Universidad de San Andrés

Departamento de Economía

Licenciatura en Economía

'Remastering economics?': evaluando la aplicabilidad del Web Scraping como fuente de datos para Latinoamérica

Autores: Orlandi, José Ignacio y Osovi Conti, Marcelo Nicolás

Legajo: 26141; 26145

Mentor: Sosa Escudero, Walter

Victoria, julio de 2018

Índice

I. Introducción

Los avances tecnológicos de las últimas décadas han representado, sin dudas, un cambio de paradigma en términos de percepción, comprensión e interacción con la realidad. Desde el más simple proceso de vinculación social hasta la más compleja de las predicciones algorítmicas tal como las conocemos actualmente son producto de la innovación, la evolución del conocimiento y de la ciencia en todos sus campos de acción.

Uno de los principales protagonistas comprendidos en este proceso ha sido el campo de lo computacional; con progresos exponenciales en materia de procesamiento y análisis de datos. Esta información, a su vez, se va tornando incrementalmente cuantiosa, en sintonía con la digitalización de la vida cotidiana: búsquedas en internet, uso de aplicaciones, interacción en redes sociales, transacciones bancarias. *Big Data*, masividad de datos, multiplicidad de observaciones, alta frecuencia de generación y recolección de información; distintas facetas de un mismo desafío: capitalizar el gran alcance de las fuentes de datos disponibles enfrentando distintos interrogantes -tanto nuevas incógnitas como la reinterpretación de antiguas preguntas -con esta nueva perspectiva, pero siempre contemplando las limitaciones y puntos débiles del análisis que de estas fuentes pudieran surgir.

Son diversos los autores que han discurrido acerca de esta materia, evaluando el potencial analítico que la *Big Data* trae aparejada y las oportunidades que tal riqueza en términos de información representa. Entre otros, Einav y Levin (2013) arguyen acerca de algunas virtudes fundamentales a considerar. En primer lugar, la disponibilidad en tiempo real de las fuentes de información, lo que puede ofrecer una ventaja a la hora de identificar patrones mientras ocurren. Por otro lado, la considerable magnitud de las bases de datos confeccionables, resolviendo esto problemas estadísticos relacionados con observaciones limitadas, además de aumentar la precisión y el *power* de los resultados obtenidos. También, los autores encuentran que la *Big Data* permite acceder a datos que envuelven el comportamiento humano, antes difíciles de obtener. Estas y otras ventajas son las que muestran el verdadero potencial de este método de recolección y análisis de datos.

En esta línea, el presente trabajo intenta valerse de esta incipiente pero rica literatura -y las técnicas asociadas a ella- para analizar cuestiones económicamente relevantes. En

particular, buscaremos profundizar en el estudio de la Paridad de Poder Adquisitivo (PPP) de las principales economías latinoamericanas[1]. A tal fin, procederemos a confeccionar una base de datos formada enteramente por precios *online* -recolectados mediante la técnica de *Web Scraping*- para luego calcular la PPP de Brasil, Chile, Colombia y México; frente a tal valor para Argentina, nuestro país de referencia.

Es preciso aclarar que este trabajo encuentra ciertas similitudes con el estudio realizado por Alberto Cavallo, W. Erwin Diewert y Robert C.: *Using Online Prices for Measuring Real Consumption Across Countries*. En él, los autores calculan la PPP con una base de datos también conformada por precios *online*, intentando basarse en el trabajo realizado por el *International Comparisons Program* (ICP) del Banco Mundial. Tal base exhibe características que resultan ser ventajosas frente a aquella utilizada por la ICP, presentando una mayor frecuencia de recolección de los datos (diaria) e incrementando sustancialmente la transparencia de los métodos de recolección de la información. En cierta medida, la presente tesis busca replicar tal estudio, ampliando su abarcatura tanto en la cantidad de productos considerados como la diversidad de categorías[2] incluidas en el cálculo del índice de interés.

Por último, el trabajo exhibirá la siguiente organización. En primer lugar se revisará brevemente la literatura precedente, considerando estudios previos que implementan la técnica de *Web Scraping* y recolección algorítmica de datos. Luego, se presentará y detallará el proceso de confección de la base de datos y el índice de precios a utilizar. Finalmente, se exhibiran los resultados obtenidos, junto con las conclusiones resultantes.

[1] El presente estudio se basará en un análisis del poder adquisitivo de cinco países, a saber, Argentina, Brasil, Chile, Colombia y México. Tal elección está fundamentada en el hecho de que estas economías son las de mayor nivel de PBI dentro de América Latina -según el FMI, para el año 2018-. Además, y como consecuencia de esto, presentan una mayor escala de comercio *online*; lo que -intuimos- otorga más representatividad a los resultados obtenidos.

[2] Rubros de productos, según la COICOP. A desarrollar a continuación.

II. *Web Scraping* en Economía

Al ser el *Web Scraping* una técnica que se ha tornado popular en los últimos años, la literatura se caracteriza por ser, en su mayoría, fronteriza. Por ello, este método no presenta una vasta cantidad de autores que la hayan utilizado y es justamente a partir de esta situación, de donde surge la intención de comentar, aunque sea brevemente, su corta pero fructífera historia. De esta manera, en el siguiente apartado, intentaremos demostrar la vasta cantidad de temas donde podemos encontrar a la *Big Data*, a partir de su capacidad para crear bases de datos acorde a los propios intereses de los investigadores.

