Editing of errors associated with linking economic data sets to a population frame

Arnout van Delden and Jeffrey Hoogland

Discussion paper (201207)

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

60083201207 X-10

Editing of errors associated with linking economic data sets to a population frame

Arnout van Delden and Jeffrey Hoogland

Summary: In this paper we concentrate on methodological developments to improve the accuracy of a data set after linking economic survey and register data to a population frame. Because in economic data different unit types are used, errors may occur in relations between data unit types and statistical unit types. A population frame contains all units and their relations for a specific period. There may also be errors in the linkage of data sets to the population frame. When variables are added to a statistical unit by linking it to a data source, the effect of an incorrect linkage or relation is that the additional variables are combined with the wrong statistical unit. In the present paper we formulate a strategy for detecting and correcting errors in the linkage and relations between units of integrated data. For a Dutch case study the detection and correction of potential errors is illustrated

Keywords: Linkage error, relations between unit types, detection and correction of errors.

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 survey 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 Van Bemmel (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[1] 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 by Van Delden and Van Bemmel (2011) 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 survey and administrative data. Total turnover and wage sums are central to estimation of the gross domestic product, from the production and the income side respectively.

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

Because the current paper and Van Delden and Van Bemmel (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.

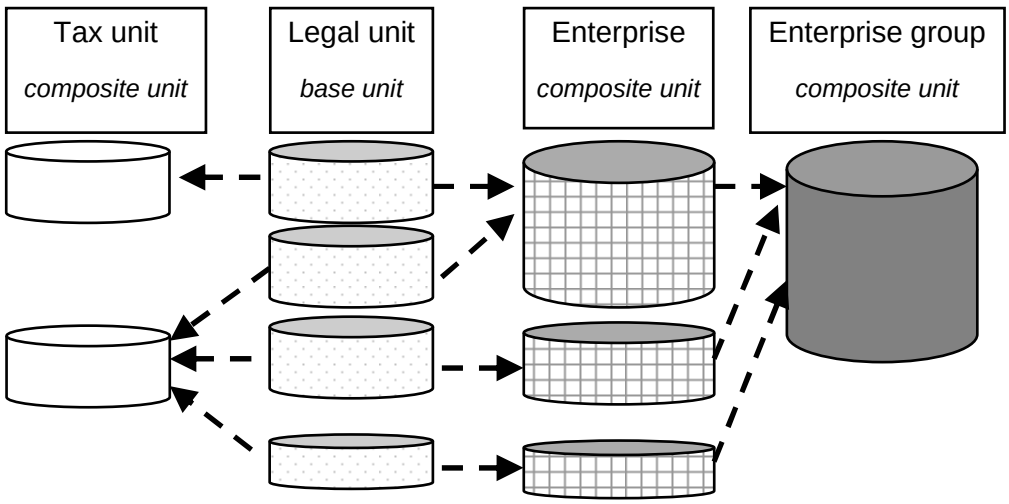

Figure *1.1.*: Different unit types in economic statistics. Each cylinder represents a single unit; arrows indicate the groupings of units.

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 (EC, 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

Within the GBR errors may occur in the relations between the unit types, which is explained in more detail in section 3. An example of a relation error is a tax unit in the GBR which is related to the wrong statistical unit. This statistical unit belongs to a certain NACE stratum. The consequence of this wrong relation may be that the tax unit 'belongs' to the wrong NACE stratum.

Also, linkage errors may occur between units in the data sets and the corresponding unit in the population frame, for example due to time delays. An example of a linkage error is a tax unit in the VAT declaration file that is wrongly not linked to the corresponding tax unit in the population frame because a new identification number is used in the VAT declaration file where the population frame still contains the old identification number.

The focus of the current paper is to describe a strategy for detecting and correcting errors in the linkage and relations between units of integrated data. For a Dutch case study the detection of potential errors is illustrated.

1.5 Outline of the paper

The remainder of the paper is organised as follows. Section 2 describes a Dutch case study. Section 3 gives a classification of the errors that are considered in the current paper. In section 4 we describe the strategy of detecting and correcting the errors. Section 5 gives an example of a preliminary test on the effectiveness of a score function that we use. Finally, section 6 concludes and gives topics for future research.

2. Description of the 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 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.

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

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.

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*.

