United Nations
Economic Commission for Europe
Statistical Division

CONFERENCE OF EUROPEAN STATISTICIANS

WP.18

Workshop on Statistical Data Collection: Riding the Data Deluge

24 April 2015

29 April – 1 May, Washington D.C., United States of America

Structuring risks and solutions in the use of big data sources for producing official statistics – Analysis based on a quality framework

Albrecht Wirthmann (Eurostat)

Abstract

An increasing number of statistical offices are exploring the use of Big Data sources for the production of official statistics. For the time being there are only a few examples where these sources have been fully integrated into the actual statistics production. Consequently, the full extent of implications caused by their integration is not yet known. Meanwhile, first attempts have been made to analyse the conditions and impact of Big Data on different aspects of statistical production such as quality or methodology. A recent task team elaborated a quality framework for the production of statistics from Big Data in the context of the Big Data project of the United Nations Economic Commission for Europe (UNECE). According to the European Statistics Code of Practice the provision of high quality statistical information is the main objective of statistical offices. Since risk is defined as the effect of uncertainty on objectives (e.g. by the International Organization for Standardization's ISO 31000) we have found it appropriate to categorise risks according to the quality dimensions they affect. The suggested quality framework for statistics derived from Big Data sources provides a structured view of quality related to all phases of the statistical business process and thus may serve as basis for a comprehensive assessment and management of risks related to these new data sources. It introduces new quality dimensions that are specific to or (of high importance when) using Big Data for official statistics, such as institutional/business environment or complexity. Using these new quality dimensions it is possible to derive risks related to the use of Big Data sources in official statistics in a more systematic way.

In the present paper we aim to identify risks induced by the use of Big Data in the context of official statistics. We follow a systematic approach of defining risks in the context of the suggested quality framework. Concentrating on the newly proposed quality dimensions we are able to describe risks that are currently not present or do not have an impact on the production of official statistics. At the same time we are able to identify current risks that will be evaluated very differently when using Big Data for producing statistics. Then we go further into the risk management cycle and provide an assessment of likelihood and impact of these risks. As the assessment of risks involves subjectivity in attributing likelihood and impact to the different risks we measure the agreement between the scores of different stakeholders given independently. Then, we propose options for reducing these risks according to the four major categories avoidance, reduction, sharing and retention. According to ISO, one of the principles of risk management should be to create value, i.e. the resources for mitigating risks should be lower than for doing nothing. Following this principle, we finally make an assessment of the possible impact of some actions on risk mitigation on the quality of the final outputs to come to a more comprehensive assessment of Big Data usage for official statistics.

Structuring risks and solutions in the use of big data sources for producing official statistics – Analysis based on a risk and quality framework

Working Paper

Prepared by Wirthmann A, Karlberg, M., Kovachev B., Reis F., *Eurostat*[1]

1. INTRODUCTION

1.1. Background

The development of “big data” has been characterised by Kenneth Neil Cukier and Viktor Mayer-Schoenberger in their article “The Rise of Big Data[2]” by the term “datafication”. Datafication is described as a process of “taking all aspects of life and turning them into data.” E.g. Facebook datafies personal networks, sensors datafy all kinds of environmental conditions, smart phones datafy personal communication and movements, wearables datafy personal conditions. This is leading to a situation of almost ubiquitous collection and availability of data.

Like many other sectors, official statistics has only started recently to discuss the issue of Big Data at strategic level. There is not yet a common and widely shared understanding of the way forward, whether this is a challenge or opportunity, whether it is small or big etc. In the framework of the High-Level Group for the Modernisation of Statistical Production and Services[3], a first SWOT analysis, accompanied by a rough risk/benefit analysis, was conducted. It was noted that “a full-blown risk analysis would also include aspects such as likelihood and impact, and perhaps also be expanded to outline strategies to mitigate and manage risks”.

While still far from a complete risk analysis, this paper aims at improving matters precisely by establishing a first structured overview. We would like to stress that this overview is to be seen as a point of departure to stimulate a general discussion within the Official Statistics Community (OSC).

1.2. Scope

This paper focuses solely on **risks** – thus excluding not only benefits but also strengths, weaknesses, opportunities and threats. This means that “the risks of

[1] Any errors and omissions are the sole responsibility of the authors; the opinions expressed in this paper are personal and do not necessarily reflect the official position of the European Commission.

[2] http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data

[3] How big is Big Data? Exploring the role of Big Data in Official Statistics: http://www1.unece.org/stat/platform/download/attachments/99484307/Virtual%20Sprint%20Big%20Data%20paper.docx?version=1&modificationDate=1395217470975&api=v2

doing nothing" (for instance, the risk of the OSC to be out-competed with other actors if it doesn't modernise) are not in scope; these are rather threats. Instead we try to highlight the risks which might occur (a) if the OSC *acts* on the opportunities offered by Big Data and starts developing or improving a particular "Big Data based official statistics product" (BOSP); (b) the risks to the new "business as usual", i.e. risks to "Big Data based" official statistics production. (As all official statistics production is associated with risks, we limit (b) to "Big Data specific" risks, i.e. risks that are non-existent or negligible for "traditional" official statistics production.)

