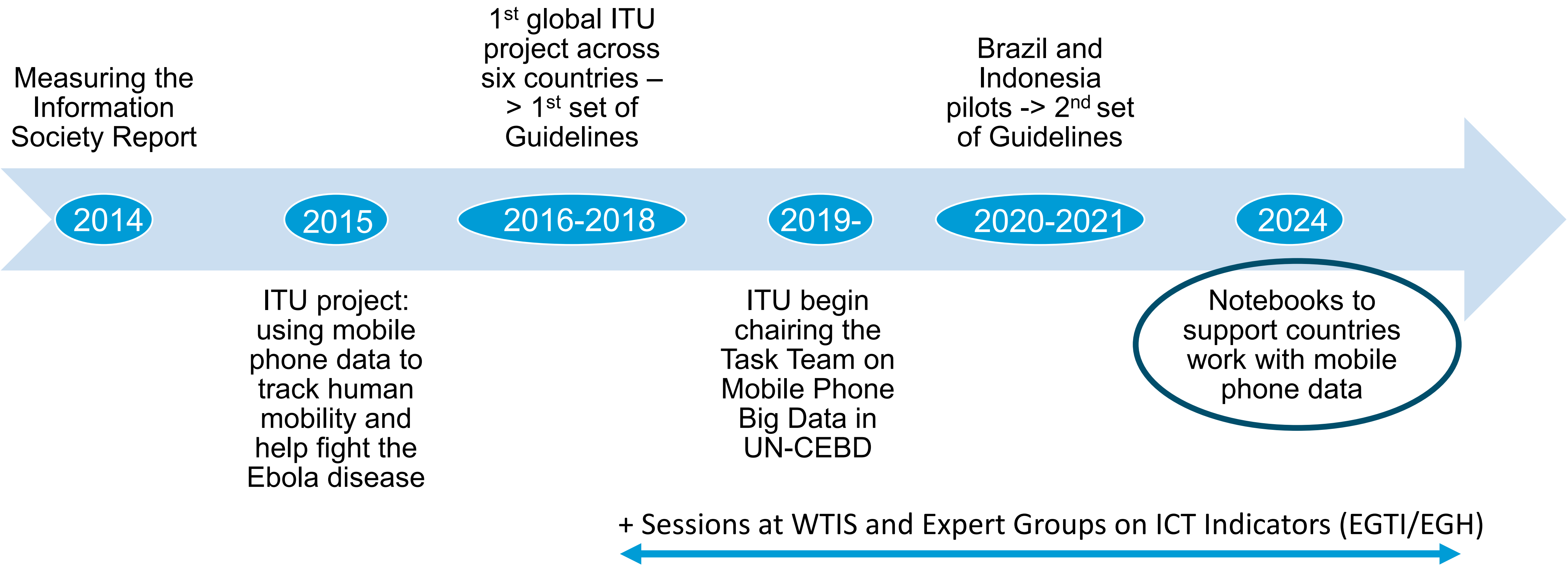
ITU notebooks for mobile phone data – from file to indicator

Fredrik Eriksson
Data Scientist, ITU

September 2024

10 years of using mobile phone data at ITU



2014

Measuring the Information Society Report

2015

ITU project: using mobile phone data to track human mobility and help fight the Ebola disease

2016-2018

1st global ITU project across six countries –> 1st set of Guidelines

2019-

ITU begin chairing the Task Team on Mobile Phone Big Data in UN-CEBD

2020-2021

Brazil and Indonesia pilots -> 2nd set of Guidelines

2024

Notebooks to support countries work with mobile phone data

+ Sessions at WTIS and Expert Groups on ICT Indicators (EGTI/EGH)

Purpose

Practical introduction to working with mobile phone data (MPD)

- Series of Jupyter notebooks “from file to indicator”
- Notebooks written in PySpark (Python API for Spark)
- The notebooks include code and documentation
- Goal is for countries to independently run the notebooks

Requirements:

