



*Aan:* **EUROSTAT Task Force on "Introduction of NACE Rev. 2.0 and base year 2005"**

*Van:* **Gert Buiten, Jarl K. Kampen & Sidney Vergouw**

*Onderwerp:* **Theory on the producing of historical time series for Short-term Business Statistics in NACE Rev.2 with an application in the industrial turn-over index in the Netherlands (1995-2008)**

*Datum:* **21st May 2008**

---

*Summary: This paper provides a discussion of the basic principles of reconstructing time series in general, after which the application of these methods in the area of Short-term Statistics and Business Surveys is handled and illustrated with an example. We conclude that it is possible to obtain reasonable approximations of historic timeseries using rather simple methodology, but the quality of the backcasted timeseries is hampered by heterogeneity of classes.*

*Key words: historic timeseries, backcasting, reclassification, NACE*

## **1. Introduction**

The agencies for official statistics in the EU-member states face a major change in the classification scheme of economic activities. The transition from NACE 1.1, introduced in 1993, to NACE 2.0 which is to be operational as from 2008, forces the statistical agencies, amongst other things, to reconsider sample plans, and revise business statistics in function of the classification. This contribution discusses possible scenarios and methodologies for the national statistical agencies for backcasting the new classification scheme in existing time series of business statistics, together with an application of the most promising methodologies. We consider two general strategies, the first requiring total or partial measurement of NACE 2.0 classifications in historic samples and/or populations (micro method), the second requiring only measurements at an aggregated level, e.g., at four level of NACE 1.1 coding (macro methods); see Moauro (2005). However, the results are applicable in any situation that an existing time series had to be backcasted because of a change of classification of sampling units.

There are several possible procedures to apply a revised classification to historical time series.

They can be divided into four main methods:

1. Use of a recoding key on published series
2. Recalculating data by recoding units at the micro level (“reconstructing”)
3. Converting or reinterpolating published series using a conversion matrix (“backcasting”, “macro-approach”)



4. Combining the micro and macro approaches, e.g., by estimating benchmarks years with a micro method and interpolating with macro techniques or by backcasting transition matrices as an intermediate step

These approaches will be discussed briefly below, together with their usefulness for application on Short-term Statistics and Business Surveys. This is followed by an application in real data. Finally, some conclusions and issues for future research are proposed.

## **2. Methodologies for producing historical timeseries**

### *2.1 Use of a key*

The method based on the use of a key is the most straightforward and simple of the four methods described in this report. The technique uses a recoding “key” with which a classification at the lowest aggregation level is directly recoded to the revised classification. E.g., the old code 4.3.2.1 is recoded to 1.2.3.4 and the historical data for 4.3.2.1. are assigned to 1.2.3.4. In its purest form, this method can only be applied if there are only 1-to-1 or many-to-1 changes from the old to the new classification. The relationship between NACE Rev. 1.1 and Rev. 2 is more complicated than that, but for a large number of series this condition is met. Especially in the area of Industry, Construction and Retail Trade there are possibilities to apply this method at least partly. The key method assures a straightforward relationship between the old and the new results, since the old data are simply transferred or projected onto the new classification. Changes in the outcomes are transparent and can easily be documented and communicated with users.

### *2.2 Micro-approach*

The micro-approach means that the revised classification will be applied to the historical time series by assigning the revised classification to each statistical unit and for every period in the time series. That is, all statistical units used for calculating the old time series are coded again according to the new classification. After that, the statistical results (averages, totals etc.) are recalculated using the same calculation routines as for the old data. In fact, the entire production process is repeated starting from the micro level, but now using the new classification. Therefore this technique is also known as the micro-approach. The method is not dependent on the type of relationship between old and new codes and can also cope with for instance 1-to-many and many-to-many relations between the old and new classification. Because of the double coding of the units according to the old and new classification, there is an exact relationship be-



tween the old and new results. In practice however, differences in the outcomes may be less transparent. In cases where outlier treatment, imputation for nonresponse, etc., have an influence on the outcomes, the differences between the old and new results are not completely related to the recoding as such. In those cases, an aggregate group that has a 1-to-1 relationship between the old and new classification may show different totals or averages. Analysing, documenting and communicating changes in the outcomes between the old and new classification is in such case more complex.

Whether these problems actually arise, depends on the survey design and the type of variable. In case of a census, these problems are smallest, since it is not necessary to use grossing up procedures or outlier treatment. In the case of sample surveys and panels however, grossing up and outlier treatment usually does have an effect on the outcomes and changes are not completely attributable to recoding. In all types of survey designs, differences may occur in cases where non-response is imputed using the average of the responding units in a specific NACE group. If the calculated variable is a total (like turnover of production in the area of Short-term Statistics), problems are probably bigger than in case of e.g., averages (like confidence data or price indices). The averages are more robust, since upward and downward differences will more or less balance out.

This method requires for each unit, information about which classification it would have had in terms of the revised classification code. At the moment of the implementation of NACE Rev. 2, this information is available. As part of the implementation, the business register has to be double coded, so for every unit at that moment both codes are available. For previous periods however, units did exist that are no longer in the business population at the implementation moment. These units have to be coded also to NACE Rev. 2. If the population is very stable, this can easily be done. If however, the dynamics are large, this will be time and resource consuming. When e.g., only one percent of the companies in the panel disappear every year, this implies that after five years only five percent of the panel has got to be replaced. This means the other 95 percent of the panel only has to be recoded once (at the starting point) and just a small part of the population a couple of times.