1.3. Structure

In Section 2, we present the frameworks related to this exercise, starting out with the obviously necessary framework for risks and risk management (Section 2.1). We also present the preliminary quality framework for statistics produced from big data (Section 2.2), since linking a quality framework to risks fulfils two objectives:

- It establishes the context for defining the risks. The defined quality dimensions together with the considered characteristics express values of an entity that are considered as important and decisive for delivering services to customers and users.
- It allows the assignment of specific risks to quality dimensions that are nested in the overarching hyperdimensions and are attached to specific phases in the production process of statistical products.

In sections 3, 4, 5 and 6 we present the risks identified so far in various contexts[4]. Here, we use the categorisation *Data Access*, *Legal Environment*, *Data Confidentiality and Security*, and *Skills*; a reorganisation according to a quality framework for statistics produced from big data (Section 2.2) should be considered as soon as that framework has reached a more final status. For each of the risks identified, we (i) provide assessments of likelihood as well as impact (as per section 2.1.3) and (ii) propose risk mitigation and management strategies (cf. section 2.1.4).

Finally, we discuss our findings and outline some next steps in Section 7

2. FRAMEWORKS

2.1. Risks and Risk Management

According to ISO 31000:2009[5], risk is defined as the "effect of uncertainty on defined objectives". This means that objectives have to be defined or be known before being able to define risks. These objectives are usually defined taking into account the institutional context of the respective organisation. Another important consideration is that risks bear the characteristic of

[4] The business case documents of the ESS Big Data project as well as on the Big Data ESSnets contain a list of risks partially related to the project and partially to using big data sources for statistical purposes. The document "A suggested Framework for the Quality of Big Data" mentions some risks related to quality dimensions.

[5] http://en.wikipedia.org/wiki/ISO_31000

uncertainty, i.e. it is not sure if the described event would occur. Risks are therefore measured in terms of probability of occurrence of an event and its consequences, i.e. the impact that the event has on achieving the defined objectives. Risk Assessment should produce more objective information which finally enables finding an appropriate balance between realising opportunities for gains while minimising adverse effects. Risk management is an integral part of management practice and an essential element of good corporate practice[6]. It is an iterative process that ideally enables continuous improvement in decision making and facilitates continuous improvement in performance.

Risks are also related to quality. The application of a quality framework should enable using opportunities provided by different sources and methodologies to achieve an output of a defined quality level in the sense that the output fulfils the needs of users. Like the risks, quality levels can be derived from an institutional environment and the objective of a certain institutions. In this context, the institutional environment defines the overall risk level that an institution is willing to bear for achieving its goals.

A risk assessment and management process can be broken into various steps that include establishing the context, the identification of risks, the analysis of risks in terms of likelihood and impact, the evaluation of the risks and finally the treatment of the risks.

2.1.1. Institutional Context

As a first step it is necessary to establish the strategic, organisational and risk management context in which the rest of the process will take place. This includes establishing criteria against which risks would be evaluated and defining the structure of the analysis.

2.1.2. Risk Identification

In the second step events should be identified that may have impact on the achievement of the defined objectives. The identification should include questions related to type of risks, timing of the event, location or how events could prevent, degrade, delay or enhance the achievement of the objectives.

2.1.3. Risk Assessment

The next step consists of determining existing controls and analysing risks in terms of likelihood as well as in terms of potential consequences. In the context of this paper, probability or likelihood of the occurrence of risks a scale from 1 (improbable) to 5 (frequent) is used. The impact of occurrence of events is measured using as well a scale from 1 (insignificant) to 5 (extreme). As illustrated in Table 1, the product of likelihood and impact renders a “risk level” ranging from 1 to 25.

Table 1: Risk Assessment

[6] Statistics Canada: 2014-2015 report on Plans and Priorities, http://www.statcan.gc.ca/about-apercu/rpp/2014-2015/s01p06-eng.htm

Impact Likelihood	1: insignificant	2: minor	3: major	4: critical	5: extreme
1: improbable	1	2	3	4	5
2: remote	2	4	6	8	10
3: occasional	3	6	9	12	15
4: probable	4	8	12	16	20
5: frequent	5	10	15	20	25

The estimated risk levels may be compared to predefined criteria in order to make up the balance between potential benefits and adverse outcomes. This enables judgements to be made about management priorities.

Table 2: Risk Index

Risk level	Risk Index
1 – 5	1-5: negligible (low priority or acceptable risk)
6 - 12	6-12: significant, tolerable (medium priority)
> 12	>12: critical, intolerable (high priority)

The priority for actions should be put on the critical risks (see Table 2), i.e. those which are likely to happen and have major to extreme impact on the objectives of the organisation.

2.1.4. Reaction to risks

The final step consists of decisions on how to react on risks. Some risks that are below a pre-defined risk level may be ignored or tolerated. For others, the costs for treating the risks might be so high that they outweigh the potential benefits. In this case the organisation may decide to abandon the related activity. Risks could also be transferred to third parties such as insurances that compensate for incurred costs. The last option would be to treat risks in defining strategies and actions that balance costs with potential benefits. This way, the organisation would decide implementing strategies for maximising benefits and minimising potential costs.

Table 3: Reactions to Risks

Treat the risk	These are actions that aim at minimising the probability or the impact of event to a level which is acceptable by the organisation. Most risks will belong to this category.
Transfer the risk	For some risks, it might be preferable to transfer the impact of an event to a third party, e.g. an insurance or body responsible for data protection. These third parties will perform an own risk analysis to decide if the risk can be tolerated and impact can be carried. Often, they demand additional measures for minimising probability or impact.
Tolerate the risk	In cases where impact would be minor or insignificant and

	the cost for treating risks would be disproportionate to the benefits, risks might simply be tolerated.
Terminate the risk	For some risks, the probability would be so high and/or the potential impact would be so extreme for the organisation that it would be preferable to terminate a certain activity. In certain cases activities are not terminated by private businesses but impact is rather transferred to the government or the activity is transferred, together with the risk to public bodies. In this case, termination is not an option for the public body.

