BIG DATA COMMITTEE

ANNUAL REPORT 2017

Istat

The report is edited by A. Righi.

A. Aubert and C. Dell'Aquila collaborated to the editing, A. Liberatoscioli broadcasted the videos in Para graph 6.

The authors of the Chapters are:

1 - G. Alleva
2 - F. R. Fuxa Sadurny, A. Righi, M.R. Simeone
3 - M. Scannapieco, D. Summa, A. Virgillito, D. Zardetto
Video: G. Bianchi, F. Scalfati, D. Summa
4 - P. Righi
5.1 - F. Polidoro
5.2 - G. Barcaroli, G. Bianchi, A. Nurra
5.3 - M. Broccoli
5.4 - D. Zardetto
6 - A. Righi

ISBN 978-88-458-1962-9

LIST OF CONTENTS

EXECUTIVE SUMMARY

This Report presents the activities concerning the use of Big Data (hereinafter “BD”) aimed at supporting the production of official statistics carried out and discussed in the Big Data Committee during 2017. The BD Committee (established by Istat with Resolution N. 4/PRES of 26/01/2016) comprises national and international experts from academia (7), international institutes (3), private companies (5), the research sector (3), public bodies (5) and some Istat members. The Committee's tasks include monitoring the activities of the Istat Big Data projects and the proposal of new projects aimed at defining new indicators based on Big Data. Furthermore, it fosters knowledge and experience sharing while promoting and supporting the creation of new partnerships.

In fact, each of the projects currently underway in Istat is supervised by the BD Committee members who work through sub-working groups (Mobile Phone Data, Internet as a Data Source and Social Media; Images, Sensors and Process-mediated Data), while two other groups have a cross-cutting character dealing with Reference Architecture and Partnerships and Privacy.

Therefore, the Report addresses all the activities carried out by the Committee, both cross-cutting and methodological. Moreover, the first results of the Istat experimental activities conducted during the latest years and referred to several different Big Data sources are presented.

After being clarified in chapter 1, the Istat and BD Committee's point of view on the characteristics and priorities of the use of Big Data in official statistics production, chapter 2 analyzes the main legal issues arising from the use of Big Data in the specific con text of National Statistical Institutes (NSIs). Chapter 3 focuses on Istat’s Reference Architecture for Internet as a Data Source for official statistics and in chapter 4 discusses a framework toward the Big Data Quality and Methodology to be applied in official statistics.

The main results of the trial carried out by Istat are described in chapter 5 and concern the experience on the use of Scanner Data for the Consumer Price Index (paragraph 5.1), the use of Internet as a data source to make estimates on the web ordering, job advertising and presence on social media of firms in the context of the ICT use of enterprises (paragraph 5.2), the trial on the use of Open Street Map for accident investigation on the Road and Motorways networks (paragraph 5.3) and, finally, the construction of the Social mood on Economy Index by social media (paragraph 5.4).

Chapter 6 recalls the short video contributions of some eminent Italian and foreign scholars, members of the BD Committee, expressing their opinion on relevant issues related to the use of BD in the production of Smart statistics ranging from the future perspectives of computer science research and the relevance of semantics to the challenges in the analysis of unstructured textual data; focus has been as well devoted to the inferential aspects of estimates with Big Data, the opportunities supplied by BD for the measurement of Sustainable Development Goals (SDGs) and the importance of partnerships for fruitful public-private experiences.

CHAPTER 1
BIG DATA FOR STATISTICAL PRODUCTION

The demand for timely, more detailed, less burdensome and costly statistics, with wider coverage, is increasing. In our digital era, data are everywhere; new sources, i.e. mobile phones, social media interactions, electronic commercial transactions, sensor networks, smart meters, GPS tracking devices, or satellite images, produce new information at an incredible speed. Digital technologies offer new opportunities for data collection, processing, storage. Increased quantity of available data discloses new opportunities, in particular about the use of administrative archives and unstructured sources, and at the same time it challenges official statistics.

New sources, among which Big Data, are gaining momentum. Since 2013, when the "Scheveningen memorandum" first formalized the need to look at Big Data sources as new sources for official statistics, the NSIs within the European Statistical System and its partners have gradually become aware of the new opportunities and challenges to official statistics offered by Big Data.

The intensified use of Big Data by NSIs calls for a new approach to the statistical production process, consistent with the statistical modernization agenda pursued by the statistical community pursues at the global level (Istat has been implementing its own modernization programme since 2014). This new approach assigns a crucial role to the integration of traditional and new sources, with a special focus on digital technologies; it requires a paradigm shift in methodology as well, from traditional sample survey-based estimates to a system applicable to a multi-source environment. Within this framework, different strategies range from Big Data combined with other sources (survey and administrative) to Big Data replacing traditional sources altogether. In a near future, blended approaches are likely to be the favourite models to enrich official statistics, by drawing data from a range of sources and producing estimates with pre-defined and transparent levels of quality and uncertainty as an absolute imperative[1].

Ensuring the quality of statistics is fundamental to preserve NSIs' credibility and reputation. The production of official statistics is anchored to internationally agreed methodologies and quality frameworks and is based on principles of professional independence and trust. As a consequence, traditional data sources, methods and techniques remain important and keep playing a key role in generating good statistics. On the other hand, if NSIs want to produce trustworthy data from new sources, they must introduce methods, concepts and tools for quality evaluation specifically addressing multi-source environments. A Big Data quality framework includes "input quality" (data access), "throughput quality" (big data methodology and estimation methods, the repository of Big Data projects and statistical production processes) and "output quality" (applications). To meet this new need, Istat is committed to create a Big Data Quality framework that will contain new inferential approaches to ensure quality and transparency of applications and results. To be fully exploited, Big Data require important investments in money and time, and, more important, in competencies. Research and innovation are the pillars of Istat's Big Data strategy, and, in more general terms, of the whole Istat's modernisation strategy. Istat Innovation Lab, a brand new infrastructure created in

1 UNECE Report of the Global Working Group on Big Data for Official Statistics Statistical Commission Forty-seventh session 8-11 March 2016
https://unstats.un.org/unsd/statcom/47th-session/documents/2016-6-Big-data-for-official-statistics-E.pdf

2017 to develop research ideas, aims at selecting and implementing projects to innovate the production processes in such context.

Since 2014, investment on methodological and thematic research has been fundamental to Istat's modernisation agenda. Research is the only possible way for Istat to acquire advanced tools and carry out in-depth quantitative and qualitative investigations. Research discloses information gaps to fill, explores new sources and increases the relevance of statistical information. In this process, high level skills and a change-oriented culture are essential. For this reason, the value of the human capital of official statisticians, as well as their continuing training, must increase.

Official statisticians, data scientists and experts from the public and private sectors benefit to a significant extent from exchanging views and experiences of statistical production with a focus on Big Data process, analysis and visualisation methods. The "learning by doing" approach appears particularly suited for building the capacity to use Big Data for official statistics. Pilot projects help testing the various approaches: several projects on a variety of Big Data sources are on-going at Istat. This report presents the first results of these pilot projects, based on a great joint effort of our researchers, the fruitful collaboration with other colleagues from Eurostat and NSIs, expertise and advice of the members of our Big Data Committee. Sharing knowledge in this field is essential, and the Big Data Committee serves as an ideal forum for a positive interaction of researchers, academics, the private sector and other stakeholders.

Two years after the launch of the Road Map of Istat's Big Data Committee (see the text), some of the new sources have been approached and significant questions on methods and quality issues are beginning to find response. Istat has completed the framework of its IT architecture in the domain of data production with Big Data sources. Several projects, using different Big Data sources for the production of statistics are currently in progress. Some of those projects are in their early implementation stage, others are in their experimental phase, others are mature for entering regular data production. The pilots are producing their first relevant results. The first is the systematic introduction of scanner data in the Consumer Price Indexes: since February 2018, these data are incorporated in the provisional estimate of the Consumer Price Index for January 2018. It is easy to understand the potential of the new sources: new indices and greater detail. Other pilots show a convergence between estimates on enterprises' ICT use, based on Internet data, and those based on our traditional survey, furthermore, OpenStreetMap can contribute to improve investigation on road accidents. Progress has been made in the production of a *Social mood on economy* index, from Social media data.

Experimentation in Istat is still in progress. The results of the pilots will be made available soon on a webpage devoted to Experimental statistics: a new section of Istat's institutional website displaying the innovative activities which have not yet reached full maturity in terms of harmonisation, coverage or methodology. The new statistical products will be monitored in terms of continuity of sources, methodologies, perception and evaluation by specific users. Only those experimental statistics that will prove to be relevant and solid will be included in the official production.

Big Data stability over time is also a key issue, hence the strategic importance of partnership agreements.

Difficulties in accessing data are the main reason why partnerships with private data providers are crucial for official statistics. Internet and Social media providers or telecommunication companies are potential counterparts, but commercial secrets and database protection rights are the most frequent reasons for private data holders to deny or limit NSIs access to their data.

A number of strategies could help solving access issues related to private data, from ad hoc partnerships with individual providers at country level to new legislation at national or EU level. Based on the experiences made so far by Eurostat and the EC, key aspects of Big Data (confidentiality, access and partnerships) could best be addressed at the global level, due to their similarities in many countries/regions and the fact that most Big Data providers are multinational companies. It is important for the statistical community to stand as a single subject when seeking partnerships, especially with data providers who operate globally (i.e., the Eurostat partnership with Google), or to develop an adequate legal framework for data access for public interest.

Countries like France and the Netherlands have started a set of initiatives to regulate big data access, and others will follow. The French Digital Act, endorsed in 2016, regulates data access rights for statistical reasons (i.e. access to data held by private entities when data concern infrastructure, like in the case of electricity and gas distribution networks). To ensure the production of the Consumer Price Index, the French Digital Act (after consultation and following a decision by the Minister of Economy and Finance) provides for mandatory electronic data collection for a category of private data (i.e. scanner data) and administrative sanctions in case of denied access. Italy has taken a first step in the same direction: the Budget Law approved in December 2017 (Law no 205/2017) introduces free access to Smart meters data (provided by "Acquirente Unico") for the permanent census.

Public-private partnerships are beneficial to all parties involved. Companies appreciate the positive impact on their reputation deriving from the institutional relevance and the public interest represented by NSIs, apart from taking advantage of solid methodologies and agreed quality standards. NSIs will evaluate the most appropriate return to potential partners willing to collaborate or to share data (e.g.: inclusion in official surveys of additional questions that could be of interest for the partners, or tailored data elaborations, etc.). Something like that is being tested by Istat with mobile phone and other private data providers. Thanks to the agreement with WIND TRE Istat will experiment models and algorithms for defining sub-populations according to official statistics classifications and quality criteria. Negotiations with VODAFONE Italia are aimed at developing statistical methods and procedures for data production with mobile phone data. The partnership with the Association of Big Retail Chains (Associazione Distribuzione Moderna) and an agreement with Nielsen Italy for the supply of scanner data allow access to data covering the entire national territory (107 provinces, 2,100 outlets of 16 Big Retail chains).

New trustworthy ways to access and share data, with sound governance frameworks and a proper legal environment, must be identified to safeguard available information and citizens' privacy.

Privacy is a core issue in handling the new sources. Istat is compliant with the national legal framework and the warranties granted by law to protect the privacy of individuals within the Italian National Statistical Programme. The EU "Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and the free movement of such data" (the so called General Data Protection Regulation - GDPR) entered into force on May 2018, will impact significantly on data production. The new Regulation, promoting privacy and data protection compliance from the initial steps of the data production process and introducing the new notion of "Privacy by design", will protect citizens' privacy even more than in the past and help identifying shared strategies and common solutions among the NSIs.

National Statistical Institutes must meet the needs of an ever-changing society and new information demands, while maintaining the quality of their products and a guiding role for a proper and competent use of statistics. The exploitation of new and unstructured sources, like Big Data, must prompt them to engage in crucial methodological endeavours on data harmonisation in

concepts, definitions and classifications. Big Data are often complex and unorganised: building a metadata structure apt to identifying the appropriate analyses and to linking data with other sources is a very complex and sensitive task. It is essential, therefore, to support research on quality standards and quality evaluation methods. Being transparent to the users is also a core issue: users should have free access to data and be provided with the necessary support to distinguish between sound and biased, incomplete or incorrect information.

In a post-truth world, data reliability is more than ever key for increasing trust and confidence in official statistics.

CHAPTER 2
LEGAL ISSUES ARISING FROM THE USE OF BIG DATA BY NATIONAL STATISTICAL INSTITUTES

NSIs has started producing statistics using Big Data sources to answer to the increasing demand for more enhanced and timely statistics, but the Big Data used by the NSIs are not originally collected or generated for statistical purposes. Only through a direct access to these new data sources, substantial progress will be possible in terms of timeliness and accuracy of official statistics, while significantly lowering the existing burden on respondents.

In the absence of specific legal provisions for the use of Big Data in official statistics, some legal issues may arise when collecting, using and combining Big Data for statistical purposes. These issues mainly relate to data ownership, personal data protection and privacy, and reputation, depending on the specific source of data: it mainly refers to personal data and confidentiality in case of data from telecommunications or utilities, and to copyright and databases protection in case of the Internet data. The following paragraphs present an overview of the legal issues which may arise from the use of Big Data.

2.1 Review of the European legal framework

Personal data and Confidentiality

Currently, in the European Union, personal data are processed in compliance with national laws transposing Directive 95/46/EC, and since May 2018, with Regulation 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (the so-called 'GDPR regulation'). This Regulation significantly affects public administrations and enterprises activities, since it provides that when developing, designing, selecting and using applications, services and products based on personal data processing, producers should take into account the right to data protection. The principles of data protection should apply to any information concerning an identified or identifiable natural person. There is the explicit introduction of 'pseudonymisation' of personal data, that is the need to keep separate the data and its ID, because the use of additional information could allow to re-identify a person. Consequently, to determine whether a natural person is identifiable, it should be taken account of all the means reasonably likely to be used to directly or indirectly identify the natural person by the controller or by others. To ascertain whether means are reasonably likely to be used to identify the natural person, all the objective factors (such as costs and amount of time required for identification with the available technology at the time of the processing) should be considered.

Moreover, the processing of personal data for purposes other than those for which the personal data were initially collected (and that is the case of the use of Big Data for statistical purposes) should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected. If the processing is needed for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller, the EU or Member State law may determine and specify the tasks and purposes for which the further processing should be regarded as compatible and lawful. Further processing for archiving purposes in the public

interest, scientific or historical research purposes or statistical purposes should be considered to be compatible and lawful. Personal data are processed by national statistical institutes and/or other national authorities in the public interest for the statistical purposes falling within the scope, and are kept in a form allowing identification of data subjects for no longer than is necessary for the sole purpose of creating Union statistics.

Derogating from the prohibition on processing special categories of personal data should also be allowed for statistical purposes. Such derogations are provided under Article 89 of the GDPR; in particular, paragraph 2 outlines the conditions under which Union or Member State's law may derogate from certain provisions.

Copyright and Databases protection

Information scraped from the Internet sources may potentially be protected by copyright, thus, to what extent NSIs can rely on exceptions in copyright legislation to use protected data for scientific/statistical purposes?

Copyright is a legal right that grants the creator of an original work exclusive rights for its use and distribution. Copyrights are considered territorial rights, which means that they do not extend beyond the territory of a specific jurisdiction. While many aspects of national copyright laws have been standardized through international copyright agreements, copyright laws vary by country.

The analysis of the legislation on copyright protection of the EU has identified several Directives[2]; furthermore, in 2014, a European Commission report entitled *'Analysis of methodologies for using the Internet for the collection of information society and other statistics'* provided a legal opinion on the legal feasibility of web scraping. The main issue focuses on the 'sui generis' database right set out in Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. This database right is considered as property right and it suggests that the act of creating a database from web scraped data could breach database rights, at least if essential parts of the database are scraped.

Nevertheless, a general reflection on the meaning of copyright for official statistics could be synthetized by underlining that there are no major problems with the copyright protection. Law No. 633 of April 22, 1941, for the Protection of Copyright, provides for an exception in case of use of the database for non-profit scientific research purposes (Art 64-sexies of law no. 633/1941 as amended). Though statistical research is not explicitly mentioned, it appears reasonable to argue that NSIs can access and use information from online databases to process this information into anonymized statistical results, even when the 'sui generis' database right protects them. Moreover, according to the statistical law, NSIs collect data for public interest purposes basically.

2 Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1479128659986&uri=CELEX:31993L0098
Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1479119938264&uri=CELEX:31996L0009
Directive 2009/24/EC of the European parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version)
http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1479309002824&uri=CELEX:32009L0024
Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2001.167.01.0010.01.ENG&toc=OJ:L:2001:167:TOC
Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights
http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1479128659986&uri=CELEX:32006L0116

Anyway, even if a possible web scraping activity made by NSIs is not forbidden, it is necessary to pay attention and analyse all the circumstances in which it happens, the type of processed data, the considered sources and the type of dissemination.