One possible extension in this area is the development of automatic recoding procedures for units that ceased to exist before the moment of the NACE Rev.2 implementation. In the case of 1-to-many splits, the percentage shares from the conversion matrix can be translated in percentage chances with which the new NACE codes can be assigned to the units. And of course, in the case of 1-to-1 conversions, the coding process simply assigns the NACE Rev.2 code from the correspondence tables.



In this respect, there is a very important difference between a census on the one hand and samples and panels on the other hand. In a census, all units are in the sample, so only the sample has to be recoded. In case of a panel or a sample, the dynamics in the other, non-observed units in the population has an influence on the outcomes. This holds especially if the variable is a total (like turnover) that has to be grossed up. The calculation of grossing up factors requires recoding of the entire population in a specific NACE code. On the other hand, if the variable is calculated as an average (like prices or confidence data), it may not be necessary to recode the entire population as long as it may be assumed that the observed sample or panel remains representative also for the new NACE group. To sum up, the micro approach can deal with all types of relationships between old and new codes, but may produce differences in the results that are not completely attributable to the reclassification. With Business Surveys usually set up as panels used for the calculation of averages, this drawback is less important in their case than for e.g., turnover statistics. From an operational point of view, this method will be more costly and difficult for older years than for recent years.

In general, backcasting time series for the NACE Rev.2 changeover is more difficult at lower aggregation levels than for the highest aggregates. Because the classification is above all expanded, many changes consist of splits of one old NACE code into several new codes. In general, many of these splits remain within the same branch, like Industry or Services. An analysis by Eurostat for the STS Short-term statistics shows that at the highest STS aggregates the new NACE groups are almost the same as the old ones. Retail trade under the NACE Rev.2 has exactly the same content as under NACE Rev.1.1. Industry and the STS aggregate “Other services” have approximately 95% the same content, Construction for about 85%. On lower aggregation levels things are more complicated, but not always. A number of 3 and 4 digit NACE Rev.1.1 codes correspond with only one new code (1-to-1 correspondences). In this case the old data can easily be used for backcasting (using the “key method”). In a limited number of cases, several old codes correspond with one new code (many-to-1 transitions). Here also old data are easily usable for backcasting. More problematic are cases where one old group is split into several new ones (1-on-many splits) or where a number of old groups corresponds with a number of new groups (many-to-many transitions). In these cases, either micro or macro methods have to be applied for backcasting.

### *2.3 Macro approach*

Opposite to the micro approach, the macro approach works at aggregate levels. The data based on the initial classification are redistributed according to the revised classification with the help



of a set of conversion coefficients. These conversion coefficients are derived from a conversion matrix. This means only one point in time is double coded according to the old and revised classification. Therefore the main advantage of this method is that it's a relatively low resource and time consumption technique. The use of conversion matrices enables this method to cope with all types of relationships between old and new groups. That means that also 1-to-many and many-to-many relationships can be handled. In the case of 1-to-1 changes, this method will act the same as the key method. In the macro method, the relationship between the old and new results is strong. Since it re-assigns the old published data to the new classification groups, the grand totals do not change and at lower aggregation levels the differences are fully attributable to changes in the classification. Analysing, documenting and communicating differences is therefore relatively simple.

Although the terms correspondence tables and conversion tables may sound like synonyms, they actually refer to different things. A correspondence table is a sorted list of NACE-codes that for each codes shows the corresponding NACE codes according to the other version of the NACE classification. This can also be represented by a cross table that is usually called a correspondence matrix. Such a matrix contains e.g., an 'x' or a '1' in every cell that can logically contain a value according to the correspondence tables. Conversion or transition tables can be made when the new classification is actually applied to a specific statistic in a country. These tables show which part of an old group corresponds to a new group, either measured in absolute values or as percentages. A cross table version of such a table is called a conversion or transition matrix. In general, this kind of table describes the transition from one version of the classification to another one. In their most basic form, they show the distribution of the number of units between the two classifications. For analytical purposes, they are usually also made with e.g., turnover or the number of employees. In order to explain to users what actually happens with the outcomes of a given statistic, transition matrices can be calculated also for the "target variables" of a statistic. In the case of confidence indicators, they may for instance show the percentage of "positive" answers for a certain question for every cell in the matrix.

Conversion matrices are defined at the aggregated level, but they can only be calculated from the micro level. Every statistical unit has to be double coded according to both version of the classification, after which a conversion table can be calculated by aggregating the macro data. In the case that old units themselves are split or merged (e.g., when all statistical units are actually derived all over again from the basic fiscal and trade register units) the basic entity for these calculations even has to be below the statistical units. Conversion matrices can be used for backcasting purposes. In that case turnover, value added and numbers of employees are the most common variables used for the calculation of conversion factors. One should be aware of differ-



ences in structure between these variables, since the structure of the used variable determines the conversion matrix. Several cases can be distinguished:

1. The conversion matrix is directly based on the statistical outcomes of the reference period. This is e.g., the case for Structural Business Statistics, that directly estimate the level of for instance turnover or production. The totals of the conversion matrix are the same as those published for the reference period.
2. The conversion matrix is based on the same variables as the “target variables”, but for a different period than the reference period. This is e.g., the case for Short-term statistics on turnover of persons employed, with the conversion matrix made for the base year of the index series and used to convert the weighting system from NACE Rev 1.1 to Rev. 2, thus producing a new conversion matrix with the weights. This is needed to aggregate the backcasted series at the lowest level of detail to higher aggregates.
3. The conversion matrix is based on a different variable as the “target variable” and on a different period. This is e.g., the case for the confidence indicators of Business Surveys and price indices in the STS area. A conversion matrix is used to convert the weighting system from NACE Rev 1.1 to Rev. 2, thus producing a new conversion matrix with the weights. Depending on the approach used, a third kind of conversion matrix may be calculated with the “target variables” for every cell, thus showing e.g., the average price or confidence level for every cell in the conversion matrix.

