DRAFT Version 4.1, 30 September2019

Ensuring data quality in the world of "Big Data" and rapidly expanding community of ecosystem accounting

Francois Soulard[1], Michael Vardon[2] and Steve May[3]

[1]Statistics Canada
[2]Australian National University
[3]Australian Bureau of Statistics

Paper for 25th Meeting of the London Group, 7-10 October 2019, Melbourne Australia

Abstract:

This paper examines the experiences of producing environmental and ecosystem accounts and different ways data quality has been assessed. This is done in order to highlight the practical issues that need to be addressed to ensure the confidence in the results needed for them to be useful in government decision making. Environmental and ecosystem accounts have been produced in some places by a range of agencies, and in all cases have necessarily involved a range of different expertise (ecology, economics, hydrology, accounting, etc) and emerging sources of "Big Data" like satellite Earth observation. The experience to date is reviewed bearing in mind seven dimensions of data quality– relevance, accuracy, timeliness, accessibility, interpretability, coherence and the institutional environment. In this, it seems that different dimensions of data quality are emphasised by different communities and are effectively traded-off, with relevance, timeliness and interpretability preferred by the user communities and accuracy and coherence preferred by the producer community. The institutional environment, which is closely associated with trust in data, is fundamental to both producers and users. It is also noted that the process for assuring data quality for environmental and ecosystem accounts has until now relied almost entirely on voluntary contributors by agencies and individuals.

Questions for the London Group to consider:

- Are the typical measures of data quality used in statistical agencies also suitable for accounts using "Big Data"?
- How has "Big Data", including remotely sensed data, been used in the production of accounts?
- What processes have been used for assessing data quality in accounts using "Big Data" sources?

Introduction

The existing data quality assurance processes used by national statistical offices and other government information agencies should be revisited for the emerging area of ecosystem accounting as well as the still maturing environmental-economic accounting more generally. In both cases, the production of accounts requires the use of data from a variety of sources and for these there are different measures of data quality. In particular, biophysical measures extracted from so-called "Big Data", and especially satellite Earth observation data, may require novel estimates of uncertainty. The uncertainty from these data arises from types of measurements that are new to most national statistical offices. In some cases, the account production in statistical offices has followed existing data quality processes, while in other cases government information agencies that have become involved in the production of accounts are using their existing procedures.

A key to this discussion is understanding what is meant by uncertainty. The Intergovernmental Panel on Climate Change (IPCC) defines uncertainty as:

> "An expression of the degree to which a value (e.g., the future state of the climate system) is unknown. Uncertainty can result from lack of information or from disagreement about what is known or even knowable. It may have many types of sources, from quantifiable errors in the data to ambiguously defined concepts or terminology, or uncertain projections of human behaviour. Uncertainty can therefore be represented by quantitative measures, for example, a range of values calculated by various models, or by qualitative statements, for example, reflecting the judgement of a team of experts (see Moss and Schneider, 2000; Manning et al., 2004) [1]."

The paper makes two main points about uncertainty:

(1) given the difficulties of providing objective metrics of uncertainty or quality in national, environmental and especially ecosystem accounts, uncertainty is easily overlooked, and;
(2) the approach and range of factors that need to be taken into account when managing and describing data quality may need to be reconsidered. In this, data quality may be described qualitatively and quantitatively.

To illustrate these points example accounts are compared against a set of seven defined quality dimensions (Table 1). It concludes with some key lessons for managing and describing uncertainty in accounting and valuation more generally.

In examining uncertainty, it is important to recognise that while there are many academic studies of ecosystems and their services, none have been institutionalised within government or repeated over a number of years. This is unlike national and environmental accounting, although similar issues can present themselves in, for example, land, water and carbon accounts. To move ecosystem accounting into regular production, countries will have to development systems to ensure the data and metadata are of sufficient quality to meet the needs of policy makers and other stakeholders. To date, ecosystem accounting, which has grown out of the traditions of both environmental accounting and academic research,

[1] IPCC Glossary of Terms used in the IPCC Fourth Assessment Report http://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr_appendix.pdf

particularly of ecosystem services, has largely relied on peer review to ensure that data quality was sufficient to meet the objectives. Reviewers look at the methods, the results, the way in which uncertainty is described and assess the degree to which the conclusions drawn can be supported by the data available. Such processes are labour intensive, reliant on expert opinion and the goodwill and availability of reviewers, as there are relatively few experts in the field of ecosystem accounting.

