Updating and monitoring the NACE code of single-establishment enterprises in the Swiss business register

Monika Ferster*

August 22, 2018

Abstract

The NACE code is an important variable for all units included in the Swiss business register. It describes the main economic activity of an establishment and is used in several statistical processes: sampling frames, sampling designs, domain extrapolation, etc. The NACE code of multi-establishment enterprises are controlled or updated by profiling information. For single-establishment enterprises, the control of NACE codes should be planned consistently to ensure that their NACE code in the register is updated in a systematic manner.

At the Swiss Federal Statistical Office we have developed and implemented a system to control the NACE code for these single-establishment enterprises. Every quarter a sample of establishments is selected to control the NACE code. The underlying sampling scheme is Poisson with sampling weights depending on establishments' sizes. There are two restrictions for the sampling design: expected quarterly sample size and length of the cycles in which all establishments of a size category should be controlled. For example, all the big establishments with at least one hundred employees should be controlled within two years. To handle the control process, we assign to each establishment a permanent random number uniformly distributed in [0;1]. Taking into account expected changes in the register, we split the [0;1]-interval for each size category into a number of disjoint selection intervals so that we can reach the demanded cycle length. Consequently, each quarterly sample will contain all the establishments of the random number subinterval belonging to this quarter and size category. Based on the permanent random number of an establishment, we know when the NACE code of the establishment has been controlled for the last time or when the next control will take place.

keywords: NACE code, single-establishment enterprises, permanent random number

1 Introduction

The NACE code describes the main economic activity of an enterprise. It is an important variable in a business register and should reflect the real situation of the enterprise as recently as possible. For monitoring reasons former controls or changes of the NACE code should be listed in the history of the NACE code of the examined enterprise.

*Monika Ferster, Swiss Federal Statistical Office, Statistical Methods Unit, Espace de l'Europe 10, 2010 Neuchâtel, Switzerland, email: monika.ferster@bfs.admin.ch

There are different ways, by which the NACE code is controlled. The NACE code of enterprises belonging to profiling / profiling light is updated regularly, but there is a remainder of mostly single establishment enterprises where the control must be planned. The control is based on quarterly samples in a way that for example the NACE code of the large enterprises is controlled within two years and the one of the very small ones within five years.

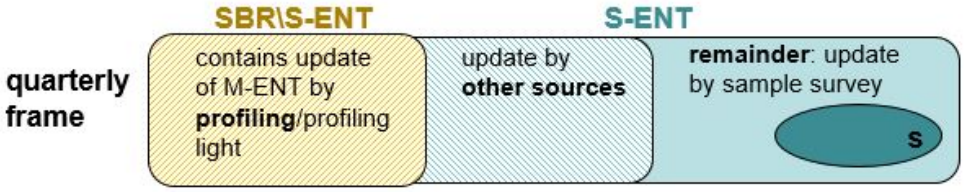
The result of a control can be that there was an error in the codification, which has to be corrected, that there was really a change in the economic activity or that the control was just a control without change. In the following we describe how the system of control.

2 Quarterly frame and NACE code control types

Each quarterly frame is built by all active Swiss enterprises in the Swiss business register (SBR). It is divided into four size classes: large (≥ 100 employees), midsize ($10-99$ employees), small ($3-9$ employees) and very small (1 or 2 employees). The first frame dates from April 2014 (second quarter) and consisted of 801,734 enterprises.

There are three types of NACE code controls or updates. The NACE code of a multi-establishment enterprise (*M-ENT*) is controlled by the so called profiling or profiling light procedure. Enterprises in the profiling are regularly updated - quarterly for profiling and yearly for profiling light enterprises. For the single-establishment enterprises (*S-ENT*) there is either a forced control of the NACE code by different sources - for example when the enterprise is recently and for the first time included in the business register - or for all other enterprises (the *remainder*) by a quarterly sample survey.

Figure 1 Types of NACE code control



Remark: We are using the words *control* or *update* of the NACE code synonymously.

3 Constraints to the sample survey

There are two competing constraints to the sample survey. On the one hand our staff responsible for the NACE code can control the NACE code of about 10,000 enterprises in a quarter. On the other hand there is a requested maximum cycle length by which the NACE code of all enterprises of a size class should ideally be updated : After eight quarterly samples, which corresponds to two years, the economic activity of all large enterprises should have been controlled. The NACE code of midsize enterprises should have undergone an update after three years, of small ones after four and of very small after five years.

It should not be forgotten that the business register is changing continuously: an enterprise can change its structure (i.e. from S-ENT to M-ENT), its number of employees and hence the size class or the way its NACE code is controlled (e.g. it was part of the profiling procedure of the past quarter and is in the *remainder* of the actual quarter).

Furthermore the number of active enterprises in the business register is not stable over time but slightly increasing.

