



Modernisation actions for compiling Prodcom Statistics

5.1.2.e

5.1.2.e

5.1.2.e

Contents

Summary	5
1.Introduction.....	7
2. Reduction of sample size (Area 3)	8
2.1 Use of alternative data sources.....	8
2.1.1 Size of the target group.....	8
2.1.2 Stability over time	9
2.1.3 Use of turnover data	11
2.1.4 Volume of sold production.....	12
2.1.5 Simulation for reference year 2016	13
2.1.6 Need for check on specialised manufacturers	16
2.1.7 Implementation plan.....	16
2.2 Optimal Prodcom sampling	18
2.2.1 Current cut-off sample for Prodcom statistics	18
2.2.2 Replacement of quality criteria	19
2.2.3 Alternative Sampling strategies for Prodcom	20
2.2.4 Results of an additional cut off based on the STS production variable	23
2.2.5 The impact for other statistics making use of Prodcom	25
2.2.6 Implementation plan	25

4.1.4



Summary

Objectives

In 2017-2018 Statistic Netherlands performed actions for Area 3 and Area 5 of the grant scheme for the modernisation of Prodcom statistics. One objectives (Area 3) is to research possibilities to reduce the Prodcom sample in order to lower the administrative burden on enterprises. Another objective (Area 5) is to develop a method for observation of sub-contracting and additional industrial services.

Results

A sizable share (about 50 per cent in 2016) of the Prodcom respondents are enterprises specialised in manufacturing a single product in the Prodcom list of products. Specialised manufacturers are likely to keep producing one and the same product for a longer period of time. Nearly 81 per cent of the manufactures, that have produced one and the same class of products for two consecutive years, are still producing this same product six years later. The turnover from industrial activities that is included in the monthly survey for STS (Short Term business Statistics) is a suitable alternative for the Prodcom value of sold production for most stable, specialised manufacturers. A comparison of results for reference year 2016 based on Prodcom data versus the alternative method based on STS showed that effects are small for most of the products. The sold production of 87 per cent of the products had a discrepancy of less than 10 per cent between the current practice and the use of STS as an alternative for the manufacturers that were specialised in one and the same product in 2014 and 2015. In few cases products either appear only when the alternative method is used or do not appear in the results when the alternative is applied while these 'new' products do exist in the 2016 Prodcom survey data.

Next to the use of alternative data sources, the use of reduced sample size will contribute to a reduction of the administrative burden. To reduce the sample size while maintaining the quality or coverage requires an improved sampling strategy. In current practice, all enterprises in manufacturing industries that have 20 or more employed persons are included in the sample. Analysis shows that this cut-off based on input variable (labour) is not cost effective for the Prodcom statistics that measures output. A reduction of 10 per cent of the coverage in STS (turnover from industrial activities) per NACE class results in a reduction of up to 40 per cent in sample size. As the distribution of data is skewed, the number of enterprises is finite and suitable regressor (cut-off variable) is available for most enterprises a cut-off sample based on STS may be an more optimal alternative to current the current practice.

4.1.4



1. Introduction

In this report Statistics Netherlands presents the results of work in preparations of the proposed changes to the regulations (FRIBS) for Prodcom statistics. Statistics Netherlands worked on both Area 3 and Area 5 of the Grant program.

Area 3 focuses on the possibilities to reduce the number of enterprises in the Prodcom sample. The objective is to reduce the administrative burden, especially for small and medium enterprises. One possibility may be excluding manufacturers that are specialized in the production of a single product from the sample for several years. If the products produced remained unchanged, the value of sold production (in Prodcom) will be similar to the turnover from industrial activities, which is collected from the manufacturers on a monthly basis in the Short Term Statistics (STS) questionnaire. Section 2.1 of this reports contains the analysis of the businesses that record one and the same product for multiple consecutive reference years. Another possible way to reduce the administrative burden is to identify a more efficient alternative to the current sampling strategy. In paragraph 2.2 the change from the number of persons employed to the turnover from industrial activities (STS) as a cut-off variable is discussed.

4.1.4

Acknowledgement

Many colleagues contributed to this report by providing crucial information or views on the subjects at hand. Special thanks to 5.1.2.e and other colleagues in the Prodcom, SBS, National Accounts and the Large and complex enterprise groups departments.

2. Reduction of sample size (Area 3)

Two approaches for reducing the sample size of the Prodcom in the Netherlands are described in this chapter. The first one (section 2.2) is the use of turnover data from STS statistics instead of the Prodcom questionnaire for manufacturers specialised in the production of a single type of goods. The second approach makes use of the STS turnover data as well. The objective is to identify a more efficient sampling strategy for Prodcom statistics. Results are presented in section 2.3. The objective is to reduce the administrative burden while maintaining the level of accuracy in the Prodcom results.

2.1 Use of alternative data sources

Work on this topic concentrates on researching the possibility to use other data sources for the observation of enterprises producing only one product (Prodcom) for multiple consecutive years. The basic idea is that for these companies, the value of the sold products is equal to the turnover from industrial activities.

2.1.1 Size of the target group

The size of the target group is the number of companies producing one and the same product for a number of consecutive years. A choice must be made on how many consecutive years of producing one product, as observed via the Prodcom survey, is enough to switch from the Prodcom questionnaire to an alternative datasource. A longer period of production of a single good will improve the quality, but a longer period will reduce the benefits as the size of the target group decreases. This trade-off between quality and potential reduction of the administrative burden is the subject of this research.

Historic data has been collected from the Prodcom statistics over the years 2008 to 2015. Based on this, the following estimation in table 2.1.1 can be made for the number of companies that will be fit for alternative observation. For comparison, the total number of observed enterprises for the Prodcom statistics was about 4.4 thousand in the year 2015.

Table 2.1.1 Expected number of companies fit for alternative observation

Length of period producing one product	Number of companies
2 years	1362
3 years	1186
4 years	1059
5 years	953
6 years	847
7 years	752

A group of about 100 to 150 additional companies meets the requirements too, but belongs to the largest and most complex enterprise groups and in the Netherlands. Statistics Netherlands requires

direct observation for these companies. Therefore these are excluded from the potential alternative data sources and thus from all the results hereafter.

The numbers shown should be interpreted as an estimation of the potential. The exact number of companies and thus the reduction of the survey sample depends on the complete observation strategy that will be needed for production. See paragraph 2.1.5 for more details on this.

2.1.2 Stability over time

Chances are however that companies, after switching to alternative data sources, start to produce a different product or multiple products. In both cases allocation to a product of the Prodcom product lists based on the historical Prodcom data will be inaccurate. This adds a potential margin of error to the alternatives of Prodcom observation. Companies can also completely cease production. Since the end of all manufacturing activities is registered in data sources other than Prodcom, this does not bring additional margin of error. These are therefore excluded from this analysis.

