Business Trends
Antti Santaharju
antti.santaharju@stat.fi

ESSnet Admin Data Seminar
18-19 March 2010

Use of monthly tax return data - transforming raw data to statistical data

Tax return data are the main data source for the turnover and wages and salaries indices which Statistics Finland publishes. The tax return data are administrative data which the Tax Administration updates every month for Statistics Finland. The use of administrative data in statistics production has indisputable advantages over sample data. The data cover almost the entire population, their collection costs and the response burden on enterprises are lower than in a sample survey. The exploitation of administrative data also carries risks which could be avoided in own data collecting. By constructing its statistics production on administrative data Statistics Finland loses the possibility to decide about the contents of the key data, and about related timetables and revisions. In the worst case the structure and timetables of administrative data could change completely without Statistics Finland being able to influence the essence of the made changes.

What is described above happened in Finland at the beginning of 2010. In January the Tax Administration implemented a tax reform in consequence of which the structure, contents, timing and coverage of the tax return data changed. The reform caused extensive needs to alter the methods and information systems built for the processing of the data.

This paper describes how Statistics Finland transforms the tax return data into statistical data with which the actual index calculations can be performed. The paper also describes how Statistics Finland controlled the changes that took place in the source data proper for the indices on turnover and wages and salaries.

Tax return data

Monthly tax return data arrive at Statistics Finland once a month from the National Board of Taxes. The data contain information about almost all commercial selling of goods and services, imports, and employers paying wages and salaries in Finland. Only small scale activity (annual turnover < EUR 8,500), public sector turnover and non-regular wage and salary payers are excluded from the tax return data. The tax return data contain information about 320,000 enterprises per reference month. The monthly tax return data have been the main source for the Finnish monthly indices on turnover and wages and salaries since 1999.

Changes in the tax return data

A tax reform implemented by the Tax Administration took effect in Finland as of the beginning of 2010. The tax reform was based on an effected reform of tax legislation which caused changes to the contents, transferring schedules and coverage of the tax return data. The most significant changes the tax reform caused to the tax return data are described in subsequent chapters.

Business Trends
Antti Santaharju
antti.santaharju@stat.fi

ESSnet Admin Data Seminar
18-19 March 2010

Structure of the data the Tax Administration delivers to Statistics Finland changed

In connection with the tax reform, the Tax Administration renewed the information systems it uses for collecting, storing, processing and transmitting the data. The structure of the raw data transmitted to Statistics Finland altered completely. Statistics Finland had to renew the methods built for the reception and processing of the data to meet the demands imposed by the new data format.

Some of the data transmitted by the Tax Administration were lost

Some of the information transmitted by the Tax Administration to Statistics Finland was lost in connection with the tax reform. Fortunately for Statistics Finland mainly auxiliary variables used in the checking of the data were lost in the reform. For instance, in the past Statistics Finland used to receive a notification of the VAT payments actually made by enterprises to the Tax Administration in addition to the VAT data the enterprises themselves had reported. New data validation methods had to be developed to replace the methods which depended on the data checking variables that were lost.

Data contents of some variables in the tax return data changed

The tax reform altered the principles according to which enterprises report their VAT and employer payments data. The numerical data reported according to the new principles are not directly comparable with the values reported prior to the tax reform. After the tax reform the tax return data had to be made comparable with the data reported prior to the reform. Otherwise the reform would have caused temporal incomparability in the time series.

Increasing amount of information on e.g. foreign trade is available from the Tax Administration

Due to the reform, more exhaustive information than before is received from the Tax Administration about certain phenomena. For instance, more exact data than before are now available about exports and imports of services. The new data have so far not been exploited in the production of statistics but they will most certainly be useful in the future in e.g. compilation of statistics on foreign trade. Before the new, more accurate data on the variables can be exploited they must be thoroughly analysed.

Some enterprises changed from monthly reporting to quarterly or annual reporting

Before the tax reform the Tax Administration transmitted all data to Statistics Finland on the basis of monthly reporting. After the tax reform small enterprises with an annual turnover of under EUR 50,000 are allowed to report their data quarterly in order to reduce bureaucracy. Annual turnover of under EUR 25,000 gives an enterprise the right to report its VAT data annually. Enterprises may decide themselves whether or no they want to opt for an extended reporting period.

According to the latest information received from the Tax Administration, 93 per cent of enterprises will report their VAT data monthly even in the future. According to estimates, these enterprises generate 99.8 per cent of

Business Trends
Antti Santaharju
antti.santaharju@stat.fi

turnover in the whole economy. Some 92 per cent of enterprises intend to report their employer payments data monthly even in the future. It has been estimated that as much as 99.9 per cent of data on employer payments will be received as monthly reported even in the future.