The units and their relations are shown in Figure *2.1*. We consider:

1. the relation between the legal unit and the enterprise
2. the relation between the base tax unit and the legal unit
3. the relations between the VAT unit and the base tax unit

During the production of the GBR relation 1 is automatically derived from ownership relations in COC and tax office data, using business rules. Relation 2 is based on matching of name, postal code and house number, which Statistics Netherlands (SN) obtains from a National Basic Business Register (BBR). Relation 3 is automatically derived from tax office data using business rules.

The linkage between data sets and population frame is split into:

4. the linkage of a VAT unit in the VAT declaration to the (identical) VAT unit in the population frame
5. the linkage of an enterprise of the survey to an enterprise in the population frame

VAT declared by VAT units are linked to the corresponding VAT units in the frame, using exact matching of identification numbers (relation 4). Likewise, survey data as obtained for enterprises are linked to enterprises in the frame using exact matching of identification numbers (relation 5).

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.

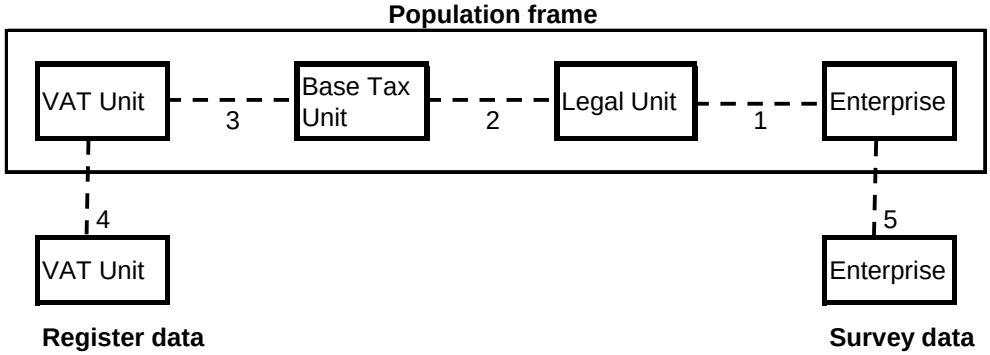

Figure *2.1.*: Relations between units types in the population frame, and the unit types of data sets that are linked to this frame.

2.3 Data

In the case study we use two types of source data. We use VAT data for non-topX enterprises. For topX enterprises we use survey 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 non-topX units, we only use observations of VAT units that are related to the target population of enterprises.

Concerning 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 data and survey data for Q4 2009 – Q4 2010 and focus on quarterly estimates. Data are stratified according to NACE 2008 classification.

3. Classification of errors in relations and in linkage between units

3.1 Overview

In the case study, errors in the relations/linkages as shown in Figure *2.1* seems most likely in relation 1 and 3. Errors in relation 1 and 3 are due to errors in the data sources used in the production of the GBR and due to time delays. For 'relation' 4 and 5 errors may occur in the exceptional case of a mistake in the identification

number. For more complex companies with more than one legal unit at the same address and postal code, errors may occur in relation 2.

In the next two sections we give a classification of errors that are considered in the present paper. We distinguish between (a) errors in the relations between different unit types within the population frame (section 3.2) and (b) errors in linkages between observations and the population frame (section 3.3). Note that the errors in the relations between unit types have been called *frame errors* by Bakker (2011) and *alignment* and *unit errors* by (Zhang, 2011).

3.2 Errors in the relations between unit types

Below we classify the different error types in order to understand the causes of errors and to assist the correction process. At SN relations between units are established in the GBR, from this GBR the population frame for a specific period is derived, see section 2.2. Any corrections in these relations are also made in the GBR and thereafter, new release of the population frame will be made.

Therefore, for the classification, we look into errors in the GBR and the consequences that they have for the population frame. We divide errors in the relations between unit types (see Figure *2.1*) into four subtypes: namely errors in relations that are present *versus* relations that are wrongly absent and errors in relations that result in coverage errors and those that do not.

Erroneous positive relations between unit types

(a) **Error leading to over coverage**. The presence of a relation between unit types in the GBR where at least one unit is non domestic, resulting in over coverage in the population frame.

For example, two VAT units are linked to an enterprise of the target population. One of the two VAT units is domestic, the other is non-domestic, see Table 3.1. According to the Dutch tax rules the non-domestic unit has to declare tax in the Netherlands and is found within the Dutch tax data.

Note that for error type (a) the relation itself may be correct, but because we wish to make statistics by country the non-domestic units should not be included.

