Using Big Data in Official Statistics: Why? When? How? What for?

GianLuigi Mazzi, Technical Director GOPA Luxembourg, [1]

CARMA 2018, Valencia

[1] Contact email: GianLuigi.Mazzi@gopa.lu

Outline

1. Introduction
2. Traditional versus Big Data Sources.
3. Delimiting and Classifying the Big Data Ecosystem.
4. Big Data Types.
5. Big Data and Key Policy Actions.
6. Macro-economic Growth and Stability
7. Labor Market Policies
8. Sustainable Development Policies
9. Providing Some “Answers”

Introduction

- The use of big data in official statistics is still an open issue.
- Policy makers request more and more information not necessarily available in traditional statistics.
  - □ Big data might fill such gaps
- But using big data is a challenging exercise for official statisticians who face 2 main obstacles:
  - □ They are often based on “unrequested information” (not collected on the basis of a robust sampling scheme);
  - □ And they are often available in unstructured form.
- Why, When, How and What for should we use these data?
- We try to provide some directions without pretending to be neither exhaustive nor conclusive, but aiming to provide an additional contribution to the ongoing debate.

Traditional versus Big Data Sources (1)

- Traditional data sources such as census, surveys, opinion pools, administrative registers typically structured as:
  - Panel: same phenomenon recorded on various units at the same time;
  - Spatial series: same phenomenon recorded in different geographical locations at the same time;
  - Time-series: same phenomenon recorded at different points in time, possibly equally spaced;
  - Combinations of the above structures: spatial time series, time series panels.
- Big data usually look different:
  - irregularly recorded
  - characterized by discontinuities and irregularities
  - appear as a pretty chaotic amount of information
  - Big data typically unstructured.

Traditional versus Big Data Sources (2)

- Well consolidated way of dealing with traditional data
  - □ availability of a large variety of tools and methods serving a number of purposes
  - □ well-tested and powerful software solutions
- Much more complex working with big data
  - □ converting them into structured data
  - □ working on original unstructured data
- Working on structured big data
  - □ existing tools and methods should be adapted to the size of data and to some big data specific features
  - □ new software developments required
- Working on originally unstructured data
  - □ developing new tools: data analytics
  - □ or improving existing techniques: data mining

Traditional versus Big Data Sources (3)

⟹ **If working with big data is so demanding and challenging, why keep insisting and not just concentrate on traditional data?**

Some more questions still to be added:

- When big data should be really beneficial for policy makers, analysts and statisticians?
- How should big data be used?
- What should big data be used for?

Some tentative and not exhaustive answers provided later:

- Stemming from the provided examples;
- Unavoidably reflecting personal views.

Delimiting and Classifying the Big Data Ecosystem

- IBM so-called 4Vs classification
  1. Volume
  2. Velocity
  3. Variety
  4. Veracity
- UNECE classification
  1. human sourced information including social networks
  2. traditional business systems
  3. internet of things
- Size based classification of big data by Dornik and Hendry
  1. Tall: not many variables but many observations;
  2. Fat: many variables and few observations;
  3. Huge: many variables and many observations.

Big Data Types (1)

- None of the classifications proposed above fully satisfactory
- Instead, propose a classification of big data by types based on their origin and generating process.
- Associating big data types to relevant policies.
- Identifying advantages and drawbacks of various big data types

Big Data Types (2)

By type and main utilization.

1. **Financial market data**: Macroeconomics, financial sector monitoring
2. **Electronic payments data**: Macroeconomics, inflation, consumers behavior
3. **Mobile phone data**: Labor market, sustainable development
4. **Sensor data and the Internet Of Things** : Sustainable development, urban monitoring
5. **Satellite image data**: Sustainable development, economic growth and land utilization
6. **Scanner prices data**: Macroeconomics, inflation, consumers behavior
7. **Online prices data**: Macroeconomics, inflation, consumers behavior
8. **Online search data**: Macroeconomics, sustainable development, human behavior
9. **Textual data**: Human sentiments, confidence, uncertainty
10. **Social media data**: Macroeconomics, sustainable development, human behavior

Big data and key policy actions (1)