The calculation of the monthly indices requires disaggregation of the data reported for varying lengths of reference period to the monthly level. For the time being the impact of the enterprises reporting on the lengthened reference periods is not significant on the index but it is important that the disaggregation methods are developed already at this stage because in the future the Tax Administration intends to raise considerably the threshold values of turnover entitling to these extended reporting periods.

Changes not yet detected

Previously implemented revisions have proven that all changes that take place in data contents cannot be detected immediately. For example, enterprises may report their data in a different way than they did prior to the revision if they interpret the inquiries implemented with the Tax Administration's new tools in a different way than previously. Efforts were made to build our information systems which process the tax return data so that we would be able to detect unexpected changes in the contents of the data promptly.

The treatment of the above-described changes in statistics production requires a consistent statistical production process. The management of the changes must be targeted at the correct stage of the statistical production process so that the changes caused by revisions can be made efficiently and the statistical production process does not become excessively complicated. The next chapters describe how the changes in the source data were treated as part of the statistical production process of the indices of turnover and wages and salaries.

Treatment of tax return data as part of the statistical production process

When the tax return data arrive at Statistics Finland they are not directly suitable for statistics production because they are administrative data formed for the Tax Administration's own purposes. The original raw data must be transformed into statistical source data that can be used in the calculation of the indices of turnover and wages and salaries.

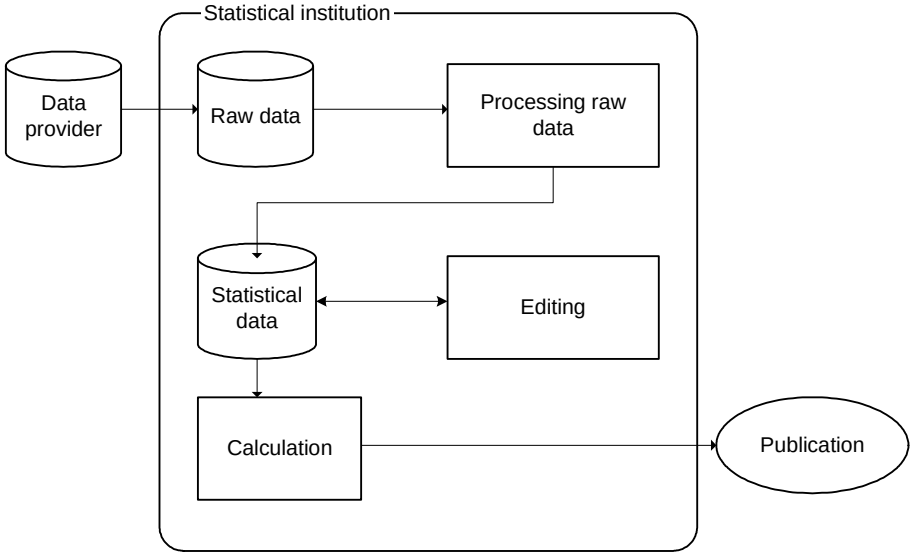
At Statistics Finland the statistical production process is divided into four sub-processes as shown in the simplified diagram of Figure 1 below: processing of raw data, editing of actual statistical data, index calculation, and publication. Each sub-process is an independent work stage which draws from the database the source data it needs, processes them as required and saves the processed data back into the database for the subsequent stages of the process.

From the point of management of the entirety, a statistical production process which is divided into separate, independent sub-processes is an efficient and straightforward solution. Changes caused by revisions in the

source data to the statistical production processes and systems can be located into a certain sub-process and the effects of the changes can be isolated from the other process stages. The effects of the changes on the entirety are taken into account in a relevant sub-process whereby the methods or information systems of the other process stages need not be altered.

The treatment of the changes that had taken place in the tax return data was located into the processing of raw data stage whereby the harm caused by the changes in the source data to the entire statistical production process became minimised. The changed source data are edited at the stage processing of raw data into statistical data comparable with the source data of the old pre-reform format. The tax reform did not require changes to be made into the other key sub-processes, such as the editing of unit level statistical data or the index calculation methods.

