MAKSWELL

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

www.makswell.eu

Work Package 2

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

Jan van den Brakel

15-03-2019

MAKSWELL

Partners:

1. Statistics Netherlands (J. van den Brakel, M. Puts, R. Willems, …)
2. Southampton University (P. Smith, N. Tzavidis)
3. Istat (F. Bacchinin, A. Ferruza, L. Di Consiglio, …)
4. Pisa University (M. Pratesi, C. Giusti)
5. Trier University (R. Münnich, F. Ertz)
6. Destatis (N. Rosinski, T. Zimmermann, K. Wichmann)

MAKSWELL

Introduction

1. Purpose
2. Inventory traditional and non-traditional data sources
3. Examples from CBDS and Istat
4. Combining survey data with non-traditional data sources
5. Non-traditional data sources a primary data source
6. 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 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

2. Inventory traditional and non-traditional data sources

- Countries: Italy, the Netherlands, Germany
- Overview for all indicators
  - Published
  - Status (official / developed)
  - Data source
  - Frequency
  - Regional detail (NUTS level)
  - Alternative (no-traditional) data sources

2. Inventory traditional and non-traditional data sources

Results Italy

Goal	Total ind.	Published	Anually	Lower fre	NUTS 0	NUTS 1 or2	Survey	Register	Other
1	7	5	5	0	2	3	4	1	
2	9	4	3	1	3	1	4	0	
3	21	13	11	2	4	9	2	19	
4	8	6	4	2	1	5	4	1	1
5	10	6	2	4	3	3	3	3	
6	9	5	1	4	0	5	4	1	
7	4	4	4	0	1	3	2	0	2
8	17	12	8	4	5	7	3	9	
9	9	6	6	0	2	4	2	3	1
10	8	4	4	0	0	4	1	3	
11	11	9	8	1	3	6	3	3	3
12	10	4	4	0	2	2	0	4	
13	6	1	1	0	1	0	0	1	
14	7	2	2	0	1	1	0	1	1
15	11	6	3	3		6	3	1	2
16	21	8	3	5	3	5	6	2	
17	25	4	4	0	3	1	2	2	

MAKSWELL

3. Examples

- 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

3. Examples

- Remote sensing data
  - Copernicus project on land use in Germany
    - Satellite data for statistical information on land use
    - Classify vegetation indices with random forest
    - Successful for main categories of land use
  - Measuring urban sprawl using satellite data (CBDS)
    - Data: MODIS Terra satellite: NDVI (250m—500m)
    - Machine learning methods to classify land use
      - Support Vector Machines
      - Random forest
      - K-nearest neighbors
      - Convolutional deep neural network

MAKSWELL

3. Examples

- Remote sensing data (cont.)
  - Detecting photovoltaic Solar panels in aerial images
    - Aerial images resolution 25 cm
    - Classification methods:
      - VGG16 convolutional neural network
      - Random forest

MAKSWELL

4. Combining survey data with non-traditional data sources

- Data sources for SDG indicators: survey data, register data, non-traditional data
- Non-traditional data as covariates in model-based inference procedures
  - Small area estimation
  - Time series models

MAKSWELL

4. Combining survey data with non-traditional data sources

Small area estimation

- Survey data modeled using
  - Area level model (Fay-Herriot)
  - Unit level model (Battese-Harter-Fuller)
- Traditional covariates: census data
  - Problem: low frequency, not available in developing countries
  - Non-traditional data sources often on higher frequencies
- Literature on predicting SDG related indicators on poverty, literacy, income, unemployment (survey data) with mobile phone data, satellite images, traffic intensity, web-scraping of online prices, etc

MAKSWELL

4. Combining survey data with non-traditional data sources

Time series models

- NSI's: repeated surveys
- Time series models: use temporal and cross-sectional correlations
  - Small area estimation
  - Nowcasting
- Structural time series models versus Box-Jenkins ARIMA / VARIMA
- Auxiliary series in structural time series models:
  - Regression component
  - Multivariate structural time series models

MAKSWELL

4. Combining survey data with non-traditional data sources

Time series models cont.

- SDG related examples:
  - Monthly unemployment using claimant counts in the UK and the Netherlands
  - Consumer confidence and sentiment from social media messages
- Advantages:
  - Improves precision (SAE)
  - Non-traditional data sources at higher frequency: nowcasting

MAKSWELL

4. Combining survey data with non-traditional data sources

Time series models cont.

- Issue: dimensionality problem
- Google trends, e.g. unemployment and search behavior on internet
- Dynamic factor models:
  - Central Banks to nowcast quarterly GDP with monthly indicators
  - Extract p common factors from n auxiliary series using PCA (p << n)
  - Multivariate structural time series model for survey series and auxiliary series
  - Common factors: dynamic trend model correlated with trend of the survey series

MAKSWELL

5. Non-traditional data sources a primary data source

- Examples for SDG indicators:
  - Satellite and aerial images to measure forest decline, urbanization, air quality
  - Sensor data for traffic intensity, air quality
  - Data measure the phenomena of interest
- General:
  - Data generating process is unknown
  - Difficult to generalize results to an intended target population
  - Selection bias (similar to non-response in survey data)

MAKSWELL

5. Non-traditional data sources a primary data source

- Methods to correct for selection bias:
  - Calibration and weighting
  - Quasi randomization or pseudo-design based methods
  - Sample matching
  - Model-based approaches
    - Super population methods
    - Informative sampling

MAKSWELL

6. Discussion

- Deliverable 2.1 and 2.2 (April 2019)
- Overview of traditional and non-traditional data sources for SDG indicators
- Examples of using non-traditional data sources for SDG indicators and official statistics (with or without survey data)
- Review of methodology for non-traditional data sources
  - Covariates in model-based inference
    - SAE -> WP3
    - Time series methods for now casting -> WP4
  - Primary source
- Deliverable 2.3: identify needs for future research (February 2020)