WP.12
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (ii): Global solutions to editing

Two Paradigms for Official Statistics Production

Prepared by Boris Lorenc, Jakob Engdahl and Klas Blomqvist, Statistics Sweden, Sweden

Abstract

The paper discusses modern systems for statistics production in the perspective of knowledge systems (or, cognitive systems). It is argued that there is a parallel between how the work on modelling statistics production systems proceeds - namely, by defining models for processes and for information - and how cognitive systems have been conceptualised in cognitive science/artificial intelligence. It should however already at the outset be noted that the information that we focus on is about subject-matter facts, not about the processes. It is the accumulated amount of that information that we refer to as knowledge (or cognitive) systems.

There is an ongoing work to integrate statistics production over subject matter areas by using same/similar tools, processes, and data structures. Also, there is an ongoing tendency to strive towards integrating register and survey data for the purpose of producing statistics. These goals are among top priorities for modern systems for statistics production. In that light, we review some evidence about possibilities of building cognitive/knowledge systems that can satisfy the ever increasing need for representation of facts in the system. We argue, on the basis of the review, that it is feasible to be inspired by other approaches within cognitive science (e.g. subsymbolic cognition, as in artificial neural networks and genetic algorithms, or distributed cognition), in addition to relying on the classical, symbolic cognition. This also includes more recent ways of structuring information, like the NoSQL approaches represented by, for instance, Google's BigTable, and the Semantic Web (e.g. the Resource Description Framework), as these ways circumvent some of the shortcomings of the classical approach.

With respect to the issues raised, we discuss in the paper how statistics production can be driven even more by user needs, and consider how modern approaches to design of knowledge systems, large data bases, and so-called Big Data might be related to the production of statistics. Regarding the field of editing, we discuss alternatives to when it could be taking place, with what (if any) purpose(s), and propose to investigate applying self-learning editing rules in the editing process. Also, opportunities for using the Bayesian inferential approach are touched upon.

I. Introduction

0. This paper has been encouraged by Li-Chun Zhang after early conversations with him on some of the approaches argued for herein. This support is thankfully acknowledged. The views expressed here are those of the authors and do not purport to reflecting the policies of Statistics Sweden.

1. An ongoing, vigorous and dedicated development has been taking place for some years within official statistics, with the aim to transform the way that statistics is produced from a stovepipe model to a model of integrated, comprehensive systems for statistics production (European Commission 2009, HLG-BAS 2011).

2. However, the development is ongoing but not completed. Thus, we still do not know whether the promises of these comprehensive systems will be fulfilled. In particular with respect to this paper, whether it will be possible to achieve:

i) better integration of administrative data and survey data,

ii) better/faster response to new or changing user needs,

among the goals that the integrated, comprehensive systems are envisaged to achieve (Axelson et al. 2011). With much of the current focus on standardisation and integration of production tools, the goals (i) and (ii) have perhaps received somewhat less attention.

3. To address the issues in §2, the question can be posed how such a system should look like so that it satisfies these requirements. While likely glossing over some developments, let us suggest that this question may be usefully considered in terms of knowledge systems or, equivalently for the purposes of this paper, cognitive systems (these being introduced and the distinctions discussed in Section III).

4. Next section establishes the motive to delve into the topic. The section after that introduces knowledge systems briefly, including some distinctions and related properties. Relevance of these properties for statistics production is discussed in Section 4. Section 5 provides some off-shots of these considerations also of possible relevance for statistics production and, in particular, editing. Section 6 offers some concluding remarks.

II. A widening content of databases for statistics production

5. An assumption for making relevant this departure into knowledge systems will be that the database that is a component of a system for production of official statistics - be it a stovepipe or a general system - plays a role of a knowledge system pertaining to that domain. Two comments are in place here:

i) the singular “database” in the preceding sentence may refer to a single or a number of interrelated databases: in other words, a database system, which in some simpler cases can be a single database;

ii) we will defer treating the question that invites itself, namely whether the database *is* the knowledge system or just an important component of it.

6. Provided that the database is to be viewed as an equivalent to a knowledge system pertaining to a domain of interest for statistics production, it is argued here on intuitive grounds that the knowledge system that is a component of a stovepipe system for statistics production contains fairly little knowledge, compared to the database that that is a component of a broad, general system for statistics production, which contains much more knowledge.

7. Again intuitively, a database that is a component of a stovepipe system for producing, say statistics on companies’ investment into IT, will contain a certain number of variables from the business register (the company’s number of employee, its NACE classification, perhaps turnover and some others) plus a number of survey variables for the surveyed part of the population on the subject matter of interest. Information which can be found in this database will typically be very well known to the team producing this statistics, as well as documented in the declaration pertaining to the statistical programme and in the metadata that pertain to this survey.

