Producing historical time series for STS-statistics in NACE Rev.2

Theory with an application in industrial turnover in the Netherlands (1995–2008)

Gert Buiten, Jarl Kampen en Sidney Vergouw

Discussion paper (09001)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2005-2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/'06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/'04–2005/'06	= crop year, financial year, etc. 2003/'04 to 2005/'06 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

Information
Telephone .. +31 88 570 70 70
Telefax .. +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax .. +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

© Statistics Netherlands, The Hague/Heerlen, 2009.
Reproduction is permitted. 'Statistics Netherlands' must be quoted as source.

6008309001 X-10

Producing historical time series for STS-statistics in NACE Rev.2

Theory with an application in industrial turnover index in the Netherlands (1995-2008)

Gert Buiten, Jarl Kampen en Sidney Vergouw

Summary:This contribution discusses possible scenarios and methodologies for the national statistical agencies for backcasting the new classification scheme (NACE Rev. 2.0) in existing time series of business statistics. We provide 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 business statistics is handled and illustrated with an example. We conclude that it is possible to obtain reasonable approximations of historic time series using rather simple methodology, but the quality of the backcasted time series is hampered by heterogeneity of classes.

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 NACE1.1 coding (macro methods); see Moauro (2005). However, the results are applicable to 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 retrapolating published series using a transition 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 Short term business statistics. This is followed by an application in real data. Finally, some conclusions and issues for future research are proposed.

2. Methodologies for producing historical time series

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. For example, 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 of 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 form 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 between the old and new results. In practice however, differences in the outcomes may be less transparent. In cases where outlier treatment, imputation for nonrespons, 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.

The micro 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 transition 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 transitions, the coding process simply assigns the NACE Rev.2 code from the correspondence tables, that is, by use of a key (see section 2.1).

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. 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-to-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 transition coefficients. These transition coefficients are derived from a transition matrix. This means that at least one point in time of double coded records according to the old and revised classification is required. The main advantage of this method is that it's a relatively low resource and time consumption technique. The use of transition 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 i.e. are consistent, and at lower aggregation levels the differences are fully attributable to changes in the classification. Analysing, documenting and communicating differences is therefore relatively simple. The micro approach may lead to problems regarding to consistency when the imputation was done 'by hand', that is using expert knowledge not available anymore.

Although the terms correspondence tables and transition 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. 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.

Transition 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 transition 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. Transition matrices can be used for backcasting purposes. In that case the number of businesses, turnover, value added and numbers of employees are the most common variables used for the calculation of transition factors. One should be aware of differences in structure between these variables, since the structure of the used variable determines the transition matrix. Several cases can be distinguished:

1. The transition 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 transition matrix are the same as those published for the reference period.
2. The transition 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 transition 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 transition matrix with the weights. This is needed to aggregate the backcasted series at the lowest level of detail to higher aggregates.
3. The transition matrix is based on a different variable than the "target variable" and on a different period. This is e.g., the case for the confidence indicators of Short term business statistics and price indices in the STS area. A transition matrix is used to convert the weighting system from NACE Rev 1.1 to Rev. 2, thus producing a new transition matrix with the weights. Depending on the approach used, a third kind of transition 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 transition matrix.

In cases 2 and 3 a separate transition 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 transition 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 transition 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 Short term business statistics, 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. This problem will be strongest in the area of services, but will also play 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 transition 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 Short term business statistics 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, transition 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 transition coefficients are obtained to convert the aggregated estimates from the initial to the revised classification. For the periods between the two benchmark 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 transition coefficients to all the periods of the time series. Just like the assumption made at the macro-approach, this assumption does not always hold. 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 short term business statistic indicators.

