Environmental expenditures by the Belgian industries in 2002.
Imputation techniques and results.

Bruno Kestemont, Statistics Belgium, October 2004.

The action has received European Comission funding coming from DG Environment
(Grant ESTAT 200271700002)

Table of contents

1. Introduction

A survey on environmental investments exists in Belgium since the year 1995 This survey is conceived as an annex or by product of the Structural Business Survey covering broader economic aspects (see Kestemont, 2000). The answer was compulsory. Current environmental expenditures where added for the first reference year 1999, with no compulsory answer from the respondents (Kestemont, 2002).

For reference year 2002, the Structural Business Survey system was completely changed in Belgium. Most of the companies must send theirs data to the National Bank (Central of Balances), and a limited number of them (the smallest ones) still receive a questionnaire from NSI Belgium. By the way, annexes on environmental expenditures where adapted to be in line with the latest Eurostat definitions.

A new Royal Decree was signed (later than foreseen in the project) and we were able to send out our questionnaires (annexe 1), on a compulsory basis, in June 2003. The novelty was the inclusion of current expenditures as compulsory annexe, and a slight adaptation of the annexes on environmental investments.

The statistical population (N) is about 700000 companies (including 1 person companies). The stratified sample (n) consisted on about 40233 companies of all size classes, in which about 23000 receiving a question on total "end-of-pipe" PAC investment (variable 21110) and total integrated PAC investment (variable 21120). 1860 industries[1] (from NACE 14-41 except 37) received the annexe on current expenditures (with the detail per environmental domain air, water, waste, soil and other), and 1853 (including NACE 37) received the annexe detailing the PAC investments per domain. Both annexed questionnaires were only sent to companies of more than 49 employees[2].

The big change in statistical procedure (getting the data through the National Bank, sending out questionnaires only to the companies not sending their results to the Bank) had an impact both on the overall answer rate to the questionnaire by companies, and to the internal (much more complicated) management for NSI. The objective of avoiding duplicate burden to the companies[3] was achieved, but on the cost on a difficult year for the statisticians.

Concerning environmental investments, some further delay was needed to get all the answers validated.

Concerning current environmental expenditures, the sample was reduced to companies with more than 49 employees. However, we observed that after 3 years of

[1] Exactly 1544 "environmental" questionnaires were joined initially to the biggest companies, but 16 additional responses (mainly SMEs) could be obtained later on, from companies being part of the overall SBS sample and that had already been surveyed previous years.
[2] This is a change with previous years, when companies of 20-49 persons or a turnover greater than 6.96 millions euros also received the annexes.
[3] Before this, the 243000 biggest ones had to send their balance to the bank, and in addition, the smallest ones and a sample of biggest had to answer our questionnaires, some times on almost the same variables.

facultative survey, the respondents did not all note that this time, it was mandatory. We get a poor answer rate that obliged us to phone and fax to most of the companies in order to reach at least 50% answer. More editing and imputation was then needed from our side. Actually, the method used for imputation and extrapolation was, as a result, more close to the methods for a facultative survey. In this report, we do not focus on the method used to extrapolate the data on environmental investments, which are described in Kestemont (2000). But the method used to estimate current environmental expenditures is described more in detail.

2. Overall survey method

The survey on environmental expenditures by industry was coupled with the Structural Business Survey (SBS), as described in Kestemont (2000).

For year 2002, the SBS changed its method of collecting data:

1- The ~243000 biggest companies have to send their balance data to the National Bank Balance Central;
2- NSI receive the administrative file and covers the "gaps" (e.g. the smallest companies, and the environmental variables) by sending a questionnaire out.

This ambitious change needed the publication of a new Royal Decree, including the revised annexes on environmental expenditure variables:

-the total end of pipe and integrated PAC investments was asked to almost all companies of the sample (as in the old decree);
-the details (per environmental domain) of this investments (variables 21110 and 21120) was asked only to the biggest industrial companies, as before;
-the annexe on current environmental expenditure became mandatory (after 3 years of facultative survey).

The help desk for PAC questions was the service "environment", while the overall management of the survey staid still under the SBS unit responsibility. The questions on PAC investments where treated within the system of SBS services. This includes automatic and manual error checking, editing and imputation[4]. Automatic checking includes the verification of totals and certain ratios (for example, PAC investment should be lower than total investment). Manual checking and phone calla to the companies occur when any part of the SBS questionnaire is missing or shows errors (e.g. errors of monetary units): the PAC questions are then corrected accordingly[5].

The annexe on current PAC expenditure was taken over by the environmental service for further calling to the companies, because the answer rate for this annexe was abnormally low. The reason for this low answer rate is that the companies (and even some surveyors of the SBS) did not catch that this time, this annexe was mandatory!

[4] For a description of the editing and imputation procedures of the SBS in Belgium, see Vekeman (2004).
[5] Typical errors are explained in Kestemont (2000, 2001).

As a result, we had to phone back much more companies than foreseen, using the same overall method as in Kestemont (2002) for a facultative survey.

3. Method for estimating current PAC expenditures

As said above, the method used to estimate the PAC investments has fundamentally not changed since Kestemont (2000). However, the method of estimating current PAC expenditures has been improved following a methodological test on 2001 data (Kestemont, 2004). We will describe this method more in detail in the current report.

4. Estimation of missing values

After calling and recalling companies by telephone and fax, we reached an answer rate of 54% for current PAC expenditures (table 1). This would have been a good result for a facultative survey, but it is a very bad result for a compulsory survey!

A simple extrapolation of the available responses to the all population would be the worse method to use, because we do not know if the population of respondents is representative of the population of non respondents (certainly not)[6].

Edition and imputation can improve the (virtual) answer rate before to start any extrapolation. Edition and imputation are generally performed on the level of the individual survey respondent level, also called the micro data level, on the basis of precautious hypothesis verified by statistical analysis described below.

4.1 Editing

An editing procedure is the process of detecting and handling errors in data, including the definition of a consistent set of requirements, their verification on given data, and elimination or substitution of data which are violating the defined requirements (Vekeman, 2004).

Manual editing was performed during the phone call procedure, when our staff not only recalled the non respondent, but also the "doubtful respondents" for which an answer seemed incoherent with an answer of the previous year or from the answer of other similar companies - for example companies answering "zero" where an expenditure was highly expected (see Kestemont, 2001). At this early stage, during nearly 1 year, we stayed on dialogue with the respondent and if there was an editing, this was practically done by the respondent itself[7]. The quantity of answers after this stage is given in table 1 of following chapter.

[6] In other words, this approach would rely on the restrictive assumption that missing data are Missing Completely At Random (MCAR).

[7] Note that for practical reasons, no encoding or computer verification could occur during this stage. Only human expert judgment with the help of the paper files of questionnaires from earlier years and same sectors is used to identify "doubtfull" answers. This means that some additional outliers are only identified later by statistical analysis. At this time of the planning, it is too late to proceed to further

4.2 Imputation

Imputation concerns essentially the completing of missing information with most plausible possible information. Omitting imputation results in an over weighting of respondents as compared to non-respondents[8].

If there is a substantial amount of auxiliary information available at the time when the estimates are to be calculated, an imputation procedure may rely on this auxiliary information. A stratified survey does have a clear advantage whenever imputation is concerned. Any information of responding companies from the same stratum can be used to calculate ratios to be used, e.g. for the breakdown of the totals of the result account of the non-responding company (see Vekeman, 2004). If there are not enough "donors" in a stratum (e.g. on NACE, level 4), it is still possible to use an enlarged stratum (e.g. on NACE level 3). This does stabilize the ratios obtained, but it definitely makes them less suitable, since the 'donor' companies will be less similar to the non-responding one.

4.2.1 Deterministic imputation

Deterministic imputation is used where only one correct value exists, as in the missing sum at the bottom of a column of numbers. A value is thus determined from other values on the same questionnaire (UN-ECE, 2000).

We performed a deterministic imputation on the level of the databank following logical principles described below:

1-a company declaring zero for the total should have zero for the details;
2-a value in "other" is supposed to cover all remaining, not detailed, expenditure: a zero is attributed to all missing domains and the total is calculated accordingly;
3-a company that answered for all domains, but not for "others" is supposed to have classified all of its expenditures: other is then estimated to be zero, and the total is calculated accordingly.

Deterministic imputation is illustrated below:

	Total	Air	Water	Waste	Soil	Other
Case 1	**0**	*0*	*0*	*0*	*0*	*0*
Case 2	**1000**	*0*	*0*	*0*	*0*	**1000**
Case 3	**3000**	**0**	**1000**	**2000**	**0**	*0*

In bold : response

telephone check, and only statistical imputation procedures are used (see an example of problematic case in chapter "deterministic imputation").

[8] In the case of, for example a stratum with 2 respondents out of 3, with a high expenditure, say 1000 and 2000, and one non-respondent with a low expenditure, say 500, the simple extrapolation would implicitly impute a value of 1500 to the non-respondent. A sound imputation technique, based on reliable auxiliary data and hypothesis, can reduce this obvious error.

In itallic : imputation

A remaining case is more problematic:

	Total	Air	Water	Waste	Soil	Other
Case 4	**3000**	?	**1000**	**2000**	?	?

In this frequent situation, a company answered something for several domains, and gave as total, the total of its detailed answers. This could be interpreted in 2 ways: 1-either the answer is fully correct and we should add zero to the remaining fields; 2-either the company declared the total of what it could identify, and we can not affirm that there is no expenditure for other or undifferentiated domains.

In this latest situation, we did NOT impute the missing values. This implies that the total answered could be considered "doubtful" (possibly underestimated) and is subject to post-editing (once we would have estimated detailed missing values – see following chapters).

As a summary, we can say that this phase of deterministic imputation consisted mainly to add missing zero's. The related SPSS syntax is available in annexe 2. The answer rate before and after deterministic imputation is given in table 1.

Table 1: Answer rate after deterministic imputation

(after elimination of dead companies)	Total	Air	Water	Waste	Soil	Other
Sample	1860	1860	1860	1860	1860	1860
Answers after recall[9]	1001	369	526	937	380	514
% answer	54%	20%	28%	50%	20%	28%
% Deterministic imputation	0%	11%	8%	2%	11%	1%
Answers after deterministic imputation	1001	577	676	967	586	526
% answers after det. imputation	**54%**	**31%**	**36%**	**52%**	**32%**	**28%**

4.2.2 Model based imputation

In order to avoid error dissemination, we avoided model-based imputation in chain. That means that only real answers and deterministic imputed values should serve as "donors" for model-based imputing missing values[10].

