Introduction to the use of Big Data for Official Statistics

Ronald Jansen

Assistant Director

Chief of Data Innovation and Capacity Branch

United Nations Statistics Division

I.   Background
    a)   Data Revolution
    b)   National – US Big Data Group
    c)   Regional – UNECE Big Data Group
    d)   Global – UN GWG Big Data Group

II.   Examples of Use Cases

III.   Data Science, New Skills and Partnerships

IV.   Program of the Symposium

2014 –   Data Revolution report

2014 –   Creation of
US Big Data group
UNECE Big Data group
UN GWG Big Data group

2014-19   International Conferences on Big Data for Official Statistics in China, UAE, Ireland, Colombia and Rwanda
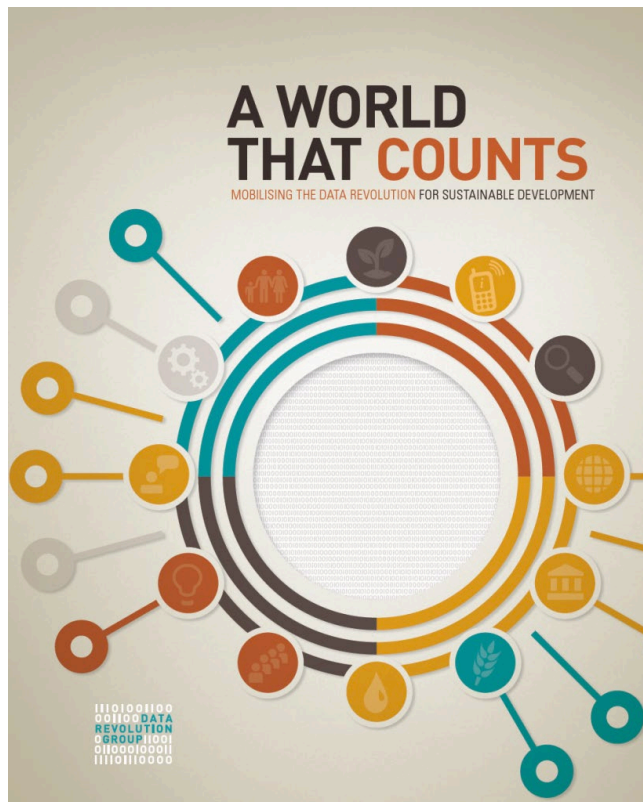
2018 –   UN Global Platform

2014

End
of
MDGs

Independent Expert Advisory Group (IEAG) gave recommendations to the UN Secretary General

- Co-Chairs **Enrico Giovannini** and **Robin Li,** CEO of Baidu
- 23 members – half from Statistics, half from outside Statistics
- **Amina Mohamed**, current Deputy Secretary General, and **Thomas Gass**, currently leading the Bern Network for Financing for Statistics, were representing the UN

2014 – Data revolution report – action items:

(i)   develop **legal, technical, privacy, geospatial and statistical standards** to facilitate openness and exchange of information;

(ii)   share technology and innovation for the common good; and create a "SDGs data lab";

(iii)   scale **investments for statistical capacity development and technology transfer**;

(iv)   mobilize global action through a World Data Forum with **global public-private partnerships** for data sharing;

2019 – What have we delivered?

(i)   GWG to develop **legal, technical, privacy, geospatial and statistical standards** for use of Big Data;

(ii)   UN Global Platform to share technology and innovation as a "SDGs data lab";

(iii)   Bern Network to scale **investments for statistical capacity development and technology transfer**;

(iv)   The 3rd World Data Forum will take place in October 2020 in Bern

US Big Data Group (2014)

Big Data Public Working Group of the National Institute of Standards and Technology ➜ really Big Data

- Fraud detection in the financial industries (banking, securities & investments, insurance)
- Persistent surveillance (object identification and tracking from high-resolution imagery or full motion video) by the US Department of Defense
- Genomic measurements
- Particle physics: analysis of Large Hadron Collider (LHC) data (Discovery of Higgs particle)
- Climate studies using the Community Earth System Model.

