APPLIED ECONOMICS IN THE DIGITAL ERA

WEBDATANET TF25

www.webdatanet.eu

Pablo de Pedraza
SEVILLE 22 SEPTEMBER 2015

Researchers & stakeholders from any different backgrounds & 34 countries

Foster scientific use of web-based data to benefit society & public interests

WG1 Quality       WG2 Innovation       WG3 Implementation

Motivation/Stimulation

Digitalization & data flowing:
- Capture, create and save more data than ever before
- The internet of things

Applied Economics have not fully tap into this opportunity

We need a strategy for publicly funded research to be part of the new realities

“What should a research agenda and strategy of academic and institutional researchers look like if applied economics is to fully benefit from Big data?”

2.2.- Identify reasons of the limited use: Big Data Characterization

2.3.- Propose potential solutions: Strategies and actions
<table>
<thead>
<tr>
<th>2.2.- Big data Characterization</th>
<th>2.3.-Strategies and actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The 5Vs in Social sciences</td>
<td>1. Tap into current opportunities</td>
</tr>
<tr>
<td>2. Case studies</td>
<td>2. Sources of error</td>
</tr>
<tr>
<td>3. Danger provision of data direct research</td>
<td>3. Claiming access to data</td>
</tr>
<tr>
<td>4. Big Data management</td>
<td>4. Stakeholder consortium</td>
</tr>
<tr>
<td>5. Big data epistemology</td>
<td>5. Networking process</td>
</tr>
</tbody>
</table>
1. The 5Vs in Social sciences
- 5vs (commercial) do not guarantee validity for social sciences
  - **Velocity**: Nowcasting (Giannone et al. 2008)
  - **Variety**: Models based on representative agents can improve
  - **Value**: short run (Mayer & cukier 2013) vs long run & societal
  - **Veracity**: *Organic data ≠ Design data*

- Good qualities for Social Sciences **BUT** not generated by any of the **existing scientific methods**: Uncertainty about generation process & **sources of error** (Coverage, Selectivity, non-response…)

- **Complementarity** with web-survey methods (ex. wageindicator.com) & experiments (Reips 2007 & Edelman 2012). BENCHMARK.
2. Case Studies
Are data able to answer research questions? Examples:

- **Prices** (Cavallo 2011 [www.bpp.mit.edu](http://www.bpp.mit.edu))
- New papers webs to identify **stock market bubbles** (Gerow & Keane, 2012)
- Twitter mood correlation with **DJIA** (Bollena et al. 2011)
- Twitter to predict **currency exchange rates** (Janetzko 2014)
- Google searches to predict: **unemployment** (Choi & Varian 2011)
  - well being during the crisis (Askitas & Zimmerman 2011)
  - **Housing markets** & mortgage delinquencies
  - **Tourists’ inflows** (Artola & Pedraza 2015)
- Wage indicator for **job insecurity** (Bustillo & Pedraza 2010) and **life satisfaction** models Guzi & Pedraza 2015)

(...)

**File drawer effect** (Rosenthal 1979), very little about **limitations** and how to overcome them (Couper 2013, Lazer et al. 2014, Butler 2013, Artola & Pedraza 2015) **not only what but why** (rumour, change in algorithms, search behavior…)

= better access+ benchmark traditional data

Problems are specific to each **research goal**
3. Danger provision of data direct research

Data from Twitter, Facebook, Google, Amazon, or eBay, Master Card, Swift, and UPS…are hidden for science or available only for few researchers.

Also the public sector produce data disregarding potential scientific use (Askitas & Zimmerman 2014: The Toll Index)

**Replicability in danger** when it could increase given ICT

**Open Data Movements** (World Bank, UN Global Pulse, OKF, DA…)

**NO SYSTEMATIC APPROACH FOR RESEARCHERS**

**Problems in opening:**
- privacy & regulations
- data keepers competitiveness
- benefits of sharing
- Cost of resources
4. Big Data management

It is possible to preserve privacy and open data conditional on specific licenses using existing technologies:

- data repositories (OAIS framework, FEDORA repository…)
- models to support traceability of data (W3c Prov model)
- web resource archiving…

Data formats, Meta data describing data sets
5. Big data epistemology
Not yet systematically used to test existing social sciences theoretical models

New tools and ways of thinking
Ex. If N=Millions everything is significant (Taylor et al 2014):
   1st approach pre-analysis with machine learning techniques
      Good: to define questions & variables at the light of theories
      Critic: correlations but not casualty may pick up things that are not there
   2nd approach Econometrics using samples drawn from Big data

   Constant interaction between THEORY AND DATA.