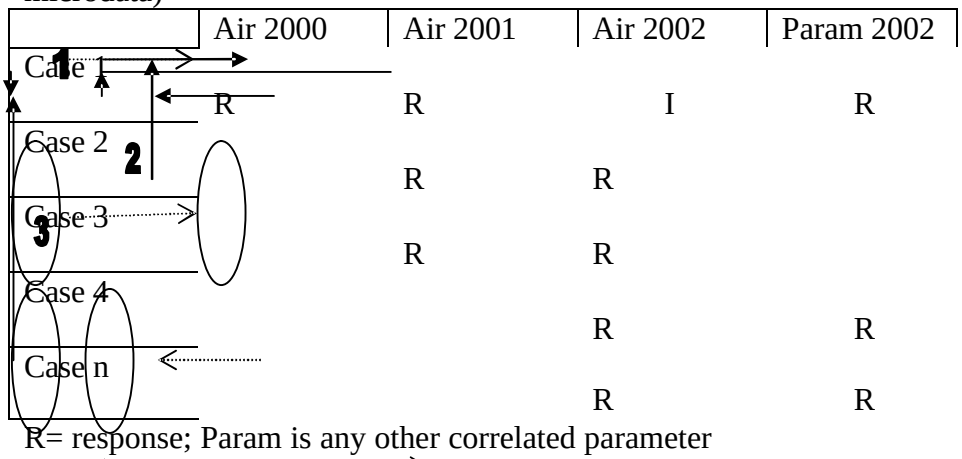
Since the outliers were verified during the phone calling procedure, we also kept them in the donors set. The risk associated with this decision is balanced by the fact that if the analysis shown a little correlation (even due to outliers), we rejected the set of donors and we looked for another method of imputation.

[9] About 45 additional answers came too late to be included in the analysis. This data will be available for temporal imputation of next years. Alternatively, it is also possible to replace the imputed value by the real value for these respondents, to analyse the strength of the imputation techniques, and to recalculate the results for year 2002, when a new publication of the all series will be planned.

[10] Deterministic imputation is secure enough. Edited answer is the "best available" estimate (with a remaining unknown response error).

The overall philosophy of model-based imputation is illustrated in figure 1 below.

Figure 1: Model-based imputation techniques (example for air data in a table of microdata)



R= response; Param is any other correlated parameter

→ = imputation ⇢ = a link as base of imputation

Suppose a table of microdata with R representing the available responses, and the empty cells the missing values. We want to impute a value in the cell "Air 2002 x Case 1". The available information R in the different other cells can be used to impute the value following different methods illustrated on the figure.

1 We call "serial imputation" the method using data from 2000 and 2001 in order to find 2002 for a specific case.

2 We call "temporal imputation using donors" the method using data 2000 and 2001 from other cases, and data from 2001 in order to find 2002 for a specific case.

Both methods are called "temporal imputation".

3 We call "sector imputation using donors" the method of using auxiliary parameters from 2001 and 2002 from other cases, and the same auxiliary parameter of 2002 in order to find 2002 for a specific case.

Other methods are available, using a combination of any of the existing information, and still avoiding chain imputation.

Figure 2 illustrate that it is virtually possible to fill many missing data if the available responses are well distributed and sufficiently correlated.

Figure 2: The potential of imputation

	Air 2000	Air 2001	Air 2002	Param 2002
Case A	R	R	I	
Case 1	I	I	R	
Case 2	R	I	I	
Case 3			I	R
Cases n to m	R	R	R	R

R= response; I= imputation

Between the numerous imputation techniques the problem is to decide which technique is more reliable and should be used first. Then, for remaining missing data, a less reliable technique could be used and so on, up to we reach the level where a simple extrapolation is preferable to any other imputation technique.

4.3 Temporal imputation

The simplest way to exploit the use of all knowledge about the individual company is to use a prior questionnaire that was completed correctly. This procedure assumes that the temporal correlation (between successive years) of a parameter is on average better than the correlation of the parameter with other parameters (of the same year) within a set of companies from the same or a neighbouring stratum.

Current expenditures should be relatively homogeneous in time, with some sudden changes (e.g. in case of a new legislation, a new investment, the creation of a new permanent job, an extraordinary consultant study etc).

If responses are present for a previous year, an average trend (deflator) between the successive years should be first estimated. There are 3 ways of estimating the trend between the 2 years:

1- Extrapolating the trend of the same variable from the same company between oldest years
2- Using the trend of the same variable from other similar companies
3- Using the trend given by other variables from the same company

For convenience reasons (auxiliary data was not available at the time of computation), we did not use the third method.

4.3.1 Analysis of temporal correlations

The correlation of the totals is shown in table 2[11].

Table 2: correlation between current PAC in 2001 and 2002[12]

	N	Pearson Correlation	Coefficient*	Std. Error	t	Sig
Air	97	,965**	1,261	,020	62,364	,000
Water	229	,876**	,964	,027	35,913	,000
Waste	509	,691**	1,112	,048	23,364	,000
Soil	72	,514**	,161	,055	2,952	,003
Other	144	,947**	1,322	,030	44,642	,000
Total	**590**	**,853****	**1,123**	**,026**	42,445	,000

** Correlation is significant at the 0.01 level (2-tailed).
* Linear Regression through the Origin: 2002 predicted by 2001.

The analysis on the available candidate-donors shows a satisfactory correlation between the 2 years[13]. The coefficient of the linear regression through the origin gives a significant default deflator for the related expenditure (all industrial sectors together)[14].

The method of temporal imputation from 2001 to 2002 current PAC expenditures is thus acceptable.

4.3.2 Serial imputation (forecasting[15] of time series)

The first method of temporal imputation is to use the model of linear evolution between 2000 and 2001, extrapolated to find 2002[16]. Thus, for each company where data was available for 2000 and 2001 but not in 2002:

$2002 = 2001 + (2001\text{-}2000)$[17]

The problem of the method is that we only rely on 2 points for estimating a thirds point, which is statistically not very robust! Outliers should here clearly be let outside the range of imputed values following this method. The limit set to acceptable results is given by the analysis of outliers in the distribution of ln(2002/2001) of the existing

[11] Before to analyse temporal correlation, we eliminated any record with a zero value in 2001 or 2002, because zeros do not allow to calculate a tendency. The SPSS syntax is in annexe 3.
[12] Calculated on original responses. The correlation calculated on edited datasets gives the same result.
[13] Correlations between 1999 and 2001 were comparable (see Kestemont, 2004)
[14] Note that the coefficient indicates a (not weighted) mean growth of current PAC expenditures in the sample: 26% for air, -4% for water, 11% for waste, -84% for soil, 32% for other and 12% for the total. It means that the companies having answered the question in 2001 and 2002, had, on an average, big changes to show. Even if these coefficients are significant, the method should thus be used with precaution: outliers should be excluded from the donor sets.
[15] Or "nowcasting"
[16] see SPSS syntax in annexe 4
[17] The second choice was to use data from 1999 and 2001, then data from 1999 and 2000 etc.

respondents for those 2 years (excluded zeros)[18]. This distribution (annexe 5) gives outliers resp. below and above the limits of –1.29 and 1.53 which corresponds to annual growth of resp. 0.28 and 4.6[19]. We only retained positive results following this method that respects this limits of annual growth.

As shown in table 3, the serial imputation on the basis of 2000-2001, for example, cannot help to impute more than 8% of missing data in our case.

Table 3: serial imputation from 2000-2001 to 2002

Serial imputation (base 2000-2001)	Total	Air	Water	Waste	Soil	Other
Sample units (2002)	1860	1860	1860	1860	1860	1860
Accepted serial imputation (limits=0,28-4,6 annual growth; positive result)	162	127	127	171	132	116
% accepted serial imputation 2000-2001	**9%**	**7%**	**7%**	**9%**	**7%**	**6%**
Responses after accepted serial imputation 2000-2001	1163	704	803	1138	718	642
Adjusted response rate	63%	38%	43%	61%	39%	35%
Additional imputation on non sample units (e.g. 20-49 classe)	110	99	126	217	97	87

Additional results can be derived from answers in 2000-2001, when some companies of 20-49 employees received the annexe on current PAC expenditure. Those additional results of imputation counted in the latest line of table 3.

4.3.3 Temporal imputation using donors

The second method of temporal imputation is based on the principle that the evolution of a variable between 2 given years (here 2001 and 2002) should be similar within each stratum. The responding units of the stratum can be used to estimate this trend. The trend is then applied to estimate the current missing value of a company. The formula is:

$$T_I = \frac{\sum_i Ct_i}{\sum_i Co_i}$$

where T_I is an estimation of the deflator (trend) of current PAC expenditure, for stratum I.

Ct_i is the current PAC expenditure from responding company i in year t.

Co_i is the current PAC expenditure from responding company i in reference year.

The missing values for each non respondent j is calculated within each stratum by:

$$Ct_j = T_I * Co_j$$

[18] We used the outliers from companion cases of the same year of study. Another solution was to use the outliers of the donors themselves. For example, the outliers of ln(2001/1999) are –1.6 and +1.9 (Kestemont, 2004).

[19] $e^{-1.29}$ and $e^{1.53}$

The deflator was calculated and applied at the lowest level possible (on the level of NACE 4 digits), if there were a sufficient number of donors in the considered stratum.

Note that the outliers were not excluded from the donor set, since we have no reason to consider these outliers as erroneous answers (all outliers had been checked by phone with the respondents).

If the total value for donors in a stratum was zero in 2001 or 2002, no deflator was calculated since this would have introduced a bias (forcing a zero deflator or an impossible deflator). In this case, we used the default deflator (coefficient) given in table 2 above for each environmental domain. Note that the default coefficient gives less contrasted results than the coefficient calculated on the level of a stratum.

Annexe 6 shows the related SPSS syntax.

The temporal imputation was computed bilaterally[20] with all available years (1999, 2000, 2001 and 2002). This gave for each missing value, a set of potential candidate values coming from difference sources of imputation. To select the best candidate missing value, we used the following preference order for year 2002:

1) – data imputed from year 2001
2) – data imputed from year 2000
3) –data imputed from year 1999
4) –data from other imputation techniques (see below)

Note that the same imputation technique can be applied backwards, e.g. using responses of 2002 and deflators of neighbouring donors to impute values for 2001[21].

The potential of this method is expressed in the table 4 for imputation on the basis of year 2001, and table 5 for the imputation on the basis of year 1999. It is limited to a maximum 21% of imputation in our case.

Table 4: Potential number of values imputed on the basis of temporal imputation (base 2001, target 2002)

Temporal imputation (base 2001)	Total	Air	Water	Waste	Soil	Other
Sample units (2002)	1860	1860	1860	1860	1860	1860
% answer after recall	1	0	0	1	0	0
% deterministic imputation	**0%**	**11%**	**8%**	**2%**	**11%**	**1%**
Responses after temporal imputation 2001	1299	710	858	1271	727	706
Response rate	70%	38%	46%	68%	39%	38%
Additional imputation on non sample units (e.g. Class 20-49)	*190*	*87*	*164*	*355*	*98*	*134*

**Data collected in 2001 but not in 2002*

[20] The reason why we did not estimate missing values from temporal series, is that the occurrence of long series is much smaller that the occurrence of a answer for one year.

