Secondary data collection

12

Piet Daas and Judit Arends-Tóth

Explanation of symbols

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

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

Table of contents

1. Introduction to the theme

1.1 General description and reading guide

Statistics Netherlands aims to produce undisputed and up-to-date statistics about Dutch society. To be able to do this, Statistics Netherlands needs up-to-date and reliable data. These could be data that the organization itself collected (primary data) or data that is gathered and maintained by others (secondary data). Mindful of the costs and administrative burden involved in the collection of primary data, Statistics Netherlands aims to maximize the use of data already collected by others for statistics production. These type of data sources are commonly referred to as secondary sources. The entire process of collecting these types of data sources by Statistics Netherlands is referred to as secondary data collection.

1.1.1 Description of the theme

The technique of acquiring secondary data sources is not unique to the statistics field. It evidently has multidisciplinary appeal, with extremely diverse academic fields drawing on the information included in secondary sources (Daas and Beukenhorst, 2008). All methods used belong to the academic discipline known as secondary research (Golden, 1976), which involves using pre-existing data for a purpose different from that for which they were originally collected. In general, three different secondary research strategies can be discerned ('t Hart et al., 2005; Golden, 1976): content analysis, secondary analysis and systematic review. The focus in *content analysis* is on the content of various forms of human communication. Frequently used sources include newspapers, books, TV images, websites and paintings. A problem with content analysis is how to satisfactorily categorize and code what is often a large volume of unstructured data. *Secondary analysis* is about using quantitative data that were previously collected by other people for a different purpose. The general methods of secondary analysis differ very little from those used for primary data sources, such as Statistics Netherlands survey data (Golden, 1976; Wallgren and Wallgren, 2007). *Systematic review* (sometimes referred to as meta analysis) combines and investigates the output of other research concerned with the same or a similar phenomenon. An example of a systematic review is an investigation into the relationship between cancer and nutrition by combining all data published on the subject in the scientific literature over the past 15 years.

Statistics Netherlands uses all three of the above secondary research methods. Without doubt the most commonly used method is secondary analysis. The data provided by secondary data sources are used mainly as input for statistics. Two examples of secondary analysis from Statistics Netherlands practice are the use of company VAT data (Tax Administration, 2008) for the short-term statistics (Statistics Netherlands, 2008b) and the use of registers containing population related data for the Virtual Census (Schulte Nordholt et al., 2004). Statistics Netherlands uses the other two secondary research methods (content analysis and systematic

review) less frequently. An example of content analysis done by Statistics Netherlands is the historical review entitled 'Een eeuw statistieken (A century of statistics)' by Erwich and Van Maarseveen (1999). An example of a systematic review of Statistics Netherlands is the publication '200 jaar statistiek in tijdreeksen (200 years of statistics in time series') by Van der Bie and Smits (2001).

1.1.2 Problems and solutions

Statistics Netherlands want to increase the amount of data provided by secondary data sources for statistics. This approach aims to lower the administrative burden of respondents and the costs of data collection. Needless to say, the cost aspect is also affected by Statistics Netherlands' secondary data acquisition expenses and the amount of work needed to transform these data to the requirements of Statistics Netherlands. Furthermore, some secondary sources, such as registers, tend to have data about a complete population, which enables the publication of extremely detailed statistics. In addition, new statistics can be published with no additional burden, in particular when combining several sources.

These benefits make secondary data sources very interesting for statistics production. Statistics Netherlands accordingly uses secondary sources for the following statistical applications:

1) in statistics production as a replacement for primary data;
2) as a sample framework and source of auxiliary information in sample design;
3) as a source of auxiliary variables for estimates;
4) as input for statistical registers (such as the Business Register and the Social Statistical Database).

Also, the data in secondary sources may be ideal for some specific statistical applications, in particular when these data sources cover an almost complete population. These data sources can be used for:

5) detailed publications (such as regional statistics);
6) publications about special (infrequently occurring) events.

Secondary sources that cover multiple time periods, or have been used by Statistics Netherlands – with hardly any changes – for a relatively long period of time, are also very suited for:

7) detailed longitudinal studies.

However, an increase use of secondary data sources by Statistics Netherlands' makes it more dependent on:

1) the existence of and access to secondary sources;
2) the fitness for use (i.e. quality) of the secondary sources available;
3) the timely delivery of secondary sources.

Problems in one or more of these dependencies can have serious implications on the statistical output of Statistics Netherlands. In the most extreme case Statistics Netherlands might no longer be able to produce some of its statistics. The above mentioned three dependencies of Statistics Netherlands and the ways developed to cope with them are discussed in this part of the Methods Series. They jointly form the Secondary Data Collection volume of this series.

1.1.3 Guide to the reader

Chapter 2 of this document explains the 'existence of and access to secondary sources'. In order to be in a position to use data from secondary sources, Statistics Netherlands obviously needs to know which secondary sources exist and are available in the Netherlands. In addition the organization must be familiar with the content of these sources. This chapter also discusses delivery agreements, and ways to contact the data source holder.

Chapter 3 covers the fitness for use of a secondary source from a statistical point of view. Statistics Netherlands will, very likely, use the data in a secondary source for a purpose different from that for which it was originally collected. This may give rise to problems. For instance, a source may define an important variable slightly differently from the one used by Statistics Netherlands. It is important that Statistics Netherlands is able to rapidly access the fitness of use of the data in a secondary source for statistics, and to pinpoint the cause of the problem. These aspects are all quality related.

Chapter 4 discusses ways to cope as well as possible with an interruption in the delivery of a secondary source. The problems that may occur are identified by using risk analysis. The combination of measures taken is referred to as a fallback scenario, which intends to minimize the effect of delivery interruptions on the output of Statistics Netherlands. Chapter 5 presents some concluding remarks.

1.2 Scope and relationship with other themes

Statistics Netherlands meets some of its statistical needs through secondary data collection, in which data collected by others is used in the statistical production process. If any, Statistics Netherlands has extremely little say in determining the units over which the data were obtained, the data collection process, and the metadata definition of the variables included (Daas and Beukenhorst, 2008). When Statistics Netherlands collects its own data or enters into a joint undertaking with an external partner, in which it retains an intensive control over the data collected, the data collected is primary and not secondary. Additional information on the way Statistics Netherlands collects its (primary) data in the 'Data Collection Strategies', 'Questionnaire Design' and 'Organization of Field Work' Methods Series themes. There is additional information on the selection of units over which Statistics Netherlands collects data in the 'Sampling Theory' and 'Panels' Methods Series themes. The processing of secondary data is not within the scope of this document,. For more information on those topics readers are referred to the 'Data editing' and

'Micro-integration' Methods Series themes, and to the Register Methodology document (Daas et al., 2005).

1.3 Place in the statistical process

The data in secondary data sources replace some of the (survey) data originally collected by Statistics Netherlands (i.e. primary data). Secondary data collection is therefore part of the input phase of the statistical process. Primary and secondary data collection jointly yield all the data Statistics Netherlands needs to produce the output required. The fact that Statistics Netherlands has hardly any influence on the secondary data collection process (Daas and Beukenhorst, 2008) also has consequences for the subsequent process steps. Account must be taken of this in subsequent processing, integrating and aggregating steps. However, these steps do not belong to secondary data collection part of the statistical process.

1.4 Definitions

Concept	Description
Administrative source	A data source that arises in the management of public or private administrative matters
Base register	A register that complies or works towards full compliance with the requirements for authentic status. Base registers are the foundation of the system of government registers in the Netherlands
Primary source	A source containing primary data
Primary data	Data collected on behalf of Statistics Netherlands and for which Statistics Netherlands has defined the conceptual and process metadata
Primary data collection	The gathering of primary data by Statistics Netherlands
Primary research	Research that uses primary data
Register	A collection of data recorded and maintained in a structured way
Secondary source	A source containing secondary data
Secondary data	Data that is collected by others (i.e. not Statistics Netherlands), used by Statistics Netherlands for producing statistics.and where Statistics Netherlands has not defined the conceptual or process metadata
Secondary data collection	The acquisition of secondary data by Statistics Netherlands
Secondary research	Research that uses secondary sources
Source	General term for a dataset (a collection of data)

1.5 General information

The Statistics Netherlands business architecture has defined a preferred sequence of the use of data sources for statistics (Huigen, 2006). The aim of the sequence is to raise efficiency, reduce costs and minimize administrative burden on companies and individuals. The preferred data source sequence in descending order is:

      i)   internal sources;

      ii)   secondary sources already in use;

      iii) new (hitherto unused) secondary sources.

Where some of the essential data is not available in any existing source, is available but cannot be used satisfactorily, is subject to extremely unstable delivery, or would involve high acquisition or processing costs (Statistics Netherlands (CBS), 2003a-b, 2008a), the data may be collected by:

      iv)   adding one or more questions to an existing survey;

      v)   conducting a new survey.

Statistics Netherlands' preference for secondary data is evident in the above list. Statistics Netherlands will only perform its own observations when information demonstrably does not exist, is unfit for statistical purposes (i.e. when the quality of existing information is inadequate), cannot be delivered in good time, or would be too expensive to use.

2. Existence of and access to secondary sources

2.1 Short description

In order to use data from secondary sources, Statistics Netherlands must be aware of the existence of appropriate sources and have permission to use the data. Agreements also have to be made with the data source holder on the delivery and any other arrangements made, such as the additional need for feed back or assistance.

2.2 Applicability

Statistics Finland recommends statistical offices to comply as closely as possible with the following conditions in order to facilitate large-scale use of data from secondary sources (Statistics Finland, 2004).

1) Legal basis
   Legislation provides a key foundation for the use of administrative data sources for statistical purposes. Data protection arrangements must be part of these provisions.

