Discussion Paper

Issues when integrating data sets with different unit types

2017 | 05

Arnout van Delden

Content

Summary

Output of official statistics is increasingly based on integration of different data sources. From 1996 onwards a System of social statistical datasets has been developed at Statistics Netherlands that aims at the fast integration of survey and administrative data on persons and households. SN has not been able to develop a similar system for the fast integration of economic datasets for tailor-made publications. One of the reasons is that economic data concern all kinds of different unit types that may be difficult to integrate with each other. The aim of the present paper is to give an overview of the kind of issues that might occur when one tries to compile new publications, in a reasonably short period of time, that concern integrating data sets with different unit types or with unit types that are of a composite nature. We held an inventory of possible issues by studying twelve different cases at SN where data of different unit types have been integrated. We also used experiences from previous studies. We structured the issues in six stage types for both the units and for the variables. Based on the inventory and its structure, we identified 22 different issues. Nine of those are related to the use of different unit types. We give some first ideas on methodologies to address those issues.

Keywords

Data linkage, microdata, social statistical database, economic data, unit types

1. Introduction

Traditionally, the output of National Statistical Institutes (NSIs) consists of a fixed, pre-scribed set of output tables year after year. This output can be made by well-regulated production processes, often based on a single input source. This input is traditionally based on sample survey data, but the last decades also administrative data have been used more frequently. See for instance (Costanzo, 2011) for the increased use of administrative data in business statistics in European countries.

There are two on-going changes to this traditional way of producing output. First of all, statistical output is increasingly based on multiple input sources, or at least NSIs move into that direction (Constanse, 2014). Second, at least at Statistics Netherlands (SN) we move towards more, tailor-made output, based on requests from governmental organisations such as municipalities and ministries. Both changes require that different data sources can easily be linked and output be constructed.

For social statistics, a system of social statistical data sets (SSD) has been developed at SN to link and integrate administrative data with survey data (Bakker et al., 2014). The SSD has four central unit types: persons, households, buildings and organisations. Persons, buildings and organisations are available in administrative data with unique identification numbers. The households not available in the data source, but they are derived from person and address information. In order to link the data sets easily, and for confidentiality reasons, anonymised identification numbers are appointed: a personal identification number (PIN), an address identification number (AIN), a household number (HIN), and an organisation identification number (OIN) (Bakker et al., 2014). Next, these numbers are used to link different data sets.

For the economic statistics, since 1978 SN has developed a general business register (Dutch: ABR) integrating various administrative data sources. At first data sources from the chambers of commerce and trade organisations were used. Since 1994 also administrative data on relations by the tax authority are an important source of input. Using these administrative data sources, relations between administrative and statistical units were derived using automatic derivation rules. In addition SN has developed methods to combine specific data sources with statistical units. For instance, SN has developed a system to combine value added tax data with statistical units and also a system to link profit tax data to statistical units.

What we would like to achieve with the economic statistics reaches even further than the current possibilities: we aim to combine a wider group of data sources, being able to derive a larger diversity in target populations, and ideally also being able to combine data for tailor-made publications, so without having a fine-tuned system as is usually the case in regular production.

In the period of 2003–2006, SN has tried to develop an Economical Statistical Database (Heerschap and Willenborg, 2006; Hoogland, 2005; Hoogland and Verburg,

2006), but did not succeed. Two of the reasons were that (a) there were not so many administrative data sets at the time and (b) it is difficult to link data sets on business statistics at micro-level to each other. Linkage of economic data is generally more difficult than that of social data, because there are many different unit types in economic data.

In recent times many more administrative data sets have become available. Also a renewed project has been started in 2016 with the ambition to develop a system in which different economic data sets can be integrated. One of the pilots of this new project is the development of an administrative unit base, see section 6.2. The idea behind this administrative unit base is that a list is created of the full set of basic units (natural and legal units) that is relevant for the output that SN produces. Next, the relations/connections between those basic units and the various administrative data sources are determined and stored. Also the relations between the basic units and the composite statistical units are determined and stored. This approach ultimately aims to facilitate the combination / integration of various data sources at SN and aims to be versatile in the sense that it can support output for different target populations.

The units in the data sets that are to be integrated, which often concern different unit types, need to be harmonised to a statistical unit type. Commonly used statistical unit types in economic statistics are the *enterprise* and the *enterprise group*, but also other types are used, depending on the statistical output, such as the *establishment* for regional information. A total of eight different statistical unit types are mentioned in the European Council Regulation 696/93 for European Statistics (EEC, 1993). Survey data are mostly directly observed through the desired statistical unit type, but this data is often combined with other sources such as administrative data. In administrative data different unit types are found, such as legal units, value-added tax units and profit tax declaration units. One of the problems that might occur is that the administrative units do not link uniquely to the targeted statistical unit type.

Besides linking economic data sets with each other, SN is also interested to combine business data with person data and business data with address information. It is good to realise that the distinction between economic and social data is not so clear cut anymore. For instance the ABR contains more than 1 million one-man businesses that can also be found within the SSD as persons. In the Netherlands, an interrelated system of base registers has been developed by the government (European Commission 2015, pp. 27–30; Valk and Spooner, 2014) that is very helpful for SN to link all kinds of data sources. Still, we would also like to combine data that is not part of this system of registers and recently different kinds of linked / integrated data have been produced, and a number of problems were encountered. The aim of the present paper is to give an overview what kind of issues might occur when integrating data sets of different unit types.

The remainder of the paper is organised as follows. Section 2 explains the approach taken with the different case studies and gives a framework to structure the issues.

Section 3 provides an overview of the issues mentioned in those case studies and of some hidden issues based on past experiences and found by analysing the framework. In section 4 we present possible methods that might be used to cope with the issues. In section 5 we summarise and discuss the results. Section 6 is an appendix in which each of the case studies is worked out in more detail. Finally, section 7 gives abbreviations.

2. Approach

2.1 Case studies

A set of case studies at SN was selected in which data sets of different unit types have been combined, a short overview is given in Table 1. The case studies varied in the unit types that were involved in the different data sets. I distinguished four groups:

- data sets that combine administrative with statistical business unit types;
- data sets that combine administrative with statistical business unit types ánd also natural persons (with one-man businesses or partnerships);
- data sets that combine administrative with statistical business unit type ánd web sites; and
- data sets with different regional unit types.

We also included two case studies in which the unit types of the data sets are the same, but where for part of the units in the data set unique linkage keys were missing. These cases were included, because they have some interesting issues that also occur in the data sets where different unit types are combined.

Note that for each publication based of the combined data sets, also a common unit type needs to be selected. For instance, in the case on bankruptcies (the third case study in Table 3) one is interested to know how many jobs are lost due to bankruptcies. The jobs are available at the level of the enterprise, bankruptcies concern legal units and the target unit also concerns legal units. The target unit type is left out of the table, because for some of the more general case studies, such as global value chain of administrative unit, different publication are made, with different unit types. More information can be found in section 6.

- For each of the case studies we asked one or more informants to give some context of the case study at hand, and to describe the issues that they were confronted with. Finally we asked the informants whether there were issues for which methodological research is needed. Descriptions of each of the case studies can be found in section 6. Our prime aim was to find topics of future research, topics that the staff at the statistical production division would come up with as being important. Therefore, we did not use a systematic list of questions, but we asked open questions.

Table 1. Different case studies

Case study	Unit types involved
	admin & statistical business unit types
Global Value Chain	Enterprise Group, Enterprise, Legal Person
Bankruptcies	Enterprise, Legal person
	admin & statistical business unit types & natural persons
Administrative Unit Base	Natural person, Administrative unit types, Statistical Business Unit types
Agricultural Census	Natural person, Base Tax Unit, Enterprise, Reporting Unit
Family Enterprises	Natural person, Website, Enterprise Group
Self-employed	Natural Person, Legal Person, Enterprise
	admin & statistical business unit types & web sites
Centre of Policy studies	Multiple studies, e.g. Enterprise, Website; Enterprise, Units of trade organisations
Top Sectors	Enterprise, Units of trade organisations, Web site
Trade Organisations	Enterprise, Unit of trade organisations, Web site
	different regional unit types
Energy Consumption	Building Unit, Address
	natural person
Usual Residence	Natural Person
System of Social Statistical DataBases	Natural Person

2.2 Framework

In addition to the open questions and the case studies, we aimed to structure the findings in a logical way. We wanted to have a structure that also offered the opportunity to verify whether there were *hIdden issues*. With hidden issues we mean issues that will occur in practice, but that were overlooked. They may be overlooked because people have found ways to cope with them in practice, or people have accepted them and believe that they cannot be handled differently.

For this structure, we looked into literature on linking data sets. For instance, de Jong (1991) gives an overview of methodology for the linkage of data sets. We also used work done in the ESSnet on Data Integration by Bakker (2011) who described various stages of processing of administrative data. This work has been extended by Zhang (2012). Finally, we used the business architecture model SABSA (Sherwood Applied Business Security Architecture), which mainly gives an IT-perspective (TOGAF-SABSA, 2011).

We divide the data into six type of states (see Figure 1), that are passed by both the units as well as the variables going from the separate data sets up to the statistical estimates from the integrated set. We will first shortly mention the different state types. Thereafter we will the apply these state types to the units and variables involved in the integration of different data sets.

The first state type concerns the concepts of units and variables, which is also referred to as the *conceptual layer* in business architecture. The second type concerns the operationalisation of the concepts. With the operationalisation we strictly mean the 'rules' to make the concepts operational in practice; the actual realisation is not included. The operationalisation fits into the *logical layer* (or information layer) of the business architecture, since it describes the information that is held in the data (not the data themselves). The third state type describes the actual realisations in terms of units and data values. Here we arrive at the *physical layer* in terms of the business architecture. The same holds for the remaining state types.

The fourth state type concerns the relations between units and the relations between variables. The fifth type concerns the derivation from actual source units and values to their statistical counterparts. The sixth type of state concerns the estimation.

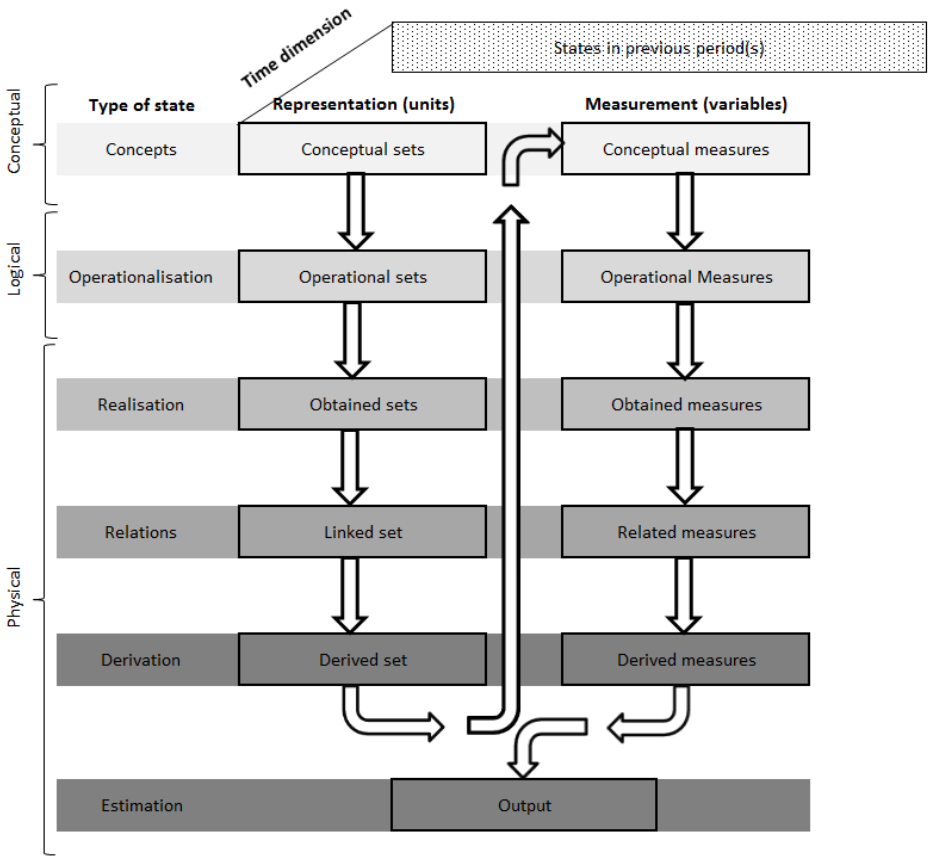

Figure 1. Structuring of states when combining data sets

Now, we will pass through these states, where we take the units of the two separate data sets as the starting point. The structure starts with looking at the *conceptual sets*, i.e. the conceptual definitions of the unit types and the populations of the data sets that are to be combined. An important question in this context is to what extent the definitions of these data sets coincide. Furthermore, one needs to know the whether the conceptual definitions of the unit types and the populations in the input data sets differ from the ones belonging to the intended output. Next, one considers how the concepts of the units in the population of the data sets at hand are translated into the *operational sets*. This latter may concern very practical issues. For instance, we have a commercial data set of business websites with identification variables that we would like to link to Dutch enterprises in a business register. It is important to know that there may be very recent websites that cannot be found in the business register because it may last up to three months before a new enterprise shows up in our business register, due to operational rules that determine whether a start-up is really economically active. Another example of an operational issue explaining why a website can erroneously not be linked to an enterprise is because the identification number that is stored along with the website is the hosting company or it is an old identification number. This is caused by the operational rules that determine which identification number is stored.

Next we look into the actual units and the obtained values of their identification variables that we can use to link the sets: the *obtained sets*. Before we can combine the data we might have to unify the formats of those identification variables. Depending on the outcome of the previous two stages, one will select a linkage method. After linkage we obtain a *linked set*. For instance one will use a 1-1 linkage method only when it is expected that both data sets cover the same population and when they are of the same size. Note that the linkage activity belongs to the physical layer in terms of IT.

Based on the linked data sets, where the relations between the units in the data sets are established, we might have to derive the set of statistical units of the intended micro data set, to arrive at the *derived set*. This is for instance the case when data are combined in order to compile the business register. After combining data on legal units and ownership relations, the statistical units are derived (as a composite unit of the underlying units). A similar process is described in the case study on the self-employed. When data sets are linked to a target population that is available in the business register, then this step "derived set of statistical units" has in fact already been done. To ease linkage of data, the units of the sources are given a unique (internal) identification number (in case that is not already present). The same holds for the derived statistical units.

After deriving the statistical units the same type of states as before are passed again, but now at the measurement side (the variables). We have now combined variables from different data sets and we are interested to estimate the statistical relation between them; some of the variables may be present in more than one data set within the combined data set. In order to arrive at the intended variables, we first look into the *conceptual measures* of variables within the data sets and those that

are needed for the intended output. We may have linked turnover data derived from value added tax data (VAT) to data on physical agricultural production (harvest per acre) and we expect the turnover to represent the sales value of the agricultural production. However, some of the agricultural companies do not have to pay tax for their sales, so we could underestimate the total sales if we do not correct for this (see case study agricultural census).

These concepts made operational in terms of the *operational measures*. The operational rules determine how we as a statistical office can differentiate between agricultural companies that are late with their tax declarations versus those that do not have to pay tax and will never declare. The next state are the *obtained measures,* that are the actual values that are reported in the data. For instance, part of the sales of economic activities are subject to a zero per cent tax tariff. This part of the sales belongs to the statistical turnover that we are interested in, but it may be underreported by companies. Then we move to the stage of the *derived measures.* This stage represents the final values of the variables, and hopefully they are close to the true values according to the target definition. We now have an integrated micro data set. Finally we will apply an estimation method to arrive at the output.