In cases 2 and 3 a separate conversion matrix will have to be calculated to describe the actual relationship between NACE Rev. 1.1 and Rev. 2 for the outcomes of the target variables in the reference period.

When applying a conversion matrix from one period also for other periods however, assumptions have to be made. Basically, one has to assume that a specific aspect of the structure from the conversion matrix remains constant over time. E.g., for turnover in the area of Short-term Statistics one may assume that for the entire time series, 60% of the value of the old group 4.3.2.1 is assigned to the new group 1.2.3.4. Translated to Business Surveys, one may assume that the balance between positive and negative answers in a new group is evenly distributed over all composing old groups. Or the other way around: that the balance in an old group is evenly distributed over all composing new groups. Of course, this kind of assumptions never holds exactly and usually becomes more disputable when the length of the series increases. This issue is especially important in the case of 1-to-many and many-to-many relationships. If one old group is split up into two or more new groups, all of these new groups may get the same level or the same development over time as the old group, depending on the type of assumption.



Therefore, this problem will be strongest in the area of services and smaller in industry, construction and retail trade.

One way to deal with this problem is to make use of experts' opinion about a specific market. Based on that knowledge one could for example assume that a particular subclass is characterized by a certain exponential growth. For instance in the case of mobile telephones, one knows that they didn't exist before a certain date and had a specific growth and growth pattern after that time. In the case of e.g., turnover, there may also be alternative data sources available for this kind of estimation (like VAT-register data). Consequently, the conversion coefficients can be adjusted to that knowledge. Besides using experts' opinions, it's also common to apply more sophisticated estimation techniques. Translation of this type of action to Business Surveys is unfortunately not easy. There are no alternative sources and there is no trend in long term development, since the surveys are designed to have the balances between positive and negative answers to hover around a certain long term average. In brief: the macro method can deal with all types of relationships between the old and new NACE groups, the old grand totals remain intact and it is relatively simple and cheap to apply. It does however, require assumptions that will never be fully met.

#### *2.4 Combining micro and macro approaches*

Micro and macro approaches each have different pros and cons. One thing that they have in common however, is that the quality of their outcomes will deteriorate if you go further back in time. There are several approaches aimed at overcoming this problem by combining both approaches. The first of these methods can be described as the benchmark/interpolation method. In this method, conversion coefficients are calculated for two different points in time using a micro method, after which the coefficients for the time points between these two are derived by interpolation using a macro method. This method can deal with all types of relationships between old and new NACE groups, including 1-to-many and many-to-many relations. It decreases the necessity to rely on assumptions as well as problems of inconsistencies between old and new grand totals. It does however, require the availability of micro data for old periods, for instance for 2000 or 1995. This technique is a combination of the micro and the macro approach. According to this method, two periods have got to be double coded. These periods are called the benchmark periods. The optimal benchmark periods are to be determined by subject matter experts. First of all, the micro data for the benchmarking periods are recoded to the revised classification. After that, two sets of conversion coefficients are obtained to convert the aggregated estimates from the initial to the revised classification. For the periods between the two bench-



mark periods, the coefficients are interpolated. For some subclasses, the evolution between the two benchmark periods might not have been linear. Therefore, a non-linear interpolating method could be used. As mentioned in the macro-approach, one could in such cases make use of experts' opinions. A possible variation of the method described consists in combining the coefficients determined for the two benchmark periods into a single set and then apply these conversion coefficients to all the periods of the time series. Just like the assumption made at the macro-approach, this assumption doesn't always hold on. However, this assumption is less crude than the assumption related to the mentioned macro-approach. For the same reasons as mentioned at the macro-approach, the benchmark/interpolation method is not directly applicable to business survey indicators.

A second possible approach is possible in cases where the micro method will be applied for e.g., three years. For a monthly or quarterly statistic, this allows to calculate 36 monthly or 12 quarterly conversion matrices. Based on such a data set, it is possible to analyse the structural patterns in the conversion matrix. Are there seasonal patterns in the conversion coefficients? In the case of 1-to-many splits, are there differences in the growth trend between the new NACE groups? Obviously, in 1-to-1 conversions there would be no structural changes at all. The outcomes of these analyses can be translated in simple models, that describe the development of the conversion coefficients for each row in a conversion matrix back in time. With this as an additional intermediate step, the macro methods can be improved in order to produce historical time series with a better quality for the earlier periods where micro data are no longer available.

### **3. Backcasting business survey indicators**

With the basis methods described above in mind, time series of Business Survey data according to NACE rev. 2 will most probably be derived by using a combination of methods: the micro method for the most recent years and macro methods for the years before that. For the most recent years, institutes may find it easiest to use the double coded register at the moment of the implementation of NACE rev. 2 and apply the micro method for a limited number of years. Micro data are usually still available in a usable form and the production system for calculating weighted averages from panel data can be ran again without much extra cost. The methodological problems of representativity, loss of data because of unit dynamics, inconsistencies with the old grand totals et cetera probably have a relatively low impact. A possibly important problem may occur where an existing code is split over a large number of new codes, like in a 1-on-5 reclassification. If the units from the old NACE group are evenly distributed over the new groups, the sample fractions of course remain the same. But if this is not the case, the sample

fractions for one or more groups may become too low to achieve representative results and/or produce unstable results over time.

