Scanner Data Sources

Michael Smedes, UNSD

Transactions Data

- Scanner Data
  - The launch of barcode scanner technology during the 1970s, and its growth in the 20th century, enabled retailers to capture detailed information on transactions at the point of sale.
  - Scanner data are high in volume and contain information about individual transactions, including date, quantities and values sold, and detailed product descriptions.

- Other Transactions Data
  - Billing Data from telecommunication providers, insurance companies etc
  - Much more but to be used in a Price Index must have product level detail

Obtaining Scanner Data

- Two main options:
  - Directly from retail businesses
  - From a third party data provider

- Considerations:
  - Cost of data set
  - Scope of items included
  - Level of item aggregation
  - Temporal coverage and detail (day, week, month)
  - Scope of store (and regional) coverage
  - Timetable to provide data
  - Ability answer data queries

Obtaining Scanner Data

- Challenges
  - Confidentiality – turnover and quantities at the store level
  - Commercial value – data is seen as a private resource
  - Trust in the NSO
  - Legal and institutional settings
  - Technology and extraction costs

Obtaining Scanner Data

- Countries so far have largely gone with direct collection, mainly because it is seen as:
  - Being the cheaper open (don't pay for data but may pay for 'set up' costs
  - More in the control of NSOs (direct relationship with the data collector)

- But, the GWG on Big Data aims to build public-private collaboration beyond what has historically been the case and is working closely with third party data providers. Potential benefits:
  - Have existing access to data (including product level data from non scanner data sources) from across many companies (negotiate once)
  - Have already undertaken some data cleaning
  - Have already undertaken some data sorting and classifying

Obtaining Scanner Data

- Direct Collection
  - Institutional Environment and culture
  - Advantages for retailers (absence of field collectors, information provided back, well run economy, recognition)
  - Statistical Law

- Build a relationship over time, initially you will probably need to accept
  - Ingesting the data in the manner they wish to provide it
  - Each companies data format
  - Each companies classification structure
  - etc

- Can look to harmonise towards common NSO design over time

Obtaining Scanner Data

- Statistical Law
  - Often uncertainty as to how 'far' the law extends with regard to Big Data as the statistical legislation is often not specific – typically legislation was designed to enable collection via a survey form
  - Compulsion can be a very useful tool in obtaining data but many countries that have pursued scanner data so far have not needed to use it – better to have powers to compel but not always essential
  - Companies sometimes prefer to be compelled (even if willing to cooperate voluntarily) because it gives them legal cover in the event something goes wrong
  - UN generic statistical law suggests NSOs should have power to require provision of big data for public policy purposes
  - New laws that are being developed in OECD countries are typically providing quite explicit and wide ranging ability for NSOs to collect big data (UK, France, NZ for example)
  - Should not over rely on legislation however, very expensive both in $ and in reputation/relationships

Third Party Data Providers

- Market Research Companies – Nielsen, GfK etc
- Retailing portals online – Amazon, eBay etc
- Have ready access to data and have already taken some steps in making data 'analysis ready'
- UN GWG on Big Data is exploring opportunities for these types of organisations to be Trusted Data Providers on the Global Platform
- Big opportunities, the major (only?) challenge is likely to be cost

Developing an Information Technology System

- Need to have tool to 'ingest' data, typically some kind of FTP. Most companies are very concerned about exposing their data through the transfer process and may go to some lengths to reduce risks.

- Do not need sophisticated software to process the data – R, SAS etc is sufficient. Global Platform can potentially provide this service.

- However does require storage space and processing power due to very large number of records, access to servers.

Classifying Scanner Data

- Stores will have their own approach to sorting or classifying data into categories.
- Typically these have been designed to meet store operational needs – stock keeping, ordering etc
- Require and approach to sorting GTINs/lowest level product group into COICOP categories: mapping store categories to COICOP, machine learning, human processing
- New area of focus for the Task Team: guidance, case studies etc

Quality Assurance

- Input Data checks – size of data set is a challenge
  - Size of the file
  - Number of products
  - Length of time series
  - Change in products
  - Number of zeros
  - …..
- Output Data checks
  - When used as a new data source only (i.e. without changing methods) output processing tasks can typically remain largely the same
  - However introducing new methods does require new thinking on output data editing, it can become more complicated to attribute causation to any particular result

- Getting data analysis ready is another focus of the Task Team

Managing Risks

- The use of scanner data changes the risks faced by an NSO in producing a fit for purpose CPI
  - Risks to quality should be reduced overall – better coverage of items and regions for example
  - Some operational risks will be lowered – for example price collector strikes or lack of staff
  - But there will be some new operational risks through the increased reliance on cooperation from private sector organisations

- Risks of non-delivery may be the largest
  - Contracts, MoUs, agreements etc in place
  - Some backup plan to collect data through the short term
  - Some additional time in the schedule to manage late delivery
  - While these risks must be considered they should not be overstated

QUESTIONS?