2.2. Quality framework

A task team comprising representatives of national and international statistical organisations developed in 2014 a preliminary quality framework for statistics produced from big data. The task team worked under the umbrella of the UNECE/HLG project, the role of Big Data in the Modernisation of Statistical Production. It extended existing quality frameworks developed for the assessment of statistics derived from administrative data sources with quality dimensions that were considered as being relevant for big data sources.

The framework distinguishes between three phases of the business process, input, throughput and output. The input phase corresponds to the “design” and “collect” phases of the GSBP, the throughput to the “process” and “analysis” phases while the output is equivalent to the “dissemination” phase.

The framework applies a hierarchical structure that was adopted from the administrative data framework developed by Statistics Netherlands[7]. Quality dimensions are nested within a hierarchical structure called hyperdimensions. The three defined hyperdimensions are "source", "metadata" and "data". Quality dimensions are nested within these hyperdimensions and are assigned to each of the production phases. For the input phase the additional dimensions “privacy and confidentiality”, “complexity” (according to the structure of the data), “completeness” of metadata and “linkability” (possibility to link data with other data) were proposed to add to the standard quality model. For each of the quality dimensions, factors relevant for their description as well as possible indicators are proposed.

In the context of this paper, risks can be deducted from these factors. For instance, factors to be considered for the quality dimension “institutional/business environment” are sustainability of the entity-data provider. The related risk would be that data would not be available from the data provider in future. Another example is related to the newly proposed quality dimension privacy and security. One important factor is “perception”, meaning a possible negative perception of the intended use of specific data sources by various stakeholders.

[7] Daas, P., S. Ossen, R. Vis-Visschers, and J. Arends-Toth, (2009), Checklist for the Quality evaluation of Administrative Data Sources. Statistics Netherlands, The Hague/Heerlen

3. RISKS RELATED TO DATA ACCESS

3.1. Lack of access to data

3.1.1. Description

This risk consists of a project charged with developing a BOSP not getting access to a necessary Big Data source (BDS).

By now, the OSC has learned the hard way that even getting out of the starting blocks and getting this access is sometimes an unsurmountable obstacle. Sometimes, it is easy to get access to a particular source – such as call data records (CDRs) it for testing/research purposes, but far harder (for legal or commercial reasons) to get access to it for production purposes.

3.1.2. Likelihood

The likelihood is highly dependent on the characteristics of the BDS. If it concerns big administrative data, it may be as low as 1, in particular if (as in the case of the traffic loop data explored by Daas *et al.*[8]) there are no personal data protection issues. If the case of a BDS held by a private entity, in particular if it is sensitive (from e.g. a data protection perspective) or valuable (from a commercial perspective), the likelihood could be very high (5).

3.1.3. Impact

The impact depends on the BOSP and on the way in which the BDS is used. If the BDS is at the very centre, the impact may be very high (4 = not possible to produce the BOSP at all), whereas it might be lower if it is still possible to produce the BOSP (albeit with lower quality) by relying on other BDSs, resulting in an impact in the range of 2-3.

3.1.4. Prevention

To reduce the risk of lack of access, prior contacts with the data provider should be taken, and a long-term agreement on data access should be established. Moreover, a comprehensive legal analysis concerning the particular combination of BDS and BOSP should be conducted. The opportunities to access the data by means of existing or future legislation should also be assessed.

3.1.5. Mitigation

If there are alternative BDSs which could be used for the BOSP, these could be explored instead.

[8] Daas, P., M. Puts, B. Buelens and P. van den Hurk. 2015. "Big Data as a Source for Official Statistics". *Journal of Official Statistics* 31 (2). (Forthcoming; publication foreseen for June 2015.)

If there is no way to produce the BOSP without the BDS, and if it is not feasible to overcome the lack of access, the endeavour has to be terminated, and the new BOSP will not see the light of day.

3.2. Loss of access to data

3.2.1. Description

This risk consists of a statistical office losing a BDS underlying a BOSP.

3.2.2. Likelihood

If the BOSP is already being produced, there’s typically a certain stability, and in some cases, the risk may be very low (1). However, in particular in the case of private actors with which insufficiently firm agreements have been established, there is nothing preventing e.g. new management from changing data provision policy, leading to a moderate risk for discontinuity (3). Moreover, if the BDS is related to a volatile activity, there is always a risk of a provider simply going out of business, and the risk may be even higher (4).

3.2.3. Impact

As the existing BOSP may be impossible to produce, a very high impact (5) would often be the case. In other cases, where the BDS is of a supplementary nature, the impact may rather be loss of quality, with an impact in the range of 2-3.

3.2.4. Prevention

The prevention strategy is similar to that for lack of access to data – but with an increased emphasis on constant vigilance also in a production setting.

Not putting all eggs in one basket (i.e. having multiple BDSs underlying each BSOP) might also be a strategy, but this may either be impractical or too costly.

3.2.5. Mitigation

If the BDS is the outcome of a volatile activity, it might be the case that a new BDS, reflecting the same societal phenomenon gradually becomes available. However, it would be too late to start “scanning the market” once the BSOP has gone offline; constant vigilance would be needed – and this might be hard to achieve.