A inicios del siglo XXI, con la revolución del internet y el comienzo del comercio online (especialmente en EEUU), surgieron algunos estudios que comenzaron a tratar la llamada *new economy business*. Este fue el momento donde comenzó a surgir la *Big Data*. Como un ejemplo, podemos encontrar a Morton et. al (2001). Ellos estudiaron el efecto del *Internet car referral services* en el precio de concesionario de los automóviles en California. Las páginas que utilizaron para su estudio mostraban información detallada sobre los autos, incluyendo sus condiciones en el mercado de ese momento como también precios de facturación y contenido editorial. Estos datos, argumentaron los autores, eran mucho más precisos, detallados y comprensivos que aquellos presentados en publicaciones offline. Fue así cómo, Morton et. al consiguieron armar una base de datos a partir de la página web Autobytel.com durante el año 1999, de más dos millones de observaciones, que incluían información del cliente, el auto deseado, la fecha en que el pedido fue realizado, el comerciante al que Autobytel.com envió el *referral* y el tiempo en el que el consumidor se interesó por el vehículo. Podemos apreciar ya como hace ya más de una década, la *Big Data* comenzaba a mostrar su potencial, no solo en la cantidad de datos, sino también en la información en sí que podía recopilar.

Por otro lado, encontramos a Baye et. al (2005), con el escrito: *Did the Euro Foster Online Prices Competition? Evidence from an International Price Comparison Site*. En él, los autores realizan un estudio donde analizan el impacto de la introducción del Euro en precios online de diferentes *retailers*. Presentando datos tanto antes como después de la adopción de una única moneda, como también información para países pertenecientes y ajenos a la Unión Europea, los autores demuestran que, luego de controlar por costos, demanda y efectos por

la propia estructura del mercado, el Euro generó un aumento en el promedio de los precios en la Eurozona de entre un 3% y un 7%. Nuevamente, es una base de datos de precios online la que posibilita este estudio. Si bien no fue armada expresamente por los autores, la página que utilizan para adquirir los datos (Kelkoo) utiliza las técnicas de *Web Scraping* para la recolección de información.

Más recientemente, encontramos a Cavallo y Rigobon (2016): *The Billion Prices Project: Using Online Prices for Measurement and Research*, mostrando cómo la Big Data permite mejorar las estadísticas y la investigación empírica desarrollada por los economistas. En particular, los autores muestran cómo los precios online pueden ser utilizados para construir índices de precios con frecuencia diaria en múltiples países, evitando sesgos que distorsionan la evidencia sobre la rigidez de los precios y sus relatividades internacionales. Para dicho propósito, realizan un *scraping* de los precios de distintos productos de los *retailers* más importantes de cada país, emulando el IPC del respectivo país. Así encuentran que sus IPC online resultan ser muy parecidos a los desarrollados por las oficinas de estadísticas nacionales.

En la misma línea, Bertolotto (2016) utiliza una base de datos muy parecida a la de Cavallo y Rigobon (2016) con el objetivo de demostrar que la velocidad de la reversión de la media del tipo de cambio real (generalmente llamada *half life*) que utiliza precios agregados es un 38% más alta que aquella que utiliza información a nivel producto. Nuevamente, la técnica del *Web Scraping* hace posible crear una base de datos de precios a nivel producto con frecuencia diaria que permita responder a esta pregunta de investigación.

Como otro ejemplo, podemos observar el trabajo realizado por Cavallo et. al (2016): *Currency Unions, Prodcut Introductions and the Real Exchange Rate*. En este escrito, los autores utilizan una base de datos de precios online para productos idénticos vendidos por cuatro grandes *retailers* en docenas de países, con el objetivo de estudiar el ratio de tipo de cambio real a nivel producto y su comportamiento agregado. En su estudio, los autores encuentran que la ley de único precio para miles de productos se mantiene relativamente bien para países dentro de las uniones monetarias. Por el contrario, aquellos que no se encuentran bajo una misma moneda, no presentan esta característica, incluso si su tipo de cambio es fijo. También, luego de reestructurar la base de datos, logran demostrar que los tipos de cambio

reales a nivel producto presentan diferencias en el momento en que los productos son primeramente introducidos, en oposición a la creencia de que esta discrepancia se relaciona con rigideces nominales o traspasos de precios heterogéneos.

Asimismo, podemos observar el trabajo realizado por Patrick Lunnemann y Ladislav Wintr (2008): *Price stickiness in the US and in Europe revisited: Evidence from Internet Prices*. Como su título lo indica, los autores realizan un estudio en el que revisan la rigidez de los precios en Estados Unidos y Europa al investigar el comportamiento de los precios online. Así, los autores encuentran que, al contrario que los hallazgos para datos provenientes de IPC, los precios online no cambian menos seguido en países de Europa frente a EEUU. Además, muestran que los precios online no son necesariamente más flexibles que sus pares offline, como también que el cambio promedio en los precios online, si bien es relativamente grande, es menor que aquellos reportados por los ICP.

Para otro ejemplo, nos encontramos con Bertolotto y Aparicio (2017), con el texto: *Forecasting Inflation With Online Prices*. En este escrito, los autores introducen índices de precios online para pronosticar el Índice de Precios al Consumo, encontrando que este método logra anticipar tendencias en las inflaciones mensuales con más de un mes de anticipación, mostrando una notable mejoría frente a aquellos desarrollados por Bloomberg. Nuevamente, es una base de datos creada a partir de la *Web Scraping* la que permite realizar este estudio.

