New Data in Macroeconomics

Alberto Cavallo
MIT & NBER

Banco Central do Brasil - November 2017
Big Data and Computer Science

- The 3vs: Volume, Variety, Variety
- Hadoop and Spark, Nosql, new database technologies
- Machine learning, Neural Networks, Deep learning

- Not what I will talk about!

For that, I recommend:

- Data Science and Big Data Analytics (MIT Professional Education)
  https://mitxpro.mit.edu/courses/course-v1:MITxPRO+DSx+3T2017/about

- Data Analysis for Social Scientists (MITX Probability, Causality, R and visualization)
  https://www.edx.org/course/data-analysis-social-scientists-mitx-14-310x-3

- Machine Learning Andrew NG - Coursera
  https://www.coursera.org/learn/machine-learning
Zvi Griliches (1985), on the uneasy alliance between economists and data:

“... we have shown little interest in improving it [the data], in getting involved in the grubby task of designing and collecting original data sets of our own. Most of our work is on “found” data, data that have been collected by somebody else, often for quite different purposes... “They”collect the data and are responsible for all their imperfections. “We” try to do the best with what we get, to find the grain of relevant information in all the chaff.”
New Data and Macroeconomics

- New technologies for data collection
  - New data, but not necessarily big

- Traditional Macro Sources
  - Statistical Offices & Multi-lateral organizations

- New Macro Sources
  - Administrative data (eg. tax, property records)
  - Scanner Data (eg. Nielsen)
  - Online Data (eg. Billion Prices Project)
  - Crowdsourced data (eg. Online survey, mobile phones)
  - Search data (eg. Google, Indeed)
  - Satellite Data (eg. lights, parking lots, tanker and crop heights)
  - Sensor data (smart phones, smart watches, IOT devices)
Each Data Source has Advantages and Disadvantages

CPI Data

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Representative sample</td>
<td>• Very costly to collect and access</td>
</tr>
<tr>
<td>• carefully-chosen goods</td>
<td>• Low frequency (monthly)</td>
</tr>
<tr>
<td>• many retailers and locations</td>
<td>• Limited number of goods and varieties</td>
</tr>
<tr>
<td>• Long Time Series</td>
<td>• Some unit values and imputed prices</td>
</tr>
<tr>
<td>• Collection of posted prices in stores</td>
<td>• Difficult international comparisons</td>
</tr>
</tbody>
</table>
Each Data Source has Advantages and Disadvantages

Scanner Data

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| • Granularity  
  • Some product details for all goods sold  
  • Transaction data  
  • Contains quantities and sometimes costs  
  • Frequency (weekly) | • High cost to collect/acquire  
  • Limited coverage (supermarkets, department stores)  
  • Data characteristics vary greatly depending on provider, location, time period, etc.  
  • Hard to compare internationally  
  • Unit values and time-averages (eg: prices are often calculated as sales/quantity in a week) |
Each Data Source has Advantages and Disadvantages

### Online Data

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Frequency (daily)</td>
<td>• Fewer retailer and locations than CPI</td>
</tr>
<tr>
<td>• Cheap to collect (but complicated)</td>
<td>• Short time series</td>
</tr>
<tr>
<td>• Granularity</td>
<td>• Not all categories of goods and services are online (not yet)</td>
</tr>
<tr>
<td>• All product details (brands, size, anything shown online)</td>
<td>• Online and Offline prices may behave differently</td>
</tr>
<tr>
<td>• All goods and varieties available for sale (census)</td>
<td></td>
</tr>
<tr>
<td>• New goods automatically sampled</td>
<td></td>
</tr>
<tr>
<td>• Easier to compare internationally</td>
<td></td>
</tr>
</tbody>
</table>
Two main uses: Prediction and Measurement

- Prediction and Forecasting
  - Use new metrics to predict a traditional variable
    - google trends and unemployment
    - twitter sentiment and stock market
  - Supervised machine learning
  - Nowcasting methods

- Measurement opportunities
  1. Improve data and methods for traditional statistics
     - Eg. online data for inflation
  2. Measure things that could not be measured before
     - Eg.
       - Mobile sensors to measure traffic patterns
       - Satellite data to measure height of crops
The Case of Argentina’s Missing Inflation
The Case of Argentina’s Missing Inflation
The Case of Argentina’s Missing Inflation

Inflation Expectations (Survey - Di Tella University)

CPI Annual Inflation

National Statistics Agency is officially "intervened"

Government starts to "monitor" the National Statistics Agency

CPI YoY
Inflation Expectations (Di Tella University)
The Case of Argentina’s Missing Inflation

Inflation Expectations
(Survey - Di Tella University)

National Statistics Agency is officially “intervened”

Government starts to “monitor” the National Statistics Agency

CPI Annual Inflation

Annual Inflation Rate (%)