2) Public approval
   The general public must have no objection to the use of 'their' data for statistical purposes. The reputation of a statistics institute as a reliable and eminent user of secondary sources is an important factor in acquiring and preserving public consent.

3) Unified identification codes
   It is vital that unified identification codes are used (for the various object types) across different sources. The identifiers enable fast data processing and give rise to fewer linkage errors. Sources without such identifiers can still be used, but costs are higher and their use will result in an increased number of errors (because of incorrect and missing links).

4) Reliable secondary data
   The secondary sources used must contain reliable data covering as much of the target population as possible. The use of these sources by multiple official organizations and the population itself increases data reliability and coverage.

5) Cooperation among administrative authorities
   Effective liaison between the authorities involved in using and maintaining the sources helps in the development of a stable and reliable system of secondary sources. The higher the level of agreements the better.

The introduction of the system of base registers in the Netherlands (Section 2.3.2) and the Statistics Netherlands Act (2003b), which provides legal permission to use data in public sources free of charge (Section 2.3.1), are excellent steps towards meeting most of the above listed conditions. Another important point is that

Statistics Netherlands must handle the data obtained with the utmost care and convey this to the outside world. To assure its continuous use in the future, no 'dents' can be tolerated in Statistics Netherlands' reputation as a reliable and prudent institute.

2.3 Detailed description

2.3.1 Scope for the use of secondary sources

Chapter 5 of the Statistics Netherlands Act (2003b) provides legal permission to use data from public sources. Statistics Netherlands may use sources maintained by the following organizations for statistical purposes free of charge:

    i)   national, provincial, and municipal government bodies, and district water boards;

    ii)   public bodies (as specified in the Joint Regulations Act (Wgr) and Article 134 of the Netherlands Constitution);

    iii)   independent central government administrative bodies.

Furthermore, in consultation with the Central Commission for Statistics, sources may be used that are maintained by:

    iv)   designated legal entities, some or all of whose duties are funded by the state or from the proceeds of statutory levies;

    v)   designated categories of companies, independent professionals, organizations and legal entities (but only if the sources mentioned under i - iv cannot deliver the data needed).

The suppliers must provide the data to Statistics Netherlands free of charge, within a specified period. An exception is made for data concerned with the Dutch banking system, which are provided by, or collected in consultation with, 'De Nederlandsche Bank' (DNB; the Dutch Bank). Article 36 of the Statistics Netherlands Act (2003b) states that the administrative burden for companies, independent professionals, organizations and legal entities must be as low as possible. The Statistics Netherlands Data Processing Decree (2003a) specifies the suppliers, the data to be provided and the delivery period.

2.3.2 Types of secondary source

Statistics Netherlands uses, or may use, a very diverse range of secondary data sources. Examples of secondary sources are the base registers[1], turnover data of the Tax Administration, survey data from the Netherlands Institute for Social Research (in Dutch: 'SCP'), price data of supermarket products and price data of houses on the internet. Some of these sources may be deemed to constitute a register or an

---

[1] Registers used throughout government that provide a comprehensive view of a given population of units, where the data are considered authentic by virtue of statutory regulation (E-overheid, 2009). Appendix A lists the thirteen base registers.

administrative source, but the distinction is unclear in some cases. Price data given on a website clearly do not constitute a register, and neither are they maintained for administrative purposes.

From the viewpoint of Statistics Netherlands' information requirement, there are three main groups of secondary data sources: statistical sources, registers (more in particular administrative registers) and 'other' sources. This categorization is based on assessing the sources against the various characteristics; see Daas and Beukenhorst (2008) for details. Figure 1 shows the various categories of secondary source distinguished. The figure also shows that the 'other' secondary source group is subdivided into administrative and non-administrative sources.

Figure 1. Secondary source categories with examples
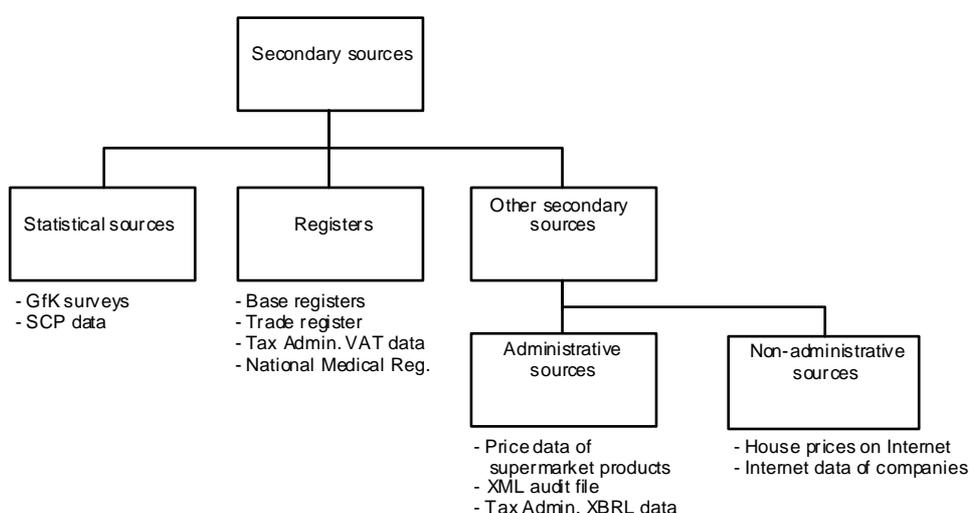
Some examples are given below for clarification. Among the statistical secondary sources that Statistics Netherlands uses are survey data collected by other organizations, such as those collected by the Netherlands Institute for Social Research (in Dutch: 'SCP'). Among the registers are the Trade Register of the chambers of commerce, the National Medical Registration (in Dutch: 'LMR') and the Municipal Personal Records Database (MPRD; in Dutch: 'GBA'). The group of registers may also be referred to as administrative registers (Daas and Beukenhorst, 2008). Base registers also belong to this group.

Base registers are special data sources that will become the foundation for the Dutch government's implementation tasks. Base registers contain data that is frequently used by the government in policy, implementation and enforcement (E-overheid, 2009). Storing data in a system of related registers is expected to help improve quality. Since governmental organizations are obliged to use the data in base registers and report any suspected errors in the data, its use will improve the quality of the data (see also Section 2.3.5). There are now thirteen designated base registers in the Netherlands (E-overheid, 2009), see Appendix A.

Next to the base registers, other registers may hold data that could be useful for some governmental organizations. When exploratory or feasibility studies reveal that using one of these registers might help reduce administrative burden, the potential reduction in burden is substantial, and that the data will be heavily used, it may be earmarked as a future base register (E-overheid, 2009). Statistics Netherlands is expected to increase the use of background and other variable data from base registers in the near future. This use will obviously depend on the rate at which the base registers become available in the Netherlands (Daas and Prins, 2007).

The last group of secondary sources is the 'other' category. Sources in this group are either administrative or non-administrative in nature. An example of an administrative source in the 'other' category is a database with prices of supermarket products, which is usually referred to within Statistics Netherlands as scanner data. An example of an administrative source in the 'other' category is a dataset with house prices collected from the internet. There are also secondary sources in the 'other' category with data that Statistics Netherlands may decide to use in the near future, such as radio frequency identification (RFID) tags, satellite-based navigation systems (GPS), mobile telephone data, and smart electricity and gas meters. There is additional information about potential statistical use of these sources in Daas et al. (2008). Statistics Netherlands is expected to start making use of data from the 'other' group of secondary sources in the near future.

Statistical registers, such as the Social Statistical Database (SSD) and the Business Register (BR), are emphatically *not* among Statistics Netherlands' secondary sources. These registers happen to be Statistics Netherlands internal products that are compiled from primary and secondary sources, and as such cannot be considered to be external sources of input data.

2.3.3 Sources available

It is the expectation that Statistics Netherlands will step up its use of secondary data. This makes it important to know what sources of information are available in the Netherlands. Therefore, Statistics Netherlands must strive to obtain the most up-to-date and complete list of potentially useful secondary sources in the Netherlands. No list of this kind is currently available in Statistics Netherlands, although there is a list of secondary sources that Statistics Netherlands uses. The Dutch Data Protection Authority (CBP, 2009) is a good place to start a search for potentially interesting but currently unused sources. Notably, this organization knows which organizations and sources hold personal data, because anyone processing data of this kind is obliged to notify the Authority. This information is held in the public notifications register for the Personal Data Protection Act. The only sources that are not included in this register are: i) those that are exempted, such as membership and payroll records, and ii) databases used by the police and judicial authorities.

2.3.4 Relationship management

Statistics Netherlands' increasing use of secondary sources necessitates good relations with the suppliers of the sources: the data source data holders. This is an activity that is in the field of relationship management. Statistics Netherlands has accordingly appointed account managers for the most important data source holders, such as the Tax Administration and the owners of the Dutch Municipal Personal Records Database (MPRD). Account managers are required both to provide and to gather information to and from the sources under the responsibility of their contacts. The duties of the account managers of Statistics Netherlands include making and monitoring agreements, managing expectations and detecting new developments (Huigen, 2006). For instance, an appointment for exploratory talks will be made with a view to establish the statistical usability of a potential source. Clear agreements must be drawn up with data source holders for sources that Statistics Netherlands decides to use, covering the delivery of the source (including metadata), the use of the data in the source, and the mutual obligations involved. The agreements must be recorded in a formal contract.