Thus, open issues in the legal framework still remain. This is the reason why recently the ESS[3] acknowledged that clearer rules for statistical offices to access privately-held data of general interest held are needed to open up new data sources and create a thriving environment for new statistical products and services.

NSIs face major difficulties in accessing new data sources: further, the absence of a specific legal basis, as well as the lack of clarity of the conditions under which access to privately-held data can be granted for the fulfilment of public service missions, reduce the possibility to move from limited experimentation to the systematic use of new data sources leading to a new generation of trusted official statistics.

There is uncertainty in relation to data ownership issues, such as property rights, protection of the 'sui generis' right for databases or trade secrets. The protection of these rights has been often advanced as an argument to refuse access to data to statistical offices, although it is possible to overcome these obstacles through constructive private partnerships. But some data holders fear possible negative impacts in terms of reputation, particularly when the data at stake are personal data.

An initiative in this field could come, at European level, in next months via a proposal of a directive or an *ad hoc* regulation (maybe in the context of the revision of the Public Sector Information (PSI) Directive through a reverse PSI approach) to be submitted to the EU Council and the European Parliament. If this were the case, statistical authorities could act as data hubs providing factual, timely and appropriate information to a wide variety of users at the service of the society as a whole. This would finally result in the opening up of new data sources for a new generation of official statistics in light of the growing European Digital Single Market (according to the proposed EU Directive).

2.2 Review of the Italian situation

There is no specific legislation in our legal order that protects and regulates the treatment of Big Data. In the absence, the general rules are applicable (see the box below).

Applicable general rules

The legislative framework for the use of data for official statistics in Italy is mainly based on the following legislative acts:

- National Statistical System (NSS) law (legislative decree no. 322, 6 Sept 1989);
- Personal data protection law (legislative decree no. 196, 30 June 2003, hereinafter referred to as Privacy Code);
- Code of Ethics and good practices for the National Statistical System (Annex A3 to the Privacy Code, *Measure n. 13 of 31 July 2002 of the* Personal Data Protection Authority, G.U. 1 Oct. 2002, n. 230, hereinafter referred to as Ethics Code);
- Database protection law (legislative decree no. 169 6 May1999 Implementation of the Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases).
- Regulation 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, which repeals Directive 95/46/CE.

3 See: ESS Vision Implementation Group (2017) Position paper on access to privately held data which are of public interest, Presented at 15th Vision Implementation Group meeting, 26 October 2017, Warszawa

When accessing, filing and processing Big Data from public and private websites, the National Institute of Statistics must comply with the following regulatory requirements:

- Istat and the other national statistical system authorities can use personal data for their institutional tasks without prejudice to compliance with the requirement that data should be relevant and not excessive.
- Personal data can be collected only for 'defined aims, which have been made explicit and legitimate, and used for other tasks that are compliant with such aims' (Privacy Code art. 11).
- 'Handling personal data for historical, statistical or scientific aims is considered as complaint with aims according to which data have been collected or handled previously' (Privacy Code art. 99). Thus, it is always possible to use for statistical purposes personal data that third parties have collected for whatever purpose.
- Istat has the obligation to inform private and public subjects, which hold or disseminate data through the various multimedia channels, about its scraping activity and the specific uses of the processed data.
- According to art. 13 of Legislative Decree no 322/1989 ("National statistical program"- hereafter PSN), all statistics of public interest in the NSS, and their related purposes, must be laid down in the PSN. This is prepared by Istat and approved by decree of the President of the Republic, after a process involving the Presidency of the Council of Ministers and the Interministerial Committee for Economic Planning (CIPE). The PSN identify all the variables that can be disseminated to meet specific information needs, at European and international level.

The most significant provisions related to the statistical treatment of Big Data are:

Privacy

- **"Processing of personal, sensitive and judicial data, within the National Statistical Program (NSP)"** - Article 4-bis Ethics Code. The NSP also reports these data (referred to art. 4, paragraph 1, letter d and e of legislative decree no 196/2003) and the surveys in which these are treated, as well as the methodologies used. The NSP is adopted, with reference to personal, sensitive and judicial data, following the approval of the Personal Data Protection Authority, according to art. 154 of the Code (Legislative Decree No. 196/2003).
- **"Electronic communications"** - The scope of the provisions in Section 123 and 126 of the Code (legislative decree no 196/2003) is restricted to electronic communications providers only. Information society service providers (such as Google, Facebook, Twitter, etc.) should apply the general provisions of the Code when processing traffic or location data. Changes to the current legal framework are expected from the e-Privacy Regulation, with a broader involvement of the aforementioned providers.
- **"Notice"** - Art. 6 of Ethics Code states that: "if the processing concerns personal data that have not been collected from the data subject and informing the latter entails a disproportionate effort compared with the right to be safeguarded – as per Section 10(4) of the Act –, the information shall be considered to have been notified if the processing is included in the National Statistical Program or else is publicized by suitable means; the latter shall have to be communicated in advance to the Italian Data Protection Authority, which may provide for specific measures and arrangements".
- **"Conservation and security measures"**- Artt. 11 and 12 of the Ethics Code refer to data conservation and security measures, which must take into account the type of treatment and the specific precautions to be adopted to prevent loss of data.

- **"Exercise of the data subject's rights"**. Art. 13 of the Ethics Code provides for a derogation from the data access, correction, updating and integration rights if such operations prove impossible because of the nature or status of the processing, or involve a clearly disproportionate effort, or where these operations do not produce significant effects either on statistical analysis or on the statistical results related to the processing. In particular, no changes shall be made if they are in conflict with statistical classifications and methodology as adopted in pursuance of international, Community and national regulations.
- **Regulation UE no 679/2016** - With this new European Regulation not only the providers of telecommunications and electronic communications services, but also public administrations, companies, public and private health structures, banking institutions, have the obligation to communicate to the Data Protection Authority the loss or destruction of personal data that may occur as a result of cyber-attacks, unauthorized access, accidents. Failure to notify involves financial penalties.

Copyright and database protection

Art. 64-sexies, Law no 633/1941 "Databases" - According to this provision, protection of copyright and related rights are not subject to the authorization by the holder of the right (referred to Art. 64-quinquies) in case of access or consultation of the database exclusively for teaching and research purposes, provided that these activities are not carried out within an enterprise, if the source is indicated and within the limits of what is justified by the non-commercial purpose pursued. Statistical research is not mentioned, but it appears reasonable to argue that Istat can use information from online databases, even when the *sui generis* database right protects them, to process this information into anonymised statistical results. Such interpretation seems confirmed by Sections 98 and 99 of the Code, according to which the processing by Istat is carried out for purposes of "substantial public interest" and in any case, compatible for whatever purpose the personal data were originally collected.

In conclusion, Big Data can be processed for statistical purposes when the non-identifiability of the data subject is assured, or the data used are not attributable to the definition of personal data. When the statistical treatment involves the processing of personal data and this processing is enclosed in the National Statistical Program, all the above mentioned legislation is applicable in the use of Big Data. There are no particular obstacles for Istat to scrape data from publicly available Internet sources or to ask the Internet sources or Internet users to provide data. In most cases, the scraping of the Internet sources would result in the processing of personal data. Hence, each study in the NSP requiring personal data treatment must specify the kind of personal data which are likely to be collected, and must be approved by the Committee for Safeguarding Statistical Information and the Italian Data Protection Authority. The processing of personal data must also respect the provisions of the Code, Annex A3 and the Statistical Legislation.

CHAPTER 3
ISTAT'S REFERENCE ARCHITECTURE FOR INTERNET AS A DATA SOURCE FOR OFFICIAL STATISTICS

3.1 Objective

Modern organizations have recognized the importance of having a defined and standard architecture able to guide the implementation of the organization's vision and to harness external drivers and changes. Since several years International and European National Statistical Institutes (NSIs) have been investing on standardization projects, which have produced several artefacts as constituting pieces of a "reference architecture" for their business.

However, when considering a reference architecture for Big Data in Official Statistics, there are few concrete results in terms of available standards, possibly due to the recent attention that the Big Data phenomenon received by NSIs. Indeed, only in 2013 the first official document that formalized the need for NSIs to look at Big Data sources as new sources for Official Statistics was published. The document is known as "Scheveningen memorandum" and it acknowledges that Big Data offer new opportunities and challenges for official statistics, for the European Statistical System and its partners.

This memorandum intends to lay down the bases for an Istat's reference architecture for Big Data. Based on existing standards, the memorandum, in particular, makes explicit reference to the Generic Statistical Business Process Model (GSBPM). In addition, it reports some findings achieved within the European project ESSnet on Big Data Pilots-I, and hence already discussed and shared with other European NSIs. In this respect, even if it is an Istat's proposal, it could be considered as a starting point for European and International work on standardizing Big Data architectures for Official Statistics. The scope of the reference architecture described below is on Internet Data, but it will be extended to further Big Data sources.

3.2 Results achieved

3.2.1 GSBPM "fitting" to Big Data

GSBPM is the standard adopted for describing production activities performed by NSIs. The principal phases of GSBPM are shown in Figure 1.

Figure 1 Principal phases of GSBPM

The fitting of GSBPM to Big Data processing requires the following changes:

- given that the GSBPM phases have a time sequence, *the two phases Design and Collect need to be swapped*. Indeed, only after the collection of a Big Data source, an actual design phase can start, aiming at setting up the statistical framework necessary to process the source.
- the Design phase of GSBPM, however, has a different meaning in the case of Big Data sources. For these sources, it is better to consider a *Preparation phase,* where, rather than doing traditional design choices (e.g. sampling design), a set of preparation activities are carried out on the Big Data source. Since the source was generated for different purposes from the official statistical analysis, it needs to be "prepared".

In summary the adaptation of GSBPM to Big Data processing will look like in Figure 2, where the above changes are highlighted.

Figure 2 Main phases of GSBPM adaptation to Big Data

3.2.2 Logical Architectural Schemes for Internet as a Data Source

Internet as a Data Source (IADS) makes reference to the set of Internet-accessible sources that can be used for collecting data considered as relevant for official statistical purposes. Among such sources, we considered: (i) Web sites and (ii) social media, specifically Twitter.

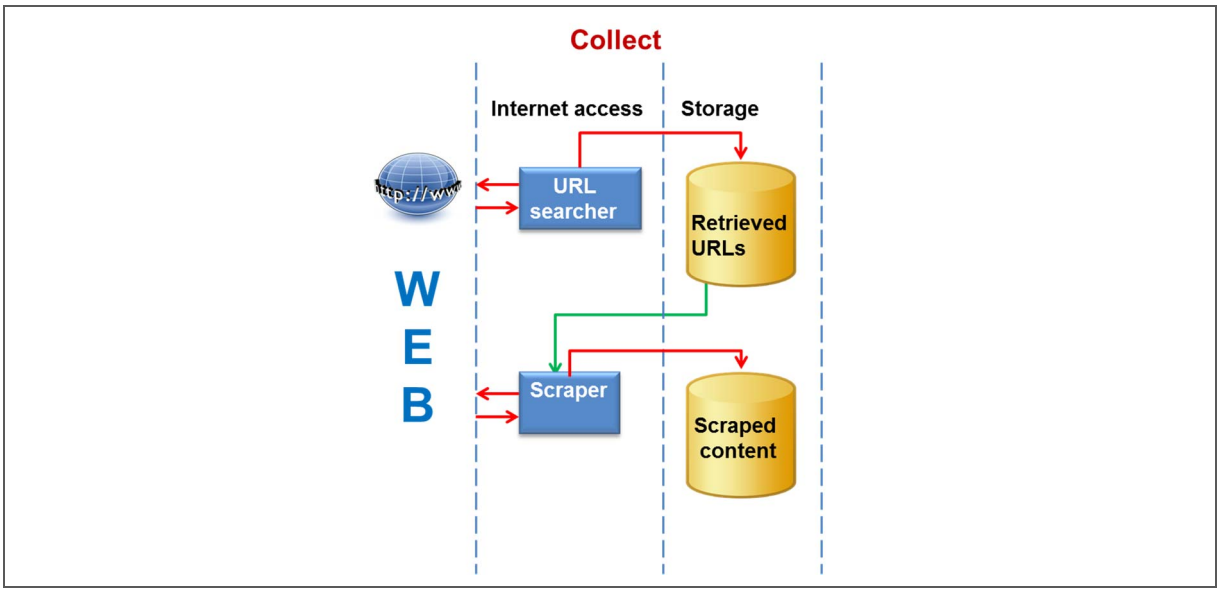
For these two source types, we will detail the Logical Architectures of the Collect, Prepare and Process phases, on which we have already had concrete pilot projects. In addition, we will describe IT Architectural Schemes as implementation examples.

3.2.3 Web sites

For Web sites, the **Collect Phase** is represented in Figure 3 and includes two sub-processes, namely: **Internet access** and **Storage**. Each sub-process has some logical blocks. The two ones for Internet access are **URL searcher** and **Scraper.** The ones considered for Storage are **Retrieved URLs** and **Scraped content**. In order to start a collection of data by scraping Web sites, firstly a list of URLs identifying the home pages of the sites to reach is needed. If this is not available, it is possible to set up a dedicated URL retrieving activity, which is what we did in an experimental pilot [4].

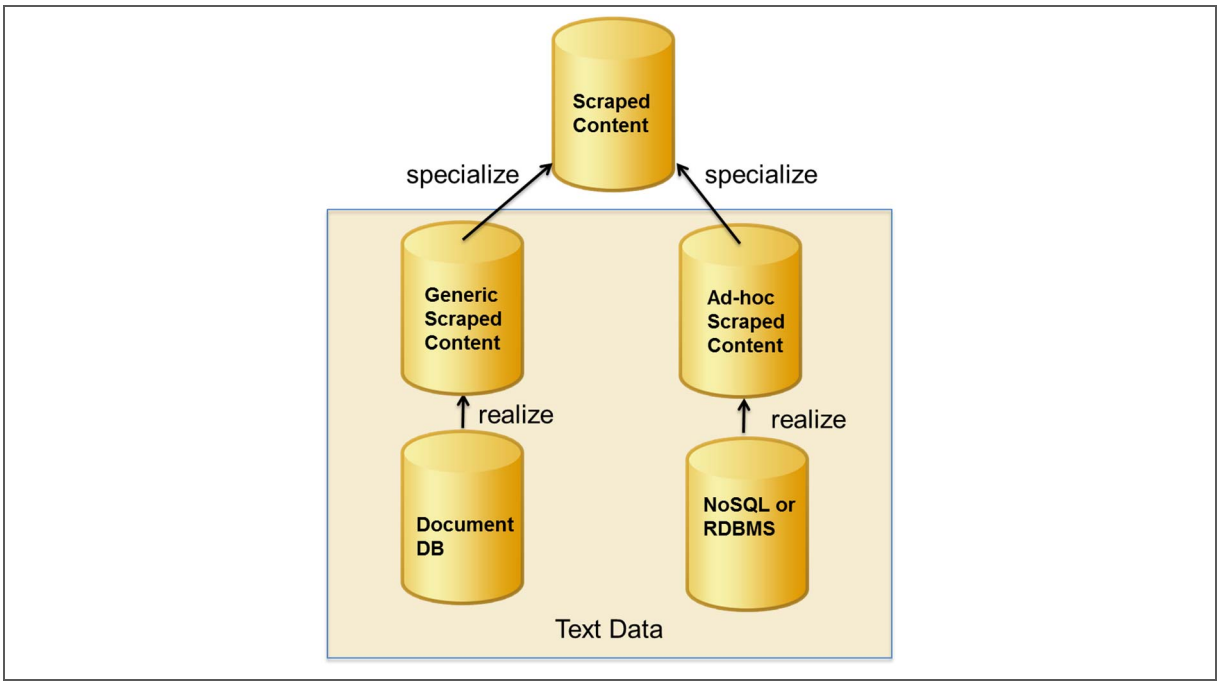
4 Barcaroli G., Scannapieco M., Summa, D. (2016). On the Use of Internet as a Data Source for Official Statistics: a Strategy for Identifying Enterprises on the Web. *RIEDS - Rivista Italiana di Economia, Demografia e Statistica - Italian Review of Economics, Demography and Statistics*, 70 (4).