4. RISK RELATED TO THE LEGAL ENVIRONMENT

4.1. Non-compliance with relevant legislation

4.1.1. Description

This risk consists of a project charged with developing a BOSP failing to take relevant legislation into consideration while doing so – thereby rendering the BOSP non-compliant with said legislation. This could concern data protection legislation, regulations concerning response burden etc.

4.1.2. Likelihood

Given the OSC's unfamiliarity with Big Data, it is not unreasonable to say that occasional (3) non-compliance could take place. The likelihood would typically be related to the BDS, since the less "sensitive" the source, the less likely it is that non-compliance occurs.

4.1.3. Impact

The impact is typically critical (4), in the sense that non-compliant production will require the BOSP to be stopped (or, if it has not yet reached the implementation phase, its development to be terminated). It could even be extreme (5) since the reputational risks resulting from non-compliant ("illegal") official statistics production might have repercussions

4.1.4. Prevention

A thorough legal analysis has to be undertaken for any BOSP – and this at several junctures (what is acceptable during a development/exploration phase might not be so during the implementation/production phase). This might in turn lead to reengineering of the BOSP to render it compliant.

4.1.5. Mitigation

Depending on the severity of the non-compliance, the first step may have to be taking the BOSP offline.

Reengineering the BOSP to render it compliant might be an option, but whether the BOSP is "salvageable" in this manner is highly dependent on the nature of the non-compliance.

4.2. Unfavourable changes in the legal environment

4.2.1. Description

New legislation relevant to a BOSP under production might be introduced, effectively rendering the BOSP non-compliant.

4.2.2. *Likelihood*

It is not completely unlikely that advocates of increased data protection succeed to introduce new requirements which directly or indirectly have repercussions on the possibility to produce a particular BOSPs. A likelihood in the range 2-3 seems to be a realistic estimate.

4.2.3. *Impact*

The impact is typically critical (4), in the sense that non-compliant production will require the BOSP to be stopped.

4.2.4. *Prevention*

A certain business intelligence has to be undertaken regularly to monitor legislative development – possibly also to influence it by putting forward the case for official statistics in relevant (e.g. consultative) fora.

4.2.5. *Mitigation*

Under the condition that proactive monitoring has been conducted, there might be time to reengineer the BOSP to render it compliant also with the new legislation from day 1 of its entry into force.

If, on the other hand, monitoring has not been conducted, so that the new legislation "arrives as a surprise" – or if the legislation is so radical that there is no way to render the BOSP non-compliant – the only option might be to take the BOSP offline.

5. RISKS RELATED TO DATA CONFIDENTIALITY AND SECURITY

5.1. Data security breaches

5.1.1. *Description*

This is the risk refers to unauthorized access to data held by statistical offices. Third parties could obtain data that is held under embargo e.g. due to release schedule[9][10]. This can be for example data that is highly anticipated by stock market investors.

5.1.2. *Likelihood*

As far as the technical aspects of protecting the IT environment within the statistical office is concerned the risk has a similar

[9] For any BOSP that is based entirely on a single BDS it is inevitable that the data would implicitly be known to the original data owner and if the methodology is transparent the derived statistics will be known as well. This situation is not addressed here but rather in the risk related to the abuse of position by owners.

[10] The data can in addition carry the risk of confidentiality breaches. This risk will be treated separately.

likelihood for BDSs as for traditional sources. However there are two additional aspects that need to be taken into account.

The first is that with some BDSs the overall risk is slightly elevated due to the fact that data security at the original owner could be compromised. This can be due to e.g. industrial espionage or hacking.

The second is that once potentially valuable data starts to be held at the office the risk of attracting malicious intent will raise. If data held is of very high business value one should be prepared to face a very high likelihood of attacks targeting the IT infrastructure so the likelihood of a breach occurring could potentially be bigger (4).

If the data held is not perceived to be of value the overall likelihood seems to be not very high – from (1) to (3) depending on data source.

5.1.3. Impact

Potential damage to reputation can be big (5). What is relevant in the case of BDSs is that if a security breach occurs at the original owner the impact on the reputation of the statistical office is expected to be lower than if the breach occurs with data that is in its custody.

On the other hand it is possible that a breach at the statistical office can have negative consequences for the original owner. In this case a high negative impact is again possible due to the damage in terms of trust between the provider and the statistical office (5).

5.1.4. Prevention

What is specific for the case of BDSs is that the security procedures of the original owner could be relevant. It is not very likely that statistical offices will get auditing powers to control this. Owners whose data is used for the production of figures with sensitive publication schedules should be informed of the consequences for official statistics of a potential security breach at their premises and asked for formal assurance that adequate security procedures are being employed.

A direct way to prevent a security breach at the owner's premises from having a big impact for the statistical office is to ensure that multiple sources are used for the same product so one compromised source would not be enough to obtain the final figure. The advantage of this approach is that more control is in the hands of the statistical office.

A way to prevent security breaches at the statistical office from having a negative impact for the original data owner is to look for a way of working that does not involve the transfer of data which is potentially sensitive from the owner's point of view to the statistical office in raw form. A possible preventive approach is the use of aggregated data. It should be kept in mind however that some forms of aggregation, e.g. ones that are designed to prevent individual members of the population from being identifiable might not be appropriate in this case. One reason for this can be the fact that the risk to the owner stems from the business value of the data which may still be substantial even after anonymisation has been achieved.