[21] The closer year will be preferred. Note that there is no interference because only real edited answers can be used as donor; there is no chain imputation.

Table 5: Potential number of values imputed on the basis of temporal imputation (base 1999, target 2002)

Temporal imputation (base 1999)	Total	Air	Water	Waste	Soil	Other
Sample units (2002)	1860	1860	1860	1860	1860	1860
Temporal imputation	399	144	285	403	148	184
% temporal imputation (base 1999)	**21%**	**8%**	**15%**	**22%**	**8%**	**10%**
Responses after temporal imputation 1999	1400	721	961	1370	734	710
Response rate	75%	39%	52%	74%	39%	38%
Additional imputation on non sample units (e.g. Class 20-49)	*580*	*122*	*273*	*540*	*137*	*186*

***Data collected in 1999 but not in 2002**

This table only show the potential of the technique, what we would have if we would only use this imputation technique. In practice, as we used various imputation techniques in an order of preference, the full potential of each technique cannot be used. For example, if we impute missing data using base year 2001, less remaining missing data can yet be imputed using basis year 1999.

4.4 Sector imputation using factors

If there is a substantial amount of auxiliary information available at the time when the estimates are to be calculated, an imputation procedure may rely on this auxiliary information. A stratified survey does have a clear advantage whenever imputation is concerned. Any information of responding companies from the same stratum can be used to calculate ratios to be used, e.g. for the breakdown of the totals of the result account of the non-responding company (see Vekeman, 2004). If there are not enough "donors" in a stratum (e.g. on NACE, level 4), it is still possible to use an enlarged stratum (e.g. on NACE level 3). This does stabilize the ratios obtained, but it definitely makes them less suitable, since the 'donor' companies will be less similar to the non-responding one.

4.4.1 Sector correlations

A model based on auxiliary variables can be used to estimate missing values. Kestemont (2004) found that the variables that are the best correlated with current PAC expenditures vary over the sectors. Good candidate explanatory variables were turnover (ESE variable 12110), value added (12150), labour (16110) and environmental taxes (30130).

At the time of writing this report, only preliminary results were available for a set of key variables[22]. The correlation between several key variables and the total current PAC expenditure is shown in table 6.

[22] Unfortunately not the value added neither the total of environmental taxes, which were good correlated for several sectors in 2001. This year (including smaller companies), the value added was

Table 6: Pearson correlation between several variables in the sample (year 2002).

Pearson Correlations

ESE code	Total	Air	Water	Waste	Soil	Other
12110Turnover	,344**	,345**	,294**	,236**	,115**	,189**
13320Wages	,605**	,516**	,489**	**,453****	,290**	**,509****
16110Labour	,549**	,495**	,475**	,408**	**,313****	,424**
30130(01)Env.Tax01	**.625****	**.558****	**.724****	.382**	.108**	.135**

** Correlation is significant at the 0.01 level (2-tailed). 500<N>1000

The best overall correlation (for all environmental domains) is found for environmental taxes (2001)[23]. Air and water expenditures are best correlated to environmental taxes, waste and others with wages, and soil with labour.

However, this depends on the sectors, as shown in table 7.

Table 7. Linear (Pearson) correlations between total current PAC expenditure and several variables in 2002.

Sector Nace 2	N valid	Value added 12110	Total wages 13320	Labour 16110	Environmental Taxes (2001). 30130
10	0	-			
14	3	**.994**	.175	.111	.533
15	143	.428**	**.562****	.531**	.109
16	7	.996**	**.999****	.227	-.209
17	75	.510**	**.613****	.594**	.426**
18	16	**.710****	.626**	.575**	.701*
19	5	**.579**	.561	.562	.710
20	21	**.771****	.667**	.634**	.393
21	33	**.697****	.570**	.573**	.441*
22	41	.033	.030	.030	-.104
23	10	-.142	-.082	-.009	.048
24	96	.710**	.715**	.723**	**.838****
25	73	.571**	**.728****	.678**	-.001
26	43	.790**	**.863****	.840**	.088
27	31	**.911****	.876**	.862**	.545**
28	95	.143	.112	.128	.184
29	59	.471**	.852**	.884**	**.960****
30	0	-			
31	33	.703**	.663**	**.728****	.199
32	14	**.933****	.923**	.928**	.065
33	14	.405	.821**	**.844****	-.129
34	31	.399*	**.484****	.457**	.084

best correlated for sectors 18 (.992**), 24 (.943**), 27 (.829**), 32 (.891**); the environmental taxes were better correlated for sectors 29 (.971**), 35 (.903**) and 41 (1.000** for 31 respondents … which means that the current expenditures of this sector are only taxes) (Kestemont, 2004).

[23] Note that the correlation for turnover was better (0.547) in 2001, when medium-sized (20-49) enterprises were included in the sample. The variable on environmental taxes for year 2002 was not available when we did this analysis.

35	6	**.976****	.796	.828*	.184
36	51	.239	.422**	.371**	**.474****
40	1	-			
41	2	-1.000**	1.000**	1.000**	1.000**
All industries	**1000**	**.344****	**.605****	**.549****	**.625****

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).
We put in bold the best choice, which is given by the highest and/or most significant correlation factor.

This analysis recalls that the method of imputation using factor should be carefully limited to those sectors and variables that shows a sufficient correlation. For example, this method is not suitable for sectors 10, 22, 23, 28, 30, 40, and 41[24].

4.4.2 Default factors

The default factor for each sector or stratum is determined by the formula (Kestemont, 2002):

$$F_I = \frac{\sum_i C_i}{\sum_i E_i}$$

where F is the default factor for stratum i
C_i is the current PAC expenditure (for domain considered) by respondent i.
E_i is the value, for company I, of the explanatory variable (e.g. turnover, 12110).

The missing values for non-respondents[25] j are calculated within each stratum (or sector) by formula:

$$C_j = F_I * E_j$$

The default factors are calculated at the lowest level possible (NACE 4), provided there are a sufficient number of donors. If no factor can be calculated at this level, a factor is taken at the level NACE 3, then at the level NACE 2.

The syntax is shown in annexe 7. As an illustration, the default factors with turnover at level NACE2 are displayed in annexe 8. They show similar great order than for

[24] The results are slightly different from the results found on data 2001. No definitive conclusion can be derived yet over time, because the sample size varied (in 2001, companies 20-49 were included). For example, in 2001, no correlation could be found for sectors 10, 16, 25, 30, 36 and 41. Sector 22 was correlated with labour, sectors 23 and 28 with value added (not available here), and sector 40 with almost all independent variables.
[25] Explanatory variables are obligatory since a long time. Moreover, an imputation on basis of administrative data is possible, which explains that those explanatory variables are generally available. If such key variables are not present in a dataset, for example if a company stopped its activity, the extrapolation (weight factors of the sample) are changed accordingly.

other years, but with some notable changes. This result confirms what we have seen before (Kestemont, 2004): that the turnover is not suitable for all pairs sector/domain and that this factor is not constant over time. It suggests the interest to make this survey annually.

The potential number of data that can be imputed following this method is shown in table 8.

Table 8: Potential number of data that can be imputed from factor imputation (base 2002)

Factor imputation (base 2002)	Total	Air	Water	Waste	Soil	Other
Sample units (2002)	1860	1860	1860	1860	1860	1860
Factor imputation	833	1261	1166	877	1265	1318
% factor imputation	**45%**	**68%**	**63%**	**47%**	**68%**	**71%**
Responses after factor imputation	1834	1838	1842	1844	1851	1844
Response rate	99%	99%	99%	99%	100%	99%
Additional imputation on non sample units (e.g. Class 20-49)	*4429*	*23520*	*23276*	*11021*	*11072*	*12210*

**PAC expenditure data not collected in 2002*

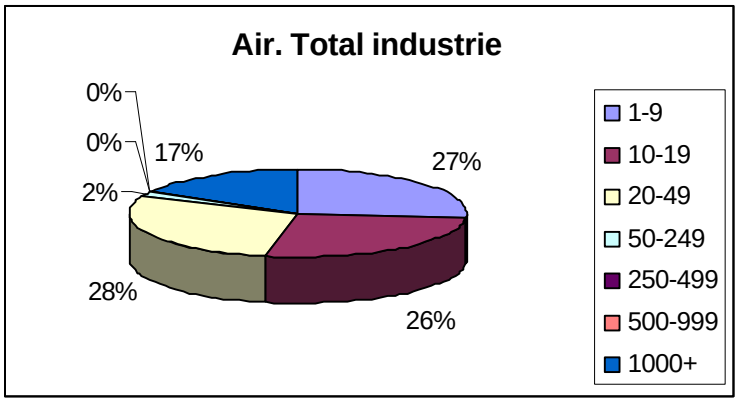
The power of the method is to be able to estimate current PAC expenditures for strata that where never surveyed over this variable, as shown in last line of table above. Small companies are concerned, as well as companies from non industrial sectors.

4.5 Trend imputation

4.5.1 Estimation for smallest companies

The method using factors can, as a rough estimate, be applied to find out possible results for smallest companies. It assumes that the factor would be constant over the size of the companies, which is far from certain in all sectors, as seen in Kestemont (2004), but plausible. For date 2001, companies of 20 to 49 employees were present in the sample. Some companies of 1 to 9 employees, but with a turnover greater than 4.96 Millions euros where also surveyed. With the available data, it was estimated that the contribution of smallest companies was far from negligible in the total current expenditures. Figure 3 gives the repartition of current expenditures for air by size of companies in 2001. In 2001, about 3/4 of current PAC expenditure would be done by companies of 1-49 persons (a class not surveyed in 2002)!

Figure 3: Current expenditures for air by size of companies in 2001.



(Source: Kestemont, 2004)

The method we used to estimate data for smallest companies in 2002 was in 2 steps. First, we used the trend of the explanatory variable identified in the analysis just above (e.g. turnover) to impute, for each individual company responding in 2001, a value of current expenditure.

The missing values for non-respondents j are calculated by formula:

$$C_j = C^\circ_j * E_j / E^\circ_j$$

C_j is the current PAC expenditure (for domain considered) by respondent j for missing year.
C°_j is the current PAC expenditure (for domain considered) by respondent j for basic year.
E°_j is the value, for company j, of the explanatory variable (e.g. turnover, 12110) for reference year.
E_j is the value, for company j, of the explanatory variable (e.g. turnover, 12110) for present year.

The difference with the method described before is that we were less conservative in using trend than in using factors, because one of the parameters used is an answer from the same company for a previous year[26]. The trend of turnover was used on the sectors where the correlation was the best for this variable, but for all remaining sectors without exception, the trend of employment was used as default trend to be applied on existing 2001 responses.

