad de grupos de viviendas idénticas, y de familias, en cada base censal, cuantificándolos según el tamaño del grupo. De este modo, puede verse el volumen de grupos repetidos que se concentran según los diferentes tamaños, e intentar así dar con un punto de corte que permita equilibrar los riesgos clasificatorios de falsos positivos y la subestimación de casos.

En el gráfico 1 puede verse la cantidad de familias y grupos con viviendas sucesivas de iguales valores detectados en el censo de 2001 y en el censo de 2010 (eje vertical), presentados según su tamaño en términos de cantidad de viviendas (eje horizontal).

Hay varias cuestiones que resaltan en el gráfico 1. En primer lugar, el censo de 2001 y el censo de 2010 muestran una distribución desigual de las repeticiones de valores en todos los tamaños de grupos de casos.

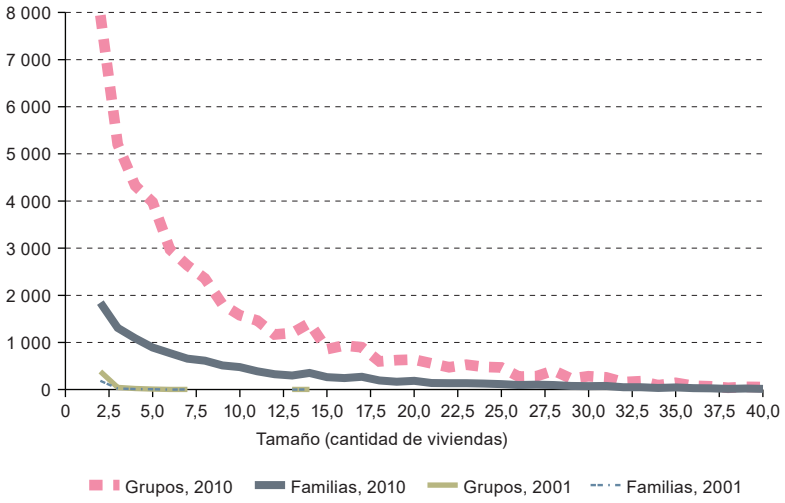
Si bien se sabía que existe un volumen atípico de datos redundantes en el censo de 2010, se puede comenzar a conocer su morfología. La distribución de grupos y familias indica, por una parte, que hay grupos extensos de viviendas repetidas (el cuadro corta la serie, pero se registraron grupos de hasta 97 viviendas). Por otra, incluso en grupos de dos viviendas (el tamaño mínimo considerado en la serie), la distancia con el censo de 2001 es muy notoria: mientras que en dicho censo hubo 384 grupos de dos viviendas, el censo de 2010 registró 7.943 (20 veces más).

[6] Si bien en el proceso de carga de 36 millones de cédulas censales cabe esperar que ocurran errores (incluida la carga repetida de una o varias cédulas), estas situaciones deberían ser muy inferiores a los casos bajo sospecha del censo de 2010.

Gráfico 1

Argentina: familias y grupos de viviendas con valores idénticos según cantidad de viviendas que participan en los grupos, censos de 2001 y 2010

(En número de familias y grupos de viviendas)



Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

En este sentido, si se tomara la cifra observada en el censo de 2001 como una referencia de la cantidad de coincidencias fortuitas y errores de carga esperables en un operativo de estas características, los grupos de dos viviendas sucesivas idénticas en ese censo son apenas un 5% de los registrados en el censo de 2010. Es decir, podrían tomarse las repeticiones de valores encontrados en el censo de 2010 como muy mayoritariamente válidas, en términos de ser coincidencias no fortuitas incluso para la línea de dos viviendas.

Sin embargo, se maximizará la prudencia del punto de corte, para conferir una robustez adicional a las estimaciones resultantes. Al analizar la caída en la curva de casos en las series correspondientes al censo de 2001, puede verse que, a partir de las cinco viviendas, se registran menos de diez grupos por categoría, y menos de cinco en los pocos casos de seis y más viviendas.

Con esa observación como referencia, se tomarán como punto de corte para analizar la información del censo de 2010 los grupos de cinco y más viviendas, excluyendo de los conteos las viviendas sucesivas coincidentes de hasta cuatro casos[7].

[7] En términos cuantitativos, para la estimación de cantidad total de pares coincidentes, esto modifica en un 10,36% los cómputos totales (422.207 pares coincidentes con cinco viviendas como punto de corte, 470.996 con dos viviendas como punto de corte). En consecuencia, con el primer punto de corte se logra un menor riesgo de falsos positivos, sin distorsionar gravemente la estimación general de los totales de los grupos.