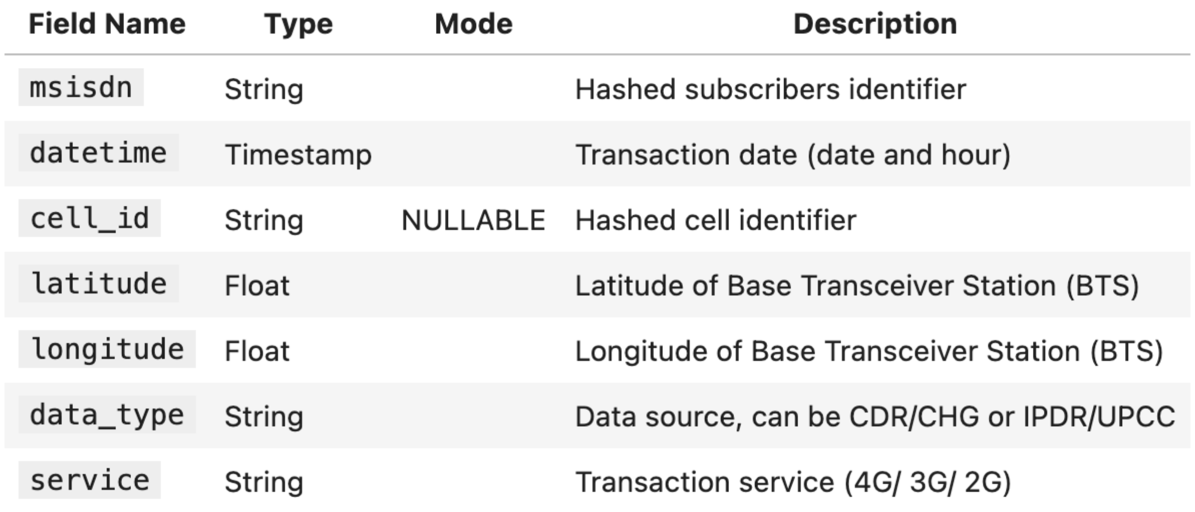
- Environment to run Jupyter Notebooks (cloud, on-premise)
- Python packages and dependencies installed (set-up code available)
- Reference data, *e.g.* geodata of local administrative units
- Staff / skills to run and adapt code to local circumstances
- Raw / synthetic MPD in CSV / Parquet (CDRs)

Call Detail Records (CDR)

- Passively generated when a subscriber:
  - Makes or receives a call
  - Sends or receives an SMS
  - Uses mobile data
- Routinely stored by MNOs for billing purposes

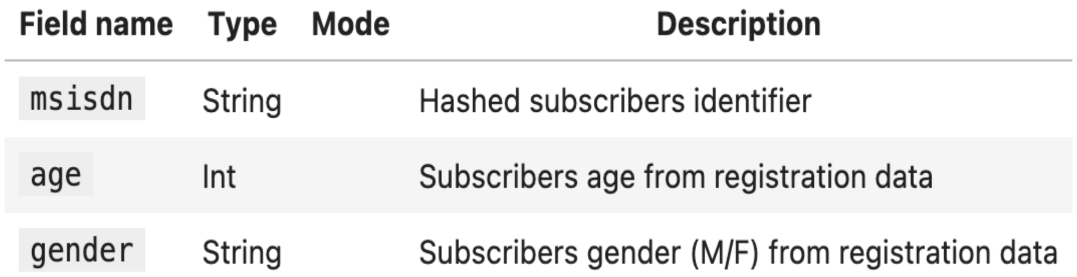
Run the Notebooks with just 7 key variables

- Mobile network operators (MNOs) collect Call Detail Records (CDRs) for billing purposes and store the connection of phones connect to nearby cell towers.



Field Name	Type	Mode	Description
msisdn	String		Hashed subscribers identifier
datetime	Timestamp		Transaction date (date and hour)
cell_id	String	NULLABLE	Hashed cell identifier
latitude	Float		Latitude of Base Transceiver Station (BTS)
longitude	Float		Longitude of Base Transceiver Station (BTS)
data_type	String		Data source, can be CDR/CHG or IPDR/UPCC
service	String		Transaction service (4G/ 3G/ 2G)

Minimum required fields.