So far, we did not give much attention to effect of the time dimension when combining data sets. Economic data can be very dynamic, especially in terms of the legal and administrative units, and in terms of changes in ownership relations. Also, in social and economic administrative data time delays occur that may affect each of the states. When one is interested in output over time, one needs to decide in terms of units and measurements which of those changes are relevant for the output and which are not. Also the linkage step can be affected by the time dimension. The time dimension may influence each of the different states of both the representation and the measurement side.

3. Overview of the issues

In this section we give an overview of the main issues that were mentioned in the case studies, amended by issues that were identified in earlier studies or that we identified ourselves. We refer to the latter as *hidden issues*.

We start with three remarks, before discussing the issues. The first remark is that there are always two conditions that need to be fulfilled when a National Statistical Institute (NSI) aims to integrate data. The first condition is that the NSI needs to have permission to combine data, for instance by informed consent in case of survey data. The second condition is that confidentiality matters within the NSI need to be arranged properly, which has been mentioned in section 6.2. The second remark is that we included all issues that were mentioned by statistical practitioners, also when they were not strictly methodological. The third remark is that we tried to relate the issues to the states in Figure 1, where we selected the state which is closest to the

cause of the issue. That is not necessarily the same location as where the issue might be solved. For instance, issues that are caused by an operational definition of the population are placed in the state *operational set*, whereas the issue is dealt with during the linkage of the units in the population (the *linked set*).

3.1 Conceptual sets

3.1.1 Disagreement about the conceptual population

The agricultural census (section 6.4) provides an example of an issue that occurs when we try to combine data sets with populations that are based on different unit types, so populations that are based on different concepts. The agricultural census is currently managed by the organisation 'RVO.nl' and the population of units that they maintain is used for the RVO survey. This RVO survey serves four topics: agricultural census data, data for manure legislation, animal health data and emission data. Some of the units have non-commercial activities, but are still obliged by law to declare data for the manure legislation, for instance a children's farm. The RVO population is defined in terms of a functional approach, i.e. on the basis of the goods and services that the units produce.

SN likes to enrich the agricultural census data with fiscal data. In order to do so, SN links the RVO survey data and the fiscal data to a population obtained from our ABR based on *enterprises* as the statistical unit type. An enterprise is based on an institutional rather than a functional approach. An enterprise is, loosely defined, a unit that has a sufficient degree of autonomy in decision making and that sells its own goods and services *to a third party*. So the latter does not include a children's farm.

The consequence of these differences in unit types is that linkage of the RVO population to the ABR leads to a set of units that are in RVO population but do not have agricultural activities according to the ABR, because the ABR records the *main* economic activity of an enterprise. Likewise, when the fiscal data are linked to the population of enterprises, we find a large group of units with agricultural activities that cannot be found in the RVO population (see description in section 6.4). So far, SN and RVO.nl did not succeed in defining one common target population. The real issue here is that SN and RVO.nl have a different conceptual target population in mind.

3.1.2 Uncertainty about the unit type

Website are a potentially interesting source of information for 'businesses'. However, a 'business' is not a well-defined concept. An NSI first needs to map each website (unit) to a statistical unit of which the concept is well-defined. Websites of businesses are very diverse in nature and it will depend on the website to which concept it maps "best". Some websites describe information of an establishment, others about a legal unit, but there are also websites that contain information on an internationally operating enterprise group.

To make it even more complicated, the identification variables that are found in a website: 'business' name, phone number, address, legal unit number (LUN) or VAT number (VATN), often found on a "contact page", may differ from the actual content of the website. For example, the content of website may be close to the statistical unit type *establishment*, whereas the LUN on the contact page may concern a holding (*enterprise*) that owns the establishment or it concerns the LUN of the company that hosts the website.

Note that a statistical unit may be linked to multiple websites, where each website gives different pieces of information about this statistical unit, for instance different products or services. A website should be considered as a unit type on its own. See the case study on measuring the internet economy (Oostrom et al., 2016; see section 6.7).

In summary, one could formulate the problem of the current section as follows: we have a target population in mind for a certain publication, for instance enterprises, and we have a data source with unit types that are related to them but they do not exactly match them. Another example is that one wishes to publish data on a population of 100 × 100 m squares and we have a data set of values per municipality. Target and source population do have a whole-part relationship but this relationship is fuzzy.

3.2 Operational sets

3.2.1 Operationalising a concept

SN has recently developed a population frame, referred to as SZO, with statistical data on self-employed one-man businesses (natural persons). It was found to be not so straightforward to operationalise which persons have business activities. A number of administrative data sets are combined to derive whether the persons have business activities.

In fact the SZO data has multiple clients within SN that each have somewhat different concepts of business activities. Therefore different operational rules are used depending on the client. Examples mentioned in section 6.5 are legal units of natural persons, the business activities of natural persons and main shareholders with legal units.

3.2.2 Discrepancies between identification variables and unit type

The operationalisation of the identification variables of the unit types is sometimes a bit too 'wide' which might lead to discrepancies between the content of the information delivered and the unit type the identification variables refers to. This has been touched upon in section 3.1.2, where this was described for websites, but there the focus was on the uncertain identity of the unit.

Continuing with the website example, Oostrom et al. (2016) found that, when linking enterprises in the Dutch ABR to the variables that are found on the contact

information of that URL, often only a part of that contact information gave an agreement. One of the reasons for a disagreement is that not all linkage key variables related to the same unit type. For instance, the address may refer to an establishment, whereas the LUN may refer to a holding which is owner of establishment. Another example is an address on a website of a one man business that refers to its home address rather than the business address, as has been found in the case study of the bankruptcies (section 6.3).

A similar situation occurs with questionnaires of the agricultural census by RVO.nl. The information in the questionnaire is hopefully about agricultural activities (acres plants with different kinds of crops for instance). The value used for the linkage keys, e.g. name and LUN, sometimes do not link to an enterprise with agricultural activities, but with the holding, which is the *owner* of the enterprise with agricultural activities. The question then remains what would be the correct enterprise for which the agricultural activities are reported.

Other known examples of differences in the operationalisations of linkage keys are that in one data set the name is stored with the full first name whereas in another data set only the initials are stored. Or in one data set special characters, like the hat in û is stored in UTF-8 while in another data set only the u is stored.

3.3 Obtained sets

3.3.1 Misspellings in linkage key values

It is well-known that data without a unique linkage key often need data preparation steps before the data can be linked. In these data preparation steps differences in formats are harmonised and one might have to deal with errors or with spelling variations in linkage variables. In the case study on energy consumption (section 6.10), variations in spelling of street names and in the house numbers with suffixes occurred. Standardisation of the spelling was needed to improve the linkage result.

In the case study on trade organisations (section 6.9) errors in the LUN of membership lists occurred. Be aware that also the LUN reported on a website may be erroneous. One reason for the latter error is that the LUN has simply been entered manually, whereby a mistake has been made. Another reason is that the LUN refers to an old number which has not yet been updated.

3.4 Linked sets

3.4.1 Linkage errors

Maybe the most elementary issue that might occur in case of linking data sets is the occurrence of linkage errors. This issue has only shortly been mentioned in the case study of the usual residents and in the case study of the SSD. Linkage errors can be of two types, namely mislinks and missed links. A mislink occurs when non-identical units are linked (in the case of linking data sets with the same unit type) or when the

linked units do not have a part-whole relationship with each other (in the case of linking data sets with different unit types). Missed links occur when units are erroneously not linked.

Mislinks cause problems when variables that are measured in the linked data sets are related to each other. For instance, suppose we link a register with salary and job information to labour force survey data, using name and postal code as linkage key variables. We might then find persons that have a full-time job according to the register variable but that are unemployed according to the labour force survey variable due to mislinks. This leads to problems when we would like to achieve a microdata set that is internally consistent, for instance by using micro-integration (Bakker, 2011). A solution to this might be to use additional linkage key variables in the linkage process (for instance house number and date of birth). This will reduce the number of mislinks but it will increase the number of missed links. One needs to decide which levels of mislinks and missed links are acceptable, leading to the next issue.

3.4.2 Probability of linkage errors

In a study on measuring the internet economy (Oostrom et al., 2016) websites of an external company were linked to enterprises within the ABR. Within the ABR, the URL of the enterprise is available as a background variable, given that the URL is known. In about one-third of the cases Oostrom et al. (2016) were very certain that the link was a correct match: that were cases where both the name of the website as well as the LUN on the website was identical to that in the ABR. In two-third of the cases this was not true. In that situation, correspondence of other variables occurred, for instance business name, the phone number and an email address. Oostrom et al. (2016) defined five classes of linkage key agreements. It remained unclear however how certain they could be that a link was indeed a correct one. It would be useful to be able to compute a probability of a correct linkage.

Notice that the situation of estimating the linkage probability is more complex than usual, since we link two populations with different unit types. In our situation we may have '1 to 1', '1 to n' and ' n to m' links between the units of different populations. In the case study of the agricultural census (section 6.4) also time delays play a role. In the agricultural census, units of the RVO.nl population are linked to fiscal units. Those fiscal units may not have responded by the time of linkage. So there it is also important that estimate the probability that a linkage is missed due to time delays.

Also in the case study of usual residence (section 6.11) the issue of estimating correct linkage probability is mentioned, only then in the context of probabilistic linkage with non-unique identifiers.

3.4.3 Consistency in linkage over different data sets

Combining data is increasingly being done within statistical offices. That means that different organisational units within a statistical office may in fact link the same data sources to each other, but they may not be aware of that. What also might happen is that those data sources are not exactly the same, and they may even consist of

different unit types, but they have many '1 to 1' relations to each other. For instance, units within VAT declarations and those within IB tax declarations are in principle different, but in practice there are have many '1 to 1' relations. In terms of the output quality of the statistical office it is desirable that when the same data sets are linked (for different outputs) that the same relations between units types at micro level are established. This is not necessarily guaranteed, for instance because different linkage variables or different linkage methods might have been used. The issue of obtaining consistency in linkage between data sets has been mentioned by the AUB case study (section 6.2). One can understand this issue as an extension for the motivation to maintain a general business register within a statistical office.

3.4.4 Combining variables at micro level without overlapping units

Sometimes, the Centre of policy studies receives a request for a tailor made publication that would require information on the relation between variables that are observed in different sample surveys with hardly any overlap between the units. In that case linkage of the units in the different data sets does not offer a solution, but statistical techniques can be used to estimate the relation between the variables in the different data sets. Note that this situation is a fundamental issue, where the relation between the variables in the different data sets can only be estimated when making additional assumptions or when additional data sources are used. SN is now unable to fulfil those requests.

3.5 Derived set

3.5.1 Missing relations

Now, given the relations between unit types within the available data sets, we might need to derive statistical units. At SN, we have a business register where most of the statistical units are derived through automatic derivation rules that make use of relations, such as ownership relations between units (Vaasen and Beuken, 2009). Also derivation of households from dwelling information and from a PR uses automatic derivation rules using information on the relations between the persons that are registered at the same address.

Now a problem occurs when there are erroneously missing relations or when there are errors in the recorded relations. Especially the large and complex units, say international companies, are prone to such errors because there is a high frequency of changes in the structure / composition of the composite units. Therefore there are different NSIs, for instance SN, that have a special business division which manually collects information on relations between legal unit and their compositions in order to construct enterprises and enterprise groups. This issue has not explicitly been mentioned in one of the case studies, so we therefore refer to it as a hidden issue.

3.5.2 Time delays

A special cause for errors in linkage between units are time delays. Time delays may result in errors in the values of the linkage keys (see section 3.1.1) but it may also result in errors in the relations between units and subsequently in the derivation of

statistical units. For instance in our register with wages and social benefits data of employees of Dutch businesses (WSR) we sometimes find large groups of employees moving from one enterprise to another, which - depending on the situation - indicates a split or merger of an existing enterprise, well before those changes are found in the ABR. This WSR example also illustrates that relations between units are dynamic and we need to have means to account for that.
A time delay in the relations between units is another hidden issue, i.e. it has not explicitly been mentioned as a problem encountered in the case studies.

3.6 Conceptual measures

So far, we considered the challenge of combining sources from the viewpoint of the units. Now we will pay attention to issues related to the target variables in the data set. Similar to the approach that we took alongside the units, we start by considering the concepts of the variables in the sources.

3.6.1 Define meaningful output

An general issue that has been mentioned related to concepts and the use of different unit types is that of the definition of meaningful target variables. For instance, in the case of the bankruptcy statistics (section 6.3) monthly changes in the number of bankruptcies are only partly indicative for the economic business cycle. This is because the number of bankruptcies also depends on the complexity of the administrative structure of business entities. So a larger number of bankruptcies from one month to another may be due to the bankruptcy of one 'central' legal unit that is followed by the bankruptcy of a large number of interconnected legal units that in fact all belong to the same "business", while the economic impact of the bankruptcy (e.g. "number of jobs lost") may be limited. The challenge then is to relate the bankruptcies to the concept of a variable that is more indicative for the economic impact.

Population dynamics of units (mergers, splits, births and deaths) are also important to take into account when a defining meaningful output; this is explained further in the next section.

3.6.2 Aggregate concepts

The first, *hidden*, issue that is related to the concepts of the variables concerns the situation that categories of a classification variables are available at the level of a base unit in the original data set, whereas we aim to produce output on a composite unit classified by the 'same' classification variable. We refer to this as deriving an *aggregate concept*. The classical example of this issue is that of the economic activity code in business statistics, the NACE code. The economic activity code is appointed to a legal unit when this unit is registered at the chambers of commerce. However, a number of publications in business statistics, such as the structural business statistics, concern output at the level of the *enterprise* which may concern one or more legal units. The issue now arises how to apply the concept of a NACE code to this composite unit. Another example concerns that of the energy use per building

classified by purpose of use (school, hospital, dwelling). Purpose of use assumes the presence of units that are homogeneous in their use, but buildings may serve multiple purposes (see case study in section 6.10).

By convention, the NACE code of an enterprise is defined as its *main activity.* This solution is laid down in Council Regulation on statistical units (EEC, 1993) and also in the use of the international economic activity classification ISIC (UN, 2008), even in the first ISIC classification from 1948 (UN, 2008) this is mentioned. In fact this solution is not entirely satisfactory.

The use of a main activity to classify statistical units can lead to sometimes unwanted effects in the case of the estimation of changes over time. Changes in in ownership relations can lead to mergers and splits of enterprises while the underlying population of base units remains the same. One can for instance have the situation that the production value in one economic industry increases and in another one decreases due to an influential merger in the population. This unit consisted of two separate statistical units with different main activities before the merger, and after the merger they are classified in either of the two main activities. Here, the merger itself, in combination with the use of a main activity, caused the increase and decrease in growth rates, while there may no change in the 'real economy'. It is therefore important to incorporate effects of population dynamics into the definition of the output, to arrive at a meaningful output.

3.6.3 Disaggregate concepts

A second, *hidden*, issue that is related to the concepts of the variables refers to the opposite situation of an *aggregate concept*. Consider for instance the event of a bankruptcy, which by law refers to a legal unit. SN publishes data on number of bankruptcies, and also specifies this *by region.* So instead of going from a base unit to a composite unit (in case of the NACE code variable) the question now is how to define a variable at a more detailed level than at which it was available originally. Now suppose that on average one bankruptcy goes along with two establishments each in a different region. Then bankruptcies by region would lead to a doubling of the number of bankruptcies. This has been solved conceptually by considering a legal unit as being composed of multiple regions and by classifying the legal unit to the 'main' region only. The issue of defining a concept for a more detailed level is referred to as a *disaggregate concept*.

