Non-survey sources of data

UN Committee of Experts on Big Data and Data Science for Official Statistics

Mr. Ronald Jansen,
Chief, Data Innovation and Capacity Branch,
Statistics Division,
UN Department of Economic and Social Affairs
New York
Overview of presentation

UN-CEBD

- Mobile Phone data for Information Society
- Scanner data/ Webscraping for Price Statistics
- Access to global private sector data
- Privacy Preserving Techniques
- Training & Skills
UN Committee of Experts on Big Data and Data Science for Official Statistics

United Nations Statistics Division
UN-CEBD Management

- South Africa
- United Kingdom
- Denmark
- Brazil
- China
- Rwanda
- UAE
- Canada
- Colombia
- Indonesia
- Mexico
- Netherlands
- African Development Bank
- Eurostat
- OECD
- UNECA
- UNECE
- UNESCAP
- UNSD
Overview of presentation

UN-CEBD

Mobile Phone data for Information Society

Scanner data/ Webscraping for Price Statistics

Access to global private sector data

Privacy Preserving Techniques

Training & Skills
Mobile Phone Data
Task Team of the UN Committee of Experts on Big Data and Data Science for Official Statistics

Introduction
The statistical community has the obligation of exploring the use of new data sources, such as Big Data, to meet the expectation of the society for enhanced products and improved and more efficient ways of working. Use of Big Data could also support the monitoring of the Sustainable Development Goals (SDGs) by improving timeliness, frequency, detail and relevance of indicators without compromising their impartiality and methodological soundness. The reports of the UN Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD) to the Statistical Commission (E/CN.3/2015/4, E/CN.3/2016/6, E/CN.3/2017/7, E/CN.3/2018/8 and E/CN.3/2019/27) provide additional background to the work of the task teams.

Mobile phones are used by large parts of the population in all parts of the world, and it is thus expected that Mobile Phone data could fill data gaps worldwide. In its 2018 "Measuring the Information Society: Report of the World Telecommunication Development Conference" the ITU acknowledged mobile data as important for measuring the SDGs.

Publications
Handbook on the use of Mobile Phone data for Official Statistics

Events
Oman's Experience in Utilizing Mobile Positioning Data for Official Statistics
Virtual Webinar  12 Apr 2021
Handbook on Information Society

Created by UNSD Catheryn Tejón on Aug 21, 2020

METHODOLOGICAL GUIDE ON THE USE OF MOBILE PHONE DATA TO MEASURE INFORMATION SOCIETY INDICATORS FOR THE SDGs

Outline

1. Introduction
2. Applications
   a. Mobile phone data in research and pilots applicable for statistics (literature review, if any)
3. Data Sources
   a. Access to administrative data (from NSO) such as land area maps of lowest administrative unit (LAU)
   b. Access to mobile phone data (here include information related to ethical/legislative/data protection laws)
   c. Data for Calibration
   d. Data extraction processes
4. Methods
   a. Concepts and definitions (including the indicators to be calculated)
   b. Quality assurance framework for producing statistics with mobile phone data
       i. Quality assurance of the data samples from MNOs (before large datasets are transferred)
       ii. Quality assurance of the raw MPO (to make sure of data integrity and reliability, and to provide guidelines to develop the methodology into scripts)
   c. Creating scripts to process the data into the necessary indicators
   d. Data processing methodology
      i. Processing of the data
   e. Quality checking of the results
   f. Calibration and Inference Method (Estimation/Extrapolation Method)
      i. Developing extrapolation models to get to total population (considering the likely scenario not all mobile operators will participate or be able to send quality data)
5. Case Study – Experiences from countries
   a. Brazil
   b. Indonesia
• SDG indicators that can be measured using mobile phone data

9.C.1 PROPORTION OF POPULATION COVERED BY A MOBILE NETWORK

17.8.1 PROPORTION OF INDIVIDUALS USING THE INTERNET
Processing Model

MNO collects and stores raw MPD and cell location information in their systems.

MNO performs raw MPD and cells data QA according to checklist provided by Positium.

MNO calculates home locations (Place of Residence) for subscribers on municipality level.

MNO joins home location information with internet usage information (mapping with cells table).

MNO provides aggregated home locations information with internet usage information.

MNO provides cell location information.