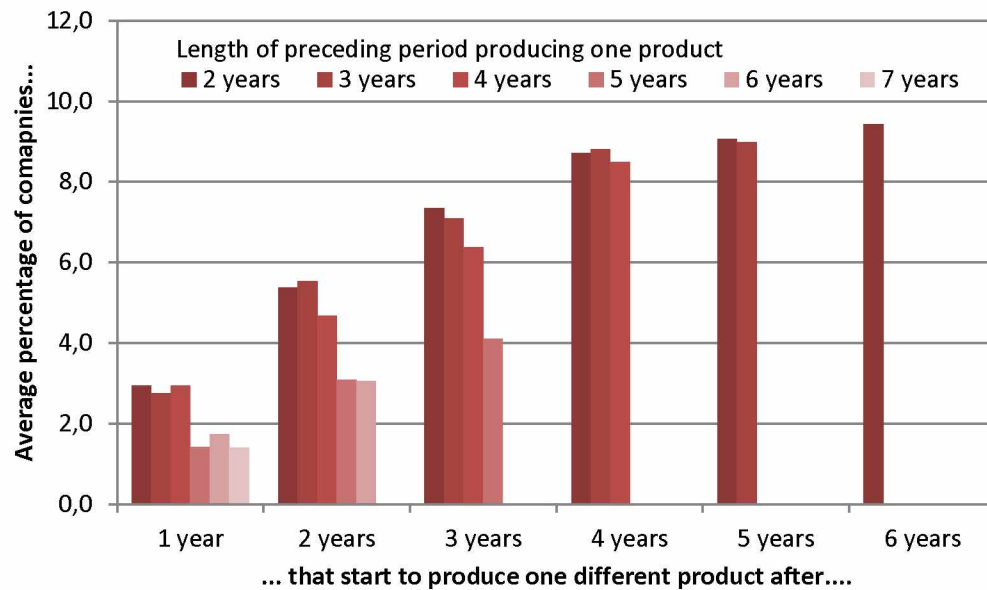
Based on historic data, an estimation is made of the number of enterprises that keeps making the same single product after excluding these enterprises from the Prodcom sample. This number is influenced by two parameters: the preceding period of producing one product required for switching, and the number of years passed after the switch. All combinations are calculated and shown in table 2.1.2.

Table 2.1.2. Expected average percentage of companies still producing the same product after stopping the observation for Prodcom

Length of period producing one product	Years passed after switch					
	1 year	2 years	3 years	4 years	5 years	6 years
2 years	93,4	88,8	85,1	82,3	81,7	80,9
3 years	94,3	89,2	86,1	82,9	82,7	
4 years	94,1	90,5	87,3	83,2		
5 years	96,3	92,8	91,1			
6 years	95,9	93,1				
7 years	97,0					

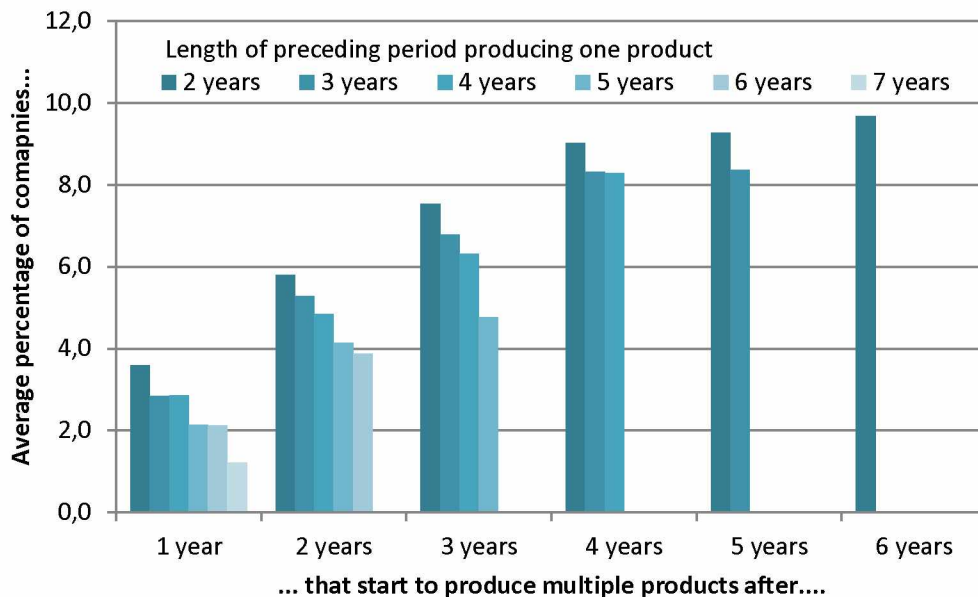
In the following graph the share of companies that start to produce one different product is shown.

Graph 2.1.1. Expected average percentage of companies producing a different product after stopping the observation for Prodcom



In the following graph the share of companies that start to produce multiple products is shown.

Graph 2.1.2. Expected average percentage of companies producing multiple products after switching to alternative data sources



These figures lead to the following findings:

- ☐ If alternative data sources are introduced, for a limited percentage of the companies the production will be contributed to the wrong Prodcom code because the enterprises change their portfolio of products.
- ☐ As expected, this margin of error grows as the years passed after the switch accumulate. However, the growth slows down over time.

- With a low threshold of two, consecutive years of specialised production still 80,9% of the companies are producing the same single product after six years. Apparently, if an industrial producer specialises in a certain product during two years, chances are high it will remain doing so for the following years.

2.1.3 Use of turnover data

As an alternative source turnover collected for the short term statistics (STS) is the most promising. The STS sample survey differentiates turnover from manufacturing from total turnover, which includes trade and other activities as well. Turnover from manufacturing will, in theory, be close to the value of sold production, which is one of the main target variables from Prodcom. The valuation of production under subcontracting will differ between Prodcom and STS (more details on this subject in chapter 3) Also, STS turnover includes production on locations outside the legal boundaries of the Netherlands if this production is under ownership and control of a Dutch resident enterprise. This production is not included in the Prodcom. The STS data is available in time and covers the enterprises of 20 or more employees, which is the current cut-off for the Prodcom sample.

For the following analysis, a scenario is chosen for the alternative observation in which companies will be added after two years of consecutive production of one and the same Product code. After two years of non-observation these specialised manufacturers receive a Prodcom questionnaire in order to check for changes in their output. For research purposes the maximum amount of available companies are taken into account for each starting year. Due to limitations in available data for the previous years, the analysis is run for the years 2014 to 2016. The join of STS and Prodcom data results in 4636 records¹.

First of all the correlation between the two variables is calculated. Since both variables are not normally distributed, Spearman's rho is used for the calculation. It gives a .95 correlation coefficient between the production value and turnover data, with a p-value of 0.00 . This means the correlation is very strong.