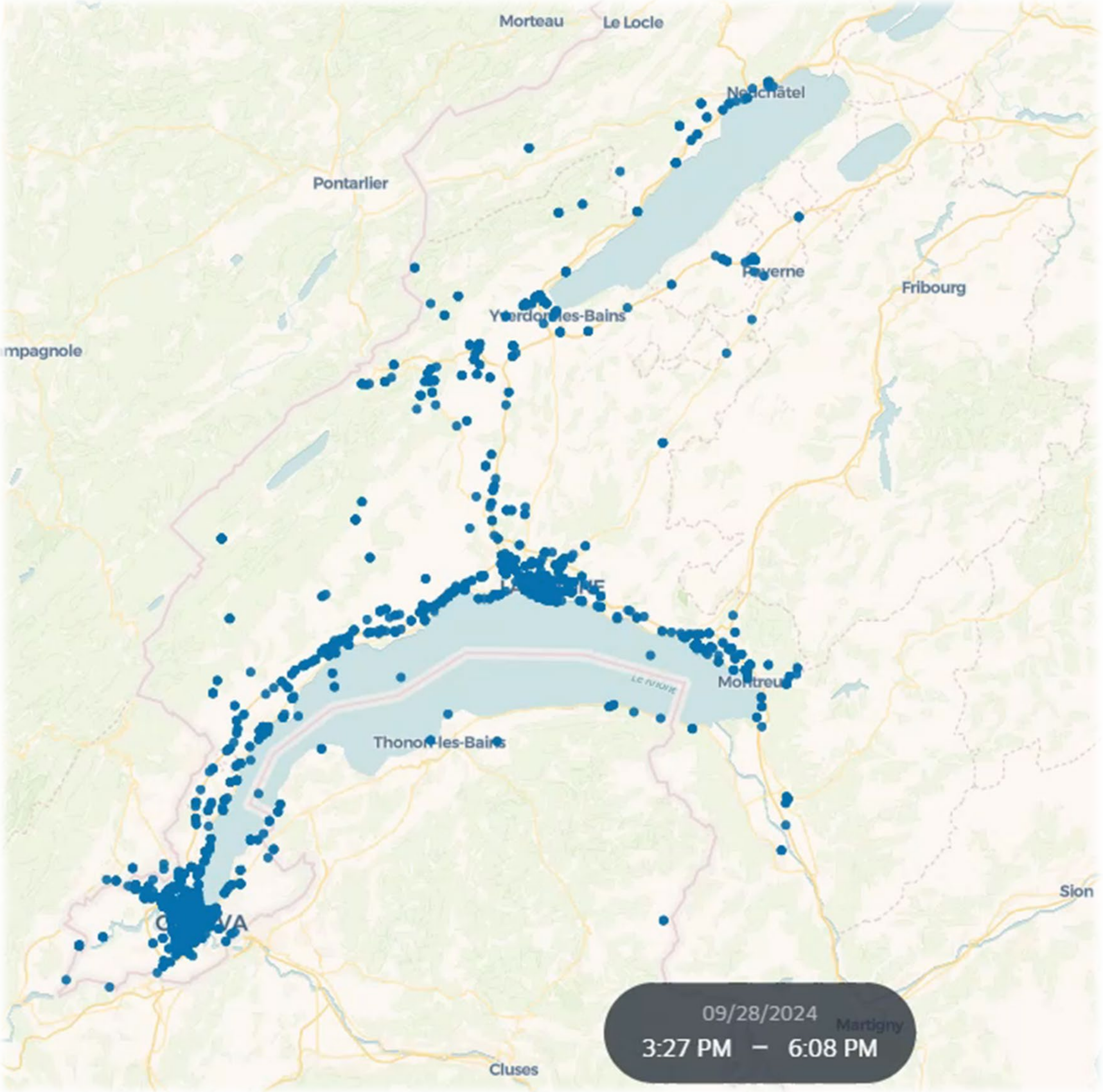
Field name	Type	Mode	Description
msisdn	String		Hashed subscribers identifier
age	Int		Subscribers age from registration data
gender	String		Subscribers gender (M/F) from registration data

Nice to have. Useful for analysis

No data yet? Use synthetic data

Artificial data that mimics the statistical patterns and properties of real-life data

- Two notebooks for creating synthetic CDRs:
  - 1st Notebook create a map of cell locations and population living near each cell
  - 2nd Notebook generate synthetic CDRs based on rules of typical human mobility patterns
- Parameters to customize rules and set probabilities for different events



Synthetic CDRs of 1000 subscribers during week of WTIS: (23 Sep – 29 Sep 2024)

Colors:
Dark red = Weekday, Light red = Weekday night,
Light blue = Weekend night, Dark blue = Weekend