Complementarities & interactions among disciplines
(Survey methods, Computer sciences, Math's and the Graph theory, Psychologists, Linguistics, …)
### 2.2.- Big data Characterization

1. The 5Vs in Social sciences

### 2.3.- Strategies and actions

1. Tap into current opportunities

#### Case Studies + Inductive reasoning feed the general approach

**Goals:**

(ex. Google searches)

**Why** well performing models may break down

Identify **complementary of tools** of any kind

Identify **complementary of data modes & methods**

**Multidisciplinary interaction:** computer sciences, experimental psychology, survey methodology, linguistics, applied mathematics, law…

(ex. [http://www.eduworks-network.eu](http://www.eduworks-network.eu))

Quicker and more flexible dissemination means **avoid File drawer effect**
2. Sources of error:
Approach big data Quality by (Quality frameworks from Statistical offices):

- Sources of error using existing and new frameworks

-Interaction/complementarities among
  - data types: (non-reactive, experiments, surveys…) to test, evaluate, complement, improve…
  - types of variables: ex subjective variables/reveled vs reported preferences

CONSTANT EVALUATION OF DATA

-Origin/activities that generates DATA: Admin, transactional, social Networks, mobile devices…
3. Claiming access to data

Better access = better data consumers’ protection (insurance companies)
Assist legal experts
Specific licenses for researchers
Musgrave’s debate (Data Taxation)
Individual data philanthropy (Ex. Quantify self)
Corporate data philanthropy (benefits of sharing, ex. Google)
Balance inequalities in Access (UN 2014)
Openness score
Build upon & contribute to open data movements (OKF, WB…)
4. Stakeholder consortium
More vigilant in fighting Scientific/Public interests

Lobby to influence legislators

Coordination of Initiatives from International institutions

Higher education curricula & interaction among disciplines
(www.webdatametrics.com)

Global infrastructure of data

Role of Statistical Offices
-Quality Frameworks (Eurostat, OECD, UNECE…): Claiming, curating, and making available while continue generating traditional (benchmark) data.
<table>
<thead>
<tr>
<th>2.2.- Big data Characterization</th>
<th>2.3.- Strategies and actions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conclusions</strong></td>
<td></td>
</tr>
<tr>
<td>Feed the general approach <strong>by inductive reasoning from case studies</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multidisciplinary networking &amp; discussions</td>
</tr>
<tr>
<td>Institutions &amp; researchers interact within a <strong>clear plan</strong> aiming benefit citizen's and consumers</td>
<td></td>
</tr>
<tr>
<td>Access to data</td>
<td></td>
</tr>
</tbody>
</table>
• Kea Tijdens, University of Amsterdam
• Prasanna Lal Das, The World Bank
• Ulf-Dietrich Reips, University of Konstanz
• Richard Freeman, Harvard University
• Muriel Foulonneau, Tudor Research Center
• Clemens A. Grünwald, Datenschutz Nord Group
• Concha Artola Central Bank of Spain
• Fernando Reis, Eurostat
• Laura Wilson, Office for National Statistics United Kingdom
• Fazel Ansari, University of Siegen
• Eva Beyvers, University of Passau.
Gracias

Thank you
Scraping the web
- Billion prices (MIT)

Data keepers open policy
- Ex. Google
- Data philanthropy (UN Global Pulse)

Legal/privacy issues

Carefully anonymized random samples

When Data belong to users
Personal data philanthropy (Ex. Quantified-self)

Open Notices:
The single question survey
Do you want to share your data?
Survey methods, questionnaire design, marketing, incentives...

Openness/Access
Information asymmetries

Business model
- Google: benefits from being open
  - Master Card
  - Others? Openness Score

Legal duties
- Not only good intentions: Data taxation
  - Public Economics debate
  - Are data a public good or a merit good?
  - Specific licenses for science

Open Data movement? Other projects
- Open Economics (Open knowledge foundation)
- Linked Open Data
- UN Global Pulse Project
- World Bank Open Data

Open Data for Social Sciences (ODSS)