Handling incompleteness after linkage to a population frame: incoherence in unit types, variables and periods

Arnout van Delden and Koert van Bemmel

The views expressed in this paper are those of the author(s)
and do not necessarily reflect the policies of Statistics Netherlands

Discussion paper (201208)

Statistics Netherlands

The Hague/Heerlen, 2012

Explanation of symbols

.	data not available
*****	provisional figure
******	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2011–2012	2011 to 2012 inclusive
2011/2012	average for 2011 up to and including 2012
2011/’12	crop year, financial year, school year etc. beginning in 2011 and ending in 2012
2009/’10–2011/’12	crop year, financial year, etc. 2009/’10 to 2011/’12 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
Grafimedia

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

Handling incompleteness after linkage to a population frame: incoherence in unit types, variables and periods

Arnout van Delden and Koert van Bemmel

Summary[1]: In this paper we concentrate on methods for handling incompleteness caused by differences in units, variables and periods of the observed data set compared to the target one. Especially in economic statistics different unit types are used in different data sets. For example, an enterprise (statistical unit type) may be related to one or more units in the Value Added Tax register. In addition those VAT units may declare turnover on a monthly, quarterly or yearly basis. Also the definition of turnover may differ from the targeted one. When tax data are linked to a population frame we may have response for only a part of the VAT units underlying the enterprise. We give an overview of different missingness patterns when VAT data are linked to enterprises and formulate methods to harmonize and complete the data at micro level.

Keywords: unit types, completion at micro level, harmonization

[1] This paper has been prepared for the ESSnet project on Data Integration

1. Background

1.1 Introduction

There is a variety of economic data available that is either collected by statistical or by public agencies. Combining those data at micro level is attractive, as it offers the possibility to look at relations / correlations between variables and to publish outcomes of variables classified according to small strata. National statistical institutes (NSI's) are interested to increase the use of administrative data and to reduce the use of primary data because population parameters can be estimated from nearly integral data and because primary data collection is expensive.

The economic data sources collected by different agencies are usually based on different unit types. These different unit types complicate the combination of sources to produce economic statistics. Two papers, the current paper and Van Delden and Hoogland (2011) deal with methodology that is related to those different unit types. Both papers deal with a Dutch case study in which we estimate quarterly and yearly turnover, where we use VAT data for the less complicated companies[2] and survey data for the more complicated ones.

Handling different unit types starts with the construction of a general business register (GBR) that contains an enumeration of the different unit types and their relations. From this GBR the population of statistical units that is active during a certain period is derived, the population frame. This population frame also contains the relations of the statistical units with other unit types, such as legal units. In the current paper we formulate a strategy for detecting and correcting errors in the linkage and relations between units of integrated data.

In the Dutch case study, after linkage, we handle differences in definitions of variables and completion of the data. After both steps, population parameters are computed. Both steps are treated in the current paper and resemble micro integration steps as described by Bakker (2011). After the computation of population parameters, an additional step of detecting and correcting errors is done as treated in the current paper.

In a next step, the yearly turnover data are combined at micro level (enterprise) with numerous survey variables collected for Structural Business Statistics. The paper by Pannekoek (2011) describes algorithms to achieve numerical consistency at micro level between some core variables collected by register data and variables collected by survey data. Examples of such core variables in economic statistics are turnover, and wages. There are also other European countries that estimate such a core variable, e.g. turnover, from a combination of primary and secondary data. Total

[2] In the current paper 'company' is used as a general term rather than as a specific unit type.

turnover and wage sums are central to estimation of the gross domestic product, from the production and the income side respectively.

Because the current paper and Van Delden and Hoogland (2011) share the same background, the current section 1.1 and the sections 1.2 and 2 are nearly the same in both papers.

1.2 Problem of unit types in economic statistics

The different unit types in different economic data sources complicate their linkage and subsequent micro integration. When a company starts, it registers at the chamber of commerce (COC). This results in a so called ‘legal unit’. The government raises different types of taxes (value added tax, corporate tax, income tax) from these “companies”. Depending on the tax legislation of the country, the corresponding tax units may be composed of one or more legal units of the COC, and they may also differ for each type of tax. Finally, Eurostat (EC, 1993) has defined different statistical unit types (local kind of activity unit, enterprise, enterprise group) which are composed of one or more legal units.

In the end, for each country, the set of unit types of companies may be somewhat different. But generally speaking, for each country, the legal units are the *base* units whereas tax and statistical units are *composite* units (see Figure *1.1*). In some countries, like France, there is one-to-one relationship between legal units and tax units and tax units are one-to-one related to statistical units. In other countries, like the Netherlands, units that declare tax may be groupings of legal units that belong to different enterprises (Vaasen and Beuken, 2009). Likewise, in Germany, tax units may declare turnover for a set of enterprises (Wagner, 2004). As a consequence, at least in the Netherlands and Germany, for the more complex companies tax units may be related to more than one enterprise. In other words, the tax and statistical units are both composed of legal units, but their composition may be different.

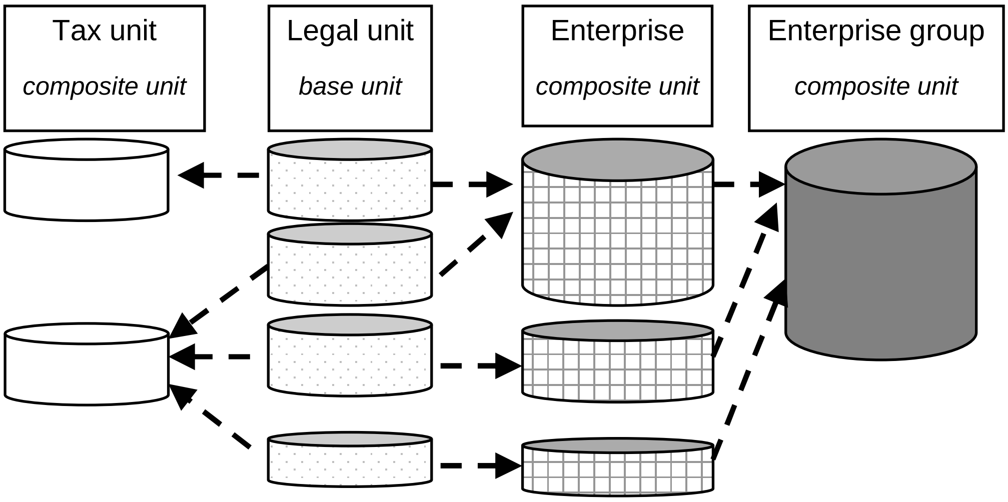

Figure *1.1.*: Different unit types in economic statistics. Each cylinder represents a single unit; arrows indicate the groupings of units.

1.3 General Business Register

NSI’s have a GBR that contains an enumeration of statistical units and the underlying legal units. The GBR contains the starting and ending dates of the statistical units, their size class (SC code) and their economic activity (NACE code). In 2008, Eurostat has renewed its regulation on a business register (Eurostat, 2008) in order to harmonise outcomes over different European countries. NSI’s also use a GBR to harmonise outcomes over different economic statistics within an NSI. In addition, the Netherlands – and other NSI’s, also added the relations between legal units and tax units to the GBR, to be able to use tax office data for statistical purposes.

1.4 Problem description

The focus of the current paper is on incompleteness of the data at the level of the statistical unit after linkage of register and survey data to a population frame. This incompleteness can be due to the absence of source data (observations) or because the source data first need to be harmonised in order to get estimated values for the statistical unit and for the intended reporting period. We can also have partially missing information if for some but not all administrative units belonging to the same statistical unit the target variable is not (yet) available.

In the current paper we distinguish three main reasons for incompleteness:

1. observations are lacking;
2. the unit types of the source data differs from the statistical unit type;
3. the definition of the source variable differs from the target variable.

Each of these three reasons is explained in detail in section 3.

The objective of the current paper is to describe methods for handling incompleteness of a variable at the level of statistical units due to incoherencies in unit types and in variable definitions of source compared to target data.

In this study we used the following starting points. Firstly, we do not aim for perfect estimates for each single statistical unit but to have accurate estimates for publication cells. Secondly, we produce outcomes for different customers where each customer wants different strata. The basic publication cells from which all those strata can be constructed are rather small. We therefore wanted a method that uses the observed data as much as possible. Thirdly, the different customers have different moments at which they want their output, varying form very early (25 days after the end of the publication period) to very late (two years after the end of the publication year). We wanted to have a single estimation method that could be used for each of the releases. Finally, the method should be able to deal with all kinds of missingness patterns. We wanted a general approach that is also useful for NSI’s with somewhat different missingness patterns. Given the third and fourth starting point we chose to use imputation, rather than weighting. We think that imputation provides flexibility to adjust the model to the corresponding missingness pattern and to deal with different publication moments.

1.5 Outline of paper

The remainder of the paper is organised as follows. Section 2 describes the Dutch case study. In section 3 we give a classification of missingness patterns, followed in section 4 by methodology to handle each of the missingness patterns. In section 5 we give an example of an accuracy test for some of the methods presented. Section 5.3 deals with special situations for which we wanted to make the outcomes more robust. Finally, in section 7 we sum up and suggest issues for further research.

2. Description of case study

2.1 Background: statistical output

In the current paper we deal with the estimation of Dutch quarterly and yearly turnover levels and growth rates, based on VAT declarations and survey data. The work is part of a project called “Direct estimation of Totals”. Turnover is estimated for the target population which consists of the statistical unit type the *enterprise*. Turnover output is stratified by NACE code × size class. An overview of all processing steps from input to output data can be found in Van Delden (2010).

The estimated quarterly figures are directly used for the short term statistics (STS). Also, the quarterly and yearly turnover levels and growth rates are input to the supply and use tables of the National Accounts, where macro integration is used to obtain consistent estimates with other parameters. Also, results are used as input for other statistics like the production index (micro data) and the consumption index (the estimates). Finally, yearly turnover is integrated at micro level with survey data of the Structural Business Statistics (SBS). Next, the combined data is used to detect and correct errors in both the turnover data as well as in the other SBS variables. Yearly turnover results per stratum are used as a weighting variable for SBS data.

In fact we deal with four coherent turnover estimates:

- net total turnover: total invoice concerning market sales of goods and services supplied to third parties excluding VAT
- gross total turnover: total invoice concerning market sales of goods and services supplied to third parties including VAT
- net domestic turnover: net turnover for the domestic market, according to the first destination of the product
- net non-domestic turnover: net turnover for the non-domestic market, according to the first destination of the product

More information on the turnover definition can be found in EC (2006). In the remainder of the paper we limit ourselves to net total turnover further referred to as turnover.

