Recent Developments in Federal Statistics – Surveys and Beyond

Frauke Kreuter
JPSM – Uni Mannheim – IAB

Melbourne June 26 ‘18
AAPOR Report on Big Data
AAPOR Big Data Task Force
February 12, 2015

INNOVATIONS IN FEDERAL STATISTICS
Combining Data Sources While
Protecting Privacy

Prepared for AAPOR Council by the Task Force, with Task Force members including:

Lill Jager, Co-Chair, Statistics Sweden
Franka Kersten, Co-Chair, PBSI at the U. of Maryland, U. of Manheim & IAB
Marcus Berg, Stockholm University
Paul Buer, RTI International
Paul Decker, National Policy Research
CJF Lange, School of Information at the University of Michigan
Julie Lare, American Institutes for Research
Cathy O’Neil, Johnson Research Labs
Ale O’Keil, HumanSec Group

Acknowledgments: We are grateful for comments, feedback, and editorial help from
Evan Ben-Porath, Jason McMillan, and the AAPOR council members.
The project on which this report is based was funded by the Federal Ministry of Education and Research under the number [16OH22064]. Responsibility for the contents of this publication lies with the author.
Source: Roberto Rigobon, Discussion on Applications and Issues with Using Commercial Data in Research, BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics November 19, 2015
Economic Indicators

Examples for online data collection (and analysis)
US Aggregated Inflation Series
(Monthly Rate, 2008 - Present)

Source: State Street, PriceStats
Employment Websites + Aggregators

**LinkedIn Economic Graph**

LinkedIn’s Economic Graph is a digital representation of the global economy based on data generated from 530 million members, 50,000 skills, 9 million employers, 10 million open jobs, and 29,000 educational institutions. In short, it’s all the data on LinkedIn.

Through mapping every member, company, job, and school, we’re able to spot trends like talent migration, hiring rates, and in-demand skills by region. These insights help us connect people to economic opportunity in new ways.

By partnering with governments and organizations around the world, we help them better connect people to opportunities.
Platforms - Crowdsourcing

- Glassdoor

Get the Ground Truth.

Premise is a data and analytics platform combined with a global contributor network that compiles direct observations for better decision making.
Observations
Netzabgabe für den 4. Juli 2014
1. IAB – Research Examples
1st Refugees
2nd Unemployment

DIE ARBEITSLOSEN VON MARIENTHAL

Die Arbeitssuchenden in einem arbeitslosen Viertel

MIT EINEM ANHANG ZUR GESCHICHTE DER PSYCHOGRAPHIE

Von E. W. F. MEYER UND HANS M. BARCK

VERLAG VON S. FRIEDRICH IN LEIPZIG 1932

The Sociography of an Unemployed Community

Marie Jahoda, Paul F. Lazarsfeld, and Hans Zeisel
IAB-SMART-Studie: Mit dem Smartphone den Arbeitsmarkt erforschen

Autoren

Sebastian Bähr
Georg-Christoph Haas
Florian Keusch
Frauke Kreuter
Mark Trappmann
PASS – Panel (10 years) + Administrative Data

Sample of households with at least one welfare benefit recipient (at reference date)
- Refreshed annually
- Surveyed annually

Random household sample of resident population
- Refreshed annually
- Surveyed annually

Opt-In
Consent in GDPR

Consent needs to be freely given.
Consent needs to be specific, per purpose.
Consent needs to be informed.
Consent needs to be an unambiguous indication.
Consent is an act: it needs to be given by a statement or by a clear act.
Consent needs to be distinguishable from other matters.
Consent request needs to be in clear, plain language; intelligible and easily accessible.
Informed Consent – Privacy - Technology

Instruction booklet with IAB-Smart app screen shots – separate and active opt-in required for all use cases

Bitte klicken Sie auf „AKZEPTIEREN“, um mit der Installation fortzufahren.

Damit alle Funktionen der App genutzt werden können, müssen Sie jeweils zustimmen:

Um die Erfassung von Nutzungsdaten zu erlauben, verschieben Sie den Schieber bitte nach rechts.