There is a trade off between sample size and cycle lengths. It has been decided that the large, midsize and small enterprises are controlled within the demanded cycle length but the cycle length of the very small enterprises is expanded in a way that the sample size of about 10,000 is maintained.

4 Random numbers

Starting with the first frame in April 2014 and ongoing with subsequent quarterly frames each enterprise entering this system receives a permanent random number uniformly distributed in the interval $[0, 1]$. The following procedure is applied to each size class:

- The interval $[0, 1]$ is divided into disjoint subintervals whose lengths reflect sampling rates of different quarters and expected changes in the Swiss business register.
- Each quarterly sample is defined by enterprises with random numbers contained in the corresponding **selection interval**.
- The quarterly samples are therefore negatively coordinated.

So we have a Poisson sampling scheme with permanent random numbers.

4.1 Calculation of selection intervals for each size class

If we suppose that the business register was static over time then for example for the midsize enterprises with a cycle length of three years the $[0, 1]$ interval whould be divided into twelve subintervals of same length $\frac{1}{12}$. In the context of sampling this interval length of $\frac{1}{12}$ can be seen as a sampling rate. If we add up all selection intervals until they cover the whole interval $[0, 1]$, we will have reached the end of the midsize cycle because every enterprise ($=$ random number) has been selected.

However, because the size of the Swiss business register is increasing slightly, the random numbers become denser and therefore the length of the selection interval slightly shorter if we want to keep about the same number of enterprises ($=$ random numbers) in a selection interval. The assumption of a constant rate of increase leads to a geometric series, where we are interested to know, when its sum is greater than 1.

With m_d as length of the the selection interval of quarter d we have the geometric series $s_k = m_1 + m_2 + ... + m_k = m_1 \cdot \frac{1-q^k}{1-q}$ with $\frac{m_{d+1}}{m_d} = q = \frac{1}{rate\_of\_increase} \forall d \in \{1, 2, ..., k\}$ and m_1 the first selection interval length or the first sampling rate. The length of a control cycle is given by the minimum value of k for which $s_k \geq 1$ ($argmin\{k; s_k \geq 1\}$).

To be able to calculate the geometric series we need values for m_1 and q. The latter is explained before and m_1 is the sampling rate for the first sample of quarter 2014_2 (April 2014). Using

the notations of Figure 1 we could define $m_1 = \frac{n_s}{N_{remainder}}$ with n_s the desired sample size of the first sample and $N_{remainder}$ the size of the remainder control type of the first frame. To be more stable we include for each size class the average of the ratios $N_{remainder} : N_{S-ENT}$ of past five quarters, denoted by $ratio$ in our calculation. Table 1 shows for each size class the values used for the calculation of the first quarter selection intervals m_1. We fixed n_s in a way, that we obtained satisfactory cycle lengths. The length of the first selection interval is then given by $m_1 = \frac{n_s}{ratio \cdot N_{S-ENT}}$.

Table 1 Starting positions for the geometric series calculation for each size class

	very small	small	midsize	large
assumptions**				
$ratio$	0.36	0.41	0.40	0.25
$rate\_of\_increase$	1.02	1.005	1.02	1.01
given (first frame of April 2014)				
population size N_{S-ENT}	332,502	141,965	44,860	2,098
final sample size n_s	4,700	3,600	1,600	100

**means of five past quarters

Figure 2 shows the different selection intervals $m_1, m_2, ...$ for midsize enterprises added up until they pass the unity. Thus the nineth selection interval m_9 signifies that the nineth midsize sample contains all midsize enterprises with random numbers in this interval. In Table 2 the two corresponding random number interval limits are the bold numbers of line 2016_1 and 2016_2.

Figure 2 random number interval limits for the midsize enterprises

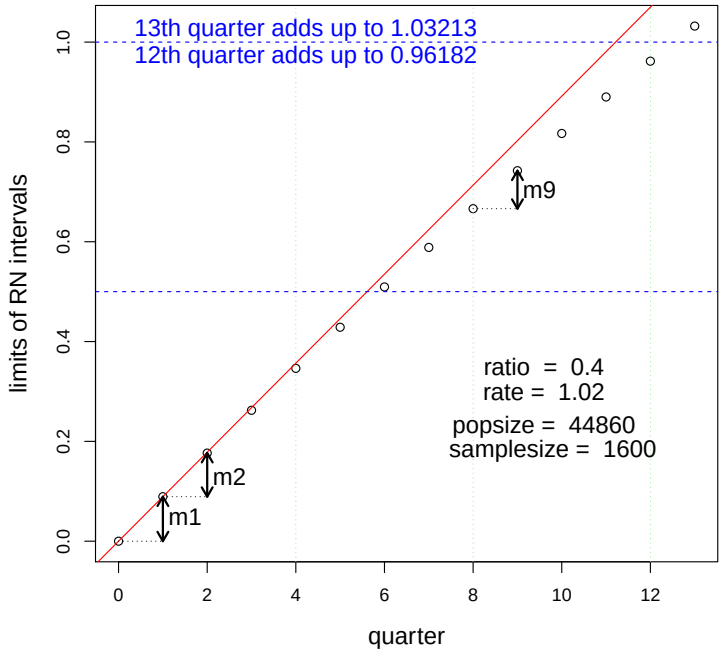


Table 2 Random number interval limits (line 2014_2 $= m_1$)