To compare the production value with turnover data even more, the differences between the two indicators are studied. Table 2.1.3 gives an overview of this. Note that the margin of error as discussed in paragraph 2.1.2 has an effect on this outcome.

Table 2.1.3. Comparison of production value and turnover

	Number of observations	Share (%)
Turnover is higher	2409	52%
Turnover is lower	1759	38%
Turnover is equal	354	8%
Either turnover or value lack	114	2%
Total	4636	100 %

First of all, this table shows that the turnover data has a great availability. Only in 46 cases (< 1% of the total) no turnover data was available. Second, using turnover data as a source seems to lead to an overestimation of the production value. For 52% of the cases the turnover value exceeds the

¹ One record holds one observation, which is a combination of a Prodcom code and a company.

production value. The median of the difference between the two indicators is 6.000 euro's in favour of the turnover, which translates into a 0,07% higher value per observation. This means the overestimation is very limited. Minimum, mean and maximum numbers are highly influenced by outliers, which are not excluded yet.

Work on this matter is not finalised, but these results indicate that turnover data is a useful administrative source for the Prodcom statistics. Section 2.1.7 presents recommendations for further research.

2.1.4 Volume of sold production

Using turnover data (STS) as a source for value of sold production (Prodcom) means no survey will be sent to the companies involved, and therefore no data will be collected on production volume. Since no useful administrative source for volume has been detected yet, these values have to be imputed. Therefore a study has been carried out about whether this is possible.

Data has been collected on the imputations used in the year 2014 to 2016. For these years, detailed information about the used imputation method has been registered in the Prodcom production system. All measurements are based on data from the provisional publication for each of the years. This is because this status contains the most imputations, thus providing the clearest impression of the situation.

First of all, the total number of observations and imputations has been calculated. They are shown in table 2.1.4.

Table 2.1.4. Overview of observations and imputations

	2014	2015	2016
Number of observations	10729	10871	11067
Number of observations with imputations	1214	1158	1315
As percentage of the total observations	11%	11%	12%
Of which both imputed value and volume	6%	7%	8%
Number of imputations of production values	789	1020	1080
Number of imputations of sold volume	1000	838	1000

These numbers are also calculated for the scenario in which alternative data sources would have been used in these years. The scenario is described in paragraph 2.1.2.

Table 2.1.5. Overview of observations and imputations of specialised producers excluded from the Prodcom survey

	2014	2015	2016
Number of alternative data sources	1760	1719	1829
As percentage of all observations	16%	16%	17%
Number of alt. observations with imputations	182	148	162
Number of imputations of production values	120	131	142
Number of imputations of sold volume	137	99	106

Statistics Netherlands uses several different methods to impute data for the Prodcom statistics. The following conclusions can be drawn on the effect of introducing STS as an alternative data source for specialised manufacturers:

- Imputing production value is not needed anymore for all administrative observed companies (~ 130 cases).
- Imputing production value for all other observations (~ 900 cases) will not be effected, since none of the imputation methods depend on data from other companies.
- The number of imputations of volume of sold production will increase (~ 1600 – 1700 cases). Currently these volumes are observed in the survey, but this information will disappear.
- For all imputations of sold volume, a method will remain available. Currently about 85% of them are based on historic data from the same company, using the sold volume in either the previous year or the year before. Depending on the observation strategy for specialised manufacturers (see paragraph 2.1.5), this data might not be available anymore. Another method of lower quality will be used, for example average numbers of the parent Prodcom group (either 8 or 6 digit).

2.1.5 Simulation for reference year 2016

So far no insight was provided on the effect of changing from a questionnaire to STS data for specialised manufacturers on the actual value of sold production nor on a product level. As the values (and volumes) are published by product this is important. If products are produced predominantly by specialised producers, the risk of misestimating on product level increases. The results presented in this section are based on the preliminary 2016 Prodcom data, the final Prodcom data for 2014,2015 and STS data for these reference years, excluding the enterprise units under large and complex holdings.

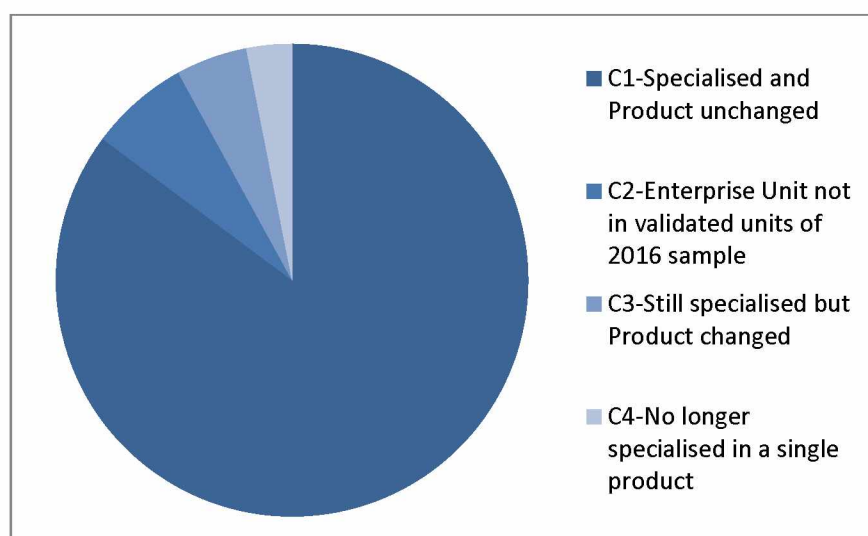
The Prodcom data for 2016 consists of 4456 enterprises that either responded or were imputed of which 418 enterprises are managed by the large and complex holdings (LCH) department. These latter enterprises are excluded from the plans for the use of STS data. Out of the remaining 4028 enterprises, a little over 50 per cent was specialised in a single product (i.e. one Prodcom class). The number of enterprises (excl LCH enterprises) that was specialised in a single product in three consecutive years (2014-2016) is 1400. The production value in 2016 of these specialised enterprises is on average similar to the non-specialised businesses; 35 per cent of number the enterprises that cover 31 per cent of the production value. The specialised producers that have a production value of over € 100 million should possibly remain in the sample, simply because of their size.

Figure 2.1.1 Size class total sold production in 2016 for manufacturers that are specialised for three consecutive years (2014-2016) versus other producers (excluding enterprises in the large and complex holdings).



In the period of 2014-2015, the number of enterprises specialised in one product (excluding LCH enterprises) was 1643. As said, 1400 remained specialised in the third consecutive year (2016). Figure 2.1.2 shows the 2016 situation for the specialised manufacturers of 2014-2015. Seven per cent of the enterprises are no longer in 2016 Prodcom data (C2). This may be the case when the number of employed persons drops below the threshold (20 persons) or in case of a bankruptcy, merger or other event in the business register. The enterprises that change to another product (C3) or start producing multiple products (C5) will be wrongly allocated to a product class when the 2016 STS data and the product class of 2014-2015 used instead of the Prodcom Survey.