Figure 1. A simplified model of administrative data processing



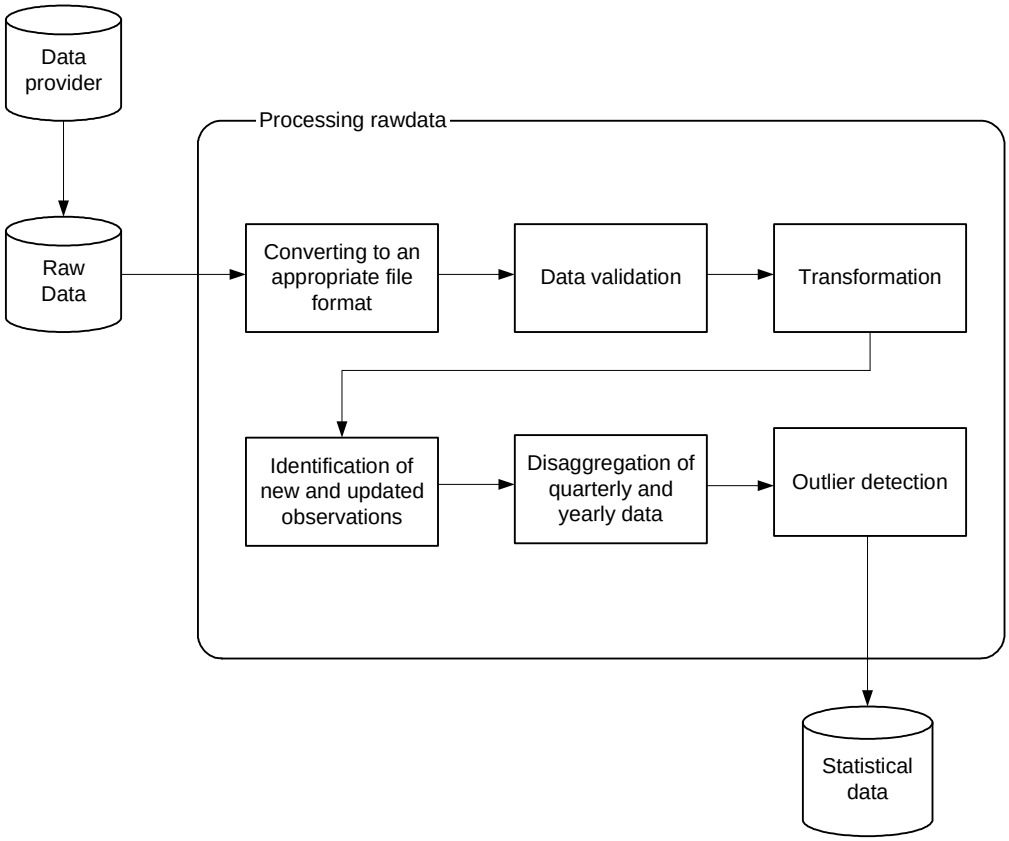
Transformation of tax return data into statistical data

The stage of processing of raw data shown in Figure 1, in which the administrative data are transformed into actual statistical data, is shown in more detail in Figure 2. The transformation of the administrative data into statistical source data can be divided into six sub-processes as shown in the following Figure.

Figure 2. A model of processing of raw data

Business Trends
Antti Santaharju
antti.santaharju@stat.fi

ESSnet Admin Data Seminar
18-19 March 2010



Converting to an appropriate file format

Statistics Finland receives the tax return data monthly from the Tax Administration in accordance to an agreed timetable. The data are supplied overnight to Statistics Finland's data transmission server which starts up automatically a Unix-script code which then starts up the SAS programs for the processing of the data. The raw data stage of the entire process is performed automatically overnight. The statistical source data with their processing details are in the database in the morning following the data transmission. The raw data supplied by the Tax Administration are converted into SAS data format because all processing of the data is performed with SAS software.

Data validation

Once the raw data have been translated into SAS data format, their validation starts. At first a check is made to ensure that all variables are in the format of the database definition supplied by the Tax Administration, the variables may only receive values according to their value range and the logical relationships between the variables are in order. For example, numerical variables may only receive numerical values whose maximum length does not exceed the number of digits specified in advance. If the variables in the data are dependent on each other a check is made to ensure that there are no logicality errors between them. Classification variables may only receive values specified in advance. Erroneous observations are printed

out by error type for the checking of error data. A summary report is printed out of possible errors.

Erroneous observations are classified into two categories, critical errors and non-critical errors. Critical errors are printed out to error data and omitted from the updating process. Non-critical errors are printed out to error data but they remain in the updating process.

The data validation process proved really useful in the testing of the revision. Thanks to the error listing, hidden problems and illogicalities in the data were caught at the testing stage of the system and critical errors were no longer present in data transmission after the introduction of the information system.

Once the "technical" quality of the data has been checked with the methods described above, summary statistics are calculated from the data. Summary values, averages, medians, lower and upper quartiles and minimum and maximum values are calculated for each variable in the entire dataset. The summary statistics are used to examine whether any major changes have taken place in the contents the data.

Transformation

When the data received from the Tax Administration have been checked in their original form, they are edited as necessary. The data supplied by the Tax Administration are transformed into the format required in the index calculation. The calculating of the indices of turnover and wages and salaries requires the data to be summed up at quite a detailed level into aggregate components that make the index calculation easier. The retention of the comparability of the data prior to and after the tax reform is also taken care of at the transformation stage. The raw data are transformed into a format that corresponds with format prior to the tax reform.

