Advances in data integration and Small Area Estimation for equitable and sustainable development

Jacques Silber

Bar-Ilan University, Israel, and

Honorary Fellow, Centro Camilo Dagum, Tuscan Interuniversity Centre, Advanced Statistics for Equitable and Sustainable Development, Italy.

Prepared for session EO675 on

Advances in data integration and SAE for equitable and sustainable development

11th International Conference of the ERCIM WG on

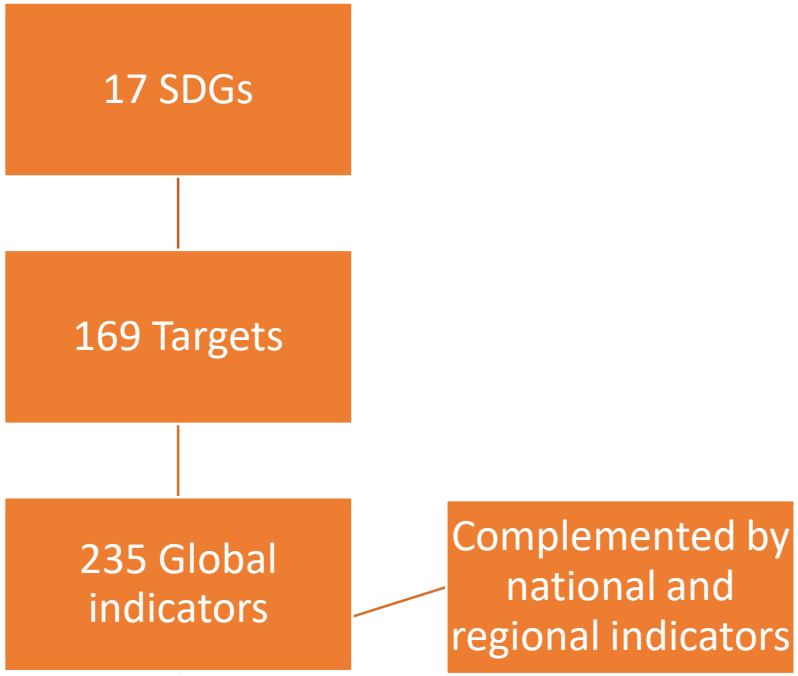
Computational and Methodological Statistics (CMStatistics 2018)

December 14-16 2018, University of Pisa, Italy.

Outline

1) On Sustainable Development Goals (SDGs)
2) Which indicators are available?
3) The need for disaggregated data
4) The need to integrate various data sources
5) Integrating administrative and survey data
6) Using Big Data to derive SDG indicators
7) Integrating census and survey data: the case of poverty mapping
8) More on Small Area Estimation (SAE) and poverty mapping
9) More on SAE and big data
10) An interesting case of SAE: the Community Based Monitoring System (CBMS) in the Philippines
11) Some concluding comments

1) The Sustainable Development Goals (SDGs)



The 17 goals


1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE AND JUSTICE STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

- SDG’s were approved by the United Nations General Assembly in July 2017.
- This list of indicators is supposed to be complemented by other national and regional indicators
- For each indicator there is one or several institutions/agencies in charge of collecting and reporting the data
- Emphasis has also been put on data disaggregation by income, gender, age, race, ethnicity, migratory status, disability, geographic location and other characteristics relevant in national contexts
- It is however clear that the long list of indicators does not cover all aspects of the goals and targets
- Moreover the data for many indicators are still unavailable, even at national level.

2) Many indicators are still unavailable, even at national level.

In fact the indicators are classified into three tiers:

- Tier I: the methodology to compute the indicators exists and the data are available

- Tier II: the methodology to compute them exists but there are few areas/regions for which the data are really available

- Tier III: no methodology has been agreed upon and information is still scarce

Illustrations

Tier 1:

- Goal 5: Achieve gender equality and empower all women and girls
- Target 5.5: Ensure women's full and effective participation and equal opportunities for leadership at all levels of decision-making in political, economic and public life
- Indicator 5.5.2.: Proportion of women in managerial positions

Tier 2:

- Goal 8: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
- Target 8.5: By 2030, achieve full and productive employment and decent work for all women and men, including for young people and persons with disabilities, and equal pay for work of equal value.
- Indicator 8.5.1: Average hourly earnings of female and male employees, by occupation, age and persons with disabilities

Tier 3:

- Goal 8: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all.
- Target 8.b: By 2020, develop and operationalize a global strategy for youth employment and implement the Global Jobs Pact of the International Labour Organization.
- Indicator 8.b.1: Existence of a developed and operationalized national strategy for youth employment, as a distinct strategy or as part of a national employment strategy.