Por último, podemos considerar el texto un trabajo ya mencionado anteriormente: *Using Online Prices for Measuring Real Consumption Across Countries* de Alberto Cavallo, W. Erwin Diewert y Robert C. En él, los autores replican la paridad de poder de compra (PPP por sus siglas en inglés) desarrollada por la *International Comparisons Program* (ICP), pero con una base de datos armada mediante la técnica de *Web Scraping*. De esta manera, utilizando precios online, Cavallo et. al muestran que estas técnicas de armado de base de datos no solamente entregan resultados muy similares al índice confeccionado por la ICP, sino que pueden mejorar dramáticamente su frecuencia y transparencia.

Así, pudimos observar cómo una sola técnica de recolección de datos puede ser utilizada con muchos fines y objetivos. Dada su enorme capacidad para generar bases de diversas formas, que se acoplen a los deseos de los investigadores, es de suma importancia tomar este

método como una valiosa fuente de información en el futuro. Sin embargo, es menester aclarar que este método no se encuentra exento de problemas. Como veremos a lo largo de este trabajo, existen varios desafíos donde incluso la *Big Data* debe encontrar métodos tecnológicos para sortear sus problemas. A modo de ejemplo, el proceso de seleccionar (*matchear*) productos entre países continúa siendo una operación laboriosa, donde el trabajo mayormente es realizado a mano, dado los altísimos costos de entrada que presentan algunos programas que intentan, no con mucha precisión, realizar el trabajo automáticamente. Esto ocurre como una consecuencia de la poca estandarización de los productos, lo que dificulta su identificación.

III. Base de datos

Recolección de los precios

Dado nuestro objetivo, la base de datos a armar debe presentar una cualidad específica: los productos deben ser idénticos entre países. Esto permite que la comparación de precios, a la hora de armar el índice de la PPP sea válido y robusto. Asimismo, es menester contar con precios al contado, que representen el verdadero costo de adquirir el producto, y no uno financiado.

De esta manera, podemos diferenciar dos grandes formas a la hora de equiparar productos para encontrar aquellos que son idénticos: por un lado, el *matching* realizado mediante el ID de los bienes, y por el otro, el *matching* manual. Sin embargo, cabe destacar, ambos modos utilizan la misma técnica de *scraping*, explicada a continuación.

Este procedimiento consta de un bot que es configurado para que pueda leer el código fuente de las páginas y extraer la información de interés que se encuentra dentro de diferentes *tags*, como por ejemplo: 2.499. En este caso, el dato deseado es “2.499”, y dado que este formato se mantiene constantemente (variando solamente el precio) además de ser único (solo presenta esta forma cuando muestra el importe del producto), podemos comandar al bot para extraer la información requerida. Con este objetivo, el algoritmo busca en todo el código de la página, hasta encontrar “span class=”salesprice””. Cuando lo hace, retira el dato que se encuentra en formato texto (es decir, aquella información entre los signos mayor o menor), y así va repitiendo el proceso

hasta que obtuvo todos los precios que se presentan en el catálogo. Luego, solamente resta que la información sea almacenada en el formato deseado. Para obtener los ID, se realiza un proceso similar.

Posteriormente, todos aquellos productos obtenidos deben sean limpiados y *matcheados*, pero cabe destacar que esta última tarea puede representar un gran desafío. Cuando el *retailer* muestra una codificación uniforme de sus productos dentro de distintos países, el *matcheo* se convierte en un trabajo simple. Dado que el ID de un mismo producto se mantiene dentro de varios países, la búsqueda de ID idénticos no contiene error alguno, y efectivamente se encuentran productos iguales. Pero cuando esto no sucede (como en la mayoría de los casos), el paso de equiparar los mismos productos se convierte en un problema.

Para este caso, la información que uno buscará obtener de la página ya no serán los ID, sino el nombre de los productos. Y una vez que se obtienen estos datos (mediante el mismo proceso explicado anteriormente), se filtrará toda aquella información considerada irrelevante y se procederá a un *matcheo* manual de los bienes en cuestión. Por ejemplo, luego del *scraping* para la tienda oficial de HP en MercadoLibre para Argentina, obtenemos lo siguiente: "Impresora Hp M426fdw Fax Duplex Wifi Multifuncion Ex 425dn". Para el caso mexicano, extraemos: "Multifuncional hp laserjet pro m426fdw laser blanco y negro". Si bien ambas muestran y describen el mismo producto, la forma en que se las presenta es totalmente diferente[3]. De aquí que el match sea algo tan complejo. En este caso, se debe filtrar toda la información innecesaria, quedándonos solamente con el código: "M426fdw", ID que no se encuentra en el código fuente, imposibilitando el proceso descripto anteriormente.

Limitaciones

Consideramos que una base de este formato puede presentar una masividad de datos sorprendente. Sin embargo, no se encuentra exenta de limitaciones que reducen su alcance.

En primer lugar, como mencionamos anteriormente, la inexistencia de una codificación uniforme de los productos puede generar problemas severos, limitando la cantidad de datos que se puedan obtener, ya que el trabajo a mano es lento y arduo. Si bien existen programas

[3] Para países que poseen distintos idiomas, naturalmente la diferencia es aún mayor.

que permiten realizar un *match* utilizando los nombres de los productos, estos presentan un gran costo de entrada, y no necesariamente conllevan consigo una precisión muy elevada.