Table 1. Brief descriptions of the seven dimensions of data quality

Dimension of data quality	Description
Relevance	The estimates produced relate to an area of interest to data users
Accuracy	The closeness of the estimate to the real answer
Timeliness	The time between the production of the estimate and the reference period
Accessibility	The way the estimates can be accessed
Interpretability	The ability to interpret the data
Coherence	The extent to which data are logically consistent in terms of definition and measurement and can be reliably combined in different ways. A part of this is ensuring comparability over time. That is changes are real and not to changes in definition or measurement
Institutional environment	The legal basis and financial the resources needed to produce estimates are in place

This paper draws on a range of material from the Australian Bureau of Statistics (ABS), and in particular the publication Quality Dimensions of the Australian National Accounts[2], Statistics Canada's Quality Assurance Framework[3], and the International Monetary Fund's (IMF) Data Quality Assessment Framework[4] as well as other references and various data quality frameworks from around the world (e.g. OECD[5] and Eurostat[6]). Statistics Canada first released the dimensions of data quality for official statistics in 2002, and revised them in 2017[7].

A key aspect of accounting that sets it apart from other systems of organising information is that, as a completely integrated system of accounts, it forces a number of checks and balances in the process of compilation. Supply must equal use and all changes between opening and closing stocks must be accounted for, even if this is through the inclusion of balancing items. Given this, the use of multiple data sources or estimation procedures forces differences to be reconciled; this is not unusual in accounting and, for example, the three different ways that gross domestic product (GDP) is produced in national accounting (which is highlighted later in the paper). The internal consistency of the accounting system sets it apart from measures of uncertainty of individual components. The aim is to maximise the usefulness of the data provided by the system as a whole (e.g. in the System of National Accounts the interactions

[2] https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/5216.0.55.002Main+Features12007?OpenDocument
[3] https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.htm
[4] https://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm
[5] https://www.researchgate.net/deref/http%3A%2F%2Fwww.oecd.org%2Fofficialdocuments%2Fdisplaydocumentpdf%2F%3Fcote%3Dstd%2Fqfs%282011%291%26doclanguage%3Den
[6] https://www.dzs.hr/Eng/international/code_of_practice_en.pdf
[7] Statistics Canada, 2002, 2017.

of labour, capital, production, consumption, accumulation, etc.), which is not necessarily achieved by maximising the accuracy of every component of the system.

Big Data and environmental accounting

The concept of "Big Data" was first coined by NASA scientists in 1997 when describing issues related to computers' capacity to handle large data sets for graphical display, i.e. visualisation (Press, 2019). Although there exist numerous definitions of the concept, Wikipedia has it as:

> "a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software" (https://en.wikipedia.org/wiki/Big_data).

"Big data" has also been called the new science of understanding and predicting human behaviour by studying large volumes of unstructured data" (Gil, 2019).

The current proliferation of sensors and other data acquisition systems is such that experts estimate that, as of 2016, some 90% of the world's data had been created in the previous two years (Lochner, 2016). This digital transformation is having a profound impact on businesses, governments and citizens. For example: business can track customer preferences and target advertising; governments can monitor movements of money and match against illegal activity, and; people have access to a wealth of information assisting them in making decisions about how to spend their time and money.

This increasing stream of "Big Data" entails a torrent of data analytics. Given the growing number of data sources and ensuing exponential growth in data points, government information agencies have never been more challenged to remain current and a trusted source of information when, one the one hand, large amounts of real-time data are available (e.g. Google consumer price index, Global Surface Water Explorer), and on the other, they are expected to provide the same rigour in vetting and qualifying with appropriate metadata and produce official estimates. The processes to do the latter, to evaluate data quality, were developed in a context of survey and other more "traditional" administrative data whose volume, variety and velocity were manageable. The circumstances have now changed.

To remain relevant, information agencies need, as a matter of survival, to increasingly rely on "Big Data". The challenge is to remain trusted by ensuring that data quality characteristics of traditional data sources are also appropriate for "Big Data" which are sometimes *in-situ* data (for example, stream gauging data) but more often than not, in the case of physical ecosystem accounting, satellite data. This circles back to the initial identification of the concept of "Big Data": visualisation and interpretation of satellite-generated data.

"Big Data" is also relevant to economic statistics, with transactional data available from a ranged of sources. Whereas once limited to government information sources, such as taxation records, economic data from a range of industries (e.g. scanner data from retail sales) are now available and part of the conundrum. While these are important, they will not be further addressed in this paper.

Water accounts and satellite Earth observation in Canada: challenges and potential

The case of the use of satellite Earth observation (SEO) to compile water accounts in Canada is a good example of data quality issues associated with the use of this new source of information.

