**The use of quality procedures through the example of a statistic based on registers**

(Claude Macchi, Dr. Bertrand Loison)

**Driven by the need to reduce production costs by decreasing the burden on respondents and the increased need of users for information, Switzerland in November 2013 published for the first time ever the preliminary results of the new business census statistic based on registers. This paper demonstrates in three sections how the Federal Statistical Office (FSO) has grown from a traditional business census using paper forms, to the integrated use of various administrative sources to produce this statistic.**

**The first section deals with the concept of this new statistic as part of an integrated system encompassing all economic statistics and meeting the challenges which result. The second section addresses the complex issue of the implementation of quality standards at every stage of the production process, i.e., the harmonisation of the definitions (variables), the continuous actualisation statistical classification of economic activities (NACE), the automation of input controls through the use of structured files (XML), the implementation of matching rules by way of a unique identifier, multi-source data, the processing of missing values and imputations, as well as the use of mathematical models for regionalisation and retropolation of the results. The third section deals with the challenges that confronted the SFO when drafting a press release presenting a new image of the economic structure of Switzerland thanks mainly to a qualitative improvement of the reference field that takes micro-enterprises into account.**

Keywords : register-based statistics, business census, quality, standardisation

Launched in 2008, developed further in 2011-12, and finalised in 2013, modernisation of the federal business census is the second major undertaking aiming to modernise Swiss official statistics following that of the census of the Swiss population completed in 2010. The main objective of this complex and ambitious project was to eliminate the need to send out 500,000 questionnaires every three or four years, by making use of administrative data and thus making it easier for the companies. Driven by the need to reduce production costs and to decrease the burden on respondents and in response to the increased need of users for statistical information, register-based statistics has made its appearance in the sphere of Swiss business statistics.

1. **From a paper-based to a register-based business census**

The FSO is not the first national statistical office to take this step. The Scandinavian countries have been producing register-based business statistics since the 1980s. Since then other Western nations, and not just in Europe, have followed the example. The trend has now found a growing number of adepts, encouraged to adopt a strategy recommended by the international organisations. The integrated approach to register-based statistical production also promotes the harmonisation of data and in this way creates a solid basis for comparing national and international statistical information, leading to a significant strengthening of their utilisation potential. After three years of intense efforts the first results of the Statistic on Enterprise Structure (STATENT) were published in mid-November 2013, bringing to an end not only the work of conceptualisation and integration but also demonstrating the viability of the concepts in concrete terms. The challenge was not to define new questions, or even to prepare questionnaires and organise the logistics, but rather to develop the basis for analysis and integration of the data contained in registers for statistical purposes. These years have been devoted to preparing, transforming, adapting and standardising business data so that it can be used for statistical production. Three major challenges had to be met.

**The challenge of integration**

Like a jigsaw puzzle the modernisation of the federal Business Census (BC) involved the collection and integration of data from the Business and Enterprise Register (BER), the National Register of Buildings and Dwellings (RBD), the Register of Business Identification Numbers (IDE), cantonal agricultural registers, the Old Age and Survivors Insurance (AHV/AVS) funds, the Federal Customs Administration, the Federal Tax Administration and the Profiling, as well as surveys for updating the Business and Enterprise Register (BER), the General Classification of Economic Activities (NOGA), and finally the Employment Statistics(ES), to create a picture that would be qualitatively coherent and utilisable for statistical production.

**The challenge of harmonisation**

In order to integrate such a large amount of multi-source administrative data it was necessary to meet certain framework conditions right from the start: 1) develop a policy for the harmonisation of statistical definitions (e.g. enterprise, full time equivalent) common to all economic statistics; 2) define XML standards for the uniform transmission of administrative data at the national level for the use of data providers (sources); 3) use official classifications (in particular NOGA/NACE) to ensure consistency within the production system; 4) be able to call upon a functional Statistical Information System (SIS) covering the entire value chain, supported by a system of metadata, sole guarantee of the traceability of the data and of the methods in the system of production. It is in this way and only in this way that the integration of data becomes possible from a professional point of view and efficient in relation to the resources committed.

**The challenge of retropolation**

Changing the statistical definitions, data sources and reference universe while making use of estimation models for the calculation of full-time equivalents (FTE) and the allocation of jobs at local level, is not simple to design from a methodological point of view but appear just very challenging to explain in a way understandable to outsiders. Establishing a link with the past to enable end users to see a link between “the new world and the old” soon came to be accepted as an absolute necessity. Retropolating data for the previous years so as to ensure the chronological series and the comparability of data pointed to two possible scenarios: 1) the retropolation is coherent, 2) the retropolation is not coherent, possibly casting doubt either on the results published in the past or on efforts to integrate the data inherent in the new register-based business census. Simply returning to the past is not possible, neither statistically nor ethically, and indeed is impossible in practice, the data being frozen in time. Simply questioning the new figures on the pretext that they are not entirely coherent does nothing to solve the problem. The solution adopted thus consisted in developing a quality control system aimed at making plausible 1) the results by economic branch, 2) the results at the micro (records) and macro (economic coherence) levels, and 3) the quality of the pairings. Although requiring much in terms of resources, this qualitative approach made it possible to eliminate ambiguities where necessary with the advantage of acquiring a profound professional knowledge of the nature of the data.