En segundo lugar, dado que esta forma de obtención de datos se basa en el comercio online, su aplicabilidad se ve limitada en países menos desarrollados, que no presentan demasiadas plataformas de este estilo. Al mismo tiempo, la canasta obtenida puede no ser, necesariamente, muy representativa del consumo del país en cuestión. Al no presentar precios de servicios, dado que sus apariciones en páginas online son poco frecuentes, la canasta obtenida se reduce principalmente a productos de consumo. En promedio, las seis categorías incluidas en este trabajo presentan entre el 7% y el 16% del consumo de cada país. Lamentablemente, hasta que una mayor cantidad de datos se encuentren disponibles online, este problema seguirá siendo un desafío.

Otro posible inconveniente puede presentarse al considerar que los precios online quizás difieran de sus pares offline. Existen trabajos que tratan este tema, como por ejemplo Cavallo (2017) donde el autor, recolectando precios tanto online como offline, para cincuenta y seis retailers en diez países: Argentina, Brasil, China, Estados Unidos, entre otros, logra demostrar que, si bien existe una dispersión, para el 72% de los casos comparados, el nivel de los precios en promedio es similar. Si bien Cavallo concluye que los precios online son una fuente representativa de los *retail prices*, incluso si la mayor parte del comercio es transaccionada offline, estas diferencias podrían afectar significativamente comparaciones a nivel precios en algunos países.

También, los precios online tienden a presentar un solo precio para todas las localidades de un país mientras que sus pares offline presentan, en general, una dispersión significativa regional. Por ejemplo, en áreas urbanas, los precios de la comida tienden a ser mayores. De todas formas, quizás estas diferencias pueden ser resueltas scrapeando negocios más localizados, aunque su aplicabilidad puede depender del país en cuestión. Seguramente, países más pobres tiendan a dificultar este accionar.

Por último, debemos mencionar que nuestra base de datos presenta una limitación temporal, como consecuencia de las fechas límites de entrega del trabajo, imposibilitando la

disponibilidad de precios distribuidos a lo largo del tiempo.[4] Sin embargo, cabe destacar, no solamente la rigidez propia de los precios en el corto plazo consideramos que juega a nuestro favor, sino que también, durante el tiempo en el que se realizó este estudio, objetos elegidos aleatoriamente no mostraron un cambio en sus precios, implicando que al menos por el lapso comprendido para la realización de este trabajo (Marzo y Abril del 2018), la PPP no debería haber variado signficativamente. Además, queremos enfatizar que no creemos que este problema le reste relevancia a nuestro aporte. Esta declaración se debe a que no solamente recolectamos información para países latinoamericanos mediante una técnica poco utilizada (que presenta ventajas ya mencionadas) sino que, además, pudimos aumentar las categorías de bienes cubiertas por el único trabajo que, hasta el momento, conocemos que haya realizado algo similar -Cavallo (2018)-.

Confección de la base de datos

Una vez finalizada la recolección de los precios, resta construir una base de datos que sea funcional al cómputo de la PPP de interés. A tal fin, se propondrá un esquema de categorización acorde a las demandas del índice de Fisher[5].

El índice en cuestión requiere agrupar los distintos productos en categorías, de las que -a su vez- se dispongan datos de ponderación en el consumo agregado de cada país. A esta demanda responden las clasificaciones aportadas por la COICOP (*Classification of individual consumption by purpose*). Estas categorías representan una clasificación de referencia publicada por la "División de Estadística de las Naciones Unidas" que separa la intención de los gastos de consumo individual incurridos por tres sectores, a saber, instituciones sin fines de lucro al servicio de los hogares, las viviendas y el gobierno.

De esta manera, la COICOP define tres niveles de estructura, crecientes en especificidad: divisiones (nivel uno), grupos (nivel dos) y clases (nivel tres). Por ejemplo, podemos encontrar:

[4] En general, las PPP se observan a lo largo de un período de al menos seis meses, estudiando sus dispersiones, movimientos, entre otras características.
[5] A profundizar en el apartado "Índice de Paridad de Poder de Compra"

		Categoría
1	Alimentos y bebidas no alcohólicas	*División - nivel uno*
1.1	Productos alimenticios	*Grupo - nivel dos*
1.1.1	Pan y cereales	*Clase - nivel tres*
1.1.2	Carne	*Clase - nivel tres*
1.1.3	Pescado	*Clase - nivel tres*

Para el caso que nos concierne, dada las limitaciones ya mencionadas que presenta el "comercio online", pudimos abarcar los siguientes grupos:

- Vestimenta y calzado.
- Ocio, espectáculo y cultura.
- Mobiliario, equipamiento y gastos corrientes en la conservación de la vivienda.
- Transportes.
- Bebidas alcohólicas, tabaco y narcóticos.

Teniendo en cuenta estos aspectos, la estructura de la base de datos quedará conformada de la siguiente manera - replicarse para cada país analizado[6] -:

ARG - PAÍS i						
Categoría "a"		**Categoría "b"**		…	**Categoría "k"**	
Producto 1	*Precio País i / Precio Arg*	Producto 1	*Precio País i / Precio Arg*		Producto 1	*Precio País i / Precio Arg*
…	*Precio País i / Precio Arg*	…	*Precio País i / Precio Arg*		…	*Precio País i / Precio Arg*
Producto Na	*Precio País i / Precio Arg*	Producto Nb	*Precio País i / Precio Arg*		Producto Nk	*Precio País i / Precio Arg*

Estadística descriptiva

Una vez definidos los lineamientos de confección de la base de datos, y concluido el proceso de *scraping* -junto con el posterior filtrado de la información-, se conformó la base definitiva con la que habrá de realizarse el análisis de interés.