UNECE Big Data Group (2014)

- Under the umbrella of the Conference of European Statisticians the UNECE Big Data Working Group developed guidance on several issues, including
  - Classification for Types of Big Data
  - Big Data Quality Framework

United Nations Statistics Division

Classification of Type of Big Data

Social Networks (human-sourced information)

1100. Social Networks: Facebook, Twitter, Tumblr etc.

1200. Blogs and comments

1300. Personal documents

1400. Pictures: Instagram, Flickr, Picasa etc.

1500. Videos: Youtube etc.

1600. Internet searches

1700. Mobile data content: text messages

1800. User-generated maps

1900. E-Mail

Classification of Type of Big Data

Business systems (process-mediated data)

21. Data produced by Public Agencies
   2110. Medical records
22. Data produced by Businesses
   2210. Commercial transactions
   2220. Banking/stock records
   2230. E-commerce
   2240. Credit cards

Classification of Type of Big Data

Internet of Things (machine-generated data)

311. Fixed sensors
    3111. Home automation
    3112. Weather/pollution sensors
    3113. Traffic sensors/webcam
    3114. Scientific sensors
    3115. Security/surveillance videos/images
312. Mobile sensors (tracking)
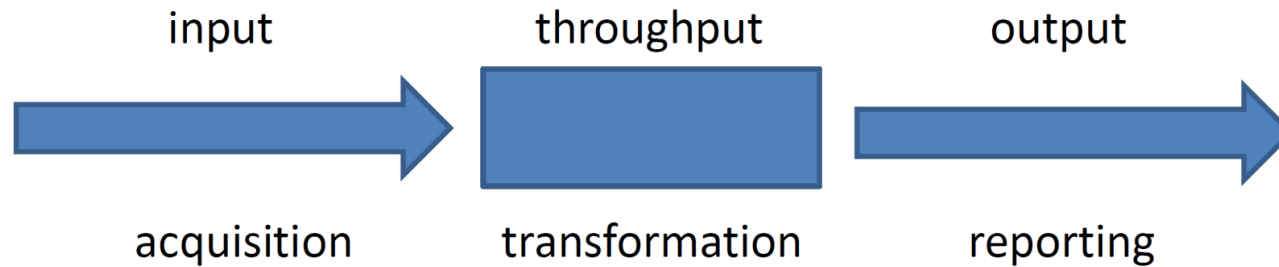   3121. Mobile phone location
    3122. Cars
    3123. Satellite images
32. Data from computer systems
    3210. Logs
    3220. Web logs

Big Data Quality Framework

Business process

input   throughput   output

acquisition   transformation   reporting

Framework

Structured view of quality for each phase

Big Data Quality Framework

Input

- Discovery stage
  - Dataset not required
  - Hyperdimension: Source and Metadata
- Acquisition
  - Dataset required
  - Hyperdimension: Data
- Approach
  - Factors to consider
  - Example of quality indicators

UN Global Working Group (GWG) on Big Data for Official Statistics

- **Created in March 2014 by the UN Statistical Commission**

- **Mandated to give direction to the use of Big Data for Official Statistics**

- **Consisting of 28 countries and 16 international organizations**

21/10/2019

UN GWG on Big Data for Official Statistics

o **Global Conferences**

- ✓ Beijing, China, 2014

- ✓ Abu Dhabi, UAE, 2015

- ✓ Dublin, Ireland, 2016

- ✓ Bogota, Colombia, 2017

- ✓ Kigali, Rwanda, 2019

2014-19 International Conferences on Big Data for Official Statistics in China, UAE, Ireland, Colombia and Rwanda

Themes

- Potential of Big Data (China), Big Data and SDGs (UAE), Public-private partnerships (Ireland), Data collaboratives and trusted data (Colombia), Working together, learning together (Rwanda)

Sessions

- Earth Observations, Mobile Phone data, Scanner data, Capacity development

Outcomes

- Reports to UNSC, Bogota declaration, Kigali declaration

2018 –   UN Global Platform – Alpha version made available by ONS, UK

o   Collaborative digital environment for trusted data, trusted methods, trusted partners and trusted learning

o   Global data sets - Automatic Identification System (AIS), Satellite data,

o   Global services - Trusted Methods Library/Service, Earth Observation Service, Location Analytics Service, Developers Service

UN GWG and its Task Teams: Intergovernmental body

UN Global Platform: Collaborative environment

UN Global Platform:
System of Hubs

✓ Hangzhou, China – NBS of China

✓ Dubai, UAE – FCSA, UAE

✓ London, UK – non-government (supported by Data Science Campus/ONS)

