

The challenges of smart data in Official Statistics

**Past, present and future
of data sources**

Challenges of Statistics

- Data to change: data produced using innovative methods
 - Increasingly daily dependence on technologies, social media and electronic transactions produces data in the most varied areas of natural and social sciences
 - They are labelled “big” for dimensions and complexity (text, diagnostic, satellite images, signals,...)
 - Paradigms of their analysis to bring benefits to the society and to everyone’s life are not clear

Evolution of Statistical literacy

- Data are the “diamond mine” of current age
- “Data” is the plural of “datum”, that means provided by an agent (the producer) at a certain point in time and space and with a medium
- Every “datum” is a representation of reality, even when it is numeric, contextual to the event, immediately available
- Data awareness and metadata (ontologies in OS)
- Data reproducibility
- Technological evolution and cultural evolution

Statistical thinking

- From **statistical literacy** to “**statistical thinking**”: numeracy, datacy and data awareness, key concepts in the art of thinking clearly
- Modern antiques: average, correlation, heterogeneity, sampling variability, selectivity, confounding effect, error propagation, systematic error, random error, uncertainty
- Statistics: multidisciplinary nature evolves into a over-disciplinary language to extract, transmit and communicate knowledge

Statistical thinking

- **Statistical thinking** is evolved and still evolving, with a focus on data awareness and data literacy, and on the uncertainties and measurement errors emerging in the analysis of modern data sources

Statistical thinking

- **Beyond calculations and measures**
- **All analytics work begins and ends with a story**
- Large proportion of statistical mistakes result from incorrect logic or interpretation despite correct numerical calculations
- Need of integrating mathematical and non mathematical knowledge
- Think in distributions and probability
- Understanding the role of statistical analysis within the greater machinery of generating scientific knowledge

Ashley Steel et al, Beyond calculations: a course in statistical thinking, 2019

Statistical thinking

- Statistical thinking is a dynamic process, which follows the evolution of times.
- It is taught and learned by following the evolution of technologies for data production and the evolution of statistics (new methods for new data: from traditional regression models to machine learning and deep statistical learning, passing through descriptive statistics, data analysis, up to the analysis of functional data) allowing longitudinal perspectives (historical series of simple and multiple data) and territorial (spatial statistics) and also joint perspectives (space-time perspectives)

The challenges of smart data

- *Data sources for statistical production. Main features, quality aspects, advantages and disadvantages of:*
 - **censuses**
 - **survey data**
 - **administrative source**
 - **big data**
- *What is a Trusted Smart Statistics?*

The challenges of smart data

Data sources for statistical production: a critical list

- Censuses
- Survey data
- Administrative sources
- Big data

Population and Housing Census: the big change and the future

- Definition and objectives of Population and Housing Census (PHC)
- The decade 2010 - 2020: period of innovation and experiment
- The impact of Covid19
- The challenges for the future

PHC: definition and objectives

Definition

- *A population and housing census (short "census") counts the entire population and housing stock of a given country and collects information on its main characteristics (geographic, demographic, social and economic, plus household and family characteristics).*

Objective

- *Census data are a rich source of statistical information, ranging from the lowest geographical divisions, covering small areas, to the national and international levels.*
- *The data collected by census help a nation, region or community make major decisions for the future. For example, the results of a census are used to distribute and allocate government funds for education, health services and delineating electoral districts. Census data can also be used for academic research or business marketing.*

PHC: the decade 2010-2020

On line information official sources



2020 World Population and Housing
Census Programme



eurostat

Your key to European statistics

- *It contains current information on national, regional and international activities related to the Programme.*
- *It presents international standards, methods and guidelines intended to assist national statistical offices and other producers of official statistics in planning and carrying out successful population and housing censuses.*
- *It is a repository of methodological reports and documents from countries.*
- *It provides information on population and housing censuses in the European member states, which are the backbone of population and social statistics.*
- *It presents the updated European legislation.*
- *It contains the major innovation from 2011 to 2021 round*
 - *The level of dissemination*
 - *The way of dissemination*
 - *The availability of data*



European
Commission

PHC: the decade 2010-2020

Main strategies

!