The account manager is also the internal Statistics Netherlands contact person for any questions and problems regarding the source, its delivery, and the source holder. Contact with the data source holder is therefore always channelled through, or follows consultation with, the account manager. As well as *ad hoc* contact through the account manager, it can be expected that there are generally at least one or more regular consultations per year between Statistics Netherlands and the data source holder. These consultations may, for instance, include annual meetings at a high (administrative) level, three-monthly user meetings, or two-monthly bilateral meetings of technical experts. Account managers will not necessarily attend all meetings of this kind on behalf of Statistics Netherlands. Needless to say, the account manager needs to be kept informed on the outcome of all meetings.

2.3.5 Feedback

Here, feedback refers to information provided back by Statistics Netherlands about the quality of the data in the source. This will often be prompted by problems with the data in the source. When there are problems with the specific data of individuals or companies, feedback will conflict with the provisions for confidentiality as laid down in the Statistics Netherlands Act (CBS), 2003b). This poses a dilemma for Statistics Netherlands. On the one hand it is important to inform the data source holder as completely as possible of any problems with data in the source, with a view to preventing future problems and improving data quality. On the other hand Statistics Netherlands is legally obliged to guarantee the confidentiality of individual data. The approach followed with the data of base registers illustrates the choice made by Statistics Netherlands.

Governmental organizations – which are obliged to use authentic data in base registers – have an *obligation* to report to the data holder any suspicion of error in the data on an individual level (Daas and Prins, 2008). For each base register, there is to be a separate obligation for providing feedback. The MPRD is currently the

only base register for which this obligation has been firmly established. Governmental organizations must report any detected or suspected errors in the MPRD back Statistics Netherlands is exempted from reporting pursuant to the Statistics Netherlands Act (2003b). The Statistics Netherlands Executive Board adopted the following resolution on the 3rd of July 2006 regarding the PR feedback obligation:

> 'Statistics Netherlands will not comply with the reporting obligations to be included in base register legislation, even if the register is not concerned with individuals or companies/organizations. Statistics Netherlands must avoid involvement in any legal effects of reporting or invasion of privacy, from either a strategic perspective or pursuant to the "confidentiality clause" (Article 37) of the Statistics Netherlands Act' (2003b).

The Personal Records and Travel Documents Database Agency (in Dutch: 'BPR') has endorsed this resolution. The BPR is the owner of the MPRD infrastructure. The BPR's support stems from their view that Statistics Netherlands does not constitute a governmental organization or administrative body within the meaning of the General Administrative Law Act. Statistics Netherlands does not actually have administrative dealings, in the sense of issuing decrees and performing other acts with an impact on individual citizens. The reporting obligation therefore has no meaning for Statistics Netherlands. From this it is to be expected that Statistics Netherlands will have no reporting feedback obligation in respect to errors in any base register whatsoever.

However, the above is not to be confused with the long-established Statistics Netherlands practice of contacting data source holders if the data supplied are unsatisfactory. When errors or ambiguity are found in the data of a unit in that source alone, Statistics Netherlands (CBS) *may* decide to approach the data holder for clarification. This may happen if a municipality sends information about planning permission for several homes stating an improbably high or low construction price and where it is absolutely necessary to be sure about the plausibility of the data. This is therefore a case of checking the internal consistency of a delivered record. Reporting may also follow problems with delivery from the source. For instance, the chain responsible for the delivery of the Benefit Entitlement Database (Tax Administration and the Institute for Employee Benefit Schemes) could never be improved rapidly without help from Statistics Netherlands. It is important for Statistics Netherlands to strictly avoid feeding back information about individual records by reporting instead on an overall level or – only in exceptional cases – with anonymized data. The privacy of individuals and companies must always be respected. Furthermore, facts that emerge from linking multiple sources from multiple suppliers may *never* be reported back. Provided the above principles are observed, feedback from Statistics Netherlands to data holders is allowable, and certainly for sources that are plagued with teething problems and other issues. This is beneficial to all.

2.3.6 Influence

For Statistics Netherlands, increased use of secondary data sources will coincide with decreasing influence over what kind of data is exactly collected. For instance, a wide range of decisions that are taken outside Statistics Netherlands, such as by data source holders, or politicians, may lead to the availability of fewer data or data of poorer quality. Statistics Netherlands must anticipate this trend as much as possible (see also Chapter 4). Moreover Statistics Netherlands must investigate the possible options for increasing its influence on the data collected in Dutch registers and other public (administrative) sources, such as the base registers. Statistics Denmark has solved this problem by arranging for a provision in the law that requires it to be consulted when new registers are set up (Statistics Denmark, 1995).

2.4 Examples

2.4.1 VAT data for short-term statistics

Parliament supported a motion tabled in 2006 by MP Charlie Aptroot calling for Statistics Netherlands to refrain from surveying small and medium-sized companies. In order to comply, Statistics Netherlands decided where possible to replace its own (primary) data collected from small and medium-sized business (<50 employees) for the short-term statistics with data from secondary sources. Companies in the short-term statistics sample must report their turnover for a given period (usually monthly or quarterly) to Statistics Netherlands. A secondary source with this kind of information is the VAT data collected by the Tax Administration (2008). With few exceptions, companies are obliged to report VAT data to the Tax Administration on a regular basis. Companies that are exempted include those with purely agricultural and healthcare operations, and small companies with annual VAT remittances below €5380 (Tax Administration, 2008). Statistics Netherlands has duly performed a preliminary study into the fitness for statistical purposes of VAT data. This study revealed the need for modifications to the delivery agreements with the Tax Administration. For instance, the agreement to deliver once every two weeks needed to be changed to once every two days, to enable Statistics Netherlands (CBS) to possess as much data as soon as possible. The primary data of the short-term statistics for small and medium-sized businesses is expected to be replaced by VAT data from the Tax Administration in due course.

2.4.2 Educational information

Statistics Netherlands is increasingly using secondary data sources with education-related data in order to satisfy the growing need for information about education. Mobility within the education system can be traced by linking databases with educational data from different years. This information facilitates the study of how secondary school and senior secondary vocational school students move to different types of education, shedding light in the process on premature school dropout rates. The sources used include education numbers files from the Information Management Group and the derived coordinated registers for each educational level

(e.g. secondary and higher education). The coordinated registers are drawn up by the Central Funding of Institutions Agency, a Ministry of Education, Culture and Sciences implementing body, in close collaboration with the customers (including Statistics Netherlands). Their purpose is to ensure that each organization uses the same sets of data. The consultation for compiling the various coordinated educational registers gave Statistics Netherlands an opportunity to help determine: i) the content of the registers, and ii) the derivation rules for the variables used. In addition to yielding registers that satisfy Statistics Netherlands requirements, this process also fosters good relations with data source owners and others involved. If any problems or questions arise, it is much easier to make contact and resolve the issues cooperatively.

2.5 Quality indicators

The increasing need for data from secondary sources is obliging Statistics Netherlands to step up its search for suitable and available sources. Work on drawing up and managing supply agreements with data source owners will also gain in prominence within the statistics process. A quality framework has been devised for assessing the quality of secondary sources (Daas et al., 2008). Part of this framework is specifically oriented to the quality aspects concerned with the supplier, the agreements and the delivery of the sources. The quality framework is covered comprehensively in Section 3.5. Table B1 in Appendix B shows the indicators for monitoring the quality of the source contact and delivery part.

3. Usability of secondary sources

3.1 Short description

A secondary source has data that are collected by others. In general, the purpose for which the data were originally collected differs from that for which Statistics Netherlands intends to use them. The nature and degree of these differences largely determine the usability (the quality) of the source for statistical purposes. This chapter discusses the main factors contributing to the usability of a source.

3.2 Applicability

Secondary sources provide extremely important information to Statistics Netherlands. Statistics Netherlands may use data in single secondary source for multiple statistics, which leads to great diversity in the requirements set on sources and data, depending on the statistic concerned. For instance, some variables are used for one statistic but not for another. Moreover, the requirements set on one and the same variable, in terms of recentness or accuracy, may differ from one statistic to another. An example is given for clarification.

Statistics Netherlands uses MPRD data about the population in numerous locations. The Population Statistics and the Labour Force Survey (LFS) both use this source, but set substantially different requirements on the recentness and completeness of the data. Population Statistics want to produce the most reliable and up-to-date as possible demographic trends in the Netherlands (Prins and Kuijper, 2007). It is therefore vital for Statistics Netherlands to obtain (and process) data that covers a wide variety of changes in the population and, at the same time, provides details of individuals as rapidly as possible. The LFS uses MPRD data mainly as a sample framework and as a source of background variables. This use affects the requirements set by the LFS on the speed at which changes are incorporated into MPRD data, because differences in some of the variables may affect the final LFS-outcomes. For instance, unlike Population Statistics, the LFS is not affected by the late inclusion or absence of births in a particular period. However, the late registration of changes of residence – within the same region – will affect the sample-based LFS but not the Population Statistics. For the LFS, the untimely registration of a change of residence may render it almost impossible for a Statistics Netherlands interviewer to contact the selected person.

3.3 Detailed description

There are several methodological issues that affect the statistical use of secondary data. These issues are part of the research field that Statistics Netherlands refers to as 'register methodology' (Daas et al., 2005), a research field distinct from sample theory used in primary data collection (e.g. through surveys). This section discusses the data collection related components of register methodology. The properties of secondary sources and the data affect the quality of use which is discussed below.

3.3.1 Technical aspects

Secondary sources may be delivered in various ways to Statistics Netherlands. The files send are often, but not necessarily, large or even 'bigger'. For instance, the MPRD is delivered in full to Statistics Netherlands once, at the start of each year, and thereafter only changes in a limited set of variables are communicated (Prins and Kuijper, 2007). The commonest delivery method for secondary sources is electronic transfer, in which the file or files are sent to Statistics Netherlands over a secure link with a transfer protocol, such as the File Transfer Protocol (FTP), a web service, or a dedicated governmental service. Occasionally a physical carrier will be delivered by mail or a courier, in the form of one or more CDs, DVDs or hard disks. Physical media are usually used only as a temporary measure in the early stages of a delivery process.

