Trusted smart statistics: the challenge of extracting usable aggregate information from new data sources

Maurizio VICHI[1] and David J. HAND[2]
[1] *Sapienza University, Rome*
[2] *Imperial College, London*

1. Introduction

The amount of data being produced and stored is growing at an exponential rate, doubling every two years or so in terms of volume, with a great variety and velocity of production. Much of this dramatic growth is attributed to changes in styles of communication due to the rise of technology and the increasing use of the internet. In particular two new forms of communication have developed: human Computer-Mediated-Communication (CMC) and the Internet of Things, (IoT) a non-human CMC.

In CMC (Herring, 1996) any human communication occurs through two or more "smart devices" having specific formats, such as instant messaging, email, chat rooms, online forums, social network services and text messaging. Communication with CMC, in contrast to the face-to-face human conversation which happens in real time (synchronous), can also be asynchronous, with parties not communicating at the same time. The by-product of CMC is the "*datafication*" (Mayer-Schoenberger and Cukier, 2013) of different aspects of the life of citizens; in fact, social media platforms tend to broadcast continuously and show communication among people on all possible topics and phenomena relating to their lives. Thus, society collectively accumulates data on massive amounts of its behaviours. If these processes are considered as an ecosystem, these data –denoted "organic data" – reflect a natural feature of this ecosystem (Groves, 2011). Thus, social actions and communications are transformed into online-quantified organic data, which can be tracked in real-time. For example, Research-Gate datifies the research network, Twitter produces a datafication of social topics, LinkedIn datifies professions, while YouTube digital datifies different behaviours of the lives of people. According to the ESS Task Force on Big Data and Official

Statistics (2018), most data in the course of the third decade of the 21st century is expected to be "organic".

The European Research Cluster on the Internet of Things (IERC) states that IoT is : a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual “things” have identities, physical attributes, and virtual personalities and use intelligent interfaces, and are seamlessly integrated into the information network. Thus, IoT produces communication between things and a datafication of actions and services for which they are used. Thus, also the process of datafication induced by IoT can be considered as an ecosystem, which produces organic data.

The automatic electronic data capture-measurement and recording which does not require human intervention induced by CMC and IoT produces data with a great velocity, variety and volume. But organic data and information/knowledge are different. To extract something useful from the data, sophisticated statistical methodologies, and algorithms, and other tools are required. Moreover, the velocity of the flow of incoming data often means that this needs to be done in real time. The integration of the data with the statistical methodologies necessary to transform these data into information/knowledge produces what can be called “smart statistics”. These provide a rich opportunity for producers of official statistics. They have the potential to yield deeper insights, more quickly, and with greater accuracy than traditional methods.

This paper looks at the potential implications of smart statistics in connection with big data, new data sources, and new ways of processing data to produce official statistics. In particular, it is concerned with the changes we are certain to see in the production and use of official statistics in the near future, as well as with what needs to be done to protect the trustworthiness and reliability of such statistics in the light of recent developments in statistics, machine learning, data mining, artificial intelligence, and related technologies.

In the next section we give a more detailed definition of smart statistics, followed by a closer look at the characteristics of the new types of data and technologies for manipulating them. For the potential to be achieved, various obstacles need to be overcome. After briefly looking at the potential benefits we examine those obstacles, making specific recommendations for how they might be tackled. A concluding section makes a final recommendation.

2. Smart statistics

The term "smart" is used increasingly widely to signify that a system has a measure of autonomous processing (or "intelligence") built into it. Familiar examples of smart systems are: "*smart cities*" to monitor and manage urban mobility, accessibility, waste management, hospitals and other community services; "*smart cars*" with advanced control systems to perceive surroundings and identify appropriate navigation paths as well as obstacles and relevant signage; "*smart farming*", for modern agriculture; "smart factory and industry 4.0" for automation in manufacturing technologies; "smart buildings" for building automation.