3) The need to disaggregate the data

As stressed by Martinez (2017), since one of the basic principles of the SDG's is to "LEAVE NO ONE BEHIND", there is a need to disaggregate the data by

- income class
- gender
- ethnicity
- geographic location
- migration status
- disability status

and other criteria…

The need for disaggregation appears in many of the SDG's targets:

Target 2.3 (Goal 2: Zero Hunger): by 2030 double the agricultural productivity and the incomes of small-scale food producers, particularly **women, indigenous peoples, family farmers, pastoralists and fishers**, including through secure and equal access to land, other productive resources and inputs, knowledge, financial services, markets, and opportunities for value addition and non-farm employment .

Target 5.4 (Goal 5: Gender equality): by 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including **persons with disabilities, indigenous peoples, and children in vulnerable situations.**

Target 8.8 (Goal 8: Decent Work and Economic Growth): protect labour rights and promote safe and secure working environments of all workers, including **migrant workers, particularly women migrants**, and those in precarious employment

But why such a disaggregation? (see, U.N. Statistics Division, 2017, and Truszczynski, 2017)

- Due to age, socio-economic status, gender, ethnicity and geography, vulnerable groups are often excluded from access to good education, health care, electricity, safe water and other critical services.
- 80% of the world's poor live in rural areas
- In 2015 85% of urban population has access to safe drinking water but only 55% in rural areas
- The lack of disaggregated data for many vulnerable groups (children, persons with disabilities, people living with HIV, older persons, indigenous peoples, migrants, refugees…) hides the extent of deprivation and disparities.
- In addition, even Censuses may have an incomplete coverage (hard to reach some populations, ethnic groups)
- And household surveys do not include institutional populations
- School based surveys are confined to children attending school
- Often administrative data cover only those enjoying services

- An estimated 250 million of the world's poorest and most marginalized people are estimated to be left out from surveys and censuses

- Data do not exist for particular disadvantaged groups such as slum dwellers, indigenous people and disabled children

- Around 70 countries across the world do not have high quality data on child mortality for the past five years

But disaggregating data implies not to ignore the following issues:

- the increasing costs of data collection and analysis

- the likely loss of data quality

- problems of confidentiality and transparency

4) The need to integrate various data sources

- Integrating data from different sources: surveys and censuses, administrative registers and new data sources (big data).

- Using big data (social media; mobile phones; scanners and image analysis) because they can generate information on aspects of life that are not captured by more traditional data sources, in particular about population groups often excluded from traditional data sources.

- Using advanced statistical methods: such as Small Area Estimation

5) Integrating administrative registers, census and survey data

The combination of survey and auxiliary data can improve the reliability of estimates without increasing the sample size of the survey. Of particular importance are administrative registers.

- Administrative registers are a source not much used hitherto
- There is a need to link census and survey data with administrative databases, as is quite common in Scandinavia.
- What are administrative data?

Administrative based data is the information collected primarily for administrative purposes. Governments collect this type of data for the purpose of registration, transaction and record keeping, usually during the delivery of a service to citizen(see, A.A. Chuwa, 2017)

These data however come from existing systems and it is not easy to use them to monitor progress in SDGs, especially given that in many developing countries the use of paper for administrative purposes is still widespread.

Introducing modern technologies in these countries and helping them using administrative data is hence of great importance.

Civil Registration Vital Statistics are here of particular importance (see, L. G. Gonzales Morales, 2017):

- From an a**dministrative and legal point of view:**

  - They provide documentary evidence and permanent record of people's legal identity, family relations and civil status

  - They secure people's rights and protect them from statelessness, early marriage, human trafficking, and other risks

  - They enable effective and efficient provision of government and social services