Statistics Netherlands must be able to open and completely read source databases and files on receipt. Reading problems may be caused by file corruption, a damaged carrier, decryption issues, or the use of a file format that is not a Statistics Netherlands standard. Any password for encrypted files must be sent separately and promptly. Finally, a completely readable file must comply with the agreed metadata definition, which unfortunately is not always the case.

3.3.2 Units

Apart from data about objects, secondary sources, in particular registers, also provide an overview of the population for a given object type. This also explains secondary sources' traditional use in statistics as sample frameworks (Erwich and Van Maarseveen, 1999). It is important for statistical purposes that the object types of units in a secondary source are clearly identifiable. Ideally the object type in the source will be identical to that required by Statistics Netherlands. Otherwise Statistics Netherlands may have problems in using the source. We present two examples, one involving an identical object type, and the other a non-identical object type.

Secondary sources with the object type 'natural person' are of interest to Statistics Netherlands. As is to be expected, this object type is indeed used by Statistics Netherlands and it also coincides with a uniquely identifiable naturally occurring unit. This unit is identified in many sources, such as the MPRD, with a Personal Identification Number (PIN). However, the PIN is not yet used universally in the Netherlands, and where it is not, it will be harder, but not impossible, to use the source (Arts et al., 2000; see Section 3.3.3).

The situation is trickier with the object type 'business'. Indeed, what exactly is a business? Is it the unit that submits returns to the Tax Administration, or the one that completes questionnaires for Statistics Netherlands? The problem with the 'business' object type is that many source data owners define it in their own way. For instance, the Tax Administration has tax entities, Statistics Netherlands has business units, and the chambers of commerce have legal entities and enterprises. The distinction is relatively unimportant for small businesses, in that for them the

units of the Tax Administration, Statistics Netherlands and the chambers of commerce tend to correspond (Aelen, 2004). The problem is mainly with large and medium-sized businesses, for which the units of the various official bodies tend to diverge. It is even possible for a 'business' to report to the Tax Administration as different tax entities for various different purposes (Aelen, 2008). This divergence makes it extremely difficult to uniquely identify large and medium-sized businesses and their (sub)divisions in the various registers, and unambiguously combine them into the units that Statistics Netherlands uses. This issue is referred to as the 'units problem', and it has a significant impact on Statistics Netherlands business statistics (Aelen, 2004, 2008). Notably, it is the large businesses that determine the economic trends in the Netherlands. Statistics Netherlands tries to resolve this difficult problem by dealing with the information on large, crucial and complex businesses at the earliest possible stage of the statistical process. This approach intends to ensure consistency between and within statistics (Verlinden, 2008). In support, a linking framework (the unit base) has been developed in which the relationship between the various individual register units for this important group of businesses is recorded and maintained. For this purpose, a units base has been constructed (Aelen, 2008).

Alongside this problem, a secondary source does not always cover the entire population of units that the statistic refers to. With a view to resolving this issue as satisfactorily as possible, all the sources that relate to similar units are combined, thereby creating a more complete picture of the population of units. There is additional information on this subject in the 'Micro-integration' part of the Methods Series.

3.3.3 Identification codes

A fundamental condition for the efficient use of secondary sources is the existence of uniform identification system for each object used. This requires object types to be clearly and consistently defined (see previous section), and identification codes to be precise, efficient, and proof against a wide variety of common errors. The precision of the identification of units depends on the object type. Precision is not an issue for some objects, such as individuals, automobiles and ships, which correspond with naturally occurring units. For other types of object (e.g. land parcels and buildings), operational or legal definitions must be sought that are both objectively observable and compatible with the objectives of the given source. The system of base registers (E-overheid, 2009) may provide a practical frame of reference for those object types. Additional information is needed for complex and dynamic objects, such as enterprises, businesses and households. In addition to clear definitions, sources must also exist that contain both identifying and relational data. The use those objects in practice a single, conscientiously managed, central linking framework (such as the units base for business objects mentioned above) is required.

The requirements on the identification codes used are more technical in nature. Account has to be taken of preventing overlap and counting errors. A self-validating

identification code, such as one that conforms to the 'eleven test' [2] for the Dutch PINs (the 'BSN'), is preferred because it can be checked for typing errors. Another issue is whether or not an identification should be free of information. For instance, the most important identification number in Denmark and Finland includes information on the date of birth and gender (Statistics Denmark, 1995; Statistics Finland, 2004). This causes few significant problems because these characteristics are almost immutable, albeit not quite (e.g. sex change). Among the disadvantages of an identification code that incorporates an attribute of the unit to be identified (e.g. a date of birth) are the risk of recognition and the fact that the code cannot be used for other object types that do not share the attribute (e.g. land parcels).

When the objects in a source have no unique identifying numbers, a unique combination of variables must be used for each of the objects in the source. For individuals this identification could be a combination of name and address (Arts et al., 2000).

3.3.4 Timestamps

Timestamps are an essential component of a data source. Timestamps that denote the existence of a unit in a source must be included. Timestamps relate to both flow (period) variables and stock variables. For instance, 'income' is an example of a flow variable, since it refers to income in a given time interval. Flow variables are sometimes also referred to as volume variables. An example of a stock variable is 'age'. It represents the situation at a point in time. Stock variables in a source may be tracked continuously or discretely. Continuous implies that the period of validity is stated, whereas discrete means that the value is known only at one or more specific points in time.

The period or periods to which data in secondary sources refers must be clearly stated. Some problems that may occur in this area are: incomplete specification of start or end period, discrepancies between recorded and actual data of events (particularly difficult with retrospective changes) and discrepancies between the intervals used by the data source holder and Statistics Netherlands.

Administrative delays at the data source holder are another time-related cause of problems. The information in a source may lag behind the real world if the data holder does not process and store the received data immediately. This problem is observed at Statistics Netherlands when, possibly very recent, survey data is linked to a secondary data source. This issue is simple to resolve if it is clear that the difference in the values of one or more variables in two linked sources is caused by administrative delay in one of them. Unfortunately, the situation is not always immediately clear.

---

[2] The 9-digit Dutch PIN (BSN) has to satisfy the following rule: (1st digit BSN x 9 + 2nd digit BSN x 8 + 3rd digit BSN x 7 + 4th digit BSN x 6 + 5th digit BSN x 5 + 6th digit BSN x 4 + 7th digit BSN x 3 + 8th digit BSN x 2 - 9th digit BSN x 1) / 11 = a whole number.

3.3.5 Variable definitions

It can happen that the definition of a variable in a secondary source incompletely agrees, or disagrees, with the definition of the variable on which Statistics Netherlands wants to publish. An example of this is the difference between the concept 'unemployment' of Statistics Netherlands and that of the job centres in the Netherlands (Van den Elshout et al., 2007). Statistics Netherlands has several ways to deal with such a problem. The value of a variable in the source can be *adjusted* to the (expected) value of the variable in accordance with the Statistics Netherlands definition, or derived by *combining data*. However, it is also possible to abandon use of the old Statistics Netherlands definition of a variable in favour of a new one. An example of this is the introduction of the 'payment concept' in the SSD (Arts and Hoogteijling, 2002). Should any of the above options be unusable for whatever reason, different sources could be sought with more information on the variable, or a survey could be conducted to collect the required information. There is additional information on this topic in the 'Micro-integration' part of the Methods Series.

3.4 Examples

3.4.1 Income Information System

Statistics Netherlands and the Ministry of Finance jointly collect information about the composition of the income and wealth of natural persons in the Netherlands. The Income Information System (IIS) was set up for this purpose. The system is managed by Statistics Netherlands. The IIS is a combination of data provided by the Tax Administration, the Supplementary Benefits Department and the Information Management Group. In the ISS the following secondary sources are combined:

- income tax returns and assessments;
- Healthcare Insurance Act returns and assessments;
- profit returns for personal and corporate income tax;
- supplementary benefit data (data on housing, care, and childcare allowance);
- wage tax data ('Fibase');
- interest payments ('Rentebase');
- dividend distributions ('Dividendbase');
- Student Finance Act records (student grants and loans);
- immovable property valuations;
- relationships management data for individuals.

The sources are linked on an individual level by means of the BSN. The Ministry of Finance uses the data for assessing the final implications of policy proposals and for evaluating tax regulations and estimates. Statistics Netherlands uses the data mainly for compiling income and wealth statistics on persons. For instance, detailed income and wealth information is delivered for each socioeconomic category and region for the Income Panel Survey (IPO). The IPO outcomes are published on StatLine (the Statistics Netherlands output base) and used in the Distribution of Wealth Yearbook, the Poverty Monitor and various other articles and press releases.

3.4.2 Social Statistical Database

The Social Statistical Database (SSD) is produced by linking data of individuals in secondary sources and Statistics Netherlands surveys on a micro level (Arts and Hoogteijling, 2002). The SSD core consists of various interlinked registers that contain demographic and socioeconomic data. Although the name suggests otherwise, the SSD core is not a single physical database, which is also not possible in view of the variety of object types involved (e.g. jobs, social benefits, individuals). The databases created for each object type are: a jobs database of employees in the Netherlands, a database of self-employed people, a database of employees working abroad, six benefit payment databases and a database of individuals plus demographic information. There are also various SSD satellites that focus on specific topics, such as the social position of ethnic minorities and crime related issues in the Netherlands. Both the SSD core and satellites are continuously improved.