Canada may have one of the most complicated water regimes on the planet. With a landmass of nearly 10 million square kilometres, with over one million square kilometre of freshwater surface, 3,500 cubic kilometres of renewable water flowing through its rivers and lakes each year, glaciers, permanent snows, permafrost areas, and a complex and diversified array of ecosystem types ranging from rainforests to dry grasslands and countless permanent and seasonal wetlands, measuring water in Canada is by no means an easy task.[8]

It had been proposed that Statistics Canada could base new estimates for water accounting items on data produced from the Global Surface Water (GSW) database.[9] The GSW, a product of the European Commission's Joint Research Centre, provides information on the water occurrence and annual water recurrence at pixel level over the past 32 years at a 30-metre resolution. Initial analysis for Canada shows that most of the water area visible at the resolution of 30 metre is captured by the new data product. This may serve various purposes such as helping to create a reference data set measuring permanent water extent.

However, specific data quality issues prevent its use as a main source of data for the water accounts. For example, the identification of seasonal water, a phenomenon that characterises large swaths of land in Canada, is depicted to a lesser degree of precision in the data product, as compared to permanent water, therefore introducing an element of non-comparability. Also, water area is obscured by floating, overhanging and standing vegetation, or hidden by infrastructure (such as bridges) and is not captured. The fact that water located below floating vegetation is not captured makes the analysis of water extent change over time problematic and overly sensitive to weather-related events such as floods and droughts. Moreover, the time periods are not long enough to capture real change in water area. Also, water bodies smaller than 30 x 30 meters are not captured due to the resolution of the input images. Tracking these small water bodies is an important aspect of the changing characteristics of water in Canada, since they are the most affected by changes in the hydrologic regime. In the end, Statistics Canada continues to produce a partial water asset account based on the geospatial modelling of *in-situ* data[10].

However, planned improvements in this specific data set, for example regarding an increased spatial accuracy, may make it a viable data source for water accounts in the future; improvements in modelling and the availability of alternate data sources may lead to enough certainty to provide reliable opening and closing stock estimates and measures of change over time. As an example of a supplementary data source, Statistics Canada is exploring the Gravity Recovery and Climate Experiment (GRACE, NASA) data. These data provide measures of change in total water mass and will be used to provide an estimate of the change over time of the total stock of water. It does not allow, however, for an evaluation of the size of the stock itself – only the changes in the stock.

---

[8] Statistics Canada, 2017b

[9] Pekel, 2016

[10] *In-situ* data, which, in this specific case, refers to sensor data from stream gauges, if often considered Big Data of sort, but by no means compare in volume to satellite Earth observation.

The use of the GSW and other global Earth observation datasets (both satellite and *in-situ*) will likely become a key part of making progress on the SEEA and especially the production of land, water and ecosystem accounts. The increasing volume of Earth observation data openly available, combined with open source solutions for processing, will enable the production of regularly updated global geospatial dataset. For example, the Canadian Space Agency has recently released over 37,000 RADARSAT-1 images acquired from 1995 to 2013 for public use, free of charge. Also, in March 2019, the first Canada-wide wetland inventory using Landsat-8 imagery and innovative image processing techniques available within Google Earth Engine was published.[11] An improved version of this preliminary wetland map, integrating optical and radar data, is currently being worked on and expected soon.

Accounting for data quality in Australia: example from the Great Barrier Reef

This section looks at the data quality issues in the ABS' Experimental Environmental-Economic Accounts for the Great Barrier Reef (ABS 2017). The experimental accounts for the Great Barrier Reef include over 100 tables, covering accounts for land extent and condition, marine extent and condition, water, carbon, biodiversity, ecosystem services and industries (e.g. agriculture, forestry and fishing). The Explanatory Notes[12] list over 70 data sources, including much "Big Data" (e.g. from remote sensing).

Information on data quality[13] for the Great Barrier Reef is summarised and arranged under the headings: institutional environment; relevance; timeliness; accuracy; coherence; interpretability, and; accessibility. The headings and information included under them is in-line with standard ABS practice.

Of particular note are the comments under the heading "accuracy", which notes the large number of data sources, and says:

> "*In the case of ecosystem accounts, it is recognised internationally that an objective accuracy measure, in the sense of proximity to a 'true value', is near-impossible to measure. The ecosystem account is a complex set of environmental and economic statistics. It combines a large number of ABS and non-ABS data sources covering various aspects of the environment and the economy to derive a number of statistical tables, including various headline measures."*
>
> and
>
> "*Given the variety of data used, and the adjustments/transformations undertaken in the compilation of the ecosystem account, an overall assessment of accuracy is necessarily subjective. It involves an assessment of underlying data, the adjustments and transformations made including assumptions used. The ABS aims to achieve best practice in each of these facets of ecosystem account compilation*."

