Overview of non-traditional data sources for the SDGs and ensuring data quality

Benjamin Rae
United Nations Statistics Division

Workshop on Data and Statistics for Evidence Based VNRs
11 December 2019, Vienna, Austria

Contents

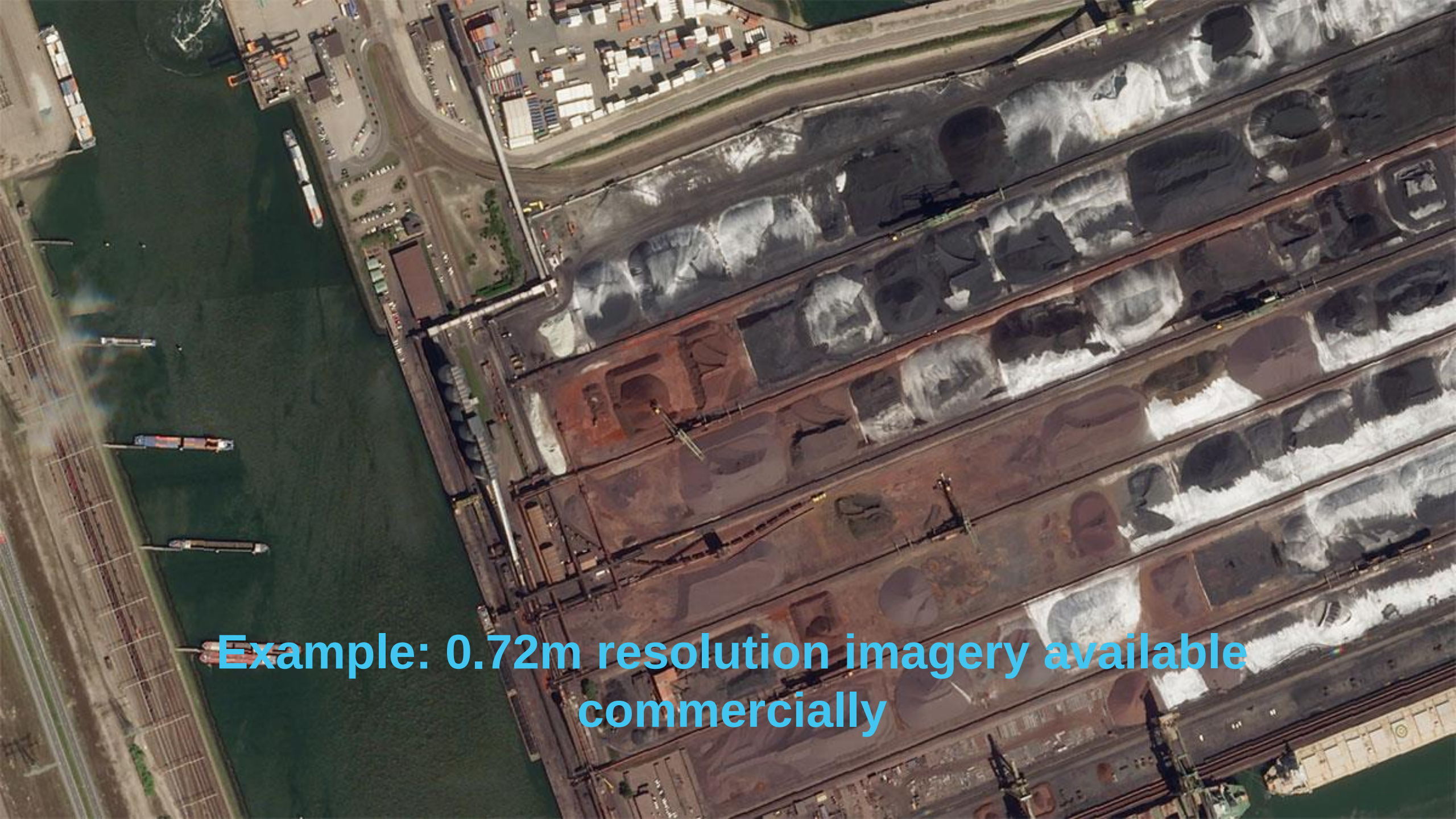
- Non-traditional data sources
- The NEW United Nations National Quality Assurance Frameworks Manual for Official Statistics
- Quality assurance of SDG indicators
- Quality assurance for statistics compiled from non traditional data sources

