Journal of Official Statistics, Vol. 29, No. 1, 2013, pp. 1–9

Discussion

Svein Nordbotten[1]

1. Introduction

I would like to thank the Editors for inviting me to comment on articles contributed to this special issue of JOS, Systems and Architecture for High-Quality Statistics Production, and congratulate the authors for their interesting and stimulating articles.

I appreciate that the invitation to the discussants provided the opportunity for "*big picture*" comments on the theme, and shall contribute with comments on system framework, data sources and collection, data organization, storage and computation, dissemination and confidentiality with some references to the submitted articles discussing these subjects. A couple of historic references are provided to illustrate previous work on related topics (Nordbotten 1954, 1975).

2. Framework for Statistical Production Systems

The editor group of John L. Eltinge, Paul P. Biemer and Anders Holmberg has prepared an article in which a framework is presented to help discussants and readers see how the different topics are interrelated (Eltinge et al., this volume). This framework (denoted **FW** in the following) illustrates the complexity of modern statistical production systems in which a data set collected for one purpose can be reused for other statistical purposes, how a technological tool can be shared by several parallel processes, how a statistical method or processing algorithm developed for one particular task can be standardized and utilized in a number of other projects, and how extensive the decision and planning tasks become in such systems.

The dynamics of a statistical system are discussed in Section 2.2 of Eltinge et al., and the authors state that *". . .some environmental factors, Z, may vary over time, and thus have an effect on quality, cost and stakeholder utility"*. I would like to elaborate on some factors which I consider extremely important in the context of modern statistical production systems.

There are two types of variations symbolized by elements in Z which should be distinguished. I will call the first type *endogenous* variation caused by feedbacks of previous states of the system itself, but predetermined for the decision maker/planner. I shall denote the second type as *exogenous* variation caused by factors outside the control of the system decision maker/planner. The first kind indicates that in deciding on actions

[1] Professor Emeritus at University of Bergen, Rieber-Mohns veg 1, N-5231 Paradis, Bergen, Norway. Email: svein@nordbotten.com

today, these can have effects on future states of the system at the same time as the effects of previous decisions are inherited by the present planner, while the second kind of variation is caused by factors outside the system and calls for information from outside the system.

Take $U = g_U$ (Q, C, R, X, Z), function (1.2.4) of the **FW**, as a starting point for the design of a separate survey. The function expresses that stakeholder utility U of a survey as determined by the vectors X and Z (Q, C and R are also assumed to be functions of X and Z), where the elements of X denote factors determined by the survey designer and elements of Z represent predetermined factors affecting the survey production.

Consider a subvector partition X_W of X with elements symbolizing data to be collected and saved. Let the partition Z_W of Z denote all historic vectors of this partition X_W, that is be a matrix in which each row is an X_W vector for a previous period

$$Z_W = X_W^{-1}, X_W^{-2}, X_W^{-3} \ldots$$

where the superscripts refer to previous periods.

Obviously, the **FW** can be used to explicitly take into consideration not only specifications for a survey being considered, but also specifications used for previous surveys. According to $Q = g_Q$ (X, Z, β_Q), functions (1.2.1) in **FW**, these historic specifications can affect the quality Q because historic data increase the efficiency of editing and create a base for time series. They can also affect the cost C by symbolizing advantages of lessons learned from previous similar surveys.

Finally, the factors symbolized by Z_W can affect the stakeholder utility. For example, statistics from a considered survey will tend to give more interesting information and a higher utility if statistics from comparable surveys are available. In general, the more statistics are available for a statistical group, the higher the stakeholder utility for this group is expected to be.

In Section 2.2, other dynamic aspects are also discussed as variations in Z vector elements due to exogenous factors outside the control of the survey designer. Let the subvector Z_E represent these elements. A typical example of such an element of the Z_E vector is the total survey budget leaving the allocation within the survey to the designer. Other factors represented in the Z_E vector can be methodological and technical innovations, and the expected demand profile for the survey results.

So far the **FW** has been considered in the context of a single survey. In a modern statistical production system, individual surveys are interrelated and decisions taken for one survey can have an impact on another. Decisions made in other surveys can of course be considered exogenous and represented by elements of the respective Z vector by each survey organization, but the interaction will then most likely be lost and the overall design will end up as suboptimal.

