

Preparing for Changes in Administrative Data for Short Term Statistics

The adoption of short term administrative data in Finnish economic statistics has become relatively broad. The shift from a purely sample based approach has been a success: the cost of collecting data has decreased and the coverage has dramatically increased, making it possible to simultaneously broaden the scope of short term statistics and decrease the burden on enterprises.

The change of approach has had an effect on work processes as well - they have become much more data driven. This paper describes experiences with one of the short term administrative data sets utilised by Statistics Finland - the VAT data. VAT declarations have been the main data source of the well over one hundred turnover indices published monthly since 1999. The importance of VAT data has even grown since the adoption of double deflation in National Accounts, as the turnover indices are now a primary source for quarterly GDP measures.

As the data providers' IT infrastructures continue to develop, the opportunities to utilise administrative data for statistics should become even more frequent and more diverse. While this is probably true in the long run, in some cases an improvement in the way the data are collected for primary use can be an impairment from the perspective of the statistician.

A commitment to the adoption of administrative or other externally collected data implies trust in the availability of the data and in the stability of the way the variables are measured in the data. The significance of the decision is increased by the associated methodological and software changes, which can be a major investment to the statistical institution and might be reversible only with a sizeable effort. Because there probably *will* be problems with the data, a statistical institution using or planning to use administrative data, especially for short term statistics, should be prepared to tackle them.

Features of administrative data

Using administrative instead of directly collected data gains coverage but loses detailed control of the schedule and the content of the data. In our experience the benefits of the VAT data by far outweigh the disadvantages but we have also noted that, due to the peculiarities of the data, the complexity of the necessary methodology has increased.

From a statistician's point of view the primary collector of administrative data is an intermediary between the respondent and the statistical institution. Thus, in terms of speed of accumulation, using administrative data is bound to be inferior to directly collected data. The definitions of variables and reporting units are probably also not optimal in administrative data - yet close enough in most cases. Here are examples of how some of these issues appear in the Finnish VAT data:

Accumulation

Finnish legislation makes it possible for an enterprise to postpone a VAT declaration for up to six months. Accordingly, the first figures published by Statistics Finland are partially based on estimation. Moreover, because even the most recent VAT data are approximately 60 days old, the first estimates are entirely based on directly collected data.

The slow accumulation of the VAT data does not mean that they are not useful for first estimates. Historical data make it possible to optimise the sample size of a direct data collection much more efficiently than would otherwise be possible. However, a proper statistical theory does not exist for this case, and the ad hoc solution we have devised has some problems.

Definition of unit of observation

While, for example, in the Finnish index for industrial output the chosen observation unit can be an establishment, an enterprise, or in some cases even a whole industry aggregate if needed, the observation unit in the VAT data is always a legal unit. This makes the VAT based statistics more sensitive to changes in the way large corporations manage their structures. The structural changes have to be dealt with at a fast pace using special methods with very limited information. While statistics which are solely based on direct data do have to deal with similar issues, the methodology needed in the VAT based statistics has, in practice, become more complex.

Definition of variables

What is considered as turnover in the VAT data is somewhat indeterminate. Another example related to the definition of unit of observation is change in business structure, which can often influence the way turnover is reported. For example, a split of a consolidated corporation's subsidiary into two new subsidiaries might cause a level shift in the reported turnover because the new subsidiaries might sell goods to each other.

The fact that the VAT data are collected for taxation purposes has an effect as well. Whether a firm's monthly sales consist of only sales of goods and services or additionally of sales of equipment used in the production of goods and services is of little concern to the tax authority. Another example is private healthcare, which is VAT exempt. Although zero VAT sales should be reported to the Tax Administration, it has little incentive to control the quality of the data.

Changes in administrative data

Like many other national statistical institutions, Statistics Finland is obliged by law to use administrative data. The agencies holding the data are also bound by law to provide them. No law, however, dictates that a government agency should keep the data content and the schedule of transmission fixed.

Changes in administrative data seem to fall under two broad categories:

- changes in legislation or in the interpretation of legislation or

- changes in the way the original data are collected, stored or transmitted.