The quarterly and yearly figures are published in different releases, as shown in Table 2.1. The quarterly releases vary from a very early estimate delivered at 30–35 days after the end of the corresponding quarter to a final estimate for SBS publication delivered April year *y*+2 where *y* stands for the year in which the target period falls.

Table 2.1. Overview of the releases of the case study

Release	Period of estimation	Moment	Explanation
Flash estimate	Quarter	30–35 days after end of target period	Provisional estimate delivered for Quarterly Accounts, STS branches with early estimates
Regular estimate	Quarter	60–70 days after end of target period	Revised provisional estimate for Quarterly Accounts and for STS
Final STS estimate	Year and corresponding 4 quarters	April *y*+1, one year after target year	The estimates of the four quarters are consistent with the yearly figure
Final SBS estimate	Year and corresponding 4 quarters	April *y*+2, two years after target year	The estimates of the four quarters are consistent with the yearly figure. The yearly figure is based on STS and SBS turnover data

2.2 Target population and population frame

The statistical target population of a period consists of all enterprises that are active during a *period*. This true population is unknown. We represent this population by a frame which is derived from the GBR. Errors in this representation are referred to as frame errors. Each enterprise has an actual and a coordinated value for the SC and NACE code. The coordinated value is updated only once a year, at the first of January and is used to obtain consistent figures across economic statistics. In the remainder of the paper we always refer to the coordinated values of SC and NACE code unless stated otherwise.

The population frame is derived as follows. First, each month, we make a view of the GBR that represents the population of enterprises that are active at the first day of the month; in short: the population state. This population state also contains the legal units, tax units and the ‘enterprise groups’-units that are related to the enterprise population at the first day of the month. Next, the population frame for a period is given by the union of the relevant population states. For example, the frame for the first quarter of a year consists of the union of the population states on 1 January, 1 February, 1 March and 1 April.

For the case study, the frame contains four unit types: the legal unit (base unit), the enterprise (composite unit) and two tax units namely the base tax unit and the VAT unit. In the Netherlands each legal unit (that has to pay tax) corresponds one-to-one to a base tax unit. For the VAT, base tax units may be grouped into a VAT unit (composite unit). So this is an extension of the more general situation of Figure *1.1*. A more extensive description can be found in Van Delden and Hoogland (2011).

As explained in Vaasen and Beuken (2009), in the case of smaller companies each VAT unit is related to one enterprise and each enterprise may consist of one or more VAT units. For the more complicated companies, referred to as topX units, a VAT unit may be related to more than one enterprise.

2.3 Data

In the case study we use two types of source data. We use VAT data for the non-topX enterprises. For the topX enterprises we use primary data because VAT units may be related to more than one enterprise. This approach is quite common, also at other NSI's in Europe (e.g. Fisher and Oertel, 2009; Koskinen, 2007; Norberg, 2005; Orjala, 2008; Seljak, 2007). For the non topX units, we only use observations of VAT units that are related to the target population of enterprises.

Concerning the VAT, a unit declares the value of sales of goods and services, divided into different sales types. The different sales types are added up to the total sales value, which we refer to as turnover according to the VAT declaration.

In the current paper we use VAT and survey data for 2008 and 2009 and the first two quarters of 2010. Data are stratified according to NACE 2008 classification.

3. Classification of missingness patterns

3.1 Introduction

We classify the missingness patterns along three main reasons for missingness:

- observations are lacking;
- the unit types of the source data differs from the statistical unit type;
- the definition of the source variable differs from the target variable.

These three main reasons are further subdivided in section 3.2–3.4. Although this classification is derived from the Dutch case study, we believe that the three main reasons of missingness also apply to other situations (variables and NSI's).

Note that we structure *patterns* not units: we do not make a classification where each unit with a pattern only falls into one class. In practice, units can have two missingness patterns simultaneously. For example: a different unit structure can coincide with a different variable meaning. The imputation method handles the missingness patterns in a certain order, as explained in section 4.6.

3.2 Missingness due to lack of observations

3.2.1 Classification

We distinguish four kinds of missingness due to lack of observations:

(1a)` Units in the population frame that did not respond (yet) but that have responded in the past.

For the quarterly data, a flash estimate is made and used in National Accounts. This flash estimate is delivered at about 30 days after the end of the period, see Table 2.1 of section 2.1. The processing time is currently around five days, so we can use data that are delivered to the tax office up to 25 days after the end of the period. At that moment 40–50% of the expected VAT units have not (yet) responded. Many of them have historical observations. Some will respond later, others are ended.

(1b) Units in the population frame with a structural lack of observations

Some structural non responding VAT units have dispensation from the tax office, because they are very small or because their activities require no tax obligations. Others may evade tax or they may be wrongly present in the frame. Also sample survey units may be structurally non respondent.

(1c) Units in the population frame that did not respond (yet) and that are new in the population frame

(1d) Units that do belong to the conceptual population but are wrongly not present in the population frame

Under coverage in the population frame is not solved by imputation but by the correction of linkages or relations between units, as explained in Van Delden and Hoogland (2011).

3.2.2 *Quantification*

To quantify the occurrence of the different patterns of missingness, we counted the number of VAT-units with a tax declaration and their corresponding turnover after linkage to the population frame over Q4 2009 and Q1 2010, for NACE-code I: “Accommodation and food service activities”. For both quarters we made two estimates: an early one and a late one (see Table 3.1) in which we “simulated” the releases as shown in Table 2.1. For units with historical turnover (missingness class 1a) we used units that had at least one declaration since January 2008. The results are shown in Table 3.2 – Table 3.5).

Table 3.2 shows that 24494 VAT units have responded at the flash estimate for Q4 2009, compared to 46059 VAT units at the final estimate, corresponding to 4.26 and 7.42 milliard Euros respectively. When we only count the turnover of those enterprises where all related VAT units have responded for the full three months at the flash estimate of Q4 2009 (see complete declarations in Table 3.2) we get only 2.95 milliard Euros.

Figures of Q1 2010 are similar, see Table 3.3. Note that units that declare on a yearly basis were included in the Q4 2009 but not in the Q1 2010 counting's. We could not include the latter because our data file was up to Q3 2010 and therefore their yearly declarations over 2010 were not present in the file.

Table 3.4 shows that 46059 VAT units have responded over Q4 2009 at the final estimate. A subset of 26628 VAT units has historical responses but did not yet respond at the flash estimate (missingness pattern 1A). At the final estimate, 26023 of the 26628 non respondents were shown to be late respondents; the other 605 VAT units never responded and (probably) were ended.

Table 3.1. Approximation of releases for Q4 2009 and Q1 2010 to estimate the frequency of missingness patterns among VAT units.

Period	Release (approximated)	Latest arrival date of declarations at tax office	Remark
Q4 2009	Flash estimate	25-1-2010	Partial response for monthly, quarterly and yearly respondents.
	Final estimate	26-8-2010*	Response (nearly) complete for monthly, quarterly and yearly respondents.
Q1 2010	Flash estimate	25-4-2010	Partial response for monthly and quarterly respondents.
	Regular estimate	26-8-2010*	Response (nearly) complete for monthly and quarterly respondents. No yearly respondents yet.

* We took the latest available date in the data file

Table 3.2. Number of VAT units and corresponding turnover for Q4 2009 and NACE “Accommodation and food service activities”.

	Flash estimate			Final estimate		
Type of declaration	Total	TopX	non-TopX	Total	TopX	non-TopX
	Number of units					
Total	24494	206	24288	46059	284	45775
Monthly	8854	139	8715	9080	139	8941
Quarterly	15159	66	15093	32536	140	32396
Yearly[1]	480	1	479	4389	5	4384
Other[2]	1	0	1	55	0	55
	Declared Turnover ($\times 10^9$ Euros)					
Total	4.26	2.28	1.98	7.42	3.94	3.48
Monthly	2.80	1.67	1.13	3.86	2.41	1.46
Quarterly	1.45	0.61	0.84	3.48	1.53	1.95
Yearly	0.01	0.00	0.01	0.08	0.00	0.07
Other	0.00	0.00	0.00	0.00	0.00	0.00
	Turnover of complete declarations ($\times 10^9$ Euros)					
Total	2.95	1.64	1.31			
Monthly	1.62	1.12	0.50			
Quarterly	1.32	0.52	0.80			
Yearly	0.01	0.00	0.01			
Other	0.00	0.00	0.00			

[1] Quarterly turnover of units that declare on yearly basis is computed as yearly turnover divided by 4. [2] Shifted calendar quarter (stagger)

Table 3.3. Number of VAT-declarations and corresponding turnover for Q1 2010 and NACE “Accommodation and food service activities”.

	Flash estimate			Regular estimate		
Type of declaration[1]	Total	TopX	non-TopX	Total	TopX	non-TopX
	Number of units					
Total	26241	202	26039	42417	278	42139
Monthly	9033	140	8893	9297	140	9157
Quarterly	17209	62	17147	33121	138	32983
Yearly	0	0	0	0	0	0
Other	0	0	0	0	0	0
	Turnover ($\times 10^9$ Euros)					
Total	4.00	2.11	1.89	6.20	3.21	2.99
Monthly	2.55	1.54	1.01	3.13	1.87	1.26
Quarterly	1.45	0.57	0.88	3.06	1.34	1.73
Yearly	0.00	0.00	0.00	0.00	0.00	0.00
Other	0.00	0.00	0.00	0.00	0.00	0.00
	Turnover of complete declarations ($\times 10^9$ Euros)					
Total	3.29	1.87	1.42			
Monthly	2.00	1.43	0.57			
Quarterly	1.29	0.44	0.85			
Yearly	0.00	0.00	0.00			
Other	0.00	0.00	0.00			

1 see footnotes of Table 3.2

Table 3.4. Number of VAT-units and corresponding turnover for Q4 2009, missingness pattern 1A[1] and NACE “Accommodation and food service activities”.

		Number of units			Turnover ($\times 10^9$ Euros)		
		Total	TopX	non-TopX	Total	TopX	non-TopX
Response at final estimate	Total	46059	284	45775	7.42	3.94	3.48
Pattern 1A	Total missing at flash estimate[1]	26628	157	26471			
	Response after flash estimate						
	Total	26023	156	25867	3.14	1.66	1.48
	Monthly	5377	80	5297	1.06	0.73	0.33
	Quarterly	17105	74	17031	2.01	0.92	1.09
	Yearly[2]	3487	2	3485	0.06	0.00	0.06
	Other[2]	54	0	54	0.00	0.00	0.00
	No later quarterly response	605	1	604			

1 VAT units with at least one historical VAT declaration since January 2008.

2 see footnotes of Table 3.2

Table 3.5. Number of VAT-units and corresponding turnover for Q1 2010, missingness patterns 1A–1C and NACE “Accommodation and food service activities”.

