Before working with MPD
Methodological decisions to make

Siim Esko
Positium
Estonia

@positium
www.positium.com
Assume you just got access to mobile phone data. What are the first questions you have in your mind before you put a team to work on the data?
- Big Data in GSBPM -

What can NSO do themselves
GSBPM new areas for NSO in terms of big data

- Conceptually – NSO; Technically - external

Specify Needs: Similar – NSO has expertise
Design: Similar – NSO has expertise
Build: Very different – Expertise is external
Collect: Very different – MNO expertise
Process: Very different – Expertise is external
Analyse: Very different – Expertise is external
Disseminate: Relatively similar – NSO has expertise
Evaluate: Similar – NSO has expertise

Relative similarity: Similar – NSO has expertise
Different – Expertise lies in external parties
Very different – Expertise is external
Relatively similar – NSO has expertise
Similar – NSO has expertise
GSBPM for cross-border tourism processing

2019

- Specify Needs
- Design
- Build
- Collect
- Process
- Analyse
- Disseminate
- Evaluate

- BPS – Statistics Indonesia
- Positium – Collection and processing design
- Positium - Build collection and processing
- Telkomsel - Collect
- Telkomsel - Process
- BPS

- BPS with Positium – QA and metadata tools
- BPS – outputs and variables
- Simple or complex are options -

Start simple or invest in advance
Basic options for methodology

Simple
- Aggregating the data “as is” or with simple filters
- Remove coverage issues with calibration

Complex
- Work on raw data
- Remove coverage issues with algorithms
Case of Indonesia – Cross-border tourism

BATAM island
MPD is 3-20x above immigration data
• Fast movers = those who cover the distance between two BTS that is only possible on a plane
• % of fast movers is very high in Batam and Bintan
• Do not enter the economic territory of Indonesia
• The bias exists all over Indonesia
• Can be countered by excluding fast movers

*Flight path from Changi airport correlates with MPD anomalies*
Inbound data extraction → Antennae location extraction → Country of origin extraction → Removal of non-representative data

Hashing IMSI/MSISDN → Data formatting → Data forwarding → Data cleaning and validation

Merging data from different MNOs → Calculation of theoretical coverage areas of antennae → Spatial interpolation of location events → Identifying country of residence

Identifying trips → Identifying entry gate → Detecting seamen → Detecting airplane passengers

Eliminating accidental border roaming → Identifying locals with foreign SIM card → Identifying foreigners with local SIM card → Identifying transit trips

Classifying trips → Aggregating the results → Extrapolating results to general population → Data updates

Output validation → Disclosure control → Result publication
Cascading of MPD data across error classes, one year

- All trips in MPD
- Tourism trips in MPD

TOTAL for 23 kabupatens

- MPD tourism: 2985615
- COR: 150379
- UE: 42387
- Fishermen: 1549
- Seamen Transit: 364142
- Sea Border Accidental Roaming: 662252
- Land Border Accidental Roaming: 642981
- Trip <80s: 2626060
- Core processes -

There are some core processes that repeat and should be uniform across different uses of the data.
Processing Data for Different Domains

Raw data

Domain specific process 1
  Domain specific process 2
  Domain specific process n

Resulting statistical indicators in domain X
Resulting statistical indicators in domain Y
Resulting statistical indicators in domain Z

Core data process 1
Core data process 2
Core data process n

Core data model

Domain specific process 1
Domain specific process 2
Domain specific process n

Resulting statistical indicators in domain X
Resulting statistical indicators in domain Y
Resulting statistical indicators in domain Z
Examples of core processes

1. Input data QA
2. Cleaning of noise
3. Trip generation
4. Home detection and usual environment

These steps are completed in a unified way for different domains
Core Data Model

MNO → Pre-Processing → Core Data Model

- Tourism
- Population
- Migration
- Transportation
- ...

Tourism statistics
Population statistics
Migration statistics
Transp. statistics
...

11/06/2019 Mobile Phone Data Training
- Quality Assurance never stops -

QA is a consistent part of every process step
• UNECE suggested framework for the quality of big data
  • Covers the 3 phases of statistical production:
    • Input – data is acquired or in the process of being acquired
    • Throughput – data is transformed, analysed and manipulated
    • Output – the resulting statistics
• UNECE suggested framework for the quality of big data
  • 3 hyperdimensions (objects which quality is assessed):
    • Source (type of data, characteristics of entity from which data is obtained, governance)
    • Metadata
    • Data
# Quality Assurance Framework

## Processing steps

<table>
<thead>
<tr>
<th>Input</th>
<th>Throughput</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
<td>Privacy and security</td>
<td>Confidentiality</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Log files</td>
<td>System independence</td>
</tr>
<tr>
<td></td>
<td>Metadata</td>
<td>Quality gates</td>
</tr>
<tr>
<td></td>
<td>Consistency</td>
<td>Steady states</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>Consistency</td>
<td>Coherence</td>
</tr>
<tr>
<td></td>
<td>Validity</td>
<td>Consistency</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>Validity</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Quality Gate 1 – Raw Data

Extraction and pre-processing

Extraction

Initial raw data

Raw data steady state (cleaned raw data)

Pre-processing

Quality Gate Report

Statistical processing

Request for Corrections

Improving Methodology & Algorithms

Methodology Algorithms

Processed Data (Data Model)

Quality Gate Report

Aggregated Data

Output

Indicators

Quality Gate Report
Common errors in raw data

- Wrong antenna coordinates or attributes
- Errors in antenna coordinates transformation
- Data gaps
- Missing data from some sub part of the system
- Time zone issues
- Incorrect format of timestamps
- Changes in continuity of the ID-s
- Duplicated records
- ...
Common errors in processing

• Process produces an error
• Process does not finish
• Process ingests erroneous data
• Process overwrites critical data
• ...
Quality Gate 3 – Output Data

Extraction and pre-processing:
- Extraction
- Initial raw data
- Raw data steady state (cleaned raw data)
- Pre-processing
- Quality Gate Report

Statistical processing:
- Improving Methodology & Algorithms
- Methodology Algorithms
- Process data
- Quality Gate Report

Output:
- Aggregated Data
- Indicators
- Quality Gate Report
Common errors in output data

If all processes run against correct methods and run correctly, output data should be sound. However,

• Low coherence to validation data

• Anomalies in the data
  • Peaks
  • Valleys
  • Gaps

• Trends that indicate a systematic change in underlying data

• New phenomena