Structure of the notebooks follows the guidelines

1. File structure and consistency check
2. Raw file processing and removal of duplicates and “robots”
3. Aggregated quality assurance checks
4. Definition “home” location
5. Calculation and visualization of final indicator
6. Validation checks

Proportion of individuals using the Internet from any location

(SDG Indicator 17.8.1)

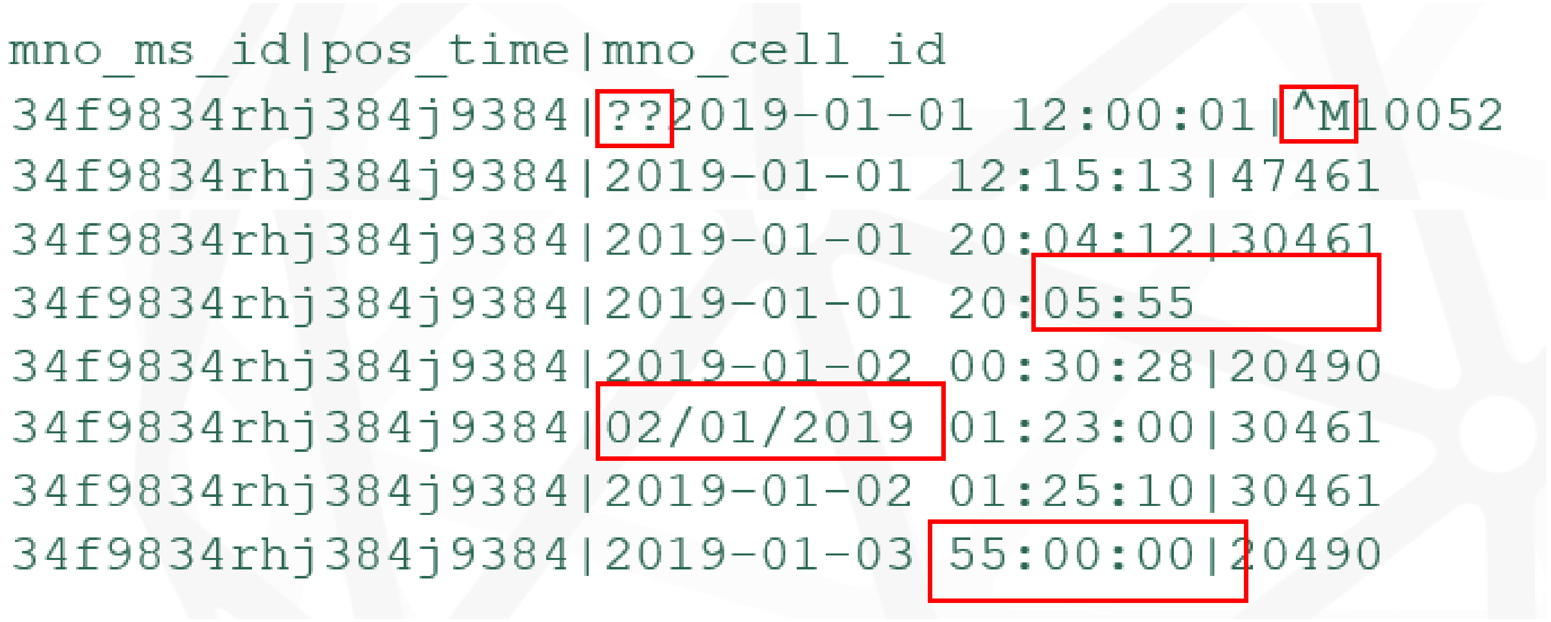
ITU Publications

International Telecommunication Union
Development Sector

Handbook on the use of mobile phone big data for measuring the information society

1. Pre-step: Check the data

- Check the consistency of the raw files
- Identify and fix elements that are possible sources of issues for processing



```
mno_ms_id|pos_time|mno_cell_id
34f9834rhj384j9384|??2019-01-01 12:00:01|^M10052
34f9834rhj384j9384|2019-01-01 12:15:13|47461
34f9834rhj384j9384|2019-01-01 20:04:12|30461
34f9834rhj384j9384|2019-01-01 20:05:55
34f9834rhj384j9384|2019-01-02 00:30:28|20490
34f9834rhj384j9384|02/01/2019 01:23:00|30461
34f9834rhj384j9384|2019-01-02 01:25:10|30461
34f9834rhj384j9384|2019-01-03 55:00:00|20490
```