2. Grupos, familias, viviendas y personas

Una primera caracterización que es posible hacer, una vez logrado un punto de corte robusto, es intentar responder cuántas "familias" existen (es decir, cuántos son los grupos de viviendas con conjuntos de "parientes" idénticos en otras localizaciones). También es pertinente preguntarse cuántas personas las componen y qué grado de repetición hay dentro de esas familias.

Para realizar y describir estas estimaciones, considérese que, dado que todos los grupos de las familias son iguales entre sí, solo las viviendas en uno de ellos estarían en situación de contabilizarse como viviendas reales a censar, siendo las demás meras réplicas de su información. Ese grupo se denominará "grupo originario", y los demás grupos se considerarán imágenes o copias suyas.

Para aproximarse a la cuantificación de viviendas y personas afectadas por las anomalías referidas en la introducción, en el cuadro 2 se distinguen tres tipos de condiciones de participación censal.

Cuadro 2

Argentina: cantidad de viviendas y personas, según participación en grupos de casos idénticos, cinco o más viviendas, censo de 2010

	Condición de participación				Total censal
	En "familias"			Viviendas únicas	
	Originarias	Réplicas	Total		
Personas					
Número	144 282	312 229	456 511	39 660 585	40 117 096
Porcentaje del total	0,36	0,78	1,14	98,86	100,00
Viviendas					
Número	43 779	96 086	139 865	11 203 614	11 343 479
Porcentaje del total	0,39	0,85	1,23	98,77	100,00

Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

También se da cuenta de las viviendas que no forman parte de ninguna "familia", es decir, que no se encuentran incluidas en grupos de viviendas de más de una aparición censal.

En términos absolutos, la cantidad de registros censales que según esta estimación serían imágenes de otros representa 312.229 personas (0,78% del total general). Respecto a las viviendas, 96.086 estarían en igual situación, es decir, un 0,85% del total. El total de personas que se encontró como participantes de al menos una "familia" fue de 456.511.

3. Distribución geográfica de las "familias"

La participación en familias tuvo una incidencia desigual según la provincia de que se tratase. Mientras que Catamarca, Entre Ríos, La Pampa, Río Negro, Neuquén, San Juan y San Luis no presentaron casos de viviendas duplicadas, en la Ciudad de Buenos Aires la incidencia de personas en familias llegó al 2% de la población, y en Salta y Chaco rondó el 0,15% (véase el cuadro 3).

Cuadro 3

Argentina: cantidad de personas en familias; cantidad, tamaño, repeticiones y extensión entre radios de las familias

Provincia	Personas en familias		Familias				Extensión entre radios de las semejanzas (*En porcentajes*)				
			Cantidad	Tamaño (*Personas*)	Repeticiones						
	Número	Porcentaje	Número	Media	Media	Máximo	Uno	Dos	Tres	Cuatro a seis	Siete a veintitrés
Ciudad Autónoma de Buenos Aires (CABA)	57 681	2,00	832	41,28	2,38	18	97,17	2,83			
Buenos Aires	256 105	1,64	5 539	48,81	4,38	103	46,08	22,84	7,99	10,10	13,00
Catamarca	-	-	-	-	-	-	-	-	-	-	-
Córdoba	20 899	0,63	294	39,83	2,28	8	83,15	15,30	1,55	-	-
Corrientes	13 240	1,33	144	47,85	2,05	3	100,00	-	-	-	-
Chaco	1 564	0,15	20	39,95	2,00	2	95,91	4,09	-	-	-
Chubut	7 056	1,39	99	36,80	2,04	3	100,00	-	-	-	-
Entre Ríos	-	-	-	-	-	-	-	-	-	-	-
Formosa	5 067	0,96	75	41,31	2,61	11	72,84	27,16	0,00		-
Jujuy	1 389	0,21	32	32,84	3,50	14	17,13	29,66	21,60	31,61	-
La Pampa	-	-	-	-	-	-	-	-	-	-	-
La Rioja	2 862	0,86	36	39,75	2,00	2	100,00	-	-	-	-
Mendoza	19 279	1,11	232	45,17	2,14	5	78,48	18,20	3,32	-	-
Misiones	1 454	0,13	12	60,58	2,00	2	100,00	-	-	-	-
Neuquén	-	-	-	-	-	-	-	-	-	-	-
Río Negro	-	-	-	-	-	-	-	-	-	-	-
Salta	1 885	0,16	24	38,33	2,04	3	64,19	35,81	-	-	-
San Juan	-	-	-	-	-	-	-	-	-	-	-
San Luis	-	-	-	-	-	-	-	-	-	-	-
Santa Cruz	2 762	1,01	37	49,54	2,24	5	80,49	19,51	-	-	-
Santa Fe	44 835	1,40	669	41,79	2,53	13	93,34	6,66	-	-	-
Santiago del Estero	18 835	2,16	274	42,23	2,50	13	94,01	5,99	-	-	-
Tucumán	84	0,01	1	42,00	2,00	2	100,00	-	-	-	-
Tierra del Fuego	1 514	1,19	25	35,84	2,12	3	84,41	15,59	-	-	-
Total	**456 511**	**1,14**	**8 345**	**46,47**	**3,70**	**103**	**65,93**	**16,26**	**4,76**	**5,76**	**4,08**

Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

Respecto al tamaño medio de los grupos familiares, el promedio general fue de 46,47 personas. En la provincia de Misiones se registró el promedio máximo, de 60,58. Esta medida da cuenta de la extensión de los grupos de personas copiados, mientras que la cantidad de repeticiones permite evaluar el número medio de veces que se encontró cada grupo (véase el cuadro 3).

La provincia de Buenos Aires presentó un número atípicamente alto de repeticiones, siendo su promedio 4,38 y su valor máximo, 103 (es decir, hay series de viviendas que pueden hallarse repetidas hasta 103 veces en los registros de la provincia). A nivel nacional, el promedio de repeticiones de los grupos familiares fue de 3,7 veces (véase el cuadro 3).

Para evaluar la distribución geográfica de los casos es preciso establecer un criterio de corte espacial para agruparlos. En este análisis se presentan los datos por provincia y se han evaluado dos subniveles administrativos de la provincia que son el departamento y el radio. No se registraron familias cuyos miembros residieran en más de una provincia, y solo en la provincia de Buenos Aires se encontraron familias que habitaban en varios departamentos a la vez (27 familias).

Como consecuencia de esto, el análisis entre distancias de los miembros de las familias se ha realizado a nivel de radio, en el cual se han detectado pares de grupos de viviendas iguales en diferentes áreas (radios) en la mayoría de las provincias. Esto permite reconstruir los radios que se encuentran en relación por poseer copias de grupos familiares en más de un radio.

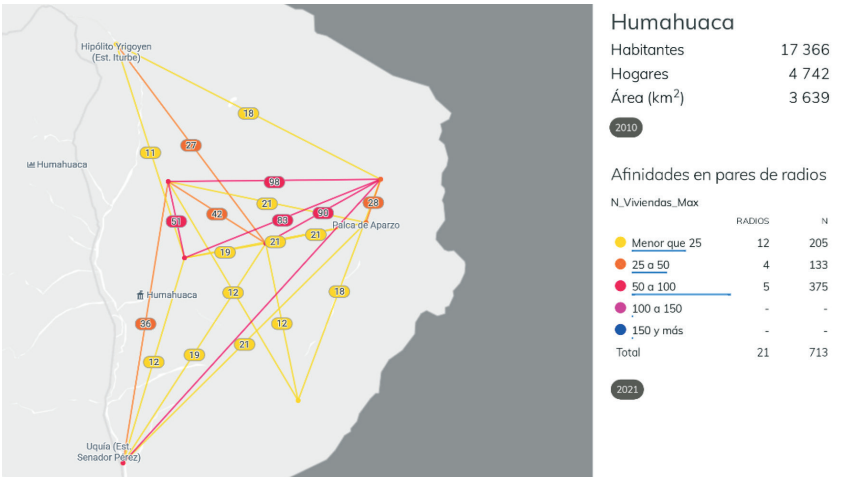
El criterio que parece haber tenido mayor prevalencia es el de la copia de hogares sobre el mismo radio. En estos casos, la operación funcionaría como una ponderación de las personas duplicadas, es decir, mientras que algunos respondientes son considerados una sola vez en su radio, otras viviendas fueron repetidas. Sin embargo, una proporción importante de viviendas duplicadas no sigue este criterio. El 34,07% de las repeticiones ocurren entre dos o más radios, siendo la Provincia de Buenos Aires el distrito que registra mayores números de radios para una misma familia. En ella, hay familias (grupos de viviendas) que llegan a estar localizadas en 23 radios diferentes (0).

En las provincias en que se detectaron familias que ocupaban varios radios, es de interés comprender su distribución geográfica. Para examinar algunos de estos casos, se elaboraron mapas con la ubicación de los diferentes grupos de réplicas que formaban cada familia. El primero que se verá es el que figuraba en la denuncia radicada por el INDEC en el caso de Humahuaca (Salta).

En el mapa 1 se muestra el departamento de Humahuaca, donde cada línea representa un conjunto de viviendas sucesivas idénticas entre sí. La existencia de redes donde los radios se encuentran interconectados refleja la transitividad de estas similitudes, tal vez derivada de la reutilización de un mismo conjunto de viviendas en varios radios. Al hacerse esto, no solamente el radio de origen y los de destino presentan zonas de información idéntica, sino que también los radios "receptores" resultan iguales entre sí, de manera total o parcial.