Figure 2.1.2 The 2016 situation of the specialised producers from 2014-2015, excluding enterprises in the large and complex holdings.



The overall effect of replacing the 2016 Prodcom data with STS data for the 1643 specialised producers in 2014-2015 is very small. The overestimation on the sum of all products is less than 1 per

cent using the level of industrial turnover from STS. The overestimation is slightly smaller when the 2015-2016 development in this STS variable is applied to the 2015 value of sold production (Prodcom) for the specialised producers. Given the results of table 2.1.3 an overestimation was expected. For seven specialised enterprises no STS data was found. These cases are for now imputed by a zero value of sold production.

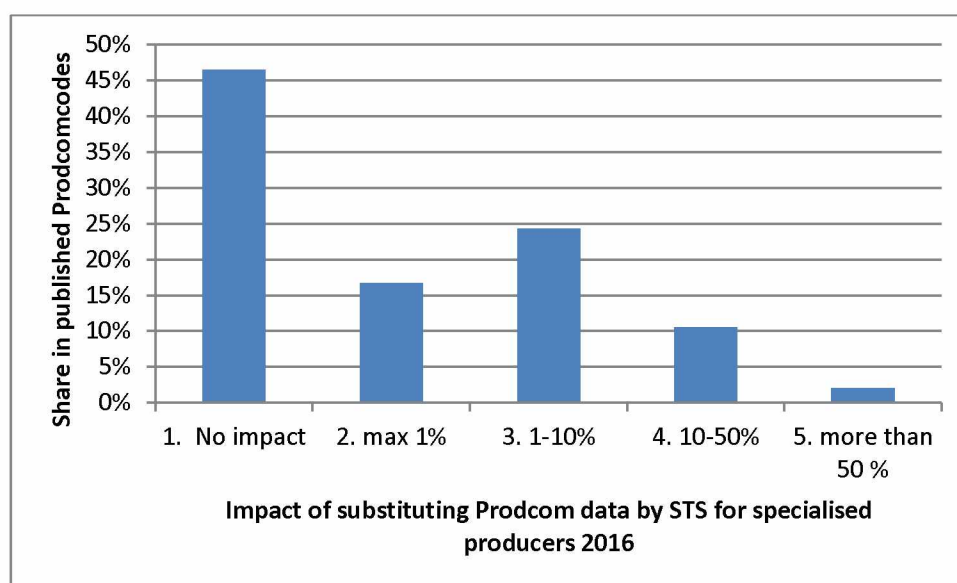
Although the overall misestimating is minimal, the effect per product in the Prodcom list may be significant. The Prodcom product list is very detailed and contains about 3,900 products². Not all products are produced in the Netherlands. Of the products that are manufactured nationally many are not published due to confidentiality. If a product is produced by very few enterprises or if production is dominated by a single large enterprise Statistics will not disclose the data unless the enterprises involved provide permission to do so. For the most detailed classes (8 digits) in the products list's hierarchy this is the case for nearly half of the classes. This complicates the analysis of the results, the most significant changes will take place when a particular product is produced only by specialised manufacturers. Many of these outcomes are confidential as these are based on less than 3 producers. Although the effect on a product level is significant, this effect may not show in the published figures.

For some products the difference between the figures based on the Prodcom data and the STS method for specialised producers is not available. These are products that are produced if the Prodcom data is used, but not when the STS method is applied or vice versa. This issue may arise when innovation results in the production of new products or when other (old) products become obsolete for example due to e.g. change in preferences, availability or prices of resources or new regulations. It is important not to have sold production of products that have become obsolete, e.g. cassette tapes or floppy disks, in the Prodcom data when the product codes from a previous year are used. In this simulation two product (each produced by more than three enterprises) classes in 2016 when the STS data is used, but have a zero production in the Prodcom table on the Statistics Netherlands website (Statline). All products that are manufactured when the Prodcom survey data is used, but not when STS data is used do not meet the confidentiality criteria for publication.

Figure 2.1.3 provides some insight in the impact on a product level of using 2016 STS data as an alternative for Prodcom survey data for specialised manufacturers, that reported one and the same product class in 2014 and 2015. The results are based on products that are produced in both cases and that meet the criteria for disclosure in 2016. For over 45 per cent of the Prodcom codes no impact of using an alternative data source was found. In 2 per cent of the cases, the absolute difference between the results from the current methodology and the alternative (STS data) was more than 50%.

² <http://ec.europa.eu/eurostat/web/Prodcom>

Figure 2.1.3 Classes of the absolute effect of substituting Prodcom survey data by STS data for specialised producers and the share in the published Prodcom codes.



2.1.6 Need for check on specialised manufacturers

From previous findings it is clear that introducing administrative data as a source from Prodcom Statistics requires an observation strategy. For producers of one and the same good, the STS approach must be checked with a control survey. The best way may be to include the Prodcom product code and description on the STS survey either every month or e.g. in the 9th month. If this option is not feasible, due to e.g. technical limitation of the IT tooling for data collection or an expected decrease in the continuity of the STS figures, the Prodcom questionnaire will be reintroduced for the specialised businesses after n (e.g. 3 to 5) years. This survey should be used to check once every n years if the Prodcom product code based on historical data is still accurate. This will however result in complicated methodological issues as the change of specialised production of good A may have changed to the production of good B or multiple products anytime in the past n years.

2.1.7 Implementation plan

The analysis in this section shows promising results. It is advised to run the alternative method based on STS data for specialised producers parallel to the operational method for several years in order to get better insights in the effects. To improve the results the alternative potential refinements to be researched aim at identifying the most important specialised producers and the ones most likely to change to multiple or another product. These actions require a fairly small amount of effort (about 160 hours) and may be taken by the Prodcom team:

- ☐ The introduction of a threshold of the value of sold production or the share of sold production by product to which specialised manufacturers contribute. The largest specialised manufacturers should possibly remain in the annual survey.
- ☐ The only other data sources that match the level of detail in the Prodcom list of products is the international trade in goods. Changes in a manufacturer's import or export portfolio may

be helpful in detecting changes in the production process and output of specialised manufacturers.

- ☐ The effect of additional criteria, for example 'companies will only enter if the turnover data (level or development) deviated within 10% of the reported production value for the previous two years'.
- ☐ Additional measures like outlier detection and filtering on the STS data.
- ☐ Put extra effort in increasing the consistency the Prodcom codes supplied by the specialised manufacturer, its NACE class in the Business Register and the products reported in the export data before cancelling the annual Prodcom questionnaire for specialised enterprises.