8. Compare this just described to a broad, general system for, say, production of business statistics. Such a system would need to be created by carefully incorporating a number of databases pertaining to the systems (stovepipes) it is to replace, thus broadening considerably the domain of such a database. While keeping the same requirement of documenting its content perhaps in the form of metadata pertaining to the broad system, is fairly unlikely that the database’s content will be well known to the production team - this is neither a requirement nor a goal. It will even less likely be known to a user. On the contrary, an automatic process needs to be envisioned.

9. However, it should still be possible - as per our assumptions in §2 - to add new survey data or administrative data, and this also in a way that meets users' needs faster and better. The latter perhaps also implying that users themselves should have access to the content of databases/knowledge systems for production of official statistics and gain an insight into what they can obtain there (and at what cost and with what quality) - that is, that they themselves be able to search, "browse" and in other ways use these systems. (Compare this with the Semantic Web, in §19.)

III. Knowledge systems: their properties and limitations

10. The complexities outlined in §9 lead us closer to knowledge systems such as those that have arisen together with the advances in cognitive science and computer technology; that is, in the preceding fifty to sixty years (but that have precursors in the 18th century and before, going back to antiquity).

11. "A knowledge base is a special kind of database for knowledge management. A knowledge base provides a means for information to be collected, organized, shared, searched and utilized" (Wikipedia "Knowledge base"). Think, then, of the whole broad, generalised system for statistics production as a knowledge base, or a cognitive system. (As indicated earlier, in §3, it is at the moment not vastly important to put forth a very precise definition.)

12. Cognitive systems can be divided roughly into those built upon computational models and other (noncomputational) models. The former are then divided into symbolic and subsymbolic models, whereas the latter category branches into a number of different approaches (like embodied cognition, socially distribute cognition, etc). (Wikipedia "Cognitive Science", section "Computational modeling").

13. As an example of a symbolic system, an early success in artificial intelligence was ELIZA, a computer program for natural language processing that interacted with the user by simulating a Rogerian psychotherapist (Weisenbaum 1966). Advances in the meantime have led to a number of more advances chatterbots, however most of them are based on pattern matching (i.e. involving little formal reasoning); and "there is currently no general purpose conversational artificial intelligence" (Wikipedia "Chatterbot").

14. The formal reasoning, just mentioned parenthetically, would involve a first-order predicate logic, or other way of formal and symbolic deduction. Whether formal reasoning can be said to be the basis for human cognition has been a debate ongoing practically since the establishment of the modern-day cognitive science in 1950s, with the outcome still likely not settled, but with a lot of indications that a viable declarative cognitive system is not the basis of natural cognition nor that it can provide a satisfactory basis for artificial cognition (or at least very difficult to achieve). We will return to this discussion later, while treating databases (in §17).

15. Another early artificial intelligence/cognitive science success - this example involving subsymbolic systems introduced in §12 - is ALVINN (Autonomous Land Vehicle In a Neural Network), an artificial neural network that in real time steers a modified vehicle on real roads (Pomerleau 1991). A recent development of this kind (not necessarily involving subsymbolic systems) is Google's driverless car, that in addition to having sensing equipment like cameras and lasers, also uses Google Street View and the GPS technology to build a better representation of its position in the environment (Wikipedia "Google driverless car").

16. However, the developments described in this section, while notable achievements, have not provided such a progress that warrant a reliance on completely automatic processes. While artificial systems can be made, they function on restricted domains and need human involvement to guarantee that appropriate corrective actions are taken when needed and, in general, a "safe fail" performance of the system exists. Further, if there will ever be such a system, it is not likely to rely completely on first-order predicate logic.

17. The model underlying the most widespread form of database management systems, that of relational database management systems (RDBMS), is a relational model, which in turn is based on first-order predicate logic (Wikipedia “Relational Model”). To the extent that the conjecture in §14 of the difficulty (or impossibility) of predicate logic to be the basis for viable automated cognition holds, this may provide a basis for considering to what extent RDBMS systems are feasible as the basis for knowledge systems on which comprehensive systems for statistics production should be built.