The syntax for trend imputation is displayed in annexe 8. Table 9 and 10 show the potential for imputation following this method on our sample 2002 . The last row concerns units that received current PAC expenditure questionnaire in 1999 or 2001 but not in 2002, but being part of the sample for the general survey in 2002[27]. For

[26] Moreover, for companies 20-49 employees, there is no bias in the "non-responses" of 2002 as compared with 2001 … because those companies were not surveyed in 2002.
[27] The corrected reference sample consists of 3117 companies (a large extract from the 1999 sample).

those units, it is possible to impute values. Respondents of 20-49 persons are typical target for this method.

Table 9: Potential for imputation using trend of auxiliary variables 2002/1999

Trend imputation (2002/1999)	Total	Air	Water	Waste	Soil	Other
Sample units (2002)	1860	1860	1860	1860	1860	1860
Trend imputation	312	245	282	321	245	236
% trend imputation	**17%**	**13%**	**15%**	**17%**	**13%**	**13%**
Responses after trend imputation	1313	822	958	1288	831	762
Response rate	71%	44%	52%	69%	45%	41%
*Additional imputation on non sample units (e.g. Class 20-49)**	*638*	*337*	*419*	*616*	*352*	*314*

**Data collected in 1999 but not in 2002.*

Table 10: Potential for imputation using trend of auxiliary variables 2001-2002

Trend imputation (2002/2001)	Total	Air	Water	Waste	Soil	Other
Sample units (2002)	1860	1860	1860	1860	1860	1860
Trend imputation	334	301	277	342	298	284
% trend imputation	**18%**	**16%**	**15%**	**18%**	**16%**	**15%**
Responses after trend imputation	1335	878	953	1309	884	810
Response rate	72%	47%	51%	70%	48%	44%
*Additional imputation on non sample units (e.g. Class 20-49)**	*241*	*256*	*294*	*419*	*256*	*241*
Responses on GD, after trend imputation	1576	1134	1247	1728	1140	1051

**Data collected in 2001 but not in 2002.*

For remaining missing values, we simply used the factors found for greater companies, as explained above.

4.6 Stratum imputation

The less desired method is very close to the extrapolation itself. Remaining missing values are imputed by taking the mean of the responses of the same stratum (the stratum being defined as a combination of one of the 5 SBS size classes and each NACE4 sector). This method avoid to be obliged to correct the weight affected to each company of the sample. It allows retaining the same weight for extrapolating the results of all parameters of the overall SBS survey.

If there are no donors in a given stratum, we take the mean of the same size class, but for nace 3 digits, then nace 2 digits.

If there are still no donors in a stratum (e.g. in strata of small companies), we just copy the eventual response given by a company for previous years (first 2001, then 2000 etc)[28].

The potential result of this process is given in table 11. The syntax is in annexe 9.

[28] Remember that the temporal methods were not available for all sectors. Few companies are still concerned with this step.

Table 11: Potential of stratum imputation

Stratum imputation	Total	Air	Water	Waste	Soil	Other
Sample units (2002)	1860	1860	1860	1860	1860	1860
Stratum imputation	837	1239	1151	873	1229	1281
% stratum imputation	**45%**	**67%**	**62%**	**47%**	**66%**	**69%**
Responses after stratum imputation	1838	1816	1827	1840	1815	1807
Response rate	99%	98%	98%	99%	98%	97%
Additional imputation on non sample units (e.g. Class 20-49)	*989*	*618*	*873*	*1015*	*786*	*740*

The table shows that even this method can not impute all missing data: for the total, 22 companies of the sample are not imputable following this method, because their stratum (even at level NACE 2) had no donor.

4.7 Potential of different methods

Table 12 compares the potential of different methods, given the initial response rate after recall.

Table 12: Potential of different imputation methods

Potential of different methods	Total	Air	Water	Waste	Soil	Other
% answer after recall	54%	20%	28%	50%	20%	28%
% deterministic imputation	0%	11%	8%	2%	11%	1%
% accepted serial imputation 2000-2001	9%	7%	7%	9%	7%	6%
% temporal imputation (base 1999)	21%	8%	15%	22%	8%	10%
% factor imputation	45%	68%	63%	47%	68%	71%
% trend imputation	18%	16%	15%	18%	16%	15%
% stratum imputation	45%	67%	62%	47%	66%	69%

Of course, these methods are not cumulative because a same data can be imputed from different methods. They can be used in cascade of preference.

4.8 Imputation following environmental taxes 2001

The latest method available is to use the variable “environmental taxes” as a proxy of current expenditure for the total and “others”. This method is quite imperfect since we know (Kestemont, 2000) that this parameter can be an overestimation of current PAC expenditures, since it contains non-affected taxes, or it can be an underestimation because other kinds of expenditures are not taken into account (salaries etc). However, “environmental taxes” is well correlated with current PAC expenditure in many sectors where the other parameters are not well correlated (e.g. sectors 19, 29, 30, 35, 40). Taxes 2001 is available for 15166 companies (1843 companies of our sample = 99%), which makes it a good candidate to fill in the ultimate gaps.

4.9 Second deterministic imputation

The various methods described above are used, one after the other, in the order of preference. It is possible that one domain have been imputed by one method, another domain by another method. It is then possible to impute deterministically the results, trying to logically find missing totals or unique gaps.

4.10 Decision tree: cascade imputation

Now that we have studied the potential of each imputation technique, we can apply these techniques one after the other, beginning with the most desired one and ending, for the latest still missing values, with the less desirable one.

The following steps where thus accomplished, in the order, and regarding available data from 1999 to 2002 (see syntax in annexe 10):

1) Edition
2) Deterministic imputation
3) Serial imputation from 2000-2001 to 2002
4) Serial imputation from 1999-2001 to 2002
5) Serial imputation from 1999-2000 to 2002
6) Temporal imputation using donors from 2001 to 2002[29]
7) Temporal imputation using donors from 2000 to 2002
8) Temporal imputation using donors from 1999 to 2002
9) Trend imputation from 2001 to 2002
10) Trend imputation from 1999 to 2002[30]
11) Sector imputation using factors (turnover, employment, wages, env. taxes)
12) Stratum imputation & imputing value of the nearest previous year
13) Imputation of environmental taxes on total current PAC exp
14) Secondary deterministic imputation

This following steps allowed to reach a 100% imputation, as illustrated in table 13.

Table 13: Answer and imputation rates.

Statistics after cascade imputation	Total	Air	Water	Waste	Soil	Other
Questionnaires recorded	1860	1860	1860	1860	1860	1860
Answers after recall	1001	369	526	937	380	514
% answer after recall	54%	20%	28%	50%	20%	28%
Cascade	853	1484	1329	916	1475	1337
% imputed	46%	80%	71%	49%	79%	72%
Answers after cascade	1854	1853	1855	1853	1855	1851
% response after imputation	100%	100%	100%	100%	100%	100%
Additional imputation on non sample units (e.g. 20-49 classe)	34721	30400	30036	28936	29954	28915

5. Extrapolation

No that we have imputed 100% of the missing values of the sample, with hypothesis that are all better that the extrapolation itself, we can extrapolate there results using the weights given to all case of the sample[31].

The syntax is given in annexe 11.

[29] Here, we suppose that the trend for a given domain in a given sector is more explicative than the trend of an economic auxiliary variable of the same company. This method is therefore used first.
[30] No auxiliary data from 2000 where available at the time of calculation
[31] this weight was corrected taking into account the non valid companies (closed or impossible to reach companies etc).

6. Results

The results are given in a separate file[32] and are summarized in annexe 12. The result given as total for the all economy suffers from the method of estimating the current expenditures for smaller companies or for non-industrial sectors, which is of a lowest quality[33].

Concerning investments expenditures, we took as a provisional answer since all answers from respondents are not yet available.

[32] The detail per size of company is given under statistical secret. The reason of giving this detail is that, following our method used for current PAC expenditure, we can consider that the most reliable results are given for class 50+ persons.
[33] Both of the latest details were not foreseen in the contract.

Références

Kestemont B. (2004). *L'enquête sur les dépenses courantes environnementales des industries belges en 2001, méthodes et résultats*. (draft report, not published, but available on demand). Statbel, Brussels, 47 pp.

Kestemont B. (2002). *Les dépenses de protection de l'environnement par les industries en Belgique - Current environmental expenditure by the Belgian industry 1999*. Statistics Belgium Working Paper N°7. Statbel. Brussels, 102 pp.

Kestemont B. (2001), « Factors affecting quality of statistics on environmental expenditures by companies in Belgium ». *International Conference on Quality in Official Statistics*, Stockholm, May 14-15. in Kestemont (2002) pp. 84-91.

Kestemont B. (2000). *Dépenses environnementales des entreprises en Belgique*. Statistics Belgium Working Paper. Statbel. Brussels, 44 pp.

UN-ECE (2000). *Glossary of terms on statistical data editing*. Conference of European Statisticians. Methodological material. United Nations. Geneva. 12pp.

Vekeman G. (2004). Editing and imputation. Rapport de stage d'actuaire. Statbel. Brussels, 54 pp.

Annexe 1 : Questionnaires

Cadre CE - Dépenses courantes consacrées à la protection de l'environnement

Numéro d'identification de l'entreprise :
Nom de la personne à contacter pour ce cadre :
Mme/M ..... Téléphone: ..... Fax: .....

Domaine de pollution	Code CEPA	Valeur EUR
	1 EPACD	2 VALCE
Total des dépenses courantes consacrées à la protection de l'environnement [a] dont:		.....
• Protection de l'air ambiant et du climat [b]	CE.01.00.00	.....
• Gestion des eaux usées [c]	CE.02.00.00	.....
• Gestion des déchets [d]	CE.03.00.00	.....
• Protection des sols et des eaux souterraines [e]	CE.04.00.00	.....
• Autres [f]	CE.09.00.00	.....

(a) Celles-ci comprennent les dépenses internes (salaires et autres), et externes (sous -traitance, redevances). Sont exclues les mesu visant à protéger le lieu de travail.
(b) Réduction des émissions ou de la concentration de polluants dans l'air, y compris les substances affectant la couche d'ozone c le climat.
(c) Collecte, transport, élimination et épuration des eaux usées. Réduction des rejets d'eaux usées. Eaux de refroidissement.
(d) Réduction de la génération de déchets d'entreprise; collecte, transport, traitement et élimination des déchets; recyclage s'il vis principalement à protéger l'environnement.
(e) Sols et eaux souterraines. Inclut la dé contamination des eaux de surface.
(f) Protection de l'environnement contre les bruits et des vibrations, protection de la biodiversité et des paysages, protection contr les radiations et autres indivisibles (études d'impact, coordination).

Votre correspondant: Madame M. Sampièri (tel 02/548 65 80)

Kader CE - Lopende uitgaven voor milieubescherming

Identificatienummer van de onderneming:
Naam van de contactpersoon voor dit kader:
Mevrouw/De heer ..... Telefoon: ..... Fax: .....