In the context of official statistics, "smart statistics" have characteristics which include:

- *large*, often very large, *raw data sets*;
- *raw data from variety of sources* including: the IoT; the CMC provided by the world wide web, social media; administrative sources; and private sector data;
- the *processing* of the data is *autonomous* and *automated* and *data-driven*;
- the *system* extracting the data will often be *interactive* and *adaptive* to novel requests;
- *statistics* can be *extractable in real time* or in as close to real time as makes sense;
- often some aspects of *data processing* will be *embedded* in the *data capture*.

The above suggests that the phrase "smart statistics" can be interpreted to mean both a *technology* including sensors or methods that allow information to be autonomously collected, tracked, processed across local and global network infrastructures and a *numerical summary* of the datafication of the reality, just as the word "statistics" itself can refer to the discipline and to specific numerical summaries.

It is important to observe that smart statistics implies the description of reality from a modern multivariate viewpoint. This means that many different aspects of the reality are observed to correctly describe it (e.g., in smart system such as an "autonomous car" several sensors such as radar, laser light, GPS, computer vision, produce different data used simultaneously to better perceive reality). In addition, data necessary for smart statistics must include time and geospatial references in order to analyse trends and convergences over time and evaluate the diffusion of the observed reality in the territory.

While the same principles apply more generally, this paper is restricted to official statistics - statistics aimed at understanding and improving society. It does not discuss, for example, statistics aimed at improving or informing a commercial operation, such as insurance decisions or marketing strategies.

For official statistics (smart or otherwise) to be useful, they must have a number of characteristics. Characteristics listed by National Statistical Institutes (NSIs) often include relevance, timeliness, consistency, coherence, availability, and accessibility, but we can add others, some of which describe the data collection and processing stages, such as transparency, trustworthiness, accuracy, security, and confidentiality.

The UK Statistics Authority's Code of Practice (UKSA, 2018), for example, is based on three pillars, each which is split into several more detailed desiderata.

- trustworthiness: confidence in the people and organisations that produce statistics and data;
- quality: data and methods that produce assured statistics;
- value: statistics that support society's needs for information.

For smart statistics, the pillars which present the most challenges will be the trustworthiness and quality pillars, since by definition smart statistics provide official bodies with less control over the raw data than do traditional official statistical sources.

In a similar vein the European Statistics Code of Practice (EU, 2011) is based on fifteen principles, amongst which are (Principle 4) a commitment to quality, (Principle 5) statistical confidentiality, (Principle 7) sound methodology, (Principle 8) appropriate statistical procedures, (Principle 12) accuracy and reliability, and (Principle 14) coherence and reliability.

Principle 14 is particularly noteworthy here since, as we shall see, it is a characteristic of smart statistics that the constituent data are often collected for purposes other than official statistics and that they come from a variety of not necessarily compatible sources. This presents novel technical challenges relating to transparency and auditability, as well as those of ensuring consistency.

Traditionally, official statistics are calculated by a National Statistical Institute from the raw data. However, embedding of processing in data capture and measurement instruments is becoming more important for two reasons. One is simply the fact that it can be done: advanced microelectronics means that preprocessing to extract signals can be done at source. For example, rather than sending raw details of radar signals, traffic monitors can process the data to count the numbers of vehicles passing a point and transmit only this count. The second reason is that it is often necessary to reduce the sheer volume of data to be transmitted. This is best accomplished by extracting the relevant signals at the earliest possible point. Having said that, one must recognise that pre-processing does mean sacrificing data, and that other aspects of the data might be useful signals for different questions.

3. New data sources and new data technologies

Official statistics are statistics collated by government agencies or other public bodies to provide a resource for governments and citizens to assist in understanding, decision-making, and planning. They cover all aspects of life, including economic, social, health, education, commercial, and so on. In the past, collecting the raw data from which to provide official statistics has been a slow and expensive process, involving exercises such as elaborate face-to-face and telephone surveys and business surveys which need to be manually completed. Increasingly, however, alternative data sources enable information to be collected without requiring any additional effort on the part of those providing the data - apart, perhaps, from giving permission that the data may be used for further purposes.

