Using administrative data in establishment surveys in Statistics Finland calculating monthly index of industrial production (IPI)

[1]Kari Rautio

Statistics Finland, Business Trends

FIN-00022 Statistics Finland

Email: kari.rautio@stat.fi

Abstract

Industrial Production Index of Finland is compiled using stratified sampling. In stratum one are the enterprises or establishments which have more than 150 employees. In stratum one there is about 300 enterprises or establishments. In stratum two there is PPS (proportional probability sampling) in enterprises or establishments more than 50 but less than 150 employees. In stratum two there is about 600 enterprises or establishments. In stratum three under 50 employees. For this admin data is used. Admin data is utilized in 31 industries of total (NACE 3-digit level). The sum of the NACE group level admin data is used as if it were one enterprise or establishment. We have 31 enterprises in stratum three. These 31 'enterprises' include about 15 000 enterprises. In admin data the publishing month is not available so estimation is needed.

Admin data are very comprehensive. It includes all enterprises that are liable to pay taxes. Data are delivered once a month to Statistics Finland from the Tax Administration. Data update and accumulate for several months after the first delivery. Therefore turnover indices may be revised after the first publication.

Comprehensive data from the previous year makes it possible to estimate the level using only a change estimate.

Change is calculated from a panel of enterprises. The panel is formed of companies with observations on the "current month" and on the corresponding month of the previous year. These three strata in NACE 3-digit level are weighted by gross value. From three digit level upward we use value added weights.

Keywords: admin data, establishment, enterprise, stratum

1. Uses of administrative data in the production of statistics

In Finland, the earliest uses of administrative data can be traced back to census collections in the eighteenth century. Modern statistical uses of administrative data started in 1970 in connection with the population and housing census. At the same time as the census collection methods were being developed, statistical exploitation of administrative data sources was also widened to business statistics. Today, approximately

[1] Kari Rautio, Statistics Finland, Business Trends, FIN-00022 Statistics Finland, kari.rautio@stat.fi

96 per cent of all basic data come from administrative registers. The remaining four per cent is covered by direct inquires. The main administrative data sources Statistics Finland uses are the Population Information System (population, buildings and dwellings), the Real Estate Information System and the Business Information System (Trade Register, Tax Administration data and Business Register).

The Statistics Act of Finland and the 2009 EU Regulation on European Statistics guarantee the access to administrative data. They oblige state authorities to provide Statistics Finland with such data in their possession that are necessary for the production of statistics. According to the Statistics Act, statistics shall be compiled using administrative sources whenever possible. The Statistics Act also gives Statistics Finland the right to access unit level administrative data with identification data and to link them for statistical purposes. The administrative micro level data may not be released to any other authorities apart from certain separately defined exceptions concerning the Business Register.

However, Statistics Finland is entitled to use the administrative registers for different kinds of statistical studies and analyses produced even for outside customers. It is also extremely important that the general public appreciates and understand the benefits from the use of administrative data for statistical purposes. In Finland, the atmosphere has always been trustful when it comes to the actions of the Government or the use of administrative data for the purpose of statistics production. Even though no serious question have been raised in the press with respect to data protection, it is important that the statistical agency always remains on guard in this respect.

In spite of the existence of a legal framework, statistical agencies still rely on the goodwill and co-operation of the keepers of administrative data. It is, therefore, necessary to participate actively in policy development. Statistics Finland is member of the Finnish Register Board Committee, which prepares strategy level definitions for a register policy in Finland. The directors of Statistics Finland and each register authority also meet at regular intervals. Statistical production based on administrative data also requires firm and committed collaboration among the relevant authorities. The use of administrative sources can be improved by working closely with the authorities.

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 response burden on enterprises are lower than in sample surveys. The administrative data are usually available free of charge or the costs are only marginal. The exploitation of administrative data also carries risks which could be avoided in own data collections. The methodological challenges to the short-term statistics using administrative data are usually timeliness and under coverage. By constructing the production of statistics on administrative data, Statistics Finland loses the possibility to decide about the contents of the key data and the related timetables and revisions. Since administrative data are originally collected for purposes other than the production of statistic, the timelines problem is difficult to overcome. Legislative amendments are problematic and usually require major efforts from Statistics Finland. In the worst case, the structures and timetables of administrative data could change completely without Statistics Finland being able to influence it at all.

2. Index of turnover in industry

The index of turnover in industry describes development in the turnover of manufacturing enterprises. Turnover for the largest enterprises in their respective industries is described with the data collected with the sales inquiry while the data on

sales obtained from the periodic tax return data are exploited to describe the turnover of other enterprises. Turnover is exclusive of value added tax. The index is calculated separately for turnover, domestic sales and export turnover.