5.1.5. Mitigation

In case a breach has occurred for data that is under the responsibility of the statistical office, the mitigating measures would be the same as for the case of traditional sources in case no negative impact for the original owner has occurred.

In case of negative consequences for the original owner the statistical office should review and strengthen its security procedures and clearly communicate and demonstrate its commitment to do so.

If a breach has occurred at the premises of the original owner then the statistical office concerned should clearly communicate the situation and insist on the improvement of the owner's security procedures. If necessary an alternative provider could be sought.

5.2. Data confidentiality breaches

5.2.1. Description

This is the risk that the confidentiality of one or more individuals from the statistical population is compromised. This can be due to an attack on the IT infrastructure, due to pressure from other government agencies or due to inadequate statistical disclosure control measures.

5.2.2. Likelihood

Similarly to the case of the data security breach risk the technical circumstances of keeping microdata do not change so much with the addition of BDSs. However also here there are caveats.

Microdata from certain data sources can have high business values so holding it would increase the likelihood of attacks.

Additionally some microdata can be potentially very useful to other government agencies e.g. law enforcement, taxation or public health related ones. In certain circumstances the commitment to the principle of statistical confidentiality may come under big pressure.

Regarding statistical disclosure control failures there is already well established practice by now. BDS might allow producing statistics for smaller subpopulations, or provide the ability of linking aggregate data from different BDS, which could increase the likelihood of risk occurrence. In addition, new sources however will require new methodological developments, so the real danger is that the disclosure control methodology is not adequately updated.

Overall with reasonable preventive measures the likelihood could be kept to reasonable levels, but since there are many different and diverse factors the appropriate evaluation here seems to be that the likelihood is high (4).

5.2.3. Impact

Potential damage to reputation can be big (5). As with the data security breach risk a breach at the statistical office can have negative consequences for the original owner. Here the impact of such an event could be potentially even bigger, especially under the assumption that current trends in public opinion continue. The damage in relations between the data provider and the statistical office is also foreseen to be very big.

5.2.4. Prevention

A sure-fire way to prevent such a risk from materialising is to not have microdata from BDSs at all (though holding other microdata still incurs the corresponding risk albeit with different likelihood and impact). Going this way would entail, similarly to the case of the data security breach risk, the need to devise other ways to exploit the data for statistical purposes. Also here the different nature of the sources would mean that new methodologies would need to be developed with the competing goals of extracting as much useful information as possible and keeping confidentiality out of danger.

In case microdata is held then IT security and access control arrangements need to be on the required level and continuously monitored. Special care needs to be taken to ensure that the new ways of getting the data are safe. Ironically such a new way could be the physical transportation of storage devices (e.g. hard disks). If this method is used then the delivery should be physically secured and encryption should be used.

5.2.5. Mitigation

The mitigating measures here are in principle the same as the ones for data security breaches. If the reason for the breach has been pressure from another government body the opportunity should be taken to strengthen the independence of the office so that similar breaches become harder in the future.

5.3. Data source manipulations

5.3.1. Description

Data providers from third parties, for example social network data or voluntarily contributed data bears the risk of being manipulated. This can be done either by the data provider itself or by third parties. For example many spurious social media messages could be generated in order to push a statistical index derived from these data in one or another way in case it is known that the index is calculated from such data.

For voluntarily contributed data it can be the case that the volunteers are representing a specific interest group with a specific agenda.

5.3.2. Likelihood

For data whose manipulation can bring big benefits the likelihood is higher. This can be data on which statistics interesting for e.g. the stock market are based. In light of the recent LIBOR and Forex scandals it could be assumed that as long as the incentive is there attempts to manipulate data would be likely.

For statistics based on voluntarily contributed data one has only to look at recent PR practices of hiring people who pretend to have a certain opinion and are paid to expressed it publicly (e.g. on internet forums) to conclude that the likelihood is not small. Overall a figure between 3 and 4 seems to be adequate.

5.3.3. Impact

A big problem with manipulations is that they can last for a long time without being detected. If a manipulation continues for a long time the impact on quality can become large. In addition damage to public trust in official statistics could also be big especially if the role of statistical offices as providers of quality data has been publicly underlined. On the other hand if a manipulation is discovered on time and then publicized this may actually improve public perception. Except in extraordinarily bad cases a maximal impact of (3) could be imagined.

5.3.4. Prevention

Performing regular benchmarking exercises with alternative sources is one possible preventive approach. These alternative sources could be traditional or otherwise. Basing the statistic on a combination of sources could prevent manipulations from having a significant impact. In cases where provider initiated manipulations are feared legal agreements could also be one approach to prevent the practice.

5.3.5. Mitigation

In terms of public relation damage the mitigating measures to be taken here are not much different from the measures to control any crisis.

In terms of data quality it would be beneficial if past data could be corrected so that even with a big delay a correct series could be produced. For this purpose regular benchmarking could be helpful. Note that the purpose of benchmarking in this case is slightly different than for the case of prevention. For prevention it is important to notice and investigate a suspicious discrepancy between the benchmark data and the BDS quickly. For mitigation purposes old benchmark data is always useful.

In addition care should be taken not to allow similar manipulations in the future – in particularly sensitive cases this could mean obtaining potentially redundant data from several providers for benchmarking purposes.

