Restructuring Statistical Processes at Statistics Netherlands

Some theory and practice on combining administrative sources and survey data through micro-databases

P. Kooiman
A.H. Kroese
Methods and Informatics Department
Statistics Netherlands
P.O. Box 4000
2270 JM Voorburg
The Netherlands
PKMN@CBS.NL
AKSE@CBS.NL

1. Introduction

At Statistics Netherlands the statistical processes are undergoing a profound restructuring. This is motivated by the rapidly increasing availability of extensive register based source data, political pressures to cut down the response burden, as well as the wish to produce more consistent outputs. Also the very high nonresponse rates in the Netherlands, especially for person and household surveys, make it necessary to exploit administrative data sources to the maximum in order to be able to master the risk of nonresponse biases. The newly designed statistical processes no longer take separate data collections or surveys as their point of departure. Instead, for a population to be described we match all available information at the micro-level, both register based and surveys, and from the population micro-database so obtained we subsequently derive a fully consistent set of estimates. In the paper we describe briefly the concepts and methodology behind this new approach, and we indicate what has been achieved so far and what problems have been encountered.

2. Four databases

The new statistical processes (as well as the new organization of our institute) are going to be built around four (types of) databases, see figure 1.

All source data are entered into our *input-datawarehouse* "Baseline". Data from different sources (administrative registers, surveys, EDI) are matched at the level of the statistical units delineating the population involved. Since the observational units are not necessarily equal to these statistical units all input-data are transformed into data about statistical units in Baseline. A number of elementary controls and edits of an administrative nature are applied at this stage of the process.

Subsequently *population micro-databases* are constructed for a number of object-types (persons, jobs, businesses, dwellings, etc.). Some of these micro-databases are interrelated. A micro-database contains all information about that object-type as present in Baseline, but now thoroughly edited and partially imputed. Moreover at this stage the input concepts underlying the data in Baseline are being translated into the output concepts underlying the statistical estimates we wish to publish. As a consequence, to construct the micro-databases, the information in Baseline

needs to be harmonized conceptually. In principle a micro-database is a flat rectangular array listing on one dimension all statistical units in the population and on the other dimension all variables that we have some information about. For many variables this information is partial only, either because it derives from incomplete registers or from sample surveys.

All estimates Statistics Netherlands considers worthwhile publishing are collected into a *library of estimates* "StatBase". The information in StatBase should be reliable (methodologically correct), conceptually and numerically consistent (no contradictions may be obtained by confronting data in StatBase) and "safe" (all data in StatBase satisfy the rules for disclosure control).

Our *output-datawarehouse* "StatLine" can best be seen as a collection of views on StatBase. In StatLine the aggregates are presented in a user-oriented way, i.e. the aggregates are arranged in (multi-dimensional) tables (called data-cubes) that reflect 'areas of interest' or 'themes'.