18. It is notable, in the context, that some of the currently thriving approaches to database construction, specifically such that encompass vast amounts of data, are non-RDBMS. The examples include Google’s BigTable, as well as solutions underlying some functions on Amazon, Twitter, and Facebook, among many others (Wikipedia “NoSQL”). The development here is going under the heading of NoSQL (i.e., no SQL: no use of SQL, a structured query language used to access RDBMS). An important property is reliance on building databases in a schema-less way. As the notion of a database schema plays the same role as the notion of theory in predicate calculus (Wikipedia “Database schema”), there is a clear connection between this property and the discussions in §§14 and 17.

19. Finally, the initiative of the Semantic Web should be mentioned. It is a “collaborative movement... that promotes common formats for data on the World Wide Web”. By formalising the inclusion of semantic content in documenting web resources, it “aims at converting the current web of unstructured documents into a ‘web of data’” (Wikipedia “Semantic Web”). Having as one of its goals creation of “a web of data that can be processed directly and indirectly by machines” (ibid., quoting Tim Berners-Lee), the initiative is closely related to the tasks that will need to be carried out when accessing for use or update the databases of comprehensive systems for statistics production. Developed in conjunction with the Semantic Web, but perhaps still relevant for official statistics production, is the Resource Description Framework, a set of specifications, originally designed as a metadata model, that is coming to be used as a general method for conceptual description or modelling of information on the web (Wikipedia “Resource Description Framework”). This is a way of including semantic, symbolic information, but not with a reliance on predicate logic or other deductive models. RDF has recently been used by data.gov.uk to publish datasets (Joinup 2011).

20. A comment perhaps appropriate for this section is that a reader might comment that we are interchangeably referring to a number of vastly different concepts as almost synonymous, while they are not that (knowledge systems, cognitive systems, artificial intelligence systems). However, what characterises them all - and what we believe allows for this kind of generalisation in a treatment with a specific goal such as this - is that they incorporate a way of representing reality, and that they aim at either support for or direct performance of automated, intelligent, human-like action. As the systems that are being built within official statistics (c.f. §1) aim at exhibiting such properties, without implying exactly how, we believe that for the present purposes it is appropriate to view these approaches as approximately similar.

Continuing with the comment, what perhaps distinguishes the most recent approaches to construction of database-like storages and functions (§ 18) could be a change of focus from that on content to that on use and function.

IV. Consequences for statistics production

21. The two paradigms alluded to in the title refer to:

i) one which assumes having a production system that corresponds to a small, limited domain, where it is possible to set up a RDBMS whose content is updated ‘manually’ (i.e. with the human aid);

ii) another which assumes a production system that corresponds to a very broad domain, where the database is built such that it allows for automated search and update, and that cannot count on having a human input for that.

That is, the paradigms are the same as in the “stovepipes” *versus* “comprehensive systems” dichotomy, but such differences are highlighted that refer to content and use of the database (in a broad sense) of the system.

22. The former paradigm (stovepipes) implies an ‘expert’ interface between the user and the system (in particular, the database); the latter (comprehensive systems) does not, by its construction principle.

23. Consequently, in order to integrate administrative and survey data, the former would both need more time (due to human involvement) and likely be more exposed to quality issues. Further, while the former would seem to require an interface that interprets the content to the user, the latter is by construction geared to allowing the user to themselves investigate its content and its fitness for use and perhaps to themselves consider cost and quality trade-offs if deliberating between sole use of administrative data and the addition of newly collected survey data.

24. A related consideration, treated recently by Laitila in a somewhat different context of evaluation of quality of administrative data, is that of collecting data for single purpose of producing statistics versus multi-purpose (Laitila 2012). In the stovepipe paradigm, data are collected for a specific survey or a set of closely related surveys; once used, the data are at best used once more, for the editing purpose of the next data collection, and at worst discarded. In the comprehensive systems paradigm, the data are “archived” (Laitila 2012), that is - in terms of the current paper - they are integrated into the database for future, and not necessarily specified, use in statistics production. This in turn invites some further reflections.

25. Survey weights are a *sine qua non* of statistic production in the so-called design-based approach to survey sampling and inference (Särndal, Swenson and Wretman 1992). Originally being inverses of inclusion probabilities, these can be thought to reflect ‘lack of bias’ in data collection: for instance, the so-called simple random sampling reflects the value that every member of the population has an equal chance of being drawn into the sample for a survey. However, in order to improve estimates for a particular survey, weights have become more complex through being tailored for particular estimates, at the same time reflecting a specific understanding of other circumstances related to the concrete data collection (e.g. nonresponse) (Deville and Särndal 1992, Särndal and Lundström 1997). By tying the weights to a specific survey or a set of related surveys, survey weights appear to be aligned with the stovepipes paradigm of statistics production. (A possibility to develop several sets of weights, each set referring to the same set of original data but tailored for different uses, might however be a way of preserving the sampling weights approach within the comprehensive systems paradigm.)