Figure 3 Collect phase for Web sites



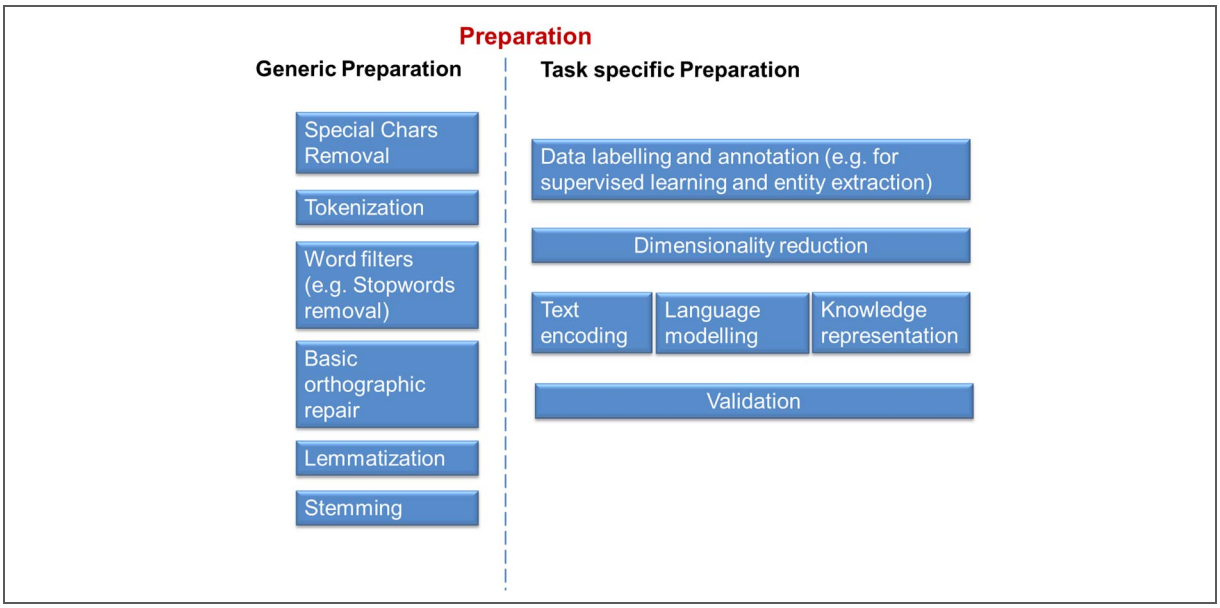
The Scraped content block can be usefully refined as described in Figure 4. Please notice that some relevant "realize" arrows permit to detail categories of technological storage solutions identified as more appropriate. So far, we focused on scraping the textual content of Web sites; an interesting extension will be the access to images that can be present on the sites.

Figure 4 Technological storage solutions for scraped content



As shown in Figure 5, the **Preparation** phase has two specific sub-processes, namely a **Generic preparation** and a **Task specific preparation**.

Figure 5 Preparation phase

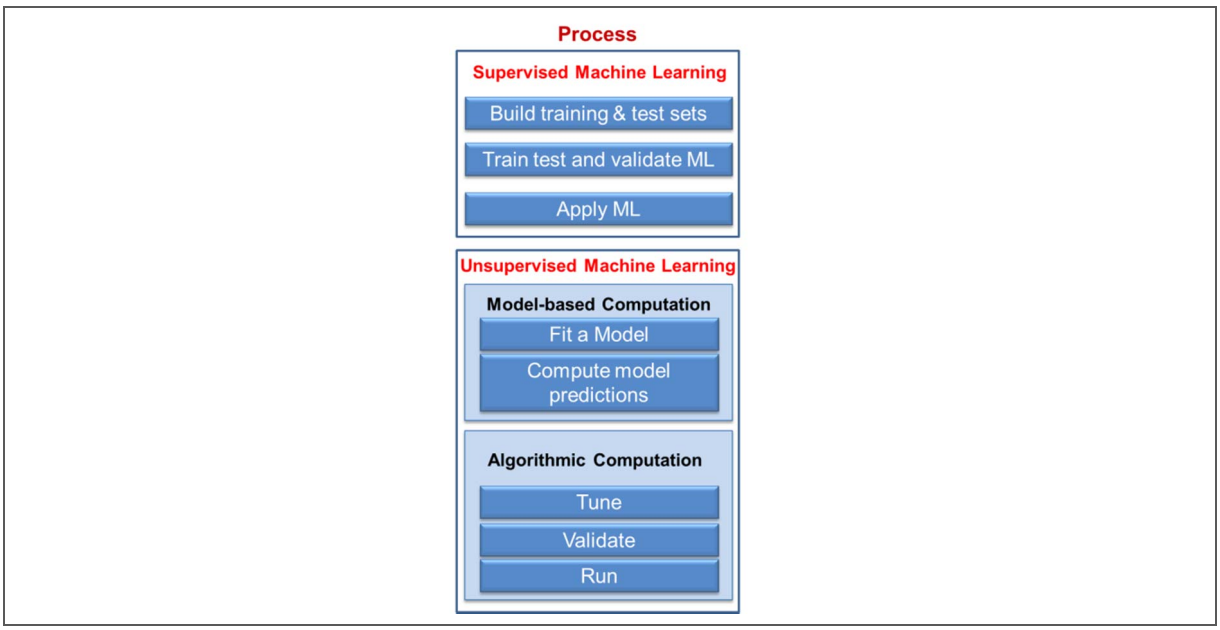


The introduction of a task-specific preparation phase is particularly interesting. As shown in Figure 5, depending on the specific processing of the text, the preparation can include a **Text Encoding** logical block, e.g. the traditional Bag of Words approach, a **Language modelling** block, e.g. recent Word Embeddings approaches or a **Knowledge Representation** block, e.g. modelling domains through formal languages for ontologies.

Finally, the Process Phase for Web sites is based on the use of a Machine Learning approach, in which either **Supervised** or **Unsupervised Machine Learning** logical blocks can be considered (see Figure 6).

For the **Unsupervised Machine Learning** it is possible to choose among logical blocks that are: Model-based Computation (Fit a model and Compute model predictions), e.g. for classification via mixture models, or Algorithmic Computation(Tune, Validate and Run), e.g. for clustering or automated reasoning.

Figure 6 Process phase



3.2.3.1 Twitter data

For Twitter data, as in the case of Web sites, the **Collect** phase has the **Internet Access** and the **Storage** sub-processes (see Figure 7). However, in this case, the Internet Access logical blocks are very much specific of Twitter APIs; these are: Rest API, Streaming API, Request Handler, for issuing requests and getting results, and Streaming Processor.

In Figure 8, the logical blocks detailing the Storage sub-process are shown. The main distinction is between the need for storing Tweets as plain text and the need to keep some structure for Tweets. Both kinds of storage are possible requirements for official statistical needs: the former could better serve tasks based on language encoding, like quality evaluation of the filters used to access Twitter data; the latter, instead, could permit queries over Tweets that are more detailed, e.g. the retrieval of geo-localized Tweets.

Figure 7 Collect phase for Twitter data

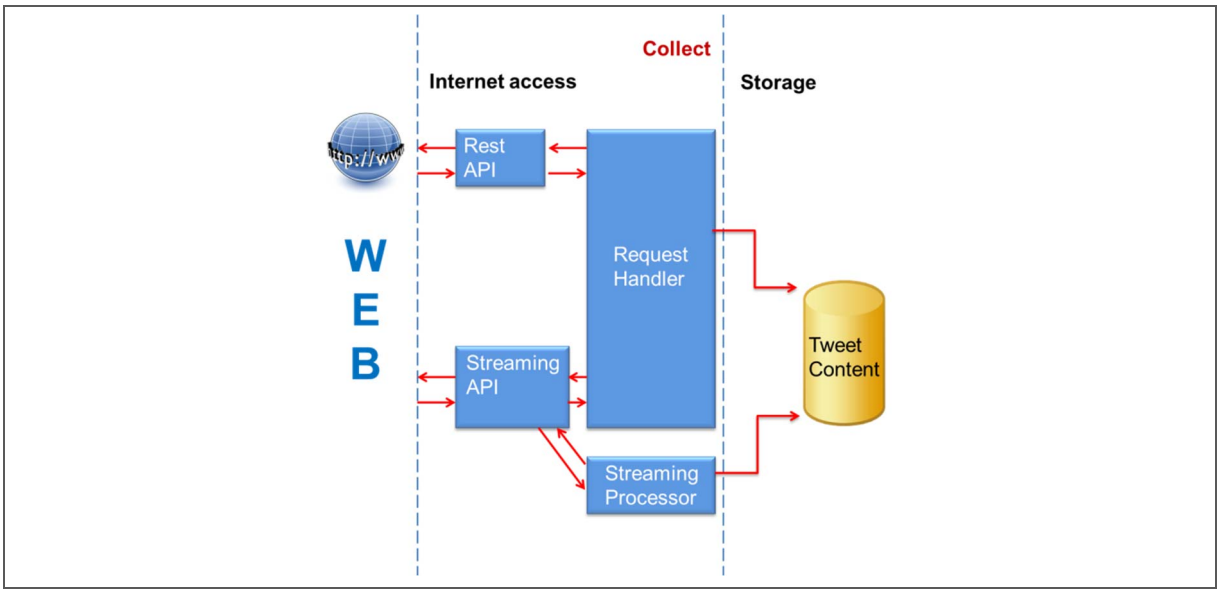
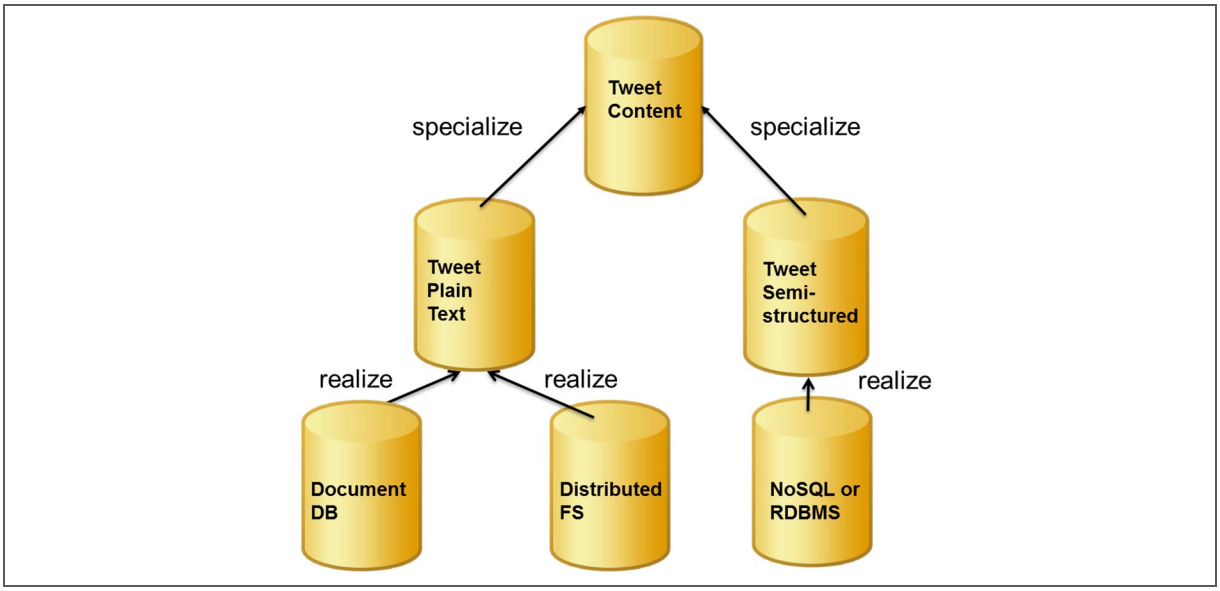


Figure 8 Technological storage solutions for Twitter data



3.2.4 IT Architectural Schemes for Internet as a Data Source

Each logical building block in the architecture presented above can correspond to different concrete implementations. An IT architectural scheme is a specific example of architecture, characterized by the choice of specific tools to implement a given building block.

Below we present an example of technical architectural schemes for the two types of Internet data, web sites and Twitter. For each phase, a set of tools is suggested.

3.2.4.1 Web sites

Collect: For URL Searcher and Scraper logical blocks, we adopted ad-hoc developed Java solutions. Similar Python-based solutions (like *Scrapy*) could be used.

A document DB like Solr or Elasticsearch is recommended. These tools allow to store and efficiently retrieve the generic scraped content from web sites, by indexing all the documents and enabling access to documents through user queries (e.g. via pattern matching). For ad-hoc web scraping, it is possible to have a relational storage or, when scaling up is a requirement, a NoSQL database; we have HBase as a solution for this latter case.

Preparation: Java or Python libraries for supporting both generic and task specific preparation. One example is the *NLTK* Python library. For task specific preparation, in the case of language modelling, the word embeddings frameworks like GloVE or Word2vec can be used. The knowledge representation logical block can be implemented via Semantic Web languages for ontology representation, in particular Web Ontology Language - OWL.

Process: R or Python libraries implementing supervised or unsupervised machine learning algorithms. Examples are the *caret* package for R and the *scikit-learn* library for Python.

3.2.4.2 Twitter data

Collect: Java, R or Python libraries for querying the Twitter API. Several options are available, like the *TwitteR* package for R or the *Tweepy* library for Python. In this case too, Solr or Elasticsearch are suited to store massive amounts of Tweets that can be efficiently retrieved through keyword filtering or by querying specific attributes of the data (for example, all tweets where language is Italian). The massive storage can be done on a NoSQL like HBase.

3.3 Lessons learnt

The logical and technological architectural schemes presented in this contribution are the result of several pilot projects that we have been running in Istat in the last years. We learnt from our practical experiences that the shown schemes are good enough for the purpose of our business. Hence, from now on we propose them as a “reference architecture” to be adopted in projects that use the Internet as a Data Source for statistical production.

The importance of a reference architecture like the proposed one is twofold:

- concrete support to technical standardization, with NSIs’ benefits ranging from avoiding vendor lock-in, to gain awareness in asset governance and capability planning.
- sharing solutions with other NSIs or organizations, thus fostering reuse and consequently gaining cost reduction and quality standards.

The video shows how the Reference Architecture has been implemented in the project described in section 5.2 on ICT use in enterprises.

CHAPTER 4
TOWARD A BIG DATA QUALITY FRAMEWORK AND METHODOLOGY IN OFFICIAL STATISTICS

4.1 Objective

Big Data (BD) expand the range of sources that have the potential to be used for official statistics. One of the challenges for National Statistical Institutes (NSIs) is assessing the quality of such data sources, and the quality of statistics produced from them.

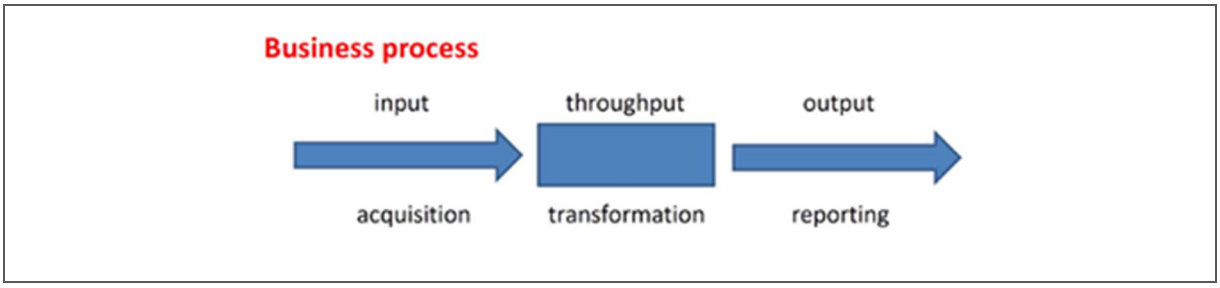
Existing frameworks generally assume NSIs have the high degree of control of the statistical process. That makes these quality frameworks unsuitable for the BD, given the external nature of these data sources. BD have not statistical purposes, they are not generated by a planned process but they simply exist. Quality issues become more prominent than statistical products generated entirely by an in-house process and the quality framework must careful consider all the phases of the business process.

Istat undertakes to define an appropriate Big Data Quality Framework (BDQF) in compliance with international standards. Quality framework is still a work in progress and Istat collaborates in international activities, such as the European Statistical System Network project, European Task Force on Big Data and the United Nation Working Group on Big Data.

4.2 Toward the Quality Framework

To describe the path for defining a BDQF, let start by introducing the proposal of UNECE 2014[5] Quality framework on Big Data. It involves three phases of the business process (figure 1): *Input* – when the BD are acquired, or in the process of being acquired (collect stage); *Throughput* – any point in the business process in which data are transformed, analysed or manipulated. *Output* – the assessment and reporting of quality with statistical outputs, it is the information about the quality of the statistical product.

Figure 1 Business process phases considered in the UNECE quality framework on Big Data 2014



[5] UNECE Big Data Quality Task Team (2014). A Suggested Framework for the Quality of Big Data. *Deliverables of the UNECE Big Data Quality Task Team*. December 2014.

***Input phase*:** data obtained from an external organization. Transparency is needed about the nature of the acquisition (UNECE, 2014). The following quality dimensions should be taken into account:

- Institutional/Business Environment (source): Sustainability of the entity-data provider, reliability status, transparency and interpretability;
- Privacy and Security (source): legislation, data keeper vs. data provider restrictions, perception;
- Complexity (metadata): technical constraints, whether structured or unstructured, readability, presence of hierarchies and nesting;
- Completeness (metadata): whether the metadata are available, interpretable and complete
- Usability (metadata): resources required to import and analysed, risk analysis;
- Time-related factors (metadata): timeliness, periodicity, changes through time;
- Linkability (metadata): presence and quality of linking variables, linking level;
- Coherence – consistency (metadata): standardization, metadata available for key variables (classification variables, construct being measured);
- Validity (metadata): transparency of methods and processes, soundness of methods and processes;
- Accuracy and selectivity (Data): total survey error approach, reference dataset, selectivity.

