Verificación de datos abiertos sobre tenencia de hogares publicados por el Ministerio de Desarrollo Social del Uruguay

Integración de Datos 2020
Docentes: Regina Motz - Edelweis Rohrer

Grupo 12
Sara Azpiroz CI: 5.265.611-6
Nicolás Nocetti CI: 3.095.751-6

Contenido

Objetivo

El objetivo que planteamos para el trabajo del curso consiste en verificar la consistencia de varios conjuntos de datos que presentan una temática relacionada y fueron publicados de manera independiente.

De esta manera, partimos de la Encuesta Continua de Hogares (ECH) publicada por el Instituto Nacional de Estadística del Uruguay (INE) y a través de la misma, tratamos de verificar varios conjuntos de datos publicados por el Ministerio de Desarrollo Social del Uruguay (MIDES), en relación a la tenencia de la vivienda en el Uruguay.

Objetivos secundarios

Como objetivo secundario, nos proponemos publicar los resultados obtenidos como datos abiertos y experimentar en la utilización del estándar RDF.

Motivación

Una de los beneficios principales que tiene el proyecto, es que la integración resultante de la comparación de datos no sólo será para consumo interno, sino que al ser publicada podrá ser accesible por todos quienes tengan interés.

En este sentido, aquellos quienes hacen seguimiento de los conjuntos de datos publicados por el MIDES tendrán una métrica sobre cuán precisos son estos en comparación con los publicados por el INE sobre la ECH.

Además, trabajar con estos datos significará para nosotros conocer un dominio nuevo (indicadores y su cálculo, terminología, catálogo de datos abiertos) al mismo tiempo que tenemos un primer acercamiento sobre los estándares para la publicación de datos abiertos, cosa con la que no hemos trabajado antes.

Fuentes

Una de las fuentes principales para el proyecto son conjuntos de datos publicados por el MIDES en el portal del Catálogo de Datos Abiertos [1], dentro de la categoría *Desarrollo Social*. Por otra parte, del portal del INE [2] se obtuvieron los microdatos anonimizados de la ECH para los años 2006 a 2018. Destacamos que los datos publicados por el MIDES tienen como fuente los datos de la ECH.

Algo a destacar de los conjuntos a trabajar, es que todos ellos han sido publicados como datos abiertos, por lo tanto entre otras cosas, constan de diccionario de datos. Contar con diccionario de datos es sumamente relevante para entender el dominio y sus entidades.

Ministerio de Desarrollo Social (MIDES)

El MIDES es uno de los principales publicadores en la categoría *Desarollo Social* del Catálogo de Datos con 212 publicaciones, siguiéndole AGESIC con 4 publicaciones. De estas 212 publicaciones de datos se eligieron aquellos que relacionan hogares con su situación de tenencia, nivel de ingresos, entre otros.

Los conjuntos son muy similares en cuanto a formato. Tienen 4 atributos, donde uno de ellos es un valor agregado. Además, todos ellos se corresponden con número de indicador que utiliza el MIDES. A continuación, detallamos los conjuntos elegidos y damos una breve descripción de sus atributos.

- C1: *Relación cuota de compra o cuota alquiler de la vivienda e ingresos del hogar según departamento. Total país* [3]. Indicador 14093.

Descripción de metadatos:
- Relación Compra Alquiler: Describe si la relación es compra o alquiler.
- Departamento: El departamento, dentro del Uruguay, asociado a la información.
- Año: El año correspondiente a la información
- Valor: El porcentaje de compra o alquiler para el departamento en el año indicado.

- C2: *Distribución porcentual de las personas según régimen de tenencia de la vivienda por tramos de edad. Total país* [4]. Indicador 12533.

Descripción de metadatos:
- Edad: Franja de edad en la que se basa la información.
- Tenencia De Vivienda: Relación del sujeto sobre la vivienda que ocupa.
- Año: El año correspondiente a la información.
- Valor: Porcentaje de personas en la franja de edad indicada con la tenencia de vivienda indicada en el año indicado.

- C3: *Distribución porcentual de los hogares según régimen de tenencia por tipo de hogar. Total país* [5]. Indicador 12471.

Descripción de metadatos:
- Tipo de Hogar: Composición del núcleo familiar del hogar.
- Tenencia De Vivienda: Relación del hogar sobre la vivienda que ocupa.
- Año: El año correspondiente a la información.
- Valor: Porcentaje de hogares con la tenencia de vivienda indicada en el año indicado.