The following secondary sources are used for the SSD core:

- Municipal Personal Records Database (MPRD);
- wage tax data ('Fibase');
- employee insurance records;
- income tax returns;
- agricultural census data;
- work incapacity schemes records;
- unemployment benefits records;
- national assistance records;
- Older and Partially Disabled Unemployed Workers Income Scheme, and Older and Partially Disabled Former Self-Employed Persons Income Scheme records;
- immovable property valuations;
- rulings on compulsory health insurance for self-employed people;
- relationship management data;
- housing statistics;
- Centre for Work and Income registrations;
- the Central Register of Higher Education Enrolment;
- tuition and course fee register;
- examination results register;
- student grants and loans records.

Data from the Labour Force Survey and employment and wages statistics are added to the above. The combination of sources yields a wealth of information on persons, jobs and social benefits for compiling statistical overviews. An important objective of the SSD is the publication of cohesive and consistent information. The SSD is accordingly the basis of many Statistics Netherlands social statistics.

The availability of several years' worth of SSD information permits objects to be tracked over time, such as to establish who have stopped receiving disability benefit, and, of those, who have found employment. These analyses enable to provide answer to questions such as 'how long a person receives unemployment benefit?' and 'whether there are any differences between population groups?'. It is also possible to obtain a considerable amount of information on small regional units and

small groups in the community because of the vast amount of information provided by the combination of all the sources in the SSD. Some examples are the number of employees entering and leaving the health and welfare care systems, and the proportion of the population of the municipality of Pekela receiving unemployment benefits.

The data in the SSD are not freely distributed because of the personal and sensitive nature of some of the information. Aggregate SSD-data is available when published in StatLine, and some institutes may obtain data for a fee, subject to consultation with and security arrangements determined by the Statistics Netherlands Centre for Policy Related Statistics.

3.5 Quality indicators

3.5.1 Quality framework for secondary sources

Statistics Netherlands has developed a quality framework for secondary sources (Daas et al., 2008), which is used in determining the *general* statistical usability of a secondary source. The framework was not set up to determine whether a source is fit for a specific purpose, such as the feasibility of replacing the unemployment figure from the Labour Force Survey with data from the Centre for Work and Income. These kinds of issues are highly specific, and are therefore outside the scope of the general framework. Moreover, the demand for specific quality measures requested by multiple users would also further inflate the framework and negatively affect its general usefulness. It is simply impossible to set up a general quality framework that satisfies all specific needs for all potential users. It can be expected that one or more supplementary, highly specific, quality measures are additionally needed (Daas et al., 2008).

3.5.2 Quality framework structure

The quality framework for secondary sources has three very distinct views, which are also called hyperdimensions (Karr et al., 2006). These hyperdimensions are: Source, Metadata, and Data (Daas et al., 2008).

Each of the three 'views' on the quality of secondary sources illustrates a different set of quality components. The three hyperdimensions do not overlap and are presented in order of increasing detail. They are all product based. Assessing the quality of a secondary source, i.e. the usability for statistics, requires evaluating the Source, Metadata and Data hyperdimensions in subsequent order.

The quality components associated with each of the Source, Metadata and Data hyperdimensions are specified as dimensions. A dimension should be viewed as a component of quality restricted to one specific characteristic. Each dimension contains one or more quality indicators, each of which refers to a specific measurable part of quality for that dimension. A quality indicator is always measured by at least one method, but there may be more than one. A measurement method is a procedure for determining or estimating the value of a quality indicator.

The connection between the various quality components identified in the quality framework is shown in Figure 2.

Figure 2. The various components of quality and their interrelationship

```
        ┌─────────────────────────┐
        │     HYPERDIMENSION      │
        └─────────────────────────┘
                    │
        n > 1       │
                    ▼
            ┌─────────────────────────┐
            │       DIMENSION         │
            └─────────────────────────┘
                        │
            n >= 1      │
                        ▼
                ┌─────────────────────────┐
                │    QUALITY INDICATOR    │
                └─────────────────────────┘
                            ▲
                1 : n       │
                            ▼
                ┌─────────────────────────┐
                │   Measurement method    │
                └─────────────────────────┘
```
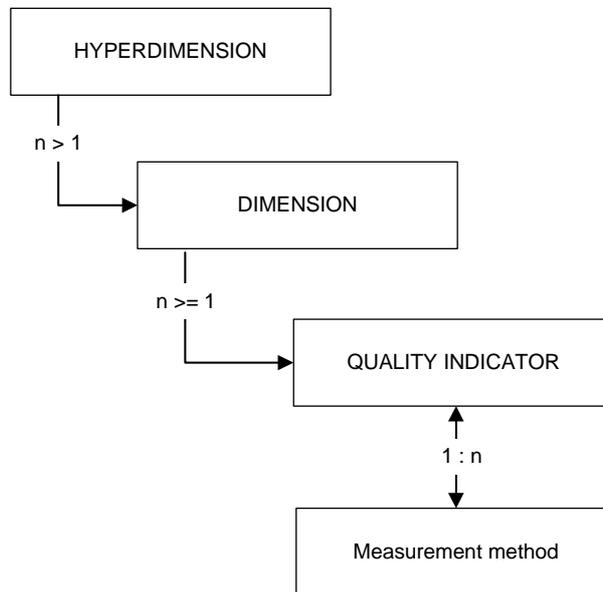
Source hyperdimension

The Source hyperdimension regards the secondary source as a single file that is delivered to Statistics Netherlands, where the focus is on the supplier and on the delivery of the file. Some examples of Source hyperdimension quality components are: arrangements on the delivery of the source, supplier data and contacts, and security and privacy agreements. The components of the Source hyperdimension are shown in Table B1 of Appendix B. The measurement methods in the Source hyperdimension are mainly qualitative. A checklist has been drawn up for assessing the quality components in the Source hyperdimension (Daas and Arends-Tóth, 2007). Evaluation of the Source hyperdimension reveals whether or not a source can potentially be used for statistics. If so, the quality components in the Metadata hyperdimension need to be evaluated.

Metadata hyperdimension

The Metadata hyperdimension is concerned with the conceptual metadata components of a secondary source, along with some process-related meta components. Some examples of quality components in the Metadata hyperdimension are the clarity and comparability of the variable and population definitions, data about the time period to which the data refer, and any checks performed on the data by the data source holder. The Metadata quality components are listed in Table B2 of Appendix B. The Metadata hyperdimension has contains qualitative measurement methods. Assessing the quality of the Metadata hyperdimension is included in the above mentioned checklist (Arends-Tóth and Daas, 2008). Evaluation of the Metadata quality indicators reveals how fit the source

is, if at all, for statistical use. If a source can be fully or partly used, the Data hyperdimension must be evaluated.

Data hyperdimension

The Data hyperdimension focuses on the quality components related to the data in the secondary source. The quality aspects components in the Data hyperdimension mainly correspond with the accuracy, completeness, time-related and integrability components of the data (Daas et al., 2011). In addition , some technical data checks are included. The Data quality components are shown in Table B3 of Appendix B. Most measurement methods in the Data hyperdimension are quantitative (Daas et al., 2011; Schulte Nordholt et al., 2011).

3.5.3 Use of the quality framework

When assessing the quality of a data source the user should first consider the Source, then the Metadata and finally the Data hyperdimension. Checklists have been developed for assessing the Source and the Metadata hyperdimensions (Daas and Arends-Tóth, 2007; Arends-Tóth and Daas, 2008). The most appropriate approach for Data is currently under investigation. Any problems encountered during the assessment of the quality components in a hyperdimension must first be solved before studying the next hyperdimension. Evaluation should stop if unfitness for statistical use is found in a hyperdimension, and any subsequent hyperdimensions disregarded. Users will obtain a definitive answer to the usability for statistics of a data source after evaluating the Data hyperdimension. It is possible (and indeed very probable) that a specific measurement is needed to answer the specific question(s) some users might have (Kuijvenhoven and Schouten, 2008).

4. Fallback scenarios

4.1 Short description

A fallback scenario is a combination of measures that enables Statistics Netherlands to deal with the unfavourable consequences of the temporary unavailability of some or all of the data provided by secondary sources. The problems are mapped out using risk analysis. Being an extremely dynamic field, it is possible that some of the details may differ in practice from the situation described in this document.

4.2 Applicability

The increasing use of secondary sources for statistics makes Statistics Netherlands more dependent on the timely availability of secondary data. This poses risks with a potentially negative impact on statistics production. The temporary or permanent unavailability of a secondary source is one example. If a secondary source is unavailable temporarily, Statistics Netherlands – often in haste – will determine how to deal with the lack of data (Frenken et al., 2008). Situations of this kind may mean that certain statistics cannot be published, at least temporarily. One solution to cope with this issue is to prepare a fallback scenario, which is a combination of measures to limit the unfavourable consequences for users. A fallback scenario must be applied only to accommodate temporary unavailability of some or all of the data in a secondary source.

The Statistics Netherlands Executive Board has decided that a risk analysis needs to be performed and a fallback scenario has to be drawn up for the statistics that are on the list of 'image-determining' statistics of Statistics Netherlands (see Appendix C) and that depend directly or indirectly on the information provided by secondary sources (Statistics Netherlands Executive Board minutes, 26 January 2009). The Statistics Netherlands image-determining statistics are those statistics that pose a substantial risk both to users' and Statistics Netherlands' reputation in the event of nontimely publication. Fifteen of the sixteen image-determining statistics depend on secondary sources (see Appendix C). It is also recommended that a fallback scenario be drawn up for all other Statistics Netherlands statistics that highly depend on secondary data.