From a statistical, demographic and epidemiological point of view

  -They record personal socio-economic characteristics of the population

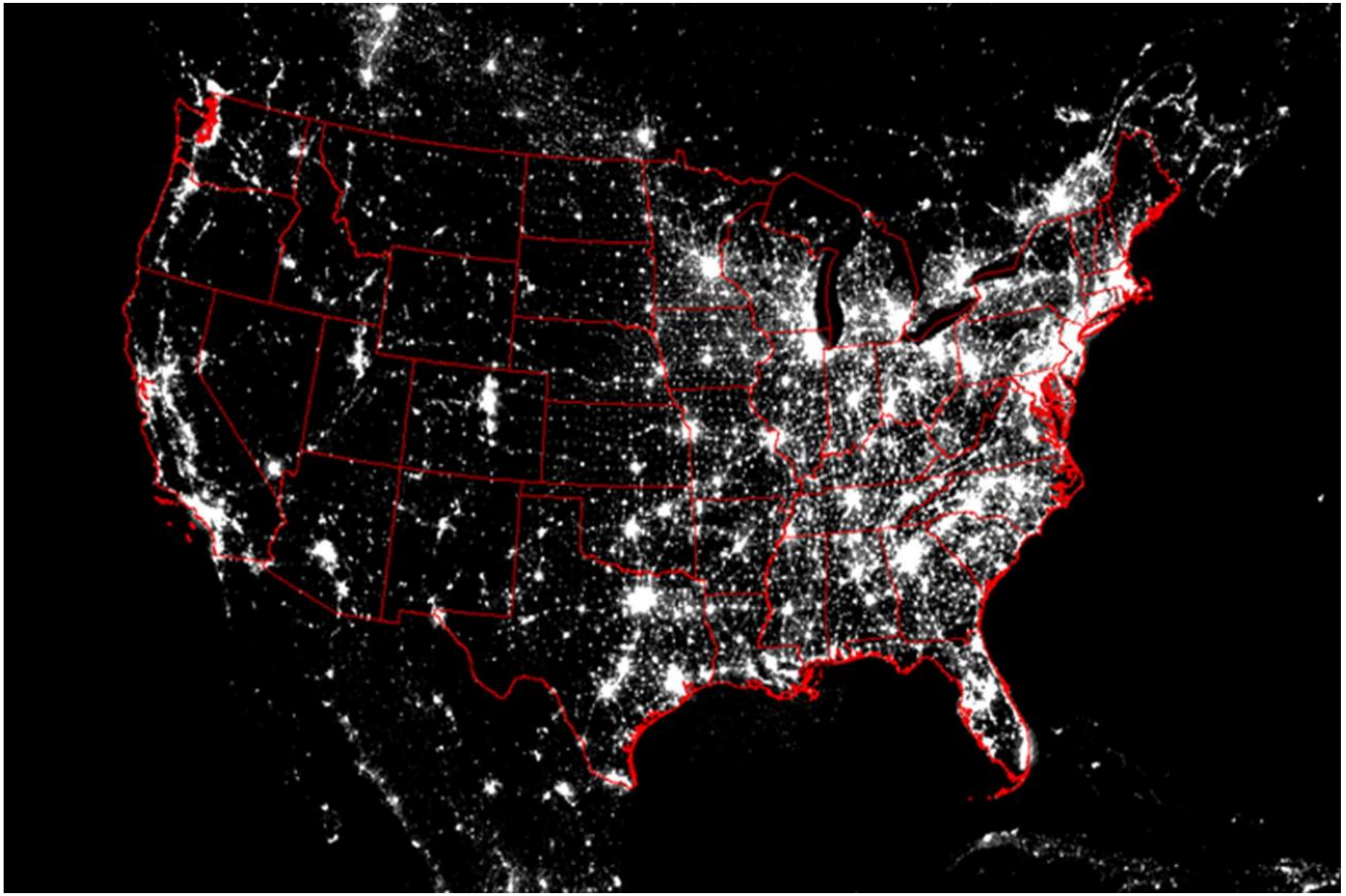
  - They empower policy and decision makers through statistical analysis of vital events by sex, occupation, education, ethnicity, etc.

6) Using Big Data to derive indicators for the SDGs

- Big data are a much less conventional data source, but they are of growing importance.
- Illustrations of the possible use of big data:
    - satellite images: (e.g. luminosity and poverty mapping)
    - mobile phone records (e.g. mapping the movement of mobile phones users can help predict the spread of infectious diseases)
    - social media data (e.g. sentiment analysis of social media can reveal public opinion on effective governance, public services delivery or human rights)

- Note however that to use the big data sources previously mentioned, there is clearly a need for training people on how to use these big data

Big data: a first illustration: Using luminosity data as a proxy for economic statistics (Chen and Nordhaus, 2011)



Big data: a second illustration: Social media fingerprints of unemployment (Llorente et al., 2014):