Positium performs QA on:
- Aggregated home data-set
- Cell locations data set
- Reference data
- Local Administrative Units data
- WorldPop population data

Positium provides methodology for calculating the SDG indicators:
- 17.8.1 Proportion of individuals using the internet
- 9.c.1 Proportion of population covered by a mobile network

Indicator calculation:
- 17.8.1 Proportion of individuals using the internet
- 9.c.1 Proportion of population covered by a mobile network

Indicator QA:
- 17.8.1 Proportion of individuals using the internet
- 9.c.1 Proportion of population covered by a mobile network
Proportion of population covered with 4G mobile network

Coverage area radius set at 10km, flat approach
Population Covered by Mobile Cellular Network (4G) in Bali Province

Land Area Covered by Mobile Cellular Network (4G) in Bali Province
Overview of presentation

UN-CEBD

Mobile Phone data for Information Society

Scanner data/ Webscraping for Price Statistics

Access to global private sector data

Privacy Preserving Techniques

Training & Skills
Introduction

The statistical community has the obligation of exploring the use of new data sources, such as scanner and web scraped data, to meet the expectation of society for enhanced products and improved and more efficient ways of working. The task team on scanner data is created as a separate team, since scanner data is one of the Big Data sources which is used more and more in national statistical systems for the calculation of price indices. Many of the price measurement issues and methods for scanner data from supermarket chains and other retailers apply also to other big data sources. For example, online prices obtained from web scraping.
Prices e-Handbook

- Glossary
- Initial considerations
  - Basic concepts
  - Sampling
  - Quality assurance
  - IT system requirements, availability of tools
- Data acquisition
  - Scanner data
  - Web scraping
  - Other
- Preparing the data for use in production of CPIs
  - Standardising the data
  - Averaging
  - Identifying unique products
  - Treatment of discounts, refunds
  - Prices and weights in scanner data
- Classification (workstream 2)
- Data filtering
  - Outlier filter
  - Dumpling filter
  - Low sales filter
- Price Indices
  - Bilateral
  - Multilateral
  - Extension methods
  - Decomposition
  - Choosing an index method
  - How to optimize computational performance
- Aggregation
  - Methods for aggregation within a data source
  - Methods for aggregation with other data sources
  - Other considerations

Recently updated

- Multilateral
  - yesterday at 6:29 PM • updated by Jacek Białek • view change
- Bilateral
  - yesterday at 5:42 PM • updated by Jacek Białek • view change
- Data requirements specification for web scraped data
  - yesterday at 3:47 PM • commented by Liam Greenough
- If in house, what strategies are available?
  - yesterday at 3:46 PM • commented by Liam Greenough
- If in house, what strategies are available?
  - yesterday at 3:40 PM • updated by Lincoln Teixeira da Silva • view change
- Data requirements specification
  - yesterday at 3:39 PM • commented by Liam Greenough
- Monitoring, validation, and plausibility checks
  - yesterday at 3:35 PM • commented by Liam Greenough
- How to obtain the data from retailers
  - yesterday at 3:11 PM • commented by Liam Greenough
- Extension methods
  - yesterday at 3:04 PM • commented by Liam Greenough

Show More
Data acquisition

Created by UNSD Clarence Lio, last modified by Tanya Flower on Apr 20, 2021

• Scanner data
  • How to obtain the data from retailers
  • Data requirements specification
  • Data sharing agreements
  • Alternative approaches
  • Monitoring, validation, and plausibility checks
• Web scraping
  • Different approaches to accessing web scraped data
  • If in house, what strategies are available?
  • Monitoring/validation/plausibility checks
  • Data requirements specification for web scraped data
  • Common technical problems
• Other
Practical guidelines on web scraping for the HICP

November 2020

*Harmonised Indices of Consumer Prices*
Prices from digital sources

Web Scraping data for:
- Clothing stores
- General Merchandisers
- Home improvement
- Electronics and Appliances

API data for:
- Airlines
- Hotels
- Car Rentals

Scanner data for:
- Food
- Personal Care
- Household operations

In-house Internet collection of:
- Travel
- Transportation
- Communications
- Furniture
- Services
# Integration of online pricing

## CPI aggregates fully priced online
- Hotel accommodations
- Air fares
- Rental cars
- Telephone services (local and cellular)
- Internet access services
- Inter-city trains and buses
- Local transit
- Driver’s licenses
- Passports
- Passenger vehicle registration fees
- Retail club memberships
- School books
- Cars

## CPI aggregates partially priced online
- Travel tour packages
- Furniture
- Mattresses
- Household textiles (sheets, towels, window coverings)
- Toys, games and hobby supplies
- Household appliances
- Cookware
- Tableware
- Tablets
- Photographic equipment and supplies
- Audio equipment
Implementing more online data in the future

Web Scraped
- Online vs instore
  - Clothing
  - Electronics

Application Programming Interfaces (APIs)
- Travel

Sharing Economy
- Netflix & Spotify
- Uber & Lyft
- Airbnb

Future Trends
- Scope definition could extend beyond country
- Growth online marketplace
- New product offers: Internet Of Things
Overview of presentation

UN-CEBD

Mobile Phone data for Information Society

Scanner data/ Webscraping for Price Statistics

Access to global private sector data

Privacy Preserving Techniques

Training & Skills
Task Team on Acquisition of Global Private Sector Data

Objectives

Approach global companies to negotiate access to their global data sources under global arrangements strictly used for statistical purposes to inform policies at national, regional and global levels, especially to advance the implementation of the 2030 Agenda for Sustainable Development.

