MAKSWELL

MAKing Sustainable development and WELL-being frameworks work for policy analysis

www.makswell.eu

Insights on statistical methodologies and new data sources for SDG and well-being indicators

Work Package 2

Methodological aspects of measuring SDG indicators with traditional and non-traditional data sources

Jan van den Brakel

Third Makswell workshop, 05-03-2020, Barcelona, Spain

MAKSWELL

Partners WP 2:

1. Statistics Netherlands
2. Southampton University
3. Istat
4. Pisa University
5. Trier University
6. Destatis

Outline

1. Purpose WP 2
2. Del. 2.1
3. Del. 2.2
4. Needs for further research: Del. 2.3
5. Discussion

MAKSWELL

1. Purpose WP2

- Overview of traditional and non-traditional data sources for SDG indicators
- Methodological development using new data sources (big data)
  - Integration of traditional and non-traditional data sources (small area estimation, nowcasting)
  - Pointing out quality aspects of big data (representativity, timeliness, …)
- Review of good and bad practices
- Identify needs for future research

MAKSWELL

2. Del. 2.1: Aspects of existing databases, traditional and non-traditional data sources and collection of good practices

- Overview of traditional and non-traditional data sources for SDG indicators, through an inventory in Italy, the Netherlands and Germany
- Review of good and bad practices:
  - Google trends to measure propensity to move (CBDS)
  - Improving the Italian CPI using scanner data
  - Mobile phone network data (day time population, mobility, tourism, migration flows, poverty, economic growth)
  - Webscraping from websites (estimating number of innovative companies, prices)
  - Social media studies (classifying messages to measure social tension and sentiment, including a nowcast excersise with the consumer confidence index)
  - Found data: estimating unmetered photovoltaic power using time series of electricity taken from the high voltage grid and time series on solar irradiance, day length, temperature, …

MAKSWELL

2. Del. 2.1: Aspects of existing databases, traditional and non-traditional data sources and collection of good practices

- Review of good and bad practices (cont.):
  - Found data: estimating unmetered photovoltaic power using time series of electricity taken from the high voltage grid and time series on solar irradiance, day length, temperature, …
  - Remote sensing data
    - Copernicus project on land use in Germany
    - Measuring urban sprawl using satellite data (CBDS)
    - Detecting photovoltaic solar panels in aerial images

MAKSWELL

3. Del. 2.2: Methodological aspects of measuring SDG indicators with traditional and nontraditional data sources

- Combining survey data with non-traditional data sources:
  - Small area estimation
  - Time series methods for now-casting
- Non-traditional data sources as a primary data source
- Evaluating sustainability through an input-state-output framework as an alternative for juxtapose a large set of indicators

MAKSWELL

4. Needs for further research (Del. 2.3)

1. Quality frame work
    - Probability samples:
        - Sampling error through variance estimation
        - Non-sampling errors: Total Survey Error (TSE) frame work
    - Non-probability data, non-traditional data, big data etc:
        - No established quality frame work
        - Extension of the TSE frame work to these new data sources: Total Error Frame-work (TEF)
        - Account for the specific features of new data sources, e.g.
            - Identification error of units of interest in fuzzy big data
            - Linkage errors
            - Feature extraction errors
            - Quantifying uncertainty under machine learning and AI alg.

MAKSWELL

4. Needs for further research (Del. 2.3)

2. Additional methodology
    - Use of non-traditional data sources in official statistics:
        - Primary data source
        - Covariates in model-based inference procedures (SAE, time series models, now-casting)
    - Primary data source:
        - Correction for selection bias
            - Methods require structured data and strong auxiliary information
            - Need for new methods that handle these issues
            - Need for methods to obtain insight in the performance of existing methods if auxiliary information is obscured with errors (linkage errors, feature extraction errors, identification errors, …)

MAKSWELL

4. Needs for further research (Del. 2.3)

2. Additional methodology (cont.)
    - Covariates in model-based inference procedures:
        - More empirical evidence to illustrate:
            - Efficiency of SAE methods compared to machine learning algorithms
            - Usefulness in the production of official statistics
        - Time series models:
            - Dynamic factor state-space model for high dimensionality issues
            - How to select relevant series without falling into the trap of data dredging?
            - How to account for time-varying state correlations?