As pointed out by Bo Sundgren, the traditional "stove-pipe" organized statistical systems are being replaced in many National Statistical Institutes (**NSI**s) by integrated statistical systems in which the work on different statistical surveys is closely integrated and interrelations represented in data collected are taken care of (Sundgren 2010). In such systems, the design decisions for a single survey must always be made with the possible effects on other parts of the total system in mind. A similar point is made by Peter Struijs, Astrea Camstra, Robbert Renssen and Barteld Braaksma in their article, in which they write "*In such a situation, one should ideally aim at total network design rather than total survey design. The network is optimised for the same dimensions as is the case when applying total survey design, but the optimisation is global rather than local.*" (Struijs et al., this volume).

Let us consider a **FW** for the whole statistical production system of a **NSI** consisting of a number of more or less simultaneous activities in different surveys. Assume that the **FW** reflects such a statistical production system restricted by an available total budget B and a set of general constraints such as limited, shared processing capacity and data archives symbolized in Z. The permissible set of solutions concerning all surveys under consideration will be represented by X vectors satisfying all methodological, system and administrative constraints as well as the cost function $C = g_C(X, Z)$. Let budget constraint be symbolized by

$$\sum C_i <= B, \text{ and all } C_i >= 0$$

where $\sum C_i$ is the sum over all allocations to surveys (and other activities) represented by elements of C, associated with respective elements of X. Similar constraints will exist for all shared and limited resources.

The overall task for the NSI managers will be to identify among the permissible vectors the vector X^O which maximizes some function of elements of vector U. The form of this function will express the significance assigned to the respective user groups served by the **NSI**.

Just as previous decisions will affect the present situation, present decisions will influence future situations. If stakeholder utilities of *future* periods from present design decisions are to be taken into account, future stakeholder utilities must also be evaluated and discounted. Moreover, these kinds of relations can be interpreted and discussed within frameworks similar to that proposed.

Above we changed the view from a framework for a particular survey to a framework for a total **NSI** production system. Statistical information demands are becoming increasingly global, regional and global cooperation agreements have been made, international statistical organizations have emerged, and the attention may soon be focused on integration of official statistical activities into a *world-wide statistical production system* and the management, trimming and optimizing of such a system. An expanded **FW** will be a no less important subject for future discussions.

3. Data Sources and Collection

The current outlook for **NSI**s is characterized by restricted budgets and a growing trend of unwilling respondents to statistical surveys. On the positive side, demand for statistical information and services is reported to be increasing in new groups of users, and technological development of significance for statistical production systems continues to improve and provide access to new or alternative data sources, making the processes more cost effective.

Allyson Seyb, Ron McKenzie and Andrew Sherett report in their article that an objective for Statistics New Zealand's 10-year programme for change is "*Maximizing the* *use of administrative data to reduce respondent load, minimize costs of data collection, and maximize data reuse*" and expect that "*By 2020, the organization's aim is that administrative data will be the primary source of information, supplemented where necessary by direct collection*" (Seyb et al., this volume).

If the data are measured in volume, Statistics New Zealand must at the moment rely heavily on sample surveys and have a great saving potential in data from administrative sources. I would guess that in many **NSI**s, data collected from administrative sources count at least for twice the volume of data collected solely for statistical use. At the same time, the cost of collecting and processing data from an administrative source is less than half of the cost of collecting and processing data per unit from a statistical survey.

The article by Lisa Thalji, Craig A. Hill, Susan Mitchell, R. Suresh, Howard Speizer and Daniel Pratt describes a developed system NIRVANA (Thalji et al., this volume). The system seems to be designed on the basis of wide experience in survey work and is focused on integration, standardization and use of paradata for improving the quality of the survey results. Particular efforts were made in the project to make the access to survey data and paradata for the managing the survey operations. From a wider perspective, the NIRVANA project illustrates the importance of working within an architecture in which any survey effort can benefit from experiences from previous surveys.