Mapa 1

Argentina: distribución de semejanzas en secuencias de viviendas, censo de 2010

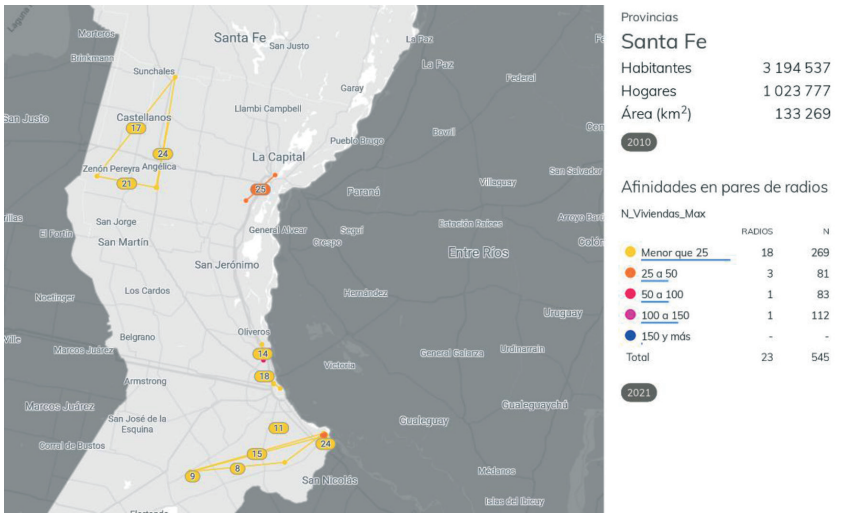


Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

Esto ocurre también en otras provincias del país, como Jujuy, Formosa o Córdoba. En la provincia de Santa Fe pueden reconocerse grupos de viviendas con "familiares" duplicados, en tamaños relativamente moderados en términos de cantidad de viviendas duplicadas, pero con distancias que exceden los 50 km ( ).

Mapa 2

Argentina: red de semejanzas entre grupos de viviendas en la provincia de Santa Fe, censo de 2010



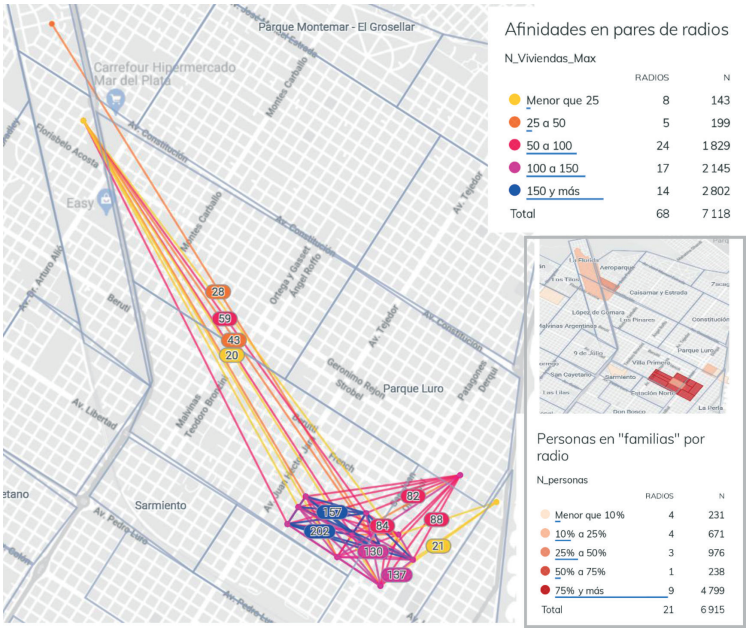
Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

La provincia que mayor incidencia de personas en "familias" mostró en el cuadro 3 fue la de Buenos Aires. Allí no solo fue mayor la cantidad de réplicas por total de habitantes, sino que fue la única provincia con mayoría de familias repartidas en más de un radio. Se analizarán varios escenarios en esta jurisdicción para ejemplificar la heterogeneidad de casos y dar cuenta de la coherencia de la información reconstruida.

En el caso de Mar del Plata, se registran radios con más del 75% de sus habitantes como réplicas de otros radios (parte inferior derecha del mapa 3).