		Number of units			Turnover ($\times 10^9$ euros)		
		Total	TopX	non-TopX	Total	TopX	non-TopX
Response at regular estimate		42417	278	42139	6.20	3.21	2.99
Pattern 1A	Total missing at flash estimate[1]	24699	146	24553			
	Response after flash estimate	19967	141	19826	2.15	1.10	1.06
	Monthly [1]	5245	66	5179	0.60	0.33	0.26
	Quarterly	14722	75	14647	1.56	0.77	0.79
	Yearly	0	0	0	0.00	0.00	0.00
	Other	0	0	0	0.00	0.00	0.00
	No later quarterly response	4732	5	4727			
Pattern 1B	Total missing at flash estimate	1541	8	1533			
	Code = 01 [2]	239	0	239			
	Code ≠ 01	1302	8	1294			
	Response after flash estimate	56	0	56	0.001	0.000	0.001
	Code = 01	56	0	56	0.001	0.000	0.001
	Code ≠ 01	0	0	0	0.000	0.000	0.000
Pattern 1C	Total missing at flash estimate	1606	4	1602			
	Code = 01	1110	1	1109			
	Code ≠ 01	148	0	148			
	Code missing	348	3	345			
	Response after flash estimate	951	1	950	0.04	0.00	0.04
	Code = 01	948	1	947	0.04	0.00	0.04
	Code ≠ 01	1	0	1	0.00	0.00	0.00
	Code missing	2	0	2	0.00	0.00	0.00

1 see footnotes of Table 3.2. 2 Code = 01: have to declare tax, Code ≠ 01: tax dispensation

Table 3.6. Number of VAT-units and corresponding turnover in the tax declaration file split into ‘linked’ and ‘not linked’ to the population frame of Q4 2009.

Linked to pop. frame of Q42009	Type of declaration[1]	STS-domain			non STS-domain	
		Total	TopX	non-TopX	TopX	non-TopX
		Number of VAT units				
Linked	Total	1132741	7306	813312	3380	308743
	Monthly	179631	3413	144700	1133	30385
	Quarterly	826622	3636	592131	2014	228841
	Yearly	119895	256	76039	232	43368
	Other	6600	1	443	1	6155
Not linked	Total	289652				
Linked later	Total	43447				
		Declared Turnover ($\times 10^9$ Euros)				
Linked	Total	327.2	147.3	144.1	11.5	24.3
	Monthly	164.8	68.8	79.1	6.6	10.2
	Quarterly	160.7	78.3	64.2	4.9	13.3
	Yearly	1.2	0.1	0.8	0.0	0.3
	Other	0.5	0.0	0.0	0.0	0.4
Not linked	Total	12.4				
Linked later	Total	4.2				

[1] see footnotes of Table 3.2

In Table 3.5 we can see that patterns 1B and 1C occur far less frequently than pattern 1A. At the flash estimate over Q1 2010 24699 units were non respondent with a historical turnover (pattern 1A), 1541 units were non respondent with a structural lack of turnover (pattern 1B) and 1606 units were non respondents that were new in the population frame (pattern 1C). Also in terms of turnover, pattern 1A is far more important than pattern 1B and 1C. Note that some of the units with a structural lack of turnover as well as some of the new units have a tax office code $\neq 0$ which means they have dispensation from the tax office.

We compared the number and turnover of VAT declarations that could be linked to the population frame of Q4 2009 versus those that could not be linked, over the *full range* of NACE codes at the final estimate, see Table 3.6 . Results show that about 3 per cent of the turnover in the declaration file could not be linked to the population frame. Since 2010, the turnover that cannot be linked to the population frame has gradually been reduced to about 1 per cent of the total declared turnover due to improvement in the population frame.

3.3 Missingness due to different unit structure

3.3.1 Classification

We distinguish between two kinds of missingness due to a different unit structure:

(2a) Observations (e.g tax declarations) are related to one enterprise group but to more than one underlying enterprise.

As explained in section 2.2 VAT declarations are mostly related to one enterprise group, but can be related to more than one enterprise underlying the enterprise group. The latter can be a problem because we wish to make estimates for strata defined by NACE codes and size classes which are properties of enterprises.

(2b) Observations (e.g tax declarations) are related to more than one Enterprise Group and also to more than one underlying enterprise.

Note that sometimes a VAT declaration is related to more than one enterprise group. This may for example occur with a unit that declares on a yearly basis and within that year the VAT unit has been taken over by a new enterprise group.

3.3.2 Quantification

We counted the occurrence of missingness patterns 2A and 2B for the Accommodation and food service activities over 2010 (Table 3.7). In Q1 2010 a total of 60775 tax declarations were linked to the corresponding population frame at the regular estimate. A total of 556 VAT declarations were related to topX enterprises. For the majority of them, 507, a VAT declaration is related to one Enterprise Group and to one enterprise, 40 declarations were related to one Enterprise Group but to more than one enterprise, and 9 were related to more than one Enterprise Group as well as to more than one enterprise. Although only 49 declarations were related to more than one enterprise this corresponded to a quarterly turnover of 2.44 milliard euros compared to the total of 3.21 milliard euros for topX enterprises. From this we can conclude that mainly tax declarations of topX entities are related to more than one enterprise. From Table 3.7 we can see that also 56 declarations were related to more than one non-topX enterprise, corresponding to a much smaller quarterly turnover of 0.02 milliard euros.

Table 3.7. Number and turnover of VAT-declarations for Q1 2010 at the regular estimate by type or relation, for Accommodation and food service activities.

Type of declaration	Number of declarations			Turnover ($\times 10^9$ Euros)		
	Total	TopX	non-TopX	Total	TopX	non-TopX
All types of relations of VAT unit to Enterprise group and enterprise						
Total	60755	556	60199	6.20	3.21	2.99
Monthly	27634	418	27216	3.13	1.87	1.26
Quarterly	33121	138	32983	3.06	1.34	1.73
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00
A VAT-unit related to one Enterprise Group and to one Enterprise						
Total	60650	507	60143	3.74	0.77	2.97
Monthly	27553	379	27174	1.61	0.36	1.25
Quarterly	33097	128	32969	2.13	0.41	1.72
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00
A VAT-unit related to one Enterprise Group but to more than one Enterprise						
Total	40	40	0	1.83	1.83	0.00
Monthly	30	30	0	0.90	0.90	0.00
Quarterly	10	10	0	0.93	0.93	0.00
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00
VAT-unit related to more than one Enterprise Group and to more than one Enterprise						
Total	65	9	56	0.63	0.61	0.02
Monthly	51	9	42	0.62	0.61	0.01
Quarterly	14	0	14	0.01	0.00	0.01
Yearly	0	0	0	0.00	0.00	0.00
Other	0	0	0	0.00	0.00	0.00

3.4 Missingness due to different meaning of variable

3.4.1 Classification

The third cause of missingness is because the meaning of the variable in the data set differs from the target variable. We subdivide this into three types:

(3a) Observations are available for a period that is longer than the target period.

In the Netherlands and in many European countries, a VAT unit may report to the tax office on a monthly, quarterly or yearly basis. Generally speaking, the larger the unit, the more frequently it has to declare its turnover to the tax office.[3] The exact rules differ from country to country (Statistics Finland, 2009).

At the end of the year we make a release where we estimate yearly and quarterly turnover, which are numerically consistent with each other. For this final estimate, we use yearly turnover observations and divided those over the four underlying quarters.

(3b) Observations are available for a period that is shifted compared to target period

Some VAT units declare tax on a quarterly basis, but the period is shifted compared to a calendar quarter. For example, units declare tax for February – April, or for March – May. Those units are referred to as "staggers". In the Netherlands staggers are rare but they occur frequently in the United Kingdom (Orchard *et al.*, 2010).

(3c) Observations are available but need to be transformed due to definition differences

In the tax declaration form, the VAT-units declare the value of products and services that have been sold, the turnover. From this declaration, the amount of tax to be paid is calculated. However, the turnover found on the tax declaration form may differ from the one defined by Eurostat (EC, 2006):

- Some units have dispensation for part of their turnover. This is the case for some branches;
- Other tax rules; for example for some activities units don't have to declare the total turnover but only the profit margin[4].
- Intra-enterprise turnover. The statistical variable turnover only consists of market-oriented sales. When an enterprise consists of two or more VAT units, the declared turnover may partly consist of deliveries of goods or services within the enterprise which is not market-oriented.

3.4.2 *Quantification*

In Table 3.2 we can see that of the total of 46059 VAT units that declared tax over Q4 2009 in the Accommodation and food service activities at the final estimate, 4389 units declared tax on a yearly basis (pattern 3A) and 55 VAT units were " staggers" (pattern 3B). In terms of turnover this corresponded to 7,42 milliard euros for the total, 0,08 milliard euros for the yearly tax reporters and less than 0,01 milliard euros for the staggers.

[3] In the Netherlands only very small units are allowed to report tax on a yearly basis. However, since July 2009, many units are allowed to report on a quarterly basis instead of on a monthly basis.

[4] This applies to trade in second-hand goods that are sold to enterprises without a tax number or to private persons.

As far as we know, the differences in the Netherlands between tax and target turnover are limited, within the domain of the STS regulation. Based on an analysis of tax data and SBS survey data over 2009, we found that for about 10 per cent of the 4 digit NACE codes within the STS domain, VAT turnover cannot be used because differences in definition are too large. For a further small number (less than 10) of 4 digit NACE codes we derive the target turnover from the statistical turnover. For the remaining nearly 90 per cent of the 4 digit NACE codes in the STS domain the VAT turnover corresponds closely to the target turnover. All those figures concern the total net turnover. For some STS domains the net turnover has to be split into sales to customers within the Netherlands (domestic) versus sales to customers outside the Netherlands (non domestic). This subdivision may sometimes be more difficult to estimate from VAT declarations.

4. Solutions to completing the variables for each type of missingness

4.1 Introduction

In section 4.2 explains at which unit level missing values are imputed. Sections 4.3 and 4.4 deal with completion, section 4.5 deals with harmonization. Finally section 4.6 explains some practical implementation issues. Some methodology to make the imputations more robust is treated in section 5.3. The methodology as described in the paper has slightly been simplified. Imputation at the level of the legal unit has been omitted because we use it only in some exceptional situations.