Examples of these alternative data sources include:

- *web-scraped data*. This typically describes an automated search through websites and the extraction and downloading of information from relevant websites. Web-scraping, as with all cutting-edge data technologies, is not without its legal questions. In this case they often relate to copyright issues.
- *administrative data*. These are data generated during the course of some administrative exercise - tax returns, health records, school tests, credit card transactions, supermarket purchases, etc - which are then stored, and which can later be used in constructing official statistics.
- *social media data*. This normally describes data collected from individuals’ interactions through platforms such as LinkedIn, Facebook, and Twitter.
- *machine generated (IoT) data*. This covers things such as automatic sensors. At a low level they might be accelerometers (e.g. as used in detecting when a car has driven over a pothole), gyroscopes (showing changing directions), pressure gauges, temperature gauges, and so on. At a higher level automatic sensors will give indications of weather conditions, vehicle congestion, crowds, crop condition, and so on.

Different sources of data have different properties. Indeed, as we discuss below, since the strengths and weaknesses of new data sources differ from those conventionally used in creating official statistics, smart statistics based on these new data sources present new kinds of challenges in ensuring reliability and trustworthiness.

In addition to new data sources, we are seeing a wider range of data *types*. Traditionally, official statistics will be based on numerical raw data (translating categories into counts, for example). But modern data capture technology means we are now also faced by text data, image data, signal data, and other types. Each of these present its own technical challenges.

To complicate things even further, great gains in understanding arise from *linking* data sets. For example, the UK's Administrative Data Research Network (ADRN, 2018) is a consortium which links data from UK government departments for the purposes of social and public policy research. But data linkage presents challenges of its own (Herzog *et al*, 2007; Christen, 2012). Even for conventional numerical databases, these include deciding on how to measure similarity between records, deduplication, variation in the way data are recorded, and different databases containing intersecting but different subsets of a population. With data of different types (e.g. combining hospital records with text data from clinical trial reports) the challenges are that much greater.

The aim of official statistics is to *describe* rather than predict. Official statistics summarise and aggregate data sets, and fit hypothesised models to data. In the context of smart statistics this may involve real-time analysis, and it might also be interactive in response to changing user requests, but the key point is that this means that the core technology will be statistics and not, for example, machine learning, which is fundamentally concerned with prediction. This matters because the predictive aim of machine learning means it emphasises *algorithms* (descriptions of how to process data) rather than *models* (descriptions of the structures that data take). Of course, some of the adaptive and interactive processes which are implicit in various machine learning algorithms will find a use in adaptive estimation, updating estimates, and interactive smart statistics systems.

Technological advance, be it from exciting new data sources or powerful new computing tools, takes place within a social context. This means that overlaying all of the above is the need to conform to legal and regulatory requirements, such as the General Data Protection Regulation, which came into force in May. Such regulations are concerned with the right to access personal data, to have it corrected, and to control how it is used. Privacy preservation, and associated tools such as anonymisation and pseudonymisation, while important to official statistics are possibly less critical than in operational domains (e.g. running a bank or hospital) because the fundamental aim in official statistics is to make statements about aggregates rather than individuals. Overall, however, the concept of "trustworthiness" in official statistics relates not merely to their accuracy, but also to their ethical nature.

The process of production of smart statistics comprises some additional phases beyond those of the classical production process of official statistics. Fundamental is the *design* of the smart system necessary to describe the reality under study. This includes the choice of methods for automatic data collection and the technologies of CMC and IoT to be used in order to guarantee large raw data sets, the raw data variety of the information (multivariate) and the autonomous and automated data processing. The data analysis phase has to guarantee the adaptive characteristic of the smart statistics, by using a statistical learning approach (Hastie, Tibshirani Friedman, 2009) where a data-driven statistical model is used to describe reality in a training data. In supervised learning the model can be a classification or a regression, in unsupervised learning, the model is a clustering producing a reduced number of prototype observations and/or a dimensionality reduction defining synthetic indicators (composite, factors) measuring complex concepts (e.g., sustainability).