Non-traditional data sources

Non-traditional data: motivation

- Nearly all market transactions are digital or will become so soon
- Costs of surveys and other traditional methods are not going down
- Yet methodologies for producing official statistics have not adapted to this “data-rich” environment
- Methodological opportunities and challenges:
  - Price and quantity data can be simultaneously collected or aggregated at the source
  - Satellite images can record changes to the earth’s surface in real time and identify unique varieties of crops and flora
  - Mobile phones can record the movement of individuals

How do we build national statistics from transactions and geospatial data in addition to surveys?


Example: 0.72m resolution imagery available commercially

Different sources of non-traditional data for official statistics

Transactions: Digital interactions with people result in large-scale collection of passive data by companies and service providers

- Interactions with digital systems leave exhaust data that can be linked to you and analyzed: mobile phones, social media, Uber, Google, Amazon

Sensing: Remote sensing, satellites, sensors, open data, tollbooth cameras, meters, POS scanners, job vacancy postings, Yelp data

- Passive observation and monitoring: counting ships in ports or cars in parking lots using satellite imagery, crop identification using drones, counting pedestrian traffic, measuring prices using point of sale data

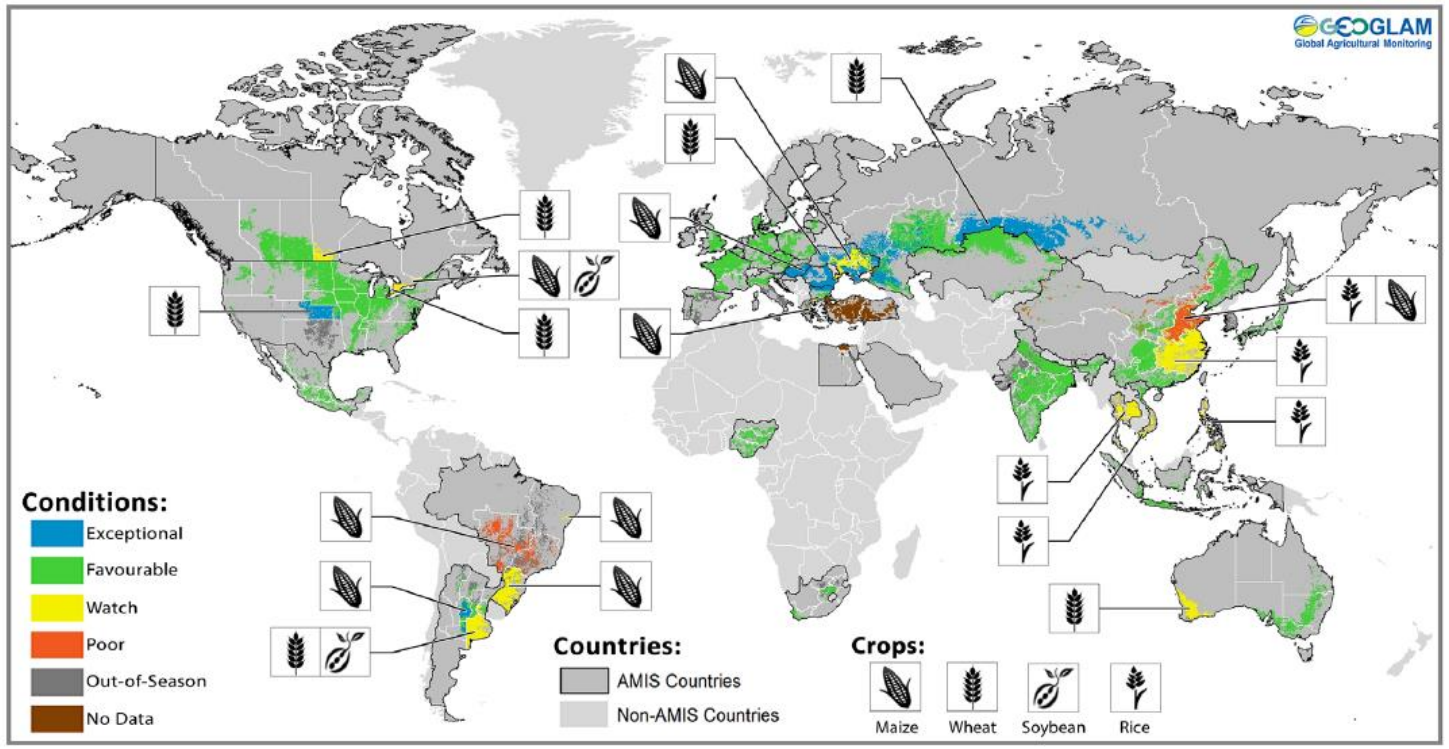
Previously unanswered questions are being answered