Deliverables

• Evaluation of priorities for global private sector data sources
  • Credit card companies (e.g. Master Card)
  • E-Commerce platforms (e.g. Amazon, Alibaba)
  • Digital intermediaries (e.g. AirBnB, Uber)
  • Mobile Network Operators (e.g. Vodafone, Telenor)
• Negotiation with providers
• After terms of access for new data source is negotiated there will be a transition of operational management to the relevant UN Big Data task team or statistical organization, including access via UN Global Platform where relevant.
• Communication
Overview of presentation

UN-CEBD

Mobile Phone data for Information Society
Scanner data/ Webscraping for Price Statistics
Access to global private sector data
Privacy Preserving Techniques
Training & Skills
Introduction

The Privacy Preserving Techniques Task Team (PPTTT) is advising the UN Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD) on Big Data on developing the data policy framework for governance and information management of the global platform, specifically around supporting privacy preserving technique.
UN Privacy Preserving Techniques Handbook

In this UN handbook, we define specific goals for privacy-preserving computation for public good in two salient use cases: giving NSOs access to new sources of (sensitive) Big Data; and enabling Big Data Collaborations Across Multiple NSOs.

Link: https://docs.google.com/document/d/1GYu6UJJ1jrR8LgooXVDsYk1s6FIM-SbOvo3oLHglFhY

Partner: The GWG Task Team on Privacy Preservation Techniques
Draft Overview and Structure PPT Handbook v2

Dashboard / ... / Regular task team

Draft Document Framework / Overview - Google Documenten

Introduction to Method General Template

We have spoken about the need for a general template for the Concepts and Settings section. We should iterate on this as a group, but as an initial starting point, I've outlined a rough template to act as a starting point. Each subsection should have a broadly similar length explanation and example.

Problem Definition

The first element to each chapter, as was suggested in the last call by Julien, would be to describe the problem being solved in abstract terms. This would be similar to the patent communities approach. Essentially, we describe what the challenge is, who are the parties involved, what do the input, what is outputted and to whom.

Example Use Case

We would then follow the problem definition by a tangible use case that is intuitive and self-explanatory. The goal would be to solidify the more abstract description with a memorable setup that the reader can likely relate too. These use cases do not need to be real examples i.e. they can be simplifications and we can avoid confusing the challenge addressed by the technology with other factors which exist in specific real-world use cases.

Overview & History

Now that the reader understands what the challenge is (i.e. the what), we can describe the mechanisms and artefacts of the technologies in the family of techniques (i.e. the how). We can discuss the different variants, history of the technology and so on. The goal of this subsection is to give the reader a good understanding of the topic at a high level and importantly direction of where to look to get a further depth of knowledge on the topic.

Security Model

To be deployed, a PPT usually needs to be accepted by the security team of an organization. All PPTs have trade-offs in terms of the trust required in other parties, the cryptographic or hardware assumptions that provide the security.

Cost of Technology

The bandwidth and computational costs, along with the rounds of communication of approaches have traditionally been a limiting factor for MPC and HE. This is not the case for other techniques such as DP and synthetic data. In both cases, floating-point numbers can cause issues (usually fixed-point values are leveraged). Other limitations, such as HEs polynomial function constraints should also be highlighted. This subsection should outline the technical barriers and limitations of the technology to the reader, perhaps identifying typical efficient and inefficient queries or protocols in the approach.
Overview of presentation

**UN-CEBD**

- Mobile Phone data for Information Society
- Scanner data/ Webscraping for Price Statistics
- Access to global private sector data
- Privacy Preserving Techniques

**Training & Skills**
Introduction

The statistical community has the obligation of exploring the use of new data sources, such as Big Data, to meet the expectation of the society for enhanced products and improved and more efficient ways of working. Big Data could also support the monitoring of the Sustainable Development Goals by improving timeliness and relevance of indicators. This should go without compromising their impartiality and methodological soundness.

Big Data is by definition different from traditional data sources currently used by National Statistical Systems (NSSs) requiring the development of new methodologies. Big Data sources pose challenges regarding methodology, quality assurance, technology, security, privacy and legal matters. This means that new skill sets are necessary. Some of which could be hired temporarily, others will need to become an integral part of the institution. It is up to the senior management to decide what will be done by the institute itself and what will be outsourced. Most likely, the statistical institute will need to build long-term partnerships with private sector, academia and research institutes to successfully work with new data sources and new technologies.
Courses.OpenMined

Our Privacy Opportunity

Beginner

7.7 hours

Foundations of Private Computation

Intermediate

60 hours

Federated Learning Across Enterprises

Coming Soon

Federated Learning on Mobile

Coming Soon
Conclusion

→ Join the Task Teams of UN-CEBD

Mobile Phone data for Information Society

Scanner data/ Webscraping for Price Statistics

Access to global private sector data

Privacy Preserving Techniques

Training & Skills
Thank you

Ronald Jansen
UN Statistics Division

jansen1@un.org
BigData@un.org
@UNBigData
https://unstats.un.org/bigdata/