### 3.1 The micro approach

For the sake of notation, we abbreviate NACE Rev. 1.1 by  $O$  and NACE Rev. 2.0 by  $S$ . The computing of time series in  $S$  on the basis of micro data in the population requires the defining of so-called aggregation matrices (see e.g., Fourtier, 2005; Kampen, 2007). The aggregation matrices consist of zeros and ones and classify the  $N_t$  businesses into the proper old respectively new classification codes at time  $t$ , where  $a_t^{O(io)} = 1$  if business  $i$  belongs to class  $o$  at time  $t$ , and  $a_t^{O(io)} = 0$  otherwise; and  $a_t^{S(is)}$  defined similarly. We then have

$$f_t^{O(o)} = \sum_i a_t^{O(io)}, \quad (1)$$

$$f_t^{S(s)} = \sum_i a_t^{S(is)}, \quad (2)$$

$i = (1, \dots, N_t)$ . This principle is generalizable to the computations of totals and means of variables. E.g., the backcasting of totals of a variable  $Y$  in  $S$  can be formulated as

$$Y_t^{S(s)} = \sum_i a_t^{S(is)} y_{it}, \quad (3)$$

with  $y_{it}$  the observed value of  $Y$  for business  $i$  at time  $t$ . Dividing (3) by the corresponding frequencies from (2) produces the means within  $S$  of  $Y$ . This methodology of backcasting however, although ideal in theory, is far from unproblematic in practice. Besides the time consuming nature of the procedure, that requires measurements of the classification in  $S$  for all business, another problem arises when there is much nonresponse in  $Y$ .

In most cases, our data will only consist of a sample of  $n_t$  businesses. Our estimators of frequencies, totals and means must be adjusted according to the inclusion probabilities of each of the businesses in the sampling schemes of the  $n_t$  sample units. The inclusion probabilities of the businesses, denoted  $\pi_{it}$ , can differ substantially from those in the  $S$  sampling design, for instance, because new neyman allocations may be specified in order to decrease sampling error. But in estimating the historical quantities (e.g., numbers, totals, means) by domain estimators

(e.g. Horvitz-Thomson estimators), the original design matrix must be used. In the case of totals,

$$\hat{Y}_t^{S(s)} = \sum_i^{n_t} a_t^{S(is)} y_{it} / \pi_{it} . \quad (4)$$

The result is consistent, but not precise, because the original inclusion probabilities were not designed for making estimates in  $S$  design. Still, it will be the best estimate feasible in practice. Inferior, but financially more attractive is to compute the aggregation matrix only once, and to use the resulting  $S$  classification at time  $t$  in all further computations:

$$\hat{Y}_t^{S(s)} = \sum_i^{n_t} a_t^{S(is)} y_{it} / \pi_{it} . \quad (5)$$

Of course, essential aspects of the dynamics of the population of businesses are lost using this approximation. As a compromise, the decision can be made to construct the aggregation matrices at a limited number of points in time. That will allow the agency to detect a possible trend in the backcasted  $S$  time series.

### 3.2 The macro approach: two alternatives

For periods where micro data are not available in a usable form or where the micro approach creates large inconsistencies with the old grand totals, macro approaches may have to be applied. As mentioned before, in the case of 1-to-1 reclassifications, the macro approach acts the same as the key method and full consistency with the old series is achieved for those groups. In general, macro approaches for statistics that measure levels (like turnover or persons employed) use estimates of the proportion of the volume of transitions between  $o$  and  $s$  at all periods of the historical time series. This requires at least one point in time  $t$  where dual coding at the level of micro data is available, producing (on the basis of the correspondence tables) a  $H_o \times H_s$  transition matrix that distributes at time  $t$  the  $H_o$  classifications in old codes to the  $H_s$  classifications in the new code. The frequencies  $f_t^{S(s)}$  of businesses in  $S$  can be computed from the frequencies  $f_t^{O(o)}$  of businesses in  $O$  by the relationship

$$f_t^{S(s)} = \sum_s p_t^{OS(os)} f_t^{O(o)} , \quad (6)$$

with  $0 \leq p_t^{OS(os)} \leq 1$  denoting the proportion of businesses in old code  $o$  that transfer to the new code  $s$ . For backcasting on the level of aggregate data, we generalise formula (3.8),

$$Y_t^{S(s)} = \sum_o p_{Y(t)}^{OS(os)} Y_t^{O(o)}, \quad (7)$$

with  $p_{Y(t)}^{OS(os)}$  a set of weights that distribute the quantities of  $Y$  in  $o$  over  $s$ . In the case of 1-to-1 and many-to-1 transitions, the corresponding weights of course equal 1 and this approach is in fact the key method (Section 2.1). In other cases, these weights have to be computed in the population or estimated from a sample,

$$p_{Y(t)}^{OS(os)} = \frac{\sum_i a_t^{O(io)} a_t^{S(is)} y_{it} / \pi_{it}}{Y_t^{O(o)}}. \quad (8)$$

Please note that these weights are in fact the shares of the components of the old NACE group that go to new NACE groups (the rows of the conversion matrix), so that the total of shares equals 1 (or 100%) for the old NACE group. One can also express the shares calculated for the components in the new NACE groups that come from various old NACE groups (the columns of the conversion matrix):

$$p_{Y(t)}^{SO(so)} = \frac{\sum_i a_t^{O(io)} a_t^{S(is)} y_{it} / \pi_{it}}{Y_t^{S(s)}}. \quad (10)$$

In practice, computation of the weights will suffer from the same restrictions as backcasting on the basis of micro level data. It may then be decided to compute the transition matrix only once, and approximate the historic time series by

$$\hat{Y}_t^{S(s)} = \sum_o p_{Y(\cdot)}^{OS(os)} Y_t^{O(o)}. \quad (11)$$

At least three possibilities exist to estimate the constant transition weights  $p_{Y(\cdot)}^{OS(os)}$ :

1. As the proportion of the number businesses that transfer from  $o$  to  $s$  in a given period,
2. As the proportion of  $Y_t^{O(o)}$  that transfers to  $s$  in a given period,
3. By means of a least squares estimator.