[6] Brasil, Chile, Colombia, México.

Es preciso recordar que los datos obtenidos refieren a ratios de precios de productos idénticos, comparados bilateralmente -contra Argentina-, de lo que se desprende un tipo de cambio implícito entre ambos países. Estos valores, a su vez, son el insumo para calcular los índices que, finalmente, darán lugar a los PLIs (instrumento de comparación de PPP entre países).

A continuación, se exhiben ciertos aspectos estadísticos de la base de datos. Es menester aclarar que, en este apartado, se han tomado valores agregados para cada país; es decir, no se ha contemplado la distinción en categorías de acuerdo con la COICOP.

Argentina - Brasil		*Argentina - Chile*		*Argentina - Colombia*		*Argentina - México*	
Media	7,136	Media	0,040	Media	0,009	Media	1,344
Mediana	6,995	Mediana	0,041	Mediana	0,008	Mediana	1,250
Desvío estándar	1,752	Desvío estándar	0,020	Desvío estándar	0,002	Desvío estándar	0,568
Coef. de variación	25%	Coef. de variación	49%	Coef. de variación	26%	Coef. de variación	42%
Mínimo	1,228	Mínimo	0,001	Mínimo	0,002	Mínimo	0,123
Máximo	16,348	Máximo	0,140	Máximo	0,020	Máximo	4,403
Nº Observaciones	1453	Nº Observaciones	629	Nº Observaciones	654	Nº Observaciones	880

Fig. 1 – Estadística descriptiva

Puede apreciarse la gran variabilidad existente en los tipos de cambio implícitos obtenidos. En particular, entendemos que este aspecto está estrechamente relacionado con la diversidad de categorías abarcadas. De hecho, podemos encontrar evidencia de esta cuestión si únicamente nos remitimos a rubros específicos: Chile, por ejemplo, presenta un tipo de cambio muy favorable en el rubro de vestimenta e indumentaria; en Brasil se repite este fenómeno en la categoría automotriz; para México, en cambio, se da una situación similar si se tienen en cuenta los electrodomésticos. En el agregado, estas diferencias se contrapesan con rubros en los que Argentina se ve favorecida en términos de precios; pero el mero hecho de que existan da lugar a un mayor nivel de variabilidad en los datos.

Este aspecto es aún más notorio en el gráfico de dispersión. No es casualidad que Chile presente múltiples observaciones sistemáticamente alejadas del promedio de los datos: todas ellas pertenecen -sin excepción- al rubro de vestimenta[7].

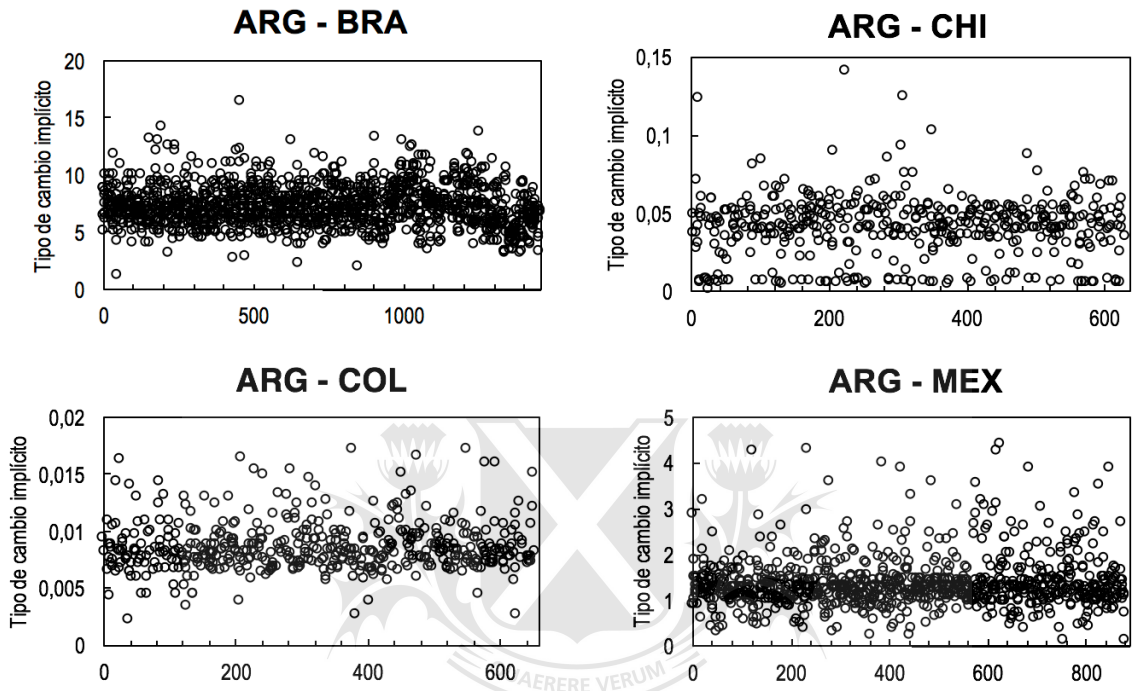

Fig. 2 – Tipos de cambio implícitos (ARS/moneda extranjera)

En particular, se manifiesta que tanto Chile como México son los países con más heterogeneidad en la dispersión de los datos (respecto de su media), con coeficientes de variación que duplican los valores para Colombia y Brasil.