(b) **Erroneous relation.** An incorrect relation between unit types in the GBR where all units are domestic.

For example a domestic VAT unit is related to the wrong enterprise. This wrong enterprise may belong to another economic activity than the correct enterprise.

Errors concerning missing relations between unit types

(c) **Error leading to under coverage.** A relation between unit types that is wrongly missing in the GBR, resulting in a domestic unit that is not covered by the population frame.

For example, an enterprise consists of two legal units, but only one of them is found in the GBR. Another example is a domestic legal unit that is present in the GBR, but is incorrectly not yet related to an enterprise.

(d) **Erroneous missing relation** An incorrect missing relation between unit types, where all the corresponding units are present in the GBR but just the relation itself is wrongly missing.

For example within the GBR, VAT unit A is related to enterprise A of the population, but should have been related to both enterprise A and B.

Table 3.1: Example of a domestic enterprise in the GBR related to a domestic and non-domestic VAT unit.

Enterprise	VAT unit	Domestic
47823457	0015720983001	Yes
47823457	0015720983002	No

3.3 Linkages errors between observations and the population frame

Likewise to errors in the relations between unit types in the GBR, errors in the linkage of data sets to the population frame can be divided into incorrect positive links (mismatches) and incorrect missing links (missed matches). In the case study we use exact matching, so we do not expect many linkage errors between observations and the population frame. Therefore we do not divide linkage errors into subtypes in the present paper.

4. Strategy of detection and correction of errors

4.1 Introduction

We distinguish between three phases in the production process where we can detect the above-mentioned errors, analyse them, and correct them if possible. The first phase is during the formation of the GBR. The second phase is just after linkage of VAT and survey data to the population frame. Those two phases focus on incorrectly missed links.

The third phase is done after the first estimation of population totals. Population totals are checked for implausible outcomes at aggregate level and next we zoom into the micro data using selective editing. In this phase we check for all possible sources of error. If a record value is suspicious we need aids to find out what type of error occurred. In section 4.2 – 4.4 we describe the three phases.

4.2 Phase 1: analysing VAT and legal units in the GBR that are not related to enterprises

Within the GBR, some relations between unit types may be absent. E.g., VAT units may, even though they might be related to legal units, not be related to enterprises. This happens e.g., due to the time delay the Dutch GBR applies when effectively introducing new enterprises. Phase one focuses on detecting *errors leading to under coverage* and correcting them. Wrongly missing relations lead to over coverage of the VAT unit population compared to the enterprise population.

To reduce the effects of this phenomenon we are thinking about two actions:

- Analyse VAT units in the GBR that are related to legal units, but are not (yet) related to enterprises. Sort these VAT units according to historical (quarterly) turnover and select the units with the largest turnover to be analysed first. Profilers should analyse these units in depth and decide upon possible relations with enterprises.
- Reduce the time delay between forming an enterprise and effectively introducing an enterprise in the GBR, by making use of information from other sources.

The first action tries to trace errors leading to under coverage and yields the introduction of new relations in the GBR. This can be effectuated in a new release of the GBR. At Statistics Netherlands it is not possible to “introduce” newly emerged enterprises in an already released GBR. The second action reduces coverage errors due to time delays.

4.3 Phase 2: analysing VAT units that cannot be linked to the population frame

Linking tax declarations to the population frame via VAT units, it turns out that not all VAT units in the tax-declarations-file can be linked to the population frame. Phase 2 tries to detect VAT units that are wrongly not linked to the population frame, this concerns *errors leading to under coverage*.

We should distinguish between two situations:

- Not possible to link a VAT-declaration to a VAT-unit in the population frame
- Not possible to link a VAT-declaration to an enterprise in the population frame

The first situation may occur e.g., when the file with the tax-declarations contains non-domestic units that are not present in the GBR. Another reason could be time delay: a new VAT-unit is already present in the tax-declaration-file but not yet present in the GBR. Again, sorting these VAT-units with respect to their turnover and profiling the units with the largest turnover, might help to reduce the effect of these linkage errors. First results for Q3 2010 show that, after sorting, 1,5 per cent of

the units with a VAT-declaration that cannot be linked to the GBR corresponds to 90 per cent of the declared turnover.

The second situation is similar to the situation as mentioned in section 4.2. However, now we have additional information from the tax declaration that profilers could use in analysing the “missing” relations.