The advantage of the latter method over the two other *ad hoc* estimators is that it is relatively easy to introduce time dependent heterogeneity in the transition weights, e.g., by letting

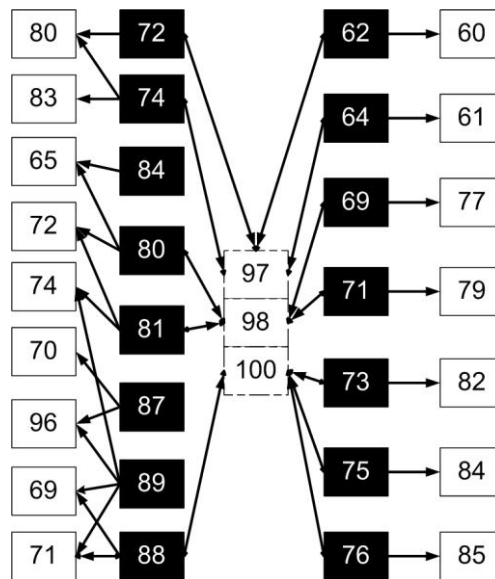
$$p_{Y(t)}^{OS(os)} = p_{Y(\cdot)}^{OS(os)} + t \cdot \theta^{OS(os)}. \quad (12)$$

A real life example of this methodology is provided in the next section.

#### 4. Estimating the evolution of Dutch business turn-over in NACE Rev. 2 (1995-2008)

##### 4.1 A description of the problem

In Industry, one of the most complex backcasting problems deals with the selection of businesses that are involved in the manufacturing, repairing, installing or maintaining of machinery, because in the new classification, and opposed to the old, the latter three activities are separated from the first. Even at two digit level, this leads to complicated transfer schemes, which in the Dutch adaptation of the publication cells to NACE Rev. 2.0, looks as follows:



In this diagram, the black boxes correspond to NACE Rev. 1.1, and the white boxes to NACE Rev. 2.0. The three boxes in the middle represent the new classes repairing, installing and maintaining. Two clusters of activities can be distinguished: one cluster, where except for the three new classes, all businesses transfer to a single new publication cell (1-to-many splits; right hand side); and one cluster, where two or more publication cells merge into several new publication cells (many-to-many; left hand side). Obviously, the backcasting of timeseries in this system may present us with several possible problems, e.g., heterogeneity.

In order to do our calculations, a database had to be prepared that besides the old timeseries of turn-over that run from January 1995 until August 2007, contained total turn-over within the new classification that run from January 2005 until April 2007. The latter series have to be constructed on the bases of microdata, and the procedure was as follows. First, for the database of companies of April 2008, for each company,

1. Assign NACE Rev. 2.0 on the basis of one-to-one transitions (i.e. use of a key);
2. Otherwise, if one-to-one transition is not applicable, assign NACE Rev. 2.0 on the basis of Prodcom;



3. Otherwise, if Prodcom supplies insufficient information, look up the company on the Internet;
4. Otherwise, apply a *best guess*.

Then sequentially, going back one month with each step,

5. Re-assign NACE Rev. 2.0 to companies with the earlier assigned NACE Rev. 2.0 if NACE Rev. 1.1 is unchanged, otherwise repeat Step 1 through 4.

The result is a database with businesses in both classifications, and estimated total turn-over in the new classification is done by applying the domain estimator of Formula (5). Although estimated, we refer to the thus obtained turn-over as the *observed* total (a number that approximates the population value to the maximal extent). This database is used to backcast a well-known short term business statistic, in this case the index of relative growth of turnover of businesses within a sector,

$$I_t^{O(o)} = I_1^{O(o)} \frac{Y_t^{O(o)}}{Y_{t-1}^{O(o)}}, \quad t > 1. \quad (13)$$

Statistics Netherlands uses the ratio of the January turnover in the baseline year and the mean monthly turnover in the baseline year  $B$  as the basic index number  $I_1^{O(o)}$ . Formula (13) then reduces to

$$I_t^{O(o)} = Y_t^{O(o)} \Big/ \frac{1}{12} \sum_{\tau \in B} Y_{\tau}^{O(o)}, \quad t \geq 1. \quad (14)$$

In the new classification, the index can be written as

$$I_t^{S(s)} = Y_t^{S(s)} \Big/ \frac{1}{12} \sum_{\tau \in B} Y_{\tau}^{S(s)}, \quad t \geq 1, \quad (15)$$

where for the period without double classifications, the historic total turn-over  $Y_t^{S(s)}$  will have to be estimated by one of the procedures proposed earlier. Of course, the change of base year using this approach is easy, and requires only replacing the denominator in (15). We discuss the performance of these procedures in our specific database in the following sections.