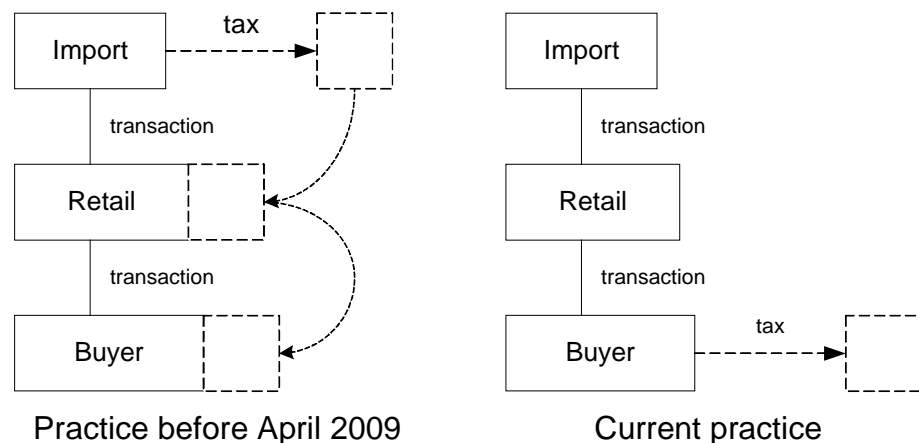
The changes we have faced have not been very big, but in some cases they have come unannounced and we have had little time to deal with them.

Some examples are given below:

Motor vehicle tax

In April 2009, Finnish motor vehicle taxation changed so that a new car's retail value no longer contains car tax. Previously the tax was included in the retail price because it was paid by the wholesaler before the car was registered; after the change it is the buyer's responsibility to pay the tax.

Motor taxation in Finland before and after April 2009



From a statistician's perspective, this caused a comparability problem. The change contributed to a 0-15% decline in the value of motor vehicle sales, depending on the branch of business and the stock of cars for which the motor vehicle tax had already been paid before April. Uncorrected, the data would have overestimated the already rapid drop in motor vehicle sales. We ended up manipulating the micro data with coefficients based on an expert opinion.

In fact the problem was not specific to the VAT data - the same issue was present in the directly collected data. However, for the direct data collection, we had the option to contact the respondents and ask for comparable figures.

VAT rates

During the nearly ten years the VAT data have been used, the VAT rates have changed several times. During 2009-2010 there will be two changes:

VAT rates in Finland

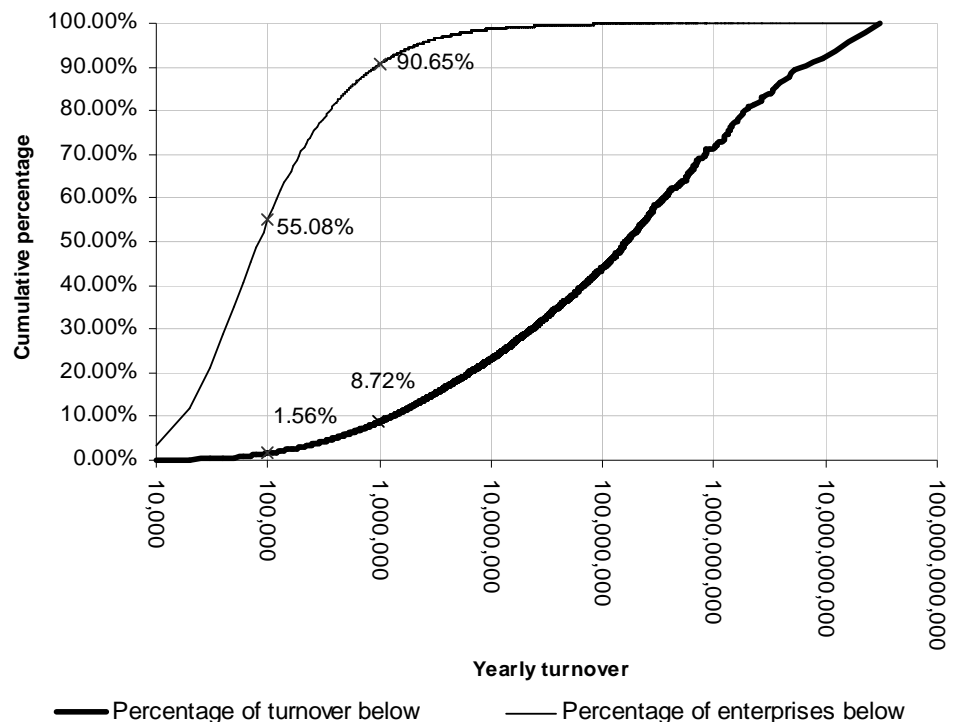
	current	1 Oct. 2009	1 July 2010
Standard rate	22%	22%	23%
Food and animal feed	17%	12%	13%
Restaurant food	22%	22%	13%
Reduced rate for books, pharmaceuticals, etc.	8%	8%	9%

