Statistics Netherlands
Kees Nieuwstad[1] (cnwd@cbs.nl),
Draft, Latest changes: 23/05/2003

"Case study" of the use of earlier versions of the statistics on industrial turnover.

1 Introduction

The call to start publishing statistics faster is getting louder. One problem, however, is having to weigh quality and timeliness, since publishing faster often goes at the expense of quality.
Here we report on a study of industrial turnover statistics where we looked at various aggregation levels and compared earlier versions to a much later version. In this way we could observe the loss of quality in the statistic if we were to publish it earlier.

The main conclusion is that working with earlier outcomes offers good perspectives. In the case we studied, the difference between the estimated turnover between the earlier and later versions was often very small.

2 Calculation method

There are a number of codes used in the turnover statistics. 36001=is the turnover in a given domestic SIC category. 36002= is the turnover in a given foreign SIC category. G stands for the estimated part. So 36001G is that part of domestic turnover (36001) that was estimated. So 36001 itself *includes* the estimated part. This knowledge can help calculate the weighted non-response rate for the sum of the domestic and foreign turnover:

$$Non-response-rate = \frac{36001G + 36002G}{36001 + 36002} * 100\%$$

The total turnover can also be calculated:

$$Total\ turnover = 36001 + 36002$$

By making these calculations at various times (we will refer to this as date picks) after sending the turnover survey, we can monitor both the non-response rate and the total turnover as a function of time. This was done with the turnover survey of September 1991. The date picks are shown in the table below.

Month	Date pick 1	Date pick 2	Date pick 3
September (0109)	31-10-01	05-11-01	30-11-01

There are two working days between pick 1 and pick 2 and eighteen working days between pick 2 and pick 3. The idea is that the data should become better with each pick. Pick 3 is closest to the "best data set possible" because virtually all data should be in by then.

[1] The views expressed are the author's and are not necessarily shared by Statistics Netherlands.

3 Analysis of non-response

The classical way to determine whether a statistic can be published at Statistics Netherlands is to look at whether the non-response rate is low enough. Usually a threshold is set for this. In this paper we set the threshold at 30%. This means that a group with a given aggregation level is considered publishable once the non-response rate for the group is less 30% or less.
The method described in paragraph 2 is used to determine the non-response for groups with different aggregation levels for each pick. The results are shown in Table 1.

Table 1. Non-response rate for the different picks of September 2001. MIG 040=Intermediate goods, MIG 050=Capital goods, MIG 060=Consumer durables, MIG 070=Consumer non-durables and MIG 090=Energy.

Group	Non-response pick 1	Non-response pick 2	Non-response pick 3
SIC 14	21%	20%	5%
SIC 15	33%	28%	16%
SIC 16	23%	23%	23%
SIC 17	31%	23%	13%
SIC 18	46%	40%	16%
SIC 19	16%	15%	13%
SIC 20	31%	28%	17%
SIC 21	20%	19%	11%
SIC 22	45%	31%	14%
SIC 23	97%	14%	0%
SIC 24	18%	15%	10%
SIC 25	46%	37%	26%
SIC 26	33%	28%	22%
SIC 27	26%	19%	12%
SIC 28	45%	37%	25%
SIC 29	42%	37%	23%
SIC 30	80%	74%	68%
SIC 31	27%	26%	19%
SIC 32	9%	8%	6%
SIC 33	36%	35%	25%
SIC 34	22%	21%	12%
SIC 35	46%	36%	21%
SIC 36	32%	29%	20%
SIC 37	35%	35%	25%
MIG 040	25%	22%	15%
MIG 050	43%	40%	28%
MIG 060	15%	14%	10%

MIG 070	34%	27%	15%
MIG 090[2]	-	-	-
Super total	37%	26%	16%

When we look at the thresholds set, the publishable sets in pick 1 are the SIC-groups 14, 16, 19, 21, 24, 27, 31, 32 and 34. The sum of the turnover of these groups is 11187345, which equals 36% of the super total of pick 1.
For pick 2 we can add SIC-groups 15, 17, 20, 23, 26, and 36 to the publishable turnover results. The turnover sum of the publishable groups is 22682736, which equals 72% of the super total of pick 2. Finally, for pick 3 we can publish all SIC-groups except SIC-group 30 on the basis of the threshold criteria. This equals 98% of the super total of pick 3.

According to the criteria the MIGs 040 and 060 could be published in pick 1. After pick 2 MIG 070 could be added. Please note that the super total can be published after pick 2 according to the criteria.