26. Discarding sampling weights is not novel in sampling theory; in fact, such a move - if done within official statistics - would bring it closer to other fields of statistics (Valiant, Dorfman and Royall 2000). Thus, potentially, a weights-less approach to statistical inference is better aligned with the comprehensive systems paradigm. In fact, as the weights are commonly a function of some additional variables considered to be of relevance for a particular data set, one may consider adding these variables into the database, and indicating in the metadata their relevance, thus enabling building of whatever weights a subsequent user might want to create. (Given that the additional variables used to construct sampling weights often are taken from relevant administrative data, these will likely already exist in the database. The question of knowing their relevance for a particular problem will be more difficult to address.)

27. The preceding parenthetic comment, as well as the whole of §26, brings in the question of building and discarding the knowledge provided by a specific set of data. Where does the building of knowledge related to a set of produced statistics begin? How is the knowledge discarded or replaced by new one? In the stovepipe paradigm, this would in theory be after the data have been collected: before the survey, we knew little about the phenomenon under investigation (*tabula rasa*), and after having collected and analysed the data, we can give answers to all pertinent questions. In fact, some knowledge is put into the weights by selecting relevant

additional variables using which to form weights, however this decision process is commonly left outside of any formal treatment, thus without rigour and transparency.

28. In a contrastive approach, we would have models that purport to represent a condensation of some knowledge, and we would update our knowledge by adding the newly gathered data. A model-based approach to survey sampling and inference is already fairly well established (Valiant et al. 2000), and there is an increasing work in bringing the Bayesian approach to official statistics (Little 2012), with which to take care of a rigorous and transparent process for bringing in and updating the knowledge within a broad system for statistics production.

V. Further considerations, regarding statistical editing

29. The issue treated by Laitila (2012), that of evaluating quality of administrative data (cf. §24), is closely related to statistical editing, the differences in principle perhaps being the level that is addressed (this difference diminishing for macro editing) and what would a possible intervention consist of (e.g. whether there is an option to go back to the data source and verify the data or not). Thus, statistical editing can be done "for a purpose", that is, in the course of producing statistics. This approach, which largely coincides within the current day editing as we know it, is the one attributable to the stovepipe paradigm.

30. A question less well addressed is how statistical editing can be done "without a purpose". That is, how to do data editing for data that are collected but without a clear purpose of their exact use for statistics production. While the authors find it sufficient for the purposes of this paper to open this question for discussion and possible further contributions, we will throw in some ideas.

31. For one, one can describe data in a standardised manner (e.g. RDF, cf. §19) and verify its adherence to the statements therein.

32. Secondly, there exist methods mostly developed on the computer-intensive side (rather than the statistical side) of data processing, like pattern matching and other similar technologies for which subsymbolic systems (like unsupervised artificial neural networks, cf. Wikipedia "Artificial neural network") could be used to signal possible data values in need of editing attention. While not representing any 'absolute truth', these systems might be useful components applicable to "general-purpose" editing.

33. Thirdly, the data might pertain to some other data already in the database, the relation perhaps described by some model. Then, deviation from the model expectation could be an indication of the need for editing (in the broad sense of editing, not necessarily implying verification and correction).

VI. Discussion

34. The main intent of this paper was to stimulate consideration of some alternative approaches within production of statistics to the ones currently employed. The paper, in particular, focused on alternatives for the database design, presenting some of their properties in the light of cognitive systems. The paper went on to relate the developments to two paradigms for statistics production, the so-called stovepipes *versus* comprehensive systems paradigm, reflecting on consequences of the two identified paradigms for statistics production generally and for editing in particular.

35. The analysis was also an opportunity to go further than usual into the developments somewhat outside the current domain of official statistics. In doing that, initiatives such as the Semantic Web were encountered, that while definitely are addressing issues of relevance to acquisition of data for official statistics production and to dissemination of produced data, are fairly rarely treated within official statistics. Other such developments of possible relevance for official statistics include RDF, metadata, and similar. We believe official statistics should get to know

better these developments and interact with them, which will likely both benefit official statistics and increase its relevance in a wider community.

36. On a closing note, this paper should not be interpreted as advocating that a particular of the two paradigms for production of official statistics is to be preferred or that a particular of the approaches to building databases is to be used therein. (In particular, it might well happen that in some years it will turn out possible to build first-order predicate logic systems for huge domains, including manipulating them automatically for search and retrieval - although the total cognitive science experience alerts us not to bee too optimistic.) Instead, its aim is just to open up for more than a single way of building databases for official statistics purposes, and to initiate discussion about ways of evaluating advantages and drawbacks of these alternative approaches.

