

Seminar on Registers in Statistics - methodology and quality 21 - 23 May, 2007 Helsinki

The VAT data in short term business statistics

Ville Koskinen
Statistics Finland
Ville.Koskinen@Stat.Fi

1. Introduction

Statistics Finland has used monthly VAT and social security payments data for compilation of short term business statistics since 1999. Using the data for statistical purposes was justified by the 1998 EU regulation for short term business statistics (STS).

The data set is big, a mammoth of 200 000 statistical units, currently almost 150 points of observation and 40 variables. It's used for compiling some 2 500 different indicators and indirectly in quarterly GDP calculations. Other uses are estimation of number of full-time equivalent employees on a monthly basis, and, potentially, short-term estimation of profitability.

An administrative source for STS has proved to be useful. On the other side, handling the data requires particular skill and somewhat different approach compared to the normal survey based case. In most uses the VAT data cannot be the *only* source we use and it must be supplemented by some proper statistical data.

Use of registers and other administrative sources is not so common in business as in social statistics. There is an endeavour e.g. by the OECD to establish a general framework on the subject, but progress has been less than swift. We don't yet know what the best practises are. We don't, for example, have a sampling theory which would really work in our particular case.

This short paper is supposed to shed some light on some of the points we feel important. It describes generally the different aspects of the usage of the VAT-data in STS.

2. Advantages

The VAT data contains all the relevant statistical units for the compilation of turnover and wage sum indicators. This makes, at least on a superficial level, aggregation of data ridiculously simple. Indices can be calculated almost directly without elaborate estimation methods.

Because all the small businesses are in the VAT data we can concentrate on the big ones. Sampling becomes much less expensive and much simpler.

2.1 Simple index calculation

In fact the entire index formula we use, excluding the base year, is just

$$i_t = \frac{X_t}{X_{t-12}} i_t,$$

where X 's are sums of the measured variable from units with *comparable* observations from period t and $t-12$. Index points for the base year are calculated using direct sums of each month. Thus the index is non-weighted (or self-weighting) chain index, which is not linked using adjacent periods but rather "year-on-year".

The choice to use of year-on-year linking is somewhat arbitrary, and has been criticised by index theory enthusiasts, but the choice can be defended because the data accumulates slowly.

Were we to link adjacent periods, the effects of missing data on the latest few periods would be greater than they currently are.

With an incomplete data, estimation of change is much easier than estimation of level. Eventually we'll acquire all the data, so the method is very robust. In a simple case an index calculated from total data using the above formula is equivalent with an corresponding index calculated directly from monthly sums of the measured variable. This, of course, does not exactly hold in any real case.

2.2 Even simpler sampling

The VAT data does not replace direct data collection. We need to produce flash estimates before the VAT data is available. Because of slow accumulation and because not all the variables are exactly what we want, we also want to have control over the measurement of at least the biggest enterprises.

It turns out that because we have total information on the *history*, we can disregard the direct collection of all the smaller businesses and use cut-off sampling. This greatly reduces the required sample size. In fact the total number of enterprises in the turnover surveys is currently as low as 2000.

Finland has relatively few big firms: the mentioned 2000 make up a major part of the economic activity. Therefore year-on-year changes in, for example, sales of the large enterprises correlate well with aggregate changes of all enterprises. Since we know from the VAT data the exact value of sales from the previous year, we get a good approximation of sales for the current year by multiplying the previous year's figure by the change calculated from the big enterprises.

A somewhat limping analogue is as follows. Suppose you're sailing. If you know your average speed and heading and *where your were at a prior point of time* you can easily and fairly accurately calculate your current position. If you want to know your exact position, you'll have to use other methods, but that requires more work and time. If you didn't know your previous location, you'd have no choice but to use a more laborious method.