Aard van verontreiniging	CEPA- code	Waarde EUR
	1 EPACD	2 VALCE
Totale lopende uitgaven voor milieubescherming [a] waarvan:		....
• Lucht- en klimaatbescherming [b]	CE.01.00.00	....
• Afvalwaterbeheer [c]	CE.02.00.00	....
• Afvalbeheer [d]	CE.03.00.00	....
• Bodem- en grondwaterbescherming [e]	CE.04.00.00	....
• Andere [f]	CE.09.00.00	....

(a) Hiertoe behoren: interne uitgaven (lonen en andere), externe uitgaven (onderaanneming, heffingen). Exclusief maatregelen inza de veiligheid en bescherming van de gezondheid op de werkplek.
(b) Reductie van emissies of concentratie va n luchtverontreinigende stoffen, inclusief stoffen die de ozonlaag of het klima aantasten.
(c) Verzamelen, vervoeren eliminatie en verwerking van afvalwater. Vermindering van afvalwaterproductie. Koelwater.
(d) Vermindering van de productie van bedrijf safval; ophaling, transport, verwerking en verwijdering van afval; recyclage indien hoofdzakelijk gericht op milieubescherming.
(e) Bodem en grondwater. Inclusief oppervlaktewaterontsmetting.
(f) Bescherming van het milieu tegen geluid en trilling, b escherming van biodiversiteit en landschappen, bescherming tegen strali en andere investeringen die niet opgesplitst kunnen worden (impact studies, coördinatie).

Uw contactpersoon: Mevr. R. Braekman (tel: 02/ 548 63 36)

Annexe 2 : SPSS 11 syntax for deterministic imputation

```
*****************************************************************************************
*DETERMINISTIC IMPUTATION
*by B. Kestemont, Statistics Belgium, June 2004
*****************************************************************************************

*****************************************************************************************
*MACRO CHKSUM
*****************************************************************************************

DEFINE !CHKSUM (R = !TOKENS (1)
                /S = !TOKENS (1)
                /A = !TOKENS (1)
                /B = !TOKENS (1)
                /C = !TOKENS (1)
                /D = !CMDEND).

*R = prefix of corrected variable (ex: e). Should be one character.
*S = prefix of the variable to be corrected (ex: c).
*A = year of missing variable to be estimated.
*B = oldest reference year used for the imputation.
*C = recent reference year for the imputation
*D = domains. Should be one character.

*create new ed variable, result of the CHKSUM
*first for the total

COMPUTE !CONCAT(!R, 't', !A, !B, !C)  = !CONCAT(!S, 't', !A, !B, !C)  .
EXECUTE .

*then the domains

 !DO !I !IN (!D).
COMPUTE !CONCAT(!R, !I, !A, !B, !C)  = !CONCAT(!S, !I, !A, !B, !C)  .
EXECUTE .

 !DOEND.


*CHKSUM.

 !DO !I !IN (!D).
*a company declaring zero for the total should have zero for the details
  -USE ALL.
  -IF (MISSING(!CONCAT(!R, !I, !A, !B, !C)) & !CONCAT(!R, 't', !A, !B, !C)=0) !CONCAT(!R, !I,
!A, !B, !C) = 0 .
  -EXECUTE .
*a value in “other” is supposed to cover all remaining, not detailed, expenditures.
*a zero is attributed to all missing domains.

  -USE ALL.
  -IF (MISSING(!CONCAT(!R, !I, !A, !B, !C)) & ~ MISSING(!CONCAT(!R, '9', !A, !B, !C)))
!CONCAT(!R, !I, !A, !B, !C) = 0 .
  -EXECUTE .
 !DOEND.
```

```
*a company which answered for all domains, but not for “others” is supposed to have.
*classified all of its expenditures: other is then estimated to be zero.

USE ALL.
IF (MISSING(!CONCAT(!R, '9', !A, !B, !C)) &  ~ MISSING(!CONCAT(!R, '1', !A, !B, !C)) & ~
MISSING(!CONCAT(!R, '2', !A, !B, !C)) &  ~ MISSING(!CONCAT(!R, '3', !A, !B, !C)) & ~
MISSING(!CONCAT(!R, '4', !A, !B, !C))) !CONCAT(!R, '9', !A, !B, !C) = 0 .
EXECUTE .

*and the total is calculated accordingly
USE ALL.
IF (~ MISSING(!CONCAT(!R, '9', !A, !B, !C)) &  ~ MISSING(!CONCAT(!R, '1', !A, !B, !C)) & ~
MISSING(!CONCAT(!R, '2', !A, !B, !C)) &
  ~MISSING(!CONCAT(!R, '3', !A, !B, !C)) &  ~ MISSING(!CONCAT(!R, '4', !A, !B, !C)))
!CONCAT(!R, 't', !A, !B, !C) = !CONCAT(!R, '1', !A, !B, !C) + !CONCAT(!R, '2', !A, !B, !C) +
!CONCAT(!R, '3', !A, !B, !C) + !CONCAT(!R, '4', !A, !B, !C) + !CONCAT(!R, '9', !A, !B, !C) .
EXECUTE .

!ENDDEFINE.


*****************************************************************************************************************
*****************************************
*PROGRAMS
*****************************************

**********************************
*on CE2002calc.sav
**********************************
*runs the macro for different years A, phase R, and domains D.
*warning: the total lenght of R+S+A+B+C+D should not be more than 8 characters.
*warning: arguments should begin with character.

SORT CASES BY
  na2 (A) .

*first CHKSUM. No need to specify a reference year B or C.

!CHKSUM R=e S=c A=02 D= 1 2 3 4 9.

*secundary CHKSUM, after temporal imputation from year B
*si R and S sont les mêmes => j'écrase les anciennes valeurs par les valeurs éditées.

!CHKSUM R=s S= s A=02 B=01 D= 1 2 3 4 9.

*R = nom de la variable éditée (ex: e).
*S = nom de la variable à éditer (ex: c).
*A = année de calcul.
*B = année de référence ancienne.
*C = année de référence récente.
*D = domaines.

*secundary CHKSUM, after serial imputation from years B and C.
*si R and S sont les mêmes => j'écrase les anciennes valeurs par les valeurs éditées.

!CHKSUM R=x S= x A=02 B=99 C=01 D= 1 2 3 4 9.

*tertiary CHKSUM, after sectoral imputation from same year
```

!CHKSUM R=s S= s A=02 B=02 D= 1 2 3 4 9.

Annexe 3: Temporal correlations between 2002 and 2001 current PAC expenditures

SPSS 11 syntax

```
*********************************************************************************
*ANALYSIS OF TEMPORAL CORRELATIONS
*by B. Kestemont, Statistics Belgium, June 2004
*********************************************************************************

*********************************************************************************
*MACRO !CORTEMP
*********************************************************************************

DEFINE !CORTEMP (D = !CMDEND).

 !DO !I !IN (!D).

-USE ALL.
-COMPUTE filter_$=(!CONCAT('ce', !I, '_cor') ~= 0 & !CONCAT('ce', !I, '_01') ~= 0).
*-VARIABLE LABEL filter_$ 'total_co ~= 0 & total_01 ~= 0 (FILTER)'.
-VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
-FORMAT filter_$ (f1.0).
-FILTER BY filter_$.
-EXECUTE .

-CORRELATIONS
  /VARIABLES=!CONCAT('ce', !I, '_cor') !CONCAT('ce', !I, '_01')
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .

 !DOEND.


!ENDDEFINE.


*********************************************************************************************************************
***
*******************************************
*ICI COMMENCE LE PROGRAMME
*******************************************

***********************************
*sur CE2002calc.sav
***********************************
*lance la macro pour différents domaines D

SORT CASES BY
  na2 (A) .

!CORTEMP D= 01 02 03 04 09.
```

Annexe 4: SPSS 11 syntax for serial imputation

```
*******************************************************************************
*SERIAL IMPUTATION
*by B. Kestemont, Statistics Belgium, June 2004
*******************************************************************************

*******************************************************************************
*MACRO SERIAL
*******************************************************************************

DEFINE !SERIAL (R = !TOKENS (1)
                /S = !TOKENS (1)
                /A = !TOKENS (1)
                /B = !TOKENS (1)
                /C = !TOKENS (1)
                /D = !CMDEND).

*R = prefix of extrap variable (ex: x). Should be one character.
*S = prefix of the variable to be edited (ex: c).
*A = year of missing variable to be estimated.
*B = oldest reference year used for the imputation.
*C = recent reference year for the imputation
*D = domains. Should be one character.

USE ALL.

*create new  variable
*temporal extrapolation, weighted by the distance between donor years.
*note that this way of whriting the formula avoid the problems linked to zeros.
*then delete the eventual outliers, arbitrary set to annual evolution greater than 50% in both
directions.
*then delete the possible negative results (because we used absolute

 !DO !I !IN (!D).
        COMPUTE !CONCAT(!R, !I, !A, !B, !C)  = !CONCAT(!S, !I, !A)  .
        EXECUTE .
        USE ALL.
        IF( MISSING(!CONCAT(!S, !I, !A)) &  ~ MISSING(!CONCAT(!S, !I, !B)) & ~
MISSING(!CONCAT(!S, !I, !C)))
        !CONCAT(!R, !I, !A, !B, !C) = !CONCAT(!S, !I, !C)+(!CONCAT(!S, !I, !C)-!CONCAT(!S,
!I, !B))/(!C-!B) .
        EXECUTE .

        DO IF (!CONCAT(!R, !I, !A, !B, !C) < 0 | (MISSING(!CONCAT(!S, !I, !A)) &  ~
MISSING(!CONCAT(!S, !I, !B)) & ~ MISSING(!CONCAT(!S, !I, !C))
        & 0.28*!CONCAT(!S, !I, !C)>!CONCAT(!R, !I, !A, !B, !C)  -!CONCAT(!S, !I,
!C)>4.61*!CONCAT(!S, !I, !C))) .
        RECODE
          !CONCAT(!R, !I, !A, !B, !C)  (ELSE=SYSMIS)  .



        END IF .
        EXECUTE .


 !DOEND.
```

```
*correct total = sum of domains if all available.
IF (~MISSING (!CONCAT('x1', !A, !B, !C)) & ~MISSING (!CONCAT('x2', !A, !B, !C)) &
~MISSING (!CONCAT('x3', !A, !B, !C)) & ~MISSING (!CONCAT('x4', !A, !B, !C)) & ~MISSING
(!CONCAT('x9', !A, !B, !C)))
!CONCAT('xt', !A, !B, !C) = !CONCAT('x1', !A, !B, !C) + !CONCAT('x2', !A, !B, !C) +
!CONCAT('x3', !A, !B, !C) + !CONCAT('x4', !A, !B, !C) + !CONCAT('x9', !A, !B, !C)  .
* de cette manière, l'imputation sur total n'est retenue qu'à défaut d'autre chose.
*remarquons que le total initial est ici éventuellement corrigé.
*Il s'agit aussi d'une partie de l''édition, ce qui explique que l'édition elle-même n'ajoute plus
grand chose.



!ENDDEFINE.


*****************************************************************************************************
***
*****************************************
*PROGRAM
*****************************************

**********************************
*on CE2002calc.sav
**********************************
*runs the macro for different years A, phase R, totals T and domains D.
*warning: the total lenght of R+A+B+C should not count more than 8 characters.
*warning: arguments should begin with character.

SORT CASES BY
  na202 (A) .


!SERIAL R=x S=e A=02 B= 00 C=01 D= 1 2 3 4 9 t.
!SERIAL R=x S=e A=02 B= 99 C=01 D= 1 2 3 4 9 t.
!SERIAL R=x S=e A=02 B= 99 C=00 D= 1 2 3 4 9 t.
```