A second possible approach is possible in cases were the micro method will be applied for e.g., three consecutive years. For a monthly or quarterly statistic, this allows to calculate 36 monthly of 12 quarterly transition matrices. Based on such a data set, it is possible to analyse the structural patterns in the transition matrix. Are there seasonal patterns in the transition coefficients? In the case of 1-to-many splits, are there differences in the growth trend between the n new NACE groups? Obviously, in 1-to-1 transitions 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 transition coefficients for each row in a transition 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 short term business statistic indicators

With the basis methods described above in mind, time series of Short term business statistic 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 representatively, 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 probabilities 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 and 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 the frequencies

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

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

$i = (1, \ldots, N_t)$. 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

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

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 nonrespons in *Y*.

In most cases, our data will only consist of a sample of n_t businesses (meaning also, that measurements of the new classification at the level of the population will be lacking). 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 in the *O* sampling design, denoted ρ_{it}, can differ substantially from those in the *S* sampling design, for instance, because in 2009, new Neyman allocations will 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,

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

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:

$$\hat{Y}_t^{S(s)} = \sum_i^{n_t} a_\cdot^{S(is)} y_{it} / \rho_{it} . \qquad (5)$$

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 (see Section 2.4). 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, macro approaches 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 volume of transitions between *o* and *s* at all periods of the historical time series. This requires at least one point in time where dual coding at the level of micro data is available, producing (on the basis of the correspondence tables) a H_O x H_S transition matrix that distributes at time *t* the classification in H_O old codes to the classification in the H_S new codes. 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

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

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*. Strictly speaking, because the parameters are estimated, all parameters should receive hats; however, for notational clarity, we omit the hats in the forthcoming explanations. For backcasting on the level of aggregate data, we generalise Formula (6),

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

with $p_{Y(t)}^{OS(0s)}$ 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,

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

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 transition 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 transition matrix):

$$p_{Y(t)}^{SO(so)} = \frac{\sum_i a_t^{O(io)} a_t^{S(is)} y_{it} / p_{it}}{Y_t^{S(s)}} . \tag{9}$$

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

$$\hat{Y}_t^{S(s)} = \sum_o p_{Y(.)}^{OS(os)} Y_t^{O(o)} \quad . \tag{10}$$

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

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

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

$$p_{Y(t)}^{OS(os)} = p_{Y(.)}^{OS(os)} + t \times q^{OS(os)} . \tag{11}$$

The least squares estimator may therefore produce better results than the *ad hoc* estimators. A real life example of this methodology is provided in the next section.

4. The evolution of Dutch business turnover 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 like the diagram in Figure 1. 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: the right hand cluster consists of 1-to-many splits, the left hand, of many-to-many splits. Obviously, the backcasting of time series 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 time series of turnover that run from January 1995 until August 2007, contained total turnover within the new classification that run from January 2005 until April 2007. The latter series have to be constructed on the basis of microdata, and the procedure was as follows. First, for the database of companies of April 2007, for each company,

1. Assign NACE Rev. 2.0 on the basis of one-to-one transitions (i.e. use of a key);

2. If one-to-one transition is not applicable, assign NACE Rev. 2.0 on the basis of Prodcom;
3. 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 turnover in the new classification is done by applying the domain estimator of Formula (4). Although estimated, we refer to the thus obtained turnover 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. Statistics Netherlands uses the mean monthly turnover in the baseline year *B* as the reference number, and the index formula is

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

In the new classification, the index can be written as

$$I_t^{S(s)} = Y_t^{S(s)} \Big/ \frac{1}{12}\sum_{t \in B} Y_t^{S(s)}, \quad t \geq 1, \tag{13}$$

where for the period without double classifications, the historic total turnover $Y_t^{S(s)}$ 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 (13). 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 the number of businesses and of turnover. 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). 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

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

As a measure of adequacy of the estimator, the backcasted total turnover using this procedure can be compared to the observed turnover in the period 1/2005 – 4/2007 obtained by Formula (4). We find that the correlation of the backcasted series and the observed series equals *r*=.988. The mean relative absolute estimation error (MRE) of the estimator, defined as

$$\frac{1}{n}\sum_t \sum_s \left| \hat{Y}_t^{S(s)} - Y_t^{S(s)} \right| \Big/ Y_t^{S(s)} ,$$