CPI YoY  Inflation Expectations (Di Tella University)
The government’s response

- Cristina Kirchner, 11/2007, Interview in Pagina 12

  With what story [relato] do we address the topic of Indec? With the relato that one day the villains from the government came to an institution that measured everything well [.]? [we need to] admit that there are political interests [.] the measurement models are not the Bible, the Coran, or the Talmut
Web-Scraping Online Data

• Every day, a *robot* downloads a public webpage, analyses its HTML code, extract price data, and stores it in a database.
My initial Computing Power

- Lenovo X60s
- Scraped 6 supermakets
- 7 hours every day

Table 1: Data Description

<table>
<thead>
<tr>
<th></th>
<th>Argentina Retailer #1</th>
<th>Argentina Retailer #2*</th>
<th>Brazil</th>
<th>Chile</th>
<th>Colombia</th>
<th>Venezuela</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ends</td>
<td>3/24/2011</td>
<td>03/20/2011</td>
<td>03/01/2010</td>
<td>03/20/2011</td>
<td>03/24/2011</td>
<td>03/01/2010</td>
</tr>
<tr>
<td>Categories</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Prices P/day (mean)</td>
<td>11,560</td>
<td>4,790</td>
<td>11,000</td>
<td>12,000</td>
<td>5,000</td>
<td>9,256</td>
</tr>
<tr>
<td>Total Products</td>
<td>26,333</td>
<td>10,929</td>
<td>21,804</td>
<td>3,5432</td>
<td>9,166</td>
<td>20,847</td>
</tr>
<tr>
<td>Price Changes</td>
<td>204,449</td>
<td>136,781</td>
<td>25,9875</td>
<td>12,0112</td>
<td>76,979</td>
<td>94,808</td>
</tr>
<tr>
<td>Market Share (Supermarkets) **</td>
<td>28%</td>
<td>n/a</td>
<td>15%</td>
<td>27%</td>
<td>30%</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Note:  *Argentina’s Retailer #1 is used in all online prices indexes reported in this paper for that country. Argentina’s Retailer #2 is used only in the robustness results shown in Figure 5. **Market shares are based on claims posted on the corporate webpages of each supermarket’s parent companies.

- Largest supermarket in each country
- Categories covered include food, beverages, and household products
Methodology

- Used data from the largest supermarket in each country, with about 11K daily products each.

- All indexes:
  - Daily
  - Include sales
  - No product substitutions
  - Use all products available in each retailer
  - Missing values within price spells are completed using the last available price for each product.

- Used standard CPI methods in these countries:
  1. price changes are obtained at the product level,
  2. then averaged inside categories using geometric means,
  3. then aggregated across categories with a weighted arithmetic mean
Methodology

- Official CPI category weights and compare online series to an equivalent official index (covering food, beverages, and household products)

Table 2: CPI Category Weights (%)

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Argentina</th>
<th>Brazil</th>
<th>Chile</th>
<th>Colombia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baby Food</td>
<td>0.27</td>
<td></td>
<td>0.08</td>
<td>0.43</td>
</tr>
<tr>
<td>Beer At Home</td>
<td>0.11</td>
<td>0.30</td>
<td>0.32</td>
<td>0.20</td>
</tr>
<tr>
<td>Distilled Spirits At Home</td>
<td>0.48</td>
<td></td>
<td>0.37</td>
<td>0.50</td>
</tr>
<tr>
<td>Wine At Home</td>
<td></td>
<td>0.83</td>
<td>0.95</td>
<td>0.12</td>
</tr>
<tr>
<td>Cigarettes</td>
<td>0.29</td>
<td>0.16</td>
<td>0.25</td>
<td>0.40</td>
</tr>
<tr>
<td>Dental &amp; Nonelectric Shaving Products</td>
<td>0.29</td>
<td>0.45</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>Deodorant/Sunscreen Preparations</td>
<td>0.63</td>
<td>0.15</td>
<td>0.21</td>
<td>0.48</td>
</tr>
<tr>
<td>Electric Personal Care Appliances</td>
<td>0.66</td>
<td>0.19</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Cosmetics Nail Preparations &amp; Implements</td>
<td>1.01</td>
<td>0.34</td>
<td>0.56</td>
<td>1.52</td>
</tr>
<tr>
<td>Baby Care Products</td>
<td>0.35</td>
<td>0.88</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Sanitary/Footcare Products</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL WEIGHT (% of CPI)</td>
<td>48.51</td>
<td>27.03</td>
<td>31.00</td>
<td>28.44</td>
</tr>
</tbody>
</table>
Online Indices seemed to work well in other countries

- Could roughly match CPIs in Brazil, Chile, Colombia & Venezuela

Fig. 1. Online and official indexes in four Latin American countries: (a) Brazil; (b) Chile; (c) Colombia; and (d) Venezuela. Notes: The daily online supermarket index is constructed with an online prices and official CPI category weights. In Venezuela, the online data has no category information and therefore the online index is built as a geometric average of all price changes observed each day. The official supermarket index is an equivalent indicator constructed as a weighted average of the “Food and Beverages” and “Household Products” official price indexes in each country.