Annexe 5: Distribution of annual growth

Distribution of ln(total current PAC exp 2002 / total current PAC exp 2002)

```
LN Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    48,00 Extremes    (=<-1,29)
     3,00      -11 .  &
     3,00      -10 .  0&
     5,00       -9 .  9&
     4,00       -8 .  2&
     7,00       -7 .  26&
     6,00       -6 .  0&&
     9,00       -5 .  059&
    14,00       -4 .  0124&&
    19,00       -3 .  112345779&
    27,00       -2 .  00234577889&
    51,00       -1 .  0001122233344455667889
    56,00       -0 .  011122233344444556678899 9
    70,00        0 .  000000001111222233334455566688889&
    41,00        1 .  00001123446678899
    37,00        2 .  00112233334555688&
    26,00        3 .  02445567999&
    18,00        4 .  001457&&
    15,00        5 .  02367&
    15,00        6 .  25668&&
     5,00        7 .  6&
    15,00        8 .  02248&&
    12,00        9 .  4699&&
    12,00       10 .  1122&
     3,00       11 .  4&
    11,00       12 .  14&&
     3,00       13 .  &
     4,00       14 .  &&
    51,00 Extremes    (>=1,53)

 Stem width:         0
 Each leaf:       2 case(s)


 & denotes fractional leaves.
```

Note that zero values in 2001 or 2002 are excluded from this analysis.
The analysis sets outliers to –1.29 and 1.53 which corresponds to annual growth of resp. 0.28 and 4.61.

Annexe 6: SPSS 11 syntax for temporal imputation using donors

```
****************************************
*Temporal imputation.
*by B. Kestemont, Statistics Belgium.
****************************************
*Annual growth is estimated by donors.
*na4 = nace à 4 digits.

*R=prefix of variable.
*A= target year.
*B= basic year.
*D=parameter.


*********************************************************************************************
*MACRO IMPTEMP.
*********************************************************************************************

DEFINE !IMPTEMP (R = !TOKENS (1)
                 /A = !TOKENS (1)
                 /B = !TOKENS (1)
                 /D = !CMDEND).

 !DO !I !IN (!D).


*select donors (responses in 2001 AND 2002).
USE ALL.
COMPUTE filter_$=( ~ MISSING(!CONCAT(!R, !I, !B)) &  ~ MISSING(!CONCAT(!R, !I, !A))).
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

SORT CASES BY
  !CONCAT('na4',!A) (A) .

AGGREGATE
  /OUTFILE= !CONCAT(!R,!I,!B,!A)
  /BREAK=!CONCAT('na4',!A)
  /!CONCAT(!R, !I, !B, 's') = SUM(!CONCAT(!R, !I, !B)) /!CONCAT(!R, !I, !A, 's') =
SUM(!CONCAT(!R, !I, !A)).

USE ALL.

MATCH FILES /FILE=*
 /TABLE=!CONCAT(!R,!I,!B,!A)
 /BY !CONCAT('na4',!A).
EXECUTE.

*calculation of deflator NACE4, where possible (not zero values).

IF (!CONCAT(!R, !I, !B, 's') ~= 0 & !CONCAT(!R, !I, !A, 's')~=0) !CONCAT('ev', !I, !A, !B) =
!CONCAT(!R, !I, !A, 's')/!CONCAT(!R, !I, !B, 's') .
EXECUTE .
```

```
*je ne tiens pas compte des cas scabreux où l'un des deux membres égale zéro
*dans ce cas et tout les autres, utilisation du coefficient résiduel par défaut pour le domaine considéré

USE ALL.
COMPUTE filter_$=(MISSING(!CONCAT('ev', !I, !A, !B)) & ~ MISSING(!CONCAT(!R, !I, !B)) & ~ MISSING(!CONCAT(!R, !I, !A))).
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

SORT CASES BY
  filter_$ (A) .

AGGREGATE
  /OUTFILE=!CONCAT('def',!R,!I,!B,!A)
  /BREAK=filter_$
  /!CONCAT(!R, !I, !B, 'd') = SUM(!CONCAT(!R, !I, !B, 's')) /!CONCAT(!R, !I, !A, 'd') =
  SUM(!CONCAT(!R, !I, !A, 's')).

USE ALL.

MATCH FILES /FILE=*
 /TABLE=!CONCAT('def',!R,!I,!B,!A)
 /BY filter_$.
EXECUTE.

*calculation of default deflator, where possible.

IF (MISSING(!CONCAT('ev', !I, !A, !B)) & !CONCAT(!R, !I, !B, 'd') ~= 0 & !CONCAT(!R, !I, !A, 'd')~=0) !CONCAT('ev', !I, !A, !B) = !CONCAT(!R, !I, !A, 'd')/!CONCAT(!R, !I, !B, 'd') .
EXECUTE .


USE ALL.
SORT CASES BY
  !CONCAT('na4',!A) (A) .



*calculation of values following model.

IF (MISSING(!CONCAT(!R, !I, !A)) &  ~ MISSING(!CONCAT('ev', !I, !A, !B)) &  ~ MISSING(!CONCAT(!R, !I, !B))) !CONCAT('es', !I,'ev') =
  !CONCAT(!R, !I, !B)*!CONCAT('ev', !I, !A, !B) .
EXECUTE .

* je garde les valeurs nulles initiales car celà est prévu dans le déflateur lui-même
*pour les donneurs, les valeurs non nulles étant, elles, accues en proportion

*initialisation de la variable d'après réponses initiales.
USE ALL.
COMPUTE !CONCAT('s', !I, !A, !B) = !CONCAT(!R, !I, !A) .
EXECUTE .

*imputation, seulement sur valeurs initiales manquantes.
IF (MISSING(!CONCAT(!R, !I, !A))) !CONCAT('s', !I, !A, !B) = !CONCAT('es', !I,'ev') .
EXECUTE .
```

```
* la variable résultat est de la forme s10299.
*il s'agit d'une proposition de valeur d'imputation pour données manquantes.
*tout le reste peut être effacé.


 !DOEND.

USE ALL.
*corriger le total = somme des domaines s'ils sont disponibles.
IF (~MISSING (!CONCAT('s1', !A, !B)) & ~MISSING (!CONCAT('s2', !A, !B)) & ~MISSING
(!CONCAT('s3', !A, !B)) & ~MISSING (!CONCAT('s4', !A, !B)) & ~MISSING (!CONCAT('s9',
!A, !B)))
!CONCAT('st', !A, !B) = !CONCAT('s1', !A, !B) + !CONCAT('s2', !A, !B) + !CONCAT('s3', !A,
!B) + !CONCAT('s4', !A, !B) + !CONCAT('s9', !A, !B)  .
* de cette manière, l'imputation sur total n'est retenue qu'à défaut d'autre chose.
*remarquons que le total initial est ici éventuellement corrigé.
*il s'agit d'une édition, ce qui limite l'apport éventuel de l'édition supplémentaire.



!ENDDEFINE.

**********************************************************************************
*FIN DE MACRO IMPTEMP.
**********************************************************************************

**********************************************************************************************************************
***
*****************************************
*ICI COMMENCE LE PROGRAMME
*****************************************

***********************************
*sur CE2002calc.sav
***********************************
*lance la macro  pour les différents domaines D, y compris le total
*je choisis de partir des réponses éditées
*Après, il faut éliminer toutes les variables intermédiaires sauf s1 à st.

SORT CASES BY
  na202 (A) .

*à partir de l'année 2001.
!IMPTEMP R= e A= 02 B= 01 D= 1 2 3 4 9 t.
```