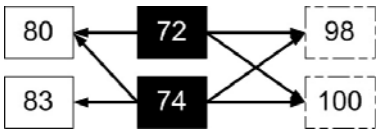
equals 1.68, and its box plot can be viewed in Figure 2. Alternatively, if we use the transitions of turnover over the same period (figures not printed in order to save space), we obtain *r*=.998 with MRE=.29 (see also Figure 2). In other words, backcasting the series by the ad hoc estimator on the basis of turnover is to be preferred over the estimator using the number of businesses.

4.3 Performance of the least squares estimators

A mean relative error of 29% 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, an independent 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 independent variables are 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. To make the procedure more clear, consider for instance, the system of publication cells in the left hand upper corner of Figure 1 (also, see Table 1),



The equation required for OLS estimation of constant transition weights (Formula 10) in this system is

$$\begin{pmatrix} Y_{t-u}^{S(80)} \\ Y_{t-u}^{S(83)} \\ Y_{t-u}^{S(98)} \\ Y_{t-u}^{S(100)} \\ \vdots \\ Y_{t}^{S(80)} \\ Y_{t}^{S(83)} \\ Y_{t}^{S(98)} \\ Y_{t}^{S(100)} \end{pmatrix} = \begin{pmatrix} Y_{t-u}^{O(72)} & 0 & 0 & Y_{t-u}^{O(74)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & Y_{t-u}^{O(74)} & 0 & 0 \\ 0 & Y_{t-u}^{O(72)} & 0 & 0 & 0 & Y_{t-u}^{O(74)} & 0 \\ 0 & 0 & Y_{t-u}^{O(72)} & 0 & 0 & 0 & Y_{t-u}^{O(74)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{t}^{O(72)} & 0 & 0 & Y_{t}^{O(74)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & Y_{t}^{O(74)} & 0 & 0 \\ 0 & Y_{t}^{O(72)} & 0 & 0 & 0 & Y_{t}^{O(74)} & 0 \\ 0 & 0 & Y_{t}^{O(72)} & 0 & 0 & 0 & Y_{t}^{O(74)} \end{pmatrix} \begin{pmatrix} p_{Y(.)}^{OS(72,80)} \\ p_{Y(.)}^{OS(72,98)} \\ p_{Y(.)}^{OS(72,100)} \\ p_{Y(.)}^{OS(74,80)} \\ p_{Y(.)}^{OS(74,83)} \\ p_{Y(.)}^{OS(74,98)} \\ p_{Y(.)}^{OS(74,100)} \end{pmatrix}.$$

Of course, for the complete set of publication cells, the equation involves 46 parameters on the right hand side which is too tedious to write here in full. Also, because we deal with a model, it is preferable to model log turnover instead of raw turnover, as a way to avoid the estimating of turnover less than zero. Using log turnover, and after estimating the model (e.g., in SPSS), we find that the correlation between observed and estimated turnover (computed of course, by exponentiation) equals r=.998, which is equal to the value obtained for the ad hoc estimator based on turnover. The MRE however, is only .24 (see Figure 2), a considerable reduction compared to the ad hoc estimator. Finally, estimation of the expanded model correcting for monotonic heterogeneity (Formula 12), yields a slightly increased r=.999 and decreased MRE=.20 (also, see Figure 2). These are still better values, but the instability of the additional 46 estimated parameters in the model made us prefer the simple OLS estimator over the latter one.