- Many applications of big data to official statistics are in the early stages of research and genesis
- In many areas the current state of the art is rapidly expanding:
  - Can big data from online platforms like Yelp help to measure local economic activity? (Glaeser, Kim and Luca 2017)
  - Can private payroll data increase the accuracy of real-time measurement of unemployment? (Cajner, Crane, Decker, Hamins-Puertolas, Kurz, 2019)
  - Can natural language processing reduce manual evaluation of survey responses and improve survey accuracy? (Baer, Jensen, Klimek, Singh, Staudt, and Wei, 2019)
- In other areas, new sources of data are leading to the proposal of new kinds of official statistics altogether:
  - Using job vacancy postings and natural language processing that firm-based surveys cannot easily collect.

In key fields, non-traditional data sources are being used to produce official statistics

- Mobility and tourism statistics produced from mobile phone data in several countries, led by Positium
- Statistics Netherlands uses supermarket scanner data to produce elements of its CPI
- Statistics Canada uses remote sensing imagery from satellites instead of a crop survey
- Phillipine Statistics Authority measures rural access to roads using geospatial information
- Using mobile phone data to estimate migration movements in several countries

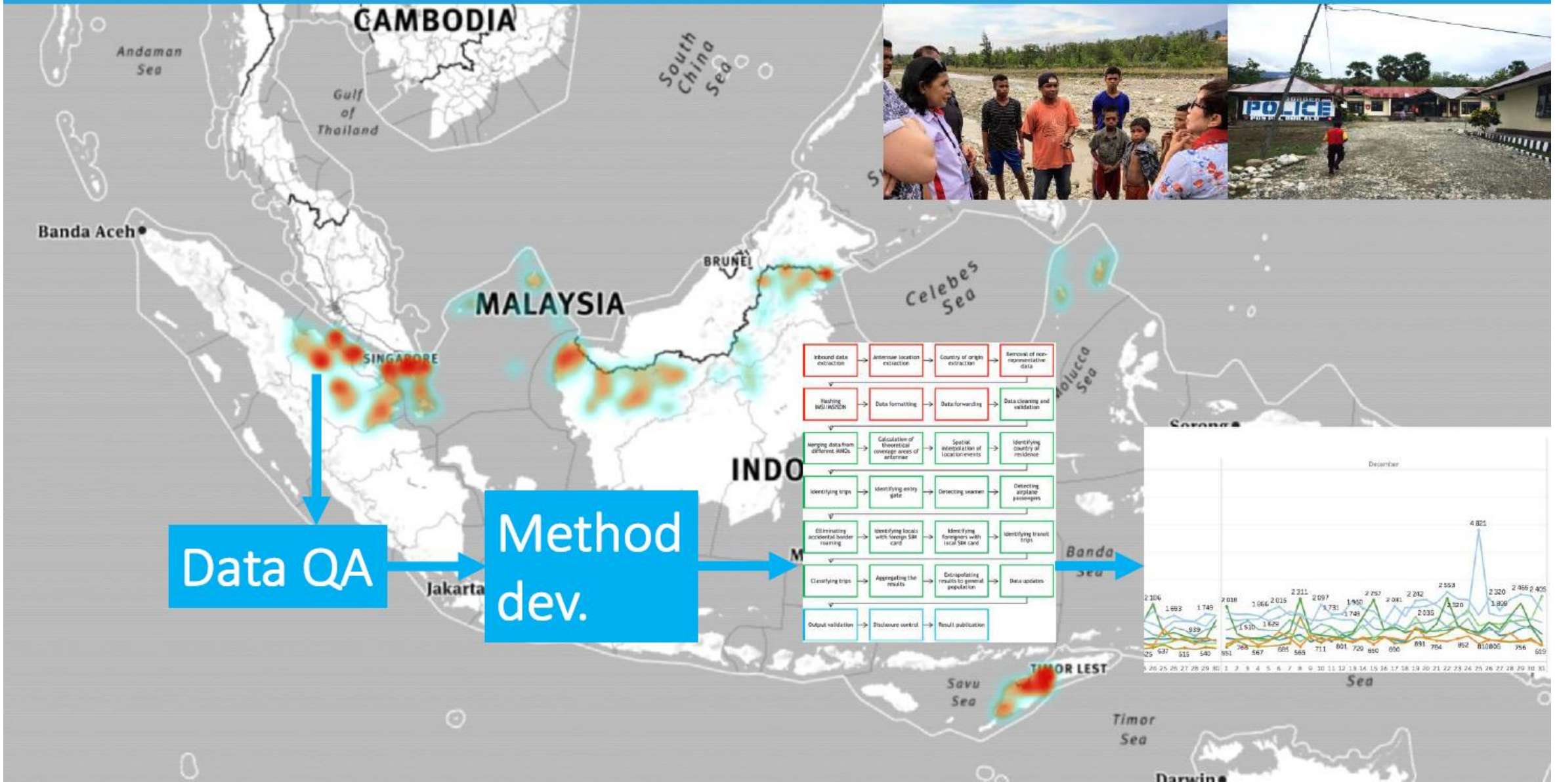
However, large scale adoption will depend on progress in methodology, software development, information architecture, and secure computation