3.7 Operational measures

3.7.1 Operationalisation at a less detailed level; consolidation

An issue that has to do with the operationalisation of a concept, may occur when financial data are collected from interconnected units. For instance in the case study of bankruptcies, described in section 6.3, one is interested in debts due to bankruptcies for say 'economic actors'. Debts of bankruptcies of general partnerships (Dutch: VOF) may be reclaimed from the private capital of business partners who then might subsequently go bankrupt also. Just adding the debts of a VOF and of

their business partners may then lead to double counting since the debt of the VOF might have been passed onto the business partners. Instead, the outcomes need to be *consolidated* first before the total can be computed; consolidation means that the internal flows are excluded from the total. This requires information on the financial flows between the VOF and the business partners, or in general between parts of interconnected business entities. The main issue that occurs here is that a variable has to be operationalised at a less detailed level than the available data.

3.7.2 Operationalisation at a more detailed level; deconsolidation

An issue opposite to that in section 3.7.1 might also occur: the targeted variable concept is needed at a more detailed level than the data at hand. This occurred for instance in the case studies on trade organisations. In addition to output at the enterprise level one would like to have regional estimates, at establishment level, of economic variables. When the information on economic variables at enterprise level are detailed to the establishment level, the values of the variables also need to be deconsolidated. That is, the internal flows between the establishments now need to be added to the figures.

When computing energy use by business type, there are some addresses for which the energy use data are linked to buildings with multiple purposes (see section 6.10). Now the question arises: how do you operationalise the variable business type in case of a multiple purpose building? One can split the building into parts that are homogeneous in business type, or one can introduce mixed categories for the variable building type (e.g. use the class retail and wholesale trade), or one can classify the building by its main business activity.

3.8 Obtained measures

3.8.1 Missing values when using data of non-standard unit types or classifications

An issue that occurs is that of missing values for an auxiliary variable or for a target variable, which especially may occur when data with non-standard unit types or classifications are used. For instance we have the target variables 'number and kind of jobs' that are lost with bankruptcies. We have data on number and kind of jobs from business data, which are collected at enterprise level. For this special publication, we need data on number and kind of jobs at the level of legal units, which is a missing data problem.

Another missing data problem occurs with the case study of the "top sectors". Top sectors are special economic domains of which the government asked SN to measure their economic performance. SN uses all kinds of economic data collected at enterprise level. But the classification variable top sectors is a non-standard one, which is not included in our ABR. This classification can only partly be derived from the NACE code. Other parts of the top sectors are derived from membership lists. For units on those lists with a website, the top sector code is checked by manually

judging the content of a website. If needed, also the entity can be directly contacted, for instance by phone. This is a very time consuming activity.

Another example of missing data is the case study of the international trade statistics (ITS) (see section 6.1). The ITS data are classified by NACE codes. This NACE code is known for all domestic units, but about 20 per cent of the units for the ITS concern non-domestic units. These non-domestic units use the Netherlands for transit of goods. Currently the relative NACE distribution of the domestic units is applied to the whole ITS population, but the question is: is there a better way to classify the non-domestic units?

3.9 Related measures

3.9.1 Conflicting information in overlapping variables

When deriving statistical target variables, we sometimes have the same variables available from different sources. For the case study of the family enterprises for instance (see section 6.5), we want to derive a variable that indicates whether an enterprise is a family enterprise or not. In order to derive that variable, we use data on family relations. In that case study it was found that family relationships from parent-child data and those from household data did not always agree. Currently, in this case SN has the decision rule that if one source indicates a family relationship that we then assume this to be true. Maybe better solutions can be developed to handle these measurement errors.

3.10 Derived measures

3.10.1 Handling (selective) missingness

In section 3.8.1 we described the situation that one has observations for (a sample of) units, but for some units the values of a target or of an auxiliary variable are missing. A slightly similar issue is that we have the values for all the variables but for a selective part of the units in the population. Continuing with the example of the "top sectors", after we have obtained the classes of top sector codes for all units in the population, we linked them to the sampled data of, for instance, the structural business statistics. But not all sampled data of the structural business statistics belong to a top sector code and not for all top sector codes do we have a random sample of observations. In fact we may have a selective set of observations (see section 6.8.2), especially for the part of the top sectors where its code is not directly related to the NACE code. The question now arises: how to make an unbiased population estimate using those observations? Is it possible to derive weights to compute the population estimates?

3.10.2 Unit type for imputation

Missing values may be imputed. An issue is at which level the data are to be imputed. For instance, Van Delden and Bemmel (2011) describe the situation where turnover from VAT-units needs to be added to arrive at values for the statistical units. They

discuss whether missing values can be best imputed at the VAT-unit level or at the level of the statistical unit.

3.11 Output

3.11.1 Output quality of the linked set

The linkage between data sets can be incomplete and linkage errors might occur. The question that was raised in the case studies at the SSD (section 6.12) and at the Centre of policy studies (section 6.7) was "what is the effect of incomplete linkage and linkage errors on the accuracy of the statistical estimates based on those linked sources?"

Another question that may arise is how we can correct for bias in output tables or the estimated relations between target variables, when the data are based on linked sets with linkage errors (e.g. Chipperfield and Chambers, 2015).

3.11.2 Relating information at different hierarchical levels

The microdata linkage project (section 6.1) shows a new interesting issue that concerns the question how to relate outcomes that are based on different unit types. Statistical output by NSIs usually concerns a single unit type at a time, for instance all kinds of target parameters for a population of enterprises and for a population of enterprise groups or for persons. An example of output for social statistics with more than one unit type are frequency tables where persons (classified by gender, age etc.) are crossed with households (classified by type of household). Likewise, in economic statistics one could make a frequency table of establishments (classified by main activity) crossed with enterprises (classified by main enterprise activity) since this gives some information on the extent to which enterprises have multiple activities.

In the microdata linkage project (section 6.1) an example of interesting output concerns the question "what do decisions at enterprise group level imply for the kind of employees that are needed and what does this mean for the educational requirements of pupils?". The question now is: how can we make output that describes these relations between unit types?

4. How to address the issues

Here, we give some first ideas on how the issues given in the overview might be addressed. In section 4.1 the issues related to the unit and their identification variables (concerning the sections 3.1–3.5) are discussed. Next, in section 4.2 we discuss how we might address the issues related to the variables (thus those of sections 3.6–3.10) that are needed to derive the output. In section 4.2 we end with the issue related to the output (section 3.11).

Table 2. Issues concerning 'units and identification variables' and potential solutions. The left column gives the subsequent states types
(*Issues with an '*' prefix stand for hidden issues and those without an '*' are the mentioned issues.*)

		Consult experts users	Profiling	Derivation rules	Data preparation	Linkage methods	Statistical Matching	Quality measure	Other data sources	Model based estimation	Measurement errors models
CON	Disagreement about the conceptual population	x									
	Uncertain identity of the unit type	x									
OPE	Operationalisation rules depend on the client			x						x	
	Discrepancies between identification variables and unit type					x					
REA	Misspellings in linkage key values				x						
REL	Linkage errors					x					
	Probability of linkage					x					
	Consistency in linkage over different data sets					x					
	Combining variables at micro level without overlapping units						x				
DER	*Missing relations		x						x		
	*Time delays		x					x	x		

4.1 Units

In Table 2 an overview is given of the issues related to units and their identification variables and the possible ways to address them.

4.1.1 Conceptual issues

The two conceptual issues mainly require a careful consideration of experts and their solution may also depend on preferences of by the users, for instance to solve a disagreement in the target population.

There is one data source that needs special attention in relation to the concept that it relates to, and that is the website. The website as such can be considered a unit type and concept on its own. On the other hand, the content of the website also refers to a concept, but which concept this is requires an analysis of the website content. Natural Language Processing techniques for text categorisation are expected to be helpful to analyse a website content. Jacobs (1992) provides an example of natural language processing techniques to trace business names and industries from texts.

4.1.2 Operationalisation rules depend on the client

Let $I_i^{\mathcal{G}} \in \{0,1\}$ define a variable that indicates whether unit i belongs to a target population $\mathcal{G}$ or not, e.g. persons that have a one-man business or not. One way to determine $I_i^{\mathcal{G}}$ is to apply derivation rules to the values of the variables at hand that as close as possible operationalise the conceptual target population. This approach requires the availability of (identifying) variables for population $\mathcal{G}$. When those variables are not available, another way to estimate $\mathcal{G}$ might be to estimate a probability $P_i^{\mathcal{G}} = P(I_i^{\mathcal{G}} = 1|\boldsymbol{x_i})$ where $\boldsymbol{x_i}$ stands for a vector of background variables. This latter approach has been used in the past at SN, to correct population size for the set of enterprises in the ABR for the presence of over coverage.

4.1.3 Discrepancies between identification variables and unit type

To my knowledge there is no existing methodology to handle discrepancies between identification variables and the unit type they refer to. A first idea is to start by linking the values of the key variables (name, address etc.) in the data set to the identification variables of the targeted unit type themselves and subsequently linking them to the identification variables of the underlying, overlying or side-lying units that are directly related to the (composite) target unit type. Then after establishing a link between a unit in the data set and the composite target unit type, an estimate has to be made of the probability that the established relation is correct.

4.1.4 Misspellings in linkage key values

Misspelling of linkage key values is a known problem for which different methods have been developed already. Pre-processing steps to prepare data for linkage on string variables are standardisation (Winkler, 1995; Harron et al., 2016) and parsing. Standardisation is cleaning words from symbols, punctuation marks etc. Parsing concerns splitting text string into logical components. After data preparation, string comparison functions can be used in the linkage step, such as the Jaro-Winkler distance, the N-grams distance and Soundex (a phonetic system). See Gu et al. (2003) for more methods.

4.1.5 Linkage errors and probability of linkage

A central issue in the present paper is to determine the relationships between the units in the data sets that are to be combined. Therefore we describe this issue more extensively and we start by explaining how relations between units in two data sets are modelled with two data sets with the same unit types and of equal size. A summary is also given in de Jong (1991). We will use the convention that the term "match" refers to the true match status of a pair of records, and the term 'link' refers to the observed status after the actual linkage process (Tuoto, 2016).

The seminal paper on estimating relations between units in two data sources is by Felligi and Sunter (1969). Their approach is as follows. Let the population of units be denoted by $i = 1, \dots, N$. Further, we have data on file X and file Y corresponding to the same population of units, both of size N. Now, we consider the situation of a 1-1 linkage, where each unit from file X is linked to a unit in file Y, but not necessarily the correct one. In addition, we have a number of identifying variables, say L, that are used to link files X and Y. We refer to those variables as linkage fields. The values for

unit i in data set X and unit k in data set Y for each of those linking fields may agree or disagree, where i and k is an arbitrary indexing of the records in files X and Y. Let the binary variable A^0_{ikl} be 1 when there is an agreement on field l for unit i in data set X with unit k in data set Y and 0 otherwise.

Further, let j denote the unknown indexing of the records in Y such that a record in X has a correct link to a record in file Y when $i = j$. We can now define a permutation matrix $\boldsymbol{P} = [\delta_{ij}]$ such that $\delta_{ij} = 1$ when unit i in data file X is linked to j in file Y, for instance

$$\begin{pmatrix} y_2 \\ y_1 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

where the vector on the right is indexed by j and on the left we have the observed linkages. If the two data sets were linked perfectly, the matrix $\boldsymbol{P}$ would have all ones on the main diagonal. Unfortunately, $\boldsymbol{P}$ is not known in practice.

Fellegi and Sunter (1969) then start estimating the probability that unit i of X and unit j of Y have an agreement on linkage field l given that they are in fact a match and the probability on an agreement given that the relation is a non-match:

- $M_{iil} = P(A_{iil} = 1 | i = j)$ the probability of agreement on a linking field, given that this is a true match, and
- $U_{ijl} = P\left(A_{ijl} = 1 \middle| i \neq j\right)$ probability of agreement on a linking field, given that this is a not true match.

A further assumption is that the probabilities M_{iil} and U_{ijl} are independent of i and j so $M_{iil} = M_l$ and $U_{ijl} = U_l$. Let $\psi = (M_1, \dots, M_l, \dots, M_L, U_1, \dots, U_l, \dots, U_L)$ be the vector of M and U probabilities. A further assumption is that the probabilities per linkage field are independent given it is a match or a non-match. Thus the total probability per record-pair (i, j) is given by M_{ii}=$\prod_l M_{iil}$ and U_{ij}=$\prod_l U_{ijl}$. This is known as the conditional independence assumption in probabilistic linkage.

Next, for the linkage process, a weight is computed for all observed pairs of records (i, k) as $W^0_{ik} = \sum_l w^0_{ikl}$ with

$$w^0_{ikl} = \begin{cases} \ln(M_l/U_l) & \text{if } A^0_{ikl} = 1 \\ \ln((1 - M_l)/(1 - U_l)) & \text{if } A^0_{ikl} = 0 \end{cases}.$$

For the 1-1 linkage process, record pairs with the largest weights are linked. In case of data sets of unequal size, one might determine a threshold above which one is very certain that the record-pair is a match. By using a large threshold value one can avoid mislinks, but simultaneously it results in missed links. To limit the number of missed links one can use a second threshold below which one is certain that the record-pair is a non-match. The linked records that are in between the two thresholds might then be investigated manually to decide whether these links are true matches or not.

Note that methods to estimate the probability of linkage have to account for the method that the data sources are linked. The latter depends for instance on the amount of overlap between the two populations. If the two data sets concern the same population, a 1:1 linkage method can be used. Another linkage situation is that the population in one data set is a subset of the other population. Information on the concepts of the population in the data sets that are to be linked and their operationalisation is needed to estimate the amount of overlap between the populations, and to estimate the occurrence of erroneous and missed linkages.

It remains to be seen whether linkage of websites and enterprises, thus the linkage of two different (but related) populations, can also be expressed using an approach that is similar to Felligi and Sunter (1969). The difference now is that the units in data set X and in data set Y refer to different unit types. The relations between those unit types are such that '1 to n' and 'n to m' relations may exist between the units in both data sets; this is determined by the differences between the concepts of the unit types (the *conceptual sets*). A further complicating factor is that there may be no agreement between unit i of X and unit j of Y on a linkage field l when it is in fact a match. Instead, the agreement may be found when linkage field l for a unit b that is a part of the composite unit i in X is compared with unit j of Y. This issue has been discussed in the context of the *operational sets*. Note that issues that occur in the conceptual sets and in the operational sets have effects on the way we can model the relations between the units in both data sets.

In some situations, we may have a sample of data for which we know whether the units are a match or not, or we may draw such a sample and determine manually whether units are a match or not. In that case, supervised classifier methods, e.g. logistic regression, naïve bayes, support vector machines and random forest, may also be used to provide estimates for the probabilities (Hastie et al., 2009).

After obtaining the probabilities, it is important to classify the obtained relations between the units of the linked data into categories such as: '1 to 0', '0 to 1', '1 to 1', '1 to n', '1 to n' and 'n to m'. These categories form input into the estimation of the relations between input units and the statistical units. The latter are subsequently used in the derivation of the measures (see for instance 4.2.3).

4.1.6 Consistency in linkage over different data sets

We are interested to develop an efficient way of linking all kinds of data sources to each other. For instance SN wishes to combine agricultural census data with VAT data. The agricultural census data concern units that are maintained by RVO.nl of which the unit types are not so clear. Probably they are a mixture of natural businesses (one-man businesses), legal units and combinations of legal units that have a common production (that comes close to the *enterprise unit*). Next, we would like to link them to fiscal VAT data. VAT units are a composite of fiscal base units.

One approach is to link the VAT data directly to the RVO.nl data. That is a complicated and time consuming approach, since we would have to make use of data

sets with relations between fiscal units and legal units, and between fiscal units and natural persons.