- C4: *Porcentaje de hogares que destinan más del 30% de sus ingresos al pago de la cuota de compra o de alquiler de la vivienda según quintiles de ingreso per cápita del hogar. Total país* [6]. Indicador 7787.

Descripción de los metadatos:
- Compra_alquiler: Describe si la relación es compra o alquiler.
- Quintiles: Quintil al cual pertenece el hogar.

- Año: El año correspondiente a la información.
- Valor: Porcentaje de hogares que destinan más del 30% de sus ingresos al pago de la cuota en el quintil y año indicado.

■ C5. *Distribución porcentual de los hogares según régimen de tenencia por quintiles de ingreso per cápita del hogar. Total país* [7]. Indicador 7809.

Descripción de los metadatos:
- Quintiles: Quintil del hogar.
- Tenencia De Vivienda: Relación del sujeto sobre la vivienda que ocupa.
- Año: El año correspondiente a la información.
- Valor: Porcentaje de hogares con la tenencia de vivienda indicada, en el quintil y año indicado.

Instituto Nacional de Estadística (INE)

El INE realiza de forma ininterrumpida desde el año 1968 la ECH. Esta brinda indicadores oficiales del mercado laboral y de ingresos de los hogares. Además constituye una base de estudios de variadas temáticas, entre ellas: salud, educación y condiciones de la vivienda.

De esta manera, el MIDES obtiene de la ECH datos sobre tenencia de vivienda, y publica agregaciones de los mismos en el portal del Catálogo de Datos abiertos. Mencionamos que obtuvimos los datos de la ECH desde el sitio del INE, en la sección destinada para la ECH [8]. Allí se encuentran publicados resultados de la misma desde el año 1990 a 2019, pero nosotros trabajaremos con los datos restringidos a los años 2006 a 2018, ya que son los que se encuentran en los conjuntos de datos del MIDES.

Abordaje de integración seguido

Dados por un lado los conjuntos de datos del MIDES y la ECH, el plan consiste en realizar una correspondencia entre los atributos presentes en los conjuntos del MIDES y las variables presentes en la ECH. Luego, realizar los cálculos sobre las variables de la ECH para obtener el dato agregado que se muestra en los datos del MIDES, para luego realizar una comparación.

Esto hará que para cada conjunto de datos del MIDES se genere un nuevo conjunto que contendrá:
- Los *atributos* del conjunto original del MIDES.
- El *valor calculado* a partir de los datos la ECH que corresponde con el valor agregado en el conjunto del MIDES. Este valor calculado se obtiene realizando un mapeo de los atributos del conjunto del MIDES a los atributos de la ECH.
- El valor *porcentaje de diferencia* que indicará la diferencia entre el valor publicado por el MIDES y el calculado por nosotros.

En este sentido, el abordaje de integración se enfoca en enlazar instancias. Finalmente, el conjunto resultado será publicado como datos abiertos.

Avance del proyecto y problemas encontrados

En esta sección describiremos el avance del proyecto según el abordaje de integración definido. En términos generales, originalmente nuestra idea era trabajar únicamente con los conjuntos C1 a C5, realizando la verificación entre ellos, agregando por algún atributo. Como esto no fue posible, posteriormente se agregó el conjunto de datos proveniente de la ECH, moviéndonos al objetivo de comparar los conjuntos C1 a C5 contra la ECH.

Etapa 1: Análisis de correspondencia semántica en los conjuntos C1 a C5

Los conjuntos C1 a C5 fueron elegidos del portal del Catálogo de datos por poseer una temática de interés. En principio no fueron elegidos con un objetivo concreto, sino que el objetivo lo definimos posteriormente. Por lo tanto, en esta primera etapa el principal problema al que nos encontramos fue: Dados unos conjuntos de datos, entender su dominio y buscar la existencia de relaciones entre ellos.

Para esto, comenzamos a estudiar los conjunto y su diccionario de datos. De esta manera intentamos encontrar relaciones entre ellos.

Allí nos percatamos de que si bien comparten varios atributos, no existía equivalencia semántica entre ellos, lo que imposibilitaba la integración de los mismos.
Con esto, toda intención de integrar algún atributo en los conjuntos y comparar resultados, queda descartada. Este problema no fue detectado inicialmente.