4.2 Level of imputation: statistical unit versus VAT unit

We analysed whether we wanted to impute at enterprise or at VAT unit level. To explain this, Table 4.1 shows an example of an enterprise is related to two VAT units. The quarterly turnover of the enterprise is given by the sum of the turnover of the two VAT units. Say we are in Q1 2010 and wish to publish data for Q4 2009 and for the whole year of 2009 In 2009 Q4 turnover of VAT unit 2 is missing. VAT unit 2 has reported for the previous quarters. This is a situation which often occurs with early estimates, see section 3.2. To complete the turnover of enterprise 1, we could impute the quarterly turnover of VAT unit 2 or we could discard the observed turnover and impute directly the total turnover for enterprise 1.

In order to make a choice we counted the number of VAT-units, classified according to the type of relation between VAT units and the enterprise within the STS-domain in the population frame of December 2009. 86 per cent of the enterprises were related to just one VAT unit and the remaining 14 per cent were related to two or more VAT-units.

Table 4.1: Part of the turnover of the statistical unit is completely missing for some quarter, but turnover is complete for historical quarters

Enterprise Id	VAT id	Quarterly turnover					Yearly turnover
		2008	2009				2009
		Q4	Q1	Q2	Q3	Q4	
1	1	102	100	105	95	103	403
1	2	27	25	30	30	?	?
Total		129	125	135	125	?	?

We compared the turnover response from VAT declarations for early and late estimates in the Accommodation and Food service activities in Q4 2009 (Table 3.2) and in Q1 2010 (Table 3.3). In Q4 2009, the VAT units related to non-topX enterprises declared 3.48 milliard euros turnover at the final estimate. For the flash estimate we can only use declarations that were sent up to 25 days after the end of the quarter. For the flash, just 1.98 milliard euros were declared by non-topX VAT units. If we further limit the declarations to those that have a complete quarterly turnover for the enterprise, it drops down to 1.31 milliard euros. For Q1 2010 we found similar results. The main reason why quarterly turnover is incomplete at enterprise level for the flash estimate is that VAT units that declare monthly have declared just two of the three monthly periods. Another reason is that some of the enterprises consist of more than one VAT unit, of which one did not yet respond.

The simplest imputation method would directly impute quarterly turnover at the level of the enterprise. In that case, the turnover of all units that did not have a complete quarterly turnover at enterprise level would be ignored and instead a value would be imputed. In the case of the Accommodation and Food service activities for Q4 2009 this would mean that we would discard for the non-topX units 1.98 – 1.31 = 0.67 milliard euros. Because a good quality of our first quarterly estimates is crucial to Statistics Netherlands we decided to use (nearly) all available turnover and impute turnover for the missing VAT declarations Thus in the case of a monthly reporter that has declared already two of the three months, we impute only the turnover of the third month.

We are aware that we could have used simpler methods, e.g. impute at the level of the enterprise and check for incompleteness on a monthly basis. In practice, this may be nearly as complicated because decision rules are needed to derive whether the reported turnover is complete or not.

4.3 Missingness due to lack of observations

4.3.1 Determine whether turnover is to be expected

When a VAT-unit that is found in the population frame has not responded yet we first have to decide whether we can expect turnover for this unit. To do so, we designed a decision tree (Figure *4.1*). The first step is that we verify whether the unit has tax dispensation, using a variable from a tax office client data base. Our experience is that those units have hardly any turnover. Therefore, we impute no turnover in those cases.

The second step is that we check how many days there are after the end of the reporting period. If we make a late estimate at which we are beyond threshold *Tr1* (see Table 4.2) we do not expect a declaration anymore, so probably the unit has ended but is falsely in the frame. If we make an early estimate, in the third step we look into the available historical declarations. If the last *Tr2* (see Table 4.2) historical periods we did not have a tax declaration, we assume that also in the current period there will be no turnover. In case there is historical turnover we go to the fourth step. If the unit has been declared inactive from the second half of the reporting period (value *Tr3* in Table 4.2) we assume that we do not expect turnover anymore. Otherwise, turnover is imputed.

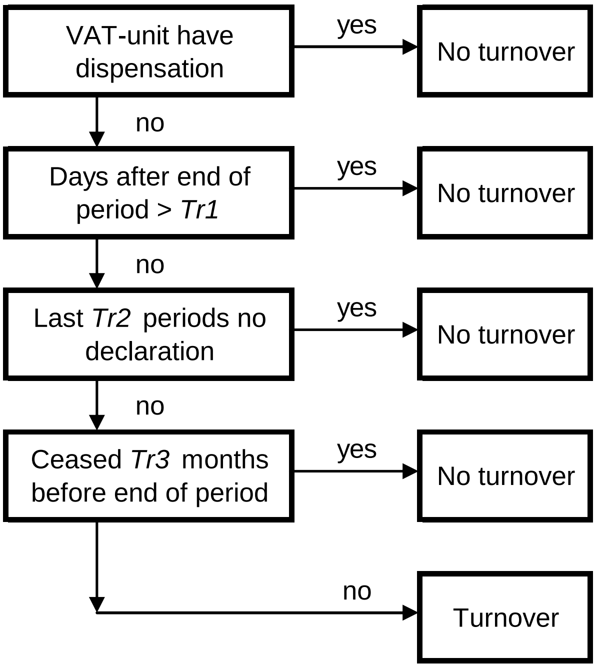
VAT-unit have dispensation
yes
No turnover
no
Days after end of period > Tr1
yes
No turnover
no
Last Tr2 periods no declaration
yes
No turnover
no
Ceased Tr3 months before end of period
yes
No turnover
no
Turnover

Figure *4.1.*Decision tree to determine whether turnover is to be expected or not.

Table 4.2. Threshold values depending on periodicity of tax declaration

Threshold value	Periodicity of tax declaration		
	Monthly	Quarterly	Yearly
Tr1	130 days	130 days	130 days
Tr2	5 months	2 quarters	2 years
Tr3	0.5 month*	1.5 month	6 months

* In fact we check whether the unit is active on the first day of the month but not on the last day of the month.

The first settings that we use for parameters *Tr1*, *Tr2* and *Tr3* are given in Table 4.3. The parameters can be determined by computing the difference between the imputed turnover at flash estimate versus the observed turnover at final estimate and likewise for the regular versus the final estimate, and minimising that difference over the publications cells.

4.3.2 Introduction to the formulas

For those VAT-units where we expect turnover, we will compute an imputation value. Also for non-responding enterprises that received a questionnaire we compute an imputation value. For both unit types we use the same formulas. In the explanation of the notation, below, we use "unit" where unit can be a VAT unit or an enterprise. The unit types will be specified when needed. As explained before, the VAT units can report on a monthly, quarterly or yearly basis. Therefore, the "period", as given below, can refer to a month, a quarter or a year. This will be specified later when needed.

We use the following notation:

O_i^t observed turnover of unit *i* in period *t*

$G^{t,s}$ ratio of turnover of period *t* compared to period *s*

$G_B^{t,s}$ ratio of turnover in a panel of units in reference group *B* of period *t* compared to period *s*

$R_{B(t,s)}$ set of units in reference group *B* that responded both in period *s* and *t*

In the following sections 4.3.3–4.3.5 we describe four imputation methods in case of lack of observations. The order in which the imputation methods are used depends on the availability of auxiliary variables and on the number of units available to estimate the imputed value. The order of the methods is described in section 4.3.6. All the formula given in section 4 are computed per stratum *h*. The size of these strata is discussed in section 4.3.6. For simplicity of notation subscript *h* is omitted.

4.3.3 Pattern 1a: Units with historical turnover values

For units that have historical turnover values, the imputed value of unit *i* for the current period *t*, $\hat{O}_i^t$, is computed as:

$$\hat{O}_i^t = \hat{G}_B^{t,s} O_i^s \tag{1}$$

where a hat stands for an estimation, *s* stands for a historical period, and $\hat{G}_B^{t,s}$ stands for the turnover ratio for period *t* and *s* in a reference group, with

$$\hat{G}_B^{t,s} = \frac{\sum\limits_{j \in R_{B(t,s)}} O_j^t}{\sum\limits_{j \in R_{B(t,s)}} O_j^s}. \tag{2}$$

We distinguish between two methods. We use either the turnover ratio of the current period compared to the corresponding period of a year ago (method A), or the ratio of two subsequent periods (method B). In Table 4.3 we specify the formula for the periodicity of response (monthly, quarterly and yearly). Responding units in the reference group that declare tax on a monthly basis are used to impute non respondents that declare on a monthly basis. Likewise, responding units in the reference group that declare tax on a quarterly basis are used to impute non respondents that declare on a quarterly basis.

For units that report monthly, we impute only turnover for the months that are missing. Quarterly turnover of units that declare on a monthly basis is obtained as the sum of the three corresponding monthly turnovers. For units that report quarterly we impute a quarterly turnover. For units that report yearly, we also impute a quarterly turnover for period *k*(*y*) which stands for quarter *k* in current year *y*, and use a growth rate based on the turnover of *k*(*y*) compared to the yearly turnover of last year (*y*–1).

Table 4.3. Specification of method A and B for periodicity of response

Periodicity of response	Duration of historical period *s*	Duration of actual period *t*	Specific notation	Method A	Method B
Monthly	month	month	*t*=*m*	*s*=*m*–12	*s*=*m*–1
Quarterly	quarter	quarter	*t*=*k*	*s*=*k*–4	*s*=*k*–1
Yearly	year	quarter	*t*=*k*(*y*)		*s*=*y*–1

4.3.4 Pattern 1b–c. Units without historical turnover

When a unit has no historical turnover we impute an average value. Again we have two methods: C and D. Method C makes use of an auxiliary variable, namely the number of working persons. Turnover of unit *i* is imputed according to:

$$\hat{O}_i^t = \hat{\bar{Z}}^t \times m_i^t \times WP_i^t \qquad (3)$$

where WP_i^t stands for the number of working persons of unit *i* in period *t*, m_i^t stands for the number of months in period *t* that unit *i* is active and $\hat{\bar{Z}}^t$ stands for the estimated average monthly turnover per working person among a reference group of respondents. $\hat{\bar{Z}}^t$ is computed as

$$\hat{\bar{Z}}^t = \frac{\sum_{j \in R^t} O_j^t}{\sum_{j \in R^t} m_j^t \times WP_j^t} \qquad (4)$$

where R^t stands for the respondents in a reference group for which turnover and number of working persons are available.