The idea behind the AUB (section 6.2), is that we start from a set of elementary, real, units from administrative data set that act as *base units.* Examples are natural and legal persons. We then derive the statistical units from those base units (using derivation rules). We now obtain '1 to 0', '0 to 1', '1 to 1', '1 to n', '1 to n' and 'n to m' relations between the statistical and the base units. Next, one links the VAT data sources to this AUB, and by using relations from the tax office between base units and VAT units, again relations like '1 to 0', '0 to 1', '1 to 1', '1 to n', and 'n to m' are obtained. Finally, one also links the RVO.nl population to the AUB. RVO.nl is probably a mixture of unit types, but one might start by linking this to the base units also. Finally, the relation(types) between VAT and RVO.nl population are obtained via the *base units*.

So the proposed strategy, which is in line with that of the ABR, is to link the source data sets to each other via a central population of base units. A practical point (in terms in implementation) is that if we centrally store the linkages between a large number of administrative data sets to the central population(s) of base units and their subsequent relation with statistical units, we can re-use this result for all kinds of possible linkages between data sets.

A second practical point is that the relations between units may change over time, especially business unit types can by highly dynamic. That implies that it is important to give a 'time stamp' when storing the relations that have been established. In fact two dates are important: the date where the relations actually refer to and the date when those relations have been estimated. The latter is important in because of the time delays that may occur.

4.1.7 Combining data without overlapping units

Combining variables at micro level that are in different data sets where those data sets have no units in common can be done by a method called statistical matching (D'Orazio et al., 2006). The general idea behind this method is that a unit in one data set is linked to a unit in the other data set based on a certain distance measure. The linked units are not the same: it is a synthetic linkage. For this synthetic linkage of the units, one makes use of auxiliary variables that both data sets have in common. The purpose of the statistical matching is that the multivariate distribution of the target variables in the fused set is close to the real (unknown) one. Usually, the conditional independence assumption is used, meaning that the target variables are independent given the values of the auxiliary variables, which is not a realistic assumption in many cases. More research is needed to alleviate this assumption, e.g. Kim et al. (2016), before statistical matching can be applied at SN. Also research is needed to judge the quality of a fused data set.

4.1.8 Missing and erroneous relations

Erroneously missing relations between units and errors in the recorded relations between units may be traced by using additional data sources that are not regularly

used in the production of the ABR. The example of the WSR data has been mentioned already. Another approach is the use of profilers, that manually check relations between units but that is a costly activity.

Time delays can be considered as a special case of missing and erroneous relations, so the use of additional data and profilers are helpful here also. Furthermore, one might retrospectively derive relations between units at different time moments to judge the impact of the errors on the outcomes as a kind of quality measure.

In section 4.1.5 we mentioned that it was important to classify the relationships of the *linked set*. The same holds for the relationships of the *derived set*. It is important to classify the obtained relations between the units of the data sources and the statistical units into categories such as: '1 to 0', '0 to 1', '1 to 1', '1 to n', '1 to n' and 'n to m'. Here we can make use of the relations found in the linked set (see 4.1.7).

4.2 Variables

In Table 3 an overview is given of the issues related to the variables that we need to estimate our target parameters and possible ways to address them.

Table 3. Issues concerning 'target variables' and potential solutions
(*Issues with an '*' prefix are refer to hidden issues and those without an '*' are the mentioned issues.*)

		Consult experts users	Profiling	Derivation rules	Data preparation	Linkage methods	Statistical Matching	Quality measure	Other data sources	Model based estimation	Measurement errors models
CON	Define meaningful output	x									
CON	* Aggregate concepts	x									
CON	* Disaggregate concepts	x									
OPE	Operationalisation at a less detailed level; consolidation								x	x	
OPE	Operationalisation at a more detailed level; deconsolidation								x	x	
REA	Missing values when using data of non-standard unit types or classifications								x	x	
REL	Conflicting information in overlapping variables										x
DER	Handling (selective) missingness								x	x	
DER	* Unit type for imputation							x			
EST	Quality of the linked set							x			
EST	Relating information at different hierarchical levels	x									

4.2.1 Conceptual issues

Similarly to the conceptual issues in the case of the units, the conceptual issues in case of the target variables require a careful consideration of experts and consultation of users. Defining meaningful output implies that the target parameters and the concepts that they try to measure are in line with each other. For instance, for the short-term statistics NSIs had to answer the question what kind of business dynamics to include when measuring short-term changes (TFRTQ, 2009). The issue of aggregate concepts has been solved so far by focussing on selecting the main category in the case of classification variables. Another option is simply to tabulate the proportions of a unit that apply to a category of a classification variables, so multiple categories may apply to a unit.

4.2.2 Operationalisation at a less detailed level

In case of economic data, operationalisation of values at a less detailed level is not only a matter of aggregating the available figures, but also of correcting for internal flows. This would ideally be addressed by getting the information from additional data sources. When that is not possible, model-based estimates might be used, but that requires that at least for a part of the population the necessary data are available. For instance, the effect of consolidation of economic variables might be assessed from data at the level of a combined unit (the enterprise) and data obtained from underlying units. At SN for instance turnover at legal unit level is available from VAT data and at the enterprise level from survey data. One might try to model effect of consolidation in relation to a number of background variables. A study by Burger and Buiten (2014) estimated that turnover was overestimated by at most 5% in half of the publication cells when data were not corrected for internal transactions.

4.2.3 Operationalisation at a more detailed level

When data is collected at the level of a composite unit but estimates are needed at a finer level, there are two possible options. The best option would be to try to obtain data at a more detailed level. For instance, SN hopes that the variable "work location" is added to the administrative (micro) data on wages and salaries per employee that SN receives. This would enable regional estimation of employment. A second option is that the values for the target parameter are estimated by using auxiliary variables. An example concerns job vacancies which are available by NACE code and we aim to regionalise the outcomes by province. This is done by using the number of jobs per province as an auxiliary variable (ter Steege, 2013). Another example is the use of a linear regression model to estimate turnover for VAT data at enterprise level for situation where a VAT units is linked to multiple enterprises so its value needs to be split (Enderer, 2008).

Note that in this latter example a classification of the relations between input and statistical unit types into the categories, '1 to 1', '1 to n' and others, is very helpful (see section 4.1.8). For instance relations of the type '1 to n' and ' n to m' are candidates to derive the required variable through a regression model.

4.2.4 Missing values when using data of non-standard unit types or classifications

This issue might be approached in the same way as with the operational issues, so by obtaining additional data or by using model-based estimation. In the special example of the "top sectors", websites are currently used to manually classify enterprises. One might try to use automatic classifier methods instead, e.g. naïve bayes and support vector machines (Hastie et al., 2009). There is also a group of unsupervised text classifiers called topic models (Blei, 2012) that try to capture one or more topics from texts and that can handle synonyms, which might be interesting to investigate.

4.2.5 Conflicting information in overlapping variables

The issue of conflicting information in overlapping variables from different sources can be handled by using priority rules like in micro-integration techniques. But when a relation is viewed as a classification variable, e.g. partner yes/no, it would be interesting to investigate whether latent class models can be used (Boeschoten et al., 2016; Di Zio et al., 2016) to predict the true value of a relationship based on data from multiple sources.

4.2.6 Handling (selective) missingness

If we have the values for all the variables but only for a selective part of the units in the population, one might compute adjusted weights. Maybe the methods used to model response propensities as a function of background variables can be useful, see for instance Schouten et al. (2009). Also, Boonstra et al. (2016) gives an overview of possible estimation methods (weighting, imputation) for different situations of data missingness in linked data sources.

4.2.7 Unit type for imputation

In this specific issue deciding at which unit type the data are imputed, one thing that might be done is to do both (for a test set) and to define a quality criterion that can be used to judge which option yields the most desired result.

4.2.8 Output quality of a partially linked set

As far as I am aware there is no ready methodology that can be used to judge whether a partially linked set is good enough for further analyses and statistical estimates. There are metrics to determine the amount of false positive matches and false negative non-matches, and one could take an audit sample to estimate those error rates accurately (e.g. Tuoto, 2016). There are also approaches under development to correct population estimates for linkage errors (e.g. Chipperfield and Chambers, 2015; Lahiri and Larsen, 2005). Research is needed to determine whether an approach to judge outcomes after linkage can be developed, an example of such a kind of analysis can be found in Di Consiglio and Tuoto (2013).

4.2.9 Relating information at different hierarchical levels

Relating information at different hierarchical levels is a relatively new subject for which there is hardly any experience on methods. Concerning the question of section 3.11.2 "what do decisions at enterprise group level imply for the kind of employees (jobs) that are needed in a country", one could relate the distribution of 'type of job'

(profession) to enterprise group characteristics such as which production functions are located in a country. Or maybe one could make an indicator that expresses the dependency of the type of jobs on the production functions of enterprise groups.

5. Discussion

We held an inventory of possible issues when linking data sets of different unit types and asked responsible employees (informants) to mention which issues they encountered and considered to be a problem. These issues were amended by hidden issues based on past experiences and also by structuring the issues into a framework. This way, we identified 22 issues, of which five were hidden. Maybe these hidden issues were not mentioned by the informants because they already found a practical solution to the problem or because they are so used to the commonly accepted solution (e.g. the use of a main category) that they did not mention it as a problem anymore.

Not all of the issues were directly related to the integration of data sets with different unit types. For instance, "misspellings in key variables" is an issue that might occur in any two data sets to be combined also when their unit types are the same. We believe that the following issues are related to different unit types:

- Concerning the units and their identification variables:
  - disagreement about the conceptual population
  - uncertain identity of the unit (type)
  - discrepancies between identification variables and unit type
  - probability of linkage *in case of different unit types*
  - missing relations between units (of different unit types)
- Concerning the target variables:
  - aggregate/disaggregate concepts
  - data needed at a higher/lower level
  - unit type for imputation
- Concerning the output:
  - relating information at different hierarchical levels.

Our study aimed to find some of the main issues that are encountered when combining sources with different unit types. We do not claim that our inventory resulted in a complete list of all possible problems, but we did find a number of important issues that should be taken care of. When an NSI has a new situation where sources with different unit types are combined, it is useful for them to consider the issues that are mentioned in our paper. This way they can anticipate on the problems that they may be confronted with.

6. Appendix A: case studies

In section 1 we expressed the possibility that businesses that underreport VAT in the first quarter of the year and over report in the final quarter such that the yearly value is correct. Although we did find some quarterly effects of this kind, we also found quarterly effects that we could not explain. In this section we consider three possible causes for the quarterly effects that we have found: (a) the presence of systematic patterns, (b) effects of new units in the observations and (c) effects from four-week reporters.

6.1 Micro Data Linkage project and Global Value Chains

Informant: Martin Luppes

6.1.1 Introduction

The idea about the micro data linkage project and the global value chains project is to learn more about the effect of decisions taken at enterprise group level on persons (their jobs and incomes) working at enterprises within the enterprise groups. Some information on this project can be found in (Nielsen and Tilewska, 2011; UNSD, 2014). This requires that information at enterprise group level needs to be related to enterprises, that in turn is related to job information and through job information, attributes of persons can be linked.

There are a number of challenges that need to be addressed in order to be able to fully exploit the potentials of all those relations sketched above.

- Companies are internationally oriented nowadays, not all companies fit within the national boundaries. Within SN for instance we have administrative data on international trade. About 80 per cent of those units are registered within the National Dutch business register (NHR), the other 20 per cent is not. That concerns for instance non-domestic units that use the Netherlands for transit of goods. For the domestic units, the target variables are classified by a number of background variables, e.g. economic activity and size class. For the non-domestic units we do not have this background information. In the current situation, we classify the total National figures within National accounts according to the distribution found among the domestic units. It would be better to use a kind of matching technique (such as a nearest neighbour imputation) on common variables to achieve a more accurate distribution over the background variables for the non-domestic units.

- We are used to describe information at one hierarchical level - thus enterprise group, enterprise, legal unit, job or person – rather than making statistics in which these levels are related to each other. How is information at enterprise group level related to that at the enterprise level? What do decisions at enterprise

group level imply for the kind of employers that are needed and what does this mean for the educational requirements of pupils?

- Relations between socio-economic data and functional data are complex. For instance, if companies shift part of their (production or supporting) processes to other locations, how does this affect the location where people live (how many dwellings are needed in which region?).

- Traditionally we have a number of variables in which we classify our data on business entities, such as NACE code and size class (by employer). There are however a number of other interesting variables to classify economic actors, such as “type of ownership’”, “type of trader”, “type of investor”, “enterprise with corporate social responsibility”, “is it a fast or slow growing company” and so on. The question then arises: can we find sources to extract / derive such information from, how do we combine the information in the already available sources in a correct way, and how do we link those to the statistical unit of interest?

If we link all kinds of micro data from multiple source on all kinds of administrative and survey data on businesses, all kinds of missingness will occur. Boonstra et al. (2016) presented ideas on how to fill those gaps, ranging from weighting to mass imputation.

6.2 Administrative Unit Base

Informant: Eric Smeets

Traditionally, SN maintains an ABR, which is a population frame for statistical business units (enterprise group, enterprise, local enterprise) which is derived from a number of administrative sources, of which the Dutch NHR is the central one. More recently, SN is also asked to published data which are based on somewhat different populations, for instance on natural persons that are self-employed (one man businesses), or data that relate business with person information.

Nowadays, many different types of administrative data are available, with all kinds of administrative units that are somehow related to businesses. Various, separate, IT-systems are running at SN to provide administrative ‘business’ data to internal users: for instance VAT data, income data, SZO data and international trade data. Since the number of sources will increase, as well as the number of cases where SN wishes to use those administrative sources, the question arises how can we derive statistical data from those sources that are valid for statistical units (Konen, 2015).

In this context SN has started a pilot study on the construction of a ‘unit environment’. This ‘unit environment’ (see Konen, 2015) contains:

a. administrative sources. Examples are VAT data, wage and salary data (WSR, Dutch: Polis), profit tax declarations for natural persons (IB) and profit tax data for

legal persons (Dutch: VPB declaration). One refers to the units in those administrative sources as source units (Dutch: BrE).

b. an administrative unit base (Dutch: AER) that contains the elementary units (natural and legal persons) with a unique identification that can on the one hand be related to the source units and on the other hand to the statistical units (Dutch: SE). These units are referred to as base units, to avoid confusion with the administrative units that are used in the ABR.
c. multiple statistical units bases: e.g. a statistical unit base for enterprise groups, one for enterprises, one for local enterprises for regional statistics and a statistical unit base for households as economic actors;
d. two connectors. One connector is the place where the relations are stored between (i) base units and the BrEs, and the second connector contains the relation between the base units and the statistical units. The relation between the base units and the statistical units is often through the legal unit.

The AER consisted of 17,5 million natural persons and 3 million legal persons, of which about 0.6 million are non-resident units. Important sources for the AER are the BRP (natural persons; including the non-resident natural persons) and the a data set (Dutch: BVR) from the tax office that describes all natural persons, legal persons and non-resident persons that have a fiscal relationship with the Netherlands. A further description can be found in Konen (2015).

There are a number of issues:

- In practice however, different administrative sources may have the same BrEs. Even data sets with different BrEs will have many '1 to 1' relations between different source units types. It is then important to have consistency in the linkage over different data sets to statistical units. Further, each of the administrative sources with economic data may have under- and over coverage (and linkage problems) when compared to a population of statistical units. It is efficient (and it gives insight) when issue like under- and overcoverage are described for multiple (related) data sources at the same time. This approach also gives more background and insight into the reasons for these coverage differences (Konen, 2015).
- Uses of these sources requires that somehow their BrEs can be related to the target population of a specific publication, in order to get clear which BrEs should be included in the publication which should not be. This also requires standardisation, so that in a next period the same criteria are used to select the BrEs.
- data confidentiality issues need to be taken into account, when working with all these administrative data sets.
- a base unit is sometimes related to multiple statistical units. That leads to the question how the values of connected to this base units can be split up over the different statistical units. This is an issue for which methodology is needed.
- time issues need to be regarded. In statistical publications sometimes different releases are computed in which errors in influential population units are corrected. One might be interested in the population of quarter x, according to

the release in quarter y. The question is how these issues are treated in an 'unit environment'.