✓ Rio de Janeiro, Brazil – National School of Statistical Sciences/ IBGE

✓ Kigali, Rwanda – NISR, Rwanda and UNECA

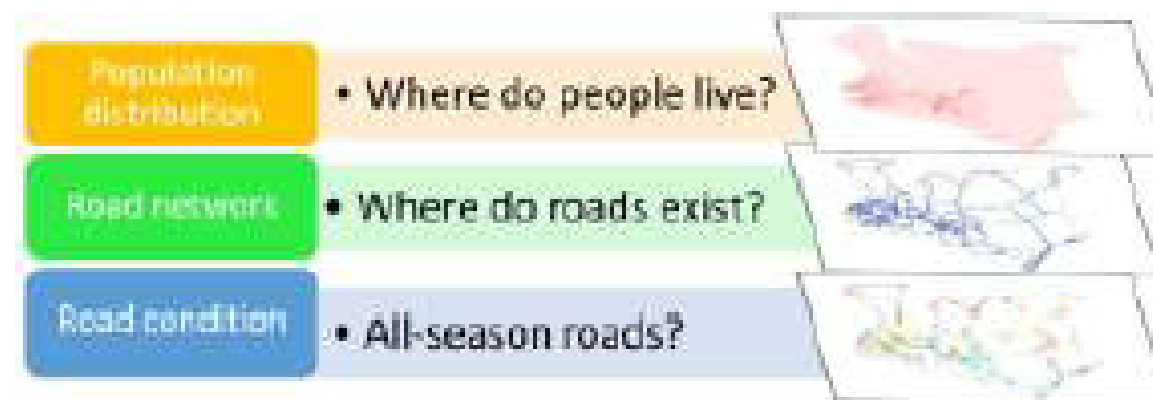United Nations Statistics Division

Examples

Use Cases of Big Data

Distance to all season road (SDG 9.1.1) with Open Street Map and Satellite data
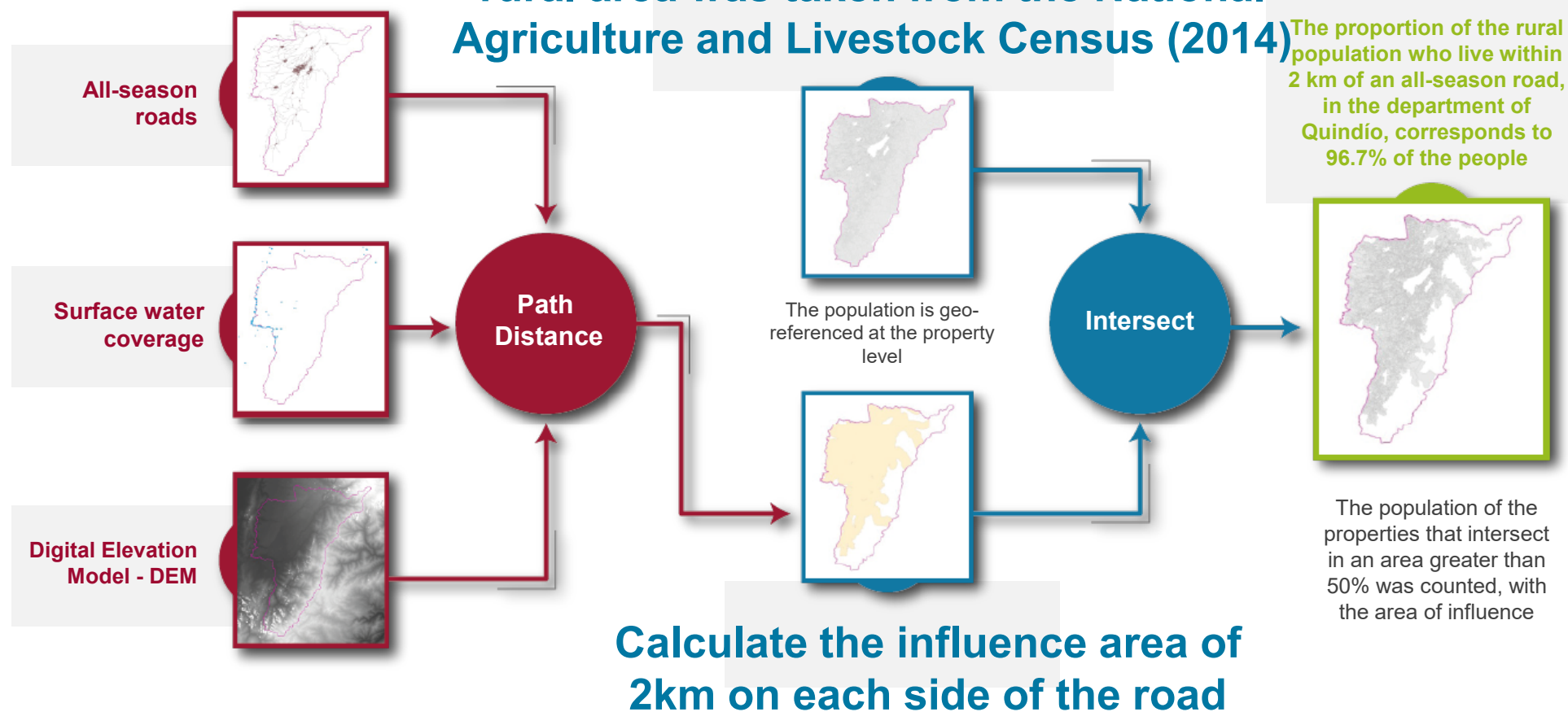
New GIS methodology for measuring RAI