In short, when the non-response goes down, the number of publishable 2-digit SIC-groups and MIGs increases.
The percentage of the turnover that can be published at the 2-digit level increases from 36% in pick 1 to 72% in pick 2 and to 98% in pick 3.

The criteria presented above are rather rigid since they only look at the response volume. In the next paragraph we look at different criteria for determining whether or not a statistic can be published.

4 Analyse of the non-response in combination with turnover

It is interesting to look at another approach than the classical Statistics Netherlands way (see paragraph 3) to see whether a statistic can be published. One can check whether the total turnover as a function of non-response converges to a certain value.
We will show the idea behind this in a brief example.
Suppose that the value of the total turnover in pick 1 is not very different from the value of the total turnover in pick 3. If so the quality of the value of the total turnover in pick 1 is virtually the same as the quality of the value of the total turnover in pick 3. In theory this makes it possible to publish earlier without loss of quality. The value of the total turnover in pick 1 can be published because the value is accurate enough. The gain in timeliness is the difference between the moments in which picks 1 and 3 are taken. The statistic could be speeded up (in theory).

In order to verify whether this argument works in practice, we extended Table 1 from paragraph 2 with five columns. This resulted in Table 2 below with the total turnover for each pick of the groups at different aggregation levels added. The table also shows how much the total turnover of picks 1 and 2 differ from the total turnover of pick 3.

[2] According to the definition MIG 090 consists of the 2-digit SIC-groups 10, 11, 12, 23, 40 and 41. Because only SIC 23 is represented in the turnover statistics data the coverage is not considered sufficient to give the results for MIG 090.

Table 2. Non-response rates in combination with the total turnover for the different picks of September.

Group	Non-response pick 1	Non-response pick 2	Non-response pick 3	Total turnover pick 1	Total turnover pick 2	Total turnover pick 3	Difference between pick 1 and pick 3	Difference between pick 2 and pick 3
SIC 14	21%	20%	5%	105147	104891	108191	-3%	-3%
SIC 15	33%	28%	16%	6679879	6753655	6770690	-1%	0%
SIC 16	23%	23%	23%	571365	578112	578112	-1%	0%
SIC 17	31%	23%	13%	446377	449043	432551	3%	4%
SIC 18	46%	40%	16%	72434	77169	74884	-3%	3%
SIC 19	16%	15%	13%	43535	43395	43151	1%	1%
SIC 20	31%	28%	17%	313394	316042	310413	1%	2%
SIC 21	20%	19%	11%	989942	989956	990005	0%	0%
SIC 22	45%	31%	14%	1902485	1938701	1793414	6%	8%
SIC 23	97%	14%	0%	2569988	2556236	2525427	2%	1%
SIC 24	18%	15%	10%	5841150	5813314	5858263	0%	-1%
SIC 25	46%	37%	26%	901264	895497	901745	0%	-1%
SIC 26	33%	28%	22%	958028	970415	962352	0%	1%
SIC 27	26%	19%	12%	1085209	1005178	997095	9%	1%
SIC 28	45%	37%	25%	2060090	1987810	1953263	5%	2%
SIC 29	42%	37%	23%	2033870	2074970	2094009	-3%	-1%
SIC 30	80%	74%	68%	578412	940780	772163	-25%	22%
SIC 31	27%	27%	19%	479016	477826	489876	-2%	-2%
SIC 32	9%	8%	6%	893555	891417	876470	2%	2%
SIC 33	36%	35%	25%	365474	362023	403507	-9%	-10%
SIC 34	22%	21%	12%	1178426	1171844	1172521	1%	0%
SIC 35	46%	36%	21%	464651	486944	616938	-25%	-21%
SIC 36	32%	29%	20%	564627	561412	577976	-2%	-3%
SIC 37	35%	35%	25%	76244	76244	76881	-1%	-1%
MIG 040	25%	22%	15%	11452038	11355874	11398174	0.5%	-0.4%
MIG 050	43%	40%	28%	5727090	6077951	6064381	-5.6%	0.2%
MIG 060	15%	14%	10%	1405229	1415199	1408642	-0.2%	0.5%
MIG 070	34%	27%	15%	10020217	10117614	9983273	0.4%	1.3%
MIG 090	-	-	-	-	-	-	-	-
Super total	37%	26%	16%	31174562	31522874	31379897	-0.7%	0.5%

When we take it that the turnover value of a SIC group can be published when it deviates less than 5% from the turnover in pick 3, then all SIC groups for pick 1 can be published except 22, 27, 30, 33 and 35. The sum of the turnover of these groups is 26778331, which is 86% of the super total of pick 1. After pick 2 it would also be possible to publish SIC group 26. This means that 88% of the super total of pick 2 is publishable.