Por otra parte, se podría presumir que existen ciertos valores extremos que eventualmente afectarían el cálculo de los indicadores de interés[8]. A fin de evaluar la influencia de los potenciales *outliers* en el cómputo del promedio, se calculó la mediana, y se encontró que no existen grandes disparidades con el valor promedio: la mayor diferencia entre ambos valores se da para México, siendo esta de 9 centavos (7% del promedio). Esta diferencia de cuantía no afecta sustancialmente los resultados, por lo que no consideramos este aspecto como un problema significativo de cara al estudio propuesto.

[7] Puntualmente, vestimenta masculina.
[8] Pues los índices de precios son promedios ponderados de las observaciones.

Finalmente, creemos que merece la pena destacar el hecho de que Brasil sea el país para el que existen más observaciones, con un número que prácticamente duplica el valor para los demás países. Esto, en parte, está vinculado con las ya mencionadas limitaciones del propio *scraping*; y tal vez, con el hecho de que la variedad de productos disponibles en el país vecino es mayor y, de este modo, también lo son las coincidencias (productos idénticos entre Argentina y Brasil).

IV. Índice de Paridad de Poder de Compra

En este apartado explicaremos y expondremos el funcionamiento del índice que se utilizará para ver la paridad de poder de compra entre Argentina y algunos países de Latinoamérica. Esta medida tiene como objetivo agregar la información individual de los precios de cada sector y cada país, con el objetivo de observar, como muestra de aplicabilidad de esta base de datos, cuánto más caro (o barato) es Argentina respecto a sus pares latinoamericanos.

Cabe aclarar, el modelo de índice que pretende replicarse, en línea con Cavallo et al. (2018) y Deaton y Heston (2010), toma como referencia a aquel confeccionado por el *International Comparison Program* (ICP) del Banco Mundial; que constituye uno de las principales medidas de paridad de poder de compra de referencia global.

Índice

Tal como sugieren Deaton y Heston (2010), la forma más conveniente de comparar el poder de compra entre países sería considerar un índice de costo de vida, que no se limite a cotejar meramente niveles de precios, sino los niveles de utilidad asociados al consumo de las respectivas canastas de bienes. Así, si los precios de un país "c" pueden expresarse como un vector $\boldsymbol{p}^c$ y las preferencias son homotéticas e idénticas en todos los países, la función de gasto puede ser expresada como $u^c\alpha(\boldsymbol{p}^c)$, para un nivel de utilidad u^c y una función linealmente homogénea no indexada en "c".

Sin embargo, esta construcción -puede pensarse- se basa en axiomas poco realistas. Como mencionamos anteriormente, se presume que las preferencias de consumo son homotéticas e idénticas entre los países, algo refutado por múltiples estudios que muestran la irrealidad de esta conjetura. En esta línea, se sugiere que no todos los países presentan los mismos gustos

y/o necesidades sobre bienes consumidos, por lo que consideramos -en concordancia con Deaton (2010)- que es más apropiado utilizar únicamente aquella fracción de la función de gasto que se encuentra libre de tales supuestos: el índice de precios $\alpha(\boldsymbol{p}^c)$.

De esta manera, y a partir de los índices de precios de cada país, proponemos definir la Paridad de Poder de Compra (PPP) de la siguiente manera. Si -arbitrariamente- se considera a un país como numerario (Argentina, por ejemplo), la PPP en pesos argentinos queda expresada como:

$$P^c = \alpha(\boldsymbol{p}^c)/\alpha(\boldsymbol{p}^{ARG})$$

por lo que únicamente restaría estimar la función $\alpha(\cdot)$ para la construcción del índice bilateral de interés.

Sean:

$$P_L^{cd} = \sum_{i=1}^{N} s_i^c \frac{p_i^d}{p_i^c} \qquad P_P^{cd} = \left(\sum_{i=1}^{N} s_i^d \frac{p_i^c}{p_i^d} \right)^{-1}$$

los índices de Laspeyres y de Paasche, respectivamente; en donde s_i^j representa la ponderación del bien i surgida de las estadísticas de consumo del país j[9], p_i^c el precio del bien i en el país c y p_i^d el precio del bien i en el país d.

Tales índices, por construcción, son ponderados por las proporciones de consumo de cada país a comparar -c o d, por caso-; y, por lo tanto, contienen un sesgo a favor de las características del consumo del país del que toman los ponderadores. Idealmente, tanto Laspeyres como Paasche deberían dar el mismo resultado, pero esto no ocurre (aunque sí son muy similares).[10] Por ello, optamos por utilizar el índice de Fisher, que calcula una media geométrica entre ambos, evitando estos posibles sesgos.

[9] $j = \{c, d\}$

[10] Sin embargo, cabe aclarar, dado que los países considerados pertenecen a una misma región, sus economías pueden considerarse como muy similares. Esto genera que Laspeyres y Paasche presenten resultados muy similares, por lo que no debe pensarse, desde nuestro punto de vista, que los sesgos existentes puedan llegan a ser muy significativos. Utilizar Fisher es más una decisión preventiva que obligada.

De este modo, obtenemos el índice de Fisher, que se define como:

$$P_F^{cd} = \sqrt{P_L^{cd} P_P^{cd}}$$

Finalmente, resta dividir el índice obtenido (P_F^{cd}) por el tipo de cambio nominal entre los países en cuestión[11]. Así obtenemos el PLI (*Price level índex*). Este índice no presenta unidades de referencia, y refleja si los precios son mayores (PLI>1) o menores (PLI<1) en cada país, relativo al de referencia (Argentina).