5.4. Adverse Public Perception of big data usage by official statistics

5.4.1. Description

Media and general public are very sensitive towards issues of privacy and use of personal data from big data sources, especially in the context of secondary use of data by government agencies taking administrative or legal measures towards citizens. Negatively perceived usages might be the positioning of speed monitoring based on analysis of navigation data[11]. The specific case of TomTom Netherlands caused a considerable drop in demand for TomTom devices and led to a decision by the company to restrict access to the data. In this specific case, the data did refer to individuals but to speed levels by road segment.

However, there could be big data sourced applications that are positively conceived by the public. One example would be applications preventing crimes such as burglary based on big data methods.

Positive as well as negative public opinions could have a strong impact on using a BDS in the context of producing official statistics.

The consequence of a negative public perception could be that

- the BDS would no longer be available to statistical offices, either through decisions of the data provider or government decisions not to use the data or
- the use of data would be restricted, possibly preventing the production if certain BOSPs;

5.4.2. Likelihood

Factors that could influence the probability of such an event or the impact of it on the production of statistics are

- the sensitivity of data, i.e. how easily persons could be identified;
- the amount of information the data reveal on individuals, e.g. increased by linking data from different sources;
- the type of data, e.g. financial transactions are perceived as more sensitive than other data;

[11] See http://www.theguardian.com/technology/2011/apr/28/tomtom-satnav-data-police-speed-traps

- the type of potential action that could be executed on the citizens, e.g. fining persons for speeding;
- unclear legal environment in which data providers and users are operating or when legal conditions conflict with public ethical opinions/standards;
- the degree of dependence on a certain data source for producing statistics; during the exploration phase, this factor might only be of minor importance. However, it might very heavily impact the production of statistics at a later stage and has therefore to be considered at the exploration phase, too. One problem might be that the final extent of data use is not known at the beginning as data sources might have the potential for serving more than one statistical domain.

The assessment of timing of adverse events is not possible, because mobilisation of the public is often triggered by publicizing events with negative impact on citizens. However, with increasing use of big data by governments and private businesses and especially with actively marketing data for other purposes than the one that triggered its original collection, it is more likely that such events would happen.

Events that strongly influence the public perception are not frequent but rather occasional (3) to remote (2). With increasing use of big data sources the probability is likely to increase, too.

5.4.3. Impact

The impact of an event depends very much on the factors that are discussed above. The impact in general is more severe for an already established production of statistical data, because the activity might have to be terminated. Impact also depends on availability of alternative data sources, although it could be that public perception does not distinguish between different data sources in case the event has materialised. In the current state of big data usage, it seems that these sources cannot replace completely traditional data sources but rather supplement existing statistics. This would decrease the impact of events. Therefore the impact of an event is considered ranging from 2 (minor) to 3 (major). During the production phase, impact could increase to 4 (critical).

5.4.4. Prevention

Preventive measures could be the definition of ethical guidelines for big data in official statistics. Ethical guidelines should be strongly based on principles like the code of practice for European Statistics or the fundamental principles of official statistics[12]. The next

[12] http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

measure would be the definition of a communication strategy that would publicize the findings of the ethical guidelines to the public and could be used to inform stakeholders on the ethical use of BDS for BOSPs.

A separate risk assessment for specific BDS could be performed to identify risks and propose preventive or mitigation actions on the basis of the ethical principles. A separate risk assessment could also include stakeholders, such as data protection agencies to ensure identification of all risks and validity of actions.

5.4.5. Mitigation

The communication strategy should also include measures in case of a growing negative public attitude. The separate risk assessment should collect positive examples of data usage and measures to prevent abuse of data, which could both be communicated via the media. In certain cases, actions might be necessarily taken at policy level and the statistical community might not be able to influence them effectively.

5.5. Loss of credibility – being no longer observation based

5.5.1. Description

Users of official statistics usually have high confidence in accuracy and validity of statistical data. This is based on the fact that statistical data production is embedded in a sound and publicly available methodological framework as well as the documentation of quality of a statistical product. In addition, most statistical data are observation based, i.e. are derived from surveys or censuses, which establish an easily understandable relationship between observation and statistical data. Using BDS which are not collected for the primary purpose of statistics bears the risk that this relationship will be lost and the users would lose trust in data from official statistics. An example related to the last round (2010) of population census relate to the fact that in some countries, statistical data were derived using multiple sources and statistical models. In a number of cases statistical data were contested by stakeholders.

5.5.2. Likelihood

The likelihood of risk incurrence depends on factors such as complexity of statistical/methodological model, plausibility of relationship between BSD and BOSP, or consistency with other statistical data. Likelihood should be in the range of 3 (occasional) to 4 (probable), meaning it would be likely to occur several times to frequently.

5.5.3. Impact

The impact of occurrence of the risk would very much depend on whether NSOs could successfully prove the accuracy and validity of the statistical data. In case this could not be achieved the impact in terms of loss of trust and credibility could also affect other statistical domains, i.e. the credibility of not only some statistical data but could put in question the organisation itself. NSOs would lose a competitive advantage towards other private organisations active in this field.

5.5.4. Prevention

Preventive actions would be to develop and publish scientifically sound methodology which is recognised by the scientific community, enrich data with metadata on quality, ensure consistency of the BOSP with non BOSP, execute strict quality assurance.

Before engaging into statistical production, BOSP could be published as experimental and stakeholders could be encouraged to contest the BOSP in order to confirm or enhance the BOSP.