Method D can be used when neither historical turnover nor number of working persons is available for unit *i* with a lacking observation. Turnover is then imputed according to the average of a reference group:

$$\hat{O}_i^t = \hat{\bar{O}}^t = \frac{1}{r^t} \sum_{j \in R^t} O_j^t \,, \tag{5}$$

where r^t stands for the number of respondents in reference group R^t. In Table 4.4 we specify the formula of methods C and D for periodicity of response (monthly, quarterly and yearly).

Table 4.4. Specification of method C and D for periodicity of response

Periodicity of response	Duration of period *t*	Specific notation
Monthly	month	*t=m*
Quarterly	quarter	*t=k*
Yearly	quarter	*t=k*

For units that declare tax on a yearly basis, method C and D is only used for the flash and regular estimate. Their imputation values are renewed at the final STS estimate when their declared yearly turnover is available. Then the units are imputed according to the method described in section 4.5.1.

4.3.5 Some exceptions to methods A–D

The above described methods cannot always be applied. We use the following exceptions:

- Units with negative turnover values in either historical period *s* or actual period *t* or both are excluded from the reference group when method A or B is applied.
- Units with negative turnover values in the actual period *t* are excluded from the reference group when method C or D is applied.
- If $\sum_{j \in R_{B(t,s)}} O_j^s = 0$ method A and B cannot be applied.
- When the composition of enterprises change between period *s* and *t*, the corresponding VAT units are excluded from the reference group, where *t* stands for the actual period and *s* for the corresponding period the previous year (method A) or for previous period (method B–D).
- The units to be imputed are excluded from method A and B when their turnover for historical period *s* is negative.

4.3.6 Order of methods A–D

The reference groups are determined by strata that are defined by the crossing response periodicity × size class (SC) × NACE code. Furthermore we use a minimum size of 20 units within a reference group in order to obtain a reasonable accurate estimate. We are aware that this is only rule a thumb; it might be better to estimate the accuracy of the imputed value and to use a minimum accuracy level.

Preliminary results have shown that growth rates (and means) differ per size class, economic activity and response periodicity. In order to impute as accurately as possible, the starting point is to use a reference stratum that is rather detailed: response periodicity × 1 digit SC code × 5 digit NACE code. If there are not enough units available we use a less detailed reference stratum. Figure 4.2 shows the methods crossed by detail of reference stratum. The cell with number 1 represents the most accurate method and the cell with number 64 the least accurate one. Depending on the available data and the number of units per reference stratum, the most accurate method is selected. The order has been filled in based on experience and expert knowledge. Section 5 describes a first test to determine whether the order is correct. After testing, the scheme of Figure 4.2 has been simplified.

Method/ Reference stratum	NACE 5-digit × SC 1-digit × Periodicity type	NACE 5-digit × SC 1 digit	NACE 5-digit × SC group	NACE 5-digit	NACE 4-digit × SC 1-digit × Periodicity type	NACE 4-digit × SC 1 digit	NACE 4-digit × SC group	NACE 4-digit	NACE 3-digit × SC 1digit × Periodicity type	NACE 3-digit × SC 1digit	NACE 3-digit × SC group	NACE 3-digit	NACE 2-digit × SC 1 digit × Periodicity type	NACE 2-digit × SC 1 digit	NACE 2-digit × SC group	NACE 2-digit
A	1	3	5	13	7	9	11	21	15	17	19	23	43	45	47	49
B	2	4	6	14	8	10	12	22	16	18	20	24	44	46	48	50
C	25	26	27	31	28	29	30	35	32	33	34	36	51	52	53	54
D	37	38	57	61	39	40	58	62	41	42	59	63	55	56	60	64

Figure 4.2 Order of methods A-D in relation to the level of detail of the reference stratum.

4.4 Missingness due to different unit structure

4.4.1 Introduction

In the topX entities, tax declarations have a many-to-one relationship with the enterprise group and within the enterprise group the declarations may be related to more than one enterprise. The question is then how to divide the turnover among the enterprises of the enterprise group. For a comparable situation, the German Statistical Office (Gnoss, 2010) uses a linear regression model where log-transformed turnover per enterprise is estimated from NACE code, from number of employees and number of local units. The resulting estimated turnover is summed up to the estimated total of a group of enterprises. The result is adjusted to the observed VAT turnover at enterprise group level.

Because in the Netherlands mainly for largest topX entities the tax declarations are related to more than one enterprise, SN chose send all topX enterprises a survey to ask for their turnover.

4.4.2 Non response in the survey of topX enterprises

Within the topX enterprises, a considerable number of VAT units is related to just one enterprise, see e.g. Table 3.7. We can use those VAT declarations in the case of non response in the survey of topX enterprises. If historical turnover is available for the non responding topX enterprise *i*, if VAT units have a many-to-one relation to this enterprise *i* and if the turnover of all those VAT-units is available then the imputed turnover $\hat{O}_i^t$ for the actual period *t* is given by:

$$\hat{O}_i^t = \hat{G}_i^{t,s} O_i^s \tag{6}$$

where $\hat{G}_i^{t,s}$ stands for the estimated turnover ratio of enterprise *i*. $\hat{G}_i^{t,s}$ is computed from responding VAT-units *j* as

$$\hat{G}_i^{t,s} = \frac{\sum\limits_{j \in R_{i(t,s)}} O_j^t}{\sum\limits_{j \in R_{i(t,s)}} O_j^s}. \tag{7}$$

where $R_{i(t,s)}$ stands for the set of units *j* that is uniquely related to enterprise *i* in period *t* and *s*. First we try impute with a yearly growth rate and if that is not possible we use a period-to-period growth rate. Note that the restrictions as described in section 4.3.5 should also be taken into account, e.g. $\sum\limits_{j \in R_{i(t,s)}} O_j^s > 0$.

When the conditions for formulas (6) and (7) are not fulfilled, we apply method A–D as described in section 4.3 but then directly at the level of the enterprise rather than at the level of the VAT units.

4.4.3 VAT declarations related to more than one non-topX enterprise

When VAT declaration *i* is related to more than one enterprise during period *t*, the observed turnover O_i^t will be divided among L related enterprises[5]. Below, we describe methods for two situations:

(I) each enterprise ℓ ($\ell = 1, \ldots, L$) is only related to VAT declaration *i*, and

(II) at least one enterprise ℓ ($\ell = 1, \ldots, L$) is not only related to VAT declaration *i* but also to one or more other VAT declarations.

Note that for the method of estimation it makes no difference whether the VAT declaration is related to more than one enterprise at one single time point or whether it relates first to enterprise A and then to enterprise B.

Situation (I) Each enterprise ℓ is related to one VAT unit

Denote R_ℓ^t as a reference population of enterprises that contains ℓ and $\hat{\bar{Z}}^t(\ell)$ as the average monthly turnover per working person for that reference population. Likewise to formula (4), $\hat{\bar{Z}}^t(\ell)$ is computed as:

$$\hat{\bar{Z}}^t(\ell) = \frac{\sum_{j \in R_\ell^t} O_j^t}{\sum_{j \in R_\ell^t} WP_j^t \times m_j^t} \tag{8}$$

where R_ℓ^t stands for the population of *j* enterprises in period *t* in a reference stratum that also contains enterprise ℓ, m_j^t stands for the number of months within a quarter that enterprise *j* is part of the quarterly population frame and WP_j^t is the number of working persons of enterprise *j* in period *t*.

Next, we make a preliminary turnover estimate of enterprise ℓ, denoted by $\tilde{O}_\ell^t$, as

$$\tilde{O}_\ell^t = \hat{\bar{Z}}^t(\ell) \times WP_\ell^t \times m_\ell^t \tag{9}$$

The final turnover estimate of $\hat{O}_\ell^t$ is obtained by calibrating the preliminary turnover estimates to the observed turnover of VAT declaration i, O_i^t:

$$\hat{O}_\ell^t = O_i^t \times \frac{\tilde{O}_\ell^t}{\sum_\ell \tilde{O}_\ell^t} \tag{10}$$

Enterprises with a negative turnover for period *t* are excluded from R_ℓ^t, but O_i^t is allowed to be negative. The determination of the reference stratum is likewise to

[5] These methods are not implemented yet in the production system

section 4.3.6. It should hold that $\sum_{j\in R_\ell^t} WP_j^t \times m_j > 0$ and $\sum_\ell \tilde{O}_\ell^t > 0$, otherwise a less detailed stratum must be taken. If the latter is not possible we do not use the VAT declaration but we impute a turnover for the non responding enterprise using the methods of “missingness due to lack of observations”.

Situation (II) Each enterprise ℓ is related to more than one VAT unit

If one or more of the enterprises ℓ is not only related to observation VAT unit *i*, but also to other VAT-units, then we try to use the whole procedure of situation I but in stead of enterprises we estimate the turnover of the related *legal units*. In the end we sum up the turnover of the legal units to the total of the enterprise. When that is also not possible – likewise to situation II – we impute at the enterprise level using the methods of “missingness due to lack of observations”.

4.5 Missingness due to different meaning of variable

4.5.1 *Pattern 3a. Observations available for a period longer than the target period*

The yearly VAT declarations are divided over the four quarters of the year by making use of the turnover distribution for a reference population, adjusted for the months that units are active during the year.

Denote R^y as the reference population with units *i’* for which we know the turnover of the whole year *y*, i.e. four quarters in case of a VAT unit that reports on a quarterly basis and 12 months for a VAT unit that reports on a monthly basis. The quarterly turnover of period *k* for the reference population, now denoted by $KO^k(R^Y)$, is computed as:

$$KO^k(R^Y) = \sum_{i'\in R^y} KO_{i'}^k \tag{11}$$

The fraction of quarterly turnover, $F_{KO}^k(R^Y)$, in quarter *k* of year *y* is given by:

$$F_{KO}^k(R^Y) = KO^k(R^Y) \Big/ \sum_{k\in y} KO^k(R^Y) \tag{12}$$

The quarterly turnover of period *k* for yearly VAT declaration of unit *i* is now estimated as

$$K\hat{O}_i^k = \frac{m_i^k F_{KO}^k(R^Y)}{\sum_{k=1}^{4} m_i^k F_{KO}^k(R^Y)} \times JO_i^y \tag{13}$$

where JO_i^y stands for the observed yearly turnover of unit *i* in year *y* and m_i^k stands for the number of months in quarter *k* that unit *i* is active.