- Finally, statistics that concern changes are often important. Usually, preliminary results are analysed to check whether they are valid. In this process of manual verification of the data, one usually wishes to understand the changes in the population. For the statistical units a whole system has been developed to trace back the dynamics of individual units (mergers, split ups etc.). How are the dynamics treated in the case of these base units?

6.3 Bankruptcies

Informant Jaap Jansen

SN publishes monthly outcomes on bankruptcies that are declared by the civilian court. For the output, the starting point of a bankruptcy is taken, not the end point which can sometimes be years later. The output concerns three unit types: bankruptcies of 'natural persons without a business', bankruptcies of 'natural persons with a one-man business' and bankruptcies of 'legal persons with business activities and institutions'. The monthly output concerns the number of bankruptcies.

Monthly publications on number of bankruptcies are also classified by region. Because a part of the legal units may concern multiple locations (e.g. establishments), SN has made the choice to select the location of the main establishment to compute the regional distribution. In contrast to SN, the Chambers of commerce (COC) count every establishment that is connected to a legal unit that goes bankrupt as a bankruptcy. COC publishes their findings also (www.ondernemersfacts.nl).

In order to compile this monthly output, SN daily receives administrative data from the civil court on declaration events. From these data, SN selects the relevant events. When it concerns a bankruptcy with business activities, the event has identifying information such as a LUN, name, address and a bankruptcy number (BIN). When it concern a bankruptcy of a natural person (without business activities), the event contains a citizen personal identification number (BSN), name, address and the BIN as identifying variables. We use the LUN, name and address as key variables to link the event to our ABR and use this to amend the data with background variables of the legal unit (including those of the one-man businesses), such as NACE code, legal form and address information (see later).

In addition to the monthly output, SN also compiles some incidental publications. For instance, SN has published on the financial impact of finalized bankruptcies (Elswijk et al, 2016). In this publication an overview is given of the financial effects of bankruptcies in terms of assets, liabilities and additional costs. In this publication, also qualitative aspects (classification variables) of bankruptcies are given, such as:

- way to initialise the bankruptcy (to the court);
- reasons for bankruptcies;
- share of disadvantaging: unlawful activities to creditors.

For this publication SN draw a sample finalized bankruptcies. The data obtained were from the civilian courts and contained official reports of receivers. A receiver is an official who is responsible for the correct execution of the financial aspects of a bankruptcy, such as payment of creditors.

An important point to consider with the financial impact of bankruptcies is to avoid a so-called 'double counting' of debts. This needs to receive special attention by SN in case of natural persons with business activities, such as with a General Partnership (Dutch: VOF). For instance, there may be a VOF with two business partners. When the VOF has debts, the receiver will try to pay the debts from the personal belongings of the two business partners. Those two business partners may then also go bankrupt and SN receives information on the debts of those persons. Those debts will overlap with the debts reported for the VOF. In order to avoid overlap, the financial situation need to be consolidated. The same holds for bankruptcies of a number of interconnected legal units. In part of the cases, the receiver provided consolidated figures in case of interconnected legal units.

SN also incidentally publishes how many jobs are lost due to bankruptcies, classified by job characteristics (such as duration of the job) and persons characteristics (such as age and sex). The methodology is described in Bloemendal (2010). To construct those data, SN links the starting dates of bankruptcies and the corresponding LUNs to the ABR to retrieve the related enterprises (statistical unit). Within the SSD (see section 6.12) information on the jobs per enterprise are available and those jobs can in turn be linked to persons. By linking the date of bankruptcy and enterprise identification number to administrative job information (including the start and ending data), SN can trace back how many jobs are lost in the proximity of the start of the bankruptcy till its end data. An issue here is that an enterprise is a composite unit that can may consist of multiple legal units. It is well possible that only a part of an enterprise goes bankrupt. Unfortunately, within the SSD it is not possible to distinguish between jobs lost due to a bankruptcy and jobs ended due to other reasons, which might lead to some bias in the results.

The following issues were found within bankruptcy statistic:

- the need to consolidate financial figures (see above)
- estimation of variables that are collected at enterprise level but are needed at legal unit level (see above);
- the linkage of a bankruptcy event with business activities to the ABR does not always give a full agreement on all of the three variables (LUN, name and address). The events that cannot be linked, and also those only a partial agreement are checked manually. Some reasons for partial agreement are:
  - the home address rather than the business address has been filled in;
  - the address concerns a multi-purpose building with multiple LUNs, and the wrong LUN has been filled in;
  - the LUN of another composite unit within the whole business structure or related units has been filled in.
- the changes of the number of bankruptcies is in itself not always meaningful. In one month there may be a single bankruptcy with a large economic impact, a next

period there may be hundred bankruptcies with hardly any impact. The latter may be due to a single initial bankruptcy that caused a number of other legal units also to go bankrupt, because is just concerns a large complicated administrative business structure.

6.4 Agricultural census

Informants: Marius Reitsema, Mando Weller

6.4.1 Background

The agricultural census at SN is compiled yearly[1] and gives core information on the structure of agriculture in the Netherlands. Examples of output are data on area's per type of agriculture (sweet peppers, tomatoes, full field vegetables, cows), labour force (regular/non regular workers) and type of agricultural activities (organic yes/no, proportion of non-agricultural activities). The agricultural census data are collected in a large survey which is held by RVO.nl. The collected data are processed at SN and results are published afterwards.

The agricultural census is a functional statistic, i.e. the agricultural data are detailed by type of agricultural product/service rather than by the main economic activity of a statistical enterprise. Thus if a farm has cattle and vegetables, but the vegetables concern a side activity, then the area of those vegetables are included in the total. A functional statistic is in principle aimed to describe the agricultural activities. Still, there is also a variable "income from non-agricultural activities" when a farm has also income from non-agricultural activities. When the agricultural actor is a legal person rather than a natural person, the financial variables that are obtained from fiscal data might include also non-agricultural activities, since we cannot always1 split those values (see section 2). Therefore the data are not completely functional.

Notice, however, that we are only interested in production for an external market. The area of vegetables grown by persons as a hobby, for instance, should not be included

Concerning these agricultural data, SN has done three activities that is related to combining sources.

- One activity is linking the population of RVO.nl to the population of the Dutch NHR to find the full target population. This activity is still on going.
- A second activity, since a long time already, SN enriches the census data with fiscal data that are obtained from tax authorities. As opposed to the census this concerns an administrative data source. This has been published for the outcomes of 2010 onwards. Combining these data sources may lead to difficulties since the

[1] Within the fiscal VPB data (see later), agricultural activities can be separated from non-agricultural activities. However in the years 2010 – 2012 the linkage process was not yet fully matured leading to the inclusion of non-agricultural activities in the data.

units in the agricultural census data and those in the fiscal data source do not always agree.
- A third activity concerns agricultural data (areas, turnover) on floriculture in the Netherlands. In the past this kind of information was published by a special public agency which has ceased to exist. SN has been asked to investigate whether turnover in floriculture can be estimated from administrative sources, and SN investigated the use of VAT.

All these three activities give raise to issues when combining the information on different sources. We start in section 2 with describing the linkage of the population of RVO to that of the NHR, because we can then introduce some terms that can be reused for the second activity (section 3). We then explain the third activity (section 4) and conclude in section 5 with some unsolved issues for which methodology might be useful in future.

6.4.2 Shift from the RVO population to that of the NHR

In the near future RVO and SN aim to base the agricultural census on the NHR population. The NHR is a national, central, base registration on business entities that are market-oriented. Advantage of the NHR is that the unit types are well defined: it concerns either legal persons or natural persons. Within the RVO population there are reporting units of which it is not always exactly clear what entities they in practice represent. The NHR is a central register of which all government organisations should make use.

The RVO population is restricted, in a governmental agricultural regulation, to all owners and keepers of animals and to all entities that obtain all or part of their income from agricultural activities. The units managed by RVO.nl that receive the census will be referred to as RVO units. Each RVO unit has a relation identification number, further referred to as the RVON.

Roughly, the population of RVO units consists of three groups:
- about 30.000 natural persons that are self-employed, a one-man businesses.;
- about 30.000 natural persons with a General Partnership (Dutch: VOF) or with a Professional Partnership (Dutch: maatschap). In those cases a business is run with one or more business partners or with one or more co-owners;
- about 4.000 legal persons, mainly private companies with limited liability (Dutch: BV).

See, COC (2014) for more explanation on the legal forms.

The NHR is maintained by the COC. Each natural or legal person that registers a legal form at the COC receives a LUN. Likewise each natural self-employed person that registers at the COC also receives a legal unit number (so self-employed is viewed as a legal form).

Within the RVO survey, part the self-employed persons report their citizen personal identification number (BSN). Note that before 2010, registration for the NHR of one-man businesses with agricultural activities was not obligatory. The other two groups

report their LUNs. So the first task at hand is to relate the identification numbers reported in the RVONs to those in the NHR.

Within the NHR we have a NACE code at four-digit level. Next we verify the economic activity of the RVONs according to the NHR: those RVO units that link to the NHR and that are involved in agricultural activities (two-digit NACE codes 01, 02 and 03) are considered to belong to the target population of the census.

There are a number of issues concerning the identification of all NHR units that together form the target population for the census:

- The objective of RVO survey is broader than wider that only that of the census. The census of RVO.nl concerns a "Combined Data Inquiry" Survey (Dutch: GDI) on four topics: the agricultural census, data for the manure legislation, animal health data and emission data (RVO, 2016). Not all of the RVO units belong to the target population of the agricultural census. For instance, a shepherd with a flock of sheep that grazes on the moors and that has mainly a recreational purpose should report its emission data for the GDI, but the flock is not meant for commercial production thus it should not be included in the agricultural census.
- The LUNs that are given on the RVO forms are not always the actual business entities that have agricultural activities, but they may be directly related to them. The activities of an agricultural entity may be split up into a number of legal units: for instance there may be a legal unit which is the owner (a holding), a legal unit that is involved in the actual producing activity, a legal unit that is involved in transport of the product, a legal unit involved that handles the trading (export) activities and a legal unit in which the pension of the owner is regulated. The LUN on the RVO form may well be the holding (with a NACE code that is unequal to agricultural activities), reporting the activity of an affiliate unit who is involved in the actual agricultural production.
- There are units on the survey where the identification number is missing. In such a case only a company name and an address are available. This name and address information may not directly be of the Legal unit that is performing the agricultural activities. The agricultural entity may have different buildings and addresses. Furthermore, the name of a holding may be very different from the legal unit with agricultural activities.
- The identifying variables of both the RVO units and the Legal units within the NHR are prone to errors. Part of the RVO population is checked carefully by RVO.nl because they may receive agricultural subsidies, but the part of the population for which this is not the case is of poorer quality. The NHR information is not always up to date. For instance, the value of the economic activity can be wrong.

The RVO population for the census in 2013 consisted of 67480 units. There were about 16 000 units present in the RVO population that could not be linked to the NHR or vice versa. This issue has not been solved yet. Up till now, the statistical division at SN responsible for publishing the agricultural census data uses the RVO population as the population frame.

Notice that up till now, we have mentioned the four-digit NACE code classification for agricultural activities. In the current publications, however, the classification of agricultural activities is more refined and also slightly different from that of the NACE classification. An example is that the census publishes on greenhouse cultivation, which is not distinguished within the NACE code classification. This implies that auxiliary data on type of agricultural activities are needed.

6.4.3 Issues in combining agricultural census data with fiscal data

The agricultural census data are combined with yearly fiscal data of the so called profit declaration, which concerns balance sheet data and profit and loss data. For natural persons with business activities the profit declarations of their business activities are a section of the IB declaration. For legal persons, the profit declaration is part of the Corporation Tax form (VPB declaration). Within SN, the profit data of both natural and legal persons are stored in a profit declaration data base (Dutch: WIA).

Entities that declare fiscal data have to use a Base Tax number (BTN) for that purpose. (Dutch: Fi-nummer, since 2010 renamed to “rechtspersoons en samenwerkingsverband informatie nummer: RSIN”). One-man businesses, general or professional partnerships or other legal forms that register itself at the COC receive a BTN for their (future) tax declarations. For one-man businesses the BTN is identical to their BSN.

The WIA data of the natural persons are made available at SN within a data set called SZO. Erkens et al. (2013) explains how this SZO data set is created. This SZO data base contains for each entity its BSN, its LUN, and BTN and background variables like size class and NACE code of the legal unit. In addition, for the WIA data of legal persons SN knows the relation between LUN, the BTN and size class and NACE code of the legal unit.

The statistical division now requests fiscal data that contains:

- WIA data of natural and legal persons that are identified by the BSN or LUN that corresponds to the RVONs of the (responding) RVO units;
- the NACE code of the selected units (at natural person or legal person level) should encompass agricultural activities (codes 01, 02 or 03);
- on the fiscal declarations itself the term “agricultural activities” is reported (this is a 0-1 indicator variable within the declaration).

One needs to make a further selection within the WIA data to obtain the correct information. This procedure is described in Weller (2015).

WIA data of natural persons with one-man businesses. Natural persons with a one-man business have to send in a one IB-declaration. It is important to realise that a natural person can be involved with multiple business entities: it have a one-man business, but in addition it can also be part of a General Partnership (Dutch: VOF) or a Professional Partnership. On the same IB-declaration, this natural person then has to declare the profit, loss and balance sheet data for each of its business entity (with the

corresponding LUN). So if an RVON is linked to a BSN, i.e. SN does not know the LUN, and the corresponding natural person has multiple business entities (LUNs), then the business entity with agricultural activities will be selected. Notice that for the WIA data the Dutch tax law has an additional "set of questions" for entities with agricultural activities. On this section, the LUN has to be reported. So, even when the NACE code of the business entity (which is a legal unit in the NHR, also for natural persons) according to the NHR is unequal to "01, 02 or 03", the data on the tax form itself can indicate that the activities of the business entity encompass agricultural activities.

WIA data of natural persons with a General Partnership (Dutch: VOF) or with a Professional Partnership. In case of natural persons with a General or a Professional Partnership, each business partner or co-owner has to hand in a separate IB-declaration. IB-declarations can be done in different ways, two options are: at the level of the whole "agricultural business unit"-level (statistical unit) or at level of the individual partner or co-owner. SN is interested in the data at the level of the "agricultural business unit". SN sorts all IB-declarations that concern the same LUNs and selects the first IB-declaration that has been done at "agricultural business unit"-level.

WIA data of legal persons. Legal persons that consist of "mother and daughters" are allowed to report a combined, consolidated VPB tax declaration of the combination of the mother and daughters. Those data are used by the tax authority to compute the amount of profit tax that has to be paid. Additionally, also the profit, loss and balance sheet data for of the mother itself, and of its daughters need to be reported. For each of those data, the corresponding BTNs are reported. The VPB-information from the mothers and the daughters are used by the tax authority to verify the fiscal information. Note that the different declarations on the same VPB form are numbered using a sequence number. The declaration with sequence number 01 is the consolidated declaration, sequence number 02 is that of the mother and sequence number 03 and higher are of daughters. SN receives the data from the tax office up till sequence number 10. (a VPB-declaration may also contain sequence numbers above 10). The sequence numbers of the daughters can vary from year to year, and are not used in the linkage process (we use the BTNs instead).