4. Benefits

There are many potential benefits to smart statistics based on new sources of data including:

- *granularity*. Alternative data sources are often very extensive. In principle (though as discussed in the next section, the practice may be different) administrative transaction data capture all the information involved in transactions, and web-scraped data can be as extensive as you like. This means that it is possible to make statements about smaller constituencies (smaller groups of people, smaller geographical regions, and so on) than would be possible with a much smaller sample.
- *timeliness*. In principle again, data can be captured essentially as it is generated - as phone calls take place, as internet purchases are made, as prices are changed on the web, and so on. This means that financial and social indices can be updated in real-time.
- *proximity to social reality compared with conventional data acquisition methods*. Since the new data sources monitor actual behaviour, summary statistics based on them tell us how people behave, not how they say they behave. Indeed the shortcomings of conventional data collection strategies was graphically revealed by a recent study showing that people underestimate the number of calories they consume by about a third (Bailey, 2018).
- *responsiveness to changing questions and circumstances.* If vast masses of very low level data are stored, then statistics aimed at follow-up questions can be readily calculated.

- *cost and effort reduction*. If the data are collected automatically, no addition effort (e.g. no survey) is necessary. A compelling example of this is the shift towards censuses based on administrative data rather than elaborate specific data collection exercises(e.g. ONS, 2018)

One of the most compelling illustrations of the power of smart statistics is the case of inflation in Argentina over the period from 2007 to 2011. Cavallo (2013) describes how inflation indices for Brazil, Chile, Colombia, Venezuala, and Argentina were calculated based on web-scraped data and compared with the official estimates. In the first four of these "online price indexes approximate both the level and main dynamics of official inflation. By contrast, Argentina's online inflation rate is nearly three times higher than the official estimate."

5. Risks and resolutions

The novel data sources and novel technologies implicit in smart statistics necessarily mean that such summaries come with risks. The development of smart statistics must go hand in hand with an understanding of the risks, and the development of strategies and tools to alleviate them.

5.1 Data quality

Quality is a perennial issue for all data analysis. Every practicing statisticians has their own horror stories about misleading statistics based on faulty data. Massive data sets and elaborate collection and processing systems have aggravated the potential risks. While one might manually examine each of a thousand data points, doing so for a billion points is infeasible. Furthermore, algorithms always give an output, regardless of the quality of the data fed in, and regardless of whether that output is meaningful or not.

Lack of control over the raw data makes things even worse. Data collected for an administrative purpose might be suitable for that purpose, but less so for deriving official statistics. Data collected from social media interactions or by web-scraping or by computer-assisted web interviewing (Hand and Vichi, 2018) are likely to have all sorts of unspecified selection distortions. Even automatic electronic measuring instruments can give faulty data - if they become detached, drift out of alignment, or fail, for example. Verification of the accuracy and validity of raw data is likely to be increasingly difficult, or even impossible, as we move into the smart data world.

Yet another layer of complication arises when data from multiple different sources are linked, merged, or fused. Each source is likely to have its own quality issues, and these will be compounded, not diluted, when data sets are combined, not least because of almost inevitable inconsistencies between data from different sources. And yet further complications arise from differences in definitions.

> *Recommendation 1:* In parallel with the introduction of each smart statistic must be a protocol for checking the veracity, authenticity, and accuracy of the raw data.

> *Recommendation 2:* Triangulation methods must be developed, whereby smart official statistics are derived via several different routes, so that consistencies and comparability of definitions can be checked.

5.2 Data provenance

Since smart statistics will often be based on data not under the statistician's control, trust in the data requires that its provenance and the meta data describing it are documented in great detail. This requires noting the nature, type, model, version, etc, of any measuring instrument, questionnaire, recording device, and so on used in the capture of the data. Moreover, all data undergo cleaning and checking prior to analysis - indeed, often this is the major part of the effort leading to a statistical conclusion. In order to have confidence in the data, and that the data represent what they purport to, every step and modification made to the raw data must be noted. In many situations, earlier versions of the data should be stored, although this will not always be possible (e.g. in dynamic situations - but then accurate time stamps must be recorded so that the path through an analysis could be reconstructed).

