Using confidential & proprietary administrative data in research

Jo Van Biesebroeck
(ERC Starting Grant: CHINA)

Workshop on Research Data Management and Sharing
ERC Executive Agency, Brussels
September 18-19, 2014
My ERC project (CHINA)

1. Build an integrated data set on the Chinese economy from detailed firm-level administrative data sources (on production & trade)

2. Investigate the impact of China’s policy reforms on its domestic economy, e.g.
   - “WTO Accession & Performance of Chinese Manufacturing Firms”
   - “Using Export Market Performance to Evaluate Regional Preferential Policies in China”
   - “Effects of Deregulation and Vertical Unbundling on the Performance of China's Electricity Generation Sector”
   - “Creative Accounting or Creative Destruction? Firm-level Productivity Growth in Chinese Manufacturing”
Integrated dataset from firm-level administrative data
  - Trade transactions at the firm-product-month-destination level (7y)
  - Supporting information on aggregate units: localities, industries,…

Some notable features:
  - Huge samples: 19 million trade transactions in 2006, 10 million firms in 2008, 2.5 million firm-year observations over time
  - Reporting is compulsory or automatic (information taken from official forms, e.g. customs declarations)
  - Wide range of information: e.g. output, sales, R&D, exports (product-destinations), location, breakdown of workforce, financial information (loans, capital), firm-type,…
My other work with administrative data

• China is unique in that I can work with the data from my office (data is now widely available)

• Experience in other countries:
  o USA: work as sworn-in federal employee in regional data center
  o Canada: work on-site in Ottawa, “no conflict of interest” clause
  o Belgium: work on scrambled data and send final programs for estimation on government servers
  o Colombia: sign statement not to share or publish any of the individual firm information (only possible for older data)
  o OECD / ISGEP: collaborative projects where researchers with access to data from one country run the same programs to conduct cross-country comparative research
  o Some countries are liberalizing access (France, Scandinavia)
Administrative data

• “New” frontier in economic research (& other social sciences)

• Examples:
  o Annual census of manufacturers (industrial organization)
  o Trade-transaction information (international economics)
  o Tax return records of individuals (public finance)
  o Take-up of social security benefits (labor economics)
  o Matching investment grants (regional economics)
  o Prodcom survey on output and value (macroeconomics)
  o Community Innovation Survey (CIS) (innovation)
Benefits of administrative data

- Large samples
- Often covers the universe of the population
- Data collection is often automatic (linked to registries)
- Reliable information
  - Fewer recall problems
  - Compulsory reporting
- Can in principle be linked to myriad sources of other information
- Standardized variables & identifiers
- Contains direct information on policy take-up
Role in open research environment

• **Features:**
  - Possible? Access only after approval by government agencies
  - Convenient? Often only accessible “on site”
  - Reliable? Owner of the data is not the one conducting the research

• **Problems:**
  - Who gets access? (discrimination v. free riding)
  - Limited transfer of program files and backup service, time limited
  - Researcher needs to rely on metadata and data collection & processing procedures of data owner.
  - How can referees gauge accuracy or reliability of results?
  - How can results be replicated?
  - How can other researchers build upon existing results?
  - How to accommodate demands for open access (grants, journals)?
**Data policy at The Economic Journal**

- The authors must provide sufficient detail […] for the reader to understand the nature of the data used and how it was constructed,

- and one of the following [*abridged*]:
  - data sets and programs that allow replication of all of the results, along with a file that describes how the data and programs can be used to replicate the results, and any manipulation that was carried out to obtain the data from the publicly available sources of the data;
  - a request for an exemption based on the grounds that the data are from commercially available or restricted access data sources, in which case programs that allow replication of all of the results, along with a file that describes how the data can be obtained or accessed, and how the programs can be used … must be provided.
  - a request for an exemption based on the grounds that the data are from a proprietary data source that is not accessible to other researchers; papers using such data are discouraged but will be considered on an individual basis by the Editor; the exemption must be requested at the time of submission; if the paper is accepted a file that describes how the data was collected and used to obtain the results must be provided.
Conclusions

• Openness and sharing is great, but at what cost?
• How to coopt governments into the research program to facilitate access and improve administrative data?
• Investing time & effort to unlock new administrative data sources is a risky endeavor for researchers. Not surprisingly there is underinvestment in this activity.
Abstract

The use of administrative data in economic research, as well as in other fields in the social sciences, has greatly increased over the years. These data sources have clear advantages. Samples tend to be orders of magnitudes larger than what researchers could collect themselves using survey instruments. Reporting the information is often compulsory by law, leading to very high response rates. When the data is taken directly from administrative registries, no individuals need to answer additional questions for the purpose of data collection, avoiding problems of imperfect recall. In many circumstances, information on the universe of the population is available and where information from different data sources (government agencies) can be connected, it can provide a very rich picture of study subjects.

Unfortunately, using administrative sources of information in one’s research comes with some drawbacks, in particular regarding data management and sharing. Due to the confidential and proprietary nature of these data, they can invariably not be shared freely with other researchers. This is a problem for the scientific community as it constrains verification of research methods and replication of results. It also makes it harder for a field of study to build upon prior research and collectively move the discipline forward. Access is often cumbersome, which hampers collaboration, slows research down, and tends to limit the sophistication of the research methods used.

On balance, there is certainly a place for administrative data sources in research, but their special nature needs to be acknowledged and accommodated in scientific journals’ data policies and granting agencies’ open access requirements.
Bio – Prof. Jo Van Biesebroeck

Jo Van Biesebroeck is a Research Professor (BOF) in Economics at the University of Leuven (KU Leuven). He obtained his PhD at Stanford University in 2001 and worked until 2008 at the University of Toronto. His fields of specialization are international trade and industrial organization. His research interests are in productivity, outsourcing, the automobile industry, and the Chinese economy; and in particular the interaction of these topics. He has advised the Canadian and Flemish government on trade and industrial policy and is regularly requested to write policy reports for international agencies (World Bank, OECD, EC, ILO, NBB, etc.). In 2009 he was the first Flemish economist to obtain an ERC grant to study the impact of China’s economic reforms. Additional information can be found at: www.econ.kuleuven.be/public/n07057/