Before 2010

Since the 2000 census round, the number of countries using traditional census methods has decreased

Year 2000: 27 traditional census
9 use register in the statistical process

Year 2010: 1 rolling census
6 register based or mixed mode
18 use register in the statistical process

2010 - 2020

!

1. *Traditional Census*
2. *Traditional Census with early update*
3. *Register-based census*
4. *Rolling census*
5. *Other methods*

PHC: the decade 2010-2020

Main strategies

1. **Traditional Census.** *All individuals are enumerated **directly** and their characteristics are registered through the completion of census forms.*
All European countries in 2000 wave

! Pressure to make greater use of information available elsewhere, lower public cooperation and participation, changing user demand and the need to control or reduce costs -> need to change some aspects of the statistical process!

2. **Traditional Census with early update.** *Two forms: long and short (Canada and USA since 2000). Use of some sampling strategies to release information (Italy, 2011). Use of some administrative information for statistical purposes, i.e. for gathering elusive population (migrants)*

PHC: the decade 2010-2020

Main strategies

3. **Register-based census.** *No field enumeration. Well established adm register (Denmark and Finland). No burden and less costly than 2. and 3. The population characteristics considered are limited to those available in the registers. This demands close cooperation between the statistical agency, register authorities and the public administration, and strict legislative oversight.*
- *Combination of register data with complete enumeration (2011, Czech Republic, Estonia, Italy, Latvia, Lithuania and Spain).*
 - *Combination of register data with existing surveys : "virtual census" (Netherlands and Slovenia in 2011).*
 - *Register data and ad hoc surveys. Sample surveys conducted ad hoc for the census, either to evaluate the accuracy and completeness of the registers or to include new variables (like in a long form) (Israel 2008, Italy 2018-2021).*
 - *Traditional enumeration with yearly updates of characteristics on a sample basis (USA, 2010)*

PHC: the decade 2010-2020

Main strategies

4. **Rolling census.** *It is based on a moving average over five consecutive years.*
- *Small municipalities (fewer than 10,000 inhabitants) are divided into five groups, and a full census is conducted each year in one of the groups. In all large municipalities, a sample survey covering 8% of dwellings is conducted each year.*
 - *After five consecutive years, the entire population in small municipalities and about 40% of the population in large municipalities has been surveyed. In all, about 70% of the population is covered in the course of the five-year cycle. This is enough to guarantee robust information at the municipality and neighbourhood levels.*

-> This method was developed mainly to improve the frequency of the data releases, and to spread over time the financial and human burden associated with the census.

It is unique to France.

PHC: the decade 2010-2020

Main strategies

3. **Other methods.** *Integration of Permanent Census and Social Surveys. The Italian project*

The strategy consists of two phases

1

The Population Census Master Sample

Yearly in Autumn (starting from 2018),
with the aim of:

- **correcting** the Register of individuals for under and over coverage
- **collecting** information for not replaceable variables