In 2016 ReCAP and the World Bank (the RAI custodian), coordinated to develop a new methodology that used geo-spatial data to measure the RAI, and trialled it in eight countries in Africa and Asia. This included using three layers of geo-spatial information to define the RAI:
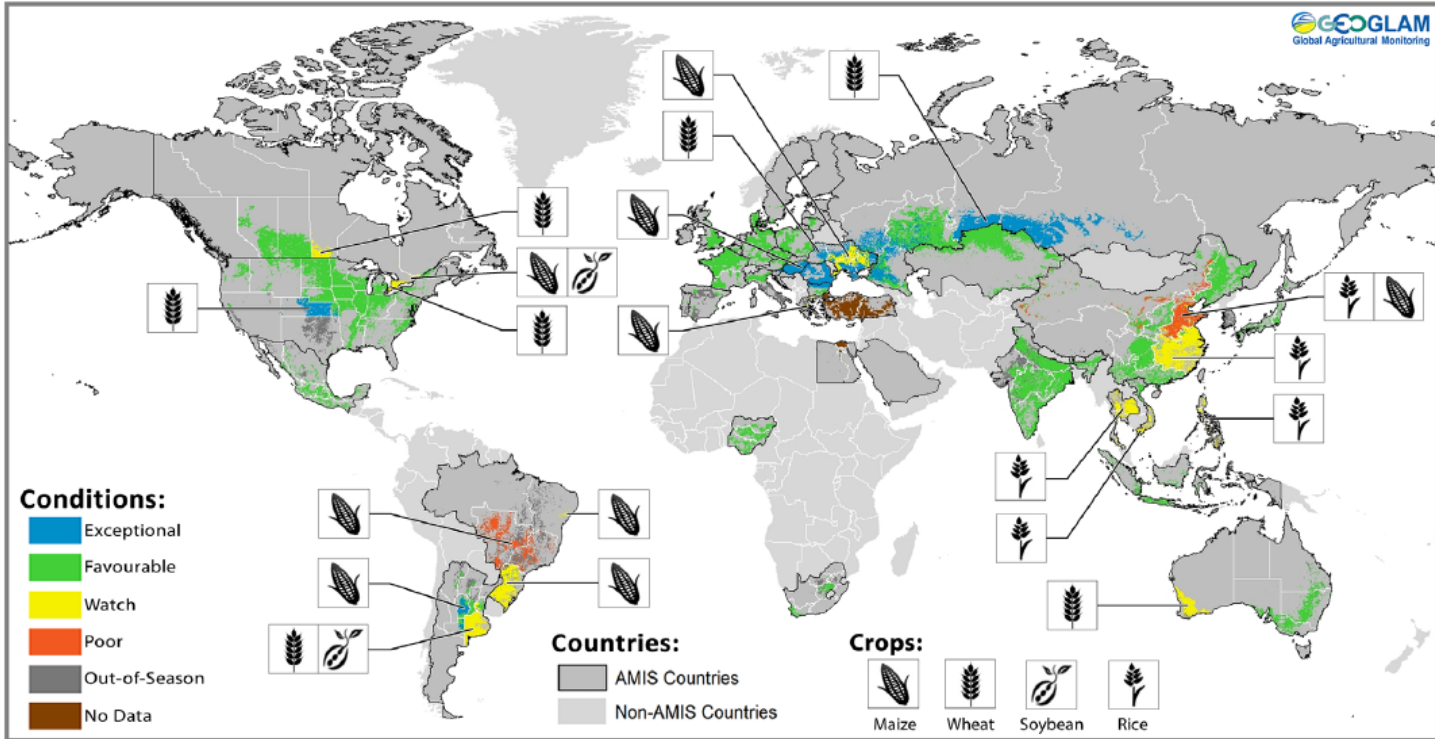- Rural Population
- Road Network Location
- Road Condition (all-season or not)

Population distribution • Where do people live?