Statistics Finland sponsored a master thesis, in which comparative performance of different sampling and estimation techniques was investigated. The result of the simulations was that a combination of cut-off sample and a year-on-year linked chain index generally gives the best accuracy. Moving to stratified sampling and/or direct estimation of levels reduces precision. The only potential improvement to the index formula suggested by the thesis was to aggregate sub-groups' change rates using fixed weights. The improvement does not, however, seem to be significant enough to outweigh the extra work required and the potential risks in the estimation of weights.

2.3 New kinds of statistics

For those interested in the deep workings of the economy, short term business statistics may seem like simpletons compared to national accounts and other of their structural counterparts. STS are fast, but have only few variables, and cannot encompass all the specifics in the economy.

This can be somewhat relieved. An all-embracing data such as the VAT data enables us to compile statistics that would not be realistically possible from any survey data.

The compilation of *regional* STS has been well established in Statistics Finland. The trick to is to use local kind of activity unit data from the Business Register to disaggregate monthly enterprise level VAT data to a regional level.

The data also allows us to look behind the index number by calculating different sub-domains' *contributions* to the aggregate year-on-year growth. Suppose you are interested in the set of

businesses A and more specifically how firms owned by foreign investors, set F , a subset of A , affects the growth. Then the contribution of the subset is simply

$$c_{F,t} = \frac{X_{F,t} - X_{F,t-12}}{X_{t-12}},$$

where the X :s with subscript F is the sum of the measured variable from the set F . The contribution of the complement set is calculated similarly.

This result generalises to any number of arbitrary sub-domains. By calculating several different contribution tables - such as firms that grow fast vs. the ones that don't, big firms vs. the small ones and so on - one can get a much improved general picture of the current business trend.

The VAT data has also contributed to some profound changes in compilation of statistics which are not directly based on it. In some cases it can be used as an excellent complementary data source. For example in some industries, new orders practically equal sales. It makes much sense to use the sales in the VAT data rather than to collect information from the firms directly. We expect that the VAT data will also have a role in the new value-based industrial production index, which will be launched in few years.

Much more would be possible than described above. The data might allow estimation of birth and death rates of businesses, estimation of distributions of growth in various dimensions, correlation analysis and so on.

3. The bleak truth

Working with the VAT data is not as blissful as one might assume by reading the previous text. A methodologist at Statistics Finland once said - after trying to find suitable imputation methods for the data - that "The VAT data is a cruel data".

In fact there is just one manifold problem, data management. The sheer size of the data set makes it hard to handle with the monthly compilation schedule. Then there are the quality issues, too.

We have no control over the definition of units or variables, and indeed we find a lot of noise in the data. This admittedly makes life more interesting for a statistician. I once called to a hardware store in a small town in eastern Finland to check their VAT figures. An old shopkeeper answered and told me that he's ending his business but still selling his stock. And so I learned how hard it is to find young people interested in continuing a hardware business in the dying municipalities of the remote parts of the country.

3.1 Editing is everything

One of the reason why we can use such a simple index formula as described in chapter 2.1 is that we invest a lot of effort in making the data fit for aggregation. During the years we have put up a heavy automatic and manual editing system for making sure that what we measure is what is actually true. To illustrate one thing we have to take into account, suppose we have a firm with a turnover series like this:

	<i>Jan.</i>	<i>Feb.</i>	<i>Mar.</i>	<i>Apr.</i>	<i>May</i>	<i>Jun.</i>	<i>Jul.</i>	<i>Aug.</i>	<i>Sep.</i>	<i>Oct.</i>
<i>turnover</i>	100	110	120	130	140	150

The series cuts off at July. In some cases like this, the firm actually has stopped operating. There might have been a fire for example. More likely though, the firm is happily keeping business as usual and just hasn't filled the VAT forms for the last four months. Or, the firm has merged into an other firm.

Of course we have no a priori information on what's going on. In our current system we automatically omit the last four observations in index calculation. This may be wrong, so if the enterprise is big enough, it'll show up on the manual editing list and we'll have a closer look.