The following is the recommended general approach to developing a scenario for dealing with the temporary unavailability of an important secondary source:

1. determine if it is feasible, in terms of time and costs, for Statistics Netherlands employees to obtain – preferably via alternative external sources – the missing data elsewhere;

2. apply a model-based approach if there is no alternative for the missing data and some of the data about the reporting period are still available. Application is subject to the plausibility of the quality of the results provided by the model;

3. notify the important users of the potential consequences of unavailability of the source;

4. postpone publication if the above options are impossible.

4.3 Detailed description

Statistics Netherlands wishes to present itself as a reliable and stable supplier of socially relevant information. Consequently, the timely publication and high quality of Statistics Netherlands' statistical information must be beyond dispute. The increasing use of secondary sources makes Statistics Netherlands more vulnerable in this respect. Fallback scenarios for statistics that use secondary data may be drawn up in order to avoid any related problems. Risk analysis can help identifying vulnerabilities. The standard process specification methodology and the development rate adopted for the Civil Service Data Information Security Decree 2007 (VIR, 2007) must be observed.

4.3.1 Defining the usability of fallback scenarios

It is unrealistic to prepare fallback scenarios for all imaginable situations. Fallback scenarios are often tailored to a specific practical situation. The best solution in any given situation will depend on what exactly is missing and the quality of the remaining information. The chosen solution must also address the costs and the time available, which will usually be short. It is therefore advisable to draw up fallback scenarios only for Statistics Netherlands statistics for which the unavailability of secondary data would have serious consequences. The early detection of potential problems increases the chance of a satisfactory response. This is why active relationship management, contact with the data source holder, is very important.

No fallback scenario has to be drawn up for a source that is permanently unavailable. For these cases, a new statistical data collection process needs to be started in order to satisfy the statistical output obligation. The transition period is expected to be lengthy. External pressure and publication obligations may necessitate the introduction of other 'creative' temporary solutions in the meantime, such as a completely model-based figure, a nowcast[3] , an expert 'guess', or use of the Delphi method[4]. It goes without saying that such a situation must be communicated clearly to the outside world. The emergency measure applied in the transition period can be viewed upon as a temporary fallback scenario.

---

[3] A nowcast is a forecast relating to the current time (or, rather, to the recent past). A nowcast produces an estimate for the period now behind us, but for which no direct statistical observation have been made. Producing a nowcast uses the same methods as for a forecast.

[4] The Delphi method is a research tool in which opinions are solicited from many experts about a topic for which there is no consensus. The answers of other experts are fed back anonymously in several rounds until consensus is reached. The method is named after the Oracle of Delphi.

4.3.2 Developing a fallback scenario

The first thing to investigate in the event of temporary unavailability of a secondary source is the possibility of achieving timely outcomes of acceptable quality in another way. The preferred approach if the costs incurred are insubstantial would normally be to obtain the missing data from alternative secondary sources, or from primary data. Where possible Statistics Netherlands always prefers to base its figures on collected data, but a modelling approach may be used if this is impossible in the short term. In view of the limited time available, a suitable model has to be previously developed and tested, and directly available If none of the proposed solutions can be used, postponement of the publication is inevitable. However, it is always possible for the director responsible to deviate – in specific cases – from the preferred priority order. In practice only a few days will be available, and in many cases there will be no choice but to take whatever appears to be the most appropriate action in those circumstances.

The above shows that there are normally two ways of achieving acceptable outcomes when faced with the temporary or partial unavailability of secondary data:

1)   use other sources (primary or secondary);

2)   estimate by modelling.

It is obviously possible to combine these options. Otherwise, the last resort is to:

3)   postpone publication.

How to handle these options is discussed below. A risk in resorting to a fallback scenario is a discontinuity in the published figures. Rectifying a discontinuity is no easy matter, and is usually not possible in the brief period in which a fallback scenario is employed. This consideration must also be accommodated as well as possible while developing a fallback scenario.

4.3.2.1 Use other sources

If it is possible to obtain the missing data from somewhere else rapidly and exactly without incurring substantial costs – from primary data or from alternative secondary sources – this is to be preferred to model-based estimation. Where possible, Statistics Netherlands always prefers to base its figures on real data. Directly approaching the contact person of a data source, for instance by telephone, is an example of a way to collect the data missing, albeit possibly with less detail (Frenken et al., 2008). Another option would be to use a different secondary source with identical, or similar, data. Once another data provider is found, little time can be spent on obtaining the data. Using a new source is likely to introduce a discontinuity in the published figures.

4.3.2.2 Model-based estimation

A model-based approach may be a solution if only some of the data have been received. Part of the data may be missing for all units, such as for the third period of a quarterly statistic. Alternatively, all the data of part of the population could be

missing, e.g. as a consequence of a defective error detection routine of the data source holder, or irreparable ICT-related storage corruption or communication problems. Essential conditions for the use of a model-based approach are:

  i)   some of the data are available;

  ii) a model has already been developed.

The research time involved in developing and testing a model and the likely unavailability of sufficiently long time series are obstacles to the use of models as fallback scenarios. Furthermore, it remains uncertain whether even a properly developed and tested model can actually be applied in any specific problematic situation. This is an additional obstacle to applying a model-based approach. Unless the model was developed with flexibility in mind – including the rapid definition and application of an ad hoc fallback scenario – the model-based approach will often prove insufficiently effective. This is an unfortunate conclusion, because in many cases it should be possible to produce a model-based estimate of acceptable quality based on knowledge and experience in the past. The creative emergency measures provided as examples at the end of Section 4.3.1 are generally also within the domain of the 'modelling' approach. However, these are weak models and are not to be preferred. If they do have to be applied, this must be clearly communicated to the users.

4.3.2.3 Postpone publication

Postponing publication is resorted to if no alternative options are available, through different methods of data collection or modelling, for achieving an acceptable outcome quality in a relatively short time. Postponement is an option only if it is foreseeable that:

  i)   the problems with the source can be resolved within a given (preferably short) time;

  ii) Statistics Netherlands' image will not be significantly harmed (public interest).

Postponement is no longer an option if problems are foreseen on either or both points. Publication of any of the image-determining statistics of Statistics Netherlands must never be postponed, and other sources or a model must be available instead.

4.4 Examples

At first glance the opportunities for using substitute sources or model-based approaches appear limited. It is often thought that rapidly switching to new secondary sources, starting or resuming primary data collection, or developing a model, will take considerable time. The change can also involve substantial expense. Several situations have arisen in Statistics Netherlands' practice that highlight the importance of fallback scenarios. It is important that the proposed measures are realistic, and some options will even be obvious. This is illustrated by several examples of fallback scenarios used by Statistics Netherlands in practice.

1. An example of using primary data collection as a substitute.

   - Problem: a central healthcare body failed to deliver the daily fees for nursing in good time.

   - Solution: hospitals were contacted by telephone to obtain the fees.

2. An example of using a secondary data source as a substitute.

   - Problem: the scanner data with the product prices of a major supermarket chain were late.

   - Solution: the prices were obtained from the supermarket chain's Internet site. This solution is preferable to imputation from the price trends of other supermarkets.

3. An example of using a model-based approach.

   - Problem: a trade association in the energy market failed to deliver information about electricity price trends in good time.

   - Solution: information that was available to Statistics Netherlands suggested that electricity prices would (retroactively) increase sharply. It was therefore considered likely that the outside world would be misinformed with an excessively low inflation rate. It was decided to estimate the expected rise. The estimate was subsequently used, which turned out to be a good (and wise) decision.

4.5 Quality indicators

The Civil Service Data Information Security Decree (VIR, 2007) sets down the required availability of information systems (including databases), risk assessments and corrective measures to be taken. This framework incorporates the issue of fallback scenarios (Statistics Netherlands Executive Board minutes, 26 January 2009). The Civil Service Data Information Security Decree prescribes standard process measures ('dependency and vulnerability analyses') that are to be evaluated and monitored regularly. The discontinuity risks to a process attributable to nontimely delivery, or unsatisfactory quality, are part of the regular evaluation prescribed in the Civil Service Data Information Security Decree.

A standard template has been created to determine the need of developing a fallback scenario for a given statistic. The risk assessment component of the template estimates whether there is any need to create a fallback scenario. Among the important aspects considered are the assessment of problems with the delivery of the source, the stability of the delivery, and the impact on the output of Statistics Netherlands. If delivery problems are likely to occur, with severe consequences for Statistics Netherlands, it is advised to draw up a fallback scenario. Because this is always the recommendation for image-determining statistics of Statistics Netherlands, it has been decided to develop fallback scenarios whenever this group of statistics depends directly or indirectly on secondary sources (report Statistics Netherlands Executive Board minutes, 26 January 2009). In all other cases the

process owner is responsible for deciding whether or not to develop a fallback scenario. The content of the template for fallback scenarios is shown in Table 1.

Table 3. Evaluation template for fallback scenario

Which statistics are involved?
- Name
- Division, sector, task force
- Uses the following secondary sources: …

General information about each secondary source
- Name of source
- Name of data source holder
- Contact person at holder
- Statistics Netherlands account manager
- Other Statistics Netherlands contacts (if any)
- What regular contacts are there between the data source holder and Statistics Netherlands?

Risk assessment
- How great is the estimated risk of the data source holder being unable to deliver the source?
- What are the consequences for Statistics Netherlands?
- How stable is the delivery of the source?