Agricultural crop production (SDG 2.4.1) with Satellite data

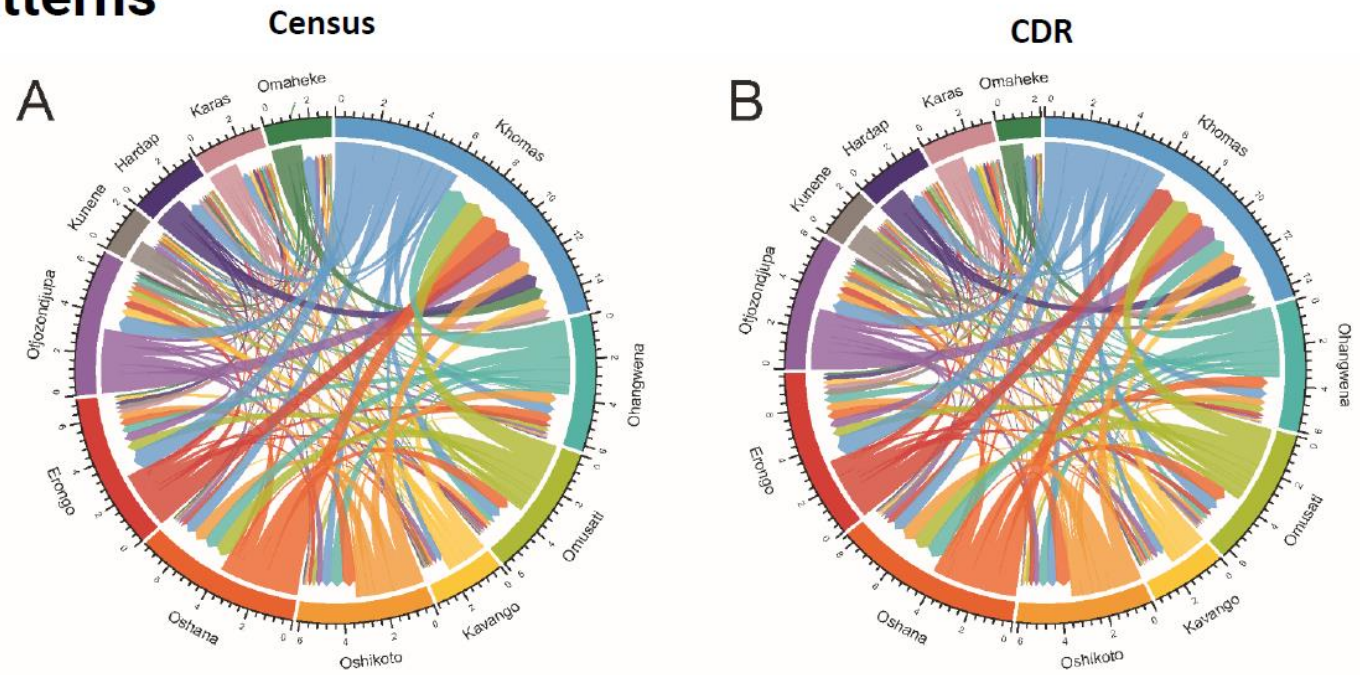
- GeoGLAM Crop Monitoring.
- EO in Service of the 2030 Agenda for Sustainable Development. Anderson *et al.* 2017