Mapa 3

Argentina: red de semejanzas entre grupos de viviendas en la ciudad de Mar del Plata, censo de 2010



Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

Al observarse la red de semejanzas (imagen principal del mapa 3) relativa a esta zona, se pueden reconocer cuatro radios (dos al noroeste y dos más cercanos, hacia el noreste) conectados con buena parte de los radios de la zona de viviendas más problemáticas (las que presentan una mayor cantidad de personas en "familias"). Puede especularse que esta zona habría operado como receptora de información de personas de esos cuatro radios, mientras que en su interior (entre sus radios) se registran semejanzas recíprocas por valores superiores a las 200 viviendas por radio. De ser así, la forma de la red podría explicarse como compatible, con una similitud parcial entre los donantes y los receptores (cada uno donó una parte) y una similitud casi total entre todos los radios reconstruidos artificialmente.

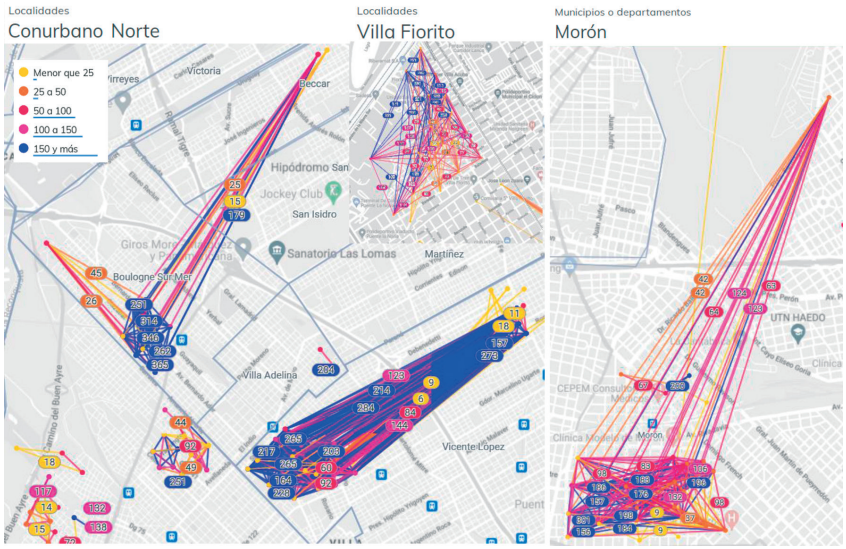
En zonas del conurbano bonaerense, se pueden encontrar situaciones en que el nivel de repetición de las secuencias de hogares parece indicar problemas serios en la captación

de los casos, ya que predominan radios con mayoría de casos que no son originales en zonas más amplias que lo señalado en Mar del Plata.

Tal es el caso de Villa Fiorito (Lomas de Zamora), un área con unos 23.000 habitantes donde más del 60% pertenece a "familias" de réplicas. Del análisis de su red de semejanzas se destaca una menor redundancia en las familias del este. Sobre esa base puede presumirse que las viviendas del oeste pudieron haber formado su representación a partir de información de las del este y que en el centro del barrio habría una enorme cantidad de viviendas redundantes (véase el mapa 3).

En otra zona del conurbano, un importante grupo de radios de Morón (entre su municipalidad y el cementerio), también dan cuenta de problemas de datos. En esa zona, los radios muestran grandes niveles de similitud entre sus viviendas, con una influencia significativa de un único radio del norte del municipio. Este radio parece haber influido notoriamente en la conformación de los radios de dicha zona, que tienen entre sí similitudes de más de 150 viviendas, con casos de más 300 viviendas idénticas ( ).

Mapa 4
Argentina: red de semejanzas entre grupos en el conurbano bonaerense, localidades seleccionadas, censo de 2010



Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

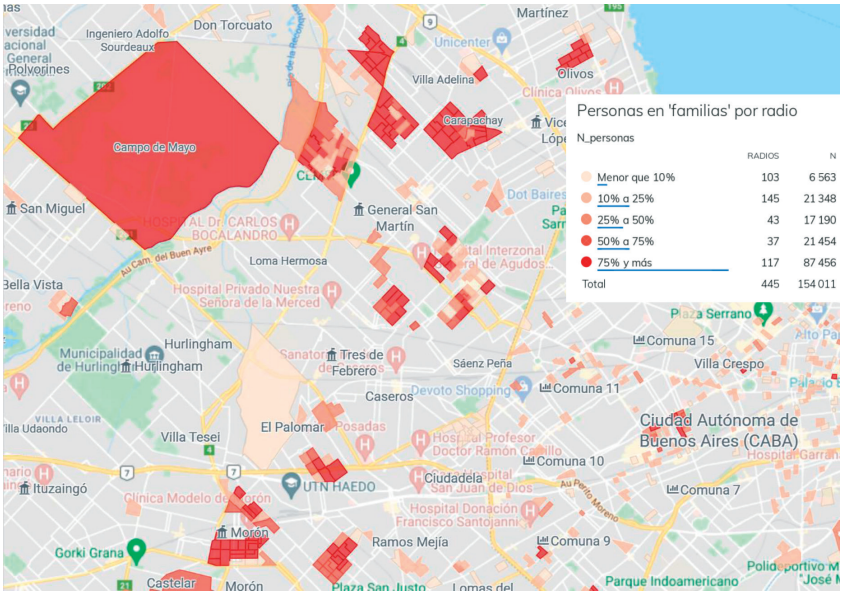
La zona norte del conurbano no es la excepción. Al contrario, presenta problemas aún más graves en lo que se refiere al volumen de repeticiones y las extensiones que recorren los casos. En esta zona, existen radios singulares que influyen en áreas completas, como es el caso de dos radios de Becar, en San Isidro, así como grupos de radios prácticamente idénticos en zonas apartadas entre sí de Vicente López (4).