Identification of new and updated observations

The Tax Administration submits to Statistics Finland monthly all possible data for the latest six months. Most of the data concerning the months prior latest month have already been updated in preceding months into the database of the statistical source data. The latest validated tax return data are compared with the data received earlier and unchanged observations are dropped out in the data editing process. The data supplied by the Tax Administration contain monthly information concerning all enterprises in Finland, so the throughput times of the programs handling the data mass are long. The identification of new and changed observations as early as possible within the process can shorten this throughput time considerably.

Disaggregation of quarterly and yearly data

The calculation of the monthly indices of turnover and wages and salaries requires the disaggregation of quarterly and annually reported data to the monthly level so that the absence of small enterprises from the calculation will not cause bias in the results.

Business Trends
Antti Santaharju
antti.santaharju@stat.fi

The data of the quarterly and yearly reporters are computationally converted to the monthly level by using the so-called common ratio method. The method was studied prior to the tax reform and found to be fairly reliable. The method can briefly be described as follows.

Let $x_{i,M}(m)$ and $x_{j,Q}(q)$ be the turnover or wages and salaries data of the monthly and quarterly reporters for month m or, respectively, quarter q. The monthly estimate for the turnover of a quarterly reporter is

$$\hat{x}_{j,Q}^{M}(m) = c_{j,Q}(m)x_{j,Q}(q),$$

where $c_{j,Q}(m)$ is the coefficient with which portions of the quarterly data are assigned to each month. The monthly coefficients of a quarter sum up together. The coefficients are calculated with formula

$$c_{j,Q}(m) = \frac{\sum_{i \in A_j} x_{i,M}(m)}{\sum_{p=0}^{2} \sum_{i \in A_j} x_{i,M}(m-p)}.$$

where A_j is the population of the enterprises (or donors) used in the calculation of the coefficient. In other words, the coefficients are calculated by selecting a population of representative enterprises, summing up their monthly data and by forming quarterly division ratios from them.

Statistics Finland tested various ways of selecting the donors A_j. Of the tested alternatives the following combinations worked best:

- **Quarterly reporters, turnover**: annual turnover not more than EUR 250,000 and enterprise belongs to the same 2-digit industrial classification level as the enterprise for which the estimate is being formed
- **Quarterly reporters, wages and salaries**: annual turnover not more than EUR 1,000,000 and same 2-digit industrial classification level as the enterprise for which estimate is being formed
- **Yearly reporters, turnover**: annual turnover not more than EUR 75,000 and same 3-digit industrial classification level as the enterprise for which estimate is being formed.

Outlier detection

Enterprise-specific erroneous and deviating observations should be detected and corrected as early on in the statistical production process as possible. So far outlier detection methods for raw data have not been implemented into our information system because they required historical data on the examined variable. We intend to concentrate on this detection of enterprise-specific errors in the near future. At the moment the data are examined in connection with the editing of the statistical data but major errors in the source data should be corrected even prior to the forming of the actual statistical data.

Batch report

The SAS program which transforms the tax return data into statistical source data produces a batch report which shows errors observed in the validation,

Business Trends
Antti Santaharju
antti.santaharju@stat.fi

numbers of new data items and data needing editing, as well as other information about the progress of the SAS batch run. The performer of the SAS batch run can check the batch report after the batch run and if no errors are observed the data can be assumed to be in order with high degree of certainty. If the data show errors, error codes are printed to the batch report from which the identification data and type of error of the erroneous observations can be ascertained.

Conclusions

The method for compiling the statistics on turnover and wages and salaries indices is comprised of separate stages or sub-processes which can be performed independently of each other. The separate sub-processes facilitated the locating of the changes that had taken place in the administrative source data into certain process stages so that the changes in the source data do not require the renewal of the entire statistical production process.

In the context of the tax reform, new methods and information systems were developed for the transformation of the raw data into statistical source data. For the checking of the data a new validation stage was developed where the source data are checked with the help of metadata - the database definition and variable interdependencies. The validation phase proved useful because with its help hidden errors in the data could be corrected in co-operation with the Tax Administration already at the testing phase, before the reception of the first actual tax return data. A notation of the errors observed in the validation is printed out on the batch report. If the program has not observed any errors, a quick glance at the batch report is enough to verify that the data are in order and unit level data need not be examined in more detail.

For the time being, all stages of the process have not yet been finalised within the renewal project. Outlier detection, which requires historical data, still has to be implemented. Its development must be addressed in the coming months. A well-developed outlier detection method can pick up changes in the contents of the data in good time before the actual index calculation.

The renewal intensified co-operation between Statistics Finland and the Tax Administration. Prior to it Statistics Finland was mainly in contact with the technical support staff of the Tax Administration who were responsible for the technology of the data transmission to Statistics Finland. The renewal enabled the establishment of improved connections between actual content experts in consequence of which we have obtained a lot of new information about the contents of the tax return data.