Etapa 2: Incorporación de un nuevo conjunto de datos - ECH

Como hemos mencionado antes, en principio se buscó relacionar únicamente los conjuntos C1 a C5, agregando según diferentes atributos y comparando los valores resultantes, pero mediante un análisis posterior nos dimos cuenta que eso no era posible. Entre otras cosas, el principal problema está en la imposibilidad de relacionar los conjuntos C1 a C5 entre sí dado que los datos publicados ya se encuentran agregados.

Buscando cómo sortear este problema intentamos encontrar un conjunto que no tuviera los datos agregados y nos encontramos con la ECH del INE, que hemos mencionado anteriormente. Estudiamos el diccionario de datos y vimos que basándonos en ella podríamos verificar al menos alguno de los conjuntos de datos iniciales.

Además, de cierta manera la ECH nos permitiría realizar un trabajo de ingeniería inversa: Agregar los datos presentados en la ECH y comparar con los publicados por el MIDES. De hecho, éstos últimos se originan de los primeros.

Así, comenzamos a verificar la posibilidad de verificar los datos obtenidos en los conjuntos a través de la ECH.

Estudiando con mayor profundidad la equivalencia semántica nos dimos cuenta que C1, C3, C4 y C5 tienen como unidad de análisis el hogar, mientras que C2 toma como unidad la persona. De esta

manera, si bien en principio parecía que C2 estaba relacionado con los demás conjuntos, no es posible integrarlo con la ECH, ya que esta se basa en los hogares.

Luego estudiamos el conjunto C3, en este se presenta el porcentaje de hogares, en función de la composición de los mismos y su relación con la vivienda, si son inquilinos o propietarios. Lamentablemente no pudimos obtener la composición del hogar a partir de la ECH por lo que este conjunto también quedó descartado para el análisis.

Seguidamente, estudiamos si era posible verificar los conjuntos C4 y C5. Estos basan sus análisis en el quintil al que pertenecen los hogares, intentamos obtener los quintiles a partir del ECH, pero no logramos encontrar una forma que fuera viable, por lo que tuvimos que descartar la verificación de los mismos.

Finalmente, optamos realizar el trabajo de verificación del conjunto C1, para el cual es clara su relación con la ECH.

Etapa 3: Mapeo de atributos de C1 con ECH y primeros resultados

Con más conocimiento sobre el dominio de los conjuntos y con una idea más definida de lo que buscábamos, partimos del conjunto C1 e intentamos mapear los atributos de este en la ECH del año 2018. De esta manera, buscamos obtener en ECH el conjunto equivalente a C1, es decir, *Relación cuota de compra o cuota alquiler de la vivienda e ingresos del hogar según departamento. Total país.*

Para realizar el mapeo nos basamos en la definición del diccionario de datos de ambos conjuntos. Mencionamos que la ECH maneja más de 100 descripciones de variables, agrupadas en categorías, donde en particular dentro de la categoría *Hogar* se tiene *Tenencia de vivienda*. Con el propósito de reflejar claridad, en la Tabla 1 presentamos una extracción del diccionario de la ECH donde se detalla la variable mencionada.

DESCRIPCIÓN DE LA VARIABLE	ECH 2018		
	NOMBRE VARIABLE	CATEGORÍAS	
		CÓDIGO	DESCRIPCIÓN Y OBSERVACIONES
D. HOGAR			
TENENCIA DE LA VIVIENDA	d8_1	1	Propietario/a de la vivienda y el terreno y los está pagando
		2	Propietario/a de la vivienda y el terreno y ya los pagó
		3	Propietario/a solamente de la vivienda y la está pagando
		4	Propietario/a solamente de la vivienda y ya la pagó
		5	Inquilino/a o arrendatario/a de la vivienda
		6	Ocupante con relación de dependencia
		7	Ocupante gratuito. Se lo permite el B.P.S.
		8	Ocupante gratuito. Se lo permite un particular
		9	Ocupante sin permiso del propietario/a
		10	Miembro de cooperativa de vivienda
	d8_2	$	Monto de la cuota de compra
	d8_3	$	Monto del alquiler (efectivamente pagado o estimado)

Tabla 1: Fragmento del diccionario de la ECH año 2018, para la variable *Tenencia de la vivienda*.

De esta manera, realizando todos los mapeos para los atributos en C1 en la ECH llegamos al resultado que presentamos en la Tabla 2. En particular, hay dos columnas para la ECH dado que el diccionario del año 2006 es diferente al de los demás.