4. Redundancia en los radios

Otra dimensión que puede ser de interés analizar, para comprender mejor la lógica de las repeticiones, es el grado en que estas se manifestaron en un mismo radio. Ello permite estimar en qué medida existen radios (zonas espaciales) que puedan estar compuestos mayoritariamente de copias o si, al contrario, las copias están distribuidas de manera relativamente dispersa por el territorio censado.

Si bien el análisis detallado de este fenómeno excede el alcance de este artículo, puede advertirse en el mapa 5 que en el Área Metropolitana de Buenos Aires la distribución de personas en "familias", de forma coherente con los mapas anteriores, se concentró en ciertas zonas de la ciudad.

Mapa 5
Proporción de personas en "familias" (no únicas) sobre el total de personas del radio para los radios con grupos de viviendas duplicadas, censo de 2010



Fuente: Elaboración propia sobre la base del procedimiento aplicado para la detección de grupos de viviendas duplicadas.

En el área representada en el mapa 5 se ven radios en que más del 50% y más del 75% de sus habitantes son copias de otras personas[8].

[8] Para permitir una exploración de los resultados, los datos construidos y referidos en este artículo pueden accederse en la cartografía titulada Análisis de viviendas idénticas (Censo 2010, Argentina) (véase [en línea] https://mapa.poblaciones.org/map/87101).

D. Conclusiones

A partir de los resultados de este análisis pueden extraerse conclusiones en diferentes niveles. En primer lugar, hay un conjunto de datos emergentes que se ha podido inferir respecto de la base de datos del cuestionario básico del Censo Nacional de Población, Hogares y Viviendas de 2010 de la Argentina, con los resultados desarrollados hasta aquí.

Se determinó que al menos 312.000 personas parecen ser réplicas de otras. Para ello se utilizaron umbrales de clasificación conservadores y se tomó como referencia la variabilidad de variables análogas del censo anterior. Se pudo verificar que los criterios con que se realizaron estas copias no fueron homogéneos en todas las jurisdicciones, ni en la intensidad (la proporción de personas afectadas) ni en los criterios de realización (si distribuir o no a las personas entre diferentes radios; cuántas veces copiar a la misma persona; cuáles eran los tamaños de los grupos copiados).

Esta heterogeneidad mostró tener efectos desiguales en los resultados. Es decir, lejos de tratarse de ajustes muestrales generales, varias zonas cargaron con buena parte de las réplicas existentes en sus distritos. No se vieron solamente operaciones de ponderación o expansión de algunos grupos de personas, sino que en lugares específicos más del 75% de los registros censales no parecían corresponder a respuestas efectivamente relevadas en la zona.

Al analizarse la distribución por provincia de estos criterios, se hizo evidente la heterogeneidad en las intervenciones. Algunas provocaban que las viviendas de la zona fueran mayoritariamente copias de zonas cercanas. En otros casos, la participación era menor, pero el origen más distante. En algunos casos se reutilizaban los grupos de viviendas pocas veces, en otros, muchas. Si bien la distribución de algunos parámetros en las viviendas duplicadas no mostró a primera vista desvíos importantes (por ejemplo, el nivel educativo de las duplicaciones no era significativamente diferente al del resto de la población), va más allá del alcance de este artículo definir los efectos de las duplicaciones en cada caso, por lo que dichos análisis no se incorporaron en la presentación de resultados.

En segundo lugar, se hace necesaria una reflexión sobre el proceder de la oficina responsable durante la producción del censo, pero también en el acompañamiento posterior de las áreas usuarias. Estas responsabilidades son difíciles de ponderar, habida cuenta de que en 2007 el INDEC fue intervenido por el Poder Ejecutivo, con el fin de distorsionar las estadísticas oficiales (Lindenboim, 2011)[9].

Esta situación se mantuvo hasta el recambio presidencial de 2015, en que una nueva dirección asumió el mando del organismo con el objetivo de poner en marcha su normalización. Al hacerlo, radicó en 2016 una denuncia penal por incongruencias en los datos censales, en que se hacía énfasis en la existencia de información duplicada.