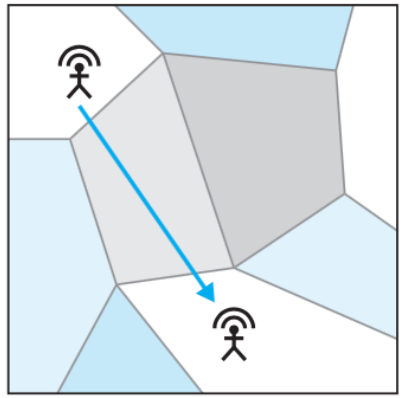
- They consider 19.6 million geolocated Twitter messages (tweets) from continental Spain, ranging from 29th November 2012 to 30th June 2013. They used four metrics:

1) Social media activity: regions with very different economical situations should exhibit different patterns of activity during the day. They hypothesized that communities with low levels of unemployment will tend to have higher activity levels at the beginning of a typical weekday.
2) Social media content: They built a list of 618 misspelled Spanish expressions and extract the tweets of the dataset containing at least one of these words. They then found a strong correlation between the fraction of misspellers and unemployment.
3) Social media interactions and geographical flow diversity: they considered all tweets mentioning another user and took them as a proxy for communication between users. They computed the diversity of communications with other areas (using entropy indices) and found that areas with large unemployment have less diverse communication patterns than areas with low unemployment.

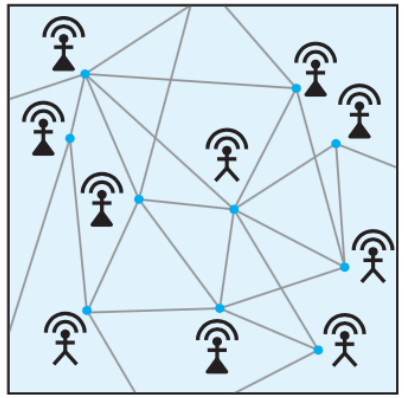
They concluded that regions exhibiting more diverse mobility fluxes, earlier diurnal rhythms, and more correct grammatical styles display lower unemployment rates.

Big data: Other illustrations of their possible use to derive indicators of sustainable development (see, J. G. Lee, 2017)

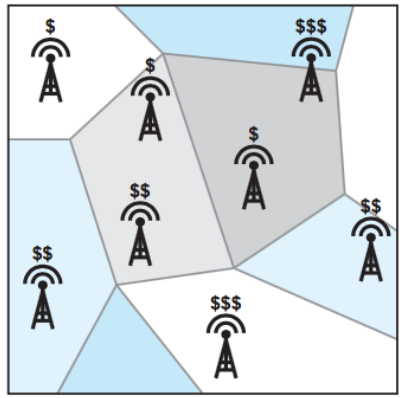
- Estimating the Indicators on Education and Household Characteristics from Anonymized, Aggregated Mobile Data



1. **MOBILITY:** As mobile phone users send and receive calls and messages through different cell towers, it is possible to "connect the dots" and reconstruct the movement patterns of a community. This information may be used to visualize daily rhythms of commuting to and from home, work, school, markets or clinics, but also has applications in modeling everything from the spread of disease to the movements of a disaster-affected population.



2. **SOCIAL INTERACTION:** The geographic distribution of one's social connections may be useful both for building demographic profiles of aggregated call traffic and understanding changes in behavior. Studies have shown that men and women tend to use their phones differently, as do different age groups. Frequently making and receiving calls with contacts outside of one's immediate community is correlated with higher socio-economic class.



3. **ECONOMIC ACTIVITY:** Mobile network operators use monthly airtime expenses to estimate the household income of anonymous subscribers in order to target appropriate services to them through advertising. When people in developing economies have more money to spend, they tend to spend a significant portion of it on topping off their mobile airtime credit. Monitoring airtime expenses for trends and sudden changes could prove useful for detecting the early impact of an economic crisis, as well as for measuring the impact of programmes designed to improve livelihoods.