Finalmente, materializamos las encuestas de todos los años de interés de la ECH en una misma base de datos y mediante una consulta SQL, generamos el conjunto resultado, donde en particular se obtiene el valor que se muestra agregado en C1. En la Tabla 3 mostramos un fragmento de los resultados obtenidos restringidos a la relación *Alquiler* y al año 2018, y en el Anexo adjuntamos la consulta SQL utilizada.

Por otra parte, es importante mencionar los atributos que componen el conjunto resultado:

- *RelaciónCompraAlquiler*, *Departamento* y *Año* son los mismos que los encontrados en C1.
- *PROMEDIO(RAI/RCI)-MIDES* es el correspondiente al llamado *Valor* en C1. Destacamos que dicho valor corresponde al indicador RAI[1] o al indicador RCI[2] según el caso.
- *PROMEDIO(RAI/RCI)-ECH* es el valor calculado por nosotros.
- *PorcentajeDiferencia* es el porcentaje de diferencia de los dos valores anteriores, calculado según la fórmula:

$$\frac{|\,PROMEDIO(RAI/RCI)ECH - PROMEDIO(RAI/RCI)MIDES\,|}{|\,PROMEDIO(RAI/RCI)MIDES\,|} \times 100\%$$

[1] El Ratio Alquiler Ingreso (rental-to-income ratio, RAI), es una medición del porcentaje del ingreso que los hogares inquilinos destinan al pago del alquiler. [9]
[2] El Ratio Cuota Ingreso (mortgage-to-income ratio, RCI) es una medición del porcentaje del ingreso que destinan al pago de la cuota los hogares que compraron su vivienda con un préstamo. [9]

C1		ECH 2006		ECH 2007-2018		
Atributo	Valor	Atributo		Descripción	Valor	Descripción
Relación Compra Alquiler	Alquiler	d7_1	d8_1	Tenencia de la vivienda	5	Inquilino/a o arrendatario/a de la vivienda
	Compra				1	Propietario/a de la vivienda y el terreno y los está pagando
					3	Propietario/a solamente de la vivienda y la está pagando
departa-mento	Montevideo	DPTO	DPTO	Departamento	1	Montevideo
	Artigas				2	Artigas
	Canelones				3	Canelones
	Cerro Largo				4	Cerro Largo
	Colonia				5	Colonia
	Durazno				6	Durazno
	Flores				7	Flores
	Florida				8	Florida
	Lavalleja				9	Lavalleja
	Maldonado				10	Maldonado
	Paysandú				11	Paysandú
	Rio Negro				12	Rio Negro
	Rivera				13	Rivera
	Rocha				14	Rocha
	Salto				15	Salto
	San Jose				16	San Jose
	Soriano				17	Soriano
	Tacuarembó				18	Tacuarembó
	Treinta y Tres				19	Treinta y Tres
año	<2006 a 2018>	AÑO	AÑO		<2006 a 2018>	
valor	<Float>	d7_2	d8_2	Monto de la cuota de compra	SI d8_1 == 5 → X= d8_3 SINO X= d8_2 AVG((X/YSVL)*100) AGRUPADO POR DPTO	El valor se calcula como el promedio por departamento del monto de la cuota compra/alquiler dividido sobre el monto total de los ingresos del hogar, multiplicado por cien.
		d7_3	d8_3	Monto del alquiler (efectivamente pagado o estimado)		
		YSVL	YSVL	Ingreso total del hogar sin valor locativo sin servicio doméstico		

Tabla 2: Mapeo de valores de atributos de C1, con las variables de la ECH.

Relación Compra Alquiler	Departamento	Año	PROMEDIO(RAI/RCI) - MIDES	PROMEDIO(RAI/RCI) - ECH	Porcentaje Diferencia
Alquiler	Montevideo	2018	24.2	25.5	5.37
Alquiler	Artigas	2018	19.6	19.29	1.58
Alquiler	Canelones	2018	20.5	20.38	0.59
Alquiler	Cerro Largo	2018	19.2	19.65	2.34
Alquiler	Colonia	2018	18.5	18.59	0.49
Alquiler	Durazno	2018	18.1	18.08	0.11
Alquiler	Flores	2018	18.3	16.98	7.21
Alquiler	Florida	2018	17.1	16.66	2.57
Alquiler	Lavalleja	2018	19.5	18.73	3.95
Alquiler	Maldonado	2018	21	22.31	6.24
Alquiler	Paysandú	2018	17.8	18.02	1.24
Alquiler	Rio Negro	2018	18.3	16.12	11.91
Alquiler	Rivera	2018	19.5	19.97	2.41
Alquiler	Rocha	2018	16.8	17.02	1.31
Alquiler	Salto	2018	18.2	18.17	0.16
Alquiler	San José	2018	17.3	17.81	2.95
Alquiler	Soriano	2018	16.8	17.37	3.39
Alquiler	Tacuarembó	2018	19.2	19.7	2.6
Alquiler	Treinta y Tres	2018	18	18.3	1.67

Tabla 3: Fragmento de resultados obtenidos luego de la integración de C1 con ECH.