[11] Amani et al. 2019

[12] See Explanatory Notes: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4680.0Explanatory%20Notes12017?OpenDocument

[13] See Quality Declaration – Summary: https://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/4680.0Quality%20Declaration02017?opendocument&tabname=Notes&prodno=4680.0&issue=2017&num=&view=

These statements are very true. They do, however, hide much about the individual sources of data and how discrepancies were identified and reconciled and point to the need to reconsider how data quality is assessed and described in the world of "Big Data" and with the rapidly expanding area of ecosystem accounting.

A key issue with the accounts for the Great Barrier Reef is recognition and acceptance by the potential users of the accounts. This is related to the institutional environment and relevance. In particular, the accounts, while recognised by key potential users, including the Great Barrier Reef Marine Park Authority and the Queensland Government, have not obviously contributed to decision-making in an area where there are many environmental concerns and economic benefits from using the reef and surrounds and it is clear that there are trade-offs. This means that while they may be recognised and relevant, they are not yet accepted and questioning of data quality is an issue that needs to be addressed.

Some of these issues were examined in an online newspaper article on the accounts by one of the authors (see Vardon, 2017). In the article, it was noted that environmental accounting offers a clear way to assess such trade-offs and will hopefully lead to better decisions. One of the other points made was that the accounts could be used by the public to hold our government and business leaders to account. This possibility no doubt makes some potential users uncomfortable and may be a reason why they have not found favour in government agencies. This may be because the ABS, as an institution, is independent of government policy and the final accounts, like the national accounts, come out on a pre-determined release date without clearance from a Minister. As such, there is little warning of unflattering information and no way of delaying the accounts to suit the needs of the relevant management agencies or Minister(s). Questioning of data and data quality is one way that accounts may be dismissed. In this, the labelling of accounts as experimental, while accurately conveying their status and it is clearly stated that one of the reasons the experimental accounts are produced is to identify ways of improving them, it also means that they can be ignored.

Discussion

Issues of SEO data quality for environmental statistics have been explored at length elsewhere[14,15]. The data quality criteria presented in Table 1, which were originally defined for evaluating the quality of a final statistical product such as survey results, were discussed from the perspective of their application to the use of SEO data. However, it is not clear that the discussion has been completely exhausted in the case of the use of SEO in environmental accounting, and particularly ecosystem accounting.

Far from dismissing the use of SEO in environmental accounting, it is however clear that each SEO data set need to be vetted from the perspective of its "fitness for purpose", and not simply adopted at face value. For example, the concept of relevance: the traditional interpretation (as in Table 1) mentions that the estimates produced relates to an area of interest to data users. However, in the case of SEO, the user needs to consider if the data also allows the question to be answered properly. In the case of surface water in Canada presented above, the SEO data may depict surface water at a certain point in time or over a certain time

[14] Australian National University, 2018
[15] United Nations et alter, 2017

period, but is it representative of other times, and more importantly, the time referred in the question? Also, is the spatial accuracy sufficient? Is the scale appropriate? Are the producer and user error rates (i.e. false positives and false negatives, also referred as omissions and commissions) acceptable? And most importantly, for environmental accountant, are measures of change over time accurate and appropriate? The data me be accurate (another dimension of data quality) for one point in time, but not comparable to another point in time; and therefore not relevant.

How well the data meet the needs of users in terms of the concept(s) measured – its fitness for purpose – need to be explored from a spatial and temporal perspective, and not only from the simple approach required for traditional sample surveys. And what is required to do so, in a way that is compliant with the goals of national environmental accounting in terms of international comparability, is an appropriate SEO Data quality framework.

The bottom line is as follows: If the data was not specifically intended or designed for the application for which it is being used, it needs to be carefully evaluated for its fitness for purpose. This implies a comprehensive assessment for its intended application. In this day and age of Big Data and Mac-statistics characterized by the rapid production and consumption of huge amounts of off-the-shelf products, the desire for a comprehensive evaluation may be limited, but more than ever required.

Luckily, as the proverb goes – "l'appétit vient en mangeant".

Conclusion

"Big Data" is here to stay and we must learn how to use it properly. Learning how to capitalize on these new data streams will take time and it will be important to document the concepts, sources and methods used so that the seven dimensions of data quality may be appropriately understood and reported against. This is essential for comparison between areas (within or between countries) as well as over time.