Other actions may be more time consuming and require consultation of the methodology department in Statistical Netherlands:

- ☐ The optimal length of the required period of consecutive production of one and the same product.
- ☐ Should the estimate of the sold production value of the specialised enterprises that are no longer included in the annual Prodcom survey be based on the value of the STS turnover from industrial activities or on the change of STS turnover and the last recorded Prodcom figure.
- ☐ The length of the period companies will be in alternative data sources before a survey will be sent to check the product portfolio and what should be done when a change is detected after several years.
- ☐ Whether or not all companies that will meet the requirements, will immediately be added to alternative data sources. If so, a cohort effect will appear for the introduction year. An alternative would be to set up a controlled inflow of companies over a few years.
- ☐ The effects on the quality of the data of the volume of sold production for the specialised manufacturers for whom the survey is abandoned.
- ☐ The feasibility and effects of the introduction of check survey that samples the specialised enterprises once every n (e.g. 3 to 5) years.

2.2 Optimal Prodcom sampling

This section starts with a description of the current sampling strategy for Prodcom and the relevant quality criteria. Followed by a brief study of literature of the relevant sampling strategy is provided. The section ends with an assessment of the impact on the coverage of the sold production from reducing the sample size.

2.2.1 Current cut-off sample for Prodcom statistics

The statistical population consists of all enterprises whose principal activity, i.e. the activity with the highest value added within the enterprise, according to business register of Statistics Netherlands is listed in section B (Mining and Quarrying) or C (Manufacturing) of the classification of economic activities in the European Community (NACE). As stated before Prodcom statistics are based on a sample. This survey includes all enterprises that employ at least 20 employed persons³. The survey is conducted every year, and data for reference year (t) is transmitted to Eurostat at t + six months. The statistical variables are:

- The physical volume of production sold during the survey period
- The value of production sold during the survey period
- For some products, the volume of total production during the survey period

Production is measured by value (in Euro) and by the volume that is appropriate for the product (kg, m², number of items, etc.).

In undertaking the Prodcom survey, there are three conditions to be met (quality criteria):

- *in each reporting country at least 90% of production in each (four digit) class of NACE must be recorded*
- *any enterprise of 20 or more employees in each NACE class should be covered.*
- *if a Member State's production in each NACE class represents less than 1% of the Community total, then data for the headings in that class does not need to be collected. In this case production should be reported as zero (Prodcom User Guide)*

NSIs are supposed to cover at least 90% of their national production. However, since no framework can provide perfect information on products and their producers, NSIs are not able to determine whether they comply with the coverage criteria, i.e. it is impossible for NSIs to know if they have achieved 100% coverage of the product, or even to know what percentage of coverage they have achieved.

Before data is transmitted to Eurostat, NSIs apply a number of micro and macro plausibility checks at different aggregation levels. Eurostat applies its own data checks and any anomalies identified are reported to the declarant country for explanation or correction.

³ The 20+ criteria must be evaluated in terms of representativeness. If no sufficient degree of representativeness can be obtained, the threshold must be reconsidered..

The Prodcom coverage criteria are based on the economic activities of enterprises according to the NACE⁴ classification. An economic activity is characterised by an input of resources, a production process and an output of products, i.e. goods or services (Eurostat, 2008). The statistical Prodcom population consists of all enterprises whose principal activity⁵ is listed in section B (Mining and Quarrying) or C (Manufacturing) of the NACE classification. The correspondence of NACE classes and Prodcom products is limited to the four digit level. Additional details in the Prodcom (six or more digits) list are not matched by the NACE classification.

The current Prodcom Regulation's quality criteria are based on the NACE classes. Although the Prodcom products, in Prodcom questionnaires are far more detailed. The Prodcom list of products is hierarchical and the six digits classes may be aggregated to four digits. Figure X shows an example. The class '10.51 Operation of dairies and cheese making' is the same for both NACE and Prodcom

Figure 2.2.1 Hierarchy in Prodcom product list⁶

10.51	Operation of dairies and cheese making	
10.51.51	Milk and cream, concentrated or containing added sugar or other sweetening matter, other than in solid forms	
10.51.51.04	Condensed or evaporated milk, unsweetened	Detail
10.51.51.08	Condensed or evaporated milk, sweetened	Detail

2.2.2 Replacement of quality criteria

Data for the Prodcom survey is collected by the NSIs usually by means of a paper or electronic questionnaire. Enterprises must match each of their products produced to one of the products on the Prodcom list, which now consists of around 3,800 products. It is clear that conducting the Prodcom survey entails an administrative burden on enterprises. It is of great concern to the Commission that the administrative burden on enterprises should be minimised. In order to reduce the burden on small enterprises, NSIs are presently not required to survey enterprises with less than 20 employees. However, there is a continuous focus on further reducing reporting requirements on Prodcom statistics. The High Level Group (HLG) of the Commission proposed recommendations in order to further reduce the burden on enterprises. This grant enables Statistics Netherlands to research the proposed simplification measures defined by the HLG. A new sample design for compiling Prodcom statistics is considered. The objective is to obtain a sufficient degree of representativeness at CPA class level, making best use of the following proposed simplifications :

- Replacement of the 90% coverage rule with a generic quality requirement
- Replacement of the 20+ employees' criterion with a generic rule that encourages NSIs to reduce the burden on small and medium sized enterprises

The optimal, new design should minimize the number of units in the annual sample, in order to reduce the administrative burden on enterprises. This research is concerned with construction of this new sample design for compiling Prodcom statistics. The research is focussed on the value of sold production rather than the volume of sold production. Like the section 2.1 on the use of alternative data, the STS turnover from manufacturing is used as a reference for sold production.

⁴ NACE is derived from the French title "Nomenclature générale des Activités économiques dans les Communautés Européennes"

⁵ All enterprises in the Netherlands are registered in the business register with a NACE code for the main activity. Additional NACE classes for other, secondary activities may be available as well.

⁶ Source: <http://ec.europa.eu/eurostat/ramon/nomenclatures>

2.2.3 Alternative Sampling strategies for Prodcom

Prodcom statistics are presently based on a non-probabilistic cut-off sample, determined by a threshold on the number of employees. Cut-off sampling is appropriate for estimating a total from a finite and highly skewed population, in which the variable of interest is covered by a small number of observations (Knaub, 2008). Furthermore cut-off sampling is considered as a low cost data collection method for official statistics, in which strongly correlated regressor data is present, and where much attention must be paid to non-sampling error and total survey error (Knaub, 2016). The regressor is the variable that determines the cut-off threshold, indicating which observation units should be included in the sample and surveyed. Some of these criteria also hold for Prodcom statistics, however, estimation of the total (of e.g. sold production by product) is not a part of Prodcom. For Prodcom there are no provisions for non-sampling error, since the total is not estimated. Non-response error is taken into account by imputation, however a validation model for this estimation method is not available.