Besides the high probability that the new tax rates temporarily increase reporting errors due to possible misunderstandings and mistakes from the tax payers' part, the change will be at least a minor technical annoyance. We do not receive turnover data, only the amounts of tax paid by tax rate, so each time a new rate is introduced, a new field has to be added to the database and all the programs reading the data updated.

Periodicity

A 'tax account procedure' will be gradually introduced into the Finnish taxation system. This means that eventually all taxpayers, including corporations and natural persons, will manage their taxes on an account in an online service resembling a web bank. Taxes will be paid on the account as a lump sum, meaning that the payer does not indicate whether the money is e.g. for VAT or social security payment. Thereafter the Tax Administration determines computationally which taxes can be considered as paid based on the VAT or other tax reports. As of 1 January 2010, the procedure will be adopted for the so called unprompted taxes, including the VAT.

Cumulative distribution of Finnish enterprises by 2007 turnover



Because payments no longer have to be linked to a particular tax, the procedure will reduce manual work at the Tax Administration. In order to further lessen bureaucracy, small businesses with a yearly turnover between EUR 25,000 and 50,000 will be allowed to pay and report on a quarterly instead of monthly basis. A yearly turnover of under EUR 25,000 permits VAT to be paid once a year.

Since there are lots of small businesses, the number of monthly VAT payments will decrease dramatically. The total monthly tax inflow, however, will be reduced by less than one per cent.

Because most of the turnover data will still be received monthly, the change has little immediate effect on statistics. In fact, although the thresholds for quarterly and yearly reporting are likely to rise, even a threshold of one million euros for monthly reporting would still be satisfactory as can be noted from the graph on the previous page.

Nevertheless, a mix of data with different reporting periods has to be dealt with. The quarterly and yearly data have to be disaggregated into monthly data, which will complicate the software for calculating short term statistics. Valuable information for quality checks is also lost because of the lump sum nature of payments, as we will no longer be able to compare the imposed VAT to the actually paid VAT.

Tax Administration's data warehousing

As the new IT system for the tax account procedure is deployed, the data used for analysis is separated from the real time taxation data. The current plan is that the analysis data will be updated once a month. Because currently the VAT data for Statistics Finland are loaded direct from the real time database, this will either mean postponing the transmission date or, which is more likely, a decrease of accumulation as the analysis data have to be updated earlier in order to keep up with the current transmission timetable. Fine tuning of the transmission dates will also become much more difficult.

Threshold for VAT liability

A Finnish business becomes VAT liable if its yearly turnover exceeds EUR 8,500. There are no plans to raise the threshold, but should such a change occur - for example in order to stimulate start-ups - the effects in the data might be hard to deal with. Some small businesses would just disappear from the data and we would have to make an estimate on how much of the missing data is caused by the threshold change and how much by closures.