Road network • Where do roads exist?

Road condition • All-season roads?

United Nations Statistics Division

The number of persons residing in the rural area was taken from the National Agriculture and Livestock Census (2014)

All-season roads

Surface water coverage

Digital Elevation Model - DEM

Path Distance

The population is geo-referenced at the property level

Calculate the influence area of 2km on each side of the road

Intersect

The proportion of the rural population who live within 2 km of an all-season road, in the department of Quindío, corresponds to 96.7% of the people

The population of the properties that intersect in an area greater than 50% was counted, with the area of influence

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

GOAL 6
"Ensure availability and sustainable management of water and sanitation for all"

TARGET 6.6
"By 2020, **protect and restore water-related ecosystems**, including mountains, forests wetlands, rivers, aquifers and lakes. "

INDICATOR 6.6.1
Percentage of change in the extent of water-related ecosystems over time

% of change in wetlands extent over time can be measured globally by earth observation based monitoring of wetlands looking at land-use, land-cover, vegetation cover, inundation frequency, biodiversity

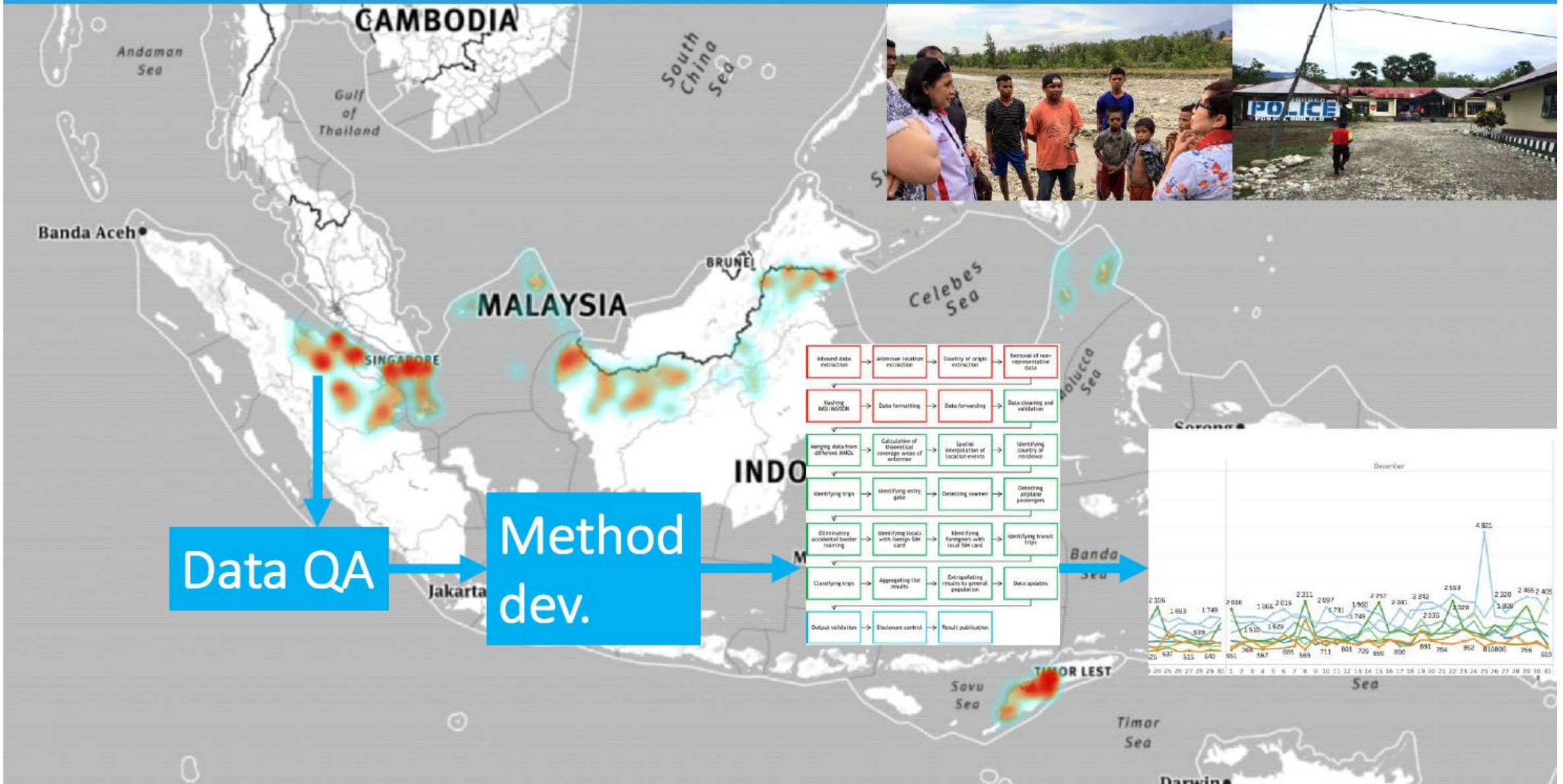Human mobility (SDG 8.9.1 and 10.7.1) and population densities with Mobile Phone data