- hidden characters

- missing delimiters

- inconsistency in data types

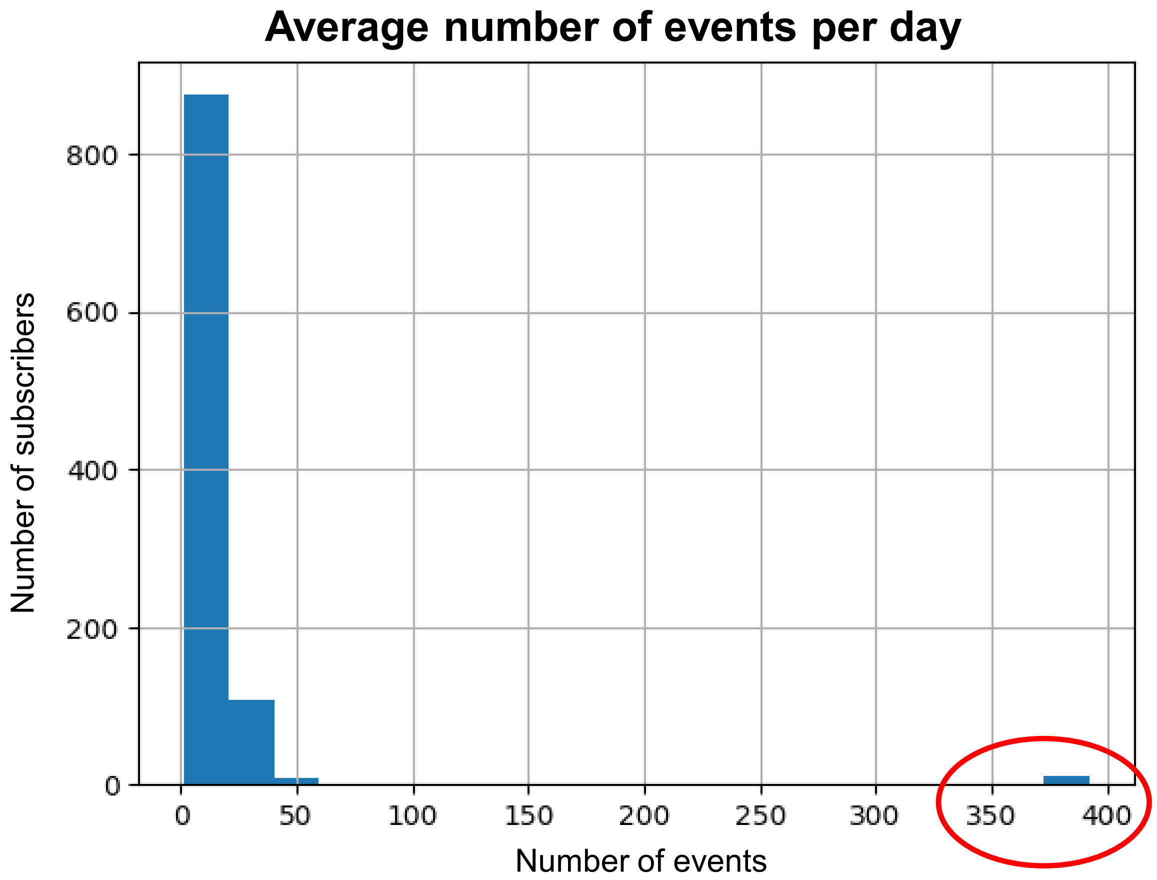
- impossible timestamps

2. Processing and cleaning CDR data

Check, clean and transform the raw MPD into a more usable format for analysis

Filtering:

1. **Remove duplicates**
2. **“Robot” filtering:** subscribers with more than *X* events per day (*e.g.* 300)
3. **Random location / “Tourist” filtering:** subscribers with fewer than *Y* events per day or Z events for the entire time period -> Not needed for this use case