1. **What about the quality of the STATENT data?**

The changeover from a survey based on a questionnaire sent to all enterprises on the national territory to a business statistic based mainly on data obtained from registers also required an adaptation of the data quality control process.

Whereas previously the information was collected by means of a single questionnaire containing the same questions for all data providers, with the new approach the quantity and diversity of data obtained from different administrative and statistical sources linked together by unique identifiers made it necessary to reorganise the quality control and ensure coherence spread across the different stages of the statistical production process itself.

**The input data**

Quality controls vary depending on the input data group.

The administrative data obtained from external sources (AHV/AVS funds, Federal Customs Administration, Federal Tax Administration), linked together by such identifiers as the personal social security number or unique business identifier are subjected to (1) technical quality tests – the introduction of a standard XML based tool for the exchange of data (sedex[[1]](#footnote-1)) made it possible to define data quality control at different levels of complexity – and (2) tests to determine plausibility, check consistency and compare with previous inputs of the same data. As much as possible the administrative data received are also compared with data that the same external sources send to other users (statistical or not).

The quality of data from other sources (surveys and statistics produced by the FSO, and statistical registers), which complete the administrative data, is controlled with the help of *longitudinal checks* (temporal comparison of data from the same source) as well as *inter-source checks* (crossover comparison of information on the same enterprises through different statistical sources).

Quality control at this first stage of statistical production is essential for all successive stages: the more the coherence between the different data is guaranteed, with input schemata that are clearly defined and documented, and the more thorough the plausibility tests have been, the higher will be the quality of the data, implicitly facilitating the tasks which follow in the framework of statistical production. The questions which arise at this point are (1) the definition of the quality of data required from the administrative data providers and (2) the consideration given by the latter to the feedback from those using the data. It should be noted that management of the administrative data has other objectives than the statistics, and it is for this reason that the needs, proposals and feedback of the users of the statistics generally take second place, being given prior consideration only when there are manifest errors in the data, or in the framework of structural or organisational revision of the administrative data themselves. From the point of view of the statistics offices it is essentially a question of following the work of the producers of administrative data very closely, seeking dialogue with them, and being included in data revision projects so as to be able to integrate their requirements in advance.

**Codification of the enterprises**

The economic activity of the enterprises concerned, a fundamental variable in the framework of STATENT, is defined on the basis of NOGA[[2]](#footnote-2), the Swiss national version of NACE Rev. 2. The enterprises are codified initially – mainly by private partners – at the time of their integration in the Business Register. The code is then validated in the framework of a statistical survey.

Unlike within the scope of the old federal Business Census (BC), where enterprises were codified definitively and exclusively on the basis of questionnaires completed by the enterprises themselves, with the STATENT the codification of enterprises is updated after this initial stage through profiling activities, specific statistical surveys (of hospitals, pension funds, etc.), professional registers and in the framework of sample surveys. The remaining enterprises – for the most part single-establishment enterprises – are updated in five-year cycles by means of direct surveys based on quarterly samples. In parallel with these activities it is planned to maximise the use of information available on the internet and to institutionalise cooperation with commercial suppliers of economic information. There are at present three data vendors that codify more than 60,000 enterprises in the commercial registers on the basis of direct contacts.

These efforts made it possible to reduce the rate of errors in the codification of the enterprises’ economic activities from the 13% recorded in the framework of the BC of 2008 to a rate of 3% (accuracy of +/- 1%) for codifications of the 2-digit level NOGA and an error rate of 6% (accuracy of +/- 2%) for codifications of the 5-digit level NOGA in 2011. The target for 2015 is to reduce codification errors to 1% for the 2-digit level NOGA and 3% for the 5-digit level NOGA.

**Assemblage and consolidation of data**

This stage involves a great many steps during which the plausibility tested and corrected input data (see above) are transformed into statistical results ready for integration into the statistical production process or for dissemination.

The integration of individual data and their aggregation involves a regular succession of cyclical processes, ending in each case with a new set of STATENT data based on corrections and updates from the preceding stage. Each new set of the data is subjected to a series of quality checks, to a great extent automated, at both the micro and macro levels, ranging from checking the consistency of nomenclatures, and individual employment values for each enterprise and economic activity, to a more global approach analysing the total number of jobs including full time jobs (calculated on the basis of a specific mathematical model) at an aggregated geographical level. Fundamental, as in the stage of data input quality control, are the *longitudinal checks* (comparison of different sets of data with the versions of previous years) as well as *inter-source controls* (of consistency with other enterprise surveys, comparison with economic statistics, etc.).

