Micro Integration

10

Bart Bakker

Statistical Methods (201108)

Statistics Netherlands

The Hague/Heerlen

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

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

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

60165201108 X-37

Table of Contents

1. Introduction to the theme of Micro integration

1.1 General description

Micro integration involves matching data from statistical units at an individual level, with the goal of compiling better information than is possible when using the separate sources. A distinction should be made between micro integration and the enrichment of individual data. No added value is provided by individual data enrichment, as only additional information is added to the individual units (Al & Thijssen, 2003).

Data from registers and surveys used for Statistics Netherlands statistics can contain a variety of errors, such as representation and measurement errors. The objective of micro integration is to improve the quality of the statistical outcomes that originate from various sources by detecting and correcting the representation and measurement errors, such that:

- the validity and reliability of the statistical outcomes is sufficient;
- only one figure is published about a phenomenon (the so-called ‘single figure concept’);
- data from different sources can be combined and correct cross-theme outcomes are published;
- correct longitudinal outcomes are published.

Representation errors are errors that involve an incomplete and/or selective description of the population about which statements are being made. There may be overcoverage or undercoverage of the population. *Completion* is used to detect and correct these representation errors. Apart from that, measurement errors can also arise in variables used for deriving elements that together form the population. To correct these errors, other micro integration techniques in addition to completion are used.

Measurement errors are errors that incorrectly describe the characteristics of the population. There can be various causes of such errors. By using information from different sources, bias in one or multiple sources can be detected and corrected. *Harmonisation* is utilised to correct measurement errors at a conceptual level, while *adjustment for measurement errors* is used to correct measurement errors at data level.

Identifying representation and measurement errors involves looking for inconsistencies. To detect representation errors, data from a source or combination of sources is ideally compared with a reference file that includes a complete list of the elements from the target population. If these do not correspond, then there is a situation of undercoverage and/or overcoverage. In many cases, however, such reference data set is not present. In this context, reference data must then be compiled during the micro integration process. Measurement errors come to light if there are inconsistencies between the scores on the variables from the combined sources.

Consistency is a quality aspect that can be considered separately from bias or reliability. A statistical outcome gains in quality purely due to the fact that it is consistent (or made consistent) with other statistical outcomes, such that a coherent description of subjects can be provided.

A method to resolve inconsistencies will be sought only if their resolution is important for the data to be published. Micro integration is usually used in a situation where different administrative sources are available for the description of the same theme or the same source for different reporting periods. These are mostly more or less comprehensive; in other words, the entire population of the administration is included therein. However, it is not essential to limit micro integration to comprehensive sources. Survey data and registers can also be integrated at micro level and, once in a while, even data from two surveys with a sufficient overlap of respondents. However, there is little experience with the correction of representation and measurement errors in the integration of survey data and registers. We therefore only use practical examples from situations involving registers.

1.2 Scope and relationship with other themes

Micro integration is distinct from editing (the theme of ‘Data editing: detection and correction of errors’), because the decision rules defined to correct various types of errors extend to various sources. In editing, the decision rules are applied within a single source.

Micro integration is different from the theme of ‘Macro integration’, because in micro integration, the data is made consistent at a micro level, while this is not the case in macro integration.

There has been a long debate over the question of whether micro integration is actually a method or not. To a significant extent, micro integration cannot be generalised, in the sense that it requires knowledge of the sources and the actual situation in society. As a result, specific micro integration rules are continually drawn up and applied. This can be seen as small steps in the total data processing procedure. The integration rules can be classified in several distinct categories, as described above. It was decided to describe these categories and include them in the Methods Series.

1.3 Place in the statistical process

The point of departure for micro integration’s place in the statistical process is the processes of the Statistics Netherlands-wide observation strategy and the Social Statistical Database (SSD, *Sociaal Statistisch Bestand*). In general, a distinction is made between the following steps in the process:

- Specifying the output based on the information needs.
- Collecting register data to cover the information needs.
- Constructing the sampling frame and sampling selection.
- Conducting additional surveys.
- Assign matching (linking) keys to the records in the registers and surveys.
- Completion: correcting for undercoverage or overcoverage of the target population.
- Harmonisation and adjustment for measurement errors.
- Weighting, consistently repeated weighting.
- Creating tables, aggregating, publishing.

The first two points do not require statistical methods, but rather good account management. The construction of the sampling frame and sampling selection fall under ‘Sampling theory’. Conducting additional surveys falls under the theme of ‘Data collection strategy’. Assigning matching keys falls under the theme of ‘Matching’. ‘Weighting, consistently repeated weighting’ falls under ‘Sampling theory’. The theme of ‘Data visualisation’ is available for the final point.

There are two components that remain:

- Completing: correction for undercoverage or overcoverage of the target population.
- Harmonising and adjusting other measurement errors.

These two components are addressed in this theme ‘Micro integration’.