Cut-off sampling can also be represented by a two strata model (Hidiroglou, 1986), where observations in the first stratum are included in the sample (take all), whereas the second stratum units are sampled (take some). The most general approach is the three stratum model which is generalized by (Benedetti et al, 2010). In this context the population is stratified in three groups, respectively take all, take some and take non. Observation units in the take all stratum are all included in the sample and surveyed, units in take some are included in the sample by probabilistic sampling, while units in the take non stratum are not surveyed.

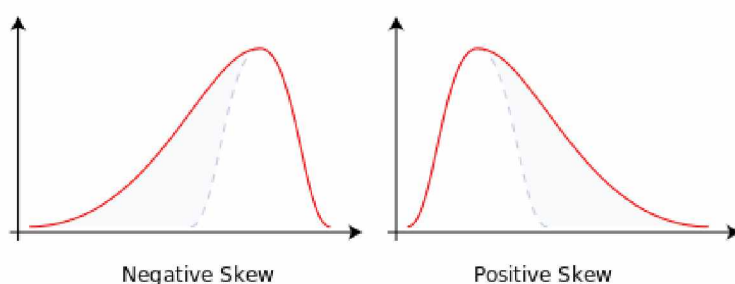
Finally, Elisson and Elvers (2001) compared cut-off sampling with simple stratified sampling. They concluded that cut-off sampling deserves more consideration, and also found that the variable that determines the cut-off threshold (regressor) has impact on the results.

From a literature point of view it is clear that cut-off sampling is appropriate for compiling official statistics. As long as strongly correlated regressor data is available and the population is both finite and skewed, cut-off sampling can be cost effective while providing good results.

2.2.3.1 Skewness

Cut-off sampling can be used for estimating a total from a highly skewed population. A skewed distribution can be considered the counterpart of a normal distribution. Contrary to a normal distribution mode, median and mean are not the same value in the centre of the distribution. A skewed distribution may have either a positive skew or a negative skew.

Figure 2.2.1 Skewed distribution negative versus positive skew. (source: wikipedia)



To examine whether cut-off sampling is appropriate for Prodcom, data from the 2013-2016 samples (only enterprises with 20 or more employees) were analysed for their distribution. The skewness⁷ for all NACE classes⁸, based on the value of sold production, is summarized in the table 2.2.1.

Table 2.2.1: Summary of skewness analysis of sold production in Prodcom deducted NACE classes

	2013	2014	2015	2016
Negative skewed NACE classes	9	4	5	6
Positive skewed NACE classes	190	187	188	189
Total number of NACE classes	199	191	193	195

The majority of NACE classes has a distribution in which a few large observations occur, and more and more smaller ones. Only a small number of classes have a distribution with a few small observations, and many large ones. Furthermore it is worth mentioning that there are no classes in which a (nearly) symmetric distribution is present. Since the 2013-2016 samples consisted of all⁹ relevant observation units, one may conclude that skewness is present in the Prodcom population. For this reason cut-off sampling is an appropriate and cost effective method (Knaub, 2008) for compiling Prodcom statistics.

The alternative of random sampling for Prodcom may not be preferred, as it may lead to unwanted effects for small samples from skewed populations (Knaub, 2008).

2.2.3.2 Turnover from industrial activities as a regressor

The number of employees served as a regressor (cut-off variable) since the first Prodcom survey. The total number of observations for the Prodcom statistics was about 4.5 thousand in the year 2016. The cut off on an input indicator such as the number of persons employed, may impose an administrative burden on enterprises that have a relative small contribution to sold production. In order to reduce the burden on enterprises, the sample size should be reduced and focused on enterprises with large contributions to Prodcom's sold production (volume ore value), while still providing sufficient representativeness. Theoretically this is possible when optimum use is made of the skewness of the distribution.

The next sections provide results of using turnover from manufacturing from the Short Term business Statistics (STS). This is considered an output variable with a high correlation to the Prodcom value of sold production (see 2.1.3) .

The STS sample for turnover data in the NACE sections B Mining and Quarrying and C Manufacturing is the same as the Prodcom in the present situation. This means that turnover from manufacturing is available, in theory, for enterprises with 20 or more employed persons. For further analysis we define this variable as the 'STS production variable', which is not a publication variable but a survey variable.

⁷ Skewness is defined as the third sample moment in statistics. A positive skewness indicates that observations are concentrated to the left of the mean value, while for negative skewness values the opposite is true.

⁸ The 4 digit NACE is based on the product codes in the Prodcom data rather than the

⁹ It is assumed that non-response error has been taken into account, and the imputation method returned representative results.

Theoretically, for an enterprise, that is not a subcontractor (see chapter 3) nor producing abroad, the sum of the sold production of Prodcom product codes should equal the STS production variable (Table 2.2.2) .

Table 2.2.2 Production of 223 products on 4 digit level.

4 digit Prodcom or NACE class	05.10	06.20	.	.	33.17	33.20
Value Sold production enterprise A	x1	x2	.	.	x221	x223

Where

$$x1, x2, \dots, x223 \geq 0$$

And

$$\text{4 digit Prodcom or NACE sum} = \sum (x1, x2, \dots, x223) = \text{STS production variable}$$

In the equation above it is assumed that a single enterprise can distribute its STS production turnover among all NACE classes. While this is theoretically true, in practice the majority of the Prodcom enterprises are contributing to a 4 digit NACE class of its main activity, making the enterprise a ‘single activity enterprise’ for Prodcom. This is a different level of specialisation than the eight digit product codes used to identify the specialised producers in section 2.1.

All enterprises in NACE sections B and C in the Netherlands are registered in the business register of Statistics Netherlands with one, or more, NACE codes. The first code belongs to their main activity, the next one(s) to the secondary activity (or activities). The STS production variable combined with the enterprise’s main¹⁰ activity can give a good prognosis in advance of what the distribution of sold production a four digit level of the product list level will be. For this it is necessary that the enterprise’s activity is correctly observed and stable from year to year as changes in output will result in outdated NACE classes in the business register. Observation of the enterprise’s activity would not be a problem since the registers are available on demand. However, the stability, or to be more precise: the changing behaviour of the enterprise’s activity, needed to be analysed. The results of this analysis for the 2014- 2015 period are summarized in tables 2.2.3. Most enterprises produce goods that are classified under one Prodcom product class (4 digit level) and these producers display a high level stability as only 3,8 per cent changed to another or multiple 4 digit product codes. Other periods tested (2013-2014 and 2015-2016) show similar outcomes.

Table 2.2.3: Output changing behaviour 2014-2015

	Number of 4 digit Prodcom codes					
2014-2015	1	2	3	4	5	6 or more
Total number enterprises	2796	820	246	90	35	27
Percentage of enterprises product portfolio changed	3,8%	14,9%	20,7%	34,4%	57,1%	74,1%

¹⁰ It was known upfront that this only holds for single activity enterprises.