4.4 Phase 3: strategy of editing after making population estimates

4.4.1 Introduction

The third phase detects all kinds of errors. In terms of errors in the relations between unit types, we expect that in this phase we can find *erroneous positive relations* making use of the estimated stratum turnover levels and changes. Strata with extremely large stratum turnover values, may have ‘errors leading to over coverage’ or ‘erroneous relations’.

4.4.2 Indicators for detection of potentially wrong population estimates

For each publication cell (combination of NACE codes) we obtain a set of population estimates concerning turnover level and yearly/quarterly growth rate. We use several indicators to assess whether population estimates are plausible (Hoogland, 2011; Van Delden *et al.*, 2010). The main indicators are

- Difference between estimated growth rate and expected growth rate;
- Turnover share of unedited potential influential errors.

The difference between the estimated growth rate and the estimated expected growth rate is too large if

$$|G_{h,r}^{k,k-s} - E(G_h^{k,k-s})| > d_h^E \text{, with}$$

$G_{h,r}^{k,k-s}$: the estimated growth rate for quarter *k* with respect to quarter *k-s* for publication cell *h* and release *r*. In practice, we mainly consider *s*=4.

$E(G_h^{k,k-4}) = G_{h,r'}^{k-1,k-5}$, that is, the estimated expected year-to-year growth rate for a specific quarter is the year-to-year growth rate for the most recent release (*r’*) of the previous quarter.

d_h^E : user-specified marginal value for the difference between the estimated growth rate and the estimated expected growth rate.

The following indicator can be used to assess micro data that is used to estimate the yearly growth rate and turnover level for publication cell *h* in quarter *k.*

$$R_h^{k,k-4} = \frac{\sum_{j \in h} V_j^{k,k-4} \max\{O_j^k, O_j^{k-4}\}}{\sum_{j \in h} \max\{O_j^k, O_j^{k-4}\}},$$

where

$V_j^{k,k-4}$ = 1, if turnover value O_j^k for enterprise *j* in quarter *k* is a potential influential error (PIE) and it is not checked or an editor indicated that the record was checked, but there was insufficient information for editing. To determine whether O_j^k is a PIE O_j^{k-4} is used as a reference value.

$V_j^{k,k-4}$ = 0, otherwise.

In the next section we explain how potential influential errors can be detected using reference values.

4.4.3 Indicator(s) for detecting potential errors at microlevel

To detect potential influential errors, score functions are used to assess the influence and risk associated with the net total turnover for an enterprise in a publication cell and quarter. A detailed description is available in Van Delden *et al.* (2010). The basic idea is described in Hoogland (2009). The turnover values O_j^k and O_j^{k-4} are used to determine the influence and suspiciousness of O_j^k. These turnover values can be either observed or imputed. We try to construct homogeneous strata in order to detect deviant turnover values and growth rates within strata. The score for influence (*I*) and suspiciousness (*S*) for a specific enterprise *j* and quarter *k* are multiplied:

$$R_j^{k,k-4} = I_j^{k,k-4} \times S_j^{k,k-4}$$

The total score $R_j^{k,k-4}$ is between 0 and ∞, and a higher score means that the net total turnover for an enterprise is considered more influential and/or suspicious. All enterprises *j* with a value

$$R_j^{k,k-4} \geq R_{\min}$$

are listed on the 'PIE list' that is shown to analysts (see below).

We give a rough description of the score functions used to assess influence and suspiciousness. To assess the influence for enterprise *j* we use the turnover values O_j^k and O_j^{k-4}, and the robust estimates of O_j^k and O_j^{k-4}. The idea is that we do not want to underestimate the influence of an enterprise if a turnover value is too small. The influence is large if the maximum value of these four turnover values is large relative to an estimate of the total turnover in the publication cell.

To assess the suspiciousness for enterprise *j* we compute partial suspicion scores. These partial scores represent suspiciousness regarding one of the features below

$S_{1,j}^{k-4}$: turnover value in quarter *k-4;*

$S_{2,j}^{k}$: turnover value in quarter *k;*

$S_{3,j}^{k}$: yearly growth rate in quarter *k*;

$S_{4,j}^{k}$: inverse of yearly growth rate in quarter *k*;

A feature is suspicious if the corresponding partial suspicion score is larger than 1. This is the case if a feature is extremely high within a stratum with enterprises, otherwise the partial suspicion score is equal to 1. For each enterprise we determine whether a feature is suspicious. A 4-digit code (I_1, I_2, I_3, I_4) is used to summarize the suspiciousness of features, see figure 4.1.