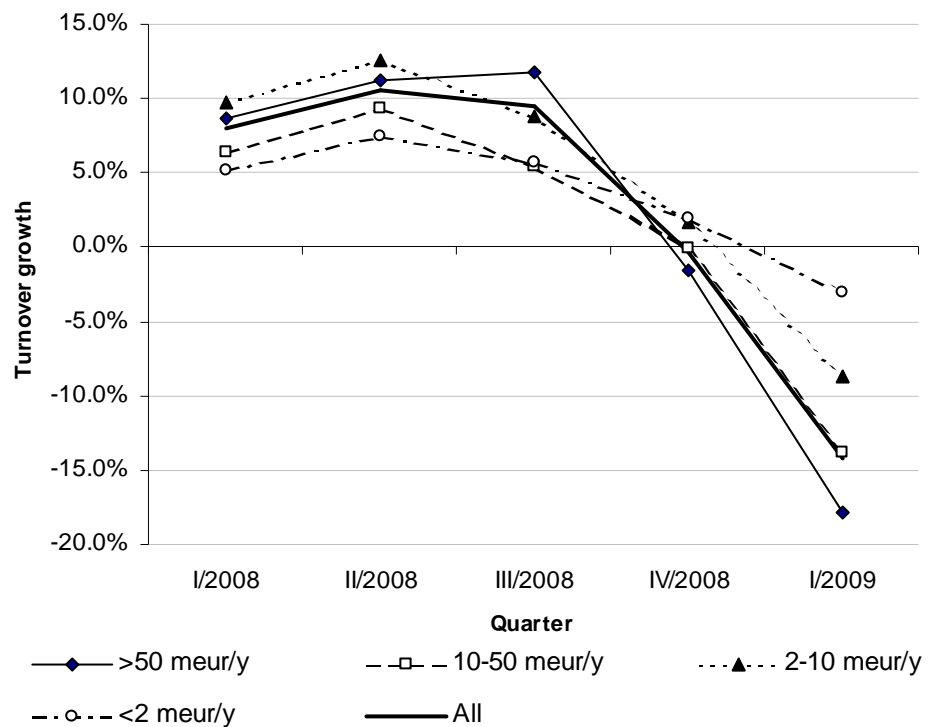
Exogenous shocks

The downturn in the second half of 2008 did not cause changes in the VAT data per se. Rather, the sudden collapse in international trade affected Finnish enterprises in such a way that the assumptions on which the direct data collection was based did not hold very well.

As illustrated on the next page, the year-on-year change in the turnover of the largest companies turned negative in the last quarter of 2008, while the turnover in medium sized companies was stagnant and that of the smallest companies was still growing. This contributed to a slight negative bias in the first monthly estimates during the last quarter of 2008 and first quarter of 2009.

The monthly direct turnover data collection is a cut-off containing the approximately two thousand largest enterprises in Finland. Such a small sample - one-tenth of the size necessary in the sample-only case - has been sustainable because the VAT data contain full historical data and the change rate for the entire population has been close to the change rate of the largest enterprises. Therefore, using historical data and a reasonably reliable change estimate has yielded good level estimates.

Quarterly year-on-year turnover growth in Finland by company size



In hindsight, a stratified sample would have prevented the bias in the first estimates. However, even in this case, rather than increasing the amount of direct data collection, we have started to look more closely at imputation methods and other ways of estimating the difference in the growth rates of the largest and smaller enterprises - yet again complicating our software for the compilation of the statistics.

How to prepare for changes in administrative data

The major challenge in using administrative data is that as a user you are at the mercy of external powers - the data provider or legislation - and that you cannot rely on the constancy of the data. Because the problem cannot be solved at its roots, the way of compiling statistics from administrative data has to be robust against inevitable changes in the data.

It is much easier to list individual problems than to list solutions that cover all possible problems. As it turns out, the list of guidelines is much shorter than the description of troubles. However, as noted before, we do feel that using administrative data is worth the effort.

Business Trends
Ville Koskinen
ville.koskinen@stat.fi

OECD STESEG Meeting
10-11 September 2009

Direct data collection

In our experience, administrative data work well with direct data, but would be a bad replacement for them. Direct data collection can also be seen as a life vest in case the administrative data should become unavailable. Therefore, the choice of basing short term statistics entirely on administrative data should be made with care and with knowledge of the risks involved.

Relations with data provider

To get information on changes in the data, collaboration with the data provider is essential. At the strategic level, the work of the government authorities keeping and using register data should be co-ordinated. In Finland an advisory board of government registers has been set up for this purpose. At the level of a particular data source, it is not enough to manage the administrative side of the co-operation. The contact person working with the data transfer and invoicing might not be the correct person to ask about changes in legislation or other driving factors or the substance of the data.