The calculation is based on estimation of change. Sums by industry are calculated from information on the enterprises with comparable data on turnover from both the examined month and the corresponding month of the previous year. These sums are used to calculate annual changes by industry. Calculation of the annual changes also takes into consideration enterprise openings and closures, as well as restructuring and change of activity. The annual changes thus obtained are used to raise the turnover index of the corresponding month of the previous year. The latest index figures may become slightly revised as the volume of data grows and enterprises report changes to their data. Due to supplementations the data will become updated in releases for over twelve months. The statistics and indices complied from the data are public under the proviso that no individual enterprise's data can be identified from them.

The indices on turnover are based on the Tax Administration's periodic tax return data and on Statistics Finland's sales inquiry. The Tax Administration's data cover all enterprises that are liable for value added tax and pay wages and salaries regularly. The material comprises data collected in connection with the value added tax and employer contribution payment and reporting procedure. The sales inquiry mainly asks about turnover at the enterprise level, but to improve purity by industry, enterprises operating in multiple industries are divided into industry-specific units. Basic information on enterprises, such as that on their industry obtained from Statistics Finland's Register of Enterprises and Establishments is also used in the index calculations.

3. Linking VAT data and survey data in industrial production index

The utilization of administrative data has made it possible to develop new ways of getting improved coverage in statistics that are not based direct on VAT data. An example of this is the revision of the production of the volume index of industrial production where the VAT data have proven as a really good complementary data source. From the quality aspect, the exploitation of administrative data also increases the coherence of economic statistics.

Statistics Finland embarked on the development of the volume index of industrial production in 2006. The reason for using the VAT data in the IPI is that the growth rates of different enterprise size categories have varied significantly in Finland. Significant proportion of small and medium sized enterprises in certain industries are left outside the sample survey. Further reasons were the introduction of deflation of data on value as the main calculation method, and improvement of the coverage of the index without significant increase of the response burden. Before any changes were made to its production, the IPI it was based only on data collected direct from a sample of 1 000 establishments. The statistics were released at the lag of only t+30 days. Since the beginning of 2010, Statistics Finland has extended the exploitation of the VAT data to the production of the IPI where they play a dual role. On the one hand, the VAT data have been used to increase the share of value data in the calculation of the index. They have been used to help the transformation of the industrial production index from a conventional product-based index to an index in which the main data type used for

volume calculations are deflated value data. In 2006, the share of value data in the industrial production index was still approximately 20 per cent. By now, the share had grown to over 70 per cent. The introduction of the VAT data was part of this change in methodology. On the other hand, and perhaps even more importantly, the VAT data are used to improve the coverage of the sample of the index of industrial production. In general, small units pose a problem for survey-based indices like the Finnish industrial output index. It is not possible to include enough small units in the sample to get sufficiently good coverage because that would impose too much response burden on the enterprises. Depending on the industrial structure, this may have considerable distorting effects. Especially problematic are industries like the Manufacture of textiles (13), Manufacture of wearing apparel (14) and the Manufacture of fabricated metal products (25), in which the share of small enterprises is large. Historically, a significant number of small enterprises have been left outside the sample from these industries. Furthermore, it is likely that the output of the excluded enterprises may develop differently from that of the sample group.

The domain of the index is divided into three strata by NACE 3-digit levels. Stratum 1 consists of all industrial establishments of enterprises with more than 150 employees. Stratum 2 consists of a PPS sample of the industrial establishments of enterprises with 50-149 employees. Depending on the 3-digit level, stratum 3 consists of all industrial enterprises (VAT data) or of a PPS sample. The VAT data are used for enterprises with fewer than 50 employees as an alternative for the value of production in selected NACE groups, where the time lag between production and delivery is short. This is an example of how the use of comprehensive administrative data prevents growth of the response burden and makes it possible to optimize sample sizes.