[9] Esta distorsión se hizo tan extrema en lo referido a la inflación que, tal como señala Lindenboim (2011), en las negociaciones de aumentos salariales entre trabajadores y empresarios se llegó a abandonar el uso del índice oficial de precios y se pactaron aumentos a más del doble del valor de la inflación oficial, sin que ello implicara aumentos del salario real.

A pesar de ello, con posterioridad a dicha denuncia, el organismo no volvió sobre la cuestión ni publicó informes sobre estimaciones o datos relativos a estas anomalías. En este sentido, el organismo continúa hasta la actualidad sin elaborar un informe público en que se detalle la naturaleza de estas incongruencias.

La ausencia de información sobre cómo se elaboran los resultados censales no solamente está reñida con recomendaciones y buenas prácticas bien establecidas, como los *Principios Fundamentales de las Estadísticas Oficiales, 1994-2013* (Naciones Unidas, 2013) o el *Código regional de buenas prácticas en estadísticas para América Latina y el Caribe* (CEPAL, 2011), sino también con la legislación vigente en el país para el cuerpo de funcionarios del Estado nacional. Desde 2015, la Ley 27275 sobre el derecho de acceso a la información pública (Albertti, Giorno y Raschia, 2017) establece que la administración pública nacional y sus organismos descentralizados deben regirse por los principios de máxima divulgación, publicidad y transparencia de sus actos.

Además, es posible señalar que la cuestión de cómo y por qué reconstruir la confianza en la estadística pública presenta antecedentes internacionales bien documentados (Poledo y García, 2012), pero requiere estrategias concretas (Martín-Guzmán, 2016) y no solamente afirmaciones autocomplacientes de las autoridades (C5N, 2022) o de sus perfiles técnicos (Poledo y García, 2021). Esto no solo se fundamenta en que las dificultades en la producción estadística argentina llevan largo tiempo, sino también en que, ya completada una nueva ronda censal en 2022, la situación es ampliamente desalentadora. El cronograma oficial preveía la publicación de resultados provisionales para enero de 2023 y resultados ampliados y definitivos para el 18 de junio de 2023. En septiembre se carece de estimaciones sobre la mayor parte de las variables, sin un nuevo calendario de publicación ni una explicación de por qué hasta el momento ha sido imposible realizar el cálculo previsto, aunque solo fuesen resultados preliminares (*La Nación,* 2023).

La imposibilidad de aplicar mecanismos de revisión abierta por pares de los procesos de producción estadística, como hacen otros institutos estadísticos (Martín-Guzmán, 2016), y la falta de metas de producción de la oficina de estadística nacional aparte de cumplir con el censo en curso, quizás sean las mayores trabas para vislumbrar un horizonte de progreso en el conocimiento censal del país. Se perpetúa así un escenario de grandes falencias, como la ausencia de cartografía censal para rondas anteriores a 2010, numerosos errores de georreferenciación correspondientes a la cartografía del censo realizado en ese año (Rodríguez, 2021 y 2022), bases inexistentes o mantenidas en forma privada para censos anteriores a 2001, viviendas duplicadas en el censo de 2010, falta de información metodológica sobre los procesamientos, zonas no censadas, entre otras.

Los censos de población son herramientas que ayudan a planificar la política pública sobre la base de información: dónde ubicar escuelas, hospitales, rutas y transporte, o dónde hacer campañas de vacunación, entre otras cosas. Por la importancia que esto les otorga, debe existir en las oficinas estadísticas una actitud acorde a estos usos, tanto en la producción como en la publicación de sus resultados y sus errores asociados.

Procedimientos como los descritos en este artículo pueden ser de ayuda para definir criterios de evaluación o verificación nuevos y más complejos, a los que puedan someterse las bases de datos censales, para así dar cuenta de las adulteraciones voluntarias y de los errores involuntarios que los datos puedan presentar[10].

Ante los evidentes fallos en el censo de la Argentina de 2010, resulta crucial para próximas rondas censales revisar y transparentar los procedimientos de recolección y producción de datos por parte del organismo nacional de estadísticas. De sostenerse la actual falta de visibilidad y apertura, la calidad, la precisión y la representatividad de la información recopilada seguirán siendo inciertas, lo que impediría su mejora por mecanismos de verificación externos o de procedimientos públicos sujetos a la consideración de expertos y de actores interesados en el mejoramiento de la calidad de la producción censal.

Bibliografía

Abbott, O. y A. Large (2009), "Measuring the level of duplicates in the 2011 Census", documento presentado en la 17ª Reunión del Comité Asesor de Metodología GSS.