Now suppose the statistician has noticed the cut in the series and calls to the firm to find out that there has been a merger. In this case we have to describe the event to the automatic editing program, which then processes the series of the participants using the rules we have given.

We've had to program elaborate rules to deal with all the little things happening inside the data. But we can't use the rules blindly: we have to carefully see where to apply them and where to just leave the data as it is. This is of course not specific to our data, but since the data set is big, the magnitude of the issue is larger.

3.2 The trouble with data processing

Currently our perhaps biggest problems relate to the fairly outdated design of our data processing system. When the data set is large and there is lot of computer processing required, one is not supposed to just have few flimsy (but lengthy) SAS-programs and Excel-macros for compiling statistics. However, this has been exactly our situation for seven years. The system was cost efficient to develop and we got the data into good use. Nevertheless, the system has ultimately become a burden.

As of now, we do all processing from the bottom up. Each time we calculate something, we start from the very rawest of the raw VAT data. This causes many problems. First of all, each computer run for recalculation of some 20-30 indices takes two to three hours. We don't have an intermediate database with edited data.

This causes an other problem - potential incoherence. Only some of the corrections to the data are shared to all the statistics. Others are specific to the statistics themselves. Each statistician sees the data differently and thus makes different choices. So we have to be very careful not to make inconsistent choices with the data.

The third problem is the extra work we have to do with statistics that have overlapping frames. Since some of the edits we have made to a firm in one place do not transfer automatically to another place, we have to redo the edits in many places.

Currently we are building a new data processing system for the VAT based statistics. The new system will address the issues mentioned here, so we don't feel that the data is unmanageable. It just needs a well designed system around it. We hope to have the new system running in early 2008.

3.3 Lack of low level quality control

The good thing about direct data collection is that the statistical office gets to decide what definitions are used for the collected variable and what the statistical units are. The data the respondents give back is an another thing, but at least there is a clarity of what the data should be.

There are some outstanding issues with the quality of administrative data. For example when the VAT legislation changes, so does all the data. Also, all the gimmicks firms do, such as changing the corporate structure every few years, mediate directly to the data. And even if there are no structural problems with the units or no changes in the legislation, there might be unwanted components in the variables, such as triangulation, or the firm may change the way how it reports trade between other firms in the same corporate group.

The problem has to be solved somehow and the way to solve it is direct data collection. Having proper figures from the biggest enterprises is absolutely necessary for us. Luckily there have not been major changes in the ways the smaller businesses are measured in the VAT data.

Nevertheless, the biggest firms in the VAT data have to be carefully checked each month, because of the problems mentioned above.

3.4 A burden on productivity

We've mentioned few major problems with the VAT data. They and other less striking issues have made our compilation cycle rather toilsome. Almost 60% of the time we use in the compilation of the VAT based statistics goes to editing. Data collection gets almost 30%, while development, analysis of the compiled data and keeping up contacts to the users have together have only roughly 10% share of the total working time. The absolute time used for editing the VAT data is also quite large.

One of the goals for the new data processing system is to make editing faster and so allow us to concentrate our efforts more on the so far in somewhat neglected parts of our work. There is also an aim to make our work in essence more interesting

4. Conclusion

The VAT data has brought us tremendous possibilities to improve our statistics. It has changed the way we think about making short term statistics and what we think good statistics could be. This has come with a price. Our tools have to be on par with what we are trying to do. Currently we have some obvious problems in the production process, but they seem solvable in the near future.

We'll supposedly see more data collected directly from external databases without being pre-processed by a respondent to suit the needs of the statistical agency. It seems that one way to go is to gather data directly by bridging business management software and data collection. This will even more change the nature of compilation of statistics. We'll have more data to work with than currently, but also much less control over it, so the challenge for economic statistics to be more adaptive to the rapid changes in the economy will be even more significant and concrete than it is today. With the VAT data we at Statistics Finland have had some tasters of the potential issues, and are prepared for tackling with them in the future.