#### 4.2 Performance of the ad hoc estimators

As a first approach, we may apply the ad hoc estimators that were proposed in Section 3.2. This requires computation of the transition matrices of number of businesses and of turn-over. See

Table 1, which shows the number of businesses that transfer over the total period of double coding (1/2005-4/2007) from the old classification (rows) to the new one (columns):

*Table 1. Transitionmatrix of number of businesses*

	80	83	65	72	74	70	96	69	71	60	61	77	79	82	84	85	97	98	100	Σ	
72	8781	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1210	907	10898
74	62	801	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55	4	922
84	0	0	1199	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1199
80	0	0	124	489	0	0	0	0	0	0	0	0	0	0	0	0	0	0	266	88	967
81	0	0	0	740	179	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	950
87	0	0	0	0	0	410	2153	0	0	0	0	0	0	0	0	0	0	0	0	0	2563
89	0	0	0	0	29	0	245	34	144	0	0	0	0	0	0	0	0	0	0	0	452
88	0	0	0	0	0	0	0	1671	64	0	0	0	0	0	0	0	0	0	119	917	2771
62	0	0	0	0	0	0	0	0	0	670	0	0	0	0	0	0	0	223	0	59	952
64	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	114	0	36	178
69	0	0	0	0	0	0	0	0	0	0	0	1597	0	0	0	0	0	0	883	363	2843
71	0	0	0	0	0	0	0	0	0	0	0	0	242	0	0	0	0	0	28	28	298
73	0	0	0	0	0	0	0	0	0	0	0	0	0	2137	0	0	0	0	119	224	2480
75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5187	0	0	0	715	162	6064
76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	373	0	0	40	29	442
Σ	8843	801	1323	1229	208	410	2398	1705	208	670	28	1597	242	2137	5187	373	337	3435	2848	<u>33979</u>	

For example, 4 companies transfer from old classification  $o=74$  to new classification  $s=100$ , but without further information, we do not know whether these are 4 different companies, or two companies observed at two points in time, etc. The proportion of companies transferring from  $o=74$  to  $s=100$  equals  $4/922$ , equal to the transition weight in Formula (11) for this particular combination of classes. As another example, we have

$$\hat{p}_{Y(\cdot)}^{OS(72,80)} = \frac{8781}{10898} = 0,8057.$$

As a measure of adequacy of the estimator, the backcasted total turn-over using this procedure can be compared to the observed turn-over obtained by Formula (5). We find that the correlation of the backcasted series and the observed series equals  $r=.988$ . The mean relative estimation error (MRE) of the estimator, defined as

$$\frac{1}{\sum_t n_t} \sum_t \sum_s (\hat{Y}_t^{S(s)} - Y_t^{S(s)}) / \hat{Y}_t^{S(s)},$$

equals 1.26, and its box plot can be viewed in Figure 1 (RelFoutAdHoc1). Alternatively, if we

use the transitions of turn-over over the same period (figures not printed in order to save space), we obtain  $r=.998$  with  $MRE=.16$  (see also Figure 1, RelFoutAdHoc2). In other words, backcasting the series by the ad hoc estimator on the basis of turn-over is to be preferred over the estimator using the number of businesses.

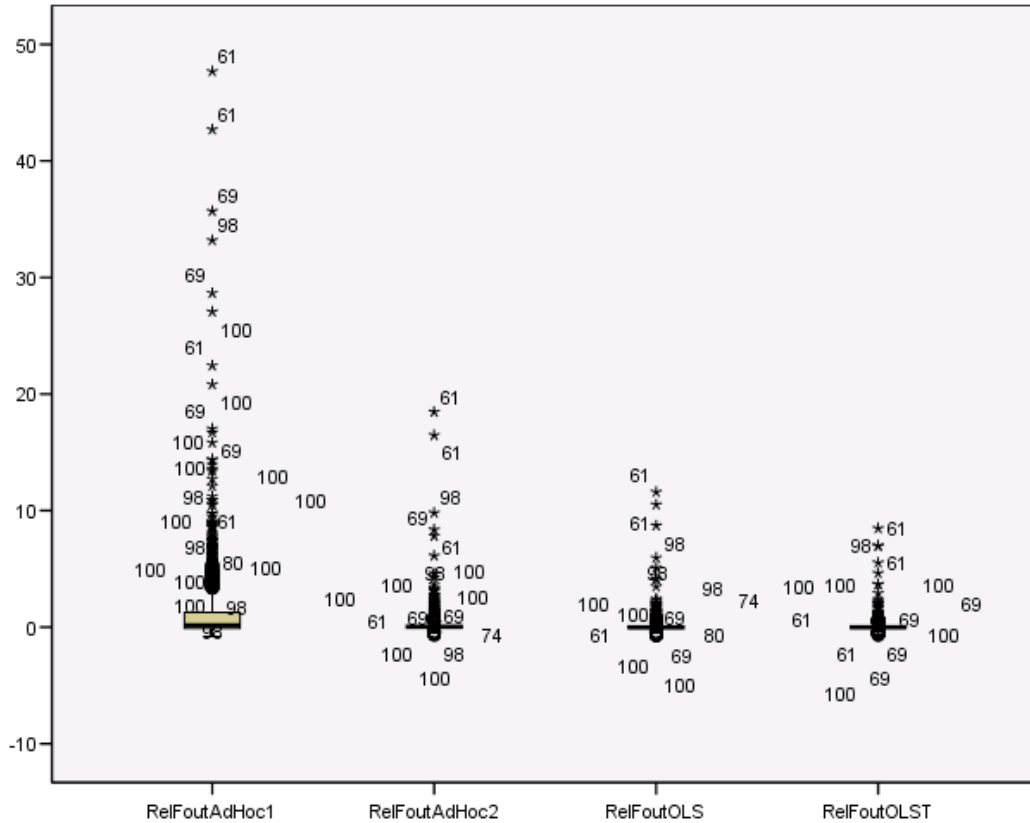


Figure 1. Box plots of mean relative estimation error

#### 4.3 Performance of the least squares estimators

A mean relative error of .16 is still considerable, making it worthwhile to explore other possibilities of estimating the transition weights. Least squares estimation is a possibility. In this approach, the dependent variable consists of the observed totals in the new classification. For each combination of classes  $o$  and  $s$  where over the period of double coding at least one (non-zero) observation exists, a dependent variable is constructed that satisfies

$$x_t^{OS(os)} = \begin{cases} Y_t^{O(o)}, & s \in o \\ 0, & s \notin o \end{cases}.$$