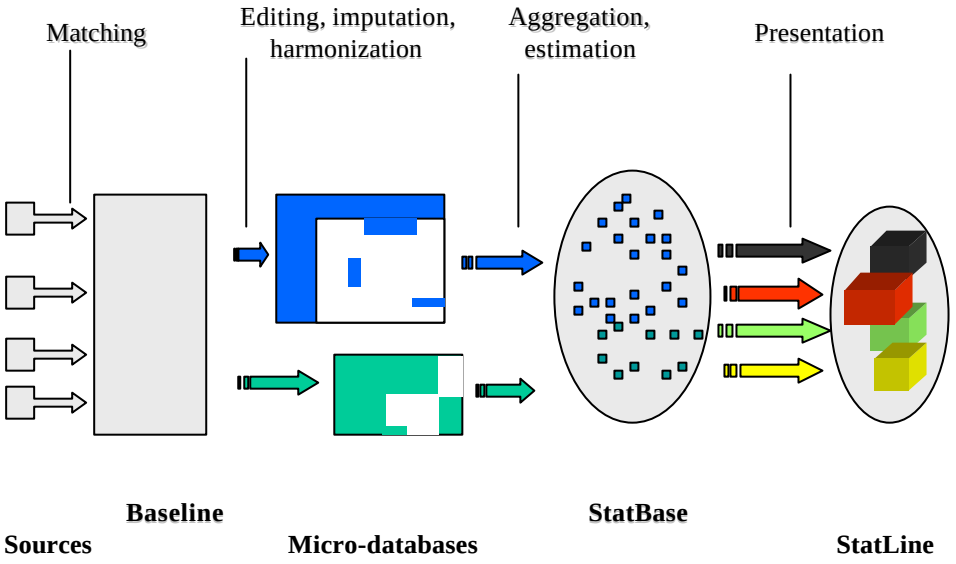

Figure 1: the four (types of) databases in the new statistical process

3. Micro-databases

3.1 Example

To make the idea of micro-databases clear, in figure 2 simplified versions of the micro-databases for persons and jobs are sketched.

The backbone of the micro-database for persons is the Municipal Base Administration, our population register containing information about sex, age, place of usual residence etc. Matched to this register are a number of surveys and incomplete registers. Examples of the surveys are the Labour Force Survey and the Living Conditions Survey; these surveys have common questions, but no common respondents. An example of an incomplete register is the register about educational attainment, constructed from records of the Employment Exchange.

The backbone of the micro-database for jobs is a register of all jobs, constructed by combining information from various administrative sources (including tax records). Among others,

it contains information about salary. The survey (partially obtained by EDI) contains information about hours of work, etc.

The two micro-databases are related as jobs belong to persons: a person can have zero, one or more jobs. At this moment we also have related micro-databases for social security benefits and dwellings.

3.2 Editing, imputation and harmonization

To construct the micro-databases, the matched data in BaseLine have to be edited, partially imputed and harmonized.

Editing of the micro-data can, in principle, be done record-wise: in detecting and correcting errors we can make use of all matched information. For example, in editing the survey in the micro-database about jobs, we can (and do) exploit matched information in the register.

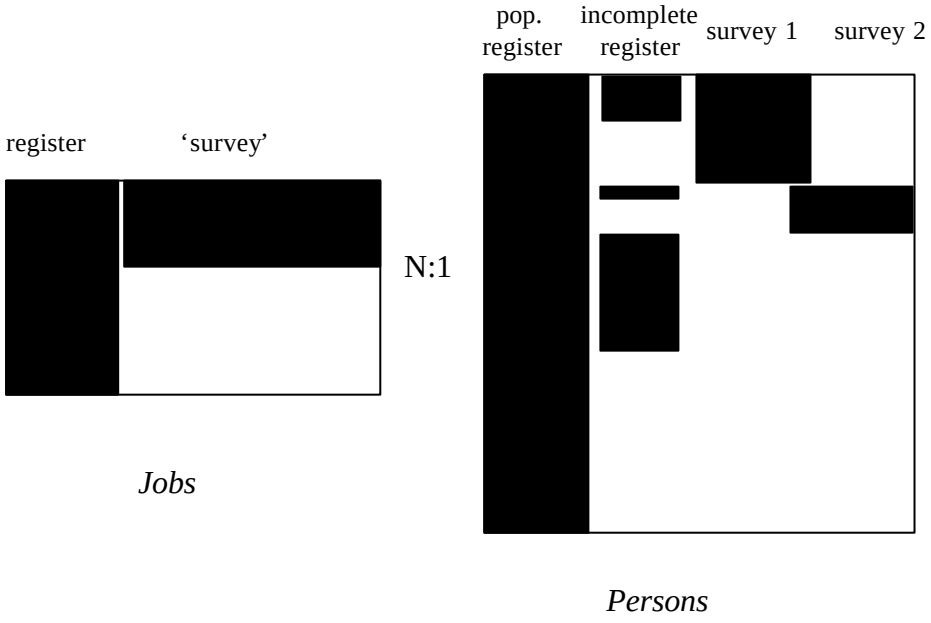

Figure 2: simplified versions of the micro-databases for persons and jobs; black indicates available data, white indicates missing data

Partial imputation is employed for item-nonresponse in a survey. It is also needed when an output variable can only partially be derived from the information in a register. An example of the latter is the variable 'position in the household'. This variable in the micro-database for persons can only partially be derived from the information in the Municipal Base Administration. For persons living together unmarried, without kids, it cannot be seen in this register whether they are just roommates or an unmarried couple. For such records 'position in the household' is being imputed.