# of quarter	quarter	very small	small	midsize	large
1	2014_2	0.039264593	0.061848794	0.089166295	0.190657769
2	2014_3	0.077759291	0.123389883	0.176584232	0.379427838
3	2014_4	0.115499192	0.184624798	0.262288091	0.566328896
4	2015_1	0.152499094	0.245555061	0.346311482	0.751379448
5	2015_2	0.188773508	0.306182189	0.428687356	0.934597817
6	2015_3				**1**
6	2015_3	0.224336660	0.366507689	0.509448017	0.116002143
7	2015_4	0.259202494	0.426533062	0.588625135	0.295610386
8	2016_1	0.293384685	0.486259801	**0.666249761**	0.473440330
9	2016_2	0.326896637	0.545689393	**0.742352335**	0.649509581
10	2016_3	0.359751491	0.604823315	0.816962702	0.823835572
11	2016_4	0.391962133	0.663663038	0.890110121	0.996435564
12	2017_1				**1**
12	2017_1	0.423541194	0.722210026	0.961823276	0.167326644
13	2017_2			**1**	
13	2017_2	0.454501057	0.780465736	0.032130292	0.336525734
...					

4.2 Extract of random number table

A unity in Table 2 signifies that in the corresponding quarter the control cycle will finish and a new one will start. That means all enterprises of this size class have been updated. All midsize enterprises will be controlled e.g. in the second quarter 2017, after precisely three years. The unity of the small ones lays in 2018_2 (not seen in table 2).

The first line of table 2 corresponds to the upper limits of the random number intervals whose random numbers are included in the sample of 2014_2. The first lower limits are 0.

4.3 Result for the first frame

Table 3 shows that the sampling fraction $\frac{4750}{119786}$ of the class of the very small enterprises is much smaller than the ones in the other classes: the very small enterprises serve as buffer to fulfill cycle length requirements of large, midsize and small enterprises.

Table 3 S-ENT: Number of enterprises by size class and type of NACE code control type

	control of NACE code by			
quarter	other sources	remainder	(in sample)	$N_{remainder} : N_{S-ENT}$
very small	212,716	119,786	(4,750)	36%
small	82,766	59,199	(3,711)	42%
midsize	27,115	17,745	(1,576)	40%
large	1,690	408	(84)	19%
total	324,287	197,138	(10,121)	38%

5 Results for the quarterly frames of 2014 & 2015

Table 4 shows the sizes of the different S-ENT NACE code control types and of the sample size for each quarterly frame in 2014 and 2015.

Table 4 Number of enterprises in quarterly frames

	control of NACE code by			
quarter	other sources	remainder	(in sample)	$N_{remainder} : N_{S-ENT}$
2014_2	324,288	197,138	(10,121)	38%
2014_3	374,197	230,633	(11,511)	38%
2014_4	386,768	230,523	(10,337)	37%
2015_1	375,315	210,315	(10,079)	36%
2015_2	377,172	214,662	(10,057)	36%
2015_3	388,543	201,953	(10,049)	34%
2015_4	386,477	206,655	(10,218)	35%

6 Monitoring the NACE code control

The first frame of April 2014 is updated and completed each quarter. This leads to the monitoring dataset which allows to look at NACE code control history. The monitoring dataset reflects always the current situation.

6.1 Updating the monitoring dataset

The monitoring dataset contains 25 numeric variables. The number of observations of the first monitoring dataset was identical to the number of observations of the first frame 2014_2. Each quarter the monitoring dataset increases by about 20'000 enterprises.

Starting with the first frame 2014_2 the monitoring dataset is quarterly updated. Each of the enterprises in it has an unique ent-Id, a random number and a status-variable ($status\_rn$) with the information, since when the ent-Id is continuously in the monitoring dataset or since when it is inactive. Furthermore there are variables which describe the situation (e.g. size class and type of NACE code control) of an enterprise when it entered this monitoring dataset for the first time. The corresponding variables of an active enterprise are also kept for the current and the last-to-current quarter. Last but not least there is information if an enterprise was drawn in the current sample the last two times and when.