Albertti, P., M. Giorno y J. Raschia (2017), "Un nuevo instrumento del estado argentino para la gestión de la transparencia. Derecho de acceso a la información pública", *Red Sociales. Revista del Departamento de Ciencias Sociales*, vol. 4, N° 06.

Blain Escalona, S. y L. Vázquez Inclán (2011), "Funciones resúmenes o hash", *Telemática*, vol. 10, N° 1.

CEPAL (Comisión Económica para América Latina y el Caribe) (2012), *Código regional de buenas prácticas en estadísticas para América Latina y el Caribe*, Santiago [en línea] https://repositorio.cepal.org/bitstream/handle/11362/16422/FILE_148023_es.pdf.

C5N (2022), "Censo 2022: Lavagna destacó que el operativo fue un éxito", 18 de mayo [en línea] https://www.c5n.com/sociedad/marco-lavagna-censo-2022-informe.

De Grande, P. (2016), "El formato Redatam", *Estudios demográficos y urbanos*, vol. 31, N° 3.

INDEC (Instituto Nacional de Estadística y Censos) (2022), *Censo Nacional de Población, Hogares y Viviendas de la Argentina. Síntesis de la planificación del Censo 2022*, Buenos Aires.

___(2016), "Información de interés público sobre el Censo Nacional de Población, Hogares y Viviendas 2010", *Gacetilla de prensa* [en línea] https://www.indec.gob.ar/indec/web/Institucional-GacetillaCompleta-107.

___(2012), "Censo Nacional de Población, Hogares y Viviendas 2010: censo del Bicentenario, resultados definitivos", *Serie B*, N° 2, Buenos Aires.

Infobae (2017), "El Censo 2010 alcanzó el 97 por ciento de la población", 4 de noviembre [en línea] https://www.infobae.com/2010/10/28/544059-el-censo-2010-alcanzo-el-97-ciento-la-poblacion/.

La Nación (2023), "A casi un año del censo, todavía son pocos los datos que se conocen: la respuesta del INDEC", 16 de mayo [en línea] https://www.lanacion.com.ar/sociedad/a-casi-un-ano-del-censo-todavia-son-pocos-los-datos-que-se-conocen-la-respuesta-del-indec-nid16052023/.

[10] El documento *Manual de revisión de datos de los censos de población y vivienda. Revisión 1* es un interesante antecedente en este sentido, el cual incluye una breve sección dedicada a la 'duplicación de registros' en su sección de 'prácticas de revisión de datos' (Naciones Unidas, 2011).

Lindenboim, J. (2011), "Las estadísticas oficiales en Argentina. ¿Herramientas u obstáculos para las ciencias sociales?", *Trabajo y sociedad: indagaciones sobre el empleo, la cultura y las prácticas políticas en sociedades segmentadas*, vol. 16.

Marshall, L. (2008), "Potential duplicates in the census: methodology and selection of cases for follow up", *Proceedings of the Section on Survey Research Methods*, vol. 1.

Martín-Guzmán, P. (2016), "Desafíos actuales en la estadística oficial", *Estudios de Economía Aplicada*, vol. 34, Nº 3.

Molinatti, F. (2017), "Las migraciones internas en Argentina: posibilidades, alcances y desafíos para su captación mediante el Censo de 2010", *XIV Jornadas Argentinas de Estudios de Población*, Santa Fe.

Naciones Unidas (2013), *Principios Fundamentales de las Estadísticas Oficiales, 1994-2013,* Nueva York, Comisión de Estadística.

___(2011), *Manual de revisión de datos de los censos de población y vivienda. Revisión 1,* Nueva York, Departamento de Asuntos Económicos y Sociales (DAES).

Poledo, M. y L. García (2022), "Lecciones aprendidas del Censo Nacional de Población, Hogares y Viviendas 2022 realizado en la República Argentina", documento presentado en la Conferencia sobre Estadísticas Europeas. Grupo de Expertos sobre Censos de Población y Viviendas. Vigesimocuarta Reunión. Ginebra, 21 al 23 de septiembre.

Rodríguez, G. (2022), *Nueva revisión de la cartografía censal del INDEC de Argentina, años 1991, 2001 y 2010*, Buenos Aires, Centro de Estudios Urbanos y Regionales (CEUR).

___(2021), "Comparabilidad retrospectiva en la cartografía censal digital del INDEC. Estado actual, avances y desafíos en Argentina y la Ciudad de Buenos Aires", *Población de Buenos Aires*, vol. 18, Nº 30.

Villa Diharce, E. (2004), "El número efectivo de grados de libertad", *Simposio de Metrología*, Centro de Investigación en Matemáticas, 25 al 27 de octubre.