In summary, when combining RVO census with VPB declarations, SN links to the relevant BTN – namely the BTN of the entity that is also present in the RVO census - within the VPB-declaration.

Unfortunately, there are also RVO units for which fiscal data cannot be found. There are at least three different cases:

- the corresponding legal unit of the RVON has not yet declared its fiscal data;
- the RVON could not be linked to a legal unit;
- the RVON has ceased to exist, or it became inactive.

For RVO units of the first two cases fiscal data will be imputed. When SN has information that the RVON ceased to exist, the fiscal data will not be imputed and

the units will be removed from the target population. This is for instance the case when the unit did not respond to the RVO survey for at least three years.

Quality control of the linked results
There are at least two ways in which the results of the linked fiscal data to the agricultural census can be checked on plausibility. First of all, for each type of agricultural production, so called "standard values" (Dutch: SO-values) can be obtained from associations that concern specific agricultural activities, e.g. horticulture and cattle breeding. Those standard values concern ratios between for instance turnover and area of production. Results from the fiscal and the census data can be compared with those standard values. More specifically, the dairy cattle is a useful group, because there is little variation in the milk production per cow. Secondly, results of the linkage process can be compared over time. For instance, the number of consolidated VPB declarations was much larger in 2013 than in 2014. Such differences are a reason to check whether the linkage process has been done correctly.

6.4.4 Issues in enriching the floriculture data with VAT data

The process of linking fiscal data to the agricultural census is already in production. In contrast, the linkage of VAT data for the floriculture, is still in an exploratory phase (Reitsema, 2016). This latter process has been worked out yet in much less detail. So in the current section, we limit ourselves to a few issues concerning the linkage with VAT data.

For the floriculture data, the Legal units are linked to VAT data by making use of relations between Legal units and VAT units. A VAT unit can correspond with one Base Tax unit, but sometimes multiple Base Tax units (with sequence number) are declaring tax together in one consolidated VAT declaration. The set of units that declare VAT together may differ from those that are on the same VPB-form. Furthermore, it is good to know that the tax offices appoints an identification number to the set of units that declare VAT: the VATN. This number (the sequence of digits) is independent from the BTN. SN receives the links between the BTNs and the VATNs from the tax office.

A VAT-declaration may concern a set of Legal units, whereas only one of them has agricultural activities and belongs to the RVO population. In that case, SN needs to estimate which part of the turnover of the VAT declaration concerns that of the relevant VAT unit. To make this selection, SN uses the fact that there are three different VAT tax rating tariffs (0, 6, 21 per cent of turnover) whereas agricultural production activities fall within the 6 per cent tariff2. When there is a (n:1) link between the VAT-unit and the Legal unit, SN only selects the turnover that is related to the 6 per cent tariff. In addition, a part of the agricultural entities make use of a special VAT regulation on agriculture. Units that are approved for this regulation do

[2] international agricultural trading activities may fall under the 0 per cent tariff. But those trading activities are usually done within a separate Legal unit. A Legal Unit with economic activity *trading* does not fall within the target population of the agricultural census.

not have to pay tax for their sales but also they do not get VAT tax refunds on their purchases. Note that the approval to the VAT regulation on agriculture implies that the unit does not have to declare VAT. SN can identify those units, because we have data from the tax office that lists those units. This is not to be confused with the 0% tariff, which is used for units involved in international trade.

Another issue is that the number of linkages between the RVO units and the VAT declarations that can really be used, is still too low. This has a number of reasons:

- for some RVONs no BTN can be found;
- some RVONs do not have to declare VAT;
- some VATs have inconsistencies in their declarations. For instance they formally declare VAT once a year, but one finds multiple declarations within a year. The reason for his behaviour can be found in the client data of the tax office.

6.4.5 Need for new methodologies

We identified a number of issues within the present case study that have not been solved.

The first point is the need for a systematic approach to judge the quality of the linkage between fiscal to survey data. What would be helpful for those involved in the actual statistical production, is the possibility for them to see the relations between the BTNs, the LUNs and the BSNs. This way, more complicated situations can be checked within the statistical production.

Secondly, there is a considerable number of RVONs that cannot be linked to BTNs and there are BTNs with agricultural activities that cannot be found in the RVO population. What would be helpful is to have the correct BSNs and LUNs of all RVONs. But for the non-respondents to the RVO census, automatic procedures might be helpful, to avoid that all those non-matched units have to be looked up manually. Some of the RVO units do belong to the full population of the “Combined Data Inquiry” of RVO, but they not belong to the agricultural census population, because they are not market-oriented. This distinction should also be made.

Thirdly, we are only interested in data concerning the agricultural activities. This requires a correct identification of the reporting unit in both the RVO and the fiscal data and it requires a separation between agricultural from non-agricultural activities. For instance a large agricultural company reported an enormous turnover in its fiscal data, but the agricultural activities were not separately mentioned.

6.5 Family enterprises

Informants: Leon Custers, Jos Erkens, Olav ten Bosch

6.5.1 Introduction

SN is interested to publish figures on family businesses in the near future. There has also been a Eurostat group, that has expressed the need for data on family business and that has developed a definition (Klimek, 2015).

A project has been started at SN to identify those entities. In terms of statistical units, the unit type is the enterprise group. Note that one enterprise group can consist of one or more enterprises, one enterprise can consist of one or more legal units. For more information on statistical and legal units in the European system see for instance Struijs (2015). The question in fact is to derive a 0-1 indicator variable family business (yes/no) as an attribute to the population of enterprise group units (Dutch: OG). This population is part of the ABR at SN.

There are the following requirements according to Eurostat (cited from Klimek, 2015):

- The majority of decision-making rights is in the possession of the natural person(s) who established the firm, or who has/have acquired the share capital of the firm, or in the possession of their spouses, parents, child or children's direct heirs;
- The majority of decision-making rights are indirect or direct;
- At least one representative of the family or kin is formally involved in the governance of the firm.
- Listed companies meet the definition of family enterprise if the person who established or acquired the firm (share capital) or their families or descendants possess 25 per cent of the decision-making rights mandated by their share capital.

SN had two different strategies to identify business activities: firstly by combining already available sources within SN and secondly by web scraping of information on the internet.

6.5.2 Using already available sources

In order to identify family businesses using available sources, SN first selects all one-man OGs within the ABR since these should also be considered as a family business. It is limited to OGs with at least two employees to exclude one-man businesses. Within that group, one has a number of selections to decide whether the top-person(s) of the OG are family of each other.

Next, SN selects the top-persons of the OG of which the legal form is "general partnership" (Dutch: VOF') of "professional partnership" (Dutch: maatschap) with two or more business partners (Dutch: vennoten) - using ownership and other relational data within the ABR - within the OG. Next SN tries to identify whether the identified business partners are family or not, using three sets of data:

- the family names of business partners and of their wives are found within the COC data ;
- information on which persons share the same household are obtained from the household data set of SN.

– parent-child and partner relations are obtained from the parent-child data set of SN. This is used as additional information to the household data, for instance to find ex-partners.

The parent-child and the household data do not always give the same answers. What SN has done so far is that if one source indicates that a business is a family business and the other source does not or the other source lacks information then the business is classified as a family business.

For the set of business partners that are identified to be family in the previous step, one then checks in the available source whether the family has the majority of the decision-making rights. For instance, the fiscal data one wages and salaries contains a variable “kind of income”, and one of the categories (Dutch: inkomenssoort 17) indicates that the person is the head of a business.
Finally, SN also identifies whether an OG is a family business in case of OGs with the legal forms *private company with limited liability* (Dutch: BV), of *public limited company* (Dutch: NV) . A whole set of methods are used, but to save space these are not described in this paper. The reader is referred to Custers and Vrolijk (2016).

A practical issue within SN is that the citizen personal identification number (BSN) has been encrypted for data confidentiality reasons, leading to a new unique SN internal PIN (Dutch: verrind nummer). Since the enterprise data contains the BSN, but not the PIN, additional steps need to be taken to link the data.

6.5.3 Using web scraped information

SN has also performed an explorative analyses to investigate to what extent it is possible to identify family businesses from the content of business web sites. Note that in practice so far, the web scraping activity, has not been used as a stand-alone method to identify family businesses. But when businesses have been identified as a family business in the procedure of section 6.5.2 and they are also identified as a family businesses by using web scraping, then we might be very sure that we appointed then correctly as a family business.

Roughly one could use three different search strategies to trace family businesses on web sites (Bosch et al., 2016):

1. Using a generic search engine, such as Google;
2. By scraping web sites of enterprises, go to the “contact page” or to the “about page” and search for the term family business (Dutch familiebedrijf)
3. Use an “aggregator site”, a site that has collected information on businesses.

In a first pilot a combination of strategy 1 and 3 was used: one searched in google in a site with phone numbers of businesses. For this site businesses deliver a short text that apparently also contains information on family businesses. On this site one used the search term “site:http://www.detelefoongids.nl familiebedrijf <gemeentenaam>”, where in <gemeentenaam> sequentially the names of all 390 Dutch municipalities were imputed. This yielded a list of 31.107 records with businesses that might potentially be a family business. From these businesses, the first page of the business

was visited (automatically) and following attributes concerning the business were retrieved: short text, name of the business, telephone number, url, image, opening hours, municipality, postal code, house number and email address.

Bosch et al. (2016) describe a number of potential amendments to the search activities: one could use strategy (2), one could use other search engine, and SN could ask for direct access to the website "www.detelefoongids.nl".

6.5.4 Issues

After direct linkage of multiple sources using identification numbers, the sources are integrated at micro level. Results at that level are sometimes incomplete and contradictory and quite a number of steps are needed. That means that micro integration techniques are needed to derive the 0-1 indicator variable. It would be interesting if we could quantify the quality of the outcome.

Combining the information of the already available sources with that of the web scraping activities would also be interesting. The web scraping data can be linked to the other sources, via the ABR on a large number of key variables of which the telephone number and url are unique identifiers, For those enterprises for which we do not know the telephone number and url, we could use the other keys trough in a probabilistic linkage method. Maybe in future also the legal unit number (LUN) and BTN can be collected, and used for linkage. After linkage, some kind of micro-integration method would be needed.

There is an interesting issue when combining website information to enterprise group data in the ABR, because these may not concern different unit types. The notion “familiebedrijf” may refer to a legal unit whereas our classification variable concerns the enterprise group. There is no methodology available yet how to tackle this issue. Let alone, that the term “familiebedrijf” as used on internet may differ from the formal Eurostat definition.

6.6 Self-employed

Informant: Jos Erkens

6.6.1 Derivation of the self-employed population

Natural persons can have some business activities. Within SN we are interested to construct a population frame with statistical data on all natural persons with business activities. This set of data is referred to as the SZO. Erkens et al. (2013) describes the process flow of the SZO. In fact we have two population frames: one for the self-employed persons that are one-man businesses or partnerships/co-owners and the other frame for director / main shareholders.

Note that ‘Business activities’ within the SZO is interpreted in a broad sense. In the SZO data, one includes self-employed entrepreneurs that need to register at the NHR

(and receive a LUN). One also includes natural persons that report certain business-like activities at their income tax declaration, but are not registered at the NHR.

The SZO data has multiple clients within SN. The exact definitions of the target populations of those clients might differ from each other. Examples are the population of legal units of natural persons (classified by legal unit attributes such as NACE code), the set of business activities of natural persons and the population of certain persons (Dutch DGA's) with legal units (classified by person attributes). Each of them requires their own delineation.

The SZO data set has been developed gradually. A number of administrative data sets are combined, e.g. profit data of the IB-declarations, VAT data, data on the tax data of houses (Dutch: WOZ). All the data are combined using deterministic linkage on identity numbers. A whole set of auxiliary variables are derived, that can be used by the internal users of the SZO data to get the data of their own target population.

The SZO data is dynamic, in the sense that new data sources can be added, and sometimes old data sources cease to exist. The SZO data are used for quarterly statistical production, starting from 2013.

6.6.2 Issues

The linkage of multiple sources is in fact a micro-integration process. For a number of variables, multiple sources are available. For instance in the SZO one estimates a 0-1 indicator variable on the presence of business activities, that is derived from variables that are present in the sources. The question arises how we can come as close as possible to the true value, and whether we can estimate the reliability of that classification variable.

6.7 Centre of policy studies

Informant: Lotte van Oostrom

Below, five cases are given of examples where the combination of sources is more complex than usual.

6.7.1 Case 1: Relating website information to enterprises

The centre of policy studies at SN (CPS) was involved in a project that estimated the extent that businesses are involved with internet economy (Oostrom et al., 2016). CBS used a data set from DataProvider that contained website information of businesses. Using the website content, enterprises were classified a number of categories of internet intensity. There were a number of issues that need to be treated in this project:

– linking the website to the ABR. In the project websites were linked to the ABR using LUN, host name, email address, phone number and address. CBS used a classification of five categories to express how certain we were about the

correctness of the linkage (very certain to very uncertain) depending on how many keys were identical and which keys are identical.

- relating the website information to the enterprise level. A first problem is that there were multiple websites that match the same LUN. A second problem is that a website may link to one enterprise that is part of an enterprise group with multiple enterprises. The question then arises whether the website content refers to the enterprise group or whether it is specific to the enterprise. Likewise, the website may link to a LUN and the enterprise my consist of multiple LUNs. CPS used decision rules to cope with this issue.

6.7.2 Case 2. Statistics using a population list from a third party

CBS received requests from trade organisations to compile data on their performance. Mostly this concerns information on wages and types of contracts and so on. The trade organisation often gives a list with LUNs. The wages and types of contract information however is available at enterprise level. In many cases one legal unit coincides with one enterprise, but the larger enterprises are related to more than one legal unit. Decision rules are used to determine if the wages and contract information that is linked to the member of the trade organisation is “close enough” to be used (for more information: see section 6.9).

The organisation Rvo.nl provides list with agricultural units for which rvo.nl has appointed subsidies. Rvo.nl is interested to have more information on the “performance of those units, in terms of innovation and investments. To that end, CBS makes uses of our publication on top sectors. An additional question is often how many enterprises that do belong to a top sector, make use of a subsidy? Data on top sectors, and on innovation and investments is at the level of the enterprise, whereas the rvo.nl units are often natural of legal persons. Like with case 2, decision rules have been made how to deal with this unit problem.

6.7.3 Case 3. Linking data from third parties with CBS data

From DUO institute (on education) CPS received for a study to estimate the time elapsed between the arrival of refugee children and the moment they go to school. CPS received a list with name, birth date, location of living (Dutch: AZC) and this need to be linked to administrative data on education. The outcome are compared with results of a survey on this topic, to investigate whether the results are consistent. In this case probabilistic linkage is needed.

Another example is that CPS was involved with a study on partnership and payment of children alimentation. Therefore we received data from a third party. In terms of identifying variables it contained variables name, gender, and birth date, and part of the data contained a personal identification number. An issue that arose here was whether the results obtained after linkage (about 95% was linked) are accurate enough.

6.7.4 Case 4. Combining two surveys for small subpopulations

The CPC received a request on combining information in the Labour Force Survey and the survey on Social Cohesion. A question might be: does the intensity of labour

participation of persons affect the level of social interaction (contact with friends, family, neighbours) of persons? In addition they were interested in specific population subgroups. In this case the number of overlapping units in both surveys is very small. This is an example where statistical matching is needed.

6.7.5 Case 5. Available information not detailed enough

CPS was involved with a study on where do people study versus where do they live? A problem are student that study at educational institutes with have multiple locations and the data only mention the main office of the institute and not the exact location.