4. Using VAT data for IPI calculation

VAT data cannot be utilised in the index for describing the development of enterprises with under 50 employees without consequences or problems. The data accumulate slowly, which causes revisions, especially in the first estimates. In the case of the industrial production index the first delivery of the VAT data is one month late. The timeliness problem has two aspects. First, even the existing data file is not fully complete when it first arrives at Statistics Finland. Some enterprises report their data late and the file continues to be updated over a period of six months. The revisions resulting from late reporting tend to focus on the second submission of the VAT data. In the first submission, an average of 11 per cent of the turnover data of enterprises with fewer than 50 employees is missing from the file. A second and more important aspect of the timeliness problem is that the data for the most recent month are still missing altogether when the industrial production index is released for the first time and have to be estimated. Despite these obstacles, the VAT data are currently used for 31 industries (Appendix 1) for which their use has been evaluated as being the most helpful. After the made revision, the publishing of the volume index of industrial production was postponed by 10 days at beginning of 2009 because this way the VAT data would be available for the second release. The release schedule still meets the requirements of the regulation of the European Union on short-term statistics.

Using VAT data in calculating the index for small and medium size enterprises in Stratum 3, we have to estimate the index for the month of publication. The estimation is based on VAT data from the previous three months. In this example NACE 282, there is about 600 enterprises. We calculate the value added of these 600 enterprises as if it was one enterprise. We also calculate the previous months of the publishing year and previous

year's value added data for the same months. In the example, we estimate the March 2016 value sum. We use a weighted sum in estimation. The weights in February 2016 were 0.4 times the value added sum for February and also the same weight to February 2015. In January 2016 and December 2015, the weights were 0.3 times the January 2016 and the December 2015 VAT data. The same weights were used to calculate the previous year and VAT data for the same months. We sum these weighted values together and divide it by the sum of the previous year's weighted value sum. Then we get the coefficient (Appendix 2) by which we multiply the value sum of March 2015 and get the estimation for March 2016. This estimation causes revisions (Appendix 3).

5. Calculating the volume index of industrial output

Method applied:
In Finland, the implementation of the Laspeyres chain index method in the calculation of the monthly index of industrial production began in 2002. We implemented the chain index method partly because of some user feedback saying that the base year index did not reflect the structural change of economy quickly enough. For example, the index undervalued electrical and electronics industry and overvalued construction sector in the late 1990s.
Chaining period is each December. Weights are changed annually. Product and establishment weights are changed every year in March. For example from March 2016 onwards we use 2014 weights (publishing January index). Monthly IPI is sample based, where the sample side varies from 100 % in the Pulp and paper industries to less than 40 % in the Manufacture of textiles. We benchmark the sample based monthly indicator to the so called total annual index. The annual index is based on annual industrial output. March 2016 we published and calculated the annual index 2014.

Calculation of elementary indices
Establishment level
The establishments report their monthly production volumes/values by product, and the lowest-level elementary index (establishment level) is calculated based on this information. Value-form information is deflated before calculations.
First, the reported current monthly production level of each product is compared against the previous year's average monthly production level of the product. Establishment's paired comparisons are then combined by summing their weighted value. The product-level weights are based on each product's production values from the previous year. The establishment-level volume index is formed by dividing the sum of the weighted paired comparisons by the total of the product weights and multiplying by 100. The calculation formula can be found below.

$$IND_{t*}^{t,m}\left(establishment_j\right) = \frac{\sum_{i=1}^{n}\left[w_i\left[\frac{q_i^{t,m}}{\bar{q}_i^{t-1}}\right]\right]}{\sum_{i=1}^{n} w_i} \times 100.$$

$IND_{t*}^{t,m}\left(establishment_j\right)$ = volume index for establishment j on year t, month m and statistical year t*.
w_i = product weight for product i
$q_i^{t,m}$ = production of product i in year t and on month m
$\bar{q}_i^{t-1}$ = average of monthly production of product i in year t-1

n = total number of products of establishment j
t^* = statistical year, i.e. the year of index calculation

Stratum level
The industry index is first calculated separately for each stratum. Stratum 1 consists of establishments employing 150 or more employees, stratum 2 of establishments with 50-149 employees and stratum 3 of enterprises with less than 50 employees (VAT data).
To form the stratum-level industry index, a sum of weighted establishment-level indices is first calculated. Either establishment weights (sum of establishment's product weights) or weights based on sampling design are used in weighting. Then, the sum of weighted establishment-level indices is divided by the total sum of establishment and sampling design weights to get the stratum-level industry index. The stratum-level industry index is calculated both for the statistical year t and for the comparison year t-1. The calculation formula can be found below.
Note: VAT data is utilized to calculate the industry index for Stratum 3.

$$IND_{t*}^{t,m}\left(industry_{aaa}^{S_k}\right) = \frac{\sum_{i=1}^{n}\left[W_i \times IND_{t*}^{t,m}\left(establishment_j\right)\right]}{\sum_{i=1}^{n} W_i}$$