Without proper measures the STS production variable will not serve as a good regressor, i.e. the cut off variable for the cut off sample, for enterprises with a more diversified output portfolio. The distribution of the STS production variable among the product in their output portfolio is not known, and the combination of their products are subject to change in subsequent years. For this reason these enterprises, on average 30 per cent of the total number of enterprises between 2013 and 2015, are excluded from further analysis.

For enterprises specialised in products in one class of the Prodcom list of products (4 digit) the STS production variable can serve as a regressor, by taking adequate measures for the small error that it will produce in the prognosis for example in the case of subcontracting or production abroad. Production abroad is out of scope for Prodcom, but not for STS turnover. In case of subcontracting the valuation of sold production may differ between STS and Prodcom, see chapter 3.

2.2.4 Results of an additional cut off based on the STS production variable

To examine what impact this new regressor can have on the sample size and final results, a simple (fictitious) top-down approach was applied to the 2013-2016 samples. The enterprises in the top P % of the STS value of turnover from manufacturing of each NACE class were sampled on the basis of the prognosis by this new regressor. Finally the Prodcom survey returns were recorded in order to analyse the influence on the final results. The first analysis is based on perfect information (PI), which assumes:

- the NACE class is correctly observed
- no error in the STS production variable

Table 2.2.4 Reduction in sample size of manufacturers with products in a single 4 digit class of the Prodcom list

	<i>Top-down percentage STS production variable</i>			
Reference year	100%	99%	95%	90%
2013	3224	2957	2383	1942
2014	3176	2910	2352	1927
2015	3127	2862	2315	1888
2016	3173	2905	2364	1934

The results in table 2.2.4 show the potential reduction in sample size by cutting off on the basis of STS production variable. The starting point is the sample according to the current practice. So the cut off based on the STS production variable is an addition to cut off of 20 or more employed persons. Although this approach is suboptimal in theory, in practice it is the only option as the STS offers no data on the turnover from industrial activities of business with less than 20 employed persons. A top down percentage of 99 per cent results in a reduction of the sample size of 8 per cent on average between 2013 and 2016. These reduction increase to 39 per cent for 90 per cent criterion.

A situation in which perfect information will occur is not likely in this context. For that reason some statistical error must be allowed in the STS production variable. In the second analysis a statistical error of at most k% is allowed, this error is the discrepancy between the STS production variable and

the value of sold production in Prodcom.. The set of errors follow an approximate normal distribution. For comparison the perfect information (PI) sample sizes are included (table 14-17).

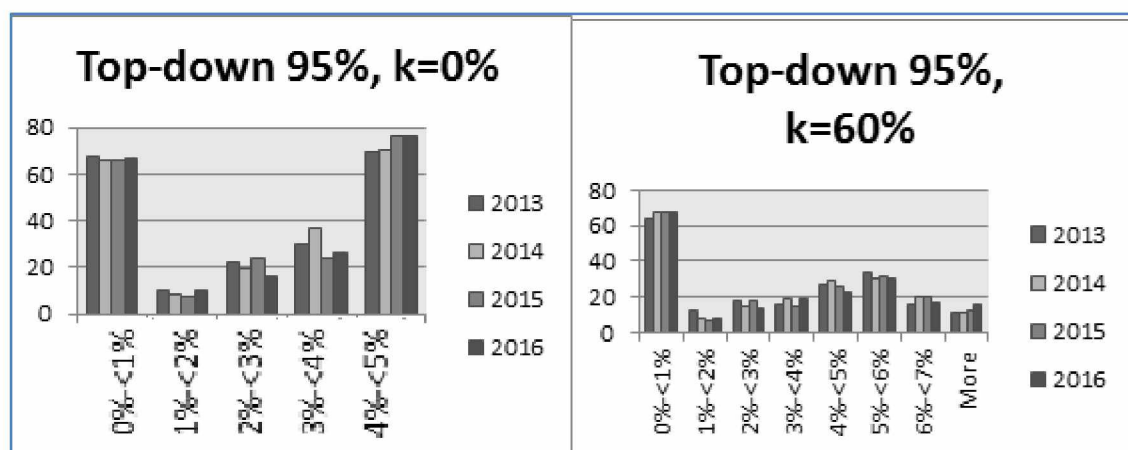
Table 2.2.5: Sample size perfect information (PI) versus a statistical error of k %

		<i>Top-down percentage STS production variable</i>		
		99%	95%	90%
2013	<i>PI</i>	2957	2383	1942
	<i>k = 60%</i>	2939	2348	1913
	<i>k = 90%</i>	2924	2340	1901
2014	<i>PI</i>	2910	2352	1927
	<i>k = 60%</i>	2893	2328	1903
	<i>k = 90%</i>	2871	2296	1878
2015	<i>PI</i>	2862	2315	1888
	<i>k = 60%</i>	2849	2284	1863
	<i>k = 90%</i>	2840	2256	1833
2016	<i>PI</i>	2905	2364	1934
	<i>k = 60%</i>	2894	2340	1900
	<i>k = 90%</i>	2877	2311	1876

Error in the STS production variable, seems to have no significant effect on the sample size (almost the same as the PI sample sizes). This result was expected due to the skewness of the population.

To examine how this change in sample size influenced the final results, the new sample quantities (at NACE class level) were compared with the original quantities. The resulting error at NACE class level for a top down percentage of 95% and error of both $k=0\%$ and $k=60\%$, is summarized in figure 2.2.2 (equivalent histograms for the remaining combinations are included in annex). The results in these figures can be seen as the loss in precision which is defined as : $100 * [(\text{Top-Down total}) / (\text{Original total})]$. A positive error indicates that the new sample observation is smaller than the original one. A negative error is not possible since the denominator is the largest quantity possible.

Figure 2.2.2 The number of 4 digit Product classes (y-axis) by classes of loss of precision (x –axis)



Of the 195 NACE classes in 2016, 68 had a loss of precision of less than 1 per cent in the case of $k=60\%$. In 8 per cent of the NACE classes this loss would have been more than 7 per cent. Comparing the results for perfect information (i.e. $k=0\%$) with the results allowing an statistical error of 60 ($k=60\%$) per cent shows that the class of the largest loss of precision (4-5%) is influenced the most by the introduction of the error. This class spreads out over a wider range and to higher maximum in loss of precision. Likely this has to do with the distribution per NACE of the STS production value in the bottom 5 per cent (of the top down value of STS production).

However, the majority of the estimators are still within the top-down error percentage, making the STS production variable a robust regressor (considering the k values of 60%). It is obvious that error in the STS variable should always be kept to a minimum when this variable is used as a regressor for the Prodcum sample.