6.8 Top sectors

Informants: Jamie Graham, Remco Kaashoek, Nino Mushkudiani

6.8.1 Background

Purpose of statistics on top sectors is to give an overview on the achievements of so called economic top sectors, relative to the achievements of all businesses in the Netherlands. This publication started in 2012, and is issued yearly with the Ministry of Economic Affairs as the contractor, see for instance Monitor TopSectoren (2015). The Monitor consists of nine top sectors, among which there are water, energy and logistics. For each of the top sectors a set of about 40 performance indicators are selected.

The central unit type to represent “businesses” is taken to be the enterprise. The reason is that the data on most of the 50 indicators are available at the level of the enterprise, with some exceptions such as the export data. A number of issues that occur when assembling the Monitor Top Sectors is described in section 2.

6.8.2 Issues in the production of the Monitor

Select target population

A key step is to appoint the set of enterprises that belong to each of the nine top sectors, further referred to as the target population of each sector. Three sources are used as input for the construction of the target set per sector. The first source, concerns enterprises within the ABR that belong to certain NACE codes. Most top sectors however, the activities cannot only be found through the NACE code.

As a second source, most top sectors also uses list of units supplied by a trade organisation (NL: “ledenlijst van een brancheverenigingen”). Those lists often contain COC identification numbers - the LUN - and the corresponding business name. SN then tries to link the LUN to the ABR, and if that does not work, the corresponding names. Assume that we find a positive link. Then the next step is to make sure that the unit has economic activities that belong to the top sector. To that end the NACE code of the legal unit (within the ABR) is consulted. If the result is unclear, then the unit is seeked on internet and the activities of the unit are compared manually with

the set of definitions the activities of the sector. Also when the unit cannot be linked to the ABR, its activities are manually looked up on the internet.

A third source are the self-employed workers.

Unit problem
Note that not all of units on trade organisation lists link 1: 1 to an enterprise in the ABR. That is in principle not a problem, as long as the economic activities of the unit that is found on the list belongs to the top sector. Then (data of) the corresponding enterprise are attributed to the respective top sector. In case of more complex enterprises that consist of multiple legal units, the LUN on a trade organisation list often concerns the holding.

SN has developed a flow diagram that describes how to handle in case of linkages that between units on the list and the enterprises within the ABR differ from 1:1. For instance it may happen that the unit on the list is an enterprise group.

Issue of different periods
The list of units supplied by a trade organisation are often of a more recent date than the publication period, where the publication period is the period for which the statistical outcomes are to be described. An implication might be that some units on the list are "born" after the publication data while other units may be missing.

Population dynamics
Large shifts in the outcomes per economic sector may occur when influential units change from NACE code. Such shifts may also occur due to changes in the members of trade organisations. SN always try to give an explanation to the users of the data when such changes occur. In addition SN will try to recompute the historical outcomes according to the new population structure, thus the outcomes as if those units were already present in the past. In this way, the changes will become better comparable over time.

Estimation issues
The data that are used for the performance indicators are obtained from different publications, e.g. the structural business statistics (SBS), national accounts (NA), and international trade statistics. The SBS outcomes are based on a sample survey and are re-weighted to the NA-outcomes, at level of a NA-grouping of economic sectors (NL: "regkols"). Outcomes on export are related to the international trade outcomes as these are considered to be more reliable.

Recall that the estimates per sector are partly based on NACE code and partly on membership lists from trade associations. For the estimates based on NACE code, the SBS-weights as computed within the original SBS statistic are used. For the trade organisation membership lists, the corresponding enterprises are linked to the SBS survey and those units are used with a weight of unity. Part of the units that are linked to the SBS, are non-respondents. The values for those units are based on imputations. There are also units on the membership list that cannot be found in the

SBS survey. So the question remains how to obtain good estimates for the units that are on the trade organisation lists.

6.8.3 Where is improvement needed?

In two key parts in the Monitor Top Sectors methodological improvements would be useful:

- methods for automatic linkage of the units in the trade organisations to the ABR ,
- automatically using information from internet (text mining) to determine whether a units core activity is within a certain top sector or not.

A final point is how to obtain good estimates for the units that are found on the membership lists of trade organisations.

6.9 Trade organisations

Informants: Arthur Giesberts, Niek Verbaan, Stephan Verschuren

Since a few years, tailor made publications for trade organisations are compiled at SN by the statistical department and/or the Centre of Policy Statistics (CPS). The overall compilation process is as follows. The trade organisation provides a membership list. The entities on the list contains names and legal unit numbers. We assume that those entities on the list coincide with the legal units for which their identification numbers are given.

The entities on the list are linked to enterprises using the most recent version of the ABR. These enterprises can be used to select the relevant economic data (see below). In addition, the enterprises are linked to the September data of the system of social statistical data bases (SSD, see section 6.12). These September SSD-data contain information on employee characteristics (see below).

The publications consist of two different kinds of target parameters. The first kind, concerns economic information: yearly turnover total of the membership population and estimation of the year-on-year quarterly turnover changes of the population. Sometimes also level and growth estimates on international trade data (import, export) are provided. The second kind concerns information on the composition of the employees of the members of the trade organisation, such as the fraction male/female, age categories etc. The second kind of data concerns figures that are based on enterprise data and concerns two subpopulations: (a) the set of legal unit-members that have a one-to-one link with an enterprise and (b) the set of all legal unit-members, including those that does not a one-to-one link with an enterprise. In nearly all of the cases, the latter group will have a many-to-one link to enterprises. An example can be found in CBS (2016).

We distinguished the following issues:

- The membership lists are not always of good quality. First of all, the information on the list may be outdated. If the linkage frequency between the membership

list and the enterprises becomes too small, SN may ask to update the list. Furthermore, also spelling errors may occur in the names, or errors may occur in the digits of the LUN.

- Sometimes the economic activity of the entities on the membership lists provided by trade organisations differs for the economic activity code we have at SN or it is more detailed. We need to link the entities of the list to the ABR to compare the economic activity code of the trade organisation with that in our ABR. So far, we keep this finer economic activity classification fixed over time: we use the classes of one membership list, and do not (try to) track effects of changes in economic activity on the outcomes.

- Some members are active on multiple economic activities. In that case, SN selects the most important one (of that member). This is done by simple derivation rules. For instance if more than half of the employers belong to one economic activity, that activity is selected to be the main one. Alternatively if a reasonable estimate for turnover per economic activity can be obtained, this can be used to find the main economic activity.

- Turnover growth is available at enterprise level, but it is estimated at legal unit-level for some publications. For those cases where multiple LUs link to one enterprise, the number of employees per legal unit is used to estimate the turnover at legal unit-level. The problem is that the quality of the employee information per legal unit becomes poorer. The reasons is that businesses (more) often register all their employees in one location and more businesses make use of pay rolling or self-employed staff.

- The outcomes on employer characteristics and those on turnover changes cannot directly be compared with each other, because the first concerns information at enterprise level and the latter at legal unit level. Therefore the employer characteristics data is presented as fractions of the total population, not in terms of absolute numbers. Fractions are also given, because membership lists are not always up-to-date.

- Some of the trade organisations provide their information at establishment level (Dutch: vestigingen). These trade organisations are also interested in regional information on establishment level. At SN we currently have limited information at regional level, we currently do not have reliable data for all enterprises on employees per establishment. The latter information would be very helpful as a background variable to make estimates of economic variables and directly to make relative frequency tables of employer characteristics per region.

- Five year figures are constructed. Depending of the agreement with the trade organisation, often a year-by-year panel method. In case of using a panel method, for each set of two consecutive years a panel of units is used in which births and deaths in the population is not accounted for. Next, the outcomes of the growth

rates are multiplied by each other. An idea is to make two figures: a panel figure and a figure that does account for population dynamics.

– There has also been a request for a publication on IT data centres. In that specific case, the list of units that should belong to the population was unclear and SN decided not to accept the request.

6.10 Energy consumption

Informant: Remko Holtkamp

6.10.1 Background

Data on consumption of gas and electricity to the level of dwellings and businesses is published yearly since 2010. The data on businesses concerns total consumption of gas and electricity stratified by 1-digit NACE and region (municipality, province, and COROP). In fact the data on businesses concern establishments (Dutch: vestigingen), since the data are linked at address-level (see later). Data on dwellings concern the average yearly consumption stratified by type of dwelling (apartment, terraced, end-terraced, semi-detached, detached), and region (municipality and province). Recently, a new publication is made in which the average yearly consumption is split classified by building type (of businesses services) is published. Building type is split into 24 categories (e.g. primary and secondary school, retail trade with and without cooling facilities, swimming pool etc.).

All data are obtained from administrative sources. The central data concerns administrative client energy data sets (CAD) (businesses and households), obtained from about 11-12 energy distribution entities (Dutch: energie netwerk beheerders). Within the CBS law it is regulated that those energy distribution entities have to deliver their CAD data to SN. The set of CAD data concerns the complete volume of energy delivery in the Netherlands. For the gas delivery, the unit within those data sets is a “gas connection point”. A gas connection point is the location where the gas meter is present. This gas meter is identified by a unique energy connection point number (Dutch: EAN). Likewise, for the electricity delivery, the unit of the data set is an “electricity connection point”, which corresponds to the location where the electricity meter is present. The electricity meter is also identified by an EAN. The EANs in turn are attributed to an address (postal code, house number and house number suffix) and by the name of the corresponding client. One address can have multiple EANs.

All three publications on energy use are compiled by combining the CAD with four sources:

– the basic registration on addresses and buildings (Dutch: BAG), which is maintained by the semi-public organisation Kadaster. The BAG Contains address (postal code, house number and house number suffix), area, use of building (dwelling, school, health care, …), and an indicator variable whether the building is empty or not;

- the ABR of SN. The ABR contains (among others) information on establishments , their addresses and economic sector.;
- a commercial data set from Locatus, that concerns information on shops: addresses, name of the establishment, whether the shop is still in use or not, and a kind of economic activity classification;
- data supplied by DataLand on area and use of a building (dwelling/business).

In addition the publication on energy consumption by type of dwelling, uses of a separate delivery from Kadaster, other than the BAG, which contains dwelling address and type of dwelling.

6.10.2 Issues in combining the sources

In a first linking round, the CAD data at EAN level are linked to the BAG, ABR, DataLand and Locatus data, using the linkage key “postal code (6 characters) and house number and house number suffix”. Next, the first aim for each EAN is to identify whether it can be classified either as a business or as a dwelling. Note that the concept “business” also includes public functions like education, health care and public administration (Dutch: openbaar bestuur). EANs with a business function are processed separately from those with a dwelling function.

With the current description we do not aim to describe the actual process, we sum up which issues are encountered that are related to the linkage of the sources:

- *Formatting issues*. The addresses in the CAD data may differ from those in the base registers (BAG, ABR). Initial linking of addresses in the CAD data to those in the BAG and in the ABR led to a matching rate of only about 80%, for a number of reasons. The first reason is that part of the EAN addresses cannot be found in the BAG or in the ABR, or they are wrong (e.g. non-existent postal codes). The second reason is that the address notation of the house numbers and their suffixes are not standardized. Holtkamp et al. (2016) explains that “Common examples are A, B, C, 1, 2, 3, I, II, III, 1st fl, 2nd fl, 3rd fl., 015, 025, 035, but in the client registers unusual cases also occur like ”next”, ”near”, ”across”, ”2 – 3”, but also ”garage”, or ”traction”, and even worse.” Holtkamp et al. (2016) uses standardisation steps to harmonize these format issues. After standardisation, the matching rate increased to over 96%.

- *Inconsistent values*. Some variables (the building is empty yes/no, use of the building, area of the building) are present in multiple sources and their values are not always identical. In those cases preference rules are used, where sources are ordered according to their reliability.

- *Identifying businesses activities in dwellings.* A small fraction of the dwelling addresses have businesses activities. When there are business activities in a dwelling, the energy use of those EANs is completely allocated to the business, in stead of to the dwelling (Holtkamp et al., 2016). These business activities are identified because for some of the dwelling addresses their energy use was found to be very high, which indicates business activities. The top 1% energy use for dwelling addresses is checked thoroughly on the presence of business activities.

This is done after the initial separation of addresses into a business or a dwelling destination. Three steps are used. In a first step the EAN client name is compared with the list of business names in the ABR. Secondly, addresses are linked to a a register that is used by municipalities for taxation of dwellings, and it indicates whether a dwelling is situated at the address or not. Thirdly, the data are linked to the Locatus data set.

- *Plausibility checks.* Randomly, 1% of the EANs are checked manually on correctness of the classification of energy use by type of business and type of dwelling. This is for instance done, by verifying the derivation of the appointed class and by seeking the address on internet. In addition, aggregated values of outcomes are compared with data from national accounts.

- *Block heating*. Sometimes very small or zero gas use values are found for a group of dwellings. In those cases may be that group of dwellings share a block heating point (Holtkamp et al., 2016). The statistical division tracks down those block heatings by using information from the BAG. Within the BAG there is a variable "BAG-address" and a variable "BAG-building" (Dutch: BAG vlakken). A BAG building may encompass a group of BAG addresses. If, within the set of BAG-addresses that belong to the same "BAG building" there is one EAN with a very large gas consumption and all others have a very low gas consumption, one assumes that this is due to block heating. Finally, the gas use at the block heating point is attributed uniformly over all dwellings in the corresponding block.

- *EAN – address – establishment linkages other than 1:1:1*. Many EAN - address – establishment linkages are 1:1:1 i.e. one EAN links to one address and that address links to one establishment. In addition there are also two other situations. First there is the situation that one EAN links to one address and at this address there is an office block with multiple establishments. In that case, one first selects the NACE code of the establishment in that office block with the largest number of people employed. Then the total energy use of this EAN is attributed to the NACE code of the largest establishment within the office block. A second situation is that of multiple EANs share the same address. In that case, the figures of those EANs are added together. If there is one establishment located at that address, the resulting total gas / electricity use is attributed to that establishment. If it concerns an office block, then again, the total energy consumption is attributed to the largest establishment.

- *Finding greenhouses*. There has been a special effort to find the energy use of greenhouses. In case of greenhouses, the energy connection point may be located in a dwelling from where the energy is transported to the greenhouse. GIS information was used to link the greenhouses to the EANs. Once these links are traced, those links are maintained until one has information that the business has ceased to exist. Once every so many years the GIS procedure will be repeated.

- *Buildings with multiple purposes.* The energy use of establishments with a service is not only classified by NACE code, but also by a separate service classification

(e.g. primary and secondary school, retail trade with and without cooling facilities, swimming pool). Some of the buildings have multiple uses, but there is only one energy connection point. In that case the whole of the energy use is allocated its main activity.

6.10.3 Need for new methodologies

We can identify a number of issues within the present case study that have not yet been solved:

First of all, all calculations are done automatically, and the scripting can be improved in efficiency. Secondly, the step of dealing with harmonisation of the addresses can be improved. Parsing – splitting the text into smaller parts – , using Ngrams – N-sequential characters – as well as string distance metrics (such as Jaro-Winkler) are available methods for harmonisation (Herzog et al., 2007) that can be used to improve the matching rate.

Thirdly, the linkage results are now evaluated by checking the 30 largest values per 1-digit NACE code in combination with a 1% random sample. Maybe a more effective sampling design is possible. An idea might be to compute a probability for a correct linkage and use that as a kind of score function. Furthermore, addresses are checked manually on the internet. Maybe part of this information can automatically be collected from internet.