Observando los resultados obtenidos que se muestran en la Tabla 3, notamos que son bastante aproximados en general. Existen diferencias importantes para algunos casos, como Río Negro y Flores, pero no sabemos con exactitud a qué puede deberse.

Comentamos que en el conjunto de datos completo, el promedio del porcentaje de diferencia es 6.11 (un valor que consideramos bajo), el valor mínimo es 0.00 (lo cuál indica que hay casos para los que el valor publicado por el MIDES y el calculado por nosotros coincide) y el valor máximo es 264.78.

Las diferencias pequeñas, de hasta 7% aproximadamente, se pueden deber a que en los datos publicados por el MIDES se deflactaron los ingresos y los precios de alquileres y cuotas de pago, para uniformizar el valor de la moneda. Para los casos en los que el error es grande, se debería realizar un estudio mayor y conocer los detalles del método de cálculo empleado por el MIDES.

Etapa 4: Publicación del conjunto resultado como datos abiertos

Para la publicación del conjunto resultado (integración de C1 con la ECH) seguimos la *Guía para la apertura y publicación de datos abiertos de Gobierno* [10], publicada por AGESIC. Allí mencionan

que existe una categorización del grado de apertura de los datos basado en estrellas, respecto a qué tan abiertos y usables son los datos ofrecidos.

De esta manera, si conseguimos que los datos estén accesibles en la web, bajo una licencia abierta y en un formato no propietario (como .csv o .xml), obtenemos la categoría de 3 estrellas. Si adicionalmente se utiliza un estándar como RDF, se obtiene la categoría de 4 estrellas. Además, AGESIC hace una serie de recomendaciones sobre los datos publicados, como por ejemplo, incluir metadatos.

Por lo tanto, decidimos publicar los datos mediante un repositorio público en *GitHub* [12], bajo una licencia GPLv3 [11] auto-generada. Además dado que los datos fuente son datos abiertos y tienen una licencia de tipo DAG de Uruguay, lo único necesario fue mencionar las fuentes y la licencia DAG, lo cual hicimos en el README del proyecto GitHub.

Finalmente, los datos fueron publicados en formato csv y xml. Además, publicamos un diccionario que describe los atributos del conjunto, formas de cálculo de valores y fuentes utilizadas.

Si bien no alcanzamos a publicar el conjunto resultado en formato RDF, a fin de experimentar en la utilización del estándar, logramos traducir el conjunto C1 del MIDES y lo publicamos.

Para la publicación del conjunto C1 siguiendo el estándar RDF, definimos el grafo que se muestra en la Figura 1.

En la misma el blank node *_:registroRAI* representa los registros de RAI para un departamento y año dados. Para representar esto, se realizó una reificación, el *_:registroRAI* es una sentencia que indica que para un departamento el promedio de RAI es un valor literal, y esa sentencia se realiza en un determinado año, también representado con un valor literal.

De forma análoga, el blank node *_:registroRCI* representa los registros de RCI para un departamento y año dados utilizando una reificación similar a la anterior.

Mencionamos que pudimos cargar la definición del grafo en *GraphDB* [13] localmente, pero no conseguimos exportar el enlace. Por lo tanto dejamos el archivo en formato N-triples en GitHub.