2.2.5 The impact for other statistics making use of Prodcum

The other departments and subjects that depend on the Prodcum data in most cases do not use data on an enterprise level. Although micro data linking is common in within Netherlands, for example in the analysis presented in this report, no structural uses of Prodcum data on an enterprise level were identified. Prodcum data is used by the National Accounts department, critical in this the quality of the results on a product by activity (NACE) level. The Producer Price Index (PPI) measures the average price development of the goods and related services that are produced by manufacturers. For weighting the individual goods to aggregates the Prodcum figures are used. The impact of changing the Prodcum sample on the PPI methodology requires further research as the PPI depends on the Prodcum data for both weighting and PPI sampling. For PPI sampling micro linking of Prodcum, international trade in goods and STS data is data per enterprise is relevant.

2.2.6 Implementation plan

In our current operations the Prodcum sample includes businesses that have a small contribution to the value of sold production in a 4 digit subgroup (NACE). To optimise the sample size per subgroup the plan is to test three stratum model. Contrary to the current practice, methodology for incrementing the sample data to an estimation for the total (value of sold) production may be introduced. Box 1 presents the methodology of such a model. Testing such a model may take about 160 hours for an experienced methodologist.

Box 1: methodology cut-off sampling and a three stratum model

Skewness in the population

Distribution of the totals within subgroups (e.g. 4 digit NACE classes) of the population shows some businesses with much higher values. In other words, the totals within the subgroups are frequently positively skewed. In order to tackle the presence of skewness in the data, we will apply the approach of cut-off sampling to divide the population into some homogenous parts. In the results in section 2.2.6. the subgroups used are the 4 digit NACE classes. Part of the research for this model is defining suitable subgroups. The number of enterprises in the 4 digits subgroups may be rather low.

Before applying the procedure of cut-off sampling we first partition our population U into three strata:

- A take-all stratum U_C whose units (businesses) are surveyed entirely
- A take-some stratum U_S from which a random sample can be drawn
- A take-nothing stratum U_E whose elements are excluded(discarded).

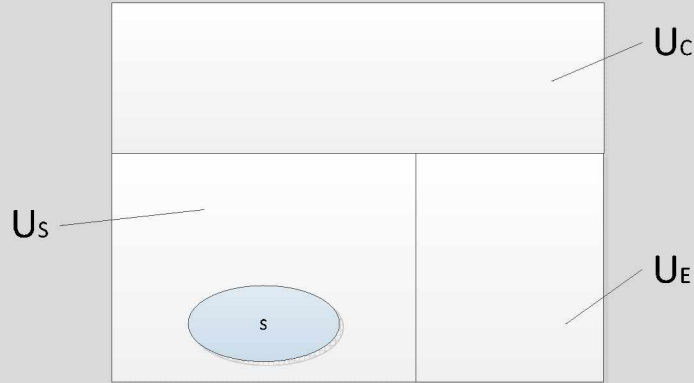


Figure B1: Partitioning of the population into three homogenous parts.

The procedure of partitioning the population consists of first determining units that must be surveyed entirely. These are usually businesses with a value of turn over or own production beyond a certain threshold. So whenever this threshold is known stratum U_C is defined. If no foreknowledge is available on a suitable threshold, the technique of *Probability Proportional to Size* (PPS) (Cochran, 1977 and Brewer et al.) can be applied. This technique is also being used within Statistics Netherlands in the process of determining the sample design for the producer price index.

Stratum U_E includes businesses whose contribution in the overall total is negligible. Statistics Netherlands has in an earlier research concluded that these are businesses with a number of employees below five.

Finally, the stratum U_S can now be defined as the complementary set of U_C and U_E within the whole population.

Estimation of totals under cut-off sampling

An estimator $\hat{t}_y$ of the overall total $t_y = \sum_{k \in U} y_k$ can be thus obtained by summing up estimators of the overall totals $\hat{t}_C$, $\hat{t}_S$ and $\hat{t}_E$ of respectively the overall totals t_C , t_S and t_E of respectively partitions U_C , U_S and U_E :

$$\hat{t}_y = \hat{t}_C + \hat{t}_S + \hat{t}_E \quad (1)$$

Although units of partition U_E have a small contribution in the overall total t_y some bias correction must be conducted to finally get an unbiased estimator $\hat{t}_y$. It can be assumed¹¹ that t_E is a constant

¹¹ This constant can be derived just before an entire survey is replaced by sampling.

proportion of the overall total t_y of the whole population. This means that there is some constant δ such that:

$$\begin{aligned}\hat{t}_y &= \hat{t}_C + \hat{t}_S + \hat{t}_E = \hat{t}_C + \hat{t}_S + \delta(\hat{t}_C + \hat{t}_S) \\ &= (1 + \delta)(\hat{t}_C + \hat{t}_S)\end{aligned}\quad (2)$$

The take-all stratum is entirely surveyed thus $\hat{t}_C = t_C = \sum_{k \in C} y_k$. The estimator $\hat{t}_S$ can be obtained by drawing a random sample s (Figure 1) from stratum S . A commonly used method is the Horvitz-Thompson estimator. This is obtained by weighting observations of the sample s by the inverse d_k of the inclusion probability π_k of units k of s :

$$\hat{t}_S = \sum_{k \in s} \frac{1}{\pi_k} y_k = \sum_{k \in s} d_k y_k$$

Because all units of the entire population are surveyed in the current Prodcum frame, it is possible to determine the inclusion probabilities π_k and consequently equation (2) above can be evaluated.

Optimal sample size under cut-off sampling

After having defined estimators of the overall total on the partitioned population, equation (2) above reveals that accuracy of the total estimator relies on the chosen sample. We utilize this dependency to choose a good sample size n_s that maximizes the accuracy of estimators of the totals.

An optimal sample size can be achieved for instance by the *Neyman Allocation* which minimizes the variance of the estimator of the overall total. Suppose that the hereinabove mentioned strata U_S , U_C and U_E have respectively sizes N_S , N_C and N_E which add up to size N of the entire population U .

Although strata U_C and U_E are predetermined before proceeding with stratum U_S , we assume stochasticity (probabilistic variation) within these strata. The *Neyman Allocation* optimal sample size is:

$$n_s = \left(\frac{N_S S_{U_S}}{N_C S_{U_C} + N_S S_{U_S} + N_E S_{U_E}} \right) N \quad (3)$$

The quantities S_{U_C} , S_{U_S} and S_{U_E} are the sample deviations of respectively U_C , U_S and U_E .

Formula (3) can now serve as a threshold for a cut-off sampling.

Other methodological issues that require additional research are:

- The relevance of enterprises of less than 20 persons employed per subgroup. Presently excluded from Prodcum and STS survey.

- Manufacturing enterprises with a diversified product portfolio that covers multiple 4 digit Prodcom product codes are excluded from the cut-off in this research. Is this a necessity, also if the value of sold production is on the lower end? Does this mean that enterprises that enter the population should receive the questionnaire in order to determine the diversification of its product portfolio.

The quality of the NACE classes (profiling) of enterprises in the Business register of Statistics Netherlands requires improvement especially for smaller businesses. The consistency between the NACE class in the business register and the products in the Prodcom data is important for sampling.