I_1 = 1, if $S_{1,j}^{k-4}$ is suspicious, otherwise I_1 = 0

I_2 = 1, if $S_{2,j}^{k}$ is suspicious, otherwise I_2 = 0

I_3 = 1, if $S_{3,j}^{k}$ is suspicious, otherwise I_3 = 0

I_4 = 1, if $S_{4,j}^{k}$ is suspicious, otherwise I_4 = 0

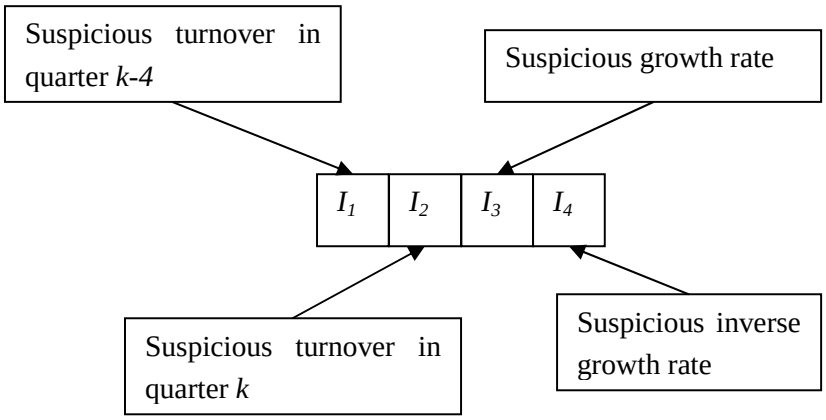

Figure *4.1*. 4-digit code for suspiciousness.

The partial suspicion scores are combined into an overall suspicion score.

$$S_j^k = \max(S_{1,j}^{k-4}, S_{2,j}^{k}) S_{3,j}^{k} S_{4,j}^{k} - 1$$

The overall suspicion score is 0 for enterprises where all partial suspicion scores are equal to 1, otherwise it is larger than 0.

An analyst decides whether enterprises on the PIE list are edited. He/she may consider additional information such as sorted tables with micro data and scatterplots showing the relation between a turnover variable in quarter *k* and the same turnover variable in quarter *k-s*. This may result in enterprises on the PIE list

that are not selected for editing or additional enterprises with potential influential errors that have been selected for editing by an analyst.

4.4.4 Correction at microlevel: determining the error type

An editor has to determine the type of possible errors and a correction if there is an actual error. In the present paper, we focus on the detection of errors in the relation between unit types and in linkage errors. There are several aids for detection of relation and linkage errors

1) partial suspicion scores
2) seasonal patterns of turnover for an enterprise
3) seasonal patterns of turnover for VAT units related to an enterprise
4) features of an enterprise, linked legal units, and linked VAT units in combination with profiler knowledge

ad 1) Partial suspicion scores

The suspicion code gives an indication of the possible error type of a potential influential error that is detected by a score function. For example, suppose that an enterprise has a suspicious turnover value in both quarter *k* and quarter *k*-4 (code 1100). This indicates a potential error in the NACE code or size class, or a possible relation/linkage error that has been present for more than a year. Suppose that an enterprise has a suspicious turnover value in quarter *k*, a 'normal' turnover in quarter *k*-4 and a suspicious growth rate (code 0110). This is an indication of a possible erroneous turnover value or a possible linkage error in quarter *k*.

ad 2) Seasonal patterns of turnover for an enterprise

Another indication of a possible relation/linkage error is a shift in turnover for an enterprise from quarter *k*-1 to quarter *k*, in combination with a change in the VAT units that are related to the enterprise in quarter *k*. An enterprise is *stable* in quarter *k* if it has the same relation with VAT units as in quarter *k*-1, otherwise it is *unstable*.

ad 3) Seasonal patterns of turnover for VAT units related to an enterprise

In Table 4.1 an example is shown of a change in the VAT units that are related to an enterprise, which seems to cause a zero turnover in quarter 3. It is plausible that VAT unit 333301 already declared turnover for the third quarter, but was (wrongly) not yet related to the enterprise 2 within the GBR.

Table 4.1. Turnover (× 1000 euros) patterns of VAT units related to enterprise 2.