In this, satellite Earth observation data products have the advantage of offering a uniform basis for global comparison. However, the data need to be carefully vetted for subnational analysis in terms of spatial and temporal precision. Users need to understand the limitations of the products and their fitness of purpose for the intended use. Global data should always be compared to national data to identify and understand discrepancies. In fact, the most important aspect to emphasize here is the need to integrate satellite Earth observation datasets with other data sources and expert knowledge to capture real change and draw valid conclusions from the accounts.

This paper is a starting point for examining data quality issues for the land, water and ecosystem accounting. It recognises that the existing processes of data quality assurance that were developed before "Big Data" are not enough. How the existing processes for statistical offices and other accounting producing agencies can be updated and/or expanded to ensure they are fit for purpose is an open question that we hope the London Group can help to answer.

References

ABS (Australian Bureau of Statistics). 2017. Experimental environmental-economic accounts for the Great Barrier Reef. Australian Bureau of Statistics Cat. No. 4680.0.55.001 http://www.abs.gov.au/ausstats/abs@.nsf/mf/4680.0

ABS (Australian Bureau of Statistics), 2009. Data Quality Framework. ABS cat. no. 1520.0., Canberra (Accessed 18 January 2018). http://www.abs.gov.au/ausstats/ abs@.nsf/lookup/1520.0Main+Features1May+2009.

Amani et al. 2019, Canadian Wetland Inventory using Google Earth Engine: The First Map and Preliminary Results, Remote Sens., 2019, 11, 842

Australian National University, 2018. Earth Observation for Environmental-Economic Accounting, Environment & Society Synthesis Workshop, 10-11 May 2018, Workshop Synthesis Report, Draft for Comment, 11 June 2018.

IMF (International Monetary Fund), 2012. Data Quality Assessment Framework – Generic Framework. IMP, Washington, D.C (Accessed 12 February 2018). http://dsbb.imf. org/Pages/DQRS/DQAF.aspx.

Keith, H. et al. 2018. Contribution of native forests to climate change mitigation – A common approach to carbon accounting that aligns results from environmental-economic accounting with rules for emissions reduction. Environmental Science & Policy. November 2018. DOI: 10.1016/j.envsci.2018.11.001

Loechner, J. 2016. 90% Of Today's Data Created In Two Years. https://www.mediapost.com/publications/article/291358/90-of-todays-data-created-in-two-years.html (Accessed 20 September 2019)

Manning, M., et al., 2004. IPCC Workshop on Describing Scientific Uncertainties in Climate Change to Support Analysis of Risk of Options. Workshop Report. Intergovernmental Panel on Climate Change, Geneva.

Moss, R., and S. Schneider, 2000: Uncertainties in the IPCC TAR: Recommendations to Lead Authors for More Consistent Assessment and Reporting. In: IPCC Supporting Material: Guidance Papers on Cross Cutting Issues in the Third Assessment Report of the IPCC. [Pachauri, R., T. Taniguchi, and K. Tanaka (eds.)].

Pekel, J.-F., Cottam, A., Gorelick N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature, vol 540, 15 decembre 2016, https://global-surface-water.appspot.com/ (accessed 27 september 2019)

Penman, J., et al. (Eds) 2000. Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories. IPCC and Institute for Global Environmental Strategies, Japan http://www.ipcc-nggip.iges.or.jp/public/gp/english/

Statistics Canada, 2002. Quality Assurance Framework. Statistics Canada, Ottawa (Accessed 12 February 2018). http://www5.statcan.gc.ca/access_acces/archive. action?loc=/pub/12-586-x/12-586-x2002001-eng.pdf.

Statistics Canada, 2017. Quality Assurance Framework. Statistics Canada, Ottawa. https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.htm

Statistics Canada, 2017b. Human Activity and the Environment: Freshwater in Canada. Catalogue Number 16-201-x.

Statistics Canada, 2018. Human Activity and the Environment: Forests in Canada. Catalogue Number 16-201-x.

United Nations, Australian Bureau of Statistics, Queensland University of Technology, Queensland Government, Commonwealth Scientific and Industrial Research Organisation, National Institute of Statistics and Geography, Statistics Canada, 2017. Earth Observation for Official Statistics, Satellite Imagery and Geospatial Data Task Team report, 5th December 2017. White cover publication, pre-edited text subject to official editing and subject to further consultation, https://unstats.un.org/bigdata/taskteams/satellite/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf (accessed 27 September 2019)

Vardon, M., 2013. Recognising and Managing Uncertainty in National and Environmental Accounting. Paper for the Valuation for Accounting Seminar, 11 November 2013. Available from: ONS and DEFRA, London, UK (Accessed 21 December 2017). https://www.researchgate.net/publication/272353660_Recognising_and_managing_uncertainty_in_national_and_environmental_accounting