In most countries, there are still many administrative data sources that have not been exploited by official statistics. In the last couple of decades, communication technology has created new and potential data sources for statistical systems. Statisticians have for some time been aware of credit card and electronic transactions as a potential source of data for statistics. Electronic recording could be a data source for rapid production of continuous short-time transaction statistics and indicators of current economic development by transaction locations. In some well-developed and organized systems, the transaction records could even be linked to records from other sources characterizing the transaction actors, and permitting new, current statistics of great interest for users in trade and industry as well as for economic policy makers. Since the recording and the transfer is done in real time and in general without "response errors", data could be ready for processing immediately and statistics be available on a continuous basis.

A number of other electronic records are being generated that can become important data sources, such as electronic records generated by electronic toll gates using Radio Frequency Identification (**RFID**), a common sight along the roads in some countries. Vehicles are electronically recorded, location and time stamped when passing the gates, and identified for toll payment by means of a connected car register. Via a car register, the type of vehicle and its attributes including its owner are extracted. These data can be important additions to the ordinary data in preparing statistics on traffic and commuting patterns.

Some countries have a permanent system of registers for such objects as persons, enterprises, real properties, vehicles, and so on within which interrelations between objects in different registers are also currently maintained (UNECE 2007a). In the car register, links to the car owners are kept up to date, in the population register links to the location of owners' dwellings are recorded, and in the enterprise register links to enterprise locations and links to employers are maintained. In the statistical production system, the interrelation links may be used to expand the description of traffic patterns to include owners' socioeconomic or industry groups.

A third fast-growing future source of administrative records, is the electronic tracking already widely used at sea. Most of the shipping fleet today is equipped with Global Positioning Systems (**GPS**) equipment permitting ships to determine their exact positions at any time. There are systems in operation combining the **GPS** with electronic position reporting via satellites, radio, mobile phone networks and so forth, permitting the transport patterns to be updated in real time. It is usual to supplement these movement records with size and type of each ship as well as its cargo, and so on. Similar systems have been in place for commercial air traffic for many years. They are now also being installed in cars as a theft precaution. In a few years, all kinds of vehicles will probably be tracked more or less continuously, providing a source for national and international transport statistics.

What kind of methodological challenges do such new sources imply? With a nearly continuous recording of events, a large mass of data will obviously be generated and a scheme for time sampling would probably be required. Since we most certainly will be interested in data permitting analysis of changes, should some kind of panel of objects be designed and observed over time? If so, how frequently should the panel be updated to take into account new objects as well as "deceased" objects (people, cars, ships)? In systems, such as those based on permanent registers, permitting the sampled electronic records to be linked to stored microdata for the objects, challenging methodological questions about the kind of estimation scheme to be applied to obtain the best results, will certainly arise.

A second potential source of data is the new social media such as open blogs, Facebook, Twitter, and so on. According to Wikipedia, there are now about 160 million public blogs in existence, of which a large proportion is maintained daily, and every minute of the day 100,000 tweets are sent. Many of these blogs and messages express facts and opinions about public affairs, and they are all open and free to use!

Can important social and economic data be systematically extracted from this stream of blogs and messages and used for the production of useful official statistics? There are blogs and tweets for most kinds of questions debated by politicians and the public in general. In the field of text analysis, a repertoire of methods and computer programs for extracting, interpreting and quantifying text content exists, and can be applied for transforming the contents of textual blogs, tweets, and so on to numerical data for statistical use.

Use of this data source could provide statisticians with data for many of the questions which we currently try to answer by special surveys. We are again challenged by the problem of selecting useful sources from a mass of sources of which a majority probably is of no interest. The solution can be to search for panels of sources proven to keep a given standard of reliability, but new methodological research and development is certainly required.

An interesting consequence of collecting data from administrative sources and continuous streams of data is that the traditional strong relationship in time between survey collection and processing of data becomes less important. Some of the peak workloads associated with censuses have already disappeared in countries producing census statistics from administrative data. Data will be collected and saved to serve future needs, and processing may be carried out when needed, relying on already collected and saved data.

4. Data Organization, Storage and Computation

In many statistical organizations, after being processed and used for primary publication microdata are saved in some kind of statistical archive (warehouse, database) for reuse in some future connection which can be specified by a sub-vector Z_W of the Z in **FW**. By saving microdata for future use, we can obtain a higher future performance because more microdata are available for each object of the population, permitting a wider set of tabulations and the creation of more realistic analytical models.