2

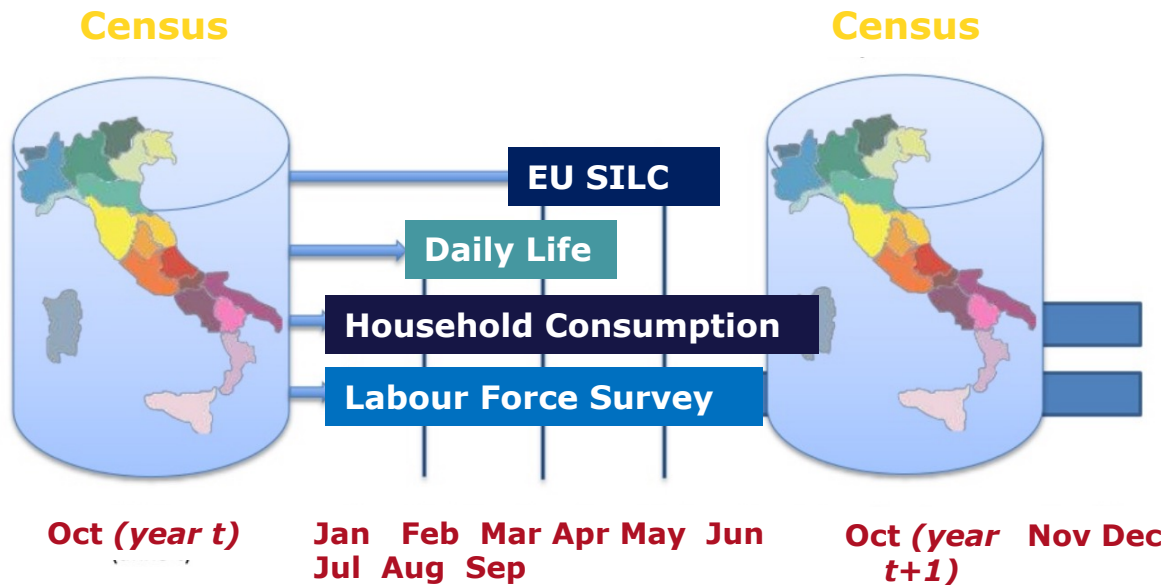
The Social Surveys

Through the year after the first phase

Sample households are selected as a **sub-sample** of those already involved in master sample

PHC: the decade 2010-2020

Other methods. Integration of Permanent Census and Social Surveys.
The Italian project



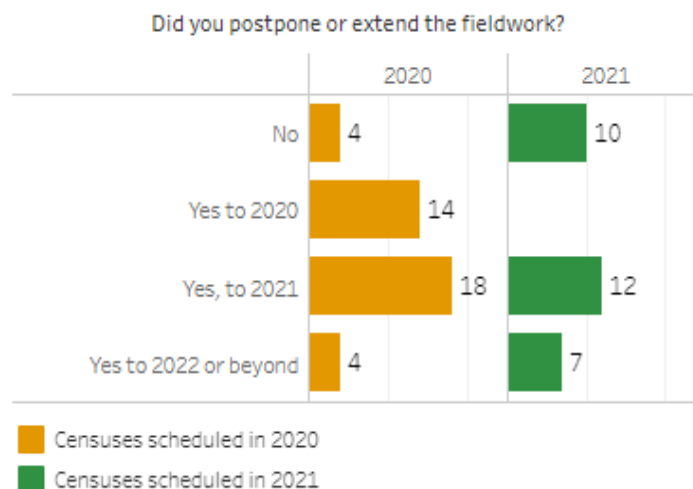
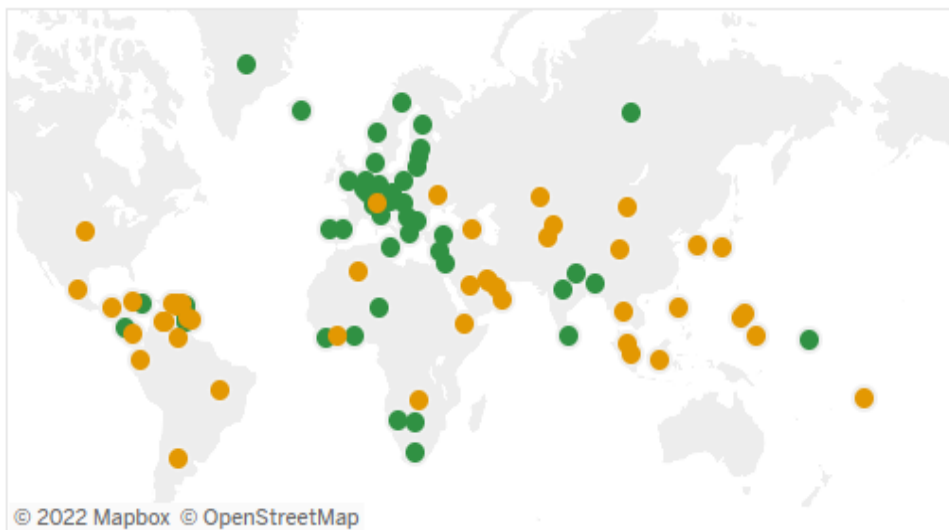
Survey	Sampled households
Daily Life	25.000
EU-SILC	30.000
Hous. Consum	20.000
LFS	250.000