MAKSWELL

4. Needs for further research (Del. 2.3)

3. Risk appetite of NSI's
    - Well-accepted approach: probability sampling in combination with design-based inference
        - Robust for model-misspecification
        - NSI is in control over
            - Minimum required accuracy of the outcomes
            - Availability of data,
    - Use of new data sources:
        - Model-based inference to correct for selectivity or as covariates in SAE and nowcasting procedures
        - More empirical research to illustrate the benefits of such methods in the context of official statistics

MAKSWELL

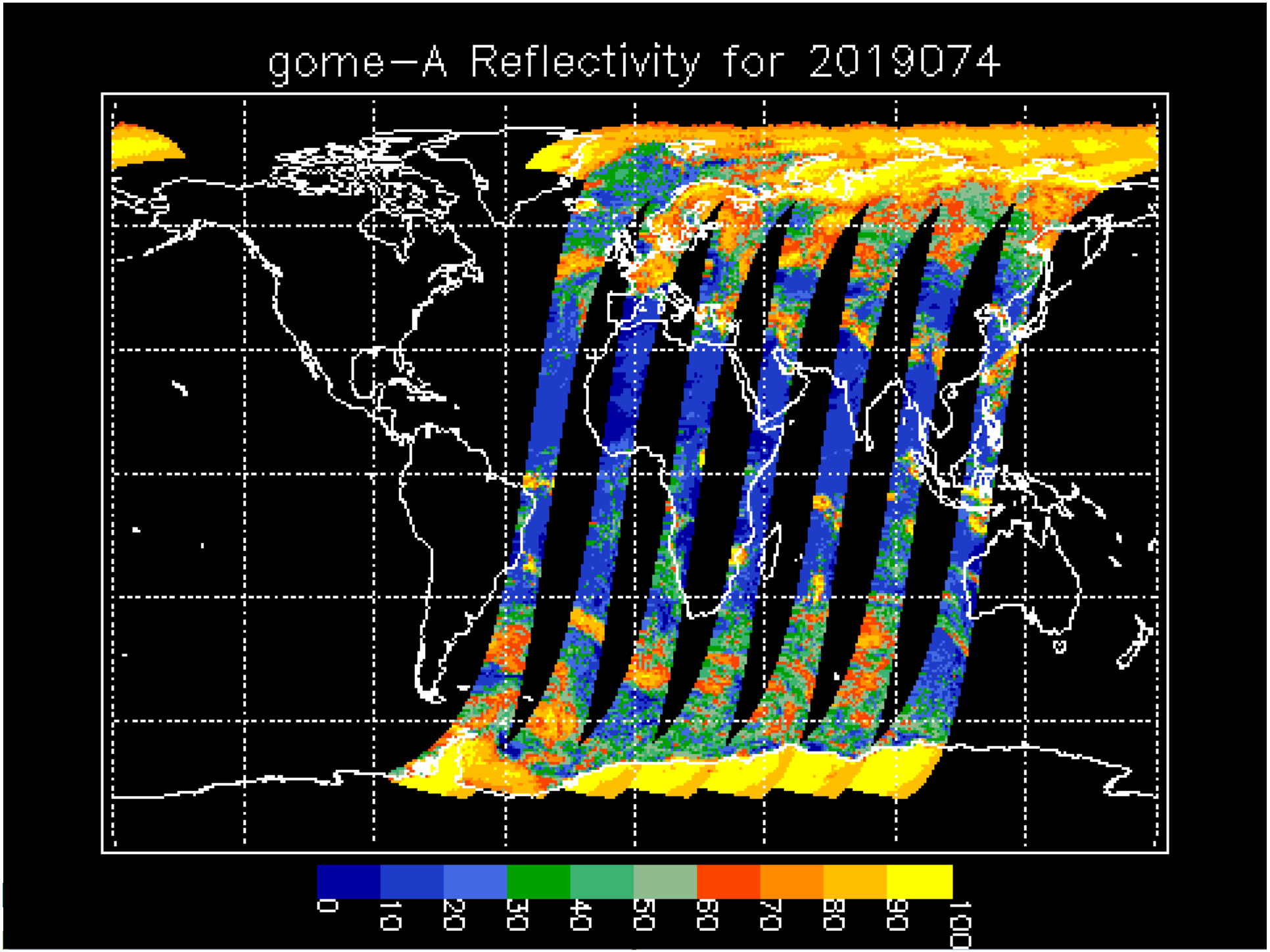
4. Needs for further research (Del. 2.3)