Domestic Tourism

- 34 Province → 514 Kabupaten
- 256 Million People
- How Many Domestic Tourism ?
- How Much Their Expenditure ? For what
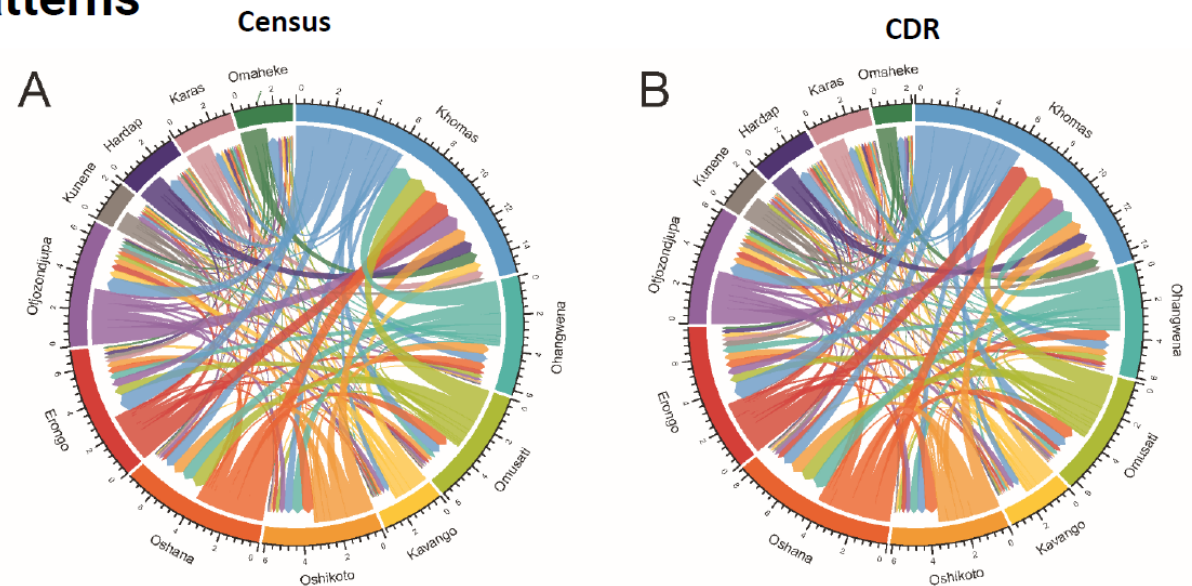- Where they come from, where they go
- Mode of Transport ?