Process information of the statistic
- Are there any alternative sources, or does any research exists which indicates that the data could be derived from a model if the source or any of the required variables are unavailable?
- Possible fallback scenarios: 1. wait; 2. model-based approach; 3. use alternative source

Summary
- Risk of untimely publication or nonpublication of the statistic
- Consequences for Statistics Netherlands
- Available alternatives

Meta-information checklist
- Update frequency of the checklist
- Date of last update
- Drawn up by:
- Signed (name and position)

---

5. Conclusion

Statistics Netherlands' increasing use of data from secondary sources leads to increasing dependency of the bureau on data collected by others. More account will have to be taken of this dependency in designing and redesigning Statistics Netherlands' statistical processes, which must be accompanied by an increasingly external focus. Secondary sources need to be sought actively, and contact must be maintained with the data source holders on management, process and expert levels. It is also vital for Statistics Netherlands to find ways to enhance its influence on the content of secondary sources, in order to ensure that sources continue to collect data that are relevant for Statistics Netherlands. Monitoring all of the above is possible with the quality framework described in this document.

6. References

Aelen, F.W.L. (2004), *Startcursus statistiekboom – Bedrijvensregister (Statistics tree initial course – Companies register)*. Statistics Netherlands, Heerlen.

Aelen, F. (2008), *HEcS+ Eenhedenproblematiek, Eindrapportage (Redesign of Economic Statistics+ Units issue), Final report*. Project document, Statistics Netherlands, Heerlen/Voorburg.

Arends-Tóth J. and Daas, P.J.H. (2008), *Kwaliteitsindicatoren van registers: de Metadata hyperdimensie (Quality indicators for registers: the Metadata hyperdimension)*. Internal report, Statistics Netherlands, Heerlen.

Arts, C.H., Bakker, B.F.M. and Lith, F.J. van (2000), Linking administrative registers and household surveys. *Netherlands Official Statistics* 15, pp.16-21.

Arts, C.H. and Hoogteijling, E.M.J. (2002), Het Sociaal Statistisch Bestand 1998 en 1999 (The 1998 and 1999 Social Statistical Database). *Sociaal Economische Maandstatistiek* 12, pp 13-21.

CBP (Dutch Data Protection Authority) (2009), Dutch Data Protection Authority website (http://www.cbpweb.nl/).

CBS (Statistics Netherlands) (2003a), Besluit gegevensverwerving CBS (Statistics Netherlands Data Processing Decree). *Bulletin of Acts and Decrees of the Kingdom of the Netherlands*, 552.

CBS (Statistics Netherlands) (2003b), Statistics Netherlands Act. *Bulletin of Acts and Decrees of the Kingdom of the Netherlands*, 516.

CBS (Statistics Netherlands) (2008a), *CBS Kwaliteitsverklaring (Statistics Netherlands Quality Statement)*, 28 April, Statistics Netherlands, Voorburg/Heerlen.

CBS (Statistics Netherlands) (2008b), *Dataverzameling (Data collection)*. Information about Statistics Netherlands surveys on the Statistics Netherlands website (http://www.cbs.nl/nl-NL/menu/methoden/dataverzameling/default.htm).

Daas, P.J.H. and Arends-Tóth, J. (2007), *Kwaliteitsindicatoren voor registers: evaluatie van de Bron hyperdimensie (Quality indicators for registers: evaluation of the Source hyperdimension)*. Internal report, Statistics Netherlands, Heerlen.

Daas, P.J.H., Arends-Tóth, J., Schouten, B. and Kuijvenhoven, L. (2008), Quality Framework for the Evaluation of Administrative Data. In: *Proceedings of Q2008 European Conference on Quality in Official Statistics*. Statistics Italy and Eurostat, Rome.

Daas, P. and Beukenhorst, D. (2008), *Databronnen van het CBS: Primaire en secundaire bronnen (Statistics Netherlands data sources: Primary and secondary sources)*. Internal report, Statistics Netherlands, Heerlen.

Daas, P., Jeurissen, E., Boonstra, H.J. and Nieuwenbroek, N. (2005), *Registertheorie: Registers en het CBS (Register theory: Registers and Statistics Netherlands)*. Internal report, Statistics Netherlands, Heerlen.

Daas, P., Ossen, S., Tennekes, M., Zhang, L-C., Hendriks, C., Foldal Haugen, K., Cerroni, F., Di Bella, G., Laitila, T., Wallgren, A. and Wallgren, B. (2011) Report on methods preferred for the quality indicators of administrative data sources. Second deliverable of workpackage 4 of the BLUE Enterprise and Trade Statistics project, September 28.

Daas, P.J.H. and Prins, C.J.M. (2007), *Programma van eisen voor basisregistraties en voor het statistisch gebruik van deze registraties (Requirements specification for base registers and for the statistical use of these registers)*. Internal report, Statistics Netherlands, Heerlen/Voorburg.

Daas, P., Roos, M. and Puts, M. (2008), *Waarnemingsinnovatie: nieuwe bronnen en mogelijkheden (Innovative data collection: new sources and new possibilities)*. Statistics Netherlands, Heerlen.

Erwich, B. and Van Maarseveen, J.G.S.J. (1999), *Een eeuw statistieken, historisch-methodologische schetsen van de Nederlandse officiële statistieken in de twintigste eeuw (A century of statistics, historical and methodological sketches of Dutch official statistics in the twentieth century)*. Statistics Netherlands, Voorburg/Heerlen.

Eurostat (2003), *Definition of quality in statistics*. Item 4.2: Methodological documents, Working group 'Assessment of the quality in statistics', Sixth meeting, Luxembourg, 2-3 October.

Eurostat (2005), *Standard quality indicators*. Quality in statistics, Seventh meeting, Luxembourg, 23-24 May.

E-overheid (2009), *E-overheid planningsoverzicht / voortgangsrapportages (E-government planning summary/ progress reports)*. Web page with progress reports of the basic e-government facilities (http://www.e-overheid.nl/atlas/planning/planning.html).

Frenken, J., Hoven, L., Prins, K. and Slootbeek, G. (2008), *CBS-beleid terugvalscenario's (Statistics Netherlands fallback scenario policy)*. Internal report, Statistics Netherlands, Heerlen/Voorburg.

Golden, M.P. (1976), *The research experience*. F.E. Peacock Publishers Inc., Itasca, Illinois, USA.

Huigen, R. (2006), *ICT Masterplan: CBS-architectuur, Business- en informatiemodel (ICT Master Plan: Statistics Netherlands architecture, Business and conceptual model)*. Statistics Netherlands, Heerlen/Voorburg.

Karr, A.F., Sanil, A.P. and Banks, D.L. (2006), Data quality: A statistical perspective. *Statistical Methodology*, 3, pp. 137-173.

Kuijvenhoven, L. and Schouten, B. (2008), *Kwaliteitsaspecten voor de hyperdimensie data (Quality aspects for the Data hyperdimension)*. Internal report, Statistics Netherlands, Voorburg.

Prins, K. and Kuijper, H. (2007), Bevolkingsstatistieken onder het persoonskaartenstelsel en het GBA-stelsel: overeenkomsten en verschillen (Population statistics under the personal card system and the Municipal Personal Register Database system: similarities and differences). *Bevolkingstrends (Population trends)* 55(1), pp 14-33.

Schulte Nordholt, E., Hartgers, N. and Gircour, R. (2004), *The Dutch virtual census of 2001*, *Analysis and methodology*. Statistics Netherlands, Voorburg/Heerlen.

Schulte Nordholt, E., Ossen, S.J.L. and Daas, P.J.H. (2011), Research on the Quality of Registers To Make Data Decisions in the Dutch Virtual Census. Paper and presentation for the 58th Session of the International Statistical Institute, Dublin, Ireland.

Statistics Denmark (1995), *Statistics on persons in Denmark, a register-based statistical system*. Office for Official Publications of the European Communities, Luxembourg.

Statistics Finland (2004), *Use of registers and administrative data sources for statistical purposes*. Best practices of Statistics Finland. Handbook 45.

't Hart, H., Boeije, H. and Hox, J. (2005), *Research methods*. 7th impression. Boom, Amsterdam.

Tax Administration (2008), *Btw algemeen (VAT general)*. Information on the Tax Administration website (www.belastingdienst.nl/zakelijk/omzetbelasting/ob01/ob01.html).

Van den Elshout, B., Jacobi, C. and Van de Valk, J. (2007), Werklozen versus niet-werkende werkzoekenden (Unemployed vs non-working job seekers). *Sociaaleconomische trends*, 3, pp. 46-53.

Van der Bie, R.J. and Smits, J-P. (2001), *Tweehonderd jaar statistiek in tijdreeksen 1800–1999 (Two-hundred years of statistics in time series 1800–1999)*. Statistics Netherlands and University of Groningen, Voorburg/Heerlen, Groningen.

Verlinden, B. (2008), *Procesgang ConGO in relatie tot gebruik registers (Proceedings of ConGO in relation to the use of registers)*. Version 27. Internal report, Statistics Netherlands, Voorburg/Heerlen.

VIR (Civil Service Data Security Regulations) (2007) Civil Service Data Information Security Decree 2007. *Government Gazette* 122, pp 1-7.

Wallgren, A. and Wallgren, B. (2007), *Register-based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology, John Wiley & Sons, Ltd., Chichester, England.

Appendix A. List of base registers

There are now thirteen designated base registers in the Netherlands (E-overheid, 2009), The designated base registers are as follow.

1) Municipal Personal Records Database (MPRD)
   Contains the personal data of residents of the Dutch municipalities. The MPRD came into operation as a base register on 1 April 2007. The MPRD and the Non-residents' Register jointly form the base register for individuals.

2) New Trade Register (NTR)
   Will contain the data of all companies, enterprises, legal entities and self-employed people in agriculture, professional partnerships, independent professionals, associations of owners and the entire public sector, including ministries, municipalities and independent administrative bodies, in the Netherlands. The NTR Act came into force on 1 July 2008.

3) Base Register of Buildings (BRB)
   The register of all buildings, residential premises, pitches and berths, including identifying and descriptive features, such as year of construction and geometry. The BRB works together with the Base Register for Addresses under the name Base Register for Addresses and Buildings (BRAB).