***Throughput phase*:** it refers to all the intermediate stages between acquisition of the data and dissemination. General principles for the quality of data in the throughput stage:

- Transformations and analysis should proceed according to theoretical principles and not be dependent on the system that is processing them;
- The data be processed through a series of stable versions that can be referenced by future process and by multiple parts of the organization;
- Definition in the business process of a checkpoint at which the quality of the data is explicitly assessed. Important features of the quality checkpoint are that the measures used to assess quality are decided in advance, and the reference time of the checkpoint as well.

***Output phase*:** It is applicable to reporting, dissemination and transparency. It is the information about the quality of the statistical product. Quality framework must be assured that the output:

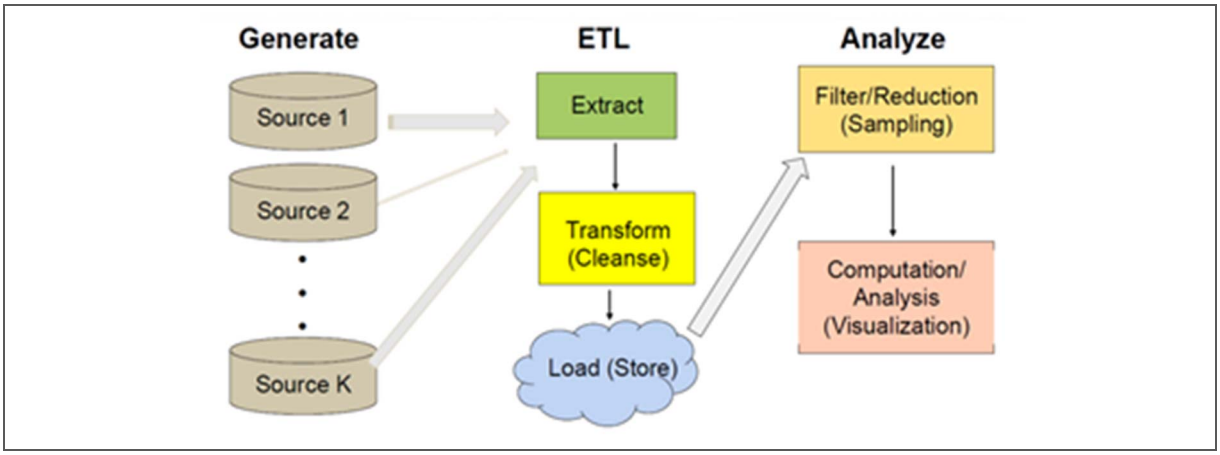
- Meet reporting criteria of the NSI;
- Provide sufficient information for the user of the data to make an informed decision regarding the output;
- Follow transparency principle around analysis and methods;
- Follow the general approach with quality dimensions, indicators and factors to consider.

In particular, NSIs are quite interested in the accuracy of the statistical output especially if the aim is to implement quantitative analysis. Accuracy is usually characterized and decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (*e.g.*, coverage, sampling, nonresponse, presence of outliers, etc.). This is the aim of the Total Survey Error approach (TSE; Groves and Lyberg, 2010[6]). TSE approach is desirable when analysing the accuracy of a potential dataset in regard to statistical analysis (UNECE 2014). The errors involved in the TSE are generally common in traditional surveys as well as in surveys using administrative data and BD. Nevertheless, for the latter, others

[6] Groves, R. M., Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly, 74,* 5, pp. 849-879.

types of error may occur since the process chain should be longer and more complex than traditional ones. The American Association for Public Opinion Research (AAPOR) Task Force on Big Data introduces the Big Data Total Error (BDTE) framework (AAPOR, 2015[7]) as an extension of the TSE and states "[TSE] … is quite limited because it makes no attempt to describe the error in the processes that generated the data. In some cases, these processes constitute a "black box" and the best approach is to attempt to evaluate the quality of the end product". AAPOR identifies the errors according to a data generating-process based on three stages: Generate, ETL and Analyse stages (figure 2).

Figure 2 Data generating-process by AAPOR (2015)



In the data generating-process, the Generate stage collapses in the input phase of UNECE proposal, while the ETL and Analyse stages collapse in the throughput phase (table 1).

Table 1 Examples of errors in the data generating process

Generate stage (Input phase)	ETL stage (part of the throughput phase)	Analyze stage (part of the throughput phase)
• Data missing or erroneous: ○ errors for environmental or technology reasons (low signal/noise ratio, lost signals) ○ Data missing for voluntary reasons • Selected data units: ○ Coverage/ sub-population, selective sources • Metadata lacking, absent, or erroneous.	• Data missing for transferring data from input to throughput phase reasons • Data Missing or erroneous for transformation process: ○ specification error (including errors in meta-data), matching error, coding error, editing error, data-munging error, data-integration error	• Errors for further transformations of the data to form more manageable or representative data sets: ○ sampling errors, selectivity errors and modeling errors ○ errors generated by Map/Reduce process ○ errors for auxiliary data missing when applying models ○ weighting errors ○ algorithmic errors

[7] AAPOR (2015). Big Data in Survey Research. AAPOR Task Force Report. *Public Opinion Quarterly, 79*, pp. 839–880.

Many quality dimensions, related to the input and the throughput in the UNECE BDQF or errors in BDTE, directly affect the accuracy of the output. Selectivity has an important impact and it cuts across the phases and stages. Selectivity according to Buelens *et al.* (2014[8]) is defined as follows: "A subset of a finite population is said to be representative of that population with respect to some variable, if the distribution of that variable within the subset is the same as in the population. A subset that is not representative is referred to as selective."

Selectivity is related to specific aspects (variables), it could be the case that a set of data that is highly selective for a subset of variables but not for all.

In the traditional survey or administrative data quality framework, the selectivity can be detected by observing the coverage of the source with respect to the target population, i.e., the *under-coverage* (a population element is missing in the frame of the source) or the *over-coverage* (the frame of the source entails duplicated, non-existent or out-of-scope elements). It may be complicated to identify the coverage because in many cases BD refer to units only indirectly related to the statistical units of interest or because of linking variables are missing for instance due to privacy issue. The lack of direct connection (linkability dimension) between the target population and the available population coming from the BD orients the choice of the methodology to be applied.

4.3 Toward the Methodology Framework

Let us define as methodology the statistical method and the inferential approach to analyze the BD. To produce high quality statistics, methodology must be planned to deal with the input and throughput quality concerns. As an example, let us consider the selectivity. It is one of the most important methodological barriers to use Big Data in official statistics (Zeelenberg, 2016[9]) because it introduces bias. Representative subsets allow unbiased inference about population quantities simply by using distributional characteristics of a variable within the subset (simple methodology), but if representativeness is not satisfied, statistical methods must be used (complex methodology) to properly treat the selectivity.

Herein below, three general classes of approaches are presented (figure 3).

Pseudo design based (Elliott and Valliant, 2017[10]): it estimates the probability of unit *k* to be in the BD source given *k* belongs to the target population *U* (inclusion probability), and it applies (pseudo) **design based approach.** Inference holds if: *i*) the model generating the inclusion probability in BD is known; *ii*) over-coverage is correctly identified. Note the pseudo-design based approach rarely is applied because of the complexity of probability estimation.

Model based (Valliant *et al.*, 2000[11]): it predicts the target *y* variable of not covered population and applies model based approach. Inference holds if: *i*) the model generating the value variable in the BD is the same in BD and in all the *U* population. The model is known; *ii*) over-coverage is correctly identified

[8] Buelens B., Daas P., Burger J., Puts M., van den Brakel J. (2014). Selectivity of Big data. *Discussion Paper nr. 11*, Statistics Netherlands.

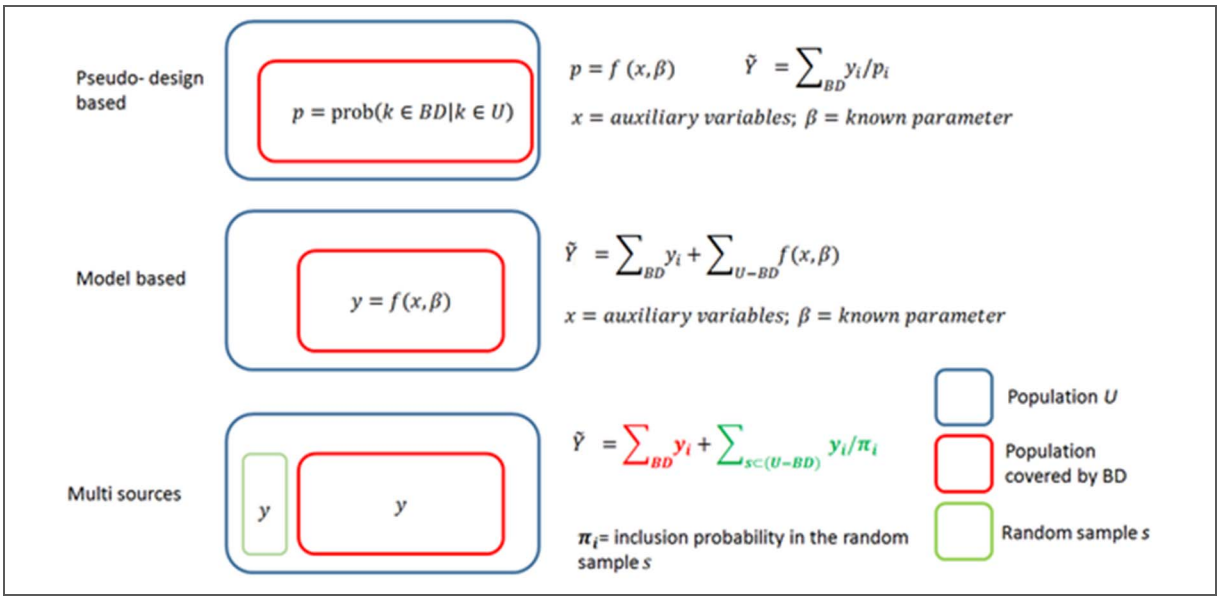
[9] Zeelenberg K. (2016). Big Data and Methodological Challenges in Official Statistics. *Discussion Paper nr.8*, Statistics Netherlands

[10] Elliott, M. R., Valliant R. (2017). Inference for Nonprobability Samples. *Statistical Science*, 32, pp. 249–264.

[11] Valliant R., Dorfman A. H., Royall R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, Wiley, New York.

Multi sources: it is not able to predict the inclusion probabilities or the values of target *y* variable. The approach requires other sources for not covered population (usually traditional survey data) and BD for covered population according to multi source approach.

Figure 3 **Synthetic definition of three general approaches to the inference when using Big Data**



Methodology helps to deal with validity concerns of BD too. Lack of validity (or relevance) occurs when the BD variable does not match the definition of the target variable. Usually validity guides the decision of the inferential approach to be applied. The lack of validity leads to systematic error if the BD variable is directly used for producing the statistics. In this case, the BD variable should be used as auxiliary information. Though the BD linkability holds, a sample of ground truth data collecting *y* target variable is performed and using the linkage between sampled units and BD units a model between *y* and the auxiliary variables in the BD is fitted. Otherwise, though linkability condition does not hold, an option could be to integrate/reconcile the official statistics (survey based or register based) with BD statistics at macro level, improving precision, timeliness and granularity (Zeelenberg, 2016).

Of course, both validity and selectivity could occur and, for instance, if the statistical process uses the BD assisted approach to calibrate the estimates, it is still important to check the selectivity concern of the auxiliary variable for defining efficient estimates.

4.3.1 Paradigm shift in statistical methodology

Methodology to figure out the BD basically uses statistical models. Commonly, the NSIs use parametric models to make inference. In the BD context these models could have problems, such as:

- high computational complexity unsuitable for high data volume;
- extreme sensitivity to erroneous data / outliers;
- extracting spurious correlation due to high data volume.

NSI must change the methodology approach considering Machine learning or algorithm based approaches that are:

1. less computational demanding;
2. scalable thanks can implicit parallelism;
3. more tolerant towards erroneous data and robust from the departures of model assumptions.

These approaches are less sophisticated than parametric approach but usually the advantages of using all the BD compensate their drawbacks.

CHAPTER 5
MAIN RESULTS

5.1 Scanner Data for the Consumer Price Index

5.1.1 Objective

Scanner data represent transaction data obtained from retail chains containing information on turnover and quantities per item code, based on transactions for a given period and from which unit value prices can be derived at item code level. They represent a crucial innovative Big Data source to estimate inflation and can be used to replace price collection in hypermarkets and supermarkets. At European level the issue of the adoption of scanner data for Harmonized Index of Consumer Price (HICP) compilation, is one of the more challenging issues. Scanner data provide several advantages deriving from the availability of detailed information concerning sales and quantity at weekly frequency, GTIN (Global Trade Item Number) by GTIN, outlet by outlet distributed throughout the entire national territory.

To access them, since the end of 2013 a stable cooperation has been established among Istat, Association of modern distribution (ADM), retail trade chains (RTCs) and Nielsen. For 2014, 2015 and 2016, scanner data of grocery products (processed food and goods for personal care and house cleaning, indeed excluding fresh food with variable weight) were collected by Istat through Nielsen for about 1400 outlets of the main six RTCs for 37 provinces. In sight of the inclusion of scanner data into price indices calculations, a probabilistic sample of about 2100 outlets of 16 RTCs covering the entire national territory was selected, for which Nielsen provided Istat with SD for years 2016 and 2017. These data were used to test statistical and IT solutions.

5.1.2 Results

For the entry into production of scanner data referred to grocery products, a two stage sample design has finally been adopted. For the first stage a probabilistic selection of a sample of 1781 hypermarkets and supermarkets has been carried out. For the second stage, a static approach to the sample design has been adopted, cutting off a sample of GTINs within each outlet/aggregate of products, covering 40% of turnover but selecting no more than the first 30 GTINs in terms of turnover. A “thank” of potentially replacing outlets (258) and GTINs (until a coverage of 60% of turnover within each outlet/aggregate, but no more than first 60 in terms of turnover) has been detected in order to better manage the possible replacements during 2018. To compile monthly indices, data referred to the first full three weeks of each month are used.

Table 1 shows the figures describing the data base implemented for December 2017 and regularly updated through scanner data and the data actually used to estimate monthly inflation since January 2018 (see the press release hyperlink- http://www.istat.it/it/archivio/208811).

The database consists of 2,039 outlets and information about turnover and quantities of 1,556,392 combination of GTINs/outlet selected in December 2017: the data currently used to estimate monthly inflation are referred to 1,781 outlets (1,370,898 GTINs/outlet) in order to keep a thank of possible replacing outlets if some of the 1,781 disappear during 2018. Moreover, information for other 1,984,181 GTINs/outlet has been selected in the 2,039 outlets as possible replacing elementary items. Taking into account all the GTINs/outlet (beyond the sample) for which Istat receives the data and that for each

GTIN/outlet weekly information is available, 45,164,106 records are stored for the 2039 outlets with reference to December and approximatively this amount of records are loaded monthly in the data base.

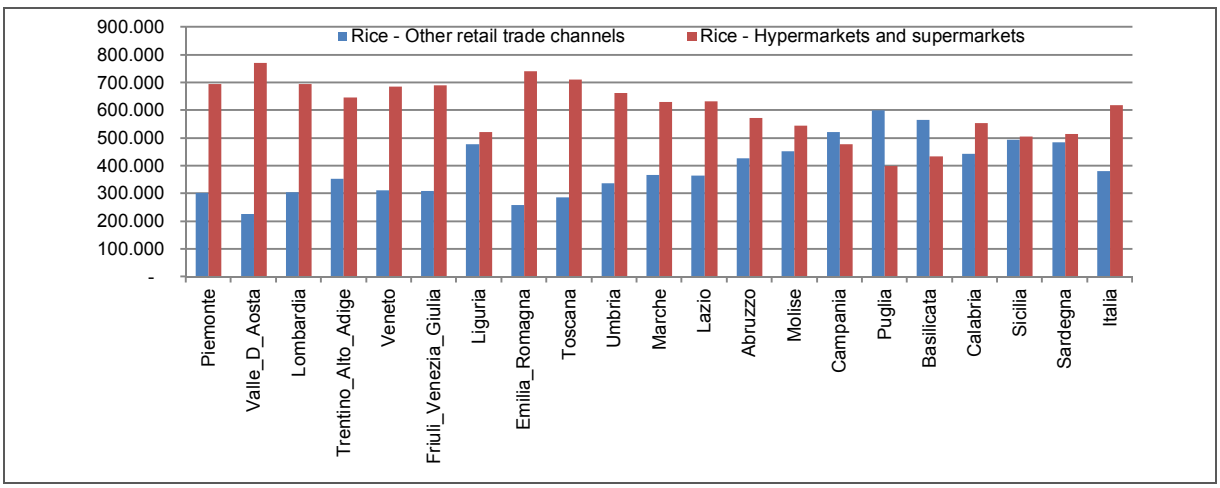
Table 1 **Monthly data base of scanner data and data used to estimate monthly inflation since January 2018**

	Total	Sample used in the compilation of CPI/HICP
Involved Outlets	2,039	1,781
GTINs/outlet	1,556,392	1,370,898
Possible replacing GTINs	1,984,181	1,760,011
Stored records	45,164,106	9,781,496
Coverage	60% coverage of turnover for each outlet/aggregate of products or 60 GTINs	

The adoption of this Big Data source within the illustrated sample design has allowed the introduction for the first time in the Italian CPI/HICP compilation of the use of sampling weights deriving from the probabilistic scheme. Therefore, for scanner data, outlet indices of grocery aggregates of products (79) are calculated as unweighted Jevons indices (geometric mean of GTINs elementary indices) and then provincial scanner data indices of aggregates of products are compiled as arithmetic mean of outlet indices of aggregate of product using sampling weights. Afterwards scanner data regional indices of aggregates of product are elaborated as arithmetic mean of provincial indices of aggregates of product using population as weights and national indices as arithmetic mean of regional indices using consumption expenditures as weights.