Note that in our example, 46 non-zero entries exist in Table 1 so that 46 dependent variables are



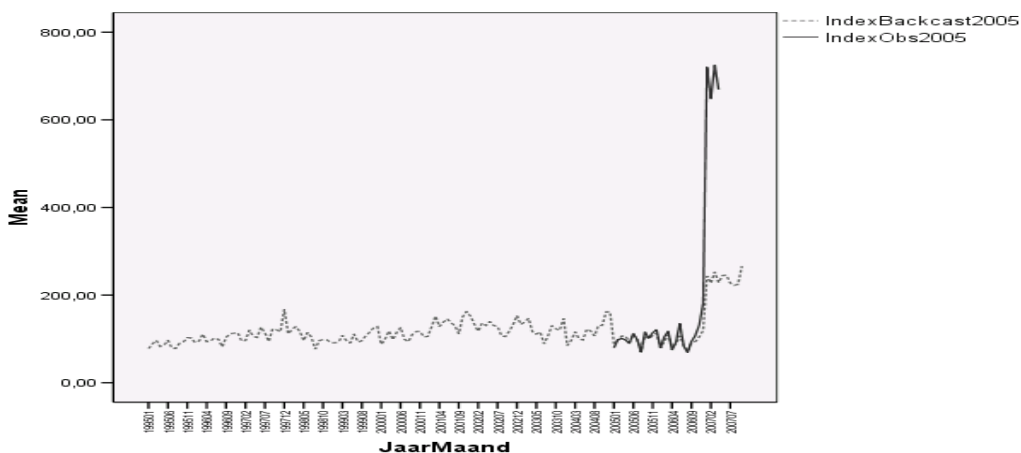
**Centraal Bureau voor de Statistiek**  
Sector DMH, MIC

constructed, because of course, the independent variables that correspond to elements in the transition matrix equal to zero, will drop out of the analysis and can be omitted prior to the analysis. Also, because we deal with a model, it is preferable to model log turn-over instead of raw turn-over, as a way to avoid the estimating of turn-over less than zero. Using log turn-over, and after estimating the model (e.g., in SPSS), we find that the correlation between observed and estimated turn-over (computed of course, by exponentiation) equals  $r=.988$ , which is equal to the value obtained for the ad hoc estimator based on turn-over. The MRE however, is only .08 (see Figure 1, RelFoutOLS), a considerable reduction compared to the ad hoc estimator. Finally, estimation of the expanded model correcting for monotonic heterogeneity (Formula 11), yields  $r=.999$  and  $MRE=.06$  (also, see Figure 1, RelFoutOLST). These are still better values, but the instability of the associated parameters made us prefer the simple OLS estimator over the latter one.

Statistics for this estimator at the level of new publication cells are the following:

	80	83	65	72	74	70	96	69	71	60	61	77	79	82	84	85	97	98	100	$\Sigma$
MRE	.11	.00	.04	.00	.02	.07	.03	.21	.03	.03	.87	.00	.01	.00	.00	.03	.09	.07	.07	<u>.08</u>
N	8843	801	1323	1229	208	410	2398	1705	208	670	28	1597	242	2137	5187	373	337	3435	2848	<u>33979</u>

Obviously, quite acceptable results are obtained in the new publication cells  $s=(83, 72, 74, 96, 71, 60, 77, 79, 82, 84, 85, 97)$ , whereas the model performs less well in  $s=(80, 65, 70, 69, 61, 98, 100)$ . Invariably, the reason for bad performance is heterogeneity, caused either by creation of too heterogeneous new publication cells, or by using data from too heterogeneous old publication cells. Consider for example, the complete time series in publication cell 70:



In this case, heterogeneity is caused by the introduction of new major company in  $s=70$  in January 2007. New publication cell  $s=96$  is fed by the same old publication cell as 70, and is therefore also affected by the heterogeneity, though to a lesser extend because there are much more companies in 96 (see Table 1).



## 5. Conclusions

It is possible to obtain reasonable approximations of historic timeseries using rather simple methodology. In many cases, simple techniques yield acceptable results, but:

1. Estimated backcasted timeseries (BTS) depend on period of double coding
2. Accuracy of BTS cannot be measured preceeding period of double coding
3. Some new publication cells suffer from scarcity
4. Some new publication cells suffer from heterogeneity

The latter problem appears to be the most serious one. Heterogeneity can have several causes:

1. Too different cyclical effects (e.g., effect of season, economical cycle, etc.)
2. Too different activities
3. Other, e.g., founding or vanishing of large company

In the above analysis, we have compared two different models. However, there is an almost infinite number of different models that can be specified to account for heterogeneity, and of which some may be more fruitful than others. Future research should shed light on this issue.

## Referenties

Fortier, S. (2005). *The conversion of historical time series according to a revised classification in the wholesale and retail sale monthly survey*. Luxemburg: Eurostat.

Kampen, J. K. (2007). *CoSBI 2008: Methodologie voor het terugleggen en backcasten*. Interne CBS-nota, Sector DMK, BPA no. DMK-2007-05-04-JKPN, CBS Heerlen.

Moauero, F. (2005). *Modelling a change of classification by a structural time series approach*. Rome: ISTAT.