More on big data and Sustainable Development (see, http://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html)

How data science and analytics can contribute to sustainable development

1 NO POVERTY
Spending patterns on mobile phone services can provide proxy indicators of income levels

2 ZERO HUNGER
Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time

3 GOOD HEALTH AND WELL-BEING
Mapping the movement of mobile phone users can help predict the spread of infectious diseases

4 QUALITY EDUCATION
Citizen reporting can reveal reasons for student drop-out rates

5 GENDER EQUALITY
Analysis of financial transactions can reveal the spending patterns and different impacts of economic shocks on men and women

6 CLEAN WATER AND SANITATION
Sensors connected to water pumps can track access to clean water

7 AFFORDABLE AND CLEAN ENERGY
Smart metering allows utility companies to increase or restrict the flow of electricity, gas or water to reduce waste and ensure adequate supply at peak periods

8 DECENT WORK AND ECONOMIC GROWTH
Patterns in global postal traffic can provide indicators such as economic growth, remittances, trade and GDP

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
Data from GPS devices can be used for traffic control and to improve public transport

10 REDUCED INEQUALITY
Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response

11 SUSTAINABLE CITIES AND COMMUNITIES
Satellite remote sensing can track encroachment on public land or spaces such as parks and forests

12 RESPONSIBLE CONSUMPTION AND PRODUCTION
Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products

13 CLIMATE ACTION
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation

14 LIFE BELOW WATER
Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities

15 LIFE ON LAND
Social media monitoring can support disaster management with real-time information on victim location, effects and strength of forest fires or haze

16 PEACE, JUSTICE AND STRONG INSTITUTIONS
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights

17 PARTNERSHIPS FOR THE GOALS
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

GLOBAL PULSE

www.unglobalpulse.org
@UNGlobalPulse 2016

There are however important issues to solve when using big data for SDGs:

- Most big data of value for SDGs such as mobile phone data and social media data are in the hands of competitive private firms
- These firms worry about sensitive information leaking out
- They are also concerned about negative public relations if controversies arise
- There is also the issue of data leaving the country
- “Pseudonymizing” data is costly
- Making data consistent so that they can be analyzed is also costly
- Making sure that the big data base obeys legal requirements is also costly

7) Integrating census and survey data: the case of poverty mapping

Small Area Estimation and Poverty Maps (van der Weide, 2017)

- Poverty maps are highly disaggregated databases of welfare (poverty and inequality, nutritional and health indicators, etc…)
- Note that the disaggregation need not be spatial (e.g. poverty of "statistically invisible" groups, such as individuals with disabilities)
- There are several reasons for the growing interest in poverty maps:
    1) As countries become richer, spatial disparities appear to be more accentuated.
    2) Simulations show that the same poverty reduction can be achieved with less than one-third of funds if targeted to the poorest communities.
    3) Poverty maps help identifying leading and lagging areas, correlating poverty with access to infrastructure, public services, education, equality of opportunity, safety nets, migration, segregation, agro-climatic conditions,…

Moreover, as stressed by van der Weide, poverty maps are useful because

- if there is decentralization of governance in some countries, this requires information on smaller administrative units
- they make small and vulnerable groups visible, such as disabled individuals, certain occupations, etc.
- they help reaching more poor households in areas with low percentage of poor

- Poverty maps at the World Bank started in the mid 1990s with
  1) publication of methodological papers (Elbers, Lanjouw and Lanjouw, 2003, *Econometrica;* Hentschel et al., 2000, World Bank; Elbers and van der Weide, 2014, World Bank)
  2) establishment of a PovMap Software (Qinghua Zhao et al.)
  3) the development of poverty maps, often in cooperation with national statistical offices, in over 70 countries

More recently the World Bank

- built poverty maps, even without a population census, using predictors derived from satellite imagery (night-time-lights, road network connectivity, greenness, pollution, etc…)
- updated poverty maps in between census years by combining population census and satellite imagery data

8) More on Small Area Estimation (SAE) and poverty mapping: a comparison of the methods

- See Guadarrama et al. (2014)

- There are pros and cons for each approach

9) More on big data and small area estimation (SAE): see Marchetti et al. (2016).

- Some local indicators could be obtained from big data (e.g.….) and then compared to those derived from SAE estimation
- If the same indicators cannot be obtained from big data and SAE, it is nevertheless possible to obtain additional covariates that can then be used in SAE (e.g. spatial information)

An illustration (see, Marchetti et al., 2016):

- The authors define the mobility M_d of an area as $M_d = (\sum_{v \in d} M_v)/V_d$

where M_v is the mobility of a given vehicle v and V_d is the number of vehicles in area d. The data on the mobility vehicles are obtained from big data using GPS. The authors then found a negative relationship between the standard deviation of the mobility of an area and the level of poverty in this area. In other words there are higher levels of heterogeneity of mobility in areas with lower levels of poverty.