Such recording practices will not be undertaken in the expectation that analyses *will* be re-run, but rather so that they could be re-run if necessary. Without this one can have no confidence that the conclusions genuinely derive from the data. These practices are necessary for transparency, accountability, and trustworthiness.

> *Recommendation 3:* Any data used to produce smart statistics should have its origins and associated metadata carefully described.

> *Recommendation 4:* All data modifications and changes should be noted, along with their reasons, so that earlier versions of data sets can be derived.

5.3 Sustainability

One of the important aspects of current official statistics is that they are consistent over time. If the same methods have been used to collect the data and the same methods used to produce a summary statistic over a period of years, then time series can be produced. These can show how things are changing, as well as how they have behaved in the past. This is useful not just for historians, but also to those tasked with formulating current policy.

Unfortunately, it is almost a necessary consequence of the definition of smart statistics that they lead to short time series, because the data collection methods change. Web-scraping relies on the structure of the web and the algorithms used to identify web sites (the Google search algorithm changes regularly); administrative data relies on the operation which generated the data being maintained over time (banking systems evolve or are upgraded); social media strategies and tools seem to change on almost a daily basis (think Bebo, Myspace, Facebook, Snapchat, Twitter, ...). Apart from changes for technological reasons, there are also commercial reasons (more attractive competitors knock out a business), regulatory reasons (often to do with privacy: at the time of writing Facebook is in the news, with its CEO having appeared before Congress in the United States), and simple societal reasons (users' preferences change).

Unfortunately, this sustainability risk is the opposite side of the coin of the benefits of smart statistics: we cannot have the latter without the former. Nonetheless, careful monitoring and recording of changes of definitions and methods where this is possible can alleviate the problems.

> *Recommendation 5:* Insofar as it is possible, consistency checks should be made over time. These should include checks of data analytic methods and of definitions. When time series show a sudden change in behaviour (level, trend, volatility) efforts should be made to understand why the changes have occurred, and in particular to provide assurance that it is not due to changes in the raw data or collection system.

5.4 Infrastructure breakdown and complexity crash

There is another important challenge which has arisen only rarely in the context of official statistics before, but which is certain to occur as we progress into the world of smart statistics. This is the vulnerability of official statistics to infrastructure breakdowns and hacking. We have seen this in other contexts - in banking systems, health records, electoral system, and others. It will become more relevant to official statistics to the extent that those are built on real-time data acquisition and fast delivery of statistics.

The importance of economic indicators means that a more sophisticated strategy for both preventing and recovering from systems failure is needed. We need to learn from the work on “safety critical systems” which has been carried out in other domains, such as nuclear engineering, aviation, and medicine. Having redundant systems is a basic notion.

It is a characteristic of smart statistics that the underlying systems they are monitoring and the data sets they are accessing are complex - the opposite of a simple survey designed to answer specific questions. Complex systems have their own vulnerabilities because of their fundamental interconnectedness and nonlinearities (see, for example, Charles Perrow’s classic work *Normal Accidents: Living with High-Risk Technologies*: Perrow (1999)). It is in the nature of these vulnerabilities that they are unseen and unexpected, but we can guard against their consequences by having robust backup and recovery plans.

> *Recommendation 6:* Regular backups should be made and fall-back systems should be in operation so that urgent official statistics will not be substantially delayed. Learn from the work on safety critical systems.

5.5 Privacy risk

While the risks to privacy arising from official statistics might not be as sharp as those arising from the collection and manipulation of data for operational purposes (e.g. smart meter data revealing when you are likely to be at home), they still exist. After all, large masses of data are collected and stored, and this is often personal, such as financial and medical data. However, once datasets have been linked there is no reason to preserve the identifying features - after all, official statistics are about aggregations, not individuals. Sophisticated methods, such as the Trusted Third Party system used by the Administrative Data Research Network, have been developed to protect individual privacy prior to and through the linkage stage.