37. Therefore, a disclaimer that the issue addressed in the paper concerned subject-matter data and its representation (or lack thereof) in knowledge systems for statistics production. There is, however, a different layer of systems for statistics production, and it concerns production processes. There exist data that pertain to this layer - the different parameters, paradata, and so on - together with descriptions of the processes themselves. These data and these processes, to the extent that they are well structured, are likely to be susceptible to being treated in a RDBMS manner. Consequently, building a formalised system for statistics production (whose basis perhaps GSIM will be, once finalised) should not be impossible, at least not on the basis of the considerations presented in this paper.

38. Further, as a general aside, rather than inventing a wheel in our own field, it might be preferable to look into the neighbouring fields and see if in any of them a wheel already exists and is usable for our own purpose; or, that people in some field might already be at work on a wheel quite similar to the one that we are needing, so learning from them or joining forces with them would make the work faster and perhaps better, especially if they are more skilled than we in making wheels. For instance, the earlier mentioned source in §19 on application of RDF to publish datasets mentions that the application uses a number of existing RDF vocabularies (Joinup 2011):

i) SKOS for concept schemes,

ii) SCOVO for core statistical structures (Joanneum n.d.),

iii) VoiD for data access,

iv) FOAF for organisations,

v) Dublin Core Terms for metadata.

While not vouching for any of them, it would be interesting to see whether or how they match the similar work initiated or existing within the official statistics domain, and what could advantageously be borrowed, used or reused (which is one of our current catchphrases) therein.

39. And finally, a question proposed for discussion. Assume we are moving away from meticulously and manually built “cathedrals” of structured knowledge and going towards “bazaars” of coexisting, loosely related datasets - not necessarily coherently structured - suitable for automatic manipulation. Are we not thereby also moving from a structural (or, equivalently, expressive) approach to knowledge management (database management) in official statistics to a functional approach, where the functions are mainly those of search and use. As a corollary, is not a change also from us (survey methodologists, in a very wide sense) being interpreters of users needs to that of us becoming enablers of users’ direct use (search, analysis) of the data? If this is a part of the changes that are about to happen, then it is probably good to be prepared.

References

Axelson, M., J. Engdahl, Y. Fossan, E. Holm, I. Jansson, B. Lorenc and L.-G. Lundell (2011). “Enterprise Architecture Work at Statistics Sweden”. In *Proceedings of Statistics Canada Symposium*

2011: Strategies for Standardization of Methods and Tools – How to get there. Ottawa, QC: Statistics Canada.

Deville, J.-C. and C.-E. Särndal (1992). “Calibration Estimators in Survey Sampling”. *Journal of the American Statistical Association*, Vol. 87, pp. 376-382.

European Commission (2009). *On the production method of EU statistics: a vision for the next decade*. COM(2009)404. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF.

HLG-BAS (2011). *Strategic Vision of the High-Level Group for Strategic Developments in Business Architecture in Statistics*. Genève, Switzerland: UNECE.

Joanneum (n.d.) The Statistical Core Vocabulary (scovo). http://sw.joanneum.at/scovo/schema.html

Joinup (2011). *Data.gov.uk uses the RDF Data Cube vocabulary to publish datasets*. http://joinup.ec.europa.eu/news/datagovuk-uses-rdf-data-cube-vocabulary-publish-datasets

Laitila, T. (2012). “Quality of Registers and Accuracy of Register Statistics”. Presented at *Q2012*, Athens, Greece. http://www.q2012.gr/articlefiles/sessions/23.1_Laitila_Quality%20of%20registers.pdf

Little, R.J.A. (2012). “Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics”. To appear in *Journal of Official Statistics*.

Lundström, S., and C.-E. Särndal (1999). “Calibration as a Standard Method for Treatment of Nonresponse”. *Journal of Official Statistics*, Vol.15, No.2, pp. 305–327 .

Pomerleau, D.A. (1991). “Efficient Training of Artificial Neural Networks for Autonomous Navigation”. *Neural Computation*, Vol. 3, No. 1, pp 88-97.

Weizenbaum, J. (1966). “ELIZA — A Computer Program for the Study of Natural Language Communication between Man and Machine”. *Communications of the Association for Computing Machinery*, 9, pp. 36-45.

Särndal, C.-E., B. Swensson and J.H. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

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