PHC: the impact of Covid19

The pandemia of Covid19 had an impact also on statistical production process.

From 2019 to 2021 many NSIs (69) planned to realize Population census.

And a lot of them (55) decided to stop and/or postpone the census operations



<https://unstats.un.org/unsd/demographic-social/census/COVID-19/>

PHC: the impact of Covid19 and the future challenges

Reasons

Difficult to realize face to face interviews

1

Difficult to recruit personnel

2

Funding limitations and constraints

3

Mobility constrictions

4

Closure of establishment

5

Challenges

To maintain high response rate using mixed data collection methods

To change the system of recruitment and payment of census personnel

To save costs (i.e. decreasing the sample for sub set of population)

To increase the use of web communication and web interviewes (web platform)

To increase the virtual meeting

PHC: ...and other challenges

Reasons

Increasing response burden -> reduction of response rate (also web response rate)

6

Challenges

To improve mixed mode and mixed data sources

Increasing institutions burden

7

- a. to respect the deadline for the expected contribution;
- b. to reinforce the statistical competencies inside the public institutions;
- c. to contrast the «routine effects» due to a frequent operation

Social and economic surveys

Definition

- ▶ A survey usually begins with **the need for information where no data – or insufficient data – exist**
- ▶ A survey is **any activity that collects information** in an **organized manner** about characteristics of interest from some or all units of a population using **well-defined concepts, methods and procedures**, and compiles such information into a **useful summary form**

Social and economic surveys: key points

- Improving the **integration** of social and economic surveys
- Addressing **non response**
- Improving focused **mixed mode surveys**
- Collecting and analysing **paradata**

Social and economic surveys

Advantages

- Good planning of objectives, questionnaire, survey frame, sampling methods
- Good estimation methods
- Good data quality

Disadvantages

- Too much costly
- Excessive response burden
- Decreasing of the response rate over time

Social and economic surveys: key points

Improving the integration of social and economic surveys

Objectives

- * to increase the cost effectiveness and relevance of survey data production
- * to achieve levels of accuracy and granularity

(Some) Strategies

- to improve the role of respondents
- to investigate the reasons of burden

Addressing non response

Objectives

- to increase the response rate
- to reduce the impact of low response rate on estimation

(Some) Strategies

- to integrate survey data using a direct record linkage
- to improve coverage using different survey (about related topic)

Social and economic surveys: key points

Improving mixed mode surveys

Objectives

- to increase the sample efficiency
- to increase the response rate

(Some) Strategies

- to evaluate the use of mixed mode in some sub phase of statistical process
- to improve the analysis of the different mode

Collecting paradata

Objectives

- to improve survey quality control (i.e. compliance of enumerators)

(Some) Strategies

- to calculate paradata directly from data collection system (in the questionnaire some specific part for the enumerator)

Administrative sources

Definition

- ▶ **Adm source.** *A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations. In a wider sense, any data source containing information that is not primarily collected for statistical purposes.*
- ▶ **Adm data.** *The data derived from an administrative source, before any processing or validation by the National Statistical Institutes (NSIs)*

**Source: ESSnet (European Statistical System net)
Admin Data, WP1 (2013)**

Administrative sources: the emerging need of data integration

Advantages

- Save money
- Reduction of response burden
- Availability of more detailed data (granularity)

Disadvantages

- Target population is different from administrative population
- Difficulties in the access to data
- Need of evaluation of data quality (sample criteria?)