Como una última aclaración, nos parece relevante comentar brevemente el trasfondo de la variable s_i. Ella representa las ponderaciones de consumo que tienen los distintos conjuntos de bienes (clasificados como se menciono en el apartado anterior), pero re escalados. En otras palabras, dado que no disponemos de una variabilidad de bienes suficientes para cubrir todas las categorías de nivel tres, tomamos como el total de la ponderación a la suma de los *shares* de consumo de aquellas clases que si presenta una disponibilidad de datos. Por ejemplo, de encontrarnos con la siguiente situación:

		Categoría	**Ponderación**	**Disponibilidad de datos**
1	Alimentos y bebidas no alcohólicas	*División - nivel uno*	100%	
1.1	Productos alimenticios	*Grupo - nivel dos*	-	
1.1.1	Pan y cereales	*Clase - nivel tres*	25%	No
1.1.2	Carne	*Clase - nivel tres*	25%	Si
1.1.3	Pescado	*Clase - nivel tres*	50%	Si

Bajo este contexto, los bienes cubiertos por "Carne" y por "Pescado" pasarán a presentar una ponderación más alta. Ahora, la suma de sus *shares* (75%) será el total (y no 100% como lo era antes). Esto significa que ahora, "Carne" pasa a presentar un 33.33% y "Pescado" un 66.67% (cuando antes era un 25% y 50% respectivamente).

[11] El tipo de cambio nominal que tomamos aquí es un promedio durante los meses en que se realizó la búsqueda de datos (Marzo-Abril del 2018).

V. Resultados

Como mencionamos anteriormente, el índice desarrollado permite observar cuánta moneda del país local es necesaria para comprar tanto como lo hace la moneda en el país numerario. Si bien no es un índice de costo de vida, el resultado nos permite tener una noción de si un país es más caro o barato que el otro. Para observar esto más claramente, calculamos los respectivos PLI.

En la siguiente tabla se exhiben los resultados:

Tabla 1 – Paridad de poder adquisitivo para el primer trimestre del año 2018 (Argentina = 1)

	PPP implícita
Argentina	1,00
Brasil	0,9212
Colombia	0,8126
Chile	0,9250
México	0,8746
Diferencia absoluta de la media	9%

Notas: PLIs bilaterales (PPP/E) cubriendo Bebidas Alcohólicas, Ropa, Electrodomésticos, Transportes, Combustibles y Consumo Recreacional (incluyendo electrónicos)

Número de productos 3640

Para un total de 3640 productos, encontramos que Argentina siempre presenta, en promedio, precios más altos que sus pares latinoamericanos.

Una posible explicación para estos resultados puede venir desde la intensidad de comercio bilateral. Uno podría pensar que aquellos países con los que Argentina comercializa más (menos) deberían presentar unos PLI más cercanos (lejanos) a uno, es decir, precios más similares (más diferentes).

Justamente, aquellos países con los que Argentina más comercializa, Brasil y Chile (16% y 4% del total comerciado, respectivamente), presentan los PLI más cercanos a uno. De la

misma manera, el país con el que Argentina menos intercambio bilateral presenta (0.90%) coincide con aquel que presenta el menor PLI: Colombia.[12]

Por otro lado, también es preciso considerar algunas cuestiones puntuales para cada país que -creemos- pueden estar estrechamente asociadas a los resultados que el estudio arrojó. En primer lugar, la política arancelaria e impositiva chilena, que tiene un rol determinante en el nivel de precios del vecino país. En particular, y considerando la procedencia extranjera de parte de los bienes considerados -preferentemente ropa, calzados e indumentaria; electrónicos y electrodomésticos- la política comercial de Chile, con menos barreras a las importaciones, probablemente sea un motivo determinante en el menor nivel de precios a nivel agregado. Si se tiene en cuenta que estas categorías representan en conjunto el 45% de la canasta agregada contemplada en el cálculo del PLI, más sentido parece cobrar esta explicación.

Para el caso de México, es menester tener en cuenta la naturaleza de la estructura productiva y comercial del país. Según datos del OEC[6], el país azteca es un importante exportador de electrónicos, electrodomésticos y autopartes; por lo que podría pensarse que tiene ventaja comparativa en estos rubros y, por tanto, precios más bajos. De hecho, esto se verifica en los datos: en las categorías que refieren a estos artículos, los precios mexicanos son notoriamente menores que sus equivalentes en Argentina. La consecuencia observable en los resultados es, justamente, un nivel de precios agregado menor.

Al considerar el caso de Colombia, podemos notar que su estructura productiva se caracteriza, en gran parte, por ser un neto exportador de petróleo. Esto conlleva a que los combustibles (que presentan una ponderación del 12%) muestren un precio relativo menor al de Argentina (dado que exhibe ventajas comparativas en este rubro). Al mismo tiempo, los datos también exponen que Colombia presenta precios menores para los bienes considerados en "Ocio y recreación". Si bien, tanto este país como Argentina son importadores de estos productos, es el grado de apertura comercial (i.e., menos aranceles), lo que genera que Colombia muestre precios menores en este rubro.

[12] Los porcentajes de comercio exterior fueron obtenidos de la The *Observatory of Economy Complexity* (OEC).