10) An interesting case of Small Area Estimation: the Community Based Monitoring System (CBMS) in the Philippines (see, Mandap, 2017)

- The Community-Based Monitoring System (CBMS) collects and processes data at the local level and integrates them in local planning and impact-monitoring.
- It promotes evidence-based policymaking and empowers communities to participate in the process.
- It was designed in 1993 to provide policymakers with a good information base for tracking the impacts of various economic reforms and policy shocks on the vulnerable groups in the society.
- A local government code was passed in 1991 in the Philippines, devolving many functions to the local governments. This evidently increased the demand for more disaggregated data.
- Moreover there has been an increasing focus on targeted interventions because of limited funds. This implies collecting information at the household level in order to identify eligible beneficiaries.
- CBMS is also likely to facilitate greater transparency and accountability in local governance

What are the key features of CBMS?

- It involves a census of all households in a community
- It involves the local government unit and some of its members will monitor the work
- The output of this project is a set of indicators assumed to measure the welfare status of the population. These indicators are selected to reflect the multidimensional nature of poverty
- A database is then established at the local level

List of CBMS poverty indicators

Dimension	Indicator	
Health and Nutrition	Proportion of children under 5 who died	Health poor
	Proportion of women who died due to pregnancy related causes	
	Proportion of children aged 0-5 who are malnourished	Nutrition poor
Housing	Proportion of households in makeshift housing	Housing poor
	Proportion of households who are informal settlers	Tenure poor
Water and Sanitation	Proportion of households without access to safe water supply	Water poor
	Proportion of households without access to sanitary toilet facilities	Toilet poor

List of CBMS poverty indicators (cont.)

Dimension	Indicator	
Education	Proportion of children 6-11 years old who are not attending elementary school	Education poor
	Proportion of children 12-15 years old who are not attending secondary school	
	Proportion of children 6-15 years old who are not attending school	
Income and Hunger	Proportion of households with income below the poverty threshold	Income poor
	Proportion of households with income below the food threshold	Income poor (extreme)
	Proportion of households who experienced hunger due to food shortage	Food poor
Employment	Proportion of persons in the labor force who are unemployed	Job Poor
Peace and Order	Proportion of persons who are victims of crime	Security Poor

Some of the additional data collected by CBMS

- physical and demographic characteristics of the village (barangay)
- service institutions and infrastructure
- disaster risk reduction and preparedness
- peace and order
- budget, revenue and expenditure
- household/member characteristics
- education
- political and community participation
- health and nutrition
- income, employment and livelihood
- housing and tenure, water sources and sanitation
- migration
- impacts of climate change

11) Some concluding comments

The SDG index and Dashboards Report 2018, published jointly by the Bertelsmann Foundation and by the Sustainable Development Solutions Network (SDSN) stresses that

- Most G20 countries have started SDGs implementation but important gaps remain
- No country is on track to achieving all SDGs
- Conflicts are leading to reversals in SDG progress
- Progress towards sustainable consumption and production patterns is too slow
- High income countries generate negative SDG spillover effects
- Inequalities in economic and social outcomes require better data.

The Bertelsmann Foundation and the Sustainable Development Solutions Network (SDSN) undertook also a preliminary assessment of government commitments to achieve the SDGs.

Among the conclusions of this assessment:

- There are considerable variations among G20 countries regarding institutionalization of the SDGs
- Countries such as Brazil, Italy and Mexico demonstrate relative high levels of institutionalization (SDG strategies, action plans, coordination units in government, etc…)
- In contrast, countries such as the United States and the Russian federation show low levels of political leadership and institutionalization of the SDGs (absence of public statements made by the head of state on how the country plans to implement the SDGs).

To make a long story short it looks like in many countries we are still far from reaching the SDGs at the national level or even from moving in the right direction. It seems even that, although the SDGs were adopted by a vast majority of countries, this was an agreement of principles which does necessarily imply a strong will to implement them.

Efforts to measure SDGs at a lower level remain hence at this stage useful but it is quite clear that SAE of SDGs are still at an experimental stage and will in the near future involve only a relatively small number of countries.

Bibliography

Chambers, R. L. and R. Dunstan (1986) “Estimating Distribution Functions from Survey Data,” *Biometrika* 73(3): 597-604.

Chambers, R. and N. Tzavidis (2006) “M-quantile Models for Small Area Estimation,” *Biometrika* 93: 255-268.