UE RECCOMANDATION OF USING ADMINISTRATIVE DATA



European
Commission

Administrative sources: the challenges

- Adopting common definitions and standards including adm data
- Adapting the statistical production process respect to data integration and data harmonization steps
- Specifying a possible new type of error

Big data (and social data): the real challenges

Definition

*"Big data is like teenage sex:
everyone talks about it,
nobody really knows how to
do it,
everyone thinks everyone
else is doing it,
so everyone claims they are
doing it"*

*[Dan Ariely, **2003** | Prof. of Psychology and Behavioural Economics at Duke University]*

*"Big Data sources can
generally be described as:
high volume, velocity and
variety of data that
demand cost-effective,
innovative forms of
processing for enhanced
insight and decision
making"*

*[Daas, Puts, Buelens, van den Hurk, **2015** | JOS, Vol. 31, no. 2]*

Big data: the real challenge

Advantages

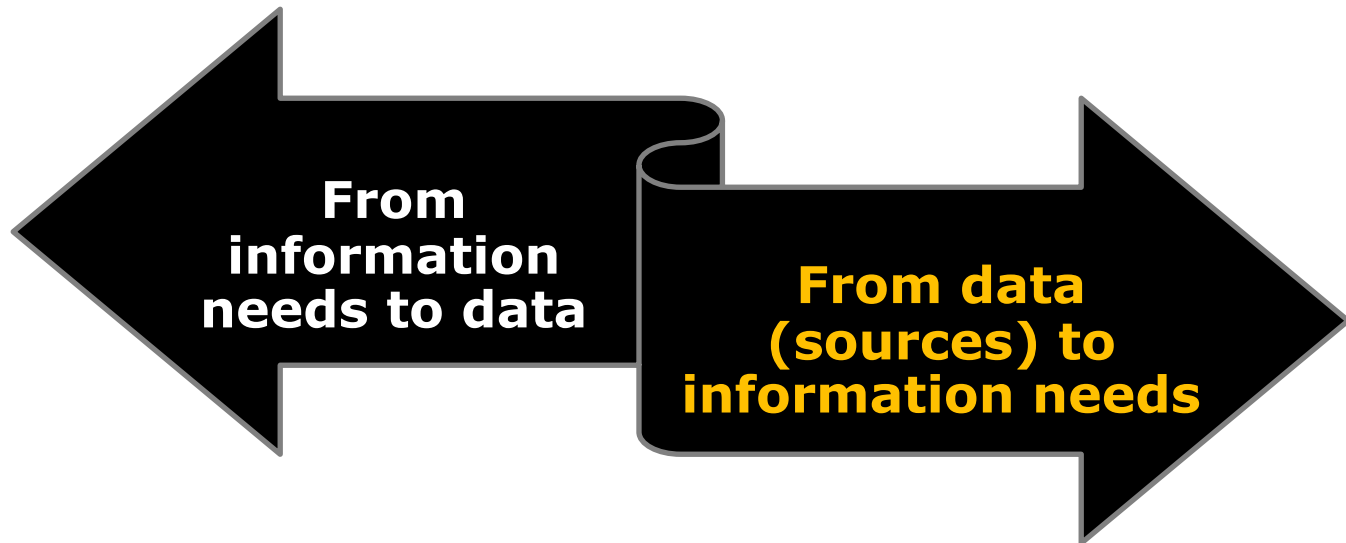
- Registration of events and behaviors
- Increase of the available information in a timeliness and cheap way
- Reduction of response burden
- Availability of more detailed data (granulaty)

Disadvantages

- Target population is different from big data population
- Definition and classification are not the same of OS
- Difficulties in the access to data
- Need of evaluation of data quality