5.5.5. Mitigation

There are two cases to distinguish. In case statistical data are contested but are of high/sufficient quality (correct/accurate), it would be sufficient to explain and communicate statistical data to the public, giving easy to understand examples.

In case data are of insufficient quality or are simply wrong, it would be necessary to launch a public analysis of the production process and the related methodological framework to identify errors and propose and implement corrective measurements. In this case, it would be much more difficult and would take longer time to regain public trust. The analysis could as well lead to the result that use of certain BDS would not be reliable enough and should be abandoned.

6. RISKS RELATED TO SKILLS

6.1. Lack of availability of experts

6.1.1. Description

The analysis of the digital trails left by people during the performance of their activities requires particular data analysis tools which are not currently the most common in official statistics. Firstly, the use of indirect evidence about people activities instead of the direct questioning in surveys requires the use of statistical models and therefore skills on model based inference and machine learning. Secondly, those digital trails consist of data which often is not in the usual table format common in survey results, with rows corresponding to a statistical unit and columns to particular characteristics of those statistical units. Digital trails are also in the form of text, sound, image and video. The extraction of relevant statistical information from these types of data requires skills in natural language processing, audio signal processing and image processing. Thirdly, these data sources tend to provide massive datasets which treatment requires a good understanding of distributed computing methodologies.

The risk of lack of availability of experts consists of upon receiving data from one of these new big data sources, the statistical office not having the possibility of processing and analysing it properly, due to its staff not having the required skills.

6.1.2. Likelihood

The likelihood of this risk will depend on three factors: 1) the particular types of skills required by each type of big data source and the probability that the statistical office will find the opportunity to explore such source; 2) The current availability of the required skills in the statistical office; and 3) the organisational culture of the statistical office.

Concerning the types of skills which may be required, it should be noted that not all sources require all the skills enumerated above. Some (e.g. Google Trends type of data) do not require distributed computing, as they come already pre-processed from the data holder, or signal processing skills, and they would mostly require skills in statistical modelling. However, there is a great variety of big data sources with most indeed requiring distributed computing, signal processing and machine learning skills. At the same time the proper exploration of these digital trails will require the need to process multiple sources. Therefore, there is a high probability that the big data sources becoming available to the statistical office will require those uncommon skills and the likelihood of this risk is very high (5).

Concerning the current availability of the required skills it will depend on the particular statistical office. Even if less common than survey methodology, model based inference is also used in official statistics in particular domains. So even if it may require some redeployment of human resources, the statistical offices could find a solution in-house. As for distributed computing skills, being mostly IT related skills, it will depend on how the IT infrastructure is managed in the organisation. Depending on how outsourced IT is, solutions could be found in the context of the existing arrangements. However, skills in signal processing and machine learning will normally not exist in most official statistical offices and the application of those skills cannot be outsourced, as they need to be applied by statistical domain experts. Therefore, from this point of view the likelihood of this risk also seems very high (5).

The organisational culture will also have an influence on the likelihood of this risk. The existence of staff with the willingness to acquire the required skills via self-learning may provide the organisation with the ability to respond to the situation of a new data source requiring skills different than usual. That will depend on the organisational culture of the statistical office, namely if it encourages staff to learn new skills and if it allows the staff time for self-learning.

Therefore, the likelihood of the statistical office not having the possibility of processing and analysing new data sources, due to lack of skills of its staff will be between probable (4) and frequent (5) depending on the self-learning culture of the organisation.

6.1.3. Impact

The statistical office not being able to process and analyse big data sources due to lack of skills of its staff may have two possible negative consequences: 1) the data source will not be explored, at least not at its full potential; 2) the source will be wrongly used.

Missing the opportunity of exploring fully the potential of a valuable big data source will have a minor impact (2) is the short-run, as statistical offices do have statistical tools to answer to current needs. However, in the long-term (and maybe even in the medium-term) the impact of losing this opportunity will have a critical impact (4) as statistical offices increasingly face the competition of private suppliers who do not have the same institutional framework which would allow them to guarantee to society the independence of the statistics produced.

However, to wrongly use the source would have an extremely negative consequence on statistical offices, as official statistics rely heavily on their reputation to perform their mission. Nevertheless, we could argue that the most important skill which, if missing, could lead to the derivation of wrong results is statistical inference, in particular model based inference, which is also the one which is less likely to be missing. Therefore the expected impact would be rather critical (4) than extreme.

6.1.4. Prevention

There are two ways in which statistical offices can pro-actively prevent this risk: 1) training; and 2) recruitment.

Statistical offices can provide the required skills to the staff by identifying in detail the skills needed to use big data sources in statistical production, by making an inventory of the existing skills of the staff, by identifying learning needs and then by providing a training courses.

Statistical offices can also recruit new staff with the required skills. This seems to have serious limitations, as statistical offices will not be able to recruit a critical mass of staff for a situation where the use of big data sources would be widespread in the office and new staff would still need several years to reach the level of experience of existing staff. However, at least some of the new staff recruited in the framework of the normal renovation of the personnel could be required to possess big data related skills.

6.1.5. Mitigation

Faced with the situation of having new big data sources available without staff with the required skills, statistical offices can mitigate the negative consequences in two ways: 1) sub-contracting; and 2) cooperation.