$IND_{t*}^{t,m}\left(industry_{aaa}^{S_k}\right)$ = volume index for Stratum k, industry aaa, year t and month m calculated in statistical year t*
$IND_{t*}^{t,m}\left(establishment_j\right)$ = volume index for establishment j, year t and month m calculated in statistical year t*
W_i = establishment weight for establishment i
n = total number of establishments belonging to Stratum k on industry aaa

Industry-level
At this stage of the index calculation process, the stratum-level indices are combined to an industry-level index. Stratum indices are weighted using information on the gross value of companies corresponding to each stratum's size. The gross value information is collected from the Structural Business Statistics information. The industry-level index is calculated for both year t and year t-1. The calculation formula can be found below.

$$IND_{t*}^{t,m}(industry_{aaa}) = \frac{\sum_{k=1}^{p}\left[G_k \times IND_{t*}^{t,m}(industry_{aaa}^{S_k})\right]}{\sum_{k=1}^{p} G_k}$$

$IND_{t*}^{t,m}(industry_{aaa})$ = new volume index for industry aaa, year t and month m calculated in statistical year t*.
$IND_{t*}^{t,m}(industry_{aaa}^{S_k})$ = volume index for Stratum k, year t and month m calculated in statistical year t*.
G_k = weight for stratum k, i.e. gross weight
p = total number of strata on industry aaa ($1 \leq p \leq 3$).

Industry-level, final
The final industry-level index is achieved by calculating the relative change between the new index for year t and t-1. The comparison year's index is multiplied by this change and thus the final industrial production index for statistical year t* is formed.

$$IND^{t,m}_{t*,final}(industry_{aaa}) = \frac{IND^{t,m}_{t*}(industry_{aaa})}{IND^{t-1,m}_{t*}(industry_{aaa})} \times IND^{t-1,m}_{t-1*,final}(industry_{aaa})$$

$IND^{t,m}_{t*,final}(industry_{aaa})$ = final industry-level index for industry aaa in year t, month m and statistical year t*.
$IND^{t,m}_{t*}(industry_{aaa})$ = new volume index for industry aaa, year t and month m calculated in statistical year t*.
$\frac{IND^{t,m}_{t*}(industry_{aaa})}{IND^{t-1,m}_{t*}(industry_{aaa})}$ = relative change of the new volume index on industry aaa for year t and month m to the new index of the corresponding month in year t-1 calculated in statistical yeat t*
$IND^{t-1,m}_{t-1*,final}(industry_{aaa})$ = final industry-level index for year t-1 and month m calculated in statistical year t-1*

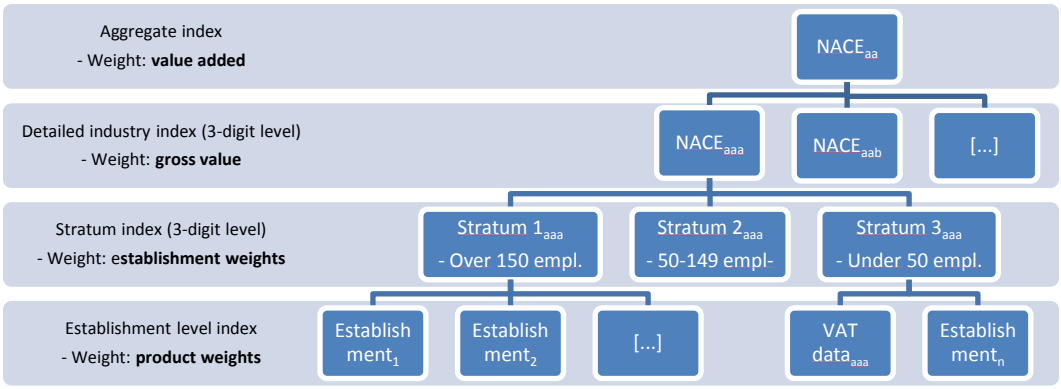
Calculation of higher-level indices
The final industry-level indices can be used to calculate varying aggregate indices. The aggregate indices are calculated for both the statistical (current) and the comparison year. In aggregation, the industry is weighted based on its value added in the Structural Business Statistics information. Correspondingly to the calculation of the final industry-level index, the relative change between the statistical and comparison year's new aggregate index is calculated. The final, corresponding aggregate index from the previous year is multiplied by this change to form the final aggregate index for the statistical year.

Source of weights
Survey data: Product weights (value of establishment's previous year's production of the product) and establishment-level weights (sum of product weights or weights based on sampling design).
Structural Business Statistics: Stratum's gross-value weights and the industry's value added.