VAT unit	Period	turnover
2222201	Quarter 1	2000
2222201	Quarter 2	2500
2222201	Quarter 3	0
3333301	Quarter 4	2200

In Table 4.2 an example is shown of an additional VAT unit that is related to enterprise 1 with a large effect on the quarterly turnover of enterprise 1. It seems that the large increase in turnover is not due to an erroneous turnover, because a large turnover is reported in each month in Q4. Features of VAT unit 3333301 and the corresponding legal unit and enterprise help to determine the type of possible error.

Table 4.2. Turnover (×1000 euros) patterns of VAT units related to enterprise 1.

VAT unit	Period	turnover
2222201	Quarter 1	20
2222201	Quarter 2	25
2222201	Quarter 3	24
2222201	Quarter 4	26
3333301	Q4, month 1	9000
3333301	Q4, month 2	10000
3333301	Q4, month 3	12000

ad 4) features of an enterprise, linked legal units, and linked VAT units in combination with profiler knowledge

An editor could check the available name and address of VAT units, and name address, legal form and number of working persons for legal units to check whether all VAT units are correctly related to the enterprise. Within the population frame information is available about changes in the composition of the enterprise (such as mergers). An event such as a merger may explain a structural shift in turnover for an enterprise and changes in the relation between the enterprise and VAT units.

5. Preliminary test on the effectiveness of score functions

We consider all enterprises that were edited within the DRT project for car trade and wholesale trade for Q4 2010. We use VAT data that were not edited before and survey data that were already edited for production of STS with the current statistical process. We investigate the relationship between the type of error and whether an enterprise is on the PIE list.

The ratio ‘number of records on PIE list that are selected for editing / total number of records selected for editing’ says something about the effectiveness of the score function used to detect potential influential errors. Assuming that an analyst makes an effort to detect potential influential errors that are not on the PIE list.

Table 5.1 shows preliminary results for edited enterprises for car trade and wholesale trade for Q4 2010. Only 76 of the 92.225 active enterprises are edited. For 90.457 active enterprises we used VAT data instead of survey data. The number of

edited records is small, because of time constraints and because part of the survey data was already edited for STS Q4 2010.

Table 5.1. Results for edited enterprises in 'car trade and wholesale trade', for the fourth quarter of 2010.

Error type	On PIE list	Not on PIE list	Total
Erroneous size class	2	2	4
Erroneous NACE code	0	0	0
Over coverage	0	1	1
Erroneous turnover value	13	23	36 (1)
Linkage error	1	5	6
Unknown error type	0	0	0
No error	7	22	29 (2)
Total	23	53	76

(1) 18 of them in survey data and 18 in tax declaration data

(2) 9 of them in survey data and 20 in tax declaration data

There were 38 enterprises on the PIE list, of which 23 were edited (Table 5.1) and 15 were not. From the 23 enterprises on the PIE list that were checked, the values of 7 enterprises were left unchanged (Table 5.1, category "no error"). Furthermore, about two third of the edited enterprises are not on the PIE list. This can mean that automatic detection of potential influential errors needs to be improved. It can also mean that some analysts select records for editing that are not influential or suspicious, or ignore enterprises on the PIE list that are influential and suspicious. The 4-digit code for suspiciousness shows that of the 38 enterprises on the PIE list, 28 enterprises have a suspicious turnover level, 7 enterprises have a suspicious turnover growth and 3 enterprises have both.

Table 5.1 shows that six of the edited enterprises contain linkage errors. Several editors indicated that it was difficult to detect linkage errors by means of the available information. That is, there might be more linkage errors in records that were selected for editing. Most of the enterprises where linkage errors are detected are not on the PIE list. We have to adjust parameters and strata for the score functions in paragraph 4.4.3 in order to detect more linkage errors.

Based on the comments that editors made we conclude that they were not always sure that data were correct, but they did not correct data in these cases. So the actual number of errors in the edited records might be larger. Furthermore, we discovered an implementation error in the computation of our suspicion scores in the production system: imputed values were wrongly not included in the PIE list. In practice, imputed values were often edited. The effectiveness of the score function as given in Table 5.1 is therefore probably underestimated.

6. Summing up and topics for further research

Statistical and public agencies face the challenge to obtain plausible population estimates from combined data sources with different units. In this paper we focused on economic data in which different unit types are used. The objective was to describe a strategy for detecting and correcting errors in the linkage and relations between units of integrated data. We used a case study concerning the estimation of quarterly and yearly turnover levels and growth rates, based on VAT declarations and survey data.