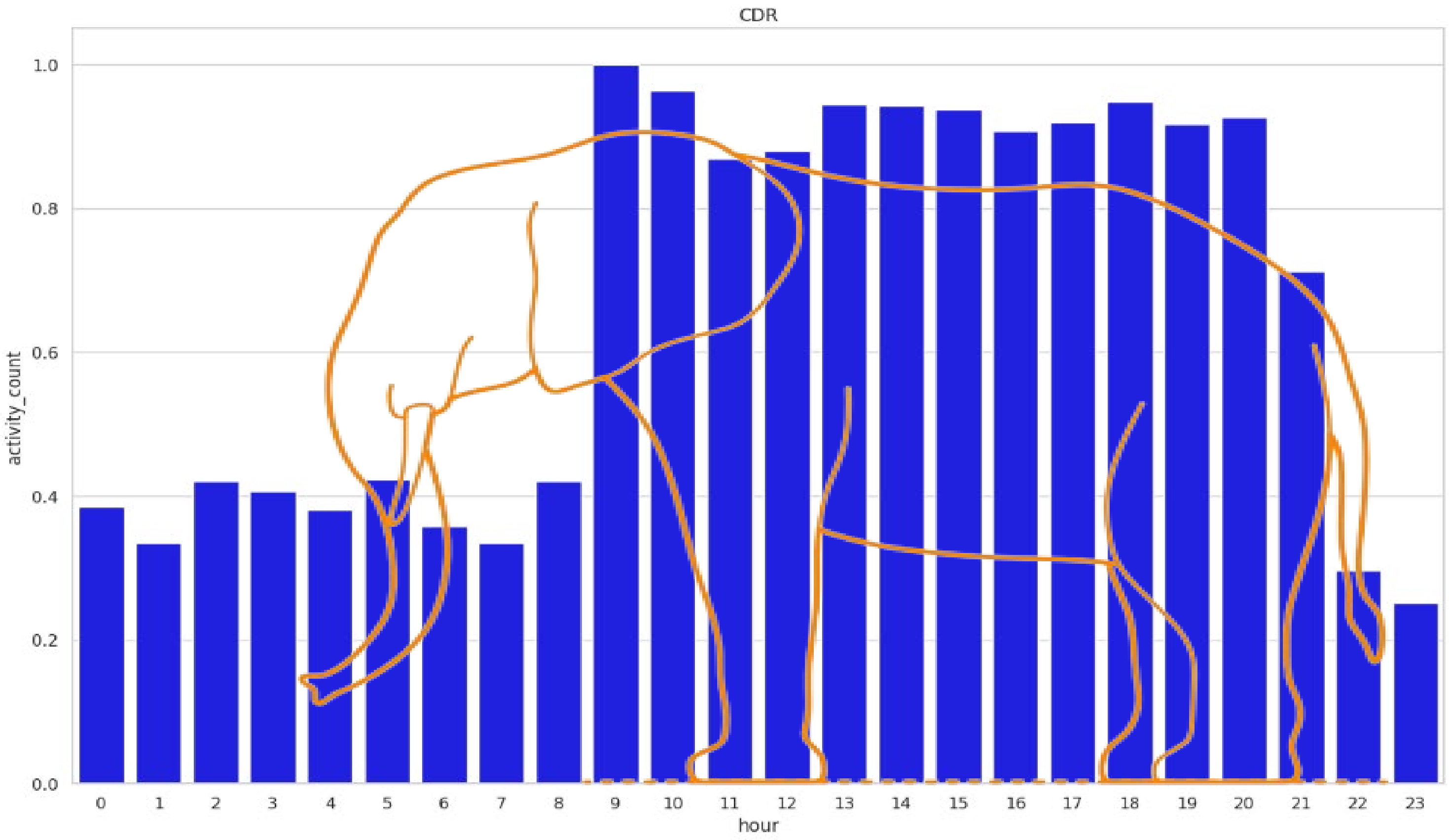


```
Number of records before robot filtering: 173481
[Stage 81:=============================================>
Number of records after robot filtering: 147751
```

3. Quality assurance

- Garbage in – Garbage out
- Identify aspects that may cause bias and affect coverage, frequency, quality and accuracy
- Output report provided in the notebooks

Diurnal distribution of subscribers' activity



4. Defining “home” cell location

- Determining a subscriber’s “home” is a crucial step for MPD uses cases -> needed to map the user to reference data
- Many methods to identify the “home” cell:
  - **The amount of activity** –> the cell with most events
  - **Consistent amount of activity** –> the cell with most activity on the highest number of distinct days
  - **Time constraints** –> the cell with most activity during nights, mornings and evening.
- ITU notebooks used a consistent time-constrained model:
  - The number events at each cell location from Mon to Thur.
  - Prioritization within three specific times (Night, Morning, Evening)