1

514

2

United Nations Statistics Division
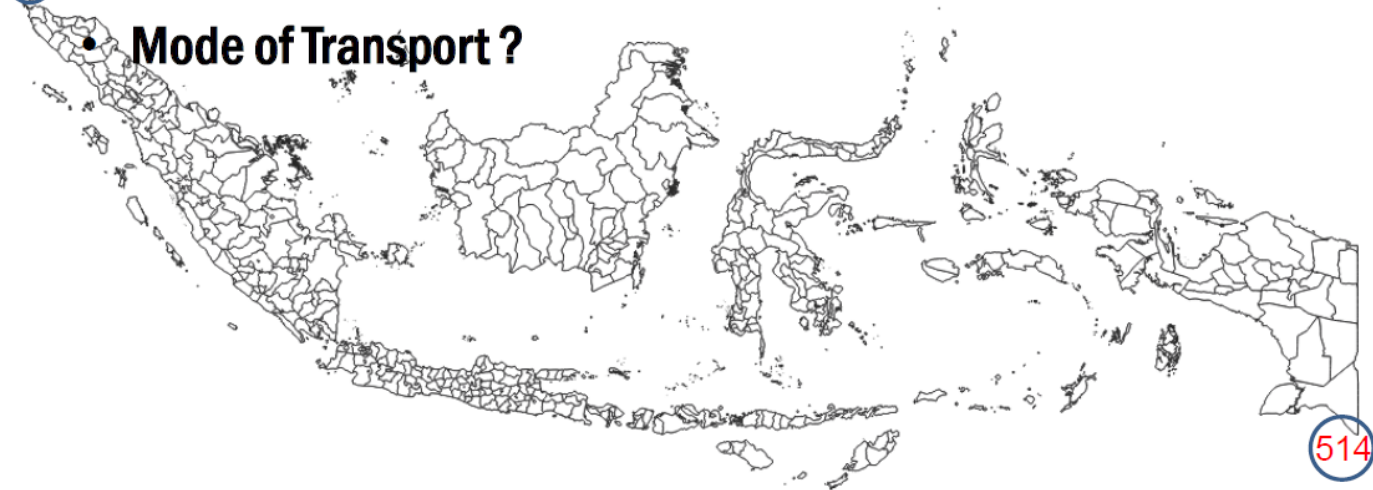
Case: Cross-Border Tourism in Indonesia

Data QA

Method dev.

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

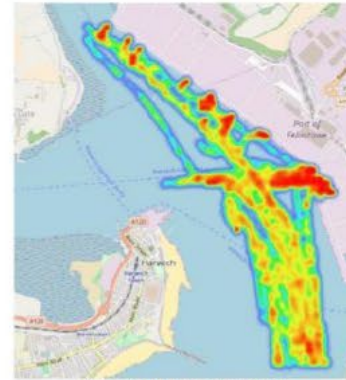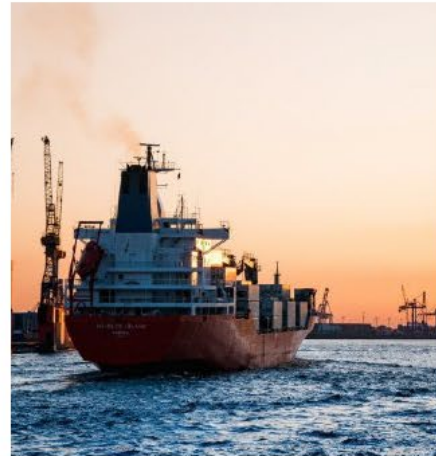**CDRs and census data show very similar migration patterns**

Note: The Zambezi region as an outliner is excluded.

Shipping Automated Information System (AIS) data

- Marine and Coastguard Agency, ORBCOMM, Global Platform
- Ship tracking data
- Port traffic frequency
- Time in port
- Real time