Potential influential errors have to detected, which can be due to errors in the relations between unit types in the GBR, from which the population frame is derived, and due to errors in linkage of data sets to the population frame. As a first result, a classification of relation and linkage errors is proposed based on our case study. We believe that the classification will also be useful for other National Statistical Institutes. We focused on errors in the relations between units, because we expect them to occur most in our case study. For the situation where errors in the linkage of a data set to a frame are likely to occur, the proposed classification can easily be extended using the categories: wrong positive linkage leading to over coverage, erroneous link, a missing link leading to under coverage and a missed link.

A second result from our study is that we split the detection of potential errors and their correction into different phases of the production process. We do so in order to find errors as early in the production process as possible although mainly errors leading to under coverage may be found in the first two phases. The use of different editing phases may also be interesting for any situation in which data are re-used. In the first phase we try to correct errors in the GBR as soon as possible, which is important as different statistics may use this GBR to derive a survey or population frame. The second phase focuses on VAT and survey data just after linkage. At SN, these data are stored in a central place, where it is made available for re-use by other statistics e.g. by international trade statistics. So also statistics that re-use data may profit from the data correction that has already been done.

Within each phase, we have some method that supports selective editing: we try to sort the errors according to their potential impact on the outcomes. In phase 1 and 2 we sort the units by turnover. In phase 3 we use a score function to detect potential influential errors.

In the first two phases we only deal with errors in the relations between units. In the third phase however, all kinds of errors may have occurred so we would like to be able to distinguish between them. We have developed several aids for the detection of a linkage error, such as a visualisation of seasonal patterns of turnover for VAT units related to an enterprise with a suspicious turnover. However, we believe that there is still a need to improve on this aspect. The strategy to distinguish linkage from other errors needs to be specified explicitly and should be automatized if possible. For instance, we can automatize the determination of the stability of the

relation between the VAT units and the enterprise. Additional indicators should be developed to determine the type of linkage error.

In terms of correcting the errors: in the case of economic statistics, so far, we need profilers to do this. Larger companies can have all kinds of constructions with underlying legal units and VAT units. Especially international companies may be very complicated and they may also be rather dynamic in terms of their underlying (base) units. In case of the smaller companies it may be useful to automatically collect internet information on name & activities. That may help to correct errors in the NACE code, for example.

In addition to the issues mentioned above, we came across several topics for future research. Additional VAT units may be linked using "loosened" matching rules as to increase the number of linked units. So far, tax and survey data are linked to the population frame using exact matching. Still, not all VAT-observations can be linked to the corresponding population frame which is partly due to time delays in the formation of the GBR. We found that we can link additional units by using just the fiscal base number, which is the core of the VAT identification number. Errors may occur more often when loosened linkage rules are used.

The efficiency of score functions might be improved. This can be done by optimizing the strata and parameters used to detect enterprises with a suspicious turnover. Another possible way to improve the efficiency of score functions is incorporate the effect of the number of working persons per enterprise on the turnover.

7. 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.*

EC, 2008, *Regulation (EC) 177 / 2008 of the European Parliament and of the council of 20 February 2008 establishing a common framework for business registers for statistical purposes and repealing Council regulation (EEC) 2186/93*.

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

Hoogland, J., 2009, *Detection of potential influential errors in VAT turnover data used for short-term statistics*, Paper for the Work Session on Statistical Data Editing in Neuchâtel, Switzerland, 5-7 October 2009.

Hoogland, J., 2011, *Editing of mixed source data for turnover statistics*, Supporting paper for the Work Session on Statistical Data Editing in Ljubljana, Slovenia, 9-11 May 2011.

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.

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.

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.

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

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. & K.J.H. Van Bemmel, 2011, *Handling incompleteness after linkage to a population frame: incoherence in unit types, variables and periods*, Deliverable for ESSnet Data Integration, work package 2.

Van Delden, A., P.P. De Wolf, R. Banning, K.J.H. Van Bemmel, A.R. De Boer, T. Carolina, J.J. Hoogland, M.P.J. Van der Loo, M.H. Slootbeek, P. Ouwehand & H. Van der Velden, 2010, *Methodology description DRT* (In Dutch), Internal report, Statistics Netherlands.

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

Zhang, L.C., 2011, *Topics of statistical theory for register-based statistics*, Paper presented at the ISI conference, Dublin 22–26 August 2011.