6.11 Usual residence

Informant: Jan van der Laan

6.11.1 Introduction

According to Gerritse et al (2016) a person is considered et be a usual resident when he or she has lived or intend to live for a continuous period of at least 12 months in their place of usual residence. A large part of the usual residence population is registered in the Dutch population register frame. The difficult part are those people that are non-registered. This are non-registered part of the usual residence is estimated from capture-recapture (CRC) method.
In order to use this CRC method three registers are linked to each other (Gerritse et al, 2016): the Dutch Population Register (PR), a register with wages and social benefits of employees of Dutch business (WSR), and police register that contains crime suspects (CSR).
In a first pass, Population Register and WSR data are linked deterministically on using a PIN. In a second pass, records are linked by probabilistic linkage using linkage variables such as birth date, postal code and house number, street name, place of residence, country of residence. Next round PR-WSR is linked to the CSR data. Again in a first pass, records are linked by a PIN. In a second pass data are linked probabilistically.

6.11.2 Issues

There are a number of issues encountered in the probabilistic linkage step:

1. We have difficulties in estimating the so called M- and U-probabilities correctly (M-probability: probability of agreement on a linkage key given a match; U-probability: probability of agreement on a linkage variable given it is not a match). This is due to a number of reasons:
   - the use of blocking variables, then the probabilities are no longer correct;
   - the use of strong assumptions (conditional independence of agreement on different variables given it is a match or given it is not a match). We know that errors in street name and postal code for instance are correlated;
   - we force 1:1 linkages, how does that affect those probabilities?
2. We would like to correct the CRC estimates for linkage errors. To that end we make use of two parameters α (type 2 error: probability on erroneously missed links) and β (type 1 error: probability on erroneous links). However, this is not an easy task, due to a number of issues:
   - how does one estimate these parameters? One solution is to use the M- and U-probabilities, but this has the problems mentioned above. Another, is to use clerical review, which in this case is not straightforward as no additional information is available next to the linkage keys.
   - the theory is developed for two registers, there is now a first extension to three registers.
   - corrections are sensitive to estimates of α and β;
   - corrections do not work with covariates. The corrections are a modification of the traditional Petersen estimator for capture-recapture. The only way to incorporate covariates is to apply the estimator within strata which puts a restraint on the number of covariates used. An alternative to the Petersen estimator is to use log-linear models in which it is easy to incorporate covariates. How to modify the estimator using log-linear model to incorporate the uncertainty in the linkage is unclear.

6.12 System of Social Statistical DataBases

Informant: Harrie van Dommelen

In most of the situations where data need to be linked in the context of the SSD a unique identifier can be used. Examples are the PIN, and an educational number. A good alternative is linkage on postal code (6 characters), house number, sex and date of birth. There are a few issues related to linkage of personal data:

Sometimes data sets have to be linked where one of the sets only contains a list of names. An example is the request concerning the education of refugees under 18 in the Netherlands (see the Case Study for the CPS). In that case you have to deal with spelling mistakes and text metric methods, such as trigrams, are needed.

Another issue concerns linkage quality. What is the effect of incomplete linkage and linkage errors on the accuracy of the statistical estimates based on those linked sources?

A final issue arises when information on relations between units is only available from a certain starting date onwards (or similarly when this information is not up to date). For instance, the BAG that contains building objects and their addresses, is available from 2011 onwards. When we need to link sources from before 2011 that are related to building objects, for instance when we are interested to know in which type of dwellings people live, we take the list of persons and their addresses of the target year (say 2010) as a starting point. Next we try to link the building objects to those addresses with the risk for erroneous links.

In the linkage of enterprises with persons there are currently some practical issues within SN because the sources are divided over the internal unit of economic statistics and that of social statistics. One of the issues concern data confidentiality: personal information needs to be anonymised and that takes time.

7. Appendix B: List of abbreviations

Table 4 provides a list of abbreviations used throughout this paper. A tick mark in column 'Dutch' means that the common abbreviation from within Statistics Netherlands is used.

Table 4. Different case studies for combining data sources with different unit types

Abbreviation	Dutch	Description
ABR	v	General business register of Statistics Netherlands
BAG	v	Basic registration on addresses and buildings
BIN		Bankruptcy identification number
BRP	v	Basic registration on Persons (including non-resident units: that are found in Dutch registrations but not living in the Netherlands)
BSN	v	Citizen personal identification number from the government
BTN		Base Tax number
BVR	v	'Management of Relations' data set from the tax office containing all natural, legal and non-resident persons that have a fiscal relationship with the Netherlands.
CAD		Client energy data
COC		Chambers of commerce
CPS		Centre of policy studies at Statistics Netherlands
CRC		Capture-recapture method
EAN	v	Unique energy connection point number
IB	v	Profit tax data of natural persons
ITS		International trade statistics
LUN		Legal unit number (Dutch: KVK-nr)
NACE		Classification of economic activity
NHR	v	National Dutch business register
NSI		National statistical institute
OG	v	Enterprise group unit type
PIN		Personal identification number which is anonymised to ensure data confidentiality (Dutch: verrind)
RVO.nl	v	Netherlands Enterprise Agency
RVON		Relation identification number of a unit of the RVO.nl population
SBS		Structural business statistics
SN		Statistics Netherlands
SSD		System of social statistical data sets
SZO	v	Self-employed one-man businesses

Table 4 (continued). Different case studies for combining data sources with different unit types

Abbreviation	Dutch	Description
VAT		Value added tax
VATN		Identification number of a unit reporting value added tax
VOF	v	General partnership
VPB	v	Profit tax data for legal persons
BrE	v	Units in the (administrative) source data
WIA	v	Profit declaration data base at Statistics Netherlands
WSR		Wages and social benefits data of employees of Dutch businesses

8. References

Bakker, B. (2011), Micro-integration: State of the art, in: ESSnet on Data Integration, Report on WP 1 State of the art on statistical methodologies for data integration, 77–107.

Bakker, B.F.M., Rooijen, J. van and L. van Toor (2014). The system of social statistical dataset of Statistics Netherlands: an integral approach to the production of register-based social statistics. Statistical Journal of the IAOS 30: 411–424.

Blei, D.M. (2012). Probabilistic topic models. Communications of the acm 55: 77–84.

Bloemendal, C. (2010). Wie verliezen hun baan bij faillissementen? CBS publication (in Dutch). Available at https://www.cbs.nl/NR/rdonlyres/8FAF5EEB-0BAC-4E8F-93A9-75C7EE89C8E0/0/2010k2v4p49art.pdf @ nakijken

Boeschoten, L., Obserski, D.L. and T. de Waal (2016). Estimating Classification Error under Edit Restrictions in Combined Survey-Register Data. CBS Discussion Paper 2016-12.

Boonstra, H.J., Rozendaal, L., Voncken R. and S. Käuser (2016). Estimation methods for linked data sources: a review for the Micro Data Linking project. SN paper.

Bosch, O. ten, Windmeijer, D. and M. Q. Le (2016). Verslag internetrobots familiebedrijven. CBS memo. (in Dutch).

Burger, J.M.S. and G. Buiten (2014). Van opvragen naar ophalen van bedrijfsinformatie: effect van niet consolideren. CBS nota (in Dutch).

CBS (2016). Maatwerk voor brancheverenigingen. ENVAQUA. Dutch Environmental and Water Technology Association. Retrieved from https://www.cbs.nl/nl-nl/achtergrond/2016/06/maatwerk-voor-brancheverenigingen-envaqua.(in Dutch)

Chipperfield, J.O. and R. Chambers (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. Journal of Official statistics 31, 397–414.

COC (2014). Starting your own business, as a self-employed entrepreneur. Leaflet from the Chambers of Commerce in the Netherlands, dec. 2014. Accessed at https://www.kvk.nl/english/starting-a-business/. 11-08-2016.

Constanse, F.C. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology 40, 137–161.

Costanzo, L. (ed.) (2011). Main findings of the information collection on the use of admin data for business statistics in EU and EFTA countries, Deliverable 1.1 of the ESSnet Admin Data, 2011.

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

Custers, L., and M. Vrolijk (2016). Identifying family businesses in the Netherlands. CBS Report.

Delden, A. van and K.J.H. Bemmel (2011). Handling incompleteness after linkage to a population frame: incoherence in unit types, variables and periods. Chapter 4 of the report on Work Package 2 of the ESSnet on Data Integration.

Di Consiglio, L. and T. Tuoto (2013). Challenges in estimation of probabilistically linked data. Proceedings of the conference on new techniques and technologies for statistics, 5-7 March 2013.

Di Zio, M., Guarnera, U. and R. Varriale (2016). Selective editing of multisource data based on latent class models. Proceedings of the ICES V conference 21-24 juni 2016, Geneva.

D'Orazio M., Di Zio, M. and M. Scanu (2006) Statistical matching: Theory and Practice. Wiley, Chichester.

Enderer, J. (2008). Is the utilization of administrative data in short term statistics an ideal standard in the conflicting priorities of user demands, response burden and budget restrictions? Proceedings of the IAOS Conference ´Reshaping Official Statistics', 14–16 October 2008, Shanghai.

Erkens, H.M.J., Voort, M.J. van der, Smeets, H.H. and J.O.P.M. Muller (2013). Procesmodel Satelliet Zelfstandige Ondernemers (SZO). CBS Nota versie 1.2. (in Dutch)

European Commission (EC) (2015). eGovernment in the Netherlands. Version January 2015, Available from https://joinup.ec.europa.eu/elibrary/factsheet/egovernment-netherlands-january-2015-v170, retrieved at 2017-02-20.

Felligi, I. and A. Sunter (1969). A theory of record linkage. Journal of the American Statistical Association 64 (328), 1183–1210.

Gerritse, S.C., Bakker, B.F.M. van der, Wolf, P.P. de and P.G.M. van der Heijden (2016). Undercoverage of the population register in the Netherlands, 2010. CBS publication.

Gu, L., Baxter, R. ,Vickers, D. and C. Rainsford (2003) Records linkage: current practice and future directions, Technical Report 03/83, CSIRO, Mathematical and Information sciences. Available at http://datamining.csiro.au.

Harron, K., Goldstein, H. and C. Dibben (eds.) (2016). Methodological developments in Data Linkage. Wiley series in probability and statistics, Chichester.

Hastie, T., Tibshirani, R. and J. Friedman (2009). The elements of statistical learning, 2nd edition, Springer, New York.

Heerschap, N. and L. Willeborg (2006). Towards an Integrated Statistical System at Statistics Netherlands. International Statistical Review, Vol. 74, pp. 357–378.

Herzog, T.N., Scheuren, F.J. and W.E. Winkler (2007). Data Quality and Record Linkage Techniques, Springer.

Holtkamp, R., Vroom, J.M. and A. Kremer (2016). Country example Netherlands: Use of the client registers for energy statistics: consumption by dwellings and businesses. Memo, Statistics Netherlands. (unpublished)

Hoogland, J. J. (2005). Micro-integratie van bedrijfseconomische gegevens: huidige praktijk en onderzoeksprojecten. Internal CBS nota. (Dutch)

Hoogland, J.J. and I. Verburg (2006). Handling inconsistencies in integrated business data. Paper presented at the Conference on European Statisticians, work session on data editing, Bonn, Germany, 25-27 September 2006, organised by the United Nations statistical commission and economic commission for Europe.

Jacobs, P.S. (1992). Joining Statistics with NLP for Text Categorization. Proceedings of ANLP-92, 3rd conference on applied natural language processing. M. Bates and O. Stock (eds.) Trento, Italy, association for computational linguistics, Morristown, NJ: 178–185.

Jong, W. de (1991)Technieken voor het koppelen van bestanden. CBS publication 1991. (in Dutch).

Kim, J.K., Berg, E. and T. Park (2016). Statistical matching using fractional imputation. Survey methodology 42, 19–40.

Klimek, J. (2015). Opinion of the Section for the Single Market, Production and Consumption on Family businesses in Europe as a source of renewed growth and better jobs European Economic and Social Committee. INT/765. Accessed at www.eesc.europa.eu, “EESC-2015-00722-00-00-AS-TRA-EN.docx”

Konen, R. (2015). POC Eenhedenomgeving. CBS nota, 2015. (in Dutch)

Lahiri, P. and M. D. Larsen (2005) Regression Analysis With Linked Data, Journal of the American Statistical Association, 100: 469, 222-230

Monitor topsectoren (2015). Methodebeschrijving en tabellenset. Centraal Bureau voor de Statistiek, Den Haag. (internetadres)

Nielsen, P.B. and Z. Tilewska (2011). Micro Data Linking - creating new evidence by utilising statistical registers. Case: international sourcing. Proceedings of the International Statistical Institute Conference, 21 - 26 August 2011, Dublin.

Oostrom, L., Walker, A.N., Staats, B., Slootbeek-van Laar, M., Ortega Azurduy, S. and B. Rooijakkers (2016). Measuring the internet economy in The Netherlands: a big data analysis. Discussion paper 2016:14, Statistics Netherlands.

RVO (2016). https:/mijn.Rvo.nl/gecombineerde-opgave. (in Dutch)

Reitsema, M. (2016). Vooronderzoek Productiewaarde Siergewassen. CBS nota. (in Dutch)

Schouten, B., Cobben, F. and J. Bethlehem (2009). Indicators for the representativeness of survey response. Survey Methodology 35, pp. 101-113

Steege, D. ter (2013). Vacaturestatistieken: Regiotoevoeging – methodebeschrijving. CBS nota. (in Dutch).

Struijs, P. (2015). Statistical Units Delineation and the Quality of Business Statistics. Paper presented at the European on Establishment Statistics Workshop. 7-9 September 2015, Poznan, Poland.

Task Force on Retail Trade Quality (TFRTQ) (2009). Non comparable changes. Document STS TF RTQ 01-2009-04. Eurostat, Luxembourg.

TOGAF-SABSA (2011). TOGAF and SABSA Integration. How SABSA and TOGAF complement each other to create better architectures. A white paper published by the Open Group. Available from www.opengroup.org

Tuoto, T. (2016). New proposal for linkage error estimation. Statistical Journal of the IAOS 32, 413–420.

United Nations (UN) (2008). International Standard Industrial Classification of all Economic activities (ISIC), Rev.4. Statistical Papers, Series M, No 4., Rev.4. UN publication, New York.

United Nations Statistics Division (UNSD) Friends of the chair group (2014). Paper on a new data infrastructure: the micro data linking approach in European business statistics. Proceedings of the International Conference on measurement of trade and economic globalisation, 29 September - 1 October 2014, Aguascalientes, Mexico. (available at http://unstats.un.org/unsd/trade/events/2014/mexico/)

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

Valk, F. van der and R. Spooner (2014). Completing the register. GeoConnexion International Magazine May 2014, pp. 23– 26. Available from www.geoconnexion.com.

Weller, M. (2015). Verrijking landbouwtelling met fiscale data 2010 – 2012. CBS memo, 25 november 2015. (in Dutch)

Winkler, W.E. (1995). Matching and record linkage, in B.G. Cox at al. (ed) Business Survey Methods, New York, J. Wiley, 355–384.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica (2012) Vol. 66, nr. 1, pp. 41–63.

Acknowledgements

We thank Danny van Elswijk, Lotte Oostrom and Wim Vosselman for providing suggestions for the case studies. Further, we thank the informants for their willingness and time to explain their case studies. Furthermore, we thank Bart Bakker, Eric van Bracht, Jan van der Laan, Johan Lammers, Sander Scholtus and Eric Smeets for their useful comments on earlier versions of the paper.
Klik hier als u tekst wilt invoeren.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2015–2016	2015 to 2016 inclusive
2015/2016	Average for 2015 to 2016 inclusive
2015/'16	Crop year, financial year, school year, etc., beginning in 2015 and ending in 2016
2013/'14–2015/'16	Crop year, financial year, etc., 2013/'14 to 2015/'16 inclusive

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

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information