Because a 5% deviation is rather large, we can also set the limit a 3% in this analysis. The difference for pick 1 is that SIC group 28 cannot be published. It also implies that the sum of the turnover of the publishable groups equals 24718241, which is 79% of the super total for pick 1. For pick 2 the turnover part that can be published on the basis of these criteria is 87%.

Finally we looked at a maximum deviation of 1% in the turnover of picks 1 and 2 with pick 3. In that case 56% of the super total of pick 1 can be published and 73% of pick 2.

SIC-groups that stand out

As we mentioned before in pick 1, the SIC-groups 22, 27, 30, 33 and 35 would not be publishable when we set the criteria at a maximum of 5% difference with pick 3. The question is why the difference is more than 5%.

SIC 33 and SIC 35 are described as *manufacture of medical, and surgical equipment and orthopaedic appliances etc. precision and optical instruments, watches and clocks,* and *the manufacture of other transport equipment (excl. motor vehicles, trailers and semi-trailers)*.
This difference may well be caused by the estimation procedure. Both SIC-group 33 and SIC-group 35 often deal with payments in instalments, which are notoriously hard to estimate.

Another special case is SIC 30, *the manufacture of office machinery and computers*. The difference with pick 3 changes from –25% in pick 1 to +22% in pick 2. This is a major switch which can be explained perhaps by the high non-response rate for picks 1 and 2: namely 80% and 74% respectively. This means that so few data are in at that time that it may be too much to expect a proper estimate.

One SIC-group that shows a clear improvement in the estimated value between pick 1 and pick 2 as compared to pick 3 is SIC 27. In pick 1 the difference is still 9%, whereas it is only 1% in pick 2. So there is an obvious quality improvement in the value of turnover between picks.
Another issue shows up here. The classical way mentioned in paragraph 2 would yield less satisfactory results for SIC-group 27 than the method described here. The non-response for pick 1 is 26%, which is below the 30% threshold. According to the classical approach the group can be published, whereas the difference with the turnover value in pick 3 is still 9%.

The last SIC group here is SIC 22 where the difference with pick 3 changes from 6% in pick 1 to 8% in pick 2. So the situation is getting worse, and the cause for this is hard to pinpoint. Probably it lies in the estimation method.

MIGs/Super total

One difference that stands out is the super total of the turnover between pick 1 and 2 is always small in comparison with pick 3: –0.7% and 0.5% respectively.

The difference of the sum of the turnover for pick 2 for MIGs 040-070 is less than 2% compared to pick 3. In pick 1 only MIG 050 deviates more than 2% from pick 3, namely –5.6%.

Graphics method

So far the data were presented in tables, but it is also possible to express the turnover of a SIC-group as a function of non-response. In this way we can check whether or not the turnover value converges as more response is received. We show two examples in Figure 1 and Figure 2.

Figure 1. Non-response as a function of total turnover for SIC 30.

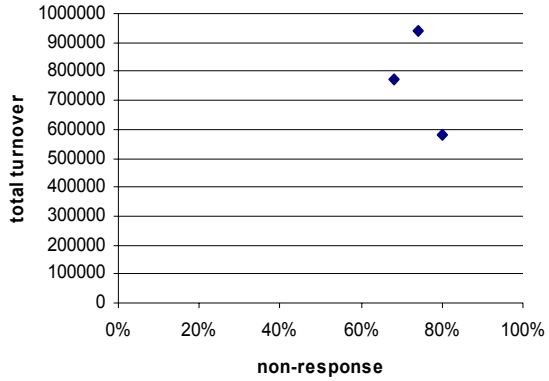
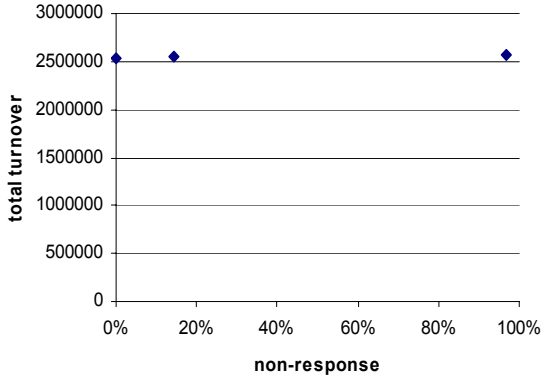


Figure 2. Non-response as a function of the turnover of SIC 23.