Online Indices seemed to work well in other countries

- Could roughly match CPIs in Brazil, Chile, Colombia & Venezuela

Figure 2: Online and Official Indexes - Annual Inflation Rate

Brazil, Chile, Colombia, Venezuela

- Online indexes able to track main inflation trends
  - Even with only 1 retailer in each country

- Matching was best in Chile and Colombia, where:
  - Supermarkets have larger market shares (27% and 30%, vs. only 15% in Brazil)
  - City concentrates population & accounts for most of the CPI (55% in Chile)

- Good news for Argentina: the supermarket used had 28% market share, Buenos Aires was 100% of CPI data
But something was wrong with Argentina...

Argentina

- In 4 years online prices increased 100% and the CPI only 25%!

Different level but similar dynamics!

- The annual inflation rates had similar dynamics over time

Argentina

How was the Government doing this?

- Many theories...
I tried lots of different things

- Alternative retailer (low cost)
- Re-weighted index
- Simpler subsistence index based on small basket
- Cell-relative imputation when change was too high
- Use only goods that had lowest ex-post inflation per category
- Use only price agreement prices (price-controlled goods)
Always higher inflation than reported by the government.

Best way to approximate official inflation?

- Take the online inflation rate and divide by 3.

Los Precios Hoy

Somos un grupo de economistas que todos los días registra los precios de un listado de productos en dos grandes supermercados de Buenos Aires y los publica en este sitio.

En esta sección mostramos un resumen de nuestras principales estadísticas, basadas en la evolución diaria de dos índices propios de inflación:

- Un Índice de Alimentos y Bebidas, equivalente al 31% del IPC.
- Un Índice de la Canasta Básica Alimentaria, como el que el INDEC usa (o usaba) para calcular el nivel de indigencia.

En la sección Novedades hacemos un análisis periódico de los datos. En las secciones de Precios, Aumentos y Caídas se pueden ver detalles de precios individuales. Finalmente, en Metodología y Preguntas explicamos cómo lo hacemos y cómo podrás ayudarnos.
In 2011, things escalated

ARGENTINA FINES MORE ECONOMISTS OVER INFLATION ESTIMATES

BY TAOS TURNER

BUENOS AIRES—Argentina’s government has fined more economists for challenging official inflation estimates in what lawyers call a violation of freedom of speech.

ARGENTINA CHARGES ECONOMISTS

By TAOS TURNER

BUENOS AIRES—Argentina’s government has filed criminal charges against the managers of an economic consulting firm, escalating its persecution of independent economists.
The rest of the world takes notice...

- **2012:**
  - *The Economist* started publishing our index every week instead of the official data.

- **2013:**
  - IMF censured Argentina for its shady inflation statistics.
The Holdouts Effect

- New CPI in Jan 2014, as the country tried to borrow again in international markets → surprisingly accurate the first month

- After holdouts ruling in NY → quickly lost credibility again...
Severity of the Statistical Crisis

- Eventually, INDEC stopped publishing:
  - Sub-indexes → things did not add up
  - Provincial data → some provinces reported much higher inflation
  - National CPI
  - Poverty index → inconsistent with announced real wage increases

Poverty is now below 5%  President Cristina Kirchner, FAO 2015

https://www.youtube.com/watch?v=dorlmCVpItY
Severity of the Statistical Crisis

- By underestimating inflation, INDEC overestimated growth

- New government in December 2016
  - could not find the hard drives where the CPI data was stored...
  - no official data for 6 months!
Since 2016, the CPI closely matches our online index

Source: www.inflacionverdadera.com - PriceStats - State Street
Since 2016, the CPI closely matches our online index

Source: www.inflacionverdadera.com - PriceStats - State Street
Since 2016, the CPI closely matches our online index

Source: www.inflacionverdadera.com - PriceStats - State Street
If you ever need historical data on Argentina’s Inflation

- Download chained monthly index from 1946 to the present at:
  http://www.inflacionverdadera.com/argentina/

Filling the Gap in Argentina’s Inflation Data*

Alberto Cavallo
MIT Sloan and NBER

Manuel Bertolotto
PriceStats and UdESA

This Draft: May 2016

Abstract

The official consumer price index (CPI) of Argentina became widely discredited after January 2007 and the National Institute of Statistics and Censuses of Argentina (INDEC) ceased its production in December 2015, leaving the country with no official measure of inflation after this date. This gap in official data poses a major problem for researchers and policymakers who need access to longer time series. In this paper, we provide a consumer price index that uses official CPI data from 1943 to 2007 and chains it to an online price index which spans from 2007 to the present. The chained index will be updated regularly on a monthly basis until a new official CPI is released.
Final Remarks

- Big Data $\Rightarrow$ huge measurement opportunity
  - New data collection tools (web, sensors, phones, satellites)
  - Build customized datasets that fit specific measurement and research needs
  - Anyone can do this!

- Next session
  - The Billion Prices Project, daily inflation in 22 countries, and online-offline comparisons