6.2 Keeping or creating a random number

Each quarterly frame is based on an extract of the Swiss business register. At the end of the frame preparing phase each enterprise has its own random number. By comparing the quarterly frame with the previous monitoring dataset it can be decided if

1. an ent-Id was as well in the monitoring dataset $\longrightarrow$ the random number is kept

2. an ent-Id was not in the monitoring dataset $\longrightarrow$ a new random number is created
3. an ent-Id has changed $\longrightarrow$ the random number of the former ent-Id is taken.

Case 1 is by far the most frequent case. Case 3 involves some hundreds of enterprises, who changed their ent-Id since last sampling (in general since three months). It could also be neglected and included in case 2. But doing so, we would loose the history of these ent-Ids. Thus we overwrite the former ent-Id with the new one and keep the information, that it is a new ent-Id of case 3. The variable $entid$ is always the current one.

6.3 Evaluating the monitoring dataset

Table 5 shows the entering in and leaving of the monitoring dataset, whereat "leaving" is not completely correct. An ent-Id no more appearing in quarterly frames is kept in the monitoring dataset with a negative $status\_rn$.

Table 5 monitoring dataset: entry-status

	status_rn	Frequency	Percent	CumFreq
no more active since	-2018_1	18463	1.66	18463
no more active since	-2017_4	17357	1.56	35820
no more active since	-2017_3	6685	0.6	42505
no more active since	-2017_2	32638	2.94	75143
no more active since	-2017_1	27560	2.48	102703
no more active since	-2016_4	9660	0.87	112363
no more active since	-2016_3	19468	1.75	131831
no more active since	-2016_2	16525	1.49	148356
no more active since	-2015_4	9367	0.84	157723
no more active since	-2015_3	20713	1.87	178436
no more active since	-2015_2	10461	0.94	188897
no more active since	-2015_1	9435	0.85	198332
no more active since	-2014_4	11961	1.08	210293
no more active since	-2014_3	15090	1.36	225383
active since beginning	2014_2	626949	56.5	852332
constantly active since	2014_3	12791	1.15	865123
constantly active since	2014_4	12833	1.16	877956
constantly active since	2015_1	18053	1.63	896009
constantly active since	2015_2	15210	1.37	911219
constantly active since	2015_3	14244	1.28	925463
constantly active since	2015_4	15372	1.39	940835
constantly active since	2016_2**	24147	2.18	964982
constantly active since	2016_3	30181	2.72	995163
constantly active since	2016_4	13991	1.26	1009154
constantly active since	2017_1	15662	1.41	1024816
constantly active since	2017_2	34618	3.12	1059434
constantly active since	2017_3	7414	0.67	1066848
constantly active since	2017_4	14597	1.32	1081445
constantly active since	2018_1	28243	2.55	1109688

**161 was omitted and 162 had double sample size

We can also identify in the monitoring dataset all the 10121 enterprises of the first sample and track, if they are still active. It appears that 1833 enterprises are no more active, 65 were only inactive during some quarters in-between and the biggest part, 8223 enterprises, are active since the beginning in 2014_2. 135 enterprises of the 10121 changed their ent-Id.

Or we want to know how many of the 232921 active remainder part enterprises in the actual quarter (2018_1) had intermediately been inactive. We see that a little bit less than 0.4% had been intermediately inactive.

Furthermor we can know for example, whether all the 408 large enterprises of the starting remainder part (in 2014_2) are already controlled. We find out that only 53 enterprises out of these 408 had never been controlled. But only 7 of these 53 are still in the remainder control type now (2018_1) and the quarter before (2017_4). And only one is still a large enterprise. This one should be examined because corresponding to the business register it was last controlled in September 2014.

7 Adaptations

A system change in the Swiss business register in 2015 hugely increased the number of very small enterprises so that the very small enterprises with full-time equivalent < 0.5 were to be neglected from 2015 on. Without this cut-off the cycle length of the very small enterprises would become too long.

In 2016 the set of enterprises for which NACE code update is obtained by other sources has been extended. This didn't change the whole process, only the criteria for the type of NACE code control had to be adapted.

A future extension of other-sources-controls could for example occur also by an implementation of a text mining procedure for some NACE codes, which checks if the website of the enterprise and its NACE code corresponds.

The random number table can easily be modified when changes happen in population (e.g. increasing of business register) or in resources (manner of how controls are done).

8 Conclusions

By the presented approach based on permanent random numbers we have a consistent managing system of the NACE code control for each enterprise in the Swiss business register. It is flexible to adapt to changing conditions as mentioned in 'Adaptations'. Quarterly sampling can be executed routinely.

With the monitoring dataset structural changes of the enterprises can be recognized. Past and future control of the NACE code is known.

[1] Ohlsson, E. (1995) *"Coordination of Samples Using Permanent Random Numbers"*, in Business Survey Methods (eds B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott), chapter 9, pages 153-169. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[2] SÄRNDAL, Swensson B., Wretman J.: *"Model Assisted Survey Sampling"*. Springer-Verlag New York, 1992.