Einwilligung in Forschungsdaten

Sie haben aber auch die Möglichkeit die App zu installieren und Punkte für die Beantwortung von Fragen zu erhalten, ohne den Nutzungsdatenzugriff zu erlauben.
## Consent to Linkage by Framing and Mode in %

<table>
<thead>
<tr>
<th></th>
<th>Phone</th>
<th>Front</th>
<th>Back</th>
<th>Total n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain</td>
<td></td>
<td>90.8</td>
<td>78.7</td>
<td>598</td>
</tr>
<tr>
<td>Loss</td>
<td></td>
<td>90.5</td>
<td>81.2</td>
<td>610</td>
</tr>
<tr>
<td>Total n</td>
<td></td>
<td>613</td>
<td>595</td>
<td>1208</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Web</th>
<th>Front</th>
<th>Back</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain</td>
<td></td>
<td>82.6</td>
<td>62.4</td>
<td>520</td>
</tr>
<tr>
<td>Loss</td>
<td></td>
<td>86.3</td>
<td>75.4</td>
<td>489</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>511</td>
<td>498</td>
<td>1009</td>
</tr>
</tbody>
</table>
Lack of understanding

<table>
<thead>
<tr>
<th>Phone</th>
<th>Consenters %correct</th>
<th>Non-consenters %correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers sent to IAB</td>
<td>88.3</td>
<td>57.8</td>
</tr>
<tr>
<td>Merged with IAB</td>
<td>93.3</td>
<td>36.7</td>
</tr>
<tr>
<td>Name/Adress saved</td>
<td>68.3</td>
<td>38.8</td>
</tr>
<tr>
<td>Result lead to you</td>
<td>63.4</td>
<td>--</td>
</tr>
<tr>
<td>IAB only access</td>
<td>85.6</td>
<td>--</td>
</tr>
<tr>
<td>Public access to identifiable data</td>
<td>87.5</td>
<td>--</td>
</tr>
</tbody>
</table>
2. Role of methodologists
Data Generating Process
Collect data and understand how data are generated through administrative and other processes.

Data Curation/Storage
Curate and manage data

Data Analysis
Variety of analysis methods suited for different data types

Data Output/Access
Communicate results and distribute data

Research Question
Formulate research goal and know which data are best suited to achieve this goal.

Data

- Designed
  - Survey
  - Experiment
- Organic
  - Administrative
    - Aspirational
    - Transactional

Source: Roberto Rigobon, Discussion on Applications and Issues with Using Commercial Data in Research, BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics November 19, 2015
3. Some new vocabulary (skills)
Data Generating Process

Data Curation/Storage

Data Analysis

Data Output/Access

Research Questions

Example: map visualization / privacy / GDPR

Example: Hadoop MapReduce; High Frequency Data; Machine Learning

Example: Record Linkage; Database Hadoop Distributed File System

Examples: geolocated social media + survey + administrative data

Examples: Behavior of interest (political participation/job searches)

Machine Learning

https://github.com/DataScienceSpecialization/courses from Roger Peng, Jeff Leek, Brian Caffo
Database Management

- **Text files and scripting language**
  - Your data is small
  - Your analysis is simple
  - You do not expect to repeat analyses over time

- **Statistical packages**
  - Your data is modest in size
  - Your analysis maps well to your chosen statistical package

- **Relational database**
  - Your data is structured
  - You will be analyzing data repeatedly over time

- **NoSQL database**
  - Your data is unstructured
  - Your data is extremely large
BD Programming – MapReduce

Foster et al. Big Data and Social Science
4. What we bring to the table
Data Generating Process

- Construct
- Measurement
- Response
- Edited Response
- Survey Statistic
- Population Mean
- Sampling Frame
- Sample
- Respondents
- Postsurvey Adjusted Data

Groves et al. 2004
Data Generating Process

- Construct:
  - Specification
- Measurement:
  - Measurement
- Response:
  - Processing
- Edited Response
- Survey Statistic
- Population Mean
  - Coverage
  - Sampling Frame
  - Sample
  - Respondents
  - Postsurvey Adjusted Data

Groves et al. 2004
Big Data Process Map

Generate
- Source 1
- Source 2
- Source K

ETL
- Extract
- Transform (Cleanse)
- Load (Store)

Analyze
- Filter/Reduction (Sampling)
- Computation/Analysis (Visualization)

Source: Paul Biemer in Japec, Kreuter et al. 2015 – AAPOR Task Force Report
Boston Street Bumps
5. Back to the IAB - Example
... typical attempts to find analysis help
Full-time employed
→ App use past 5pm

Part-time employed
→ App use at noon

Schlickman et al. (DataDiggers) App Nutzerverläufe statt Surveys? Eine Machbarkeitsstudie. DataFest Germany, Mannheim 2015, http://sswml.uni-mannheim.de/Teaching/DataFest%20Germany/DataFest%20Germany%202015/
Summary

1. Methodologists need to help teams find or create the right data for a specific purpose
2. It is easy to overlook what is missing. Privacy and confidentiality even more important
3. Methodologists should make use of all available data, and they can (with a few new skills)
4. Frameworks can help assess quality of single and combined data sources
5. Working in teams and with unlikely partners can accelerate
THANK YOU!

fkreuter@umd.edu

Twitter: @fraukolos