Statistics for the simple OLS estimates 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**	**Σ**
MRE	*.34*	*.14*	*.17*	*.06*	*.10*	*.25*	*.25*	*.50*	*.19*	*.18*	*.64*	*.04*	*.11*	*.01*	*.01*	*.01*	*.22*	*.29*	*.25*	*.24*
N	*8843*	*801*	*1323*	*1229*	*208*	*410*	*2398*	*1705*	*208*	*670*	*28*	*1597*	*242*	*2137*	*5187*	*373*	*337*	*3435*	*2848*	*33979*

Obviously, quite acceptable results are obtained in the new publication cells s=(77, 82, 83, 84, 85) whereas the model performs less well in s=(69, 70, 80, 96, 98, 100) where MRE's are above average. 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, as displayed in Figure 3. In this case, heterogeneity is caused by the introduction of a large 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 extent because there are much more companies in 96 (see Table 1). In this particular case, a decision must be made whether the data used for estimating the transition matrix should exclude the newly introduced large company.

5. Conclusions

It is possible to obtain reasonable approximations of historic time series using rather simple methodology. The estimator based on least squares estimation performs better than the ad hoc estimators, but it requires a longer period of double coding if it is to be applied fruitfully (in fact, if only one point in time with double coding exists, the least squares estimator coincides with the ad hoc estimator based on turnover). In many cases, simple techniques yield acceptable results, but:

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

The latter problem appears to 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.

References

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.

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

Table 1. Transition matrix of number of businesses (January 2005 – April 2007)

	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	1210	907	*10898*
74	62	801	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	*1199*
80	0	0	124	489	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	31	*950*
87	0	0	0	0	0	410	2153	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	*452*
88	0	0	0	0	0	0	0	1671	64	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	223	0	59	*952*
64	0	0	0	0	0	0	0	0	0	0	28	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	883	363	*2843*
71	0	0	0	0	0	0	0	0	0	0	0	0	242	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	119	224	*2480*
75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5187	0	0	715	162	*6064*
76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	373	0	40	29	*442*
Σ	*8843*	*801*	*1323*	*1229*	*208*	*410*	*2398*	*1705*	*208*	*670*	*28*	*1597*	*242*	*2137*	*5187*	*373*	*337*	*3435*	*2848*	*33979*