1.4 Definitions

Concept	Descriptions
SSB	Social Statistical Database (Sociaal Statistisch Bestand)
Micro integration	A method where data from statistical units is matched at individual level with the goal of compiling better information than is possible when using the separate sources. The quality relates to validity, reliability and consistency (cross-theme and longitudinal). Micro integration techniques are: completing, harmonising and correcting for measurement errors.
Completing	Using a system of decision rules, correcting for the undercoverage and overcoverage of the target population of a study.
Harmonising	Using a system of decision rules, bringing data under a single common denominator. This denominator is the concept that you want to measure.
Correcting for measurement errors	Using a system of decision rules, resolving inconsistencies in data to improve the quality.

2. Completion

2.1 Representation errors

A target population is determined at the start of a study. This target population is the population about which data is collected and outcomes are presented in the survey. The extent to which the population studied deviates from the target population is called the total representation error. In the case of a study based on administrative data sources, the total representation error can consist of the following components:

- Undercoverage and overcoverage of the target population because the target population of the administrative data deviates from the target population of the statistical study.
- Undercoverage because not all elements from the target population of the administrative data are included in the administrative records.
- Overcoverage because elements that are not part of the target population of the administrative data are (still) present in the data.
- Undercoverage caused by the inability to match elements from the administrative data that are part of the target population.
- Overcoverage caused by the match/mismatch of elements from the administrative data that are not part of the target population.

Several examples will illustrate the above.

Example 1. The target population of the study concerns students in higher education that are part of the Dutch population on 1 October 2008. The most appropriate source to define the population is the so-called Single Figure File for Higher Education (*Eéncijferbestand Hoger Onderwijs*). However, this relates to subsidised education in the in the Netherlands. Consequently, students that study abroad (mainly in Belgium or Germany) and students that follow private higher education are not counted. This means that there is undercoverage of the target population. The file also includes students who live abroad and study in the Netherlands. This creates overcoverage of the target population.

Example 2. The target population of the study concerns job seekers who do not have a job at the end of September 2008. The most appropriate administrative source for this is the administrative records of the “UWV werk bedrijf”, the Dutch employment office. Because of an administrative delay, people who have since found a job are not immediately deregistered as job seekers. In practice, the larger the unemployment rate, the longer the administrative delay: when there are more clients, there is quite simply less time to catch up with the delay. In this case, this administrative delay leads to overcoverage of the target population.

Example 3. The target population of the study consists of the people who are part of the Dutch population on 1 January 2008 and who are suspected of committing a crime in 2007. The most appropriate administrative source for this is the combination of the structural count of the Municipal Personal Records Database (GBA; Gemeentelijke Basisadministratie persoonsgegevens) on 1 January 2008 and the police Identification Service System (*HerKenningsdienstSysteem - HKS*). Due to matching problems and errors, however, there is a situation of both overcoverage and undercoverage. Matches were missed because the identifying information was too limited to make a match. In addition, due to writing errors or other reasons, matches were made between the data of two different people.

The completion method corrects for the total representation error. In addition, the measurement errors arising due to mismatches are also corrected in the adjustment for measurement errors. It is sufficient that corrections are made for overcoverage and undercoverage, regardless of the underlying cause.

2.2 Correction for undercoverage

There are different methods to correct for undercoverage. We make a distinction among the following:

- Combining different sources to create a complete list of all elements from the population.
- Assigning a weight to the elements observed, so that the total target population is represented in the outcomes.
- Using a form of unit imputation so that the total target population is represented in the outcomes.

The second and third correction methods are discussed in detail in other components of the Methods Series, in Bethlehem (2007) and Israëls, Pannekoek and Schulte Nordholt (2007) respectively. For this reason, this report discusses only the form of completion where different sources are combined.

The correction for undercoverage starts by defining the target population as precisely as possible. For example, in the first example of undercoverage, the target population can be defined as: “students following higher education that live in the Netherlands”. The undercoverage that results from the use of the Single Figure File for Higher Education can be partially completed by utilising the information from another source: the registration of the students that receive student grants and loans for a higher education programme. Student grants and loans can also be provided for private education and education followed in Belgium or Germany. It is true that not all students missing from the Single Figure File for Higher Education will be found in this way, but a large part of them will.

In the case of the people suspected of crimes, another method is used to reduce the undercoverage. It is known that there are missed matches due to people moving house. This is in addition to the fact that the people suspected of crimes have an interest in not being registered under their correct personal details. This often leads, at least temporarily, to an inability to match their data. Here also, this involves an administrative delay: in many cases, after a period of time, the correct personal details will be included in the registration. With retroactive effect, the older details from the same suspect will then be matched with this. By using the most recent volume of the police HKS, the older volumes can be completed.