At each territorial level, for grocery products it is carried out the weighted aggregation of indices referred to scanner data (hyper and supermarkets) with indices referred to other channels of retail trade distribution: the weights for the aggregation are estimated using qualitative information on the shopping habits of consumers coming from Household Budget Survey (HBS). In graph 1, it is showed an example, referred to a specific product (Rice), of the weights of the retail trade channels considered (hypermarkets and supermarkets/others) in different regions (the differences between regions in the North, with the exception of Liguria, and regions in the South are clear).

Graph 1 **2018 CPI/HICP regional weights of different retail trade channels for Rice**



5.1.3 Lessons learnt and perspectives

In the process towards the entry into production of scanner data, some crucial issues emerged. The first one concerns the role of partnerships: the collaboration with the modern distribution and the representative association (Association of Modern Distribution, ADM) was fundamental to obtain scanner data. The collaboration was set up on the common understanding of the benefits that the information on inflation would have achieved from scanner data and of the opportunity for the modern distribution to identify its own role and contribution to the price general index. Nevertheless, this collaboration still presents some risks because it is only partially defined; thus, in the perspective of future developments, a further step is needed, aimed at setting a more formal agreement.

Another crucial issue is represented by the huge IT problems in computational terms derived from the large scale adoption of scanner data. This problem was at the base of the choice for 2018 of a static approach to GTINs sampling and it is still a key point to deal with to move towards a dynamic approach, whose adoption is the main goal for 2019. The introduction of scanner data referred to retail trade outlet with surface between 100 and 400 square meters, on the one hand, and to fresh products with variable weight, on the other hand, is planned for 2020.

5.2 Internet as a Data Source : ICT use of enterprises: web ordering, job advertising and presence on social media

5.2.1 Objective

A multi-source approach (based on a combined use of survey, administrative and BD sources) should allow to overcome usual limits of each single source, in particular those affecting Big Data.

This multi-source approach requires a shift in the paradigm of statistical inference. The traditional one followed by NSIs is usually based on design-based survey sampling theory and model-assisted inference. The new one (algorithmic-based inference) is derived by data science: the emphasis is on the exploration of all available data, seeking information that has not been extracted so far; models have to be evaluated no longer by their interpretability, but rather by their capability to correctly predict values at unit level, and to use them for estimating the parameters of interest.

Istat has experimented this new approach in order to obtain a subset of the estimates currently produced by the sampling “Survey on ICT usage and e-Commerce in Enterprises”, yearly carried out by Istat and by the other member states in the EU. Target estimates of this survey include the characteristics of websites used by enterprises to present their business (for instance, if the website offers web ordering facilities; job vacancies; presence in social networks). To produce these estimates, data are collected by means of traditional questionnaires.

An alternative way is to make use of Internet data, i.e. to collect data by accessing directly the websites, processing the collected texts to individuate relevant terms, and modelling the relationships between these terms and the characteristics we are interested in estimating. To do that, the sample of surveyed data plays the role of a training set useful to fit models that can be applied to the generality of enterprises owning a website. Administrative data (mainly contained in the Business Register) are used to cope with representativeness problems related to BD source. The sequential application of web scraping, text mining and machine learning techniques allows to obtain auxiliary variables suitable for applying a prediction approach and produce estimate that can be compared to the survey ones.

In terms of quality (accuracy), the impact of the new estimators is both positive (reduction of the variability and of the bias due to sampling variance, to total non-response and to measurement errors in the survey) and negative (model bias and variance). Whenever the quality of estimates obtained by means of this new approach reveals to be not lower than the ones produced by the traditional process, the former has to be preferred, as it allows not only to produce aggregate estimates, but also to predict individual values, useful for instance to enrich the information contained in registers.

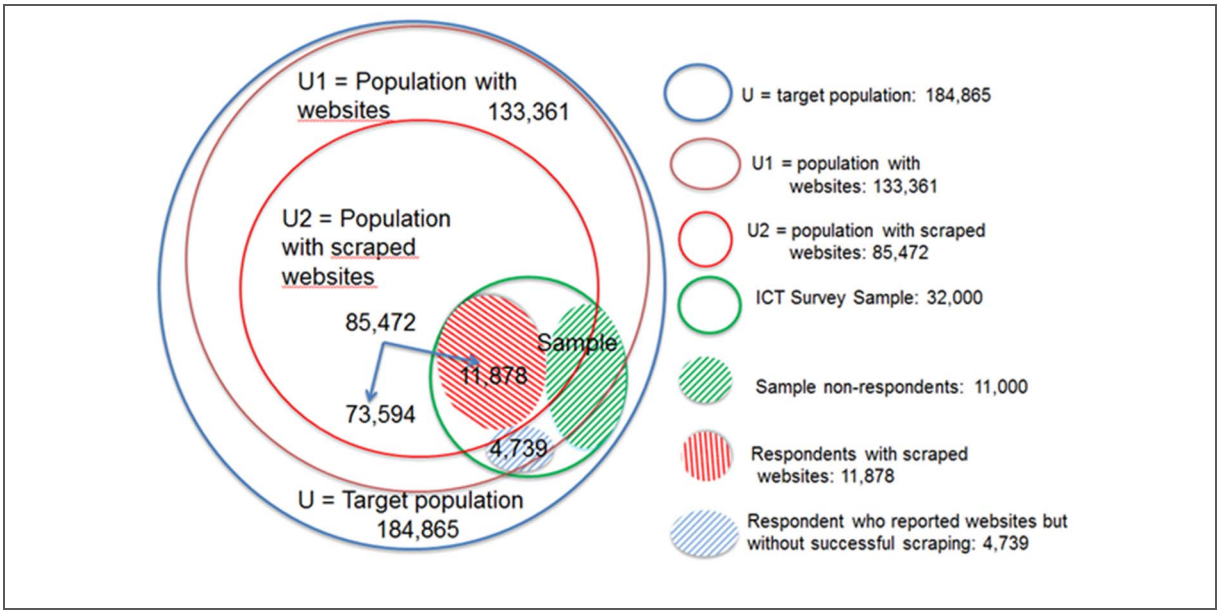
5.2.2 Results achieved

A complex procedure has been developed in order to:

1. get the websites address (Uniform Resource Locator) potentially for all enterprises included in the population of reference (URL retrieval);
2. access websites with available URL and scrape their content (web scraping);
3. process the content of the scraped websites in order to identify the best predictors for the target variables (text mining);
4. fit models (machine learning) in the subset of enterprises where both Internet data and survey data were available (considering survey data as the true values) and predict the values of target variables for all the enterprises for which the retrieval and scraping of their websites was successful.

Figure 1 reports the different subsets of the population of interest (enterprises with at least 10 persons employed operating in various economic activities of manufacture and non-financial services), involved in the overall procedure:

Figure 1 Subsets of the population of interest



The "Survey on ICT usage and e-Commerce in Enterprises" produce on a yearly basis a set of estimates reporting rates of web-ordering, job advertising and presence on social media declared by enterprises that own or make use of websites. In particular enterprises are asked to answer to filter question about having own web site of Internet page. This filter question does not refer specifically to

the ownership of the website, but to the use of a website by the enterprise to present its 'business'. It includes not only the existence of a website which is located on servers belonging to the enterprise or located at one of the enterprise's sites, but also third party websites (e.g. one of the group of enterprises to which it belongs i.e. website of the parent company or holding company). However, it does not include any presence of the enterprise on the web (for example the presence of the enterprise with e.g. its name or its contact information in online yellow pages are not included in this variable). Moreover enterprises on e-marketplaces where they have the possibility to advertise themselves, quote prices for ad hoc services etc. are not enterprises that are considered to have a website.

These estimates are available for the total population, and for different domains of interest, among which:

- Cross-classification by *Size classes of persons employed* (4) and *Economic macro sectors* (4) (16 different sub-domains);
- *Administrative Regions* (21 different domains);
- *Detailed economic activities* (27 domains).

Together with the current estimation method (*design based / model assisted*), alternative estimates have been calculated by adopting two different estimators: a *full model based* one and a *combined* one. The characteristics of the three different estimators are reported in the following table.

Table 1 Estimators

Estimator	Formula	Weighting	Description
Design based / model assisted	$\hat{Y} = \sum_r y_k w_k$	$\sum_{k=1}^{r} w_k = N_U$	w_k weights are obtained by calibration procedure of basic weights (inverse of inclusion probabilities) making use of known totals in the population in order to reduce the bias due to non-response and the variability due to sampling errors
Model based	$\hat{Y} = \sum_{U^2} \tilde{y}_k w'_k$	$\sum_{k=1}^{U^2} w'_k = N_{U^1}$	The estimate of the total number of enterprises offering web ordering facilities on their websites is given by the count of the predicted values $\tilde{y}_k$ for all units for which it was possible reach their websites (population U^2), calibrated in order to make them representative of all the population having websites (U^1).
Combined	$\hat{Y} = \sum_{U^2} \tilde{y}_k + \sum_{r^1} (\tilde{y}_k - y_k) w''_k + \sum_{r^2} y_k w'''_k$	$\sum_{k=1}^{r^1} w''_k = N_{U^2}$ and $\sum_{k=1}^{r^2} w'''_k = N_{U^1 - U^2}$	Estimates are produced by summing three components: 1. the counting of predicted values in the subpopulation U^2 of units for which it was possible to scrape and process corresponding websites; 2. an adjustment based on the consideration of the differences between the r^1 reported values and the predicted values (expanded to the same subpopulation U^2); 3. the counting of observed values for the r^2 respondents that declared a website, that was not found nor scraped, expanded to the whole subpopulation $U^1 - U^2$.

Once computed, the 3 different sets of estimates can be compared. For instance, considering web-ordering the results are reported in Table 2. The first column indicates the domain for which the estimates are calculated. The absolute values of sample units, population, and websites offering web-ordering facilities are listed. Current design-based estimates together with lower and upper limits of corresponding confidence interval are reported. Finally, model based and combined estimates are shown (highlighted in red when they lay outside the design based confidence intervals).

Table 2 Web-ordering estimates comparison

DOMAIN		Design based estimate	Lower limit C.I.	Upper limit C.I.	Model based estimate	Combined estimate
Size class of persons employed						
cl1	from 10 to 49	14.57	13.32	15.83	15.22	13.8
cl2	from 50 to 99	15.96	13.83	18.08	16.23	15.1
cl3	from 100 to 249	17.91	16.04	19.78	17.71	17.38
cl4	from 250 and more	25.72	23.78	27.65	23.25	26.04
Economic macro sectors and size classes						
M1cl1	Manufacturing (C) 10-49	10.04	8.08	11.99	11.06	9.88
M1cl2	Manufacturing (C) 50-99	12.09	8.87	15.3	14.8	14.29
M1cl3	Manufacturing (C) 100-249	15.69	12.6	18.77	15.76	15.38
M1cl4	Manufacturing (C) 250+	24.18	21.06	27.3	22.65	21.09
M2cl1	Energy (D,E) 10-49	8.69	6.54	10.84	9.73	11.51
M2cl2	Energy (D,E) 50-99	10.5	5.98	15.03	11.55	9.73
M2cl3	Energy (D,E) 100-249	13.89	8.95	18.84	15.04	11.79
M2cl4	Energy (D,E) 250+	18.79	11.86	25.72	16.97	14.55
M3cl1	Construction (F) 10-49	2.92	2.03	3.81	5.54	5.02
M3cl2	Construction (F) 50-99	3.1	0.29	5.91	5.32	4.28
M3cl3	Construction (F) 100-249	2.05	0.3	3.81	5.19	5.19
M3cl4	Construction (F) 250+	8.12	1.09	15.16	10	8.75
M4cl1	Non-financial services 10-49	20.28	18.26	22.3	20.26	18.4
M4cl2	Non-financial services 50-99	21.76	18.36	25.16	19.36	17.68
M4cl3	Non-financial services 100-249	21.76	19.03	24.48	20.89	20.82
M4cl4	Non-financial services 250+	28.32	25.56	31.07	24.85	31.51
Nace economic activities						
naceict0	activities not included in ICT Sector (defined in terms of NACE as 261, 262, 263, 264, 268, 465, 582, 61, 62, 631, 951)	15.13	13.94	16.31	15.54	14.25
naceict1	activities included in ICT Sector	10.97	8.17	13.77	14.88	13.65
naceist01	manufacture of food products, beverages and tobacco products	19.4	12.86	25.94	17.04	14.82
naceist02	manufacture of textiles, apparel, leather and related products	16.05	9.2	22.91	13.85	11.93
naceist03	manufacture of wood and paper products, and printing	12.45	6.55	18.36	13.21	11.3
naceist04	manufacture of coke and refined petroleum products, of chemicals and chemical products, of basic pharmaceutical products and preparations, of rubber, plastic and of other non-metallic mineral products	10.44	6.85	14.02	11.78	11.73
naceist05	manufacture of basic metals and fabricated metal products, except machinery and equipment	5.94	3.02	8.85	7.65	7.25
naceist06	manufacture of computer, electronic and optical products	9.47	4.94	13.99	11.98	9.73
naceist07	manufacture of electrical equipment and of machinery and equipment n.e.c.	5.62	2.86	8.38	10.45	8.88
naceist08	manufacture of transport equipment	16.68	3.04	30.32	12.49	14.72
naceist09	manufacture of furniture, other manufacturing, and repair and installation of machinery and equipment	8.84	4.57	13.11	11.79	11.27
naceist10	electricity, gas steam, air conditioning supply, water supply, sewerage, waste management and remediation activities (d-e)	9.87	8.03	11.71	10.77	11.5
naceist11	construction	2.94	2.07	3.81	5.54	5
naceist12g	wholesale and retail trade and repair of motor vehicles and motorcycles	20.39	18.98	21.81	20.32	20.28
naceist15	transport and storage, except warehousing and support activities for transportation (h except 53)	14.16	6.57	21.75	11.47	10.9
naceist16	postal and courier activities	26.13	16.37	35.89	14.16	18.26
naceist17	accommodation	82.57	77.37	87.78	71.77	68.71
naceist18	food service activities	23.63	14.59	32.67	22.23	15.48

Table 2 *(continued)* **Web-ordering estimates comparison**

DOMAIN		Design based estimate	Lower limit C.I.	Upper limit C.I.	Model based estimate	Combined estimate
Nace economic activities						
naceist19	publishing activities	62	45.62	78.39	49.21	49.44
naceist20	motion picture, video and television programme production, sound recording	15.62	2.97	28.27	23.63	17.64
naceist21	telecommunications	21.45	13.71	29.19	20.44	20.8
naceist22	IT and other information services	8.82	5.65	11.99	12.89	12.45
naceist23	real estate activities	11.08	5.78	16.39	13.68	13.99
naceist24	professional, scientific and technical activities except veterinary activities	5.27	1.57	8.97	10.33	7.5
naceist25	administrative and support service activities except travel agency, tour operator and other reservation service and related activities (N except 79)	4.83	2.93	6.73	8.4	7.33
naceist26	travel agency, tour operator and other reservation service and related activities	44.2	31.8	56.59	44.19	47.71
Administrative Regions						
REG01	PIEMONTE	11.96	7.46	16.46	13.77	13.6
REG02	VALLE D'AOSTA	16.8	6.43	27.17	21.77	20.76
REG03	LOMBARDIA	11.76	10.42	13.1	14.38	13.07
REG05	VENETO	14.72	12.22	17.22	16.67	15.8
REG06	FRIULI-VENEZIA GIULIA	17.17	5.62	28.73	14.23	14.67
REG07	LIGURIA	11.39	5.96	16.83	14.86	12.02
REG08	EMILIA-ROMAGNA	12.63	9.89	15.36	15	14.9
REG09	TOSCANA	14.55	10.3	18.8	15.91	14.35
REG10	UMBRIA	24.23	20.35	28.1	16.34	15.43
REG11	MARCHE	20.37	7.51	33.23	16.58	14.04
REG12	LAZIO	16.62	12.47	20.77	16.02	13.79
REG13	ABRUZZO	17.41	9.08	25.74	13.87	14.23
REG14	MOLISE	14.06	4.08	24.03	12.41	15.17
REG15	CAMPANIA	15.87	10.82	20.91	14.4	14.33
REG16	PUGLIA	20.32	14.46	26.18	14.61	12.21
REG17	BASILICATA	12.02	4.34	19.7	13.78	8.34
REG18	CALABRIA	20.4	10.93	29.87	17.47	10.05
REG19	SICILIA	19.17	6.95	31.4	16.7	12.56
REG20	SARDEGNA	14	7.85	20.14	14.93	15.29
REG21	Provincia Autonoma Bolzano	31.43	24.93	37.92	29.38	26.64
REG22	Provincia Autonoma Trento	19.51	16.87	22.14	22.78	23.21
Total		14.97	13.81	16.13	15.51	14.22

For web-ordering estimates a graphical comparison is shown in Figure 2. The dashed lines define the area delimitated by the lower and upper limits of the confidence intervals calculated in correspondence of each design based estimate.