5. Indicator calculation

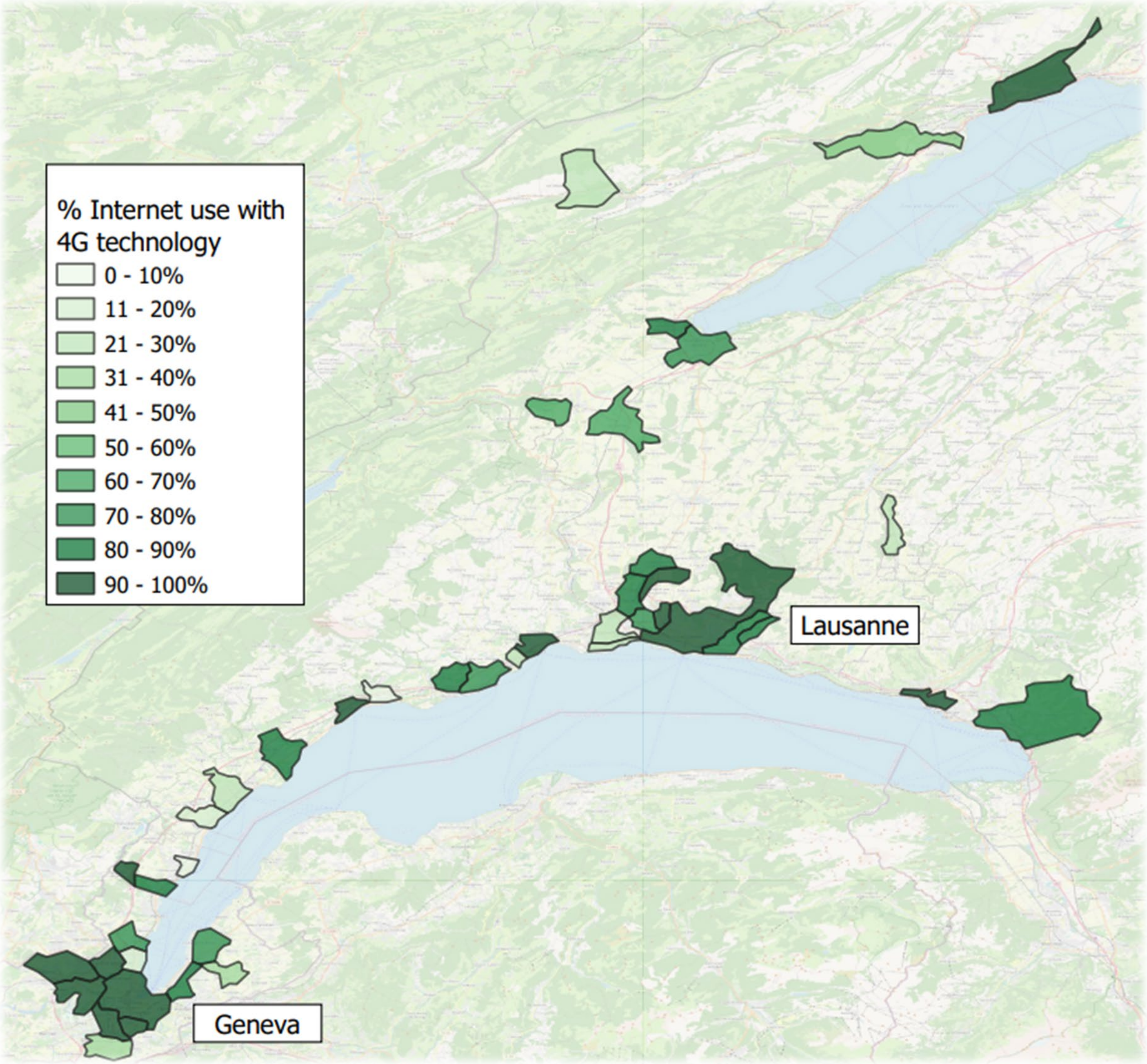
- A person with **any data event** is considered using the Internet
- The **proportion of individuals using the Internet** for an area is calculated as:

$$proportion\ of\ people\ using\ internet\ (lau2) = \frac{data\ users\ home\ count}{home\ count}$$