4) Base Register for Addresses (BRA)
   Contains all addresses of addressable premises (residential premises, houseboat and other berths, and caravan and other pitches) in the Netherlands. The BRA and the Base Register for Buildings work together under the name Base Register for Addresses and Buildings (BRAB).

5) Topography Register (TR)
   The register that holds the digitized data of geographical features (e.g. RoadSection, WaterSection, RailwaySection, Area) to a scale of 1:10,000. The TR works closely with the Cadastral Register. The Act incorporating the BRT came into force on 1 January 2008. Country-wide topographical databases to a scale less than 1:10 000 is available from 1 January 2010.

6) Cadastral Register (CR)
   The register of each parcel in the Netherlands to a scale of 1:10,000. A parcel may be a building, but could also be a field of corn. The CR works closely with the Topography Register. The Act incorporating the CR came into force on 1 January 2008. Country-wide databases to a scale less than 1:10 000 is available from 1 January 2010.

7) Vehicle Register (VR)

Contains the registration particulars, data about the vehicle and the personal data of the vehicle holder. Is also known as the Registration Number Registers. The Act incorporating the VR came into force on 1 July 2008.

8) Base Register for Wages, Labour Relations and Benefits (BRWLB)

Contains data about wages, social benefits and employment contracts. Will be based on part of the existing Benefit Entitlement Database.

9) Real Estate Valuation Base Register

Contains the value and importance of premises subject to the 'Valuation of Immovable Property' (in Dutch: 'WOZ'). These are inhabited or uninhabited properties or parts thereof. The legislative proposal for this base register, together with that for the Base Income Register was adopted by parliament in 2008.

10) Base Income Register (BIR)

Contains the income data of individuals. The legislative proposal for the BIR, together with that for the Real Estate Valuation Base Register was adopted by parliament in 2008.

11) Non-residents' Base Register (NRBR)

Contains the data of individuals who are not resident in the Netherlands but have some relationship with the country, e.g. temporary employment in the Netherlands. At the time of registration the nonresident (subject to satisfactory proof of identity) will be assigned a Citizens Service Number (in Dutch: 'BSN'). The NRBR and the PR jointly form the base register for individuals.

12) Base Register for Large Scale Topography

Will contain topographical data (to a scale of between approximately 1:100 and 1:5000) of important topographical features (e.g. buildings, roads, bridges, rail lines, watercourses, dykes and high tension pylons) in the Netherlands. A standard background map is also envisaged.

13) Subsurface Base Register (SBR)

This register contains data about surveys and rights of use of spatial (geological) and infrastructure subsurface (soil) features. A geological model and a soil map is drawn up for the entire Netherlands.

More information is available on the base register website (link: http://www.e-overheid.nl/onderwerpen/stelsel-van-basisregistraties/basisregistraties).

Appendix B. List of quality indicators for secondary sources

Table B1. Dimensions, quality indicators and measurement methods for Source

DIMENSIONS	QUALITY INDICATORS	MEASUREMENT METHODS
1. Supplier	1.1. Contact	- Name of data source
		- Data Source Holder (DSH) contact info
		- National Statistical Institute (NSI) contact info
	1.2. Purpose	- Reason for use of data source by DSH
2. Relevance	2.1. Usefulness	- Importance of source for NSI
	2.2. Possible use	- Potential statistical use of data source
	2.3. Information demand	- Does the data source satisfy information demand?
	2.4. Response burden	- Effect of data source on response burden
3. Privacy and security		
	3.1. Legal provision	- Basis for existence of data source
	3.2. Confidentiality	- Does the Personal Data Protection Act apply?
		- Has use of data source been reported by NSI?
	3.3. Security	- Manner in which the data source is send to NSI
		- Are security measures required (hard-/software)
4. Delivery	4.1. Costs	- Costs of using the data source
	4.2. Agreements	- Are the terms of delivery documented?
		- Frequency of deliveries
	4.3. Punctuality	- How punctual can the data source be delivered?
		- Rate at which exceptions are reported
		- Rate at which data is stored by DSH
	4.4. Format	- Formats in which the data can be delivered
	4.5. Selection	- What data can be delivered?
		- Does this comply with the requirements of the NSI
5. Procedures	5.1. Data collection	- Familiarity with the way the data is collected
	5.2. Planned changes	- Familiarity with planned changes of data source
		- Ways to communicate changes to NSI
	5.3. Feedback	- Contact DSH in case of trouble?
		- In which cases and why?
	5.4. Fallback scenario	- Dependency risk of NSI
		- Emergency ,measures when data source is not delivered according to arrangements made

Table B2. Dimensions, quality indicators and measurement methods for Metadata

DIMENSIONS	QUALITY INDICATORS	MEASUREMENT METHODS
1. Clarity	1.1 Population unit definition	- Clarity score of the definition
	1.2 Classification variable definition	- Clarity score of the definition
	1.3 Count (numeric) variable definition	- Clarity score of the definition
	1.4 Time dimensions	- Clarity score of the definition
	1.5 Definition changes	- Familiarity with occurred changes
2. Comparability	2.1 Population unit definition comparison	- Comparability with NSI-definition
	2.2 Classification variable definition comparison	- Comparability with NSI-definition
	2.3 Count (numeric) variable Definition comparison	- Comparability with NSI-definition
	2.4 Time differences	- Comparability with NSI reporting periods
3. Unique keys	3.1 Identifying keys	- Presence of unique keys - Comparability with unique keys used by NSI
	3.2 Unique combinations	- Presence of useful combinations of variables
4. Data treatment (by DSH)	4.1 Checks	- Population unit checks performed - Variable checks performed - Combinations of variables checked - Extreme value checks performed
	4.2 Modifications	- Familiarity with data modifications - Are modified values marked and how? - Familiarity with default values used

Table B3. Dimensions, quality indicators and measurement methods for Data

DIMENSIONS	QUALITY INDICATORS	MEASUREMENT METHODS
1. Technical checks		
	1.1 Readability	- Accessability of the file and data in the file
	1.2 File declaration compliance	- Compliance of the data in the file to the metadata agreements
	1.3 Convertability	- Conversion of the file to the NSI-standard
2. Accuracy		
Objects	2.1 Authenticity	- Legitimacy of objects
	2.2 Inconsistent objects	- Extent of erroneous objects in source
	2.3 Dubious objects	- Presence of untrustworthy objects
Variables	2.4 Measurement error	- Deviation of actual data value from ideal error-free measurements
	2.5 Inconsistent values	- Extent of inconsistent of combinations of variable values
	2.6 Dubious values	- Presence of implausible values or combinations of values for variables
3. Completeness		
Objects	3.1 Under coverage	- Absence of target objects (missing objects) in the source
	3.2 Over coverage	- Presence of non-target objects in the source
	3.3 Selectivity	- Statistical coverage and representativeness of objects
	3.4 Redundancy	- Presence of multiple registrations of objects
Variables	3.5 Missing values	- Absence of values for (key) variables
	3.6 Imputed values	- Presence of values resulting from imputation
4. Time-related dimension		
	4.1 Timeliness	- Time between the end of the reference period and receipt of the source
	4.2 Punctuality	- Time lag between the actual and agreed delivery date
	4.3 Overall time lag	- Overall time difference between the end of the reference period and the moment the NSI has concluded that the source can be used
	4.4 Delay	- Extent of delays in registration
Objects	4.5 Dynamics of objects	- Changes in the population of objects (new and dead objects) over time
Variables	4.6 Stability of variables	- Changes of variables or values over time
5. Integrability		
Objects	5.1 Comparability of objects	- Similarity of objects in source -at the proper level of detail- with the objects used by the NSI
	5.2 Alignment of objects	- Linking-ability (align-ability) of objects in source with those of NSI
Variables	5.3 Linking variable	- Usefulness of linking variables (keys) in source
	5.4 Comparability of variables	- Proximity (closeness) of variables

Appendix C. List of image-determining statistics (2008)

Name of statistic (abbreviation)	Secondary sources used
Population statistics	- Municipal Personal Records Database (MPRD)
Social security and Reintegration by municipalities	- Social security benefits
Consumer price index (CPI)	- Scanner data from supermarkets
Quarterly economic growth	- Indirectly dependent (via STS, ITG, LFS, quarterly employment and CPI)
Labour Force Survey (LFS)	- Municipal Personal Records Database, Centre for Work and Income, Tax Administration sources, Benefit Entitlement Database
International Trade in Goods (ITG)	- VAT data
Income Panel Survey and Regional Income Survey	- Tax Administration sources, Student grants and loans
Short-term statistics (STS)	- VAT data
Government deficit and debt, EMU criteria	- Benefit Entitlement Database
Producer prices	- Independent of secondary sources
Structural Business Statistics (SBS)	- VAT data, corporate income tax
Social Statistical Database (SSD), satellite Integration	- Indirectly dependent (via SSD core)
Safety Monitor	- Municipal Personal Records Database, Tax Administration sources
Apportionment criteria	- Municipal Personal Records Database, Address Coordinates Netherlands
Ministry of Social Affairs and Employment/ Ministry of the Interior and Kingdom Relations	- Top10Vector database, Topography Register
Quarterly employment	- Benefit Entitlement Database
Monthly unemployment	- Indirectly dependent (via LFS).

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Secundaire waarneming				
1.0	23-04-2009	First Dutch version	Piet Daas Judit Arends-Tóth	Ivo Beuken Dirkjan Beukenhorst Pim Geomini Jos Jacobs
English version: Secondary data collection				
1.1E	16 -02-2012	First English version	Piet Daas Judit Arends-Tóth	