Because administrative data usually have analytic uses inside the maintaining organisation, it might be useful to try to find synergy there. With business data, another option is co-operation with interest groups, i.e. the representatives of the businesses filling in tax reports.

Continuous analysis of the received data

A change in the administrative data provider's software might introduce errors to the data. Alternatively, a legislative change affecting only a small number of enterprises might cause an unnoticed change in the data making them partly disparate with historical values. It is good to do a basic analysis on the raw data each time they are received. The things to follow are accumulation and basic statistics on the measured variables by industry and the way the data provider flags values.

Importance of definitions

Because the measures and measurement units in administrative data might not be what is actually needed, there is a danger that the definitions which you aim for become forgotten or fuzzy. This can lead to problems when a definition in administrative data changes - if you do not keep it clear what your targeted measures and units are, you might end up just following the data and reducing the relevance of the statistics.

Data cleansing and data processing

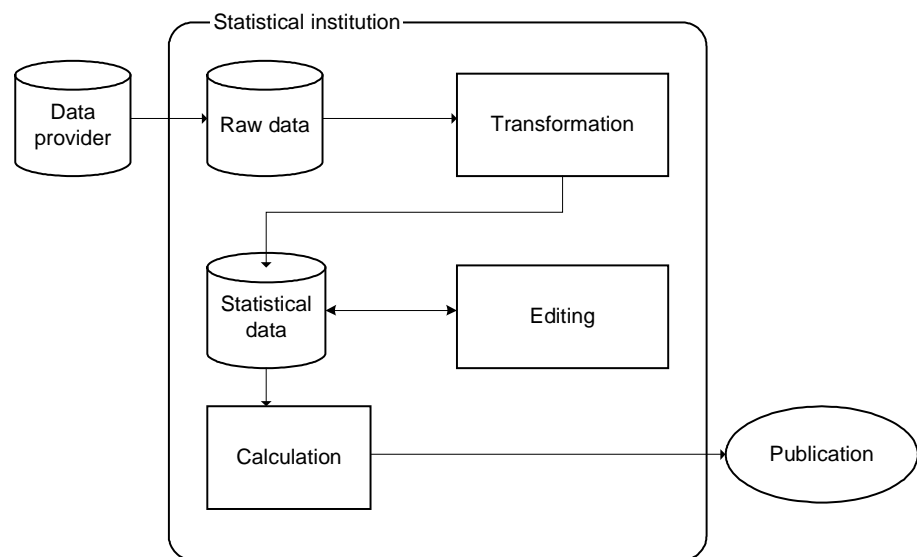
Over one-half of the source code in the software used for compiling the Finnish turnover indices is used for the cleaning up of data. This has been a change of focus from the traditional way short term statistics are made, making parts of the work process resemble those in the business register or structural business statistics, which have been utilising larger data sets and administrative data much longer.

Perhaps an unexpected side effect of this has been that because the software used for short term statistics may have to be updated at a short notice, it is no longer sufficient to have separate people focusing on economics, statistics and programming, but people with a combination of all these skills are now required.

Data management

Since administrative data have to be cleansed, the distinction between raw and statistical data should be made in data storage as well as conceptually. Put differently, calculations for statistics should be based on a stable data model and the underlying administrative data should be transformed to fit that. The received data should be stored with little modifications - perhaps by just converting to an appropriate file format or reading the data into a database - and the necessary metadata indicating the contents of contents should be clearly linked to each historical variation of the data.

A simplified model of administrative data processing



This model has the benefit of isolating issues caused by changes in raw data. When the raw data change, only the transformation into statistical data has to be updated. The same reasoning also means that the compilation software should be designed as modular, and since the software is going to change anyway, it is not a bad idea to use version control for the source code.

Conclusion

The goal of this paper has been to demonstrate that it is useful to prepare oneself for changes when using administrative data for short term statistics. This does not mean that anyone should be discouraged - the virtues of administrative data are indisputable - and since it is more probable that the trend is towards an increase rather than a decrease in the use of not only administrative data but also other kinds of harvested data in statistics, restraining oneself from at least looking into administrative data might be short-sighted.