- Disaggregation by **technology**: most advanced used technology (2G/3G/4G)

Example: *proportion of people using the Internet with 4G technology =* $\frac{4G\_home\_count}{home\_count}$

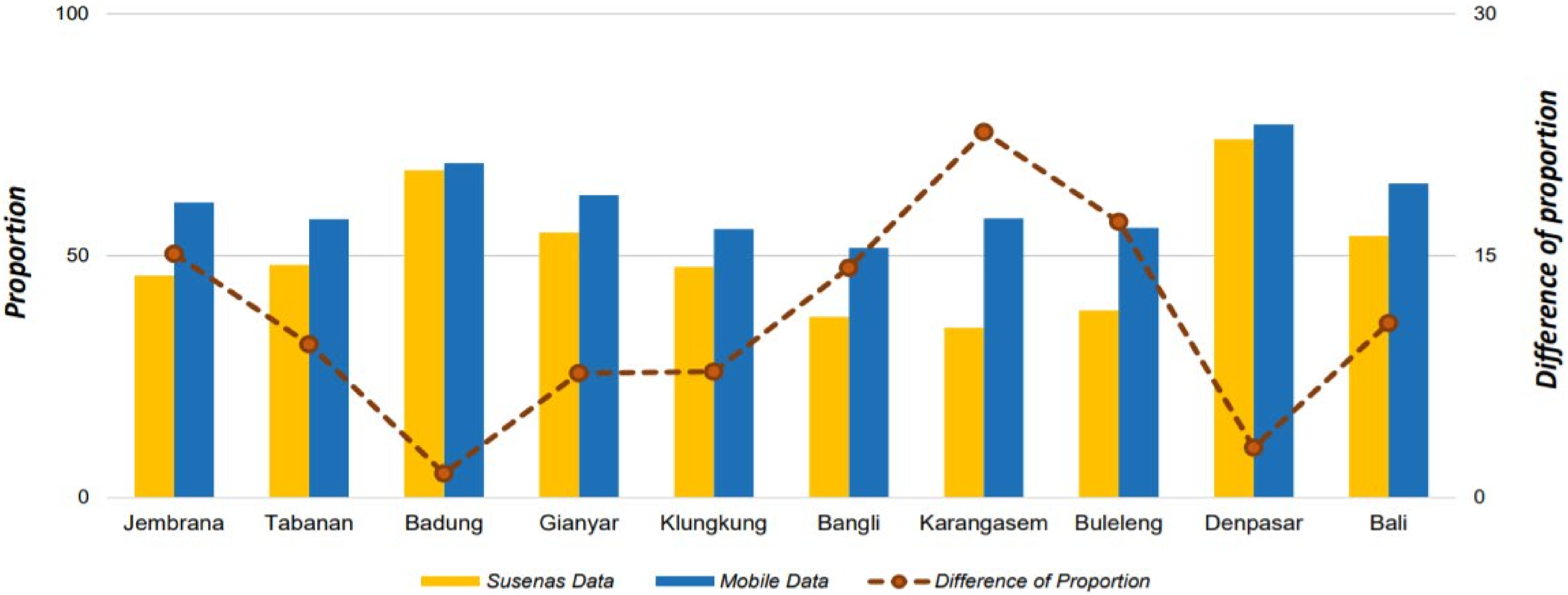
- Disaggregation by **age and gender** (if data is available)



Note: The colours in this map are for illustrative purposes only and do not reflect real Internet use through 4G technology in Swiss Romandie.

6. Validation against available survey data

Comparison of % internet users in Bali, Indonesia between mobile phone study and survey data



Source: BPS Indonesia. The proportion differences of internet users (17.8.1) in Bali, Indonesia, between mobile phone data and survey data from the National Socio-Economic Household Survey (Susenas)

Next steps:

- Share and test notebooks with countries using real data
- Finalize and upload notebooks to ITU GitHub repository
- Continue development of synthetic data
- Guidelines for post bias-adjustment – MPD is not random sample
- Conduct workshops and trainings (with the regional hubs and partners)
- Work with partners in integrating MPD as one of the data sources in countries

We are here to work with you if you want to learn more

Thank you!