However, it is not always possible to correct for undercoverage. After all, administrative or other data must be present to perform this type of correction. This data is not always available. However, it is possible to approximate the scope of the undercoverage by estimating the total population using a survey. This serves to give an impression of the error caused by the undercoverage. This can then be added to the data as a quality characteristic.

2.3 Correction for overcoverage

Overcoverage of the statistical target population is corrected by eliminating the elements that do not belong to the population. To do this, these elements must first be identified. An exact operational definition of the target population is therefore essential. An example of this is: the people included in the 2008 UWV file (version of 1 April 2010, which is corrected for the administrative delay to 2009 inclusive) and who have the value ‘yes’ for the variable SEARCHINGFORWORK on 1 October 2008. A distinction can be made between the following situations in this context:

- The definition of the target population can be operationalised in the volumes of a single source. In that case, this action is relatively simple. In the above case, the administrative delay that caused the overcoverage was deleted from the version of 1 April 2010.
- The definition of the target population cannot be operationalised in a single source, and another source is needed to delete the elements that cause the overcoverage. The overcoverage caused by the delayed deregistration from the UWV of people who have already found a job is corrected by matching the UWV file at an individual level with a file with data about employees. The elements that are included both in the employee file and the UWV file have already found work and should be deleted. The administrative delay in the UWV file is corrected for in this way.

However, it is not always possible to correct for overcoverage. This is, for example, the case for overcoverage caused by mismatches. The number of mismatches can probably be estimated (see Arts, Bakker and Van Lith, 2000, for example), but exactly how much overcoverage this leads to is not established in this way. After all, part of the mismatches may still actually belong to the target population. Because, in such a case, details from two different people have been matched, this leads to errors that are comparable with measurement errors. Another disadvantage is that the elements that cause the overcoverage cannot be identified.

2.4 Applicability

The correct use of the method requires knowledge of the area to be studied and the sources used in the study. The method can be used by formulating and programming decision rules.

Identifying undercoverage is no easy task. Only if there is a full frame that can serve as a reference file, it is possible to determine whether there is total coverage of the target population. However, this type of reference file is usually not present. In many cases, this file is compiled while completion is taking place. Additional research is needed for the check of whether the method of completion has actually resulted in coverage of the target population. For example, a focused survey can be used for this purpose.

Overcoverage is also not always easy to trace. In many cases, overcoverage due to an administrative delay can be corrected over time or by using another source. However, this is not the case for overcoverage caused by mismatches.

If it has been demonstrated or if there is a strong suspicion of unacceptable amounts of undercoverage or overcoverage, it can be decided not to use this data for the target population in question. Standards are difficult to provide for this purpose. What is acceptable for one study may not be acceptable for another.

3. Harmonisation and adjustment for measurement errors

3.1 Measurement errors

Measurement errors occur when characteristics of the population are not described correctly. We call this *bias*. Many different types of measurement errors occur in registers and surveys. Groves et al. (2004) divide measurement errors in surveys into three groups: errors made in operationalising concepts, errors made in the observation in the survey, and errors made in the processing of the survey. The errors in administrative data correspond for the most part to the errors in the surveys, because in many cases one or more survey techniques are used to fill administrative databases. For example, electronic tax or other forms can be considered as a form of Computer Assisted Web Interviewing (CAWI) or Paper and Pencil Interviewing (PAPI), the intake interview at the UWV as a form of Computer Assisted Personal Interviewing (CAPI), etc. (Bakker, Linder and Van Roon, 2008).

However, several important differences can also be identified. The first difference is that the scope of the errors depends on such things as the importance that the register holder places on the quality of the information. If the information is essential for the proper performance of the register holder’s core tasks, then this person will check the information more carefully than if the information were not important for this purpose. It is known, for example, that the start and end dates of jobs derived from personal tax returns are not of high quality. The reason for this is that the level of the tax amount to be paid is not dependent on this, and the Tax Administration also does not check this. An example of a situation where information is properly checked is for the information from Education Number Files (*Onderwijsnummerbestanden*) used to pay for education, because an audit is performed on these files. We can suitably assume that the quality after the audit is better than before: schools have an interest in a large number of students, while the auditors of the Ministry of Education, Culture and Science have an interest in a smaller number of students. The healthy tension between these two parties serves to effectively monitor the limits of the definitions and the application thereof (Bakker, 2003, pages 129-130).

A second difference is that a distinction must be made between the measurement errors that are made in the operationalisation and the measurement of the administrative concept and the extent to which these administrative concepts can be translated into statistical concepts. In the conceptualisation and operationalisation of the statistical concepts, this administrative information must often be converted from one situation to another. Errors can also be made in this process.

A third difference is the administrative delay. This is a measurement error that is specific for administrative data: an event is registered later – sometimes much later – than it took place. As a result, these events are not counted at a reference time that occurs before the time of registration. However, this does not always have to be problematic for the statistical outcomes. If the scope of the events and the extent of the administrative delay