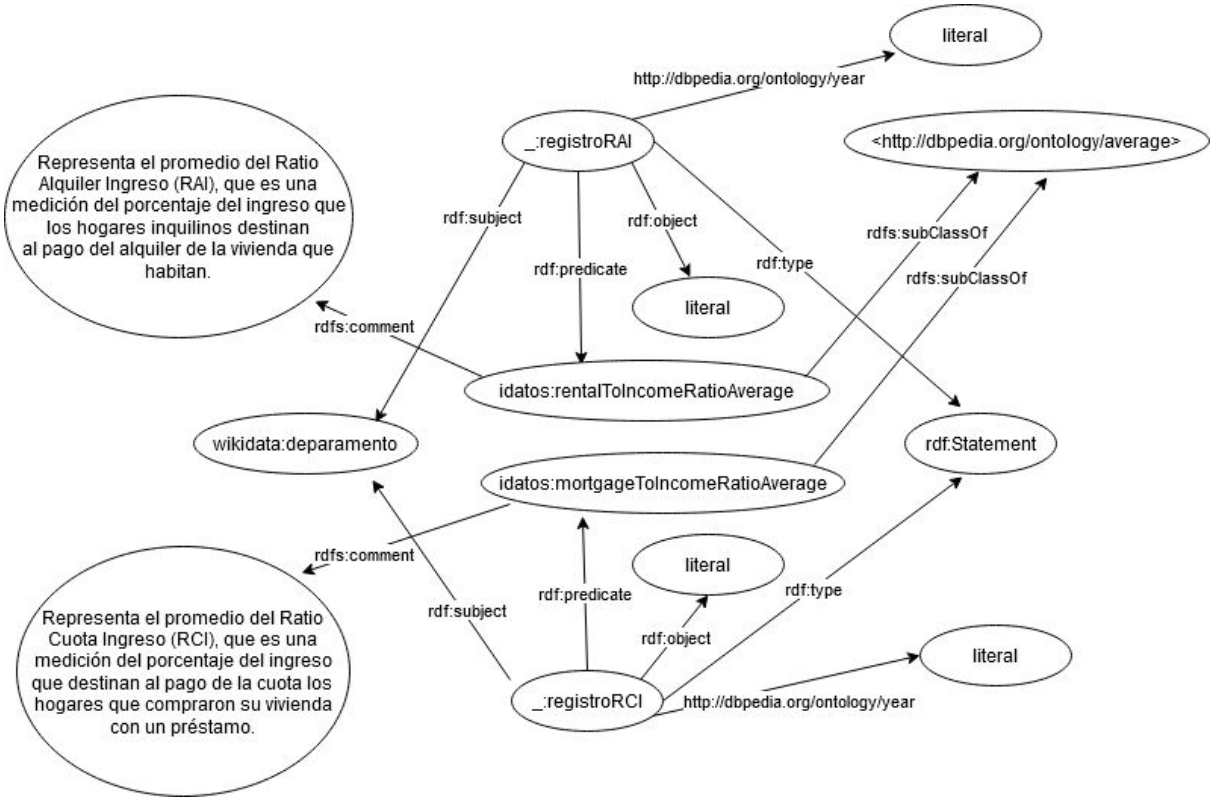

Figura 1: Grafo RDF

Dimensiones de calidad

Como hemos venido mencionado, los datos con los que trabajamos presentaron los siguientes beneficios:

- Fueron publicados como datos abiertos, bajo una licencia abierta, lo cual nos permitió publicar nuestro trabajo sin restricciones normativas.
- Contaban con diccionario de datos, lo que permitió entender la semántica de los atributos y la forma de cálculo de los valores mostrados.
- En el caso de los publicados por el MIDES, referenciaba a la ECH, lo cual facilita su trazabilidad.

Sin embargo, detectamos los siguientes inconvenientes para el caso de los conjuntos del MIDES:

- Pese a ser publicados como datos abiertos, están bastante agregados, lo cual de cierta manera incumple con las buenas prácticas que recomiendan publicar los datos lo más *crudo* posible.
- No encontramos grandes problemas en cuanto a la correctitud sintáctica o semántica. Lo que sí notamos, es que llaman distinto a atributos que significan lo mismo (por ejemplo *Compra_alquiler* y *RelaciónCompraAlquiler*).
- Además, para los valores calculados utilizan como nombre de atributo *valor*, y creemos que sería mejor utilizar un nombre más significativo, mencionando sobre qué indicador universal está basado el cálculo (por ejemplo, para el caso de C1, llamar *valor* como el *promedio(RAI/RCI) por departamento*)
- Luego de realizada la comparación con los valores que calculamos a partir de la ECH, si bien el promedio del porcentaje de diferencia es 6.11 (un valor que consideramos bajo), no nos

resulta trivial entender a qué se debe. Consideramos una causa de esta diferencia la deflactación de los valores monetarios, pero no podemos asegurarlo. Aun así esto no podría explicar aquellos casos que presentaron diferencias porcentuales mayores. En este sentido, si el MIDES tomó para los cálculos otras variantes sería bueno que sea aclarado, así como si se deflactaron los valores monetarios. A fin de que sea posible reproducir el mismo resultado.
- La ECH tiene datos desde 1990, sin embargo en los diferentes conjuntos del MIDES solo están los datos desde 2006 a 2018. por lo que los datos del MIDES no son completos, y al faltar el año 2019, tampoco están frescos.