4. Remote sensing data
    - Successful in absence of high quality official data on well-being, poverty and SDGs
    - But also for European situations? Requires high quality remote sensed data!
    - Integration of remotely sensed data in the frame-work of SDGs and well-being measurement is predominantly experimental
    - Need for a comprehensive overview of datasets and methods to facilitate their use in official statistics

MAKSWELL

4. Needs for further research (Del. 2.3)

4. Remote sensing data (cont.)
    - Statistical purpose is predictions for well-being and SDGs:
        - spatial comparisons
        - temporal comparisons
    - Requires:
        - combining multiple images
        - comparing images of the same area over time
    - Issues:
        - spatial image inconsistency
        - temporal image inconsistency
        - computational power and data storage capacity


gome-A Reflectivity for 20190074
0
10
20
30
40
50
60
70
80
90
100

MAKSWELL

4. Needs for further research (Del. 2.3)

4. Remote sensing data (cont.)
    - Research:
        - methods that reduce spatial and temporal image inconsistency
        - better understanding how these error sources influence/increase uncertainty of predictions for poverty, well-being and SDGs
        - methods that account for or reduce the influence of these error sources
        - methods that quantify the increase of uncertainty on predictions for well-being, poverty and SDGs due to these error sources
        - research into methods to validate the reliability of constructs for poverty and well-being derived from remote sensed data

MAKSWELL

4. Needs for further research (Del. 2.3)

4. Remote sensing data (cont.)
    - Data and hardware issues
        - processing images for large areas and longer periods require considerable computational power and data storage capacity
        - availability of the required AI hardware and AI knowledge is not standard available at national statistical institutes
        - satellite images of sufficient quality or resolution are only commercially available
        - to evaluate methods, geo-coded unit level data on income and poverty are required
        - to facilitate wider development of satellite-based applications an Europe wide computing infrastructure should be established

MAKSWELL

4. Needs for further research (Del. 2.3)

5. Deep learning
    - Important for analyzing satellite and aerial images
    - Example from Del 2.1: counting solar panels and land use
    - Further research:
        - better understanding error sources to evaluate uncertainty of estimation results
        - how to create training sets, test sets and validation sets to maximize model generalizability to intended target populations
        - minimize the effect of errors in the annotating process
        - the impact of class imbalance on uncertainty measures
        - methods to quantify how these errors contribute to estimation uncertainty
        - model interpretability and transparency
        - data science and hardware knowledge for computationally intensive methods

MAKSWELL

5. Discussion

Further needs for research

- Quality frame work for non-traditional data sources
- Needs of a new/extended framework to measure wellbeing and SDG indicators in terms of data sources, methodology, and quality requirements
- Extension of methodology to correct for selection bias
- Applications of SAE that uses non traditional data sources as covariates
- Time series methods
- Risk appetite of NSI's: empirical applications in the context of official statistics
- Methods for processing remote sensing data
- Deep learning

- Deliverable 2.3: Planned deadline February 2020