Year of weights
Product and establishment weights from year t-2 are introduced yearly in March (JAN publication).
For example in 2016 from JAN industry weights are from 2014 (t-2).

For volume indices: at what level is seasonal adjustment performed
Seasonal adjustment is performed at 3-digit level.

Figure 1. Calculating the monthly IPI



Summary

VAT data does not depict the value of monthly production (e.g. inventories, sales of capital goods). The data are not available for the current month so we have to use estimation (Appendix 2).

Estimation can cause substantial source of revision (Appendix 3). This is not a good thing considering that the focus is always on the latest release. The latest release typically gets a lot of media coverage but revisions are almost never mentioned. In addition, the VAT data are revised due to increased accumulation of data. Also changes in the business cycle emphasise problems in estimation.

On the other hand, similarly problems with sample selection are emphasised because expansions and contractions may affect companies of different size in different ways, and at different points in time. There is some indication that the estimates are currently biased.

Using administrative data instead of directly collected data gives us many advantages, like coverage, but at same time, we lose detailed control of the schedule and content of the data. Using administrative data both solves problems and causes new problems. The key lies in finding the right balance. Statistics Finland has an ongoing project where short-term statistics are being renewed. At the beginning of 2017, Statistics Finland will renew its short-term statistics. In the renewal of the volume index of industrial output the use of periodic tax return data will be increased. A majority of industries (3-digit level) will be produced directly with the help of periodic tax return data. The remaining industries (3-digit level) are produced with the help of an inquiry and the use of periodic tax return data for enterprises with fewer than 50 employees is increased. This means increases revisions but improves coverage.

References

Appelqvist, J., (2012). *Experiences with estimating administrative data using a time-series model: An interim report*, deliverable ESSnet WP4.
Koskinen, V., (2009). *Preparing for Changes in Administartive Data for Short Term Statistics*. OECD STESEG Meeting, 10–11 September 2009.
Koskinen, V., (2007). *Managing administrative data in statistics*.
Paavilainen, P., (2011). *Efficient use of administrative data in the production of economic statistics in Finland.*
Rautio, K., (2013). *Calculating monthly index of industrial production.*

Appendix:

1. VAT data are currently used for these 31 industries in IPI

NACE	Label	Number of Enterprises
102	Processing and preserving of fish, crustac. and molluscs	138
103	Processing and preserving of fruit and vegetables	133
107	Manuf. of bakery and farinaceous products	676
109	Manuf. of prepared animal feeds	66
139	Manufacture of other textiles	602
141	Manuf. of wearing appare, except fur apparel	971
181	Printing and service activities related to printing	902
201	Basic chemicals, fertilisers	38
205	Manufacture of other chemical products	59
222	Manufacture of plastic products	502
232	Manufacture of refractory products	7
236	Manufacture of articles of concrete, cement and plaster	208
237	Cutting, shaping and finishing of stone	220
251	Manufacture of structural metal products	1304
255	Forging, pressing, stamping and roll-formint of metal	97
256	Treatment and coating of metals: machining	2041
257	Manuf. of cutlery, tools and general hardware	225
259	Manufacture of other fabricated metal products	770
262	Manufacture of computers and peripheral equipment	49
263	Manufacture of communication equipment	53
265	Manuf. of instruments and appliances for measuring testing and navication	195
282	Manufacture of other general-purpose machinery	577
289	Manufacture of other special-purpose machinery	472
301	Building ofships and boats	265
310	Manufacture of furniture	924
321	Manufacture of jewellery and the others	429
323	Manufacture of sports goods	157
325	Manufacture of medical and dental instruments and supplies	123
329	Manufacturing n.e.c.	239
331	Machines of metal products, of industry	1975
332	The machines of the industry and equipment	429
	Total number of enterprises	14846

2. Example of the industrial production index and the VAT data in industry 282

Example of using VAT data in one industry (282):										
Industry	282									
Estimating	Value sum	Comparable value sum								
2016/03	81059764	86219496	2015/03							
2016/02	59865779,5	58122116	2015/02							
2016/01	52212912,9	57757647	2015/01							
2015/12	46525579,3	54671656	2014/12							
2016/02	23946312	23248846	2015/02							
2016/01	15663874	17327294	2015/01							
2015/12	13957674	16401497	2014/12							
	53567859	56977637	0,940156							
			Coefficient by which we multiply 2015/03 value sum and we have estimated value 2016/03							

3. Figure of revisions using VAT data in nace 282