Given the above, it is perhaps surprising that individuals are more willing to divulge data to commercial operations than to government. One possible explanation is that the gain they receive in exchange for the data is more immediate and obvious: they gain access to the social network or some other service. In contrast, the gains from giving the data to official statisticians are less obvious: more effective schools, better hospitals, a more efficient economy and so on do not directly and obviously link to you telling a statistician how old you are, your medical history, and your education.

> *Recommendation 7:* An organisation is trustworthy only to the extent that it protects the privacy of its users: methods to ensure the privacy and safety of data used to produce smart statistics must be put in place.

6. Conclusion

Smart statistics based on new data sources and new methods of analysis have properties not possessed by more traditional official statistics - such as timeliness and adaptability. If alternative suppliers of statistics provide products which do have these properties then any reservations about other aspects - perhaps less confidence in accuracy, or that the statistics match a formal definition - must be balanced against them. Put bluntly, if producers of official statistics do not respond by adapting in terms of speed, timeliness etc etc, if they do not produce smart statistics, then they risk being replaced by other suppliers who might have a less reliable product.

At bottom, the particular strength of National Statistical Institutes is their trustworthiness. This is a primary factor distinguishing them from alternative sources which might claim to produce comparable indices: commercial financial operations producing inflation indices, estate agencies producing house price indices, insurance companies producing population statistics, and so on. Such organisations might have reservations about divulging their data and methods - understandably, since their business edge might be based on proprietary sources and tools.

We suggest that trusted official statistics are critically important in the so-called “post truth” era, where appeals to emotion risk undermining objective facts. Effective policies must be based on ground truths, not wishful thinking. In turn, to be trusted, statistics must be trustworthy, meaning that that the raw data and the statistical methods applied to those data must merit trust. And as a final step, to merit trust the data and methods must be transparent: observers must be able to check and verify this trustworthiness. This leads to our final recommendation:

Recommendation 8: An independent oversight body or regulator should be created to resolve disagreements, to look into failures, etc. This body should have the power to explore the sources of all data which are fed into the production of official statistics.

7. References

[1] ADRN (2018) https://www.adrn.ac.uk/research-impact/research/ Accessed 5th May 2018.

[2] Bailey R. (2018) https://datasciencecampus.ons.gov.uk/2018/02/15/eclipse/ Accessed 5th May 2018.

[3] Cavallo A. (2013) Online and official price indexes: measuring Argentina's inflation. *Journal of Monetary Economics*, **60**, 152-165.

[4] Christen, P. (2012) Data Matching—Concepts and Techniques for Record Linkage,Entity Resolution, andDuplicateDetection. Data-Centric Systems and Applications. Springer, Berlin.

[5] EU (2011) http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15 Accessed 4th May 2018.

[6] Groves R. (2011) "Designed Data" and "Organic Data", Director's Blog, United States Census Bureau, https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html

[7] Hand D.J. (2018) Statistical challenges of administrative data. *Journal of the Royal Statistical Society, Series A*, **181**, 1-24.

[8] Hand D.J. and Vichi M. (2018) The role of communication in statistical science and the strategies of communication for statistics users. To be presented at the ESS DGINS Conference, Bucharest, 10-11 October.

[9] Hastie T, Tibshirani R., Friedman J. (2009) The Elements of Statistical Learning, Springer-Verlag.

[10] Herzog, T., Scheuren, F.,Winkler,W.E. (2007) Data Quality and Record Linkage Techniques. Springer, Berlin.

[11] Mayer-Schönberger, V., & Cukier, K. (2013) Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

[12] ONS (2018) *Administrative Data Census Project*. https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject Accessed 5th May 2018.

[13] Perrow C. (1999) *Normal Accidents: Living with High-Risk Technologies*. Princeton University Press, Princeton.

[14] UKSA (2018) https://www.statisticsauthority.gov.uk/wp-content/uploads/2018/02/Code-of-Practice-for-Statistics.pdf Accessed 4th May 2018.