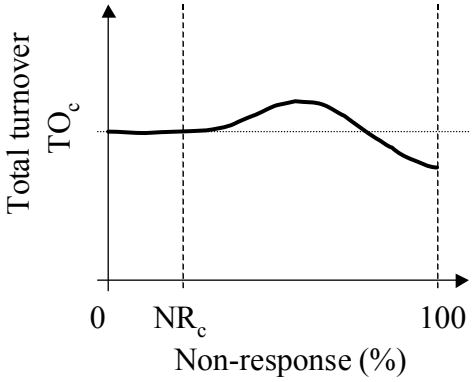
These two figures show the two extremes. SIC group 30 is an example of a poorly converging SIC-group. The non-response drops from 80% to 68% between pick 1 and pick 3. So only 12% more response is coming in. Moreover the turnover value fluctuates substantially as we mentioned before.
SIC 23 is an example of a SIC-group that converges fast. The non-response falls from 97% in pick 1 to 0% in pick 3. The turnover value only differs 1% between pick 1 and pick 3. The turnover difference between pick 2 and 3 is 0%.

Practical methodological problem

The theoretical method described above has one disadvantage. In reality the method cannot be used directly to determine whether a given SIC-group can be published. When pick 1 takes place nothing is known about pick 3, so it is impossible to make the comparison described above. It is only possible to determine with hindsight which SIC-groups would have been publishable. This is not terribly useful, so the question remains how the method could be applied in practice. The solution may be as follows.

Suppose that there are results of the turnover statistics available at any given moment. In that case you can make a graph for all SIC-groups of the turnover in the SIC-group at any moment versus the non-response of that SIC-group at any moment (Compare Figure 1 and 2). We show a stylised plot in Figure 3 as an example.

Figure 3. Stylised plot of the total turnover as a function of non-response.



The problem that nothing is known about pick 3 can be solved by determining the non-response value NR_c whereby the total turnover is sufficiently converged to TO_c. In that case a turnover value of sufficient quality would be published when the non-response value is NR_c .
One way of determining NR_c is to look at past results. When the NR_c is determined for a large number of data sets, it is possible to get a best NR_c value for each SIC-group at which publication is possible. However, this still requires more empirical study.

Although it may be difficult in practice to determine NR_c the results in paragraph 4 indicate that this method may be quite successful. After all there are many 2-digit SIC-groups where the total turnover hardly changes between pick 1 and pick 3. This means that there must be a convergence to TO_c and this in turn means that working successfully with earlier results and increasing the timeliness of the statistic both look promising.

5 Comparing the methods

In this paragraph we compare the percentages of publishable turnover (at the 2-digit SIC level) from paragraphs 3 and 4. The results are shown in Table 3.

Table 3. Percentage of publishable turnover (at the 2-digit SIC level) for the different picks and for the different methods.

Method	Pick 1	Pick 2	Pick 3
Classical way at SN, threshold value 30%	36%	72%	98%
Deviation turnover, maximum 5%	86%	88%	
Deviation turnover, maximum 3%	79%	87%	
Deviation turnover, maximum 1%	56%	73%	

What is remarkable is that only 36% of the total turnover could be published in pick 1 when the classical way is used, whereas the use of an alternative method creates the possibility to publish far more. With a 5% deviation maximum, about 86% of the super total of the turnover in pick 1 can be published. With a 3% deviation it is still possible to publish 79% and with the 1% that is still 56%.

The gain produced by the alternative method is far smaller in pick 2. The classical SN way would yield 72% of publishable outcomes and the three alternative methods yield 88%, 87% and 73% respectively.

6 Conclusions and recommendations

Conclusions

- Working with earlier outcomes offers good perspectives because in the case we studied the difference between the estimated turnover between the earlier and the later results is often quite small.

- Graphs with turnover as a function of non-response can be powerful data analysis tools.

Recommendations

- The study was carried out on one data set only. There is no good second data set available right now that would be fit for this kind of research. It would be advisable to collect such kinds of data sets for follow-up studies.

- This kind of research can be applied to many similar types of turnover statistics. It is interesting to make a similar study of these statistics.

- The application of a more advanced method should not mean any let up in trying to get a good response rate. A good response rate remains one of the pillars of good statistics.