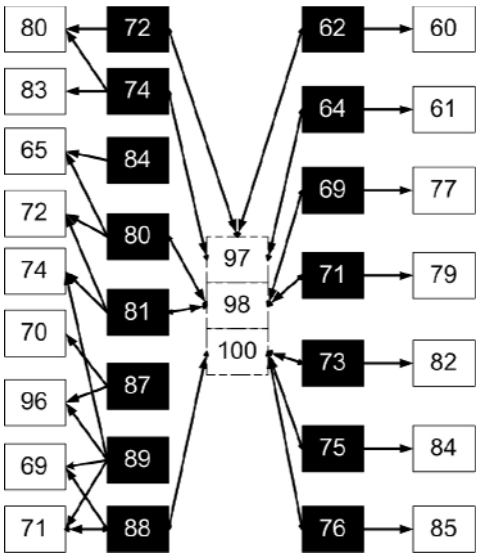


Figure 1. Example of a diagram of NACE Ver. 1.1 – NACE Ver. 2.0 transfers

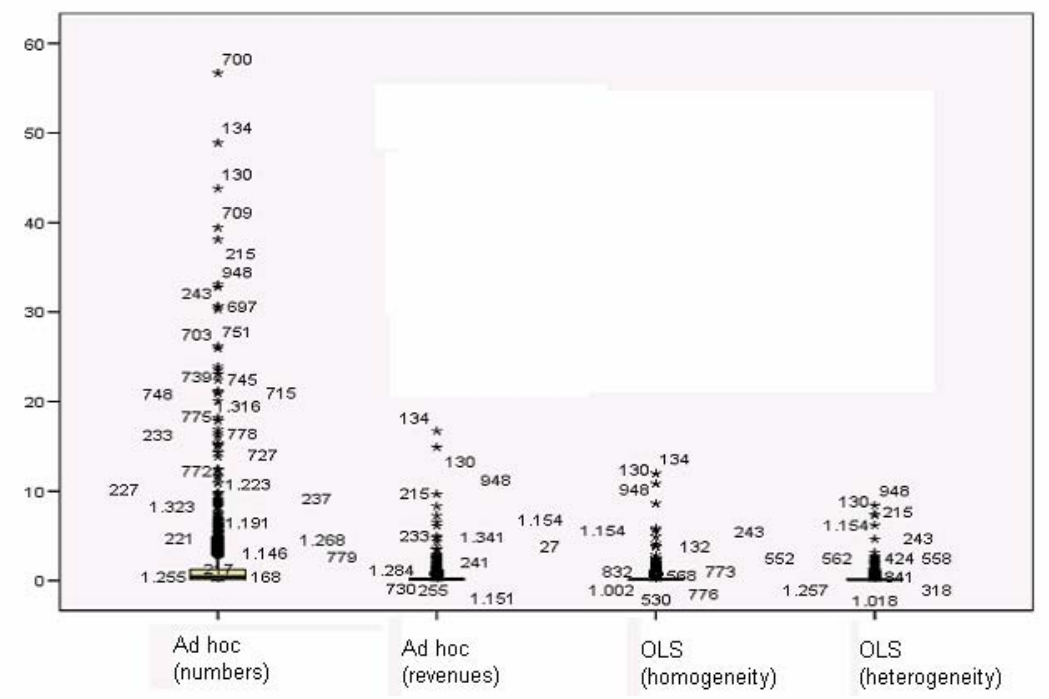


Figure 2. Box plots of mean relative estimation error (MRE)

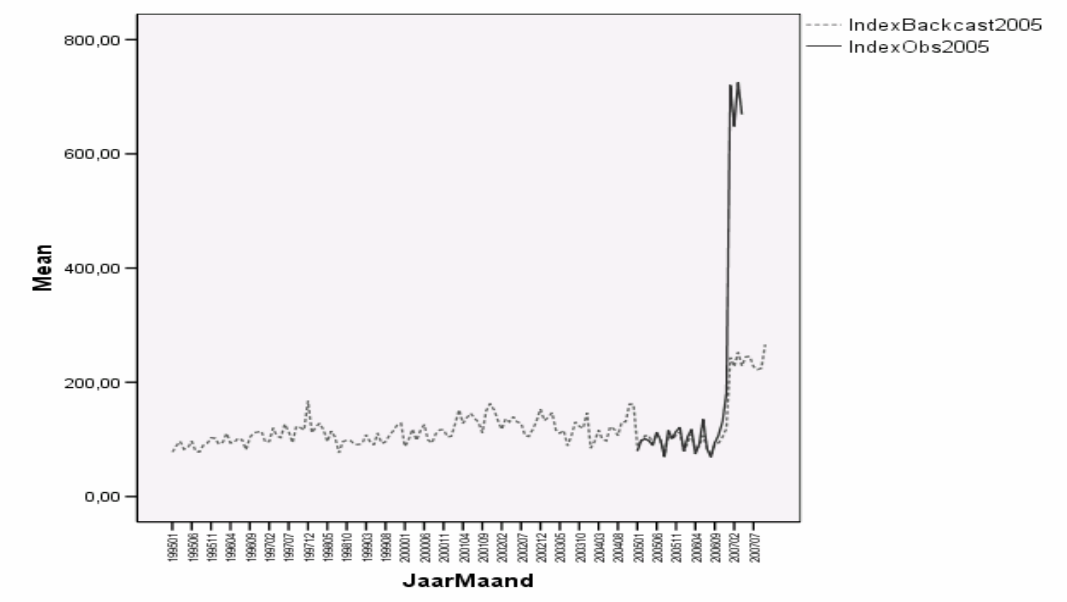


Figure 3. Turnover index for new publication cell s*=70 (1995-2008)*