Some special rules apply to the imputation method for pattern 3a. The stratum level containing the reference population is determined according to the scheme for method A as described section 4.3.6. Units are excluded from R^y when their quarterly turnover is negative. VAT units that are related to enterprises with a changing VAT unit composition during the year are also excluded from the reference group. The observed JO_i^y of the unit to be imputed is allowed to be smaller than 0. Also, units to be imputed are excluded from this method when they belong to an enterprise with a changing composition of VAT units during the year. When $\sum_{k \in y} KO^k(R^Y) = 0$ or when $\sum_{k=1}^{4} m_i^k F_{KO}^k(R^Y) = 0$ the method cannot be used. In those cases that the method cannot be applied, method A–D of section 4.3 is used.

4.5.2 *Pattern 3b. Observations available for a period that is shifted compared to the target period*

Some VAT-units declare tax for a three months period that is shifted compared to a calendar quarter. So far, this concerns less than 1 per cent of all VAT units that declare on a quarterly basis. Therefore, we use a simple correction. We attribute a shifted three months value to the calendar quarter that overlaps most, in terms of calendar days, with the shifted period.

4.5.3 Pattern 3c. Observations with differences in definition

VAT turnover may differ from target turnover. Below we describe solutions for two issues.

Issue (I) VAT declaration patterns

Some VAT-units have remarkable temporal VAT turnover patterns. For example, a unit declares exactly the same turnover during three subsequent quarters followed by a different turnover value in the fourth quarter. For those remarkable patterns we first sum up the turnover of the four quarters to a yearly turnover. Next, we estimate the quarterly turnover as described in section 4.5.1. These pattern corrections can only be done after the declarations for the whole year have been received, which corresponds to the final STS estimate.

Issue (II) Definition differences

For some VAT-units VAT turnover deviates from the target definition. We estimate the target turnover of period *t* of VAT unit *i* from the VAT data using a linear transformation:

$$\hat{O}_i^t = \hat{a}^y O_i^t(*) + \hat{b}^y / c \qquad (14)$$

where *t* can stand for a month, a quarter or a year depending on the response periodicity and *c*=1 for yearly turnover, *c*=4 for quarterly turnover and *c*=12 for a monthly turnover.

We estimate the parameters $\hat{a}^y$ and $\hat{b}^y$ using SBS survey data and VAT data at the level of the *enterprises* for historical year *y** (= *y*–2). We use only enterprises that are active during the whole year and for which we have response for both the SBS and the VAT; the latter needs to be complete for all underlying VAT units. The parameters in formula (14) are estimated by applying a linear regression to enterprises *j* within a stratum:

$$O_j^{y*}(SBS) = \hat{a}^{y*} O_j^{y*}(VAT) + \hat{b}^{y*} + e_j^{y*}, \qquad (15)$$

where e_j^{y*} stands for the residual of enterprise *j* in year $y*$. Parameters are estimated per stratum, where a stratum corresponds approximately with 4 digit NACE level. In fact base strata are chosen from which all output can be made.

The residuals in formula (15) are minimized using weighted least squares, where the weights are defined by $w_j^{y*} = 1 / \pi_j^{y*} O_j^{y*}(VAT)$. π_j^{y*} stands for the inclusion probability of unit *j* in year *y** of the SBS survey and is included so the regression represents the population. The component $1 / O_j^{y*}(VAT)$ is included because we assume that the variance is proportional to the size of the enterprise. We compute the standard error of the estimated parameters, accounting for the sampling design. When $\hat{a}^{y*}$ is not significantly different (t-distribution) from 1, we use $\hat{a}^{y}$=1 otherwise $\hat{a}^{y} = \hat{a}^{y*}$. Likewise, when $\hat{b}^{y*}$ is not significantly different from 0, we use $\hat{b}^{y}$=0 otherwise $\hat{b}^{y} = \hat{b}^{y*}$.

For NACE codes with a poor correlation between SBS and VAT data, i.e. a correlation coefficient smaller than 0.7 on log transformed data, target turnover cannot be estimated from VAT turnover. For those NACE codes we use sample survey data instead.

4.6 Some practical implementation issues

In the current section we mention some practical implementation issues. The first issue is that some of the missingness patterns described in section 4.3 – 4.5 can occur simultaneously. For example a unit that declares tax on a quarterly basis can be a non-respondent for the flash estimate. Simultaneously, this VAT unit can be related to two enterprises.

The different patterns of missingness are handled in the following order:

1. The VAT declaration patterns are corrected.
2. The target turnover is estimated from the VAT turnover. After step 1 and 2, the VAT turnover is comparable to the turnover of the survey data: both comply with the target turnover. In any imputation after step 2 that is done at the level of the enterprise, its source (VAT or sample data) is no longer relevant. Harmonisation is done before completion in order to have more enterprises available for the reference populations.

3. Turnover of VAT-units that declare on a yearly basis is divided over the four quarters of the year. Step 3 is only needed for the final STS and SBS release. For the flash and regular quarterly release this step is skipped.
4. Missing observations are imputed.
5. Turnover of step 3 and 4 is divided over two or more enterprises in case the VAT unit is related to more than one enterprise.

Furthermore, there are many implementation issues that concern the treatment of auxiliary information, such as the reporting periodicity of the VAT units, and the actual number of working persons and NACE code of the enterprise. We also deal with cases where auxiliary information is missing and cases with conflicting auxiliary information because it varies from period to period or by source.

5. Example of a test of accuracy of imputation methods[6]

5.1 Data and methodology of the test

5.1.1 General setting and data set

In this section we give an example of an accuracy test of the imputation methods. At Statistics Netherlands it is crucial to have a small difference between early and final quarterly estimates. In line with that, we test imputation methods in the case of “lack of observations” as given in section 4.3. We use VAT data and compared imputed values at an early date with observed values at final response. We limited ourselves to non-topX units.

We took VAT data from Q1 2008 – Q2 2010, where each quarter was linked to the population frame of the corresponding quarter. We removed non-domestic VAT units, VAT units linked to topX enterprises, VAT units that did not link to enterprises and VAT-units that linked to more than one enterprise. We also removed extreme values: about 500–1000 per quarter. The resulting data file contained 13.5 million records.

We show test results for the imputation of Q1 2010. The tests are done at 2- and 3-digit NACE level within the domain of the STS statistics. At 2-digit NACE level there are five very small strata, namely 06, 12, 36, 97, 99. Imputation results of those small strata were poor compared to the others. Those five strata were excluded from the results shown below.

Note that in the evaluation we have used two restrictions

[6] This section has slightly been modified compared to Chapter 4 of WP2 of the ESSnet Data Integration. Results of Table 5.2 and Table 5.3 were recomputed to exclude outliers in 2010Q1 and 2009Q1. In the original paper only outliers on 2010Q1 were excluded.

- we included only those units for which the quarterly turnover is complete at the final estimate.
- we included only those units that could be imputed by methode A–D

5.1.2 Indicators

To explain the indicators, we first introduce some notation:

$O^k_{h,early}$ total observed turnover of VAT declarations in quarter *k* and stratum *h* at the early estimate;

$\hat{O}^k_{h,early}$ total imputed turnover in quarter *k* and stratum *h* at the early estimate;

$O^k_{h,final}$ total observed turnover of VAT declarations in quarter *k* and stratum *h* at the final estimate.

Within each stratum, only those units that fulfil the restrictions mentioned in section 5.1.1 are included.

We evaluate the imputation results using three (base) indicators. The first base indicator is the relative difference for quarter *k* between the total turnover per stratum *h* based on observations and imputations at the early estimate and the total turnover at the final estimate:

$$D^k_h = 100\left(\frac{O^k_{h,early} + \hat{O}^k_{h,early}}{O^k_{h,final}} - 1\right) \qquad (16)$$

The second indicator is its absolute value denoted by $\left|D^k_h\right|$.

The third base indicator is the overall relative difference over all strata, given by

$$D^k = 100\left(\frac{\sum_{h=1}^{H}(O^k_{h,early} + \hat{O}^k_{h,early})}{\sum_{h=1}^{H} O^k_{h,final}} - 1\right) \qquad (17)$$

5.1.3 Description of the tests

<u>Test 1</u>

In test 1 we concentrate on method A and B of section 4.3. Section Figure 4.2 shows the crossing ‘method ´ reference stratum,’ a cell in this crossing will be referred to as a sub method. Each sub method has been given a number which shows the order that is used in production, as explained before in section 4.3.6. In test 1 we test the accuracy of the sub methods 1–24 at a registration date corresponding to roughly 50 percent response. The registration date is the date that the VAT-declaration of a company is registered at the tax office. As indicators we computed the average and the median of $\left|D^k_h\right|$ over the strata *h* at two and three digit NACE level.

To have a fair comparison between methods, we included only those records that could be imputed by all methods. The minimal size of the reference population was set to 20 units. We included only those strata *h* that fulfilled the above condition. Furthermore, we included 2-digit NACE levels, with non-response for both quarterly and monthly reporters.

Test 2

In test 2 we analysed the accuracy of the imputation in production given the order of the sub methods in Figure 4.2 at two registration dates corresponding to roughly 50 and 75 per cent response (see Table 5.1). As indicators we computed (1) D^k, (2) the average, median, 10 and 90 percentile of D_h^k, and (3) the average and median of $\left|D_h^k\right|$. As strata we took the two and three digit NACE level.

Table 5.1. Response rate at two registration dates for Q1 2010.

Registration date	Response (%)	
	Quarterly reporter	Monthly reporter, 3^{e} month
28 April 2010	58	54
30 April 2010	77	75

5.2 Test results and first conclusions

Test 1

Table 5.2 shows that indicator values for imputation accuracy of units that report on a monthly basis are much smaller than of those that report quarterly. This is simply because the indicators are computed for quarterly turnover. At the chosen registration date (28-4-2011) most monthly reporters have already declared turnover for the first two months of the quarter and about 50% of the units have reported for the third month.

Table 5.2. Average and median of $\left|D_h^k\right|$ (in per cent) for different imputation methods, Q1 2010 at 50% response.