The same distributions are reported also in the case of job advertisements (Figure 3) and presence in social media (Figure 4).

Figure 2 **Web-ordering estimates comparison (dotted lines represent limits of confidence intervals of design based estimates)**

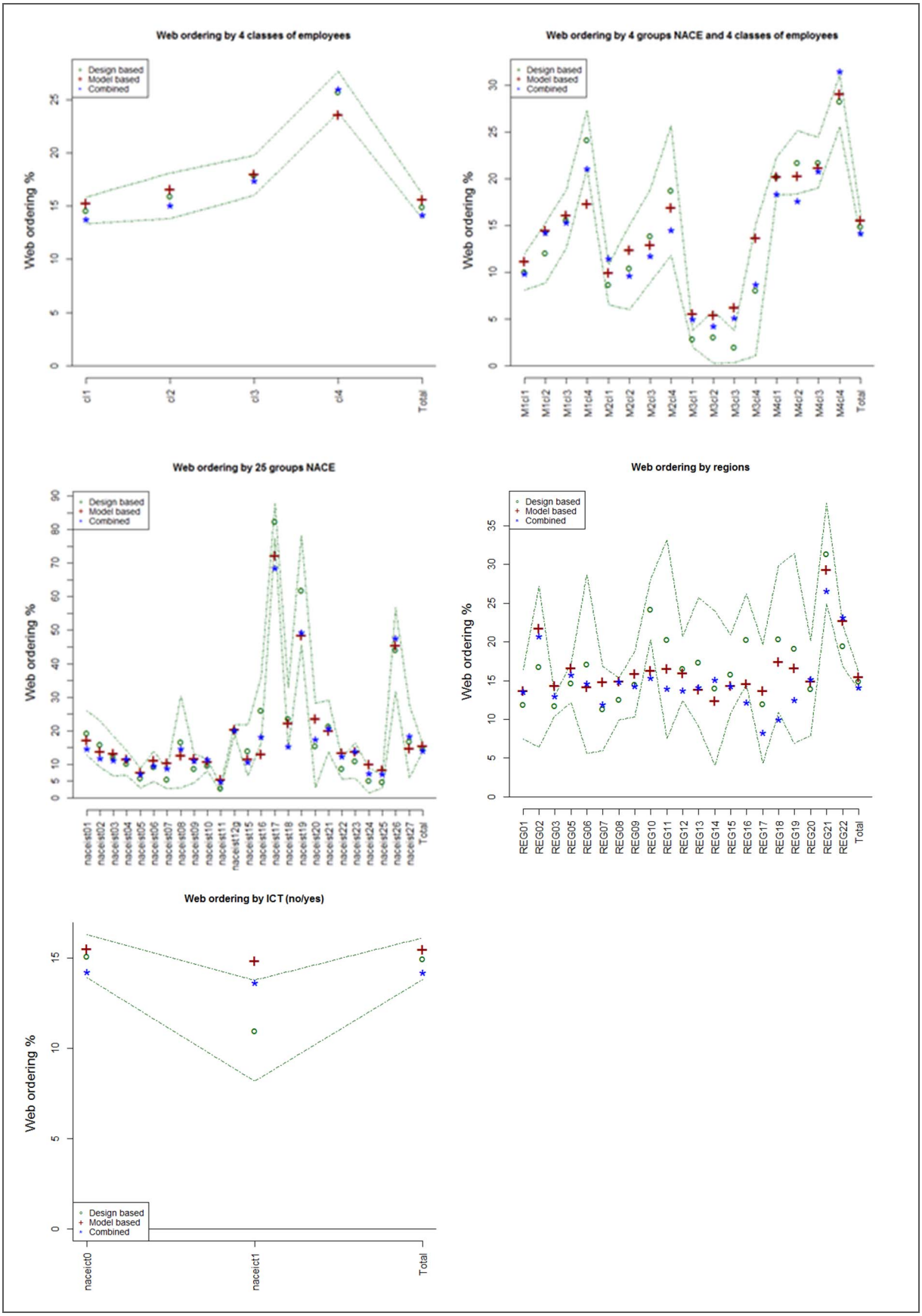


Figure 3 **Job advertisements estimates comparison (dotted lines represent limits of confidence intervals of design based estimates)**

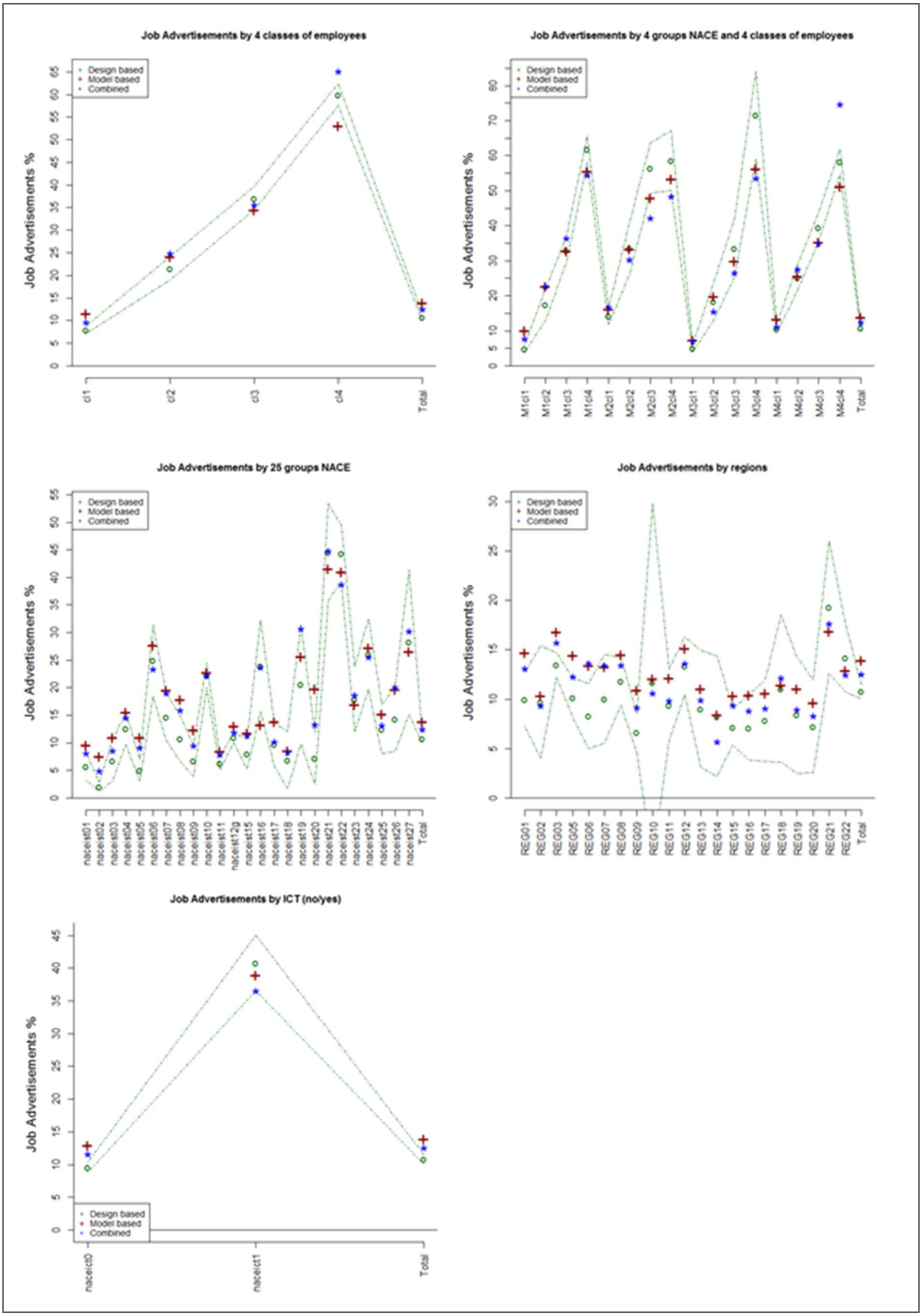
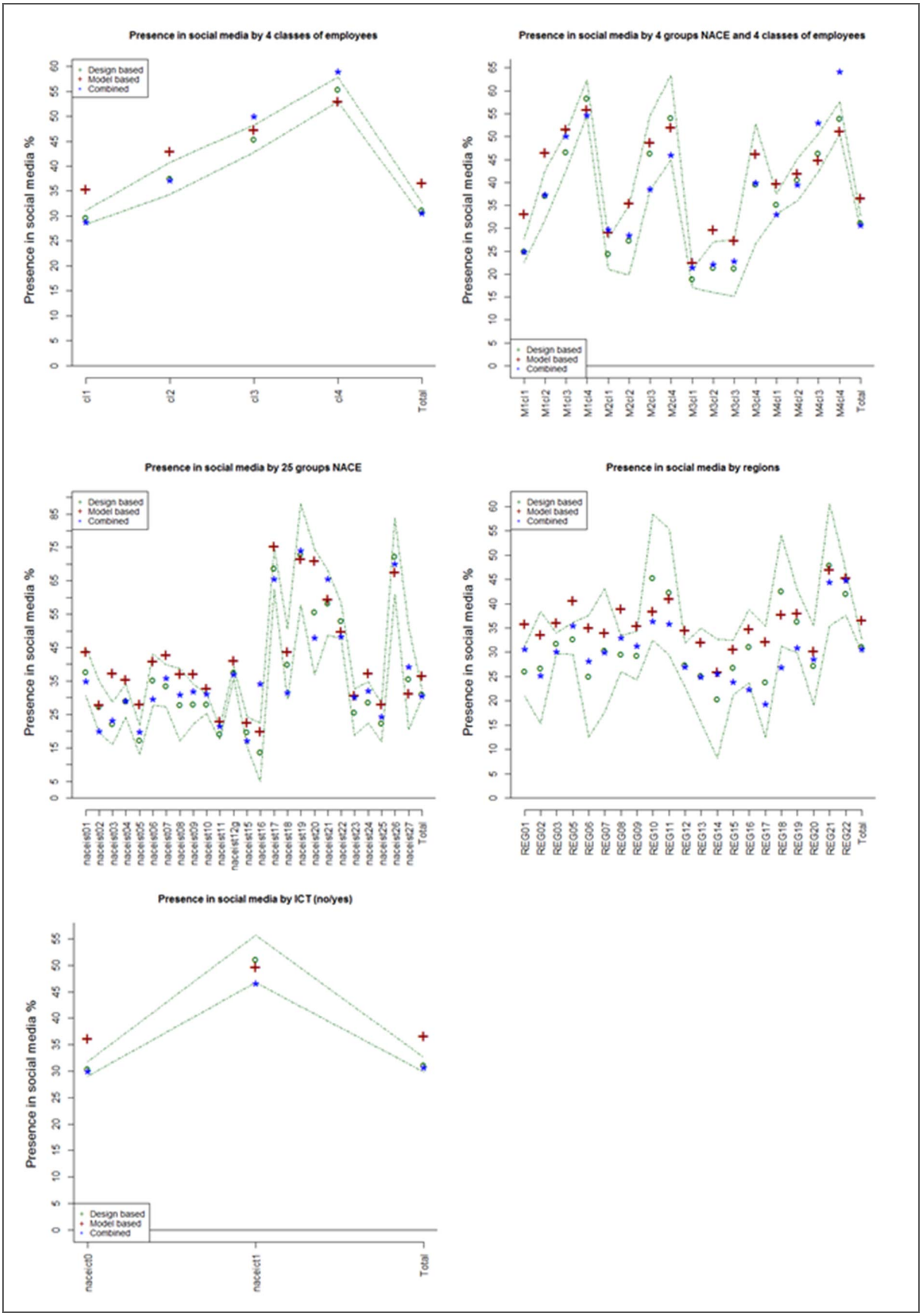


Figure 4 **Presence in social media estimates comparison (dotted lines represent limits of confidence intervals of design based estimates)**



5.2.3 Lessons learnt

A first analysis of the estimates related to web-ordering, job-advertisements and presence in social media rates, obtained with the two alternative estimators, compared to the estimates produced by the official survey, allows some preliminary conclusions:

The three different sets are not incoherent. For instance, considering web-ordering the estimates for the total are well inside the confidence interval of the survey estimate, and this is the same for many values in the different domains.

Looking at coherence as one important dimension of quality, both combined estimates and full model based estimates can be considered as equally acceptable. But two considerations can be made.

1. The second component of the combined estimator is based on an assumption of perfect correctness of reported values, and considers predicted values as errors when they do not coincide with the reported ones. But controls have been carried out when fitting models, and in half of the cases in which predicted values were contradictory with reported ones, this was not due to model fault, but to response errors. So, this assumption does not always hold. In any case it would be advisable to deepen this phase also by returning to the respondents to verify if it is an error in response or if, for example, the model has evaluated the content of a site different from that one considered by the respondent.
2. If a medium-term aim is to make multi-annual frequency of the questions in the survey related to the websites characteristics (as Eurostat envisaged), then the combined estimator cannot be applied, as it relies on the current availability of reported values from the survey, and the full model based estimators remains the only alternative. In this case, there would be an issue in time series analysis due to problems in comparability between survey estimates and model based ones.

The main flaws of the model based estimator are in the presence of:

- prediction errors;
- under-coverage of the population of enterprises owning websites, part of which has not been reached by web scraping.

As for the first, taking into consideration the presence of response errors in the test set, once eliminating them by manual inspection, the accuracy of the model predictions increases to more than acceptable levels (around 90% for web ordering, about the same for the other two variables), in any case comparable with the accuracy of survey data.

As for the second, pseudo-calibration allow to limit the bias, especially when the difference in the values of the parameters in the two sub-populations is not high, as it is the case.

5.3 Internet as a Data Source: use of Open Street Map for accident investigation on the road and motorways networks

5.3.1 Objective

Road Safety Performance Indicators (RSPI) give a multidimensional approach for accident investigation concerning roads, vehicles and persons involved. Combining the use of statistical surveys, administrative Geographical Information Systems (GIS) and Big Data (BD) sources the result gives new elements on planning infrastructure solution, applying policies to reduce deaths and serious injuries, reducing social costs to the community and estimating efficiency and effectiveness of safety initiatives.

Preventing road trauma on public roads is a core responsibility for government, its agencies and stakeholders. It requires a common and shared responsibility. The scale of the road safety challenge and the diversity of the effects of road traffic injury underline the importance of exploring synergies among the decision makers of the road network.

Road traffic injury is a major global public health problem. More than 1.24 million people die each year on the world's roads. Many more suffer permanent disability, and between 20 and 50 million suffer non-fatal injuries. These are mainly vulnerable road users and involve the most socio-economically active citizens.

In socio-economic terms, countries around the world are paying a high price for motorised mobility. Country estimates indicate that the value of preventing road death and injury is equivalent to between 1% and 7% of Gross Domestic Product. A European Commission target for Road Safety program 2011–2020 is planning of halving deaths in road traffic crashes.

Nowadays there is a clear information bias as regards the appropriate reference denominators to be placed as basis in construction of statistical indicators linked to road accidents. Resident population is used as a common proxy for exposed at risk in a specific geographical area, but not always an appropriate solution, especially in the light of the seasonal nature of road accidents and concentration, in some periods of the year, in specific locations.

The estimate traffic flows would undoubtedly provide a key to understanding integrated and innovative phenomenon. Resident population does not mean present population at the time of the event. Vehicle fleet can be another administrative source that gives a more accurate information, but the characteristic of the phenomenon implies a deductible distortion on measures due to the mobility of road users.

The purpose of the project is to expand statistical information with the supply of traffic flows, measuring the frequencies of the events in order to be interpreted not only as absolute values, but also as probability of being involved in the accident, taking into account the different exposure to risk.