Case: Cross-Border Tourism in Indonesia



Human mobility (SDG 8.9.1 and 10.7.1) and population densities with Mobile Phone data

CDRs and census data show very similar migration patterns



Note: The Zambezi region as an outliner is excluded.

Current challenges

- Different owners of data
- Confidentiality concerns (corporate secrecy and individual privacy)
- Different data vocabularies and definitions

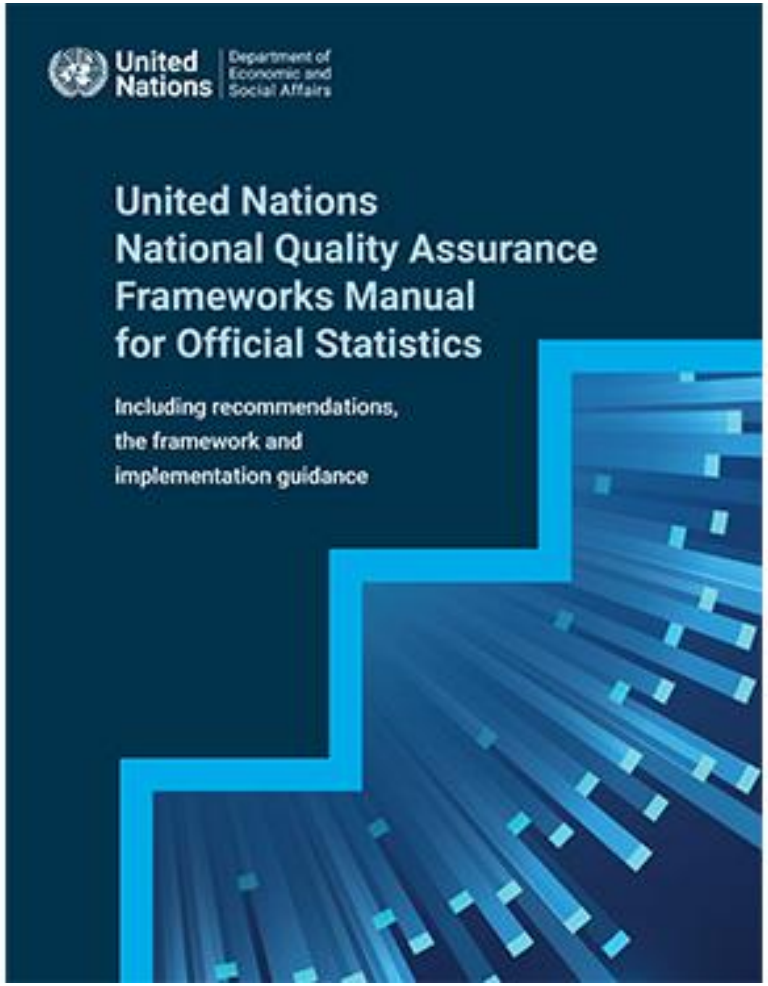
These all make integration challenging

- Current statistical methodology research often lags behind data availability
- Data gravity means that analysis often must be done close to the data source, which requires access; yet…
  - Current research on secure computation and encrypted analysis is in early stages and is not generally scalable to large volumes—meaning it is still hard to share large volumes of confidential data among untrusting parties