Para concluir esta sección, destacamos lo siguiente sobre la ECH:
- Habían conjuntos de datos que manejaban diferentes diccionarios de datos (por ejemplo el diccionario de 2006 es diferente al diccionario de 2007 a 2018), por lo que al migrar a un nuevo diccionario sería bueno actualizar los formatos anteriores, a fin de mantener la consistencia.
- De todas maneras, los datos publicados son completos y mantienen una granularidad fina. No se encuentran agregados lo que aporta un valor mayor. Además, se encuentran publicados en diferentes formatos, tanto abiertos (.dbf y .dat) , como propietarios (.sav).

Resultado final

Como mencionamos en la sección *Avance del Proyecto - Etapa 4*, la implementación final consiste en la publicación del conjunto integrado, siguiendo las normas para la publicación de datos abiertos. En este sentido, el proyecto está visible en la web, cuenta con una licencia abierta, y está publicado en diferentes formatos (.xml y .csv). Por otra parte, se publicó el conjunto C1 original del MIDES siguiendo el estándar RDF (en formato .nt). Además se incluyeron metadatos y una imagen del grafo RDF para facilitar su comprensión.

Se puede acceder al proyecto mediante el link: https://github.com/nnocetti/IDatos

Mencionamos que el conjunto resultado mantiene la misma información que el conjunto original C1 (mismos atributos y el mismo nivel de agregación), pero lo extendimos agregando el dato calculado a partir de la ECH, y el porcentaje de diferencia. Por lo tanto, podemos decir que mantiene la granularidad. Además, tanto en los metadatos como en el README del proyecto, se referencia a las fuentes originales, por lo que promovemos la trazabilidad de los datos con los que trabajamos

Destacamos además, que el abordaje consistió en trabajar con los datos materializados, descargando los mismos desde las fuentes y cargandolos en una base de datos. Además, el modelo común de integración fue uno relacional, dado que dicha carga de datos se hizo a una base de datos relacional. Para todo esto se utilizó la herramienta de *Microsoft Access* [14].

En cuanto a la mantenibilidad, dado que los datos con los que trabajamos tienen la característica de ser históricos (es decir, son *fotos* de la realidad tomadas en diferentes años), los datos publicados no van a cambiar, pero sí será necesario actualizar el conjunto añadiendo información nueva (es decir, agregando nuevas *fotos*).

Conclusiones

En este proyecto partimos de conjuntos de datos que fueron elegidos por tener una temática relacionada y que nos pareció interesante, y luego mediante análisis posteriores logramos establecer un objetivo concreto.

Durante ese proceso nos dimos cuenta que no es trivial el dominio de los datos, el cuál nos llevó su tiempo comprender. Además, comprobamos que no siempre que se tienen conjuntos con temáticas similares, se pueden realizar agregaciones para comparar. En nuestro caso, vimos que si bien partimos de conjuntos que parecían fáciles de relacionar y verificar, no hay que desestimar la forma en la que éstos se presentan, dado que el hecho de que contengan datos agregados imposibilita cualquier agregación posterior.

Por otro lado, en este proyecto tuvimos un primer acercamiento al trabajo con datos abiertos y a la publicación de los mismos, así como tambíen al estándar RDF. Los datos abiertos son muy útiles por su potencialidad de ser procesados por máquinas, pero no quita que siempre será necesario un análisis humano previo. En este sentido es de gran utilidad el uso de diccionarios y referencias a las fuentes utilizadas para clarificar cualquier incertidumbre.

Como trabajo a futuro, sería interesante por un lado estudiar en detalle las causas que originan las diferencias encontradas, y además extender la verificación para aquellos conjuntos que fueron planteados inicialmente, los cuales no llegamos a verificar. También sería útil mostrar los resultados obtenidos de una manera más amigable, por ejemplo a través de gráficas.