Annexe 7. SPSS 11 syntax for sector imputation using factors

```
****************************************
*Factor imputation.
*by B. Kestemont, Statistics Belgium.
****************************************
*Default factors are estimated by donors.
*na4 = nace à 4 digits.
*na2 = nace à 2 digits.

*R= prefix of source variable.
*A= year.
*D= environmental domain.

*calculated on GD, after merging all available responses


*************************************************************************************
*MACRO FACTOR.
*************************************************************************************

DEFINE !FACTOR (R = !TOKENS (1)
                    /A = !TOKENS (1)
                    /D = !CMDEND).



*initialisation of estimate .
 !DO !I !IN (!D).
COMPUTE !CONCAT('gd',!I,!A) = !CONCAT(!R,!I,!A) .
EXECUTE .
 !DOEND.



*do first for nace4, then nace3, then nace2.
 !DO !J = 4 !TO 2 !BY -1.
 !DO !I !IN (!D).


*select donors (responses in both variables).
USE ALL.
COMPUTE filter_$=(~ MISSING(!CONCAT(!R, !I, !A)) &  ~ MISSING(!CONCAT('v12110',
!A))).
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

SORT CASES BY
  !CONCAT('na',!J,!A) (A) .

AGGREGATE
  /OUTFILE= !CONCAT ('aggr',!I, !J, !A,'.sav')
  /BREAK=!CONCAT('na',!J,!A)
```

```
  /!CONCAT('exp',!i,!j,!A) = SUM(!CONCAT(!R, !I, !A)) /v12110_1 = SUM(!CONCAT('v12110',
!A))
  /!CONCAT('emp',!i,!j,!A) = SUM(!CONCAT('v16110', !A))
  /!CONCAT('sal',!i,!j,!A) = SUM(!CONCAT('v13320', !A))
  /!CONCAT('tax',!i,!j,!A) = SUM(v3013001)
  /N_BREAK=N.

USE ALL.

MATCH FILES /FILE=*
 /TABLE= !CONCAT ('aggr',!I, !J, !A,'.sav')
 /BY !CONCAT('na',!J,!A).
EXECUTE.

*calculation of factors NACE, where possible (obs: zero values of !CONCAT('exp',!i,!j,!A)
included).
*obs: if denominator is zero => missing value.
COMPUTE !CONCAT('fgd',!I,'ca', !J) = !CONCAT('exp',!i,!j,!A)/v12110_1*1000000 .
COMPUTE !CONCAT('fgd',!I,'em', !J) = !CONCAT('exp',!i,!j,!A)/!CONCAT('emp',!i,!j,!A) .
COMPUTE !CONCAT('fgd',!I,'sa', !J) = !CONCAT('exp',!i,!j,!A)/!CONCAT('sal',!i,!j,!A)*1000000
.
COMPUTE !CONCAT('fgd',!I,'ta', !J) = !CONCAT('exp',!i,!j,!A)/!CONCAT('tax',!i,!j,!A) .
EXECUTE .


*warning: following selections only valid for year 2002.
*estimation following turnover.
IF (MISSING(!CONCAT('gd',!I,!A)) & (!CONCAT('na2',!A) = "18" | !CONCAT('na2',!A) = "19" |
  !CONCAT('na2',!A) = "20" | !CONCAT('na2',!A) = "21" | !CONCAT('na2',!A) = "27" |
  !CONCAT('na2',!A) = "32" | !CONCAT('na2',!A) = "35"))
  !CONCAT('gd',!I,!A) = !CONCAT('v12110', !A) * !CONCAT('fgd',!I,'ca', !J)/1000000 .
EXECUTE .

*estimation following employment.
IF (MISSING(!CONCAT('gd',!I,!A)) &  (!CONCAT('na2',!A) = "31" | !CONCAT('na2',!A) = "33"))
  !CONCAT('gd',!I,!A) = !CONCAT('v16110', !A) * !CONCAT('fgd',!I,'em', !J) .
EXECUTE .

*estimation following salaries.
IF (MISSING(!CONCAT('gd',!I,!A)) & (!CONCAT('na2',!A) = "15" | !CONCAT('na2',!A) = "16" |
  !CONCAT('na2',!A) = "17" | !CONCAT('na2',!A) = "25" | !CONCAT('na2',!A) = "26" |
  !CONCAT('na2',!A) = "34"))
  !CONCAT('gd',!I,!A) = !CONCAT('v13320', !A) * !CONCAT('fgd',!I,'sa', !J)/1000000 .
EXECUTE .

*estimation following taxes.
IF (MISSING(!CONCAT('gd',!I,!A)) & (!CONCAT('na2',!A) = "24" | !CONCAT('na2',!A) = "29" |
  !CONCAT('na2',!A) = "36"))
  !CONCAT('gd',!I,!A) = v3013001 * !CONCAT('fgd',!I,'ta',!J) .
EXECUTE .



*Now we use different parameter for different domains, allover the remaining sectors.
*estimation following taxes for air and water.
!IF (!I = 1 | !I=2 | !I=t) !THEN.
IF (MISSING(!CONCAT('gd',!I,!A)))
  !CONCAT('gd',!I,!A) = v3013001 * !CONCAT('fgd',!I,'ta',!J) .
EXECUTE .
!IFEND.
```

```
*estimation following wages for waste and others.
!IF (!I = 3 | !I=9 ) !THEN.
IF (MISSING(!CONCAT('gd',!I,!A)))
  !CONCAT('gd',!I,!A) = !CONCAT('v13320', !A) * !CONCAT('fgd',!I,'sa', !J)/1000000 .
EXECUTE .
!IFEND.

*estimation following labour for soil.
!IF (!I = 4) !THEN.
IF (MISSING(!CONCAT('gd',!I,!A)))
  !CONCAT('gd',!I,!A) = !CONCAT('v16110', !A) * !CONCAT('fgd',!I,'em', !J) .
EXECUTE .
!IFEND.

 !DOEND.
 !DOEND.



*Correction of total (preference initial>calculated>estimation on total).

COMPUTE gdtemp= !CONCAT ('gdt',!A).
EXECUTE .
COMPUTE   !CONCAT ('gdt',!A) = !CONCAT(!R,'t',!A).
EXECUTE .
IF (MISSING (!CONCAT ('gdt',!A)))
  !CONCAT('gdt',!A) = !CONCAT ('gd1',!A) + !CONCAT ('gd2',!A) + !CONCAT ('gd3',!A) +
!CONCAT ('gd4',!A) + !CONCAT ('gd9',!A) .
EXECUTE .
IF (MISSING (!CONCAT ('gdt',!A)))
  !CONCAT('gdt',!A) = gdtemp .
EXECUTE .


!ENDDEFINE.


*results: fgd1em4, 3, 2= estimator domain 1 for last year calculated nace 4, 3, 2.
*results: gd102= estimate domaine 1 for year 02.


*********************************************************************************************
*Program.
*********************************************************************************************


!FACTOR R= e A= 02 D= 1 2 3 4 9 t.
```

Annexe 7: Current PAC expenditures per million turnover in 2002

Sector	Air	N1	Water	N2	Waste	N3	Soil	N4	Other	N5
14	198	3	0	3	161	3	23	3	237	3
15	159	85	1983	120	2057	151	250	94	404	78
16	7	4	3	4	228	7	9	5	651	4
17	32	44	1914	59	1460	80	171	46	309	42
18	0	5	277	5	466	15	448	6	1153	5
19	0	3	222	4	1571	5	0	3	472	3
20	11	10	103	11	1508	22	89	11	548	10
21	48	21	2429	26	2235	34	127	22	365	19
22	318	38	7	38	445	56	17	40	235	36
23	26	9	299	11	2433	11	10	9	148	9
24	1369	79	3543	88	3197	97	162	71	1483	68
25	241	40	268	44	2344	75	588	40	978	35
26	826	31	623	33	2013	44	152	29	419	24
27	2851	19	3936	22	4515	31	294	19	721	17
28	372	56	1011	64	1160	105	381	57	407	53
29	89	34	349	42	689	67	65	35	166	32
31	200	20	221	20	841	36	172	19	259	18
32	45	6	94	8	382	15	98	5	116	5
33	0	7	0	7	875	15	0	7	175	7
34	137	19	271	20	815	30	163	21	692	18
35	1195	7	432	6	730	9	67	5	107	5
36	86	34	445	38	1037	56	357	35	248	32
40	,	,	,	,	,	,	1336	1	,	,
41	0	2	15955	2	2208	2	4350	2	25431	2

Annexe 8: SPSS syntax for trend imputation using donors

```
****************************************
*Trend imputation.
*by B. Kestemont, Statistics Belgium.
*August 2004.
****************************************
*Default factors are estimated by donors.
*na2 = nace 2 digits.

*R= prefix of source variable.
*A= year.
*B= base year.
*D= environmental domain.

*calculated on GD, after merging all available responses


*************************************************************************************
*MACRO TRENDI.
*************************************************************************************

DEFINE !TRENDI (R = !TOKENS (1)
                /A = !TOKENS (1)
                /B = !TOKENS (1)
                /D = !CMDEND).


*initialisation of estimate .
 !DO !I !IN (!D).
COMPUTE !CONCAT('sm',!I,!A, !B) = !CONCAT(!R,!I,!A) .
EXECUTE .
 !DOEND.


!IF (!B= 99) !THEN.

*estimation following trend of turnover: all remaining sectors.

 !DO !I !IN (!D).
USE ALL.
IF (MISSING(!CONCAT('sm',!I,!A, !B)) & ~MISSING(!CONCAT(!R,!I,!B)) &
!CONCAT('v12110', !B) ~=0)
  !CONCAT('sm',!I,!A, !B) = !CONCAT(!R,!I,!B) * !CONCAT('v12110', !A) / !CONCAT('v12110',
!B)  .
EXECUTE .
 !DOEND.

!ELSE.
*warning: following selections only valid for year 2002.

 !DO !I !IN (!D).

*estimation following trend of turnover.
USE ALL.
```

```
IF (MISSING(!CONCAT('sm',!I,!A, !B)) & ~MISSING(!CONCAT(!R,!I,!B)) &
!CONCAT('v12110', !B) ~=0 & (!CONCAT('na2',!A) = "18" | !CONCAT('na2',!A) = "19" |
  !CONCAT('na2',!A) = "20" | !CONCAT('na2',!A) = "21" | !CONCAT('na2',!A) = "27" |
  !CONCAT('na2',!A) = "32" | !CONCAT('na2',!A) = "35"))
  !CONCAT('sm',!I,!A, !B) = !CONCAT(!R,!I,!B) * !CONCAT('v12110', !A) / !CONCAT('v12110',
!B) .
EXECUTE .

*estimation following trend of employment: all remaining sectors.
USE ALL.
IF (MISSING(!CONCAT('sm',!I,!A, !B)) & ~MISSING(!CONCAT(!R,!I,!B)) &
!CONCAT('v16110', !B) ~=0)
  !CONCAT('sm',!I,!A, !B) = !CONCAT(!R,!I,!B) * !CONCAT('v16110', !A) / !CONCAT('v16110',
!B)  .
EXECUTE .

 !DOEND.

!IFEND.

*Correction of total (preference initial>calculated>estimation on total).

COMPUTE smtemp= !CONCAT ('smt',!A, !B).
EXECUTE .
COMPUTE   !CONCAT ('smt',!A, !B) = !CONCAT(!R,'t',!A).
EXECUTE .
USE ALL.
IF (MISSING (!CONCAT ('smt',!A, !B)))
  !CONCAT('smt',!A, !B) = !CONCAT ('sm1',!A, !B) + !CONCAT ('sm2',!A, !B) + !CONCAT
('sm3',!A, !B) + !CONCAT ('sm4',!A, !B) + !CONCAT ('sm9',!A, !B) .
EXECUTE .
USE ALL.
IF (MISSING (!CONCAT ('smt',!A, !B)))
  !CONCAT('smt',!A, !B) = smtemp .
EXECUTE .




!ENDDEFINE.


*results in the form: sm102.

*********************************************************************************************
*Program.
*********************************************************************************************
*note: for 2000, I do not have the auxilliary data.

!TRENDI R= e A= 02 B= 01 D= 1 2 3 4 9 t.

!TRENDI R= e A= 02 B= 99 D= 1 2 3 4 9 t.
```