Por último, tenemos a Brasil. De la misma forma que Chile, este país presenta una política arancelaria más beneficiosa para la importación de indumentaria, lo que conlleva a que los precios agregados del rubro "Artículos de vestir y calzado" sean significativamente menores que en Argentina. También, el rubro "Transporte" muestra un resultado similar. Esto se debe a que Brasil, no solo es un gran productor de petróleo a nivel internacional (generando menores precios en los combustibles), sino que también exhibe una industria automotriz muy desarrollada (más grande que la argentina), presentando consecuentemente ventajas comparativas en la elaboración de este tipo de productos. Al considerar que, en conjunto, estas dos categorías representan más del 50% de la canasta agregada utilizada para el cálculo de los PLI, la explicación parece cobrar más sentido.

VI. Conclusión

En este trabajo intentamos convalidar la aplicabilidad de la *Big Data* en algún tema de relevancia económica. Bajo dicho objetivo, elegimos calcular la paridad de poder de compra entre Argentina y diversos países latinoamericanos. Nuevamente mencionamos, la relevancia de esta tesis se encuentra en el armado de la base de datos y no tanto en su efectiva aplicación.

De esta manera, en el primer capítulo realizamos una revisión de literatura donde pudimos apreciar las diferentes aplicaciones que tuvo la *Big Data* a lo largo del siglo XXI, encontrando y promoviendo numerosas respuestas a partir de la riqueza de sus bases de datos.

En el segundo capítulo comentamos tanto la forma en que se recogieron los datos, como también la confección de la base de datos y las limitaciones que esta base (y en general, la *Big Data* en su totalidad) presenta a la hora de su aplicación concreta. Notamos que el *match* de los productos no es un tópico a destratar dada su severa dificultad y, al mismo tiempo, su gran importancia. También, entre otros temas tratados, apreciamos la representatividad restringida que presentan este tipo de bases.

En el tercer capitulo comentamos cómo armamos el índice de precios, el por qué de su elección, como también las categorías que logramos cubrir a partir del *scraping*. Vimos que tanto Paasche como Laspeyres podían generar algún tipo de sesgo, por lo que elegimos el índice de Fisher que resulta ser una media geométrica entre ambos.

Finalmente, en el cuarto y último capítulo observamos los resultados obtenidos del cálculo: Argentina presenta precios promedios más caros que sus pares latinoamericanos. Propusimos varias explicaciones para comprender este resultado, dentro de las que encontramos la intensidad de comercio bilateral, como también las características intrínsecas de cada país (sus barreas arancelarias, ventajas comparativas, entre otros aspectos).

Como posibles futuras aplicaciones encontramos relevante poder agregarles un espectro temporal a los datos, en pos de poder analizar las persistencias de los desvíos, como también los movimientos del índice. También poder sortear algunas de las limitaciones antes mencionadas, como la diferencia de precios por localización o incrementar la representatividad de los datos, agregado información sobre los precios de distintos servicios en los casos que se pueda.

VII. Referencias bibliográficas

Bertolotto, M. (2016) Matching Distortion and Mean-Reversion Properties of the Real exchange Rates. *Universidad de San Andres*, Buenos Aires, Argentina.

Bertolotto, M. y Aparicio, D. (2017) Forecasting Inflation With Online Prices. *Massachusetts Institute of Technology*, Massachussets, Estados Unidos.

Cavallo, A. (2016) Scraped Data and Sticky Prices. *MIT & NBER*, Massachussets, Estados Unidos.

Cavallo, A., Diewert, W., Robert, C., Feenstra, R. y Marcel, P. (2018) Using Online Prices for Measuring Real Consumption Across Countries. *Massachusetts Institute of Technology*, Massachussets, Estados Unidos.

Cavallo, A., Neiman, B. y Rigobon, R. (2014) Currency unions, product introductions, and the real exchange rate. *The Quarterly Journal of Economics*, 529-595.

Deaton, A. y Heaston, A. (2010) Understanding PPPs and PPP-based National Accounts. *American Economic Journal, 2(4),* 1-35.

Departamento Administrativo Nacional de Estadística de Colombia (DANE). Recuperado de http://www.dane.gov.co.

Einav, L. y Levin, J. (2014) Economics in the age of big data. *Science*, *346*

Ellison, G. y Ellison, S. (2005) Lessons About Markets from the Internet. *Journal of Economic Perspectives*, *19*(2), 139-158.

Fondo Monetario Internacional (FMI). Recuperado de http://www.imf.org/external/spanish/index.htm.

Instituto Brasileiro de Geografia e Estatística (IGBE). Recuperado de https://ww2.ibge.gov.br/home/.

Instituto Nacional de Estadísticas de Chile (INE). Recuperado de http://www.ine.cl.

Instituto Nacional de Estadística, Geografía e Informática de México (INEGI). Recuperado de http://www.inegi.org.mx/default.aspx.

Instituto Nacional de Estadística y Censos de la República Argentina (INDEC). Recuperado de https://www.indec.gov.ar.

Meyer, E., Schroeder, R. y Taylor, L. (2014) Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? *Big Data & Society*, 1-10.

Morton, F., Zettelmeyer, F. y Silva.-Risso, J. (2001) Internet Car Retailing. *The Journal of Industrial Economics*, *49*(4), 501-519.

Polidoro, F., Giannini, R., Lo Conte, R., Mosca, S. y Rossetti, F. (2015) Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS* 31*,* 165-176.

The Observatory of Economic Complexity. Recuperado de https://atlas.media.mit.edu/en/.

United Nations, DESA (1999): Classifications of Expenditure According to Purpose, Statistical Papers, Series M No. 84, ST/ESA/STAT/SER.M/84.