Referencias

[1] AGESIC. (s. f.). Catálogo de Datos Abiertos. Catálogo de Datos Abiertos. Recuperado 5 de diciembre de 2020, de https://catalogodatos.gub.uy/

[2] INE. (s. f.). *Instituto Nacional de Estadística del Uruguay*. Recuperado 5 de diciembre de 2020, de https://www.ine.gub.uy/inicio

[3] MIDES. (2020, 23 mayo). *Relación cuota de compra o cuota alquiler de la vivienda e ingresos del hogar según departamento. Total país* [Conjunto de datos]. Catálogo de Datos Abiertos. https://catalogodatos.gub.uy/dataset/mides-indicador-14093

[4] MIDES. (2020, 1 mayo). *Distribución porcentual de las personas según régimen de tenencia de la vivienda por tramos de edad. Total país.* [Conjunto de datos]. Catálogo de Datos Abiertos. https://catalogodatos.gub.uy/dataset/mides-indicador-12533

[5] MIDES. (2020, 20 mayo). *Distribución porcentual de los hogares según régimen de tenencia por tipo de hogar. Total país* [Conjunto de datos]. Catálogo de Datos Abiertos. https://catalogodatos.gub.uy/dataset/mides-indicador-12471

[6] MIDES. (2020, 20 mayo). *Porcentaje de hogares que destinan más del 30% de sus ingresos al pago de la cuota de compra o de alquiler de la vivienda según quintiles de ingreso per cápita del hogar. Total país* [Conjunto de datos]. Catálogo de Datos Abiertos. https://catalogodatos.gub.uy/dataset/mides-indicador-7787

[7] MIDES. (2020, 20 mayo). *Distribución porcentual de los hogares según régimen de tenencia por quintiles de ingreso per cáptia del hogar. Total país* [Conjunto de datos]. Catálogo de Datos Abiertos. https://catalogodatos.gub.uy/dataset/mides-indicador-7809

[8] INE. (1990–2019). *Encuesta Continua de Hogares* [Conjunto de datos]. Sitio web del INE. https://www.ine.gub.uy/web/guest/encuesta-continua-de-hogares1

[9] ANV. (2013). *Informe Mercado Inmobiliario. ÁREA FINANCIAMIENTO Y MERCADO INMOBILIARIO*. https://www.anv.gub.uy/sites/default/files/2019-10/InformeMercadoInmobiliario_20131016.pdf

[10] AGESIC. (2018). *Guía para la apertura y publicación de datos abiertos de Gobierno*. Buenas Prácticas. https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/sites/agencia-gobierno-electronico-sociedad-informacion-conocimiento/files/documentos/publicaciones/guia_de_apertura_y_publicacion_diseno.pdf

[11] Free Software Foundation, Inc. (2007, 29 junio). *GNU General Public License*. GNU Operating System. Recuperado 6 de diciembre de 2020, de https://www.gnu.org/licenses/gpl-3.0.html

[12] GitHub, Inc. (2020). *GitHub*. GitHub: Where the world builds software. Recuperado 6 de diciembre de 2020, de https://github.com/

[13] Ontotext. (s. f.). *GraphDB*. GraphDB Downloads and Resources. Recuperado 6 de diciembre de 2020, de https://graphdb.ontotext.com/

[14] Microsoft. (2020). *Microsoft Access*. Software y aplicaciones de base de datos. Recuperado 6 de diciembre de 2020, de https://www.microsoft.com/es/microsoft-365/access

Anexo

Consulta SQL para la generación del conjunto resultado

```
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2019 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2019_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2019 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2019_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2018 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2018_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2018 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2018_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2017 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2017_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2017 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2017_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2016 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2016_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2016 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2016_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2015 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2015_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2015 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2015_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2014 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2014_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2014 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2014_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;
```

```
UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2013 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2013_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2013 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2013_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2012 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2012_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2012 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2012_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2011 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2011_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2011 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2011_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2010 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2010_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2010 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2010_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2009 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2009_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2009 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2009_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2008 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2008_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2008 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2008_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2007 AS Año, AVG((D8_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2007_Terceros
WHERE D8_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
```

```
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2007 AS Año, AVG((D8_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2007_Terceros
WHERE (D8_1=1 OR D8_1=3) AND (YSVL <> 0)
GROUP BY Departamento;

UNION
SELECT 'Alquiler' AS RelaciónCompraAlquiler, Departamento, 2006 AS Año, AVG((D7_3/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2006_Terceros
WHERE D7_1=5 AND YSVL <> 0
GROUP BY Departamento
UNION
SELECT 'Compra' AS RelaciónCompraAlquiler, Departamento, 2006 AS Año, AVG((D7_2/YSVL)*100) AS PROMEDIO(RAI/RCI)-ECH
FROM H_2006_Terceros
WHERE (D7_1=1 OR D7_1=3) AND (YSVL <> 0)
GROUP BY Departamento;
```