It will be useful to distinguish between microdata and macrodata stored in the statistical archive. A cube or **OLAP**-approach, as adopted by Statistics NZ and many other **NSI**s, has proved to be a good strategy for storing macrodata and serving requests for special statistics. Systematic storage for microdata is, however, needed for responding to requests which cannot be served by the cube storage or for serving researchers' requests for access to microdata. Such user needs can be served by a cumulating system of cleaned microdata files on which it is possible to remove unwanted records, link object records from different files, directly and by reference, and to store the results as a new generated microdata file ready to serve future requests. The cost of such system readiness is data redundancy, since the same data will appear in more files. An alternative approach, considered and discussed for decades, is to record each atomic data element and retrieve the data for each request (Sundgren 1973); the "single-cell" approach of Statistics NZ seems to be a partial realization of Sundgren's model.

For effective macrodata and microdata storage and retrieval, a well-developed metadata system is crucial. It must be automatically maintained and updated for each transaction/operation on data in storage. Efficient reuse of data is completely dependent on a well-structured metadata system from which users can obtain information to form their requests. We can think about representation of metadata system effectiveness in the **FW** as symbolized by an element of the Z_W vector indicating the coverage of the metadata system in the past, while a corresponding element of the X_W subvector expresses a planned improvement of the metadata system.

There is another important issue linked to integrating microdata. The more data you have collected about an object, the easier it will in general be to identify errors in the object data. This creates a dilemma: What to do when editing of newly collected data for an object reveals errors in old data included in the archive and already used in several productions? If the **NSI** policy is not to correct microdata in the archive because of desired consistency between earlier published statistics and the archived microdata, the producer will have to ignore errors in previously released statistics even though new statistics then may not be consistent with statistics previously published. On the other hand, if the **NSI** policy is to improve the quality of all statistics released, the producer must correct data in the archive, rerun all productions based on the corrected data set, publish correction messages and notify all users who have received statistical services based on the erroneous data. My guess is that the latter policy in the long run is unrealistic, and that **NSI**s must admit that errors always can appear in statistics. A compromise will be to make corrections in data less than *N* years old, and keep the newer data files in a preliminary storage before transfer to the archive.

In Nordic and some other countries, the introduction and maintenance of official central registers for population, enterprises, dwellings/properties, and so on including unique and permanent object identifiers have been a political goal in order to reduce inconvenience to the public and government registration efforts, and save expenses. The official registers are also in some cases available for private use, for example by banks and insurance companies. This fact makes data from administrative sources particularly useful for the **NSI**s, since the data records received from administrative organizations are pre-identified and data from different sources can be linked as if they originated from a huge single source. Because of the permanence of the identifiers, records from different years or periods can be linked and microdata time series created from all contributing sources (Nordbotten 2010).

Except for estimation methods used in the context of sample surveys, the computation of statistics has been dominated by simple aggregation and has remained unchanged for centuries. An important discussion is going on among survey statisticians about design-based versus model-based statistics. Is the time due for a general paradigm shift? In statistical processes like data collection, editing and imputation, attribute values are frequently considered realizations of stochastic variables. Why not consider any collected data set, for example a population census, as a random sample of a superpopulation and use the huge amount of existing paradata to develop quality measures for the estimates published? In such a scenario, a cell-based storage structure as outlined by Seyb et al. can be interesting component.

5. Dissemination and Confidentiality

The objective of official statistical production systems is to provide, with their allocated resources, statistical information requested by a wide set of user groups. Traditionally, this objective was fulfilled by publishing statistical tables, frequently designed before data were collected and based on the needs expressed by user groups dominated by state authorities. In more recent times, statistical information systems have been regarded a common service to users from all organizations, industries, education, research, media and the general public by responding to special requests.

The dissemination service can be provided in different ways, as discussed by Seyb et al. There are, however, two aspects, *instant online service* and *access to microdata*, which are of particular interest in the context of modern statistical production system architecture (UNIECE 2007b). They have in common that both create disclosure concerns.