- Policy makers in different areas need complete, reliable and often timely information for designing, implementing and monitoring policy initiatives
  - Using qualitative and quantitative information
- Information set based on traditional sources
  - surveys, opinion pools, administrative data
- Available information no necessarily in line with policy makers expectations
  - lack of information in some areas or domains
  - available information not enough reliable
  - information made available too late
- Using alternative source: big data
  - filling existing information gaps
  - allowing for designing and implementing new indicators
  - helping in identifying new patterns and relationships useful for policy making activities

Big data and key policy actions (2)

- Focusing on a few relevant policies
  - □ macro-economic growth and stability
  - □ labor market policies
  - □ sustainable development policies

- Describing how big data can contribute in building-up the information sets required for each policy
  - □ associating user needs with specific big data types

Macro-economic Growth and Stability (1)

- Producing earlier estimates of GDP household consumption and retail sales turnover
  - □ financial market data, electronic payment data, online search data (Google Trends data)
  - □ Koop and Onorante (2013), Baldacci et al (2016) and Buono et al (2018)
- Producing weekly/daily indicators for GDP households consumption and retail sales turnover:
  - □ financial market data, electronic payment data
  - □ Galbraith and Tkacz (2007) Stock and Watson (2002a), Giannone, Reichlin and Small (2008) and Aprigliano et al (2016).
  - □ Complex models (data reduction tools and mixed frequency models).
- The real benefit of producing high frequency indicators still to be evaluated:
  - □ Better monitoring of the economic situation versus risk of confusing users.

Macro-economic Growth and Stability (2)

Measuring inflation

- We already have timely indicators .....
- ..... that could be improved using scanner price data and online price data;
- Improving the overall quality of the inflation measurement
  - □ improved coverage
  - □ up-to-date weights at a highly detailed level
- Can even replace the HICP (experiments in Netherlands, Luxembourg etc.).
- Potential for nowcasting and/or forecasting inflation

Macro-economic Growth and Stability (3)

- New information provided by big data
  - □ Volatility and market micro-structure studies
    financial market data
  - □ Consumer behavior in relation to unexpected events
    electronic payment data
  - □ Analysis on the price behavior at industry and regional level
    scanner price data
  - □ Consumer sentiment and consumer confidence
    social media data
- Providing information on economic uncertainty based on texts from newspapers, twitter and etc.
  - □ text mining and text analytic techniques

Labor Market Policies (1)

- Labor market indicators usually based on periodic surveys
  - implementation problems in developing countries
  - some persisting drawbacks in terms of timeliness also in developed countries
- Big data can complement traditional information system
  - labor force surveys cannot be discarded
- Role of big data can significantly differ according to the degree of development of countries
  - mostly used for improving timeliness in developed countries;
  - providing basic market information in developing countries.

Labor Market Policies (2)

- Improving employment/unemployment estimates:
  - □ online search data (Google Trends), mobile phone conversation, mobile phone position;
  - □ D'Amuri and Marcucci (2012) and Tuhkuri (2016) investigate the power of big data in nowcasting and forecasting Unemployment data by using Google Trend.
- More detailed pictures at geographical level
  - □ Toole et al. (2015) forecasted the employment at regional and European countries levels by using the call duration information and changing behavior in social communication related to the employment status; innovative approach based on Bayesian classification models.
- Possibility of obtaining information already during the current period
- Encouraging results obtained in several developed or developing countries (European union and Africa)

Labor Market Policies (3)

New information provided by big data:

- individual employment status
  - □ using mobile phone data validated by household survey data
- measuring the effect of employment shocks on individual behavior
  - □ mobile phone and GPS data
- improving the match between job vacancy and labor demand
  - □ disseminating information by SMS
- using big data from online job searching portal to assess demand for workforce skills and observing job-search behavior and improving skills matching
  - □ online data search

Sustainable Development Policies (1)

- Complexity of the sustainable development policies:
  - 17 goals, 169 targets, 230 indicators
- Traditional information system unable to ensure the follow up of all goals and targets
  - persistent differences among developing and developed countries
  - lack of traditional information particularly relevant for some goals: difficulties in measuring complex phenomena (poverty, well-being)
- Several big data types can help filling the information gap
  - ensuring a better follow up of goals and targets.