The aim of this work is to give an assessment of risk for the entire Italian road network for all the different road agencies that are responsible. A small percentage of roads account can solve a large percentage of deaths and serious injuries; the task is to identify such routes as a priority identifying a high-risk crash locations.

The tool of the assessment of crash data with the geographical localization of road accidents is also a big effort that Istat and Automobile Club of Italy (ACI) are developing to identify high risk crossing the information with road networks and points of traffic.

5.3.2 Methodology

To reach the mentioned purpose, a Geographic Information System (GIS) is used. GIS is a system designed to capture, store, manipulate, analyse, manage, and present spatial or geographic data. GIS applications are tools that allow users to analyse spatial information, edit data in maps, and present the results of all these operations. In order to relate information from different sources, GIS uses spatial location as the key index variable. Just as a relational database containing text or numbers can relate many different tables using common key index variables, GIS can relate otherwise unrelated information by using location as the key index variable. This key characteristic of GIS has begun an alternative frontier on producing statistical information. Any variable that can be located spatially using an x, y, and z coordinates representing, longitude, latitude, and elevation, respectively. These GIS coordinates may represent other quantified systems of territories (polygons), road networks (lines) and point of traffic (points). Join attributes by location is the algorithm that takes an input vector layer and creates a new vector layer that is an extended version of the input one, with additional attributes in its attribute table. The additional attributes and their values are taken from a second vector layer.

A spatial criteria is applied to select the values from the second layer that are added to each feature from the first layer in the resulting one (Chart 1).

Chart 1 **Graphical representation of the join attributes by location algorithm**

To measure the road network length, **Open Street Map system** is applied. Open Street Map is a collaborative project aimed on creating free content maps of the world. The project aims at a collection world of geographical data, with the main purpose of creating maps and cartography.

The key feature of the geographic data present in OSM is having a free license, the Open Database License. It is therefore possible use them freely for any purpose with the only constraint of mentioning the source. Everyone can contribute by populating or correcting data.

The maps are created using the data recorded by portable GPS devices, aerial photographs and other free sources. Most of the Android and iOS GPS navigation software on portable devices are powered by OSM as WisePilot, Maps.me, NavFree, Scout etc.

The Open Street Map vector layers, used in this work, daily updated and free downloadable data, concern road graphs and points of traffic.

The total amount of the road arches in Italy are 3,490,212. By geographical territory: 905,953 in North East, 999,451 in Nord West, 642,777 in Centre, 565,582 in South and 376,449 in Islands.

The Point of Traffic (POT) detected in Italy are 228,243. Divided by geographical territory: 60.128 in North East, 105,509 in Nord West, 29,109 in Centre, 16,904 in South and 16,593 in Islands. (Chart 2 and Chart 3)

Additional shapes available are Buildings, Land Use, Natural, Places, POWF (Point of Worship), POIS (Point of interest), Railways, Transport, Water and Waterways.

Chart 2 **Italian road graph network**

Chart 3 **Italian points of traffic location**

To build congruent indicators and harmonize the classification of road types by arch length in OSM and the location of road accidents, a “bridge matrix”, containing the corresponding attributes in both classification, is produced .

A link table with the aggregation of all 8,090 local administrative units territory, at 2011, to the 7998 municipalities at 2016, included in the Italian territory is built with an upgrade of Census Map localities shapes referred to 2011 (51,227 localities) to municipalities 2016. The choice of the localities shapes is due to the harmonization need of the complete roads graph to the “road type classification” used by the road accidents survey, defining a corresponding and fitted aggregation with the values of road accident survey variable in urban, rural and motorways road classification.

An innovative method, measuring the length in meters of the road graph, is given by the information on the number of carriageways of each road arch of OSM.

The different data sources integration lead to a more comprehensive understanding of the road safety problem and to address effective actions to face connected problems. Many stakeholders can apply new road safety policies using accurate and easily comprehensive road safety data. A meaningful target setting requires competent data collection and analysis of risk using the collected data.

Any improvement to infrastructure can contribute substantially to reduce deaths and serious injured road users. Many high severity crash types can be eliminated with the effective use of infrastructure. Just few localized infrastructure investments produce such high benefits as infrastructure measures targeted at making road safety improvement, road infrastructure, often the single most significant factor that contributes to the severity outcome of a crash. Moreover, good examples of infrastructure policy can be reused. Somewhat standard, guideline and tool are a mechanism to translate policy into action.

As regards data used to build indicators on the road accidents risk, they are currently collected in Italy by Istat Road accidents survey, referred to all road accidents resulting in deaths (within the 30th

day) or injuries, involving at least a vehicle circulating on the national road network and documented by a Police authority.

To carry out comparisons and evaluate the rates performance, three different indicators were calculated.

Road accidents (numerators) by road type, geographical details and other variables and dimensions, were used out of different denominators:

- GIS computing (Census Map + Open Street Map road graph at 1st January 2017) expressed by length in meters per carriageway;
- ACI Vehicle fleet (Automobile Club of Italy) all motorized vehicles excepted trailers on 31st December 2016;
- Resident population (demo.istat.it) on 31st December 2016.

The set of road indicators is developed by accidents, vehicles involved, killed and injured persons every 100 kilometres of carriageway in the province.

To make possible a ranking and an evaluation of calculated indicators has been used a generalized tool developed by Istat.

Two generalized tools are available for the analysis and benchmarking of results produced by different composite indicators[12] :

- RankerTool is a desktop software at https://www.istat.it/en/methods-and-tools/methods-and-it-tools/analyse/analysis-tools/ranker
- i.Ranker instead is a web application at https://i.ranker.istat.it

5.3.3 Results

The first result produced was the process of the three group of indicators: road accidents out of resident population, vehicle fleet and the new road network length.

The set of road indicators is developed by accidents, vehicles involved, killed and injured persons every 100 kilometres of carriageway in the province. Looking at the main results, a maximum risk exposure for motorways and urban infrastructure has been recorded mainly in large centres. For the rural roads, the medium-sized provinces are more affected by prevention measures.

The set of vehicles fleet indicators is developed by accidents, vehicles involved, killed and injured persons every 100 thousand registered vehicles in the province. Concerning key results, on the motorway sector, the low incidence of vehicles registered in the province with the presence of

12 The steps on the computing process in divided in three phases:
- **standardization of elementary indicators,** the standardization aims to make the indicators comparable as they are often expressed in different units of measurement and may have different polarities. Therefore, it is necessary to bring the indicators to the same standard, reversing the polarity, where necessary, and turning them into pure, dimensionless numbers;
- **aggregation of standardized indicators, itis the combination of all the components to form the synthetic index (mathematical function);**
- **validation of the synthetic index, it consists in verifying that the synthetic index is consistent with the general theoretical framework. In particular, the ability of the index to produce stable and correct results (robustness) and its discriminating capacity must be assessed.**

The three methods applied on this work are: MZ - Arithmetic Mean of the z-scores; MR - Relative index; MPI - Mazziotta-Pareto Index (MPI) (De Muro et al. 2010)
The methodological note and the user guide is available online at:
http://www.istat.it/en/files/2014/03/RANKER-manuale.pdf
https://i.ranker.istat.it/wr_guida.htm
https://i.ranker.istat.it/wr_guida_notametodologica.htm

important infrastructural nodes and seasonal factors amplifies the distortion of the indicators distribution. In urban areas, commuting is not highlighted in the construction of the vehicle fleet. In the rural area, this category of indicators does not show the presence of a dense network of consular roads on the territory.

The set of population indicators is developed by accidents, vehicles involved, killed and injured persons every million residents in the province. The low number of inhabitants by province, in the infrastructural nodes of the motorway network, explains the values of the indicators. Port areas, transit areas and production settlements in their urban areas do not moderate the values of the indicators according to the resident population. In rural areas, the low resident population amount greatly influences the correct computing of the indicators showing a biased incidence of risk, without valuating the real traffic flows.

On the set of road indicators, the application of the three methods shows a high correlation between the values of the three distributions, confirmed also by the respective indexes of cograduation. MZ - Arithmetic Mean of the z-scores is the method that better fits the rankings with the MR Relative index and MPI - Mazziotta-Pareto Index.

Chart 4 **Italian road length MZ indicator ranking**

Table 1 **Road length indicators ranking and values as a synthetic extract of the first eleven provinces**

PROVINCE	MZ	MR	MPI-	MZ	MR	MPI-
Torino	14	11	13	0.6416	0.3375	106.0732
Vercelli	48	45.5	50	-0.017	0.2116	99.1453
Novara	36	40	35	0.1605	0.2289	101.4378
Cuneo	91	92	91	-0.6395	0.0902	93.5663
Asti	71	61	70	-0.2648	0.1702	96.9966
Alessandria	55.5	54	55	-0.1287	0.1884	98.5395
Aosta	103	103	102	-0.8391	0.0508	91.5172
Imperia	39	41	38	0.0967	0.2248	100.6247
Savona	22	19	22	0.4706	0.3014	104.0491
Genova	6	4	7	1.2715	0.4707	109.4584
La Spezia	42	38	41	0.0711	0.2355	100.3471

Table 2 Cograduation matrix among the methods applied on road length indicators

Ranks	MZ	MR	MPI-	Average
MZ	1.0000	0.9958	0.9980	**0.9979**
MR	0.9958	1.0000	0.9942	0.9967
MPI-	0.9980	0.9942	1.0000	0.9974

Table 3 Correlation matrix among the methods applied on road length indicators

Values	MZ	MR	MPI-	Average
MZ	1.0000	0.9948	0.9960	**0.9969**
MR	0.9948	1.0000	0.9945	0.9964
MPI-	0.9960	0.9945	1.0000	0.9968

The set of vehicles fleet indicators using the same methods gives very different rankings of the synthetic indexes.

MZ - Arithmetic Mean of the z-scores is not the best method but anyway the difference with the relative index (MR) is acceptable.

Chart 5 Italian vehicle fleet MZ indicator ranking

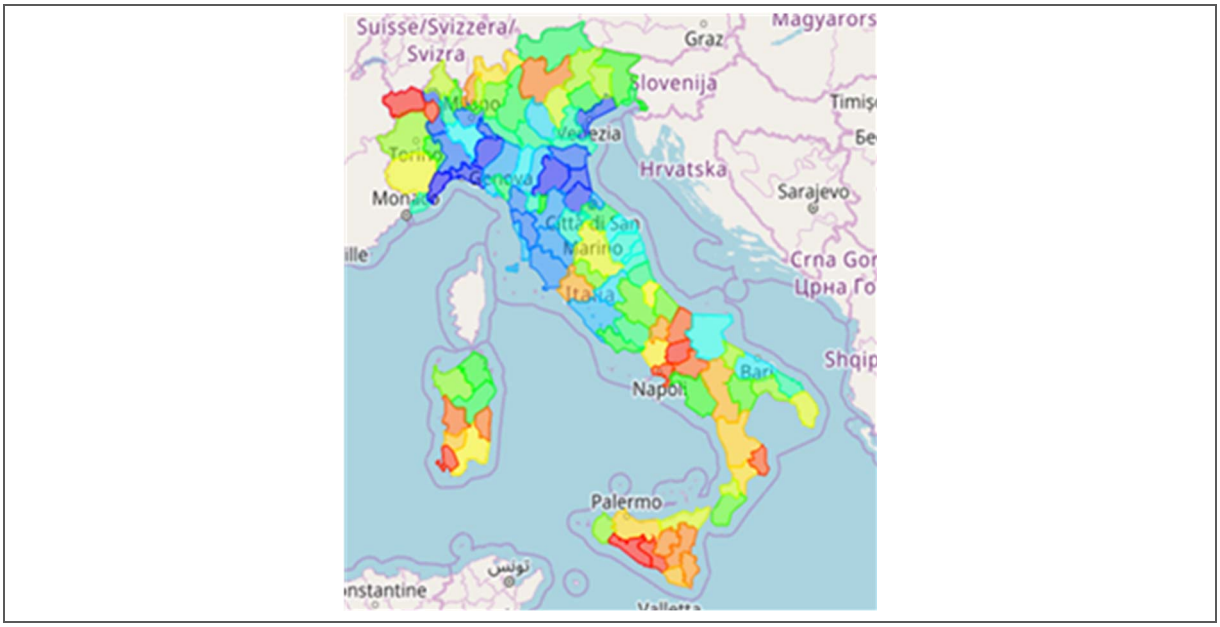


Table 4 Vehicle fleet indicators ranking and values as a synthetic extract of the first eleven provinces

PROVINCE	MZ	MR	MPI-	MZ	MR	MPI-
Torino	70	74	73	-0.1878	0.2623	97.3081
Vercelli	8	13	12	0.7045	0.4262	105.3578
Novara	40	41	38	0.1674	0.3327	101.4093
Cuneo	81	80	78	-0.3135	0.2508	96.5824
Asti	57	55	52	-0.0094	0.3001	99.5173
Alessandria	11	11	7	0.6703	0.4279	106.237
Aosta	109	109	109	-1.0611	0.1066	89.1809
Imperia	37	39	39	0.2196	0.3356	101.3224
Savona	1	1	1	1.2361	0.5369	109.5656
Genova	4	6	28	0.8225	0.4441	102.8264
La Spezia	23	26	30	0.4339	0.3781	102.5037

Table 5 **Cograduation matrix among the methods applied on vehicle fleet indicators**

Ranks	MZ	MR	MPI-	Average
MZ	1.0000	0.9980	0.9862	0.9947
MR	0.9980	1.0000	0.9866	**0.9949**
MPI-	0.9862	0.9866	1.0000	0.9909

Table 6 **Correlation matrix among the methods applied on vehicle fleet indicators**

Values	MZ	MR	MPI-	Average
MZ	1.0000	0.9984	0.9857	0.9947
MR	0.9984	1.0000	0.9867	**0.9950**
MPI-	0.9857	0.9867	1.0000	0.9908

The set of population synthetic indexes are similar as ranking to those for vehicle fleet. MZ - Arithmetic Mean of the z-scores is the method that better fits the rankings with the MR and MPI-

Chart 6 **Italian resident population MZ indicator ranking**

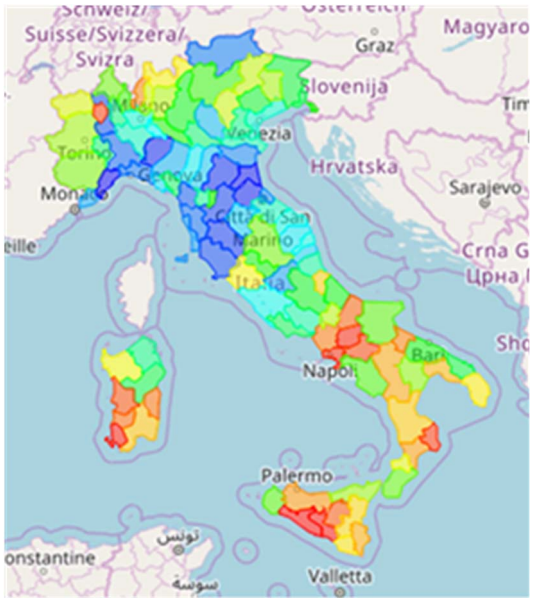


Table 7 **Resident population indicators ranking and values as a synthetic extract of the first eleven provinces**

PROVINCE	MZ	MR	MPI-	MZ	MR	MPI-
Torino	72	77	72	-0.2167	0.2505	97.0752
Vercelli	9	11	11	0.7949	0.4289	106.1896
Novara	42	44	39	0.1218	0.3157	100.9608
Cuneo	68	66	66	-0.1865	0.2656	97.7868
Asti	44	45	41	0.1165	0.3118	100.7817
Alessandria	11	12	8	0.7584	0.4287	107.128
Aosta	79	79	75	-0.2783	0.2464	96.718
Imperia	19	19	21	0.4857	0.3828	103.8206
Savona	1	1	1	1.5918	0.5768	112.7116
Genova	13	13	30	0.7105	0.4179	102.2049
La Spezia	25	30	33	0.376	0.3566	102.052

Table 8 Cograduation matrix among the methods applied on resident population indicators

Ranks	MZ	MR	MPI-	Average
MZ	1.0000	0.9980	0.9866	**0.9949**
MR	0.9980	1.0000	0.9862	0.9947
MPI-	0.9866	0.9862	1.0000	0.9909

Table 9 Correlation matrix among the methods applied on resident population

Values	MZ	MR	MPI-	Average
MZ	1.0000	0.9988	0.9988	**0.9992**
MR	0.9988	1.0000	0.9983	0.9990
MPI-	0.9988	0.9983	1.0000	0.9990

The application of different weighting criteria leads to very divergent results. The analysis according to the road infrastructures allows purifying a component of mobility of the phenomenon. The seasonal factor due to a more objective measurement also improves the concept of exposure to the risk of being involved in a traffic accident.