Annexe 9: SPSS 11 syntax for imputation using stratum mean

```
*********************************************************************************************
*MEANI IMPUTATION
*by B. Kestemont, Statistics Belgium, June 2004
*********************************************************************************************

*********************************************************************************************
*MACRO SERIAL
*********************************************************************************************

DEFINE !MEANI (R = !TOKENS (1)
                    /A = !TOKENS (1)
                    /B = !TOKENS (3)
                    /D = !CMDEND).

*R = prefix of base variable (ex: e). Should be one character.
*A = year of missing variable to be estimated.
*B = oldest reference years used for the imputation.
*D = domains. Should be one character.

*on élimine d'abord les entreprises non valides
*(même si elles ont donné un résultat pour les questions environnementales)
*afin de pouvoir élaborer les statistiques pour le présent rapport
*(nombre de valeurs imputées au sein des entreprises valides)

FILTER OFF.
USE ALL.
SELECT IF( ~ MISSING(!CONCAT ('poids', !A)) & !CONCAT ('poids', !A) ~= 0).
EXECUTE .

*Extrapolation par strate

SORT CASES BY
  !CONCAT('na4', !A) (A) .


!DO !I !IN (!D).

USE ALL.
COMPUTE !CONCAT ('m', !I, !A) = !CONCAT (!R, !I, !A) .
EXECUTE .

*Pour les différentes classes de NACE.
!DO !K = 4 !TO 2 !BY -1.

*moyenne par strates de classe 0 à 5.

!DO !J = 0 !TO 5.

USE ALL.
COMPUTE filter_$=( ~ MISSING(!CONCAT (!R, !I, !A))& ~ MISSING(!CONCAT ('poids', !A))
& !CONCAT ('poids', !A) ~= 0 & !CONCAT ('classe', !A) =!J).
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
```

```
EXECUTE .

AGGREGATE
  /OUTFILE= !CONCAT('moyenne',!I,!J,!K,!A,'.sav')
  /BREAK=!CONCAT('na', !K , !A)
  /!CONCAT ('m', !I, !J, !A) = MEAN(!CONCAT (!R, !I, !A))
  /N_BREAK=N.

USE ALL.
MATCH FILES /FILE=*
 /TABLE= !CONCAT('moyenne',!I,!J,!K,!A,'.sav')
 /RENAME (n_break = d0)
 /BY !CONCAT('na', !K , !A)
 /DROP= d0.
EXECUTE.

*affecter valeur de la moyenne de strate suivant les classes de taille.
USE ALL.
IF (MISSING(!CONCAT ('m', !I, !A)) & ~ MISSING(!CONCAT ('poids', !A)) & !CONCAT
('poids', !A) ~= 0 & !CONCAT ('classe', !A) =!J)
  !CONCAT ('m', !I, !A) = !CONCAT ('m', !I, !J, !A) .
EXECUTE .

!DOEND.
!DOEND.


*Pour les quelques entreprises qui restent, affecter la valeur de 2001 à 1999.

!DO !L !IN (!B).

USE ALL.
IF (MISSING(!CONCAT ('m', !I, !A)) & ~ MISSING(!CONCAT ('poids', !A)) & !CONCAT
('poids', !A) ~= 0)
  !CONCAT ('m', !I, !A) = !CONCAT (!R, !I, !L) .
EXECUTE .

!DOEND.


!DOEND.

*correct total = sum of domains if all available.
IF (~MISSING (!CONCAT('m1', !A)) & ~MISSING (!CONCAT('m2', !A)) & ~MISSING
(!CONCAT('m3', !A)) & ~MISSING (!CONCAT('m4', !A)) & ~MISSING (!CONCAT('m9', !A)))
!CONCAT('mt', !A) = !CONCAT('m1', !A) + !CONCAT('m2', !A) + !CONCAT('m3', !A) +
!CONCAT('m4', !A) + !CONCAT('m9', !A)  .
* de cette manière, l'imputation sur total n'est retenue qu'à défaut d'autre chose.
*remarquons que le total initial est ici éventuellement corrigé.
*J'obtiens D variables de forme m102.

!ENDDEFINE.

*********************************************************************************************
*Program.
*********************************************************************************************


!MEANI R= e A= 02 B= 01 00 99 D= 1 2 3 4 9 t.
```

Annexe 10: Syntax for cascade imputation

```
*************************************************************************************
*CASCADE IMPUTATION
*by B. Kestemont, Statistics Belgium, June 2004
*************************************************************************************

*************************************************************************************
*MACRO CASCADE
*************************************************************************************


DEFINE !CASCADE (R = !TOKENS (1)
                 /A = !TOKENS (1)
                 /D = !CMDEND).

*R = prefix of base variable (ex: e). Should be one character.
*A = year of missing variable to be estimated.
*D = domains. Should be one character.


SORT CASES BY
  !CONCAT('na4', !A) (A) .


!DO !I !IN (!D).

*initialisation
USE ALL.
COMPUTE !CONCAT ('v', !I, !A) = !CONCAT (!R, !I, !A) .
EXECUTE .

*Serial imputation from 2000-2001 to 2002.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('x', !I, !A, '0001') .
EXECUTE .

*Serial imputation from 1999-2001 to 2002.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('x', !I, !A, '9901') .
EXECUTE .

*Serial imputation from 1999-2000 to 2002.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('x', !I, !A, '9900') .
EXECUTE .

*Temporal imputation using donors from 2001 to 2002.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('s', !I, !A, '01') .
EXECUTE .

*Temporal imputation using donors from 2000 to 2002.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('s', !I, !A, '00') .
EXECUTE .

*Temporal imputation using donors from 1999 to 2002.
```

```
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('s', !I, !A, '99') .
EXECUTE .

*Trend imputation from 2001 to 2002.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('sm', !I, !A, '01') .
EXECUTE .

*Trend imputation from 1999 to 2002.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('sm', !I, !A, '99') .
EXECUTE .

*Sector imputation using factors.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('gd', !I, !A) .
EXECUTE .

*Stratum imputation using means.
IF (MISSING(!CONCAT ('v', !I, !A)))
  !CONCAT ('v', !I, !A) = !CONCAT('m', !I, !A) .
EXECUTE .

!DOEND.



*Environmental taxe on total.
IF (MISSING(!CONCAT ('v', 't', !A)))
  !CONCAT ('v', 't', !A) = v3013001 .
EXECUTE .

!ENDDEFINE.

*the result is of the form: v102.
*ready for a latest deterministic imputation.
*then for extrapolation.
*********************************************************************************************
*Program.
*********************************************************************************************


!CASCADE R= e A= 02 D= 1 2 3 4 9 t.
```

Annex 11: Syntax for extrapolation

```
**********************************
*Extrapolation
**********************************

*Classes de taille
**********************

*d'abord créer les classes de taille suivant l'EU, à partir de l'emploi total 1611002

STRING classeEU (A2).

USE ALL.
IF (v1611002 =0) classeEU = "01" .
EXECUTE .

USE ALL.
IF (v1611002 =1) classeEU = "45" .
EXECUTE .

USE ALL.
IF (v1611002 >=2 &v1611002<=4) classeEU = "51" .
EXECUTE .

USE ALL.
IF (v1611002 >=5 &v1611002<=9) classeEU = "48" .
EXECUTE .

USE ALL.
IF (v1611002 >=10 &v1611002<=19) classeEU = "04" .
EXECUTE .

USE ALL.
IF (v1611002 >=20 &v1611002<=49) classeEU = "07" .
EXECUTE .

USE ALL.
IF (v1611002 >=50 &v1611002<=249) classeEU = "52" .
EXECUTE .

USE ALL.
IF (v1611002 >=1000) classeEU = "22" .
EXECUTE .

USE ALL.
IF (v1611002 >=500 &v1611002<=999) classeEU = "17" .
EXECUTE .

USE ALL.
IF (v1611002 >=250 &v1611002<=499) classeEU = "14" .
EXECUTE .

*attention, il n'y a pas de classe là où il n'y a pas de valeur pour v1611002)

*Extrapoler
*************
*1000 euros!
```

```
************

COMPUTE air = ev102*poids02/1000 .
EXECUTE .
COMPUTE water = ev202*poids02/1000 .
EXECUTE .
COMPUTE waste = ev302*poids02/1000 .
EXECUTE .
COMPUTE soil = ev402*poids02/1000 .
EXECUTE .
COMPUTE other = ev902*poids02/1000 .
EXECUTE .
COMPUTE c21140 = evt02*poids02/1000 .
EXECUTE .

COMPUTE c2111001 = v211001*poids02/1000 .
EXECUTE .

COMPUTE c2112001 = v212001*poids02/1000 .
EXECUTE .

COMPUTE verif = c21140 - air - water - waste - soil - other .
EXECUTE .


*aggréger par classe de taille Eurostat et NACE 2 digits
*********************************************************************
*d'abord, ne sélectionner que ceux où il y a des valeurs

USE ALL.
COMPUTE filter_$=( ~ missing (v1611002) & ~ missing (c21140)).
VARIABLE LABEL filter_$ ' ~ missing (v1611002)  & ~ missing (totalt) (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .


AGGREGATE
  /OUTFILE='U:\données\Dépenses\entreprises\ResultGDClasseEUNace2.sav'
  /BREAK=na202 classeeu
  /i1211002 = SUM(c1211002) /i2111001 = SUM(c2111001) /i2112001 = SUM(c2112001)
  /iairt = SUM(air) /ieaut = SUM(water) /iwastet = SUM(waste) /isoilt = SUM(soil) /
  /iothert = SUM(other) /i21140 = SUM(c21140)
  /N_BREAK=N.
```

Annexe 12: Private PAC exp. in Belgium, Keur (2002)

				21140				
NACE2	**21110 (prov)**	**21120 (prov)**	**21140**	Air	Water	Waste	Soil	Other
10	0	0	0	0	0	0	0	0
14	112	267	498	42	113	166	23	155
15	49115	154032	202447	4127	107863	60574	11399	18484
16	72	151	2351	20	17	644	23	1647
17	4001	1512	30793	750	11682	13417	2372	2573
18	0	0	3246	0	231	1156	242	1617
19	37	0	788	0	60	595	0	133
20	261	1542	22482	222	599	14000	206	7455
21	1691	625	25878	141	10321	13556	510	1351
22	73	366	9585	3009	155	4138	223	2061
23	884	3447	109390	520	5383	100942	205	2339
24	21087	55653	371858	118874	96732	98421	13087	44744
25	2812	6128	48230	1949	1586	17091	3591	24013
26	7264	3062	27627	3154	2948	14243	1073	6209
27	6191	23014	108718	21094	36181	37613	5554	8276
28	1800	1271	29763	3247	6740	13785	2272	3720
29	859	1062	13762	606	2459	7057	331	3309
30	0	0	57	0	0	60	1	0
31	2367	1129	6475	608	925	2996	577	1369
32	609	364	5033	309	955	2547	223	1000
33	16	0	1409	0	20	1009	0	380
34	13149	3048	36104	2237	4307	12812	2920	13828
35	865	244	4376	1320	930	1652	104	370
36	2778	1173	9364	620	999	5288	1781	677
37	515	1404	2604	0	0	2271	61	145
40	10683	62686	93809	12656	34182	22166	1742	23062
41	59	10399	78318	0	55121	2227	2909	18060
45	831	819	1999	0	0	0	0	0
50	122	7	835	0	0	0	0	0
51	0	0	4875	0	0	67	0	0
52	0	0	3608	0	0	0	0	0
55	1	0	1860	0	0	0	0	0
60	68	24685	1107	0	0	0	0	0
61	0	0	2	0	0	0	0	0
62	0	0	185	0	0	0	0	0
63	12	2374	2120	0	0	0	0	0
64	0	0	1631	0	0	0	0	0
65	0	0	6	0	0	0	0	0
67	0	0	27	0	0	0	0	0
70	0	0	2314	0	0	0	0	0
71	0	24	866	0	0	0	0	0
72	0	0	108	0	0	0	0	0
73	215	0	56	0	0	0	0	0
74	0	6406	1718	0	0	0	0	0
80	0	0	0	0	0	0	0	0
85	0	0	0	0	0	0	0	0
90	242698	250	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0
93	0	5	0	0	0	0	0	0

Table of contents