Statistical offices can sub-contract the data processing and analysis of new big data sources to other organisations which provide these types of services. This seems to be a viable solution, as a new sector of enterprises specialised in processing these types of data is emerging. However, this is a solution which itself has some risks, as the statistical office would have less control of the production of possibly sensitive statistical products. It is a solution which also has the disadvantage that it does not allow the staff of the statistical office to learn and acquire the required skills.

Cooperation with other organisations which have staff with the required skills and who would also have an interest in the exploration of the big data source seems to be a more promising solution. This cooperation could take the form of joint projects with staff from the statistical office and staff from the other organisations in equal footing, sharing their knowledge. This would have the advantage of not only mitigate the risk of lack of skills, but also allow the staff of the statistical office to acquire those skills.

6.2. Loss of experts to other organisations

6.2.1. Description

This risk consists of statistical offices losing their staff to other organisations after they have acquired big data related skills.

6.2.2. Likelihood

The likelihood of this risk will depend on two factors: 1) existing attractive opportunities in organisations outside official statistics; 2) working conditions at the statistical offices.

Concerning opportunities in organisations outside official statistics, the likelihood of this risk seems probable (4). There is a high demand for people with big data related skills in the private sector and also in other public sector organisations. After acquiring big data skills, official statisticians will have the comparative advantage of being experienced statistics practitioners. Besides the specific big data related skills, other organisations need data scientists with also more traditional skills, such as users' needs assessment and development of key performance indicators (KPI) which are common between official statisticians. Additionally, it is expectable that the staff who will be more willing to acquire new skills will also be the one who would also be more open to a change in career and leave the statistical office.

Concerning the working conditions in the statistical offices, it will obviously depend mostly on the particular office. However, statistical offices in general still offer attractive professional possibilities to people quantitatively minded. Statistical offices offer the largest range of possible domains to work and the largest variety of data to work with. This would mitigate somehow the likelihood of the risk of statistical offices loosing their staff to occasional (3).

6.2.3. Impact

The impact of this risk would be the same as for the risk of not having staff with the appropriate skills in the first place. Therefore, the impact would be critical (4) as argued above.

6.2.4. Prevention

The only possibility for statistical offices to prevent this risk seems to be to provide attractive working conditions to their staff. This is true in general for its entire staff. However, in the particular case of staff open to learn new skills, namely big data related skills, working conditions could be improved by providing learning opportunities where they could develop their professional interests. Statistical offices could also pay particular attention to be open to innovative new projects and ideas involving new big data sources coming from statisticians working in the several statistical domains. Finally, the prevention of loosing staff to other organisations in the sequence of their big data skills, will depend on a good identification of the staff able and willing to work with such data, and on the provision of good opportunities for their professional development.

6.2.5. Mitigation

The mitigation of this risk would be done as for the risk of not having the staff with the appropriate skills: 1) sub-contracting; and 2) cooperation.

7. DISCUSSION

From this first overview, it is obvious that it is impossible to establish a single likelihood or impact for a given “big data risk” – typically, both measures depend heavily on the Big Data source as well as on the “Big Data based official statistics product” involved.

We therefore conclude that the logical next step in this endeavour is to proceed by means of example – taking a number of possible pilot projects (each involving a combination of one or more BDSs and one or more BDOSs) as the point of departure, and – for each such pilot – striving to assess likelihood and impact for each risk.

To this end, we are on the verge of launching a stakeholder survey, trying to gauge the OSC’s assessment of likelihood, impact (and possible prevention/mitigation actions) concerning a number of possible pilots – and to seek OSC suggestions concerning risks that we have not included in this paper.

8. References

UNECE (2014), "A suggested Framework for the Quality of Big Data", Deliverables of the UNECE Big Data Quality Task Team, http://www1.unece.org/stat/platform/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2

UNECE (2014), "How big is Big Data? Exploring the role of Big Data in Official Statistics", http://www1.unece.org/stat/platform/download/attachments/99484307/Virtual%20Sprint%20Big%20Data%20paper.docx?version=1&modificationDate=1395217470975&api=v2

Daas, P., S. Ossen, R. Vis-Visschers, and J. Arends-Toth, (2009), Checklist for the Quality evaluation of Administrative Data Sources, Statistics Netherlands, The Hague/Heerlen

Dorfman, Mark S. (2007), Introduction to Risk Management (e ed.), Cambridge, UK, Woodhead-Faulkner, p. 18, ISBN 0-85941-332-22)

Eurostat (2014), "Accreditation procedure for statistical data from non-official sources" in Analysis of Methodologies for using the Internet for the collection of information society and other statistics, http://www.cros-portal.eu/content/analysis-methodologies-using-internet-collection-information-society-and-other-statistics-1

Reimsbach-Kounatze, C. (2015), “The Proliferation of “Big Data” and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis”, OECD Digital Economy Papers, No. 245, OECD Publishing. http://dx.doi.org/10.1787/5js7t9wqzvg8-en

Reis, F., Ferreira, P., Perduca, V. (2014) "The use of web activity evidence to increase the timeliness of official statistics indicators", paper presented at IAOS 2014 conference, https://iaos2014.gso.gov.vn/document/reis1.p1.v1.docx

Even if not explicitly mentioning risks, this paper in fact approaches the many risks associated to the use of web activity data for official statistics.

Eurostat (2007), Handbook on Data Quality Assessment Methods and Tools, http://ec.europa.eu/eurostat/documents/64157/4373903/05-Handbook-on-data-quality-assessment-methods-and-tools.pdf/c8bbb146-4d59-4a69-b7c4-218c43952214