Harmonization of the input information takes a lot of effort. An example where this has to be done is the variable 'educational attainment' in the micro-database for persons. Information about this variable can be found in the Employment Exchange register and in both the Labour Force Survey and the Living Conditions Survey. Definitions of these variables (for example time of observation, classification employed) may differ considerably, though, and (priority) rules have to be derived as to how the corresponding output variable can be obtained.

3.3 Estimation

The edited, imputed and harmonized interrelated micro-databases provide the starting point of the formal estimation stage of the statistical process. Statistical aggregates fit for publication are

derived and will be stored in StatBase. Currently it is not yet clear whether StatBase will physically be a separate database or be integrated with the output datawarehouse StatLine.

Estimation on the basis of a number of related micro-databases is a complex issue. In order to construct a large number of reliable and consistent estimates, a sophisticated estimation technique has to be applied. The estimation procedure should be able to construct a fixed set of estimates from a fixed set of micro-databases. It should also be possible to add additional estimates after the fixed set of estimates has already been published. These new estimates should be consistent with everything that has been published before.

The procedure we have developed at Statistics Netherlands is based on weighting. Consistency between estimates is obtained by calibration techniques, where new estimates are calibrated with respect to earlier estimates. The procedure is formally described in Renssen et al (2001); an application in practice is described in Kroese et al. (2000). There is a large demand for small area statistics; the availability of the large micro-databases fosters such a demand.

This year, some statistical publications will be based on the system of related micro-databases, in particular our publication on the Structure of Earnings Survey. To this end, the micro-databases of persons and jobs will be used. Next year, we plan to meet the Eurostat Census requirement by estimating a set of Census-like tables from the set of micro-databases.

4. Conclusions

It is to be expected that the number of publications based on the system of micro-databases will grow fast. The richness of the matched information makes it possible to confront all kinds of variables and, hence, obtain very flexible publications. Nonresponse correction can be done much more adequately since there is a lot of auxiliary information around that can be used to this end. The micro-databases also are an important framework for cost-savings, as they facilitate greatly the use of administrative registers as a source for statistical estimates.

Last, but not least, estimation from a fixed set of micro-databases makes it easier to get fully consistent estimates in our output-datawarehouse StatLine. Users are more and more interested in flexible analyses in which various statistical results are confronted. Overall data consistency greatly facilitates that job.

REFERENCES

Keller, W., Bethlehem, J., Willeboordse, A., and W. Ypma (1999), "Statistical processing in the next millenium," *Proceedings of the XVIth Annual International Methodology Symposium on Combining Data from Different Sources, May 1999, Canada*.

Kooiman, P., A.H. Kroese, and R.H. Renssen (2000), "Official Statistics: an estimation strategy for the IT-era," *Proceedings of XIVth Compstat meeting, Utrecht August 2000*, pp. 15-26.

Kroese, A.H., and R.H. Renssen (1999), "Weighting and imputation at Statistics Netherlands," *Proceedings of the IASS conference on Small Area Estimation, Riga August 1999*, pp. 109-120.

Kroese, A.H., R.H. Renssen and M. Trijssenaar (2000), "Weighting or imputation: constructing a consistent set of estimates based on data from different sources," *Netherlands Official Statistics*, **15**, pp. 23-31.

Renssen, R.H., A.H. Kroese, and A.J. Willeboordse (2001), "Aligning estimates by repeated weighting," *CBS-report H 491-01-TMO.*