United Nations National Quality Assurance Frameworks Manual

SUSTAINABLE DEVELOPMENT GOALS

United Nations National Quality Assurance Frameworks Manual



The Manual

- respond to the ***new data ecosystem***,
- fueled by technological advances,
- and new demands for detailed and timely data for policymaking,
- in the context of the 2030 Agenda for Sustainable Development

United Nations National Quality Assurance Frameworks Manual

Five core recommendations

#1 Integrate Fundamental Principles of Official Statistics in the legal and institutional frameworks

#2 Include the requirement for quality assurance in the statistical legislation

#3 Establish national quality assurance framework (NQAF); all national statistical system (NSS) members to commit to quality assurance

#4 Base or align your NQAF with international or regional quality frameworks

#5 Implement NQAF at the NSO, throughout the NSS and for data and statistics produced outside the NSS, as appropriate

Quality assurance of SDG indicators (Chapter 8)

Quality assurance of SDG indicators (Chapter 8)

- Challenges of assuring quality of data and statistics for the SDG indicators
- Roles of the different entities participating in this task
- Requirements and elements to be assured of special importance in this context according to the UN NQAF levels:
  - Managing the NSS
  - Managing the statistical environment
  - Managing the statistical processes
  - Managing statistical outputs

Quality assurance of SDG indicators (Chapter 8)

Challenge of assuring the quality of SDG indicators

- The production of the SDG indicators may ***involve all members of the NSS*** as well as new or ***non-traditional statistics producers*** and data providers;
- The set of identified global SDG indicators is ***large and diverse;***
- Countries are establishing their own ***national indicator frameworks*** according to their priorities;
- Global SDG indicators when adopted in 2017 were at ***different stages of methodological development***;
- The ***disaggregation*** of the global SDG indicators is a major challenge for countries.

Quality assurance for statistics compiled from non-traditional data sources (Chapter 7)

Quality assurance for statistics compiled from non-traditional data sources (Chapter 7)

List of other data sources (no classification)

- Cross-country sample surveys by supra-national organizations or international enterprises;
- Data compiled and maintained by private professional organizations or business associations or non-profit institutions in general;
- Data and records compiled and maintained and/or owned by enterprises that cover large parts of the population Earth observation and remote sensing;
- Thematic mapping and monitoring systems (e.g., field-monitoring stations for water quality, air pollution etc.);
- Research/scientific and pilot studies;
- Citizen generated data

Note: *New data sources can be often associated with other data sources but may as well belong to statistical or administrative data sources, depending on national circumstances.*

Quality assurance for statistics compiled from non-traditional data sources (Chapter 7)

- **Main advantage**: opportunity to overcome resource limitations, to allow much more frequent and timely reporting, provide more objective information and, most importantly, to be able to generate data on phenomena that are difficult or impossible to capture with traditional statistical and administrative data sources.

- **Major quality challenges**:
  - The limited access to other data sources and legal challenges regarding its access (as it can also be the case for administrative sources) requires arrangements with the data providers (e.g. government agencies, private sector and research institutions); lack of knowledge about the existence of such data; and sustainability of the source over time;
  - Incoherent use or lack of use of statistical standard concepts, definitions and classifications that put the accuracy, reliability, coherence and comparability of the resulting statistics in question;

Quality assurance for statistics compiled from non-traditional data sources (Chapter 7)

- **Major quality challenges (continued)**
  - Providers of data (which may be the owner or holder of the data) are not subject to and do not adhere to the Fundamental Principles of Official Statistics (FPOS) and associated statistical quality principles such as professional independence and quality commitment;
  - Utilizing data for statistical purpose may potentially put the confidentiality and privacy of individuals, households and businesses at risk depending how detailed data is being published;
  - Data of sources such as those from mobile phones or social media are not representative of the entire population and may cause serious selection bias when used for statistical purposes;

Thank you

For further information:

https://unstats.un.org/unsd/methodology/dataquality/

https://unstats.un.org/bigdata/