Method / sub method	NACE 2 digit				NACE 3 digit			
	Quarterly reporter		Monthly reporter		Quarterly reporter		Monthly reporter	
	Avg	P_{50}	Avg	P_{50}	Avg	P_{50}	Avg	P_{50}
A								
1	1.37	0.83	0.23	0.13	1.52	0.92	0.51	0.23
3	1.19	0.80	0.23	0.13	1.48	1.08	0.51	0.23
5	1.16	0.71	0.23	0.08	1.55	0.95	0.45	0.20
7	1.34	0.84	0.22	0.10	1.48	0.87	0.49	0.21
9	1.17	0.65	0.22	0.10	1.43	0.92	0.49	0.21
11	1.09	0.68	0.23	0.12	1.53	0.93	0.43	0.22
13	1.21	0.64	0.25	0.14	1.51	0.81	0.47	0.21
15	1.24	0.84	0.21	0.12	1.51	0.90	0.49	0.20
17	1.14	0.72	0.21	0.12	1.48	0.92	0.49	0.20
19	1.14	0.64	0.23	0.14	1.61	1.05	0.44	0.22
21	1.16	0.61	0.25	0.15	1.50	0.85	0.46	0.21
23	1.23	0.67	0.25	0.16	1.58	0.90	0.45	0.22
B								
2	1.16	0.57	0.49	0.17	1.39	0.78	0.72	0.24
4	1.03	0.66	0.49	0.17	1.24	0.95	0.72	0.24
6	1.05	0.79	0.43	0.19	1.37	1.05	0.63	0.25
8	1.19	0.66	0.46	0.17	1.42	0.83	0.67	0.23
10	1.03	0.67	0.46	0.17	1.24	0.95	0.67	0.23
12	1.05	0.82	0.42	0.19	1.39	1.09	0.61	0.27
14	1.30	0.97	0.35	0.17	1.58	0.93	0.57	0.26
16	1.07	0.66	0.44	0.15	1.32	0.86	0.65	0.22
18	1.01	0.72	0.44	0.15	1.27	1.07	0.65	0.22
20	1.03	0.82	0.40	0.19	1.41	1.12	0.58	0.25
22	1.31	1.07	0.35	0.15	1.63	1.01	0.55	0.24
24	1.34	1.08	0.34	0.15	1.66	1.09	0.53	0.20

Table 5.2 shows that the average and median values of $\left|D_h^k\right|$ are larger at 3 digit than at 2 digit NACE level. Table 5.2 also shows that for quarterly reporters the results for method A, using a yearly growth rate, is more accurate than method B, using a period-to-period growth rate. For monthly reporters at 2 digit level, results of method A tend to be better than for method B but at three digit level their was no clear trend.

In addition, Table 5.2 shows some patterns in the performance among the sub methods for quarterly reporters at 2 and 3 digit level:

- sub methods that use strata that differentiate among periodicity type perform slightly less than those that do not; A (1 vs. 3; 7 vs. 9; 15 vs. 17); B(2 vs. 4; 8 vs. 10 and 16 vs. 18).
- sub methods that use a size class group perform nearly as good as than those that use a 1-digit size class; A (5 vs. 3; 11 vs. 9; 19 vs.17); B (6 vs. 4; 12 vs.10; 20 vs. 18).

For monthly reporters differences among sub methods were much smaller and such patterns were not found.

Test 2

At 2- and 3-digit NACE level (

Table 5.3) most indicators for imputation accuracy clearly improved from 50 to 75 per cent response. The difference between $P_{10}(D_h^k)$ and $P_{90}(D_h^k)$ is smaller at 75 than at 50 per cent response and also average and median values for $\left|D_h^k\right|$ are smaller. Note that the average value for $\left|D_h^k\right|$ is larger than its median, that means that the distribution of $\left|D_h^k\right|$ across strata *h* is skewed to the right. The average for D^k at 2-digit NACE level and 75 per cent response is minus 0.37 per cent, whereas D^k at the total STS domain was found to be minus 0.68 per cent (not shown). This - difference is because when computing the average, all strata have the same weight.

Table 5.3. Imputation accuracy for Q1 2010 at two NACE levels and starting dates.

NACE level	Response	D_h^k (in per cent)				$\left\|D_h^k\right\|$ (in per cent)	
		Avg	P_{10}	P_{50}	P_{90}	*Avg*	P_{50}
2 digit	50%	-0.73	-5.44	0.03	3.37	3.43	1.52
	75%	-0.37	-2.94	0.52	2.65	2.29	1.09
3 digit	50%	0.90	-6.72	0.31	10.83	5.44	2.53
	75%	1.21	-2.60	0.47	5.33	3.34	1.11

5.3 Some remarks

The results presented here are meant to illustrate how an imputation method can be tested. First test results suggest that the initially designed order as given in Figure 4.2 might be improved. Results in Table 5.2 suggest for example that for method A and to a lesser extent for method B, reference strata can start at 3 digit NACE level rather than at 5 digit level. That may be useful for reference strata with a limited number of units, as estimated growth rates may then be more sensitive to outliers. In practice we made already some adjustments to Figure 4.2. For instance, in case of method D, that imputes an average turnover, size class strata are no longer merged.

In Table 5.2 we presented average values for $\left|D_h^k\right|$, with all strata having a weight of one. In fact $\left|D_h^k\right|$ was found to be smaller with larger strata. When you wish to focus on improving imputation results for the larger strata, a weighed average of $\left|D_h^k\right|$ may be used with turnover as weights.

Finally we wish to remark that test results might depend on the method that has been used to remove extreme values. Preferably, imputation methods are tested with data that have been edited in production.

6. Improving robustness in special situations

6.1 Negative turnover values

The use of negative turnover values in the reference group can cause implausible imputation values. Therefore, we excluded them from the reference group as explained in more detail in section 2. Furthermore, negative historical values of a unit to be imputed can cause implausible imputation values when this unit is imputed using a turnover ratio of the actual to the historical period (method A or B). Therefore, those units are imputed using an average value based on the actual period (method C or D).

6.2 Stable versus unstable unit structure

Implausible imputation values can occur when in the reference group the VAT units that are related to enterprises change during the periods that are used to compute a reference turnover ratio or mean. This is best explained by giving an example.

Table 6.1 shows three VAT-units that belong to the same enterprise. The columns stand for the subsequent months of 2010. VAT unit 2 ended in month 5 and VAT unit 3 is new from month 4 onwards in the population frame. We can see that the enterprise has shifted turnover from VAT unit 1 and 2 to VAT unit 3.

Imagine that we make an early estimate for Q2 2010 and some units that declare on a monthly basis did not yet declare their tax over June. To impute the missing values according to method B, we compute a turnover ratio of June to May for a reference group. Say that unit 3 has responded at the early estimate but units 1 and 2 have not. If we would include unit 3 in the reference group of method B, we would overestimate the turnover ratio because we did not account for the fact that turnover is shifted from VAT unit 1 and 2 to VAT unit 3.

Table 6.1. Monthly turnover of three VAT units related to the same enterprise.

VAT Id	Monthly turnover (× 1000 euros)									
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
1	10.0	13.0	12.0	14.0	11.0	2.0	3.0	2.0	3.0	2.0
2	0.8	0.8	1.1	0.4	0.3					
3				0.0	0.0	12.0	13.0	15.0	14.0	12.0

To avoid implausible imputation values, VAT-units that belong to enterprises with a changing VAT unit composition during the period to be imputed, are excluded from the reference groups, see also section 2.

Note that implausible imputation values can also occur when we impute at the level of the VAT unit and the VAT unit *to be imputed* belongs to an enterprise with a changing composition of VAT-units. From

Table 6.1 follows that if we would use the historical monthly values (month 4) of unit 1 and 2, and unit 3 has already responded that we would overestimate the total turnover. We wish to make these imputations more robust by using the following rules:

- Enterprises where the composition of the underlying VAT units in the historical turnover differs from that in period *t* are imputed at the level of the enterprise.
- Enterprises where the composition of the underlying VAT units in the historical turnover differs from that in the actual period *t* are excluded from the reference group.

6.3 Late versus ended respondent

Figure *4.1* of section 4.3.1 presents rules to decide whether turnover is to be expected in case of lacking observations. For estimates that are made before threshold *Tr1*, we have a deterministic decision scheme that assumes turnover is ether expected or not depending on historical observations (*Tr2*) and the period that the unit is active according to the population frame (*Tr3*). This deterministic approach might lead to a bias in the estimation of turnover. Alternatively to using *Tr2* and *Tr3*, we could impute all units and multiply them by a ‘proportion late respondents’. This proportion represents the fraction of VAT units that did not respond at the time of the current release but does respond before *Tr1*. These proportions should be determined for a group of homogeneous units. One could think of NACE code, periodicity of response, size, number of months without historical observations etcetera.

Results in Table 3.4 about the Accommodation and food service activities showed that from the 26628 units that did not respond 25 days after the end of quarter Q4 2009, 26023 units (98 per cent) did respond at the final estimate. The remaining 605 units (2 per cent) were probably ended.

6.4 Dealing with frame errors in smallest size classes

6.4.1 Description of methodology

In preliminary tests we sometimes found extreme imputation values in the smallest enterprises (SC 0–2, up to two working persons), because there were units in the reference group with an unexpectedly large turnover. There are different reasons for those extreme turnover values. Firstly, some units truly have a very large turnover combined with only 0–2 working persons. This is especially the case with units dealing with “royalties” of artists and in the case of holdings. Both dominate in certain NACE codes. Secondly, the SC of the unit may be wrong due to a missing relation between the enterprise and a legal unit that does have a certain number of working persons. Also, the SC may be wrong because when the number of working persons is unknown a value of 0 working persons is taken during derivation of the GBR.

To give an idea about the frequency of this problem, we counted the number of VAT units with a quarterly turnover larger than 10 million euros in SC 0. Averaged over Q1 2008 – Q2 2010, there were 77 of those extreme VAT units per quarter on a total of 47 thousand VAT-units per quarter for SC 0.

To avoid implausible imputation values in the smaller size classes, we compute an imputation SC that can deviate from the coordinated SC in the frame. The imputation SC is derived as follows.

Firstly, for enterprises in small size classes (SC 0 and 1) we check whether the actual number of working persons of the enterprise corresponds with the coordinated SC. If the actual number working persons corresponds to a much larger size class (SC 6 and larger), the imputation SC is based on the actual number of working persons. The values for the lower size classes (SC0 and 1) and the upper ones (SC 6 and larger) have been taken from the next step.