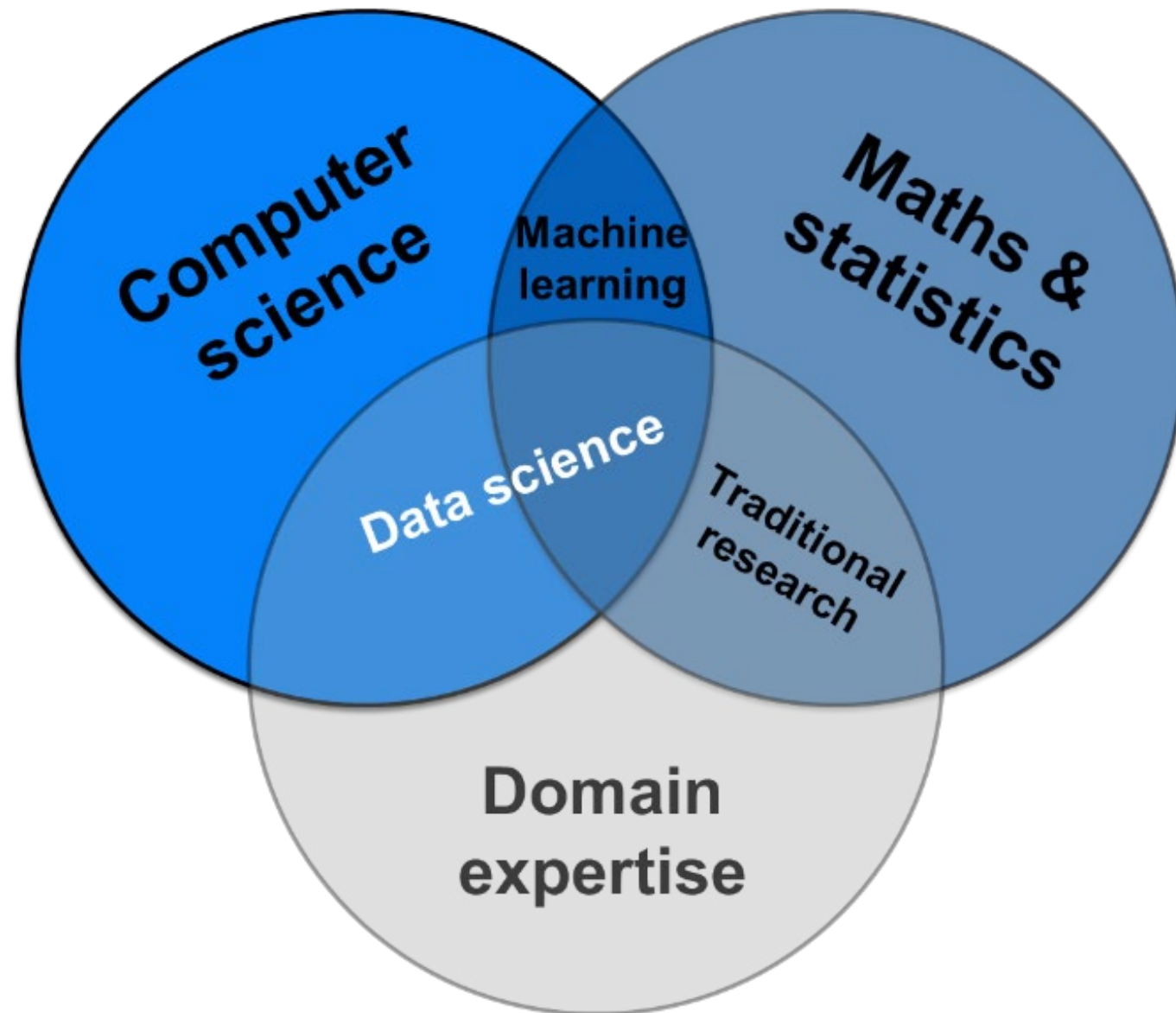Monday 12 December 2016

Monday 19 December 2016

Sunday 25 December 2016

Sunday 1 January 2016

Transport statistics, CO2 emissions and illegal fishing with AIS data

Data Science, New Skills and Partnerships

Data Science for Official Statistics

The mission of the ONS Data Science Campus is to work at the frontier of data science and Artificial Intelligence – building skills and applying tools, methods and practices – to create new understanding and improve decision-making for public good. **It defines data science as "*applying the tools, methods and practices of the digital and data age to create new understanding and improve decision-making*".**

How are Big Data projects using Data Science (AI and Machine Learning)

- Machine learning to identify agricultural crops from satellite images – need for ground truthing

- Machine learning to estimate extent of fresh water surfaces from satellite images – need for ground truthing

- AI to recognize patterns of shipping routes using AIS data for measuring transport and illegal fishing

- AI to recognize patterns of human mobility using mobile positioning data to estimate commuting, domestic tourism or internal migration

Why a Data Science facility with a national statistical office?

- to harness and exploit large digital datasets and data streams,

- to develop and test algorithms, which lead to statistics and insights,

- to develop new skills in the task force of the statistical office, as well as attracts partner communities to work with the statistical office.

Why a Data Science facility with a national statistical office?

- **Partnerships with** Private sector, academia, research institutes and civil society

- **Achieve the promise** of timely, more frequent and more granular data to inform and achieve the sustainable development goals and targets.

UN GWG and Partnerships

- Extending partnerships to private sector, academia, civil society, donor community:

- Positium, Flowminder, OneSoil, Sinergise, Azavea, Planet, Telenor, GSMA, Algorithmia, Esri, AWS, Google, Microsoft, Alibaba, Global Partnership for Sustainable Development Data, Sustainable Development Solutions Network

Program

Wednesday - morning

- UN Global Working Group and its Task Teams

- Introduction to the Big Data Centre

Wednesday - afternoon

- Use of Satellite data for official statistics

Thursday - morning

- UN Global Platform

- Use of Big Data for SDG indicator

Thursday - afternoon

- Use of Mobile Phone data for official statistics

Friday - morning

- Use of Big Data for Economic Statistics

- Use of Big Data for Official Statistics

Friday - afternoon

- Quality Assurance while using Big Data

- Round Table on way forward for using Big Data in Asia and the Pacific

Thank you