This work, which takes about two and a half months, makes it possible to adapt and refine results, especially in certain economic areas. Although on the one hand the data is generally subject to only marginal changes (e.g. in the framework of the STATENT 2011 it was possible to correct the number of enterprises by 1.34% and of jobs by 0.55%), at more disaggregated, communal and regional levels and at various levels of economic activity, the difference in results can at times be significant. Tests at the stage of assemblage and consolidation, notably in the framework of the first versions of STATENT, made it possible to discover areas such as the public sector in general as well as some of the NOGA positions covering various activities that are not clearly specified (residual categories) where there were major deficiencies (erroneous attribution of economic activity codes and problems of consistency at the level of data input) requiring considerable corrective efforts.

**Integration of the regional statistical offices into the data quality process**

The regional administrations are now fundamental partners in the framework of the production procedures and are now actively integrated in the stages of validation of STATENT data. Once the data are in an almost finalised state they are forwarded to the statistical services of the 26 cantons and 7 major cities in the form of (1) data aggregated by commune, economic activity and jobs, and (2) individual data with information on the names, economic activity and jobs for each enterprise. On the basis of their own knowledge and experience, and within the limits of their possibilities, the partners verify the coherence and plausibility of the results provided, transmit the missing enterprises (for the most part local units) and any corrections regarding changes in names and addresses. Integration of the feedback from the regional statistical services initiates the final stage of determining the plausibility of the STATENT data, making the final adjustments and imputations as well as geocoding, thus concluding the quality control process stage and freeing the data for dissemination.

1. **Facing challenges in the Press Release**

Table 1 presents the final statistical results and specifically the differences at the national level between the traditional BC of 2008 and the STATENT of 2011. While the changeover from the traditional BC to STATENT made a more comprehensive understanding of the Swiss economic fabric possible, with the addition of units and jobs that had been excluded from the BC's statistical scope, it has been far from easy to communicate the statistical results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **RE 2008** | **STATENT 2011** | **Deviation [thousands]** | **Deviation [%]** |
| **Number of enterprises** | 384,000 | 564,000 | 180,000 | 47% |
| **Number of establishments** | 452,000 | 639,000 | 187,000 | 41% |
| **Number of jobs** | 4,196,000 | 4,847,000 | 651,000 | 16% |
| **FTE** | 3,513,000 | 3,897,000 | 382,000 | 11% |

Table: Final Results

It has not been easy to communicate deviations, in some cases considerable, to the media notably in relation to the statistical input of very small observation units (micro enterprises and jobs with a low level of activity). In order to ensure the correct transmission of results in the public domain it was necessary to meet two challenges.

**The challenge of complexity**

Modernising without being critical of the past above all requires the creation of a common knowledge base, making it possible to adapt the message to the scientific knowledge of the users of official statistics, be they stakeholders at the national, cantonal or communal level. In concrete terms the changes in definitions and in the reference universe and the use of imputation models for calculation of the FTEs and for the allocation of employment at the local level made it necessary to provide information and training at the same time so as to ensure correct utilisation and interpretation of the results without undermining the credibility of official statistics. Simplifying the changes so as to make the message acceptable and understandable turned out to be more a communication than a statistical challenge.

Indeed, while the traditional BC – based on counting – had the advantage of making a 1 :1 link possible between the enterprise and its characteristics (NOGA code, jobs, localisation) it had the disadvantage of eliminating certain enterprises (micro enterprises, …) from the field of observation thus presenting an incomplete picture at the national level. While a register-based census – using imputation models and therefore more abstract for explanation – leads to results that are less certain in terms of employment at the local level, it does have the advantage of ensuring much greater completeness at the national level. How to explain this dilemma and eliminate all fears without undermining credibility is above all a communication task rather than a statistical one.

**The challenge of communication**

The fact that the publication of the first provisional results for the 2011 reference year went well is essentially due to the considerable use made of traditional communication tools: 1) adapting the message to the groups of users (researchers need to understand the mathematical model, news agencies want to know if previously published results were false,…); 2) explaining the change of concept just as much as the new figures (local politicians want to know why the definitions have changed and what the advantages are in concrete terms, while researchers will be more interested in questioning the relevance of the definitions adopted and what impact they have had on the new results); 3) educate simply (taking the knowledge level of the target audience into account) and inform openly targeted regional partners, journalists and administrative services; 4) ensure complete transparency as to the method so as not to undermine the credibility of public statistics.

In conclusion, modernising is not itself a problem. The real challenge is ensuring comparability with the past. Furthermore, the ever greater use of administrative data for the production of public statistics also calls for ever greater skill in the processing of large volumes of data, placing the data warehouse concept at the hub of the statistical production system as the only guarantee of the continuing consistency of the data used by statisticians.

1. <http://www.bfs.admin.ch/bfs/portal/fr/index/news/00/00/02.html> [↑](#footnote-ref-1)
2. <http://www.bfs.admin.ch/bfs/portal/en/index/infothek/nomenklaturen/blank/blank/noga0/vue_d_ensemble.html> [↑](#footnote-ref-2)