Secondly, for the remaining enterprises, we compare their historical quarterly turnover with the median value per SC. If the turnover per enterprise is considered to be too large, the imputation SC will be larger than the coordinated SC. This is done as follows. Denote L_ℓ as the set of size classes of the small enterprises for which we compute an imputation SC. Some of those will be assigned a new, larger, imputation SC. The set of these larger imputation size classes are denoted by L_u. Note that subscript ℓ stands for lower and u for upper, with $\ell < u$. Let *sc* denote an individual size class and $O_{sc,med}^{k-1}$ the median quarterly turnover per SC of period *k*–1. Now we compute the smallest value of $O_{sc,med}^{k-1}$ within L_u, and the largest value of $O_{sc,med}^{k-1}$ within L_ℓ. For enterprise *j* we now check the conditions:

$$O_j^{k-1} > \min_{sc \in L_u} \{O_{sc,med}^{k-1}) \text{ and } O_j^{k-1} > \max_{sc \in L_\ell} \{O_{sc,med}^{k-1}) \quad (19)$$

If the conditions in formula (19) are not fulfilled, the imputation SC of enterprise *j* for period *k* equals the coordinated SC. Otherwise, the imputation SC is the SC for with distance d_j^{k-1} is minimal, with

$$d_j^{k-1} = |\ln(O_j^{k-1}) - \ln(O_{sc,med}^{k-1})| \quad (20)$$

and use as imputation SC of enterprise *j* for period *k* that SC for which d_j^{k-1} is minimal.

To get an idea about the threshold value that we needed to take for L_ℓ and L_u, we conducted a small test with real data (see appendix). This resulted in L_ℓ=1 and L_u=6.

7. Summing up and topics for further research

We have described a methodology to handle incompleteness due to differences in unit types, variable definitions and periods of observed data compared to target ones. We use a regression method to harmonise differences in definition and use imputation for completion. Our method of mass imputation can only be used to estimate a limited number of (related) core variables. Although we have described the methodology only for net turnover, we have implemented it for four related turnover variables; therefore we have some additional steps. When many variables need to be completed, mass imputation is not suitable because it is hard to impute values at unit level that are plausible for all possible combinations of variables.

We use an imputation method that tries to use all the available observations and impute only the “missing parts”. For each missing value we try to make use of ‘the best’ available auxiliary information. We took that approach to produce early estimates of good quality and to have good results for relatively small domains. The approach is flexible: models can be adapted if accuracy tests show that the quality of the imputations for certain missingness patterns is not good enough. Other NSI’s may have different unit types and likewise their missingness patterns may be different. Still, the general approach can also be used by other NSI’s.

The method presented in the current paper can be cumbersome to implement, mainly because we use various types of auxiliary information depending on the pattern of missingness. If an NSI is not interested in using all observed register data, the method can be simplified considerably by always imputing quarterly turnover at the level of the enterprise.

We see some possible refinements of the current approach. Firstly, we could use a weighted combination of a yearly (method A) and a period-to-period (method B) turnover ratio. Tests so far showed little difference in the quality of method A and B, but maybe the results of a composite estimator are more robust for small strata.

A second refinement would be to correct somehow for the difference in the turnover ratio of respondents to non respondents, for example by using time series techniques. Such a correction will only be an improvement when this difference is more or less stable over time.

A third optional refinement is to include the effect of the number of VAT units that is related to the enterprise into the imputation model of the reference group and of the recipient unit. So far, our experience is that the variation in turnover among VAT units that are related to the same enterprise is large. Therefore we do not expect that this will improve the results. Fourthly, rather than using a fixed order of imputation methods × strata, we could re-compute the order of the methods based on historical data – as part of the production process. Re-computation of the preferable imputation method during the production process is done by Finland (Koskinen, 2007).

In the case study presented, we first link the observed turnover of VAT-units to statistical units (enterprises), then complete turnover at the level of the enterprises and finally add up to obtain total turnover of the stratum population. Alternatively, we might have estimated the stratum totals directly from completing the turnover of VAT units, thus ignoring the relation with the enterprises. Problem with the direct approach is that we need to classify the units according to economic activity. At SN, the VAT units are classified by economic activity at the tax office, referred to as a 'branch code'. Preliminary research showed that this branch code deviates from the coordinated NACE code and is not good enough to determine small NACE strata. However, research may be done to find out whether the branch code is good enough to determine the total for the whole STS domain. For late estimates, when all VAT-units have responded, we could add up turnover to the STS domain. This total could then be used as a restriction to the imputations at micro level. Such a restriction may improve the quality of late estimates.

8. References

Bakker, B.F.M., 2011, *Micro integration: the state of the art*, Chapter 5 in: Report on work package 1 of ESSnet on data integration.

EC, 1993, *Council regulation (EEC) No 696/93 of 15 March 1993 on the statistical units for the observation and analysis of the production system in the Community*.

EC, 2006, *Commission regulation (EC) No 1503/2006 of 28 September 2006 implementing and amending Council Regulation (EC) No 1165/98 concerning short-term statistics as regards definitions of variables, list of variables and frequency of data compilation*.

Eurostat, 2008, *STS requirements under NACE Rev. 2., 28 August 2008, STS requirements based on CR 1165/98, amended by CR 1158/2005*.

Fisher, H. & J. Oertel, 2009, *Konjunkturindikatoren im Dienstleistungsbereich: das mixmodell in der praxis Statistisches Bundesamt*, Wirtschaft und statistiek **3**, 232-240.

Gnoss, R., 2010, Powerpoint sheets with a method from German Federal Statistical Office Destatis, as obtained from Ronald Gnoss.

Koskinen, V., 2007, *The VAT data in short term business statistics*, Seminar on Registers in Statistics - methodology and quality 21 - 23 May, 2007 Helsinki

Norberg, J., 2005, *Turnover in other services*, Statistics Sweden, Final Technical implementation Report. 8 June 2005.

Orchard, C., K. Moore & A. Langford, 2010, *National practices of the use of administrative and accounts data in UK short term business statistics*, Deliverable of work package 4 the ESSnet ‘Use of administrative and accounts data in business statistics’.

Orjala, H., 2008, *Potential of administrative data in business statistics: a special focus in improvements in short term statistics*, IAOS Conference on Reshaping Official Statistics, Shanghai, 14–16 October 2008.

Pannekoek, J., 2011, *Models and algorithms for micro integration*, Deliverable for ESSnet Data Integration work package 2.

Seljak. R., 2007, *Use of the tax data for the purposes of the short-term statistics*, Statistical Office of the Republic of Slovenia. Seminar on Registers in Statistics - methodology and quality 21–23 May, 2007 Helsinki.

Statistics Finland, 2009, *STS mini-workshop on 8–9 June 2009*, Meeting with experts from Statistics Netherlands, Statistics Sweden, UK National Statistics, Statistics Estonia, Statistics Finland and National Board of Customs as part of the MEETS project.

Vaasen A.M.V.J. & I.J. Beuken, 2009, *New Enterprise Group delineation using tax information*, UNECE/Eurostat/OECD BR seminar, Luxembourg 6-7 October 2009 (Session 2).

Van Delden, A., 2010, *Methodological challenges using VAT for turnover estimates at Statistics Netherlands*, Conference on administrative simplification in official statistics (Simply2010), 2–3 December Ghent 2010, Belgium.

Van Delden, A. & J. Hoogland, 2011, *Editing after data linkage*, Deliverable for ESSnet Data Integration work package 2.

Wagner, I., 2004, Schätsung fehlender Umsatzangaben für Organschaften im Unternemensregister (in German), Destatis, Statistisches Bundesambt, Wiesbaden.

Appendix. Test of methodology to correct for frame errors in size class

To get an idea about the threshold value for L_ℓ and L_u, we conducted a small test with real data. We took VAT data of Q1 2008 – Q2 2010 (10 quarters) and selected the domestic, non-topX VAT units and computed their quarterly turnover.

For each unit within size class (sc) 0 and 1 we determined whether they get a newly imputed size class or whether they keep their original size class when we take L_u to be 5, 6, 7 or 8 and when we compute the median quarterly turnover of a size class at 2- or 3-digit NACE code.

Next, we assumed that when the quarterly turnover of a VAT unit was larger than a threshold value, O_{tr} ., this indicates that the original size class the unit is wrong (i.e. a frame error). We computed results for three threshold values, namely 1, 5 and 10 million euros. For each unit we determined whether its size class was larger than this threshold or not.

Finally, for each quarter, we counted the number of units classified according to 'new size class' versus 'original size class' crossed by '$\leq O_{tr}$' versus '$>O_{tr}$'. We took the total outcome for 10 quarters. The results give an *indication* for false and true negatives and false and true positives at different values of L_ℓ and L_u with the median computed at two NACE levels.

Table 8.1 shows an example of the test results for size class 0, with the median computed at 2-digit NACE code and L_ℓ=0. Table 8.1 shows that the smaller L_u the more units were assigned to a new size class. Also, the smaller the value for O_{tr}, the more units had a value $>O_{tr}$. Assuming that the number of size class errors in the frame is limited, we considered O_{tr} = 10 million euros to be most realistic value. When the median is computed at 3-digit NACE level (not shown) the number of VAT units with a new size class is slightly smaller, but then we have more cases where we do not have the minimum of 10 units to compute the median.

We considered a false negative (cell 'old sc' × '$>O_{tr}$') to be much more severe than a false positive (cell 'new sc' × '$\leq O_{tr}$'), because assigning a new size class will probably lead to a reasonable imputation value. Based on the results we selected L_u=6: this leads to a limited number of false negatives and we avoid that the number of false positives becomes much larger than the number of true positives. In size class 1 (not shown) there was also a considerable number of units larger than O_{tr}, namely 229 at O_{tr}=10 million euros, with only 1 false negative. We therefore selected L_ℓ=1.

Table 8.1 Test of methodology to correct for errors in size class 0.

	L_u =8		L_u =7		L_u =6		L_u =5	
	$> O_{tr}$	$\leq O_{tr}$	$> O_{tr}$	$\leq O_{tr}$	$> O_{tr}$	$\leq O_{tr}$	$> O_{tr}$	$\leq O_{tr}$
$O_{tr} = 10 \times 10^6$								
New sc	581	290	759	913	767	2776	767	7403
Old sc	191	470736	13	470113	5	468250	5	463623
Total	772	471026	772	471026	772	471026	772	471026
$O_{tr} = 5 \times 10^6$								
New sc	753	118	1083	589	1358	2185	1389	6781
Old sc	651	470276	321	469805	46	468209	15	463613
Total	1404	470394	1404	470394	1404	470394	1404	470394
$O_{tr} = 1 \times 10^6$								
New sc	869	2	1574	98	2541	1002	4328	3842
Old sc	4535	466392	3830	466296	2863	465392	1076	462552
Total	5404	466394	5404	466394	5404	466394	5404	466394