Chen, X. and W. Nordhaus (2011) “Using luminosity data as a proxy for economic statistics,” *Proceedings of the National Academy of Sciences* 108 (21) 8589-8594. Available at https://doi.org/10.1073/pnas.1017031108

Chuwa, A. (2017) “Building Confidence in the Administrative - based Data - A Case of United Republic of Tanzania,” International Conference on Sustainable Development Goals Statistics, Manila.

Elbers, C., J. O. Lanjouw and P. Lanjouw (2003) “Micro-Level estimation of Poverty and Inequality,” *Econometrica* 71(1): 355-364.

Elbes, C. and R. van der Weide (2014) “Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality,” Policy Research Working Paper 6962, Development Research Group, Poverty and Inequality Team. The World Bank, Washington.

Fay,R. and R. Herriot (1979) “Estimates of income for small places: An application of the James-Stein procedures to census data,” *Journal of the American Statistical Association* 74: 269-277.

Giusti, C., S. Marchetti, M. Pratesi and N. Salvati (2012) “Robust Small Area Estimation and Oversampling in the Estimation of Poverty Indicators,” *Survey Research Methods* 6(3): 155-163.

Gonzalez Morales, L. (2017) “Integration of Different Data Sources for SDG Monitoring: Increasing Usability of CRVS,” International Conference on Sustainable Development Goals Statistics, Manila.

Guadarrama, M., I. Molina and J. N. K. Rao (2014) “A Comparison of Small Area Estimation methods for Poverty Mapping,” *Statistics in Transition new series and Survey Methodology* 17(1): 41-66.

Hentschel, J., J. O. Lanjouw, P. Lanjouw and J. Poggi (2000) “Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty

A Case Study of Ecuador,” *World Bank Economic Review* 14(1): 147 – 165.

Horvitz, D. G. and D. J. Thompson (1952) “A Generalization of Sampling Without Replacement from a Finite Universe,” *Journal of the American Statistical Association* 47(260): 663-685.

Jiang, J., T. Nguyen and J. S. Rao (2011) “Best Predictive Small Area Estimation,” *Journal of the American Statistical Association* 106(494): 732-745.

Lee, J. G. (2017) “Big Data Revolution for Sustainable Development and Humanitarian Action,” International Conference on Sustainable Development Goals Statistics, Manila.

Llorente, A., M. Garcia-Herranz, M. Cebrian and E. Moro (2015) “Social media fingerprints of unemployment,” *PLoS One* 10(5).

Marchetti, S., N. Tzavidis and M. Pratesi (2012) “Non-parametric bootstrap mean squared error Estimation for M-quantile Estimators of Small Area Averages, Quantiles and Poverty Indicators,” *Computional Statistics & Data Analysis* 56(10): 2889-2902.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S, Rinzivillo, L. Pappalardo and L. Gabrielli (2015) “Small Area Model-Based Estimators Using Big data Sources,” *Journal of Official Statistics* 31(2): 263-281.

Martinez, A. (2017) “A Snapchat on Analytical Tools for Disaggregating SDG Data,” International Conference on Sustainable Development Goals Statistics, Manila.

Martinez, A. (2017) “How data science and analytics can contribute to sustainable development,” International Conference on Sustainable Development Goals Statistics, Manila.

Min, Y. (2017) “Data Disaggregation and the SDGs: An Overview,” International Conference on Sustainable Development Goals Statistics, Manila.

Molina, I. and J. N. K. Rao (2010) “Small area estimation of poverty indicators,” *The Canadian Journal of Statistics* 38(3): 369-385.

Reyes, C. M. and A. B. E. Mandap (2017) “Monitoring the Sustainable Development Goals (SDGS) at the Local Level Through the CBMS,” International Conference on Sustainable Development Goals Statistics, Manila.

SDG Index and Dashboards Report 2018. Gobal responsibilities. Implementing the Goals. Bertelsmann Stiftung and Sustainable Development Solution Network.

Truszczynski, M. (2017) “Disaggregated SDG Monitoring by a Wider Use of Administrative Registers,” International Conference on Sustainable Development Goals Statistics, Manila.

Tzavidis, N., S. Marchetti and R. Chambers (2010) “Robust estimation of Small-Area Means and Quantiles,” *Australian and New Zealand Journal of Statistics* 52(2): 167-186.

United Nations “Big Data for Sustainable development,” Available at http://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html

Van der Weide (2017) “Poverty Mapping at World Bank,” International Conference on Sustainable Development Goals Statistics, Manila.