In their article, Tom Krenzke, Jane F. Gentleman, Jianzhu Li and Chris Moriarity discuss challenges in planning and preparations for a web-based Online Analytical System (**OAS**) (Krenzke et al., this volume). The objective of the **OAS** is to serve researchers and other users with special needs with data from the National Center for Health Statistics (**NCHS**). The authors are well aware of the problem of balancing user service and disclosure risk. To provide the data requested, the data have to pass the Statistical Disclosure Control (**SDC**) treatments. How to control the risk for disclosure subject to the anticipated attack strategies of an intruder are discussed and solutions outlined. Even though the discussion is limited to the survey environment of the **NCHS**, the discussion also can be instructive for statisticians working under more general conditions.

It can be interesting to consider disclosure in the context of the **FW**. Let disclosure risk for a specified type of microdata be defined as

$$C_d = p_d \times D_d$$

where risk C_d can be regarded as a cost element in vector C of the system, p_d is the probability for disclosure of the considered microdata and D_d is the damage suffered by the concerned physical person(s) or owner(s) of a harmed object due to this disclosure. Reducing C_d means usually restricting the available statistical information (reducing U, P, Q) and increasing the treatment cost, another element of C. A discussion of the disclosure risk in terms of objective estimates for the probability of disclosure and the size of the damage if the data are disclosed is rarely seen and would be useful.

How to provide access for researchers (and why not also other users?) to microdata files without disclosing confidential data? This has become an extremely important question when real-time online service systems are being considered. Several approaches, including removal of all identifying elements, distorting data by random errors, authorization arrangements and so on, have been proposed and tried. They have all their pros and cons.

One strategy can be that the online researchers (users) are given access to "invisible" microdata files, that is, are restricted to seeing shadow views of the relevant microdata files. In shadow views, the viewer cannot see any attribute data, but whether attribute values are present or not, whether a value was observed or imputed, and so forth. The users can be permitted to create new microdata files by operating on existing files, form new files by integrating existing files, study their shadows and carry out analytic computations on data based on a repertoire of approved analytical programs. The disclosure problems of researchers' online reuse of microdata would then be limited to the screening of analytical results and be similar to those discussed in the article by Krenzke et al., and the same **SDC** treatment could be relevant.

6. A Final Comment

Any architecture for statistical production systems has to be designed in accordance with statistical (and other relevant) legislation. Several ideas discussed above would in some countries require modifications of existing statistical laws. Some NSIs have the advantage of one or more official identification systems. Do they have the required basis for legal access to administrative data from all organizations using the systems? Are the legal basis and the conditions for an NSI to store and integrate microdata for an indefinite time period satisfactory? Is the legislation safeguarding a confidential transfer and treatment of microdata up to date? Does the legislation permitting access to microdata reflect a fair balance between the public needs for information and required confidentiality? Statistical laws need periodic modifications to satisfy social and economic development and represent a significant premise for the statistical system architecture. Statistical legislation should therefore be a part of future discussions of statistical system architecture frameworks.

7. References

Nordbotten, S. (1954). Theory of Production and Production of Statistics, Translation of a paper Produksjonsteori og statistikkproduksjon in Norwegian published in Stimulator, 7, 28–31. Available at: http://nordbotten.com/articles

Nordbotten, S. (1975). On the Design of Statistical Information Systems. In Systemering 75, Festskrift till Børje Langefors, J. Bubenko and M. Lundeberg (Eds), Lund: Studentlitteratur. Available at: http://nordbotten.com/articles

Nordbotten, S. (2010). The Statistical Archive System 1960–2010: A Summary. Nordisk Statistikermøde i København 11.–14. august 2010. København. Available at: http://nordbotten.com/articles/

Sundgren, B. (1973). An Infological Approach to Data Bases. Stockholm University and Statistics Sweden, Stockholm. Available at: http://sites.google.com/site/bosundgren/my-life/AnInfologicalApproachtoDataBases.pdf

Sundgren, B. (2010). The Systems Approach to Official Statistics, Chapter 18 in Official Statistics in Honour of Daniel Thorburn, 225–260. Available at: http://officialstatistics.wordpress.com

UNECE (2007a). Register-Based Statistics in the Nordic Countries. Review of Best Practices with Focus on Population and Social Statistics. United Nations Economic Commission for Europe. Available at: http://www.unece.org/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

UNECE (2007b). Managing Statistical Confidentiality & Microdata Access. Principles and Guidelines for Good Practice. New York and Geneva: United Nations.