This first result leads us towards the right direction to relate road accidents to traffic flows for a correct measurement of the phenomenon. The 0.5079 cograduation value between Road and Population ranking and also the 0.6006 between Road and Fleet ranking can be considered as an improvement towards the correct risk to exposure for road users and useful to give further information to stakeholders in order to assess prevention actions in specific and localized infrastructure.

It was essential to start from the knowledge of the registry of the national road graph at localities level to reach soon the final indicator of “vehicles per kilometer” per road arch.

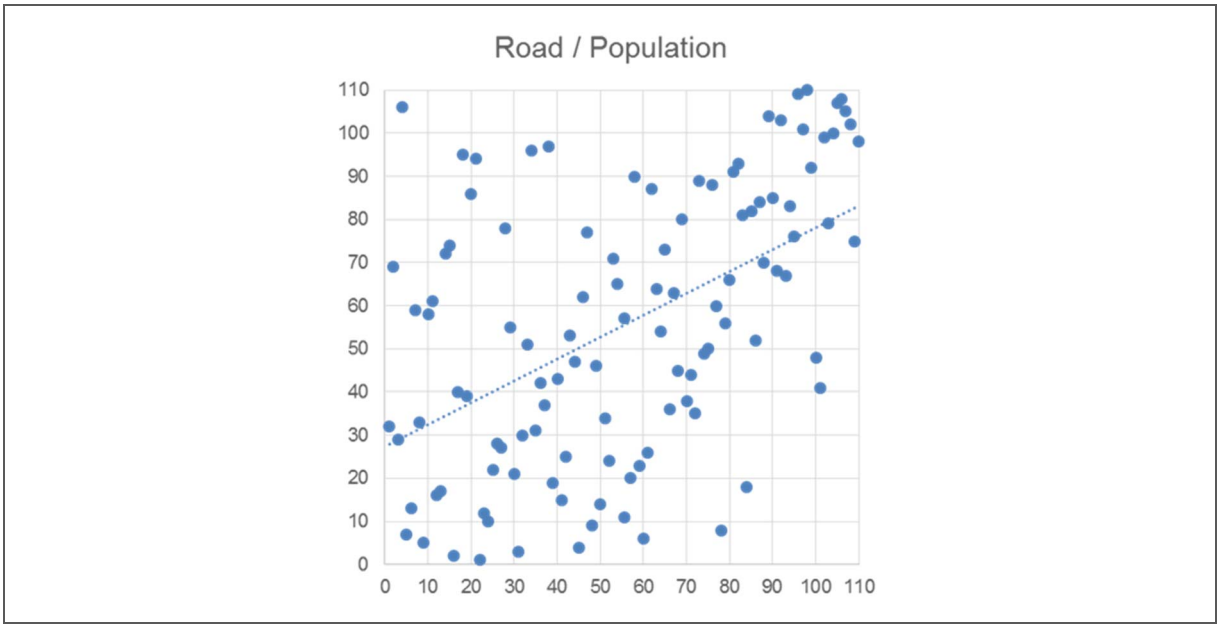
Table 10 The 5 best and worst provinces on the ranking distributions

Rank	MZ Road	MZ Fleet	MZ Population
1	Milano	Savona	Savona
2	Monza e della Brianza	Bologna	Ravenna
3	Roma	Piacenza	Forli
4	Napoli	Genova	Piacenza
5	Firenze	Forli	Bologna
...	...	...	...
106	Carbonia - Iglesias	Carbonia - Iglesias	Napoli
107	Crotone	Caltanissetta	Benevento
108	Oristano	Benevento	Carbonia - Iglesias
109	Isernia	Aosta	Caltanissetta
110	Ogliastra	Agrigento	Agrigento

Table 11 Cograduation matrix among the indicators of road length, vehicle fleet and resident population

Cograduation	Road	Fleet	Population
Road	1.0000	0.6006	**0.5079**
Fleet	0.6006	1.0000	0.9432
Population	0.5079	0.9432	1.0000

Chart 7 **Plot of ranking position of the Italian province among Road and Population ranking**



5.3.4 Lesson learnt

The length of the road network gives for sure a consistent first set of information concerning the different territories. The first output of the project, in line with the process of modernization of Istat statistical production, is the focus on the exploitation of existing administrative sources, the scouting of new sources and the analysis of integrated and auxiliary data. The basis of the renewal in the statistical production is to upload any source integration; even any new technique implemented and applied methodology. Every small change that overall effect becomes a process of improvement of the quality of the statistical information provided by Istat.

The final aim of this project is to be able to reach the measure of "daily average theoretical vehicles" that give the statistical information of the estimated number of vehicles that theoretically every day (of the month, of the year or of the period considered) travel the entire network, road or road arch considered.

The computational process is obtained as a relationship between kilometres travelled on the section under examination (in a month, in the year or in the reference period) and the length in kilometres of same stretch multiplied by the number of days; It's a quantitative measure of the extent to which the network is used, of the motorway or section concerned. The next step of the process is to identify the road arcs involved in heavy traffic intensity (POTs) by implementing the construction of new synthetic indicators.

The next step of the process is to identify the road arcs involved in heavy traffic intensity (POTs) by implementing the construction of new synthetic indicators.

5.4 Social Mood on Economy Index

5.4.1 Objective

Nowadays, more and more people all over the world use social media platforms to express their feelings and ideas, as well as to share or debate opinions on virtually every conceivable topic. As a consequence, the interest towards social media as a means of 'measuring' public's mood continues to increase significantly.

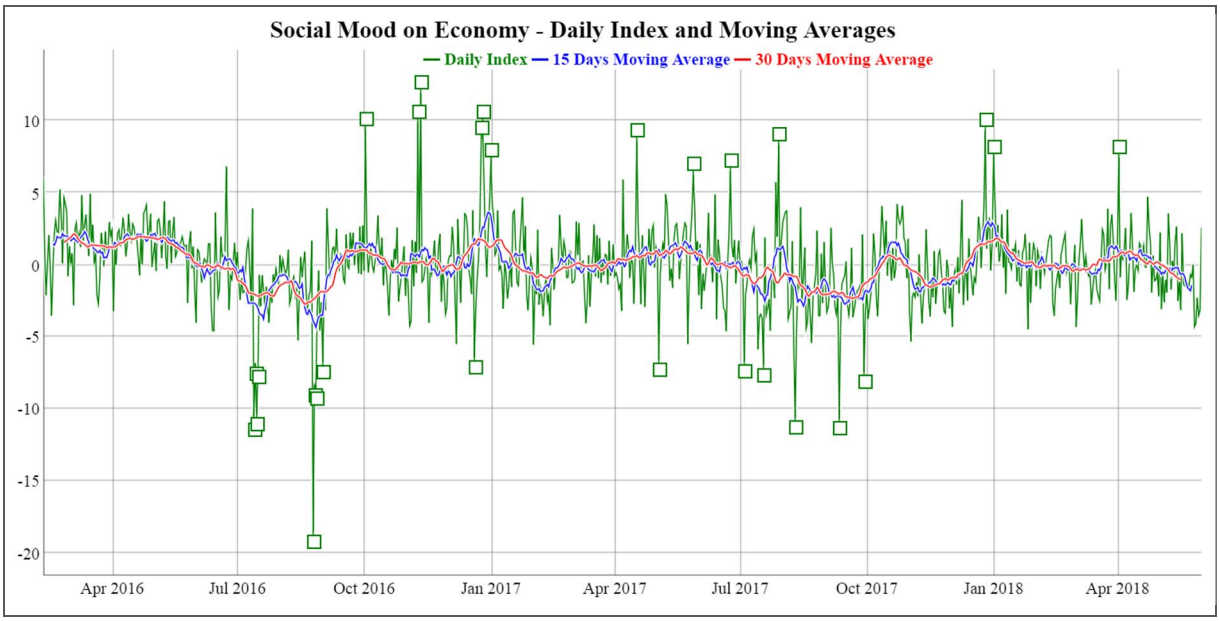
The Italian National Institute of Statistics (Istat) is currently investigating whether social media messages may be successfully exploited to develop *domain-specific* sentiment indices, namely statistical instruments meant to assess the Italian mood about specific topics or aspects of life, like the economic situation, the European Union, the migrants' phenomenon, the terrorist threat, and so on.

To this end, Istat researchers have developed procedures to collect and process only social media messages containing at least one keyword belonging to a specific *filter*, namely a definite set of relevant Italian words. Domain-specific filters have been designed by subject-matter experts with the aim of filtering out since the beginning messages that would very likely turn out to be off-topic for the intended statistical production goal.

5.4.2 Results

Istat provides here a preview of its experimental *social mood on economy* index, which is still in progress. At the moment the index only uses Twitter as a source, but further social media might be taken into consideration in the future. The new statistical instrument has been devised to enable high-frequency (e.g. daily) measures of the Italian sentiment on the state of the economy like those in the following chart A (see also the dynamic graph).

Chart A **Time series of the Social Mood on Economy Index**



This interactive plot shows the daily time series of the *social mood on economy* index (green line), along with the corresponding 15-days (blue line) and 30-days (red line) moving averages. The index is normalized so as to take values in [-100, 100]: the higher the value, the better the mood. Significant peaks and valleys of the daily index have been annotated and are highlighted with a small square: just hover the mouse over the square and a tooltip will display the dominating topic(s) emerging from the tweets of the day (when several topics are listed, they are sorted by decreasing prevalence).

Topics characterizing notable valleys in the daily index can be classified into two broad groups: 'disasters and terrorism' and 'welfare and economy'. Examples of the first group are the Central Italy earthquake (24th August 2016), the Livorno flood (10th September 2017), the Andria-Corato train collision (12th July 2016), the truck attack in Nice (14th July 2016) and the Christmas market attack in Berlin (19th December 2016). The second group involves debates and worries about poverty (14th July 2016), women's retirement regulation (9th August and 28th September 2017), youth unemployment (31st August 2016 and 3rd July 2017) and general unemployment (2nd May 2017). As for the peaks of the index, they are either linked to holydays or, again, to welfare and economy. Remarkable examples include school teachers' expectations about the removal of the triennial mobility constraint introduced by law 107/2015 (8th and 10th November 2016) and the debate triggered by Pope Francis' claim that the goal of universal employment should be considered superior to the one of universal income (27th May 2017).

5.4.3 Methods and data processing pipeline

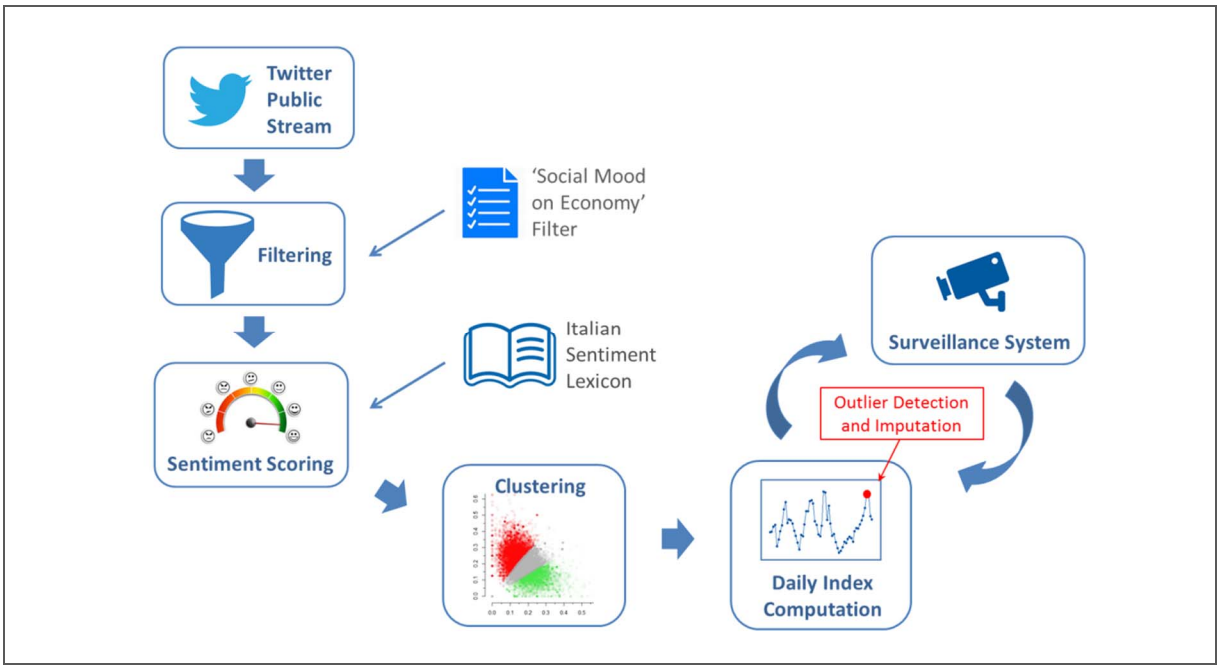
Twitter's Streaming Application Programming Interface (API) is used to collect samples of public tweets matching a filter made up of 60 relevant keywords (some are actual words, some are phrases). A subset of these keywords has been borrowed from the questionnaire items of the Italian Consumer Confidence Survey, a monthly sample survey that collects data in the first two weeks of each month and disseminates results by the end of the month. However, the phenomenon tracked by the *social mood on economy* index and consumer confidence only partially overlap. Still, the index can detect and promptly point out events that influence consumer confidence but occur after the interview period of the Consumer Confidence Survey: the Central Italy earthquake of 24th August 2016 is a striking example.

To compute daily index values, all the tweets collected in a single day (about 40,000, on average) are processed as a single block. Messages are first cleaned and normalized, then undergo a sentiment analysis procedure. For each tweet, positive and negative sentiment scores are calculated. To this end, message words are matched against entries of an Italian Sentiment Lexicon, namely a vocabulary whose lemmas are associated to pre-computed sentiment scores. Atomic scores of matched words are then averaged to yield tweet-level scores. Subsequently, tweets are clustered according to their sentiment scores into three mutually exclusive classes: Positive, Negative and Neutral tweets. Lastly, the daily index value is derived as an appropriate central tendency measure of the score distribution of the tweets belonging to the Positive and Negative classes (see chart B).

Special care has been devoted to make the index robust against possible contaminations by off-topic tweets that might pass the filter. To this end, a surveillance system has been put in place, which continuously searches for anomalous values in the daily index time series by means of two independent and complementary outlier detection routines. Daily values detected as potential outliers cause the system to automatically generate a set of dedicated diagnostic reports. These are then sent to human reviewers in charge of deciding whether the detected values are actually proper data points

or truly anomalous. The latter case typically arises when an off-topic tweet that happened to pass the filter becomes "viral" on Twitter. Being re-twitted and quoted thousands of times in a day, viral tweets may have an unduly impact on the daily index and introduce bias. As a consequence, all the daily index values classified as truly anomalous are eventually imputed via linear nearest-neighbour interpolation.

Chart B A schematic representation of the processing pipeline of the Social Mood on Economy Index



The *social mood on economy* index is solely based on samples of public tweets. The index uses only unlinked anonymized data. No selection of Twitter's users is made during sampling and only the textual content of collected tweets is processed. Upon collection, tweets are never linked to Twitter's users: Istat does not know who has sent them. Daily index values result from the aggregation of numeric scores attached to tens of thousands of messages. This is an irreversible process: no tweet can be reconstructed by analysing the index.

5.4.4 Lesson learnt and future work

The *social mood on economy* index is still under development. Istat researchers are currently exploring more sophisticated natural language processing techniques and alternative approaches to sentiment analysis. Further work will be devoted to the study of the time evolution of the index. For instance, it will be investigated whether the index correlates with, or even can help to predict, other economic indices that are officially released by Istat at low-frequency (i.e. on a monthly, quarterly or annual basis). The hope is that the *social mood on economy* index could either improve the performance of Istat's forecasting models, or enrich existing statistical products (e.g. the Equitable and sustainable well-being - BES), or even be disseminated as new statistical output in its own right.

CHAPTER 6
COMMENTS OF BIG DATA COMMITTEE MEMBERS ON MAJOR BIG DATA ISSUES

Some of the BD Committee members have been interviewed on issues relating to the use of Big Data for production in Official statistics. On the webpage, thus, some short videotape contributions of eminent Italian and foreign scholars expressing their opinion on relevant issues related to the use of BD are available.

Carlo Batini, professor of Computer Sciences, Systems and Communications Department at University of Milano Bicocca, elaborated on the Perspectives of computer science research and the relevance of semantics.

Piet Daas, senior methodologist and data scientist Statistics Netherlands (CBS), reflected on the importance of partnerships in the experience of CBS Centre of BD statistics.

Michail Skaliotis, head of Task force on Big Data at Eurostat, presented the perspective of the use of BD for advances in Smart statistics.

Gero Carletto, lead economist and manager of the Living Standards Measurement Study and Development Data Group at World Bank, explained the role of Big Data for the future of Sustainable Development Goals (SDGs).

Monica Pratesi, president of Italian Statistical Society, explained the Inferential aspects of estimates with Big Data.

Stefano Iacus, professor of Statistics at University of Milan, elaborated on the challenges in the analysis of unstructured textual data.