## PARKING ZONE

### 5.2 Complications in confidence indices

For confidence indicators backcasting is more complicated than for turn-over indices. The former type of qualitative statistics basically has three answer categories for each question: a positive one (“increased”), a negative one (“decreased”) and a neutral one (“remained the same”):

$$y_i = \begin{cases} -1 \Rightarrow \text{increased,} \\ 0 \Rightarrow \text{remained the same,} \\ 1 \Rightarrow \text{decreased} \end{cases} \quad (16)$$

For publication purposes, weighted percentages per answer category and per question are calculated. Subsequently, the balance of positive and negative answers is calculated for each question, as the main variable. A weighted average of balances for a selected number of questions may then be calculated in order to arrive at e.g., producers confidence or an economic sentiment indicator, but that step will not be dealt with here. So in the case of confidence surveys, for a given NACE Rev. 2 aggregate  $S$  the individual answers of firms  $i$  in the same answer category  $c$  are weighted by a “inner weight”  $w_{it}$  and aggregated per type of answer, after which the total weight per answer category is divided by the total weight of all answer types:

$$I_{jt}^{O(o)} = \sum_{i \in o \cap j} w_{it} y_i / \omega_{jt}^{O(o)} = Y_{ij}^{O(o)} / \omega_{jt}^{O(o)}, \quad (17)$$

For the calculation of overall aggregates, additional “outer weights”  $W_t^{O(o)}$  are necessary. In that case, the results for a lower aggregate of NACE can be treated in a similar way as the individual answers within such a NACE group:

$$J_t^{O(o)} = \sum_j W_{jt}^{O(o)} I_{jt}^{O(o)} / W_t^{O(o)}. \quad (18)$$

The aggregated confidence indicators data according to NACE Rev. 1.1. are transferred to the new classification by using a conversion table of the weighting scheme to calculate weighed averages of the old data for every NACE Rev. 2 publication aggregate. First, the “outer weights” of the old NACE groups, used to aggregate the results per NACE code to higher aggregates, have to be distributed over the elements of these groups that go to different new NACE groups by means of a transition weight  $p_{W(tj)}^{OS(os)}$ , which yields

$$W_{tj}^{S(s)} = \sum_o p_{W(tj)}^{OS(os)} W_{jt}^{O(o)}. \quad (19)$$



But the inner weights also need correction,

$$\omega_{ij}^{S(s)} = \sum_{i \in s \cap j} \omega_{it} = \sum_o p_{\Omega(tj)}^{OS(os)} \omega_{jt}^{O(o)}, \quad (20)$$

with

$$p_{\Omega(tj)}^{OS(os)} = \sum_{l \in s \cap o \cap j} \omega_{lt} / \omega_{jt}^{O(o)}. \quad (21)$$

The index in the new classification may be written as

$$I_{jt}^{S(s)} = Y_{tj}^{S(s)} / \omega_{jt}^{S(s)}, \quad (22)$$

with

$$Y_{tj}^{S(s)} = \sum_{i \in s \cap j} w_{it} y_i = \sum_o p_{Y(tj)}^{OS(os)} \sum_{i \in o \cap j} \omega_{it} y_i = \sum_o p_{Y(tj)}^{OS(os)} Y_{tj}^{O(o)}, \quad (23)$$

and

$$p_{Y(tj)}^{OS(os)} = \sum_{i \in o \cap s \cap j} w_{it} y_i / Y_{tj}^{O(o)}. \quad (24)$$

Finally, the expression of the aggregate index value as defined analogously to Formula (18) is

$$J_t^{S(s)} = \frac{\sum_j \left\{ \sum_o p_{W(tj)}^{OS(os)} W_{jt}^{O(o)} \left( \frac{\sum_o p_{Y(tj)}^{OS(os)} Y_{tj}^{O(o)}}{\sum_o \sum_{l \in o \cap j} p_{\Omega(tj)}^{OS(os)} \omega_{jt}^{O(o)}} \right) \right\}}{\sum_j \sum_o p_{W(tj)}^{OS(os)} W_{jt}^{O(o)}}, \quad (25)$$

which has only the several transition weights that are a function of the old classification and that need to be estimated.

This approach implicitly assumes that confidence is evenly distributed within all components of the old NACE aggregate. In other words: every element of the old NACE aggregate that corresponds with a different new NACE aggregate has the same confidence as the other components. Statistics Netherlands has tested this approach on an existing data set of the Industry Survey. The method produces results that are consistent with the overall totals of the old series. The underlying assumption however, of all elements of an old NACE aggregate having the same confidence does not hold for the Dutch situation.



**Centraal Bureau voor de Statistiek**  
Sector DMH, MIC

A second adaptation of the general macro approach to confidence indicators accepts that confidence levels (e.g., the percentage of positive answers) may differ within the components of an old NACE group. It calculates the ratio between confidence of every cell in the transition matrix and the confidence of the corresponding entire old NACE group (row in the matrix). For an old period  $t-u$  this gives,

$$PC_{t-u}^{c,S(s)} = \sum_s (OW^{OS(os)} (PC_t^{c,OS(os)} / PC_t^{c,O(o)}) PC_{t-u}^{c,O(o)}) / \sum_s OW^{OS(os)} . \quad (15)$$

This method implicitly assumes that the ratio  $(PC_t^{c,OS(os)} / PC_t^{c,O(o)})$  is stable over time. The confidence levels of the cells  $PC_t^{c,OS(os)}$  composing a given new NACE aggregate thus differ, but all cells coming from the same old aggregate still follow the same pattern over time. Statistics Netherlands also tested this approach on an existing data set of the Industry Survey. The method produces results that are consistent with the overall totals of the old series. A comparison with the results from a micro approach for several periods showed that ratio  $(PC_t^{c,OS(os)} / PC_t^{c,O(o)})$  was .... over time.