Big data: the real challenges

Change of paradigm

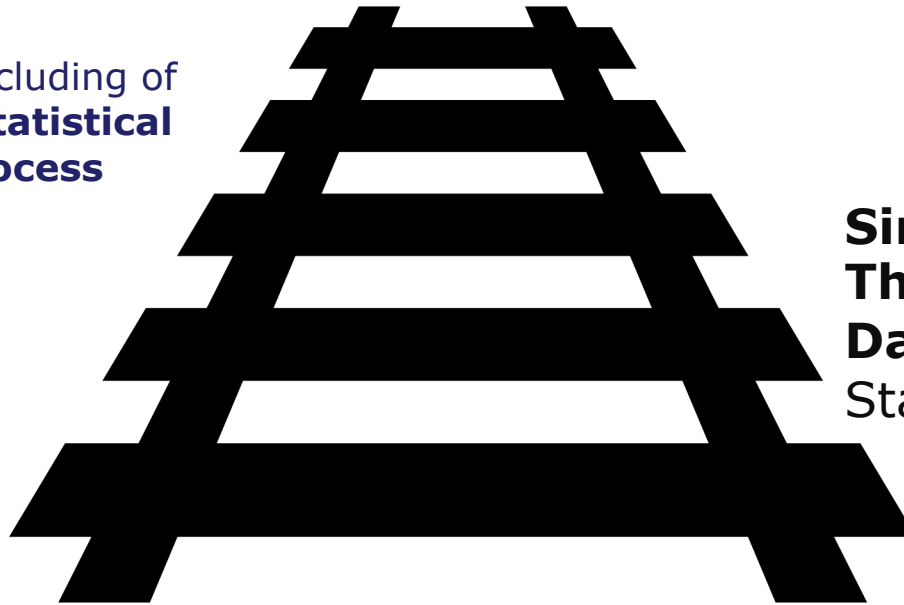


(HOW) TO ADD STATISTICAL VALUE TO BIG DATA?

Trusted smart statistics

The short way towards Trust smart statistics

Since 2010 - Including of
**Adm data in statistical
production process**
(i.e. census)



Since 2014 -
**Thinking about «Big
Data for Official
Statistics»**

**Today - Reaching trust and smart
statistics, including new data sources
respect the data quality criteria for OS**

Trusted smart statistics

Official statistics measures life, and when life is changing [official statistics] changes as well [M. Kotzeva, 2018, speech of 13° CNS, <https://youtu.be/zeCRLizkKko>]

Trusted smart statistics is the contemporary step of Official statistics.

- *It involves **more data sources and data input** respect to traditional vision*
- *It implies necessarily **new expertises** for processing and engineering new data sources and new data input*
- *It impacts on the general process of statistical production, starting from **an active participation of external stakeholders and (big) private companies**, not just in terms of information needs, but also as active subjects of data model*
- *Official statistics has to maintain the quality in data **production including more actors and factors in its statistical production process, opening its quality to the world, ever in the respect of the mission to KNOWLEDGE OF THE WORD through data***

References

- Braaksma B., Zeelenberg K., Big Data in Official Statistics, Discussion Paper, January 2020
- Davies W. «How statistics lost their powers – and why we should fear what comes next » <https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy> (The Guardian, 19 January 2017)
- Radermacher W., Official Statistics 4.0: Verified Facts for People in the 21st Century, Springer International Publishing, 29 dic 2019
- Ricciato F., Wirthmann A., Giannakouris K., Reis F., Skalioti M., Trusted Smart Statistics: motivations and principles, Statistical Journal of IAOS , 35 (2019), 589-303
- The Economist 7th February 2020, America's census looks out of date in the age of big data, <https://www.economist.com/international/2020/01/20/americas-census-looks-out-of-thdate-in-the-age-of-big-data>
- United Nations, Principles and Recommendations for Population and Housing Censuses, 2014