Sustainable Development Policies (2)

As examples, we focus on some Sustainable Development Goals:

- Goal 01: End poverty in all its forms everywhere
- Goal 03: Ensure healthy lives and promote well-being for all at all ages
- Goal 08: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
- Goal 11: Make cities and human settlements inclusive, safe, resilient and sustainable

SDG 1: End poverty in all its forms everywhere

- Traditionally measured by surveys:
    - surveys often ignore people outside the traditional household
- Using call data records (mobile phone) can produce good estimates of poverty
    - in absence of reliable surveys
    - selecting appropriate keywords related to the poverty status
- Using complex methods based on mobile phones data supported by other information such as:
    - statistical survey
    - night lights satellite images
    - automated recognition of roofing materials of new structures (sensor)
    - trade estimation via real-time postal traffic analysis
    - trends in retail purchasing behavior
- The last approach, ideally provides the most reliable estimation of poverty almost in real time:
    - disaggregated at regional and local level

SDG 3: Ensure healthy lives and promote well-being for all at all ages

- Well-being is a complex multidimensional phenomenon not easily measurable
  - involving qualitative and quantitative variables
  - surveys on individual well-being
- Alternative well-being measures based on Google trend data obtained by choosing well-being related keywords
  - good approximation of survey based measures
  - better timeliness
  - data available weekly
  - possibility to derive a total measure of well-being
- Other aspects related to the well-being:
  - good health: mapping the movement of mobile phone users to predict the spread of infectious diseases
  - well-being of women and girls: social media data, mobile phone data

SDG8: Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all

- Macroeconomic indicators generally reliable especially in developed and emerging countries
    - some lack of information observed at sub country level
- Satellite data can help in providing a detailed mapping of economic development
    - using lighting nighttime luminosity (combining satellite images with population distribution)
- Mobile phone data can provide detailed picture of the employment status and work conditions
    - complemented also by web search engine data
- Mobile phone also as an instrument for helping unbanked people in accessing financial services
    - micro-credit facilities

SGD 11: Make cities and human settlements inclusive, safe, resilient and sustainable

- Traditional information not particularly developed at urban level (small area estimates can only partially fill the gap)
- Since more than 40% of global population is planning to live in urban area by 2030, reliable and complete information is needed
- Monitoring urban congestion and traffic situation
  - □ traffic lights data and other sensor data
- Mapping mobility patterns
  - □ mobile phone data
  - □ identifying poverty area within cities: satellite imagery data
- Monitoring pollution
  - □ IOT sensor data
- Land use and cover (satellite data)

Providing Some "Answers" (1)

Why keep insisting on big data and not just concentrate on traditional data?

- Big data has an impressive potential in terms of information they contain
  - □ not yet fully discovered, understood and used
- They can influence several aspects of our daily life and habits
  - □ decision making processes, understanding the society, providing an almost real-time and more granular picture of our community
- They can positively influence sciences
  - □ observational sciences, medicine, economy and finance, social sciences, education and more

- But big data have to be known, understood and interpreted carefully
  - □ avoiding misleading or erroneous conclusions
  - □ big data are an information set far to be perfect

Providing Some "Answers" (2)

When big data should be really beneficial for policy makers, analysts and statisticians?

- Filling gaps in traditional information sets
    - timeliness, relevance, coverage (cross-sectional level, geographical level)
- To complement non-adequately developed traditional information systems
    - developing and under-developed countries
- Do helping measuring complex phenomena
    - poverty, well-being

Providing Some "Answers" (3)

How should big data be used?

- As much as possible complementing traditional information systems
  - □ replacement is definitely premature
  - □ exceptions: developing and under-developed countries
- Within a robust rationally build and methodologically sound environment
  - □ adequate infra-structure for hosting and treating big data
  - □ adapted traditional statistical and econometric tools do deal with large and often sparse datasets
  - □ developing new tools dealing with unstructured data: Data mining, Data analytics
- Having clearly in mind big data potential but also its limitations and drawbacks
  - □ non based on a robust sampling frame
  - □ based on often non-requested information

Providing Some "Answers" (4)

What should big data be used for?

- Improving timeliness
  - □ nowcasting/forecasting
  - □ advanced estimates (already available during the reference period)
  - □ proving higher frequency estimates (weekly/daily)
- Providing information on key phenomena when traditional data sources are lacking
  - □ unemployment/employment, inflation
- Providing indicative measures of complex phenomena
  - □ material deprivation
- Providing information on individual and collective behavior and feelings
  - □ indicators measuring changes in behavior
  - □ sentiment indicators
  - □ confidence indicators

Thank You!!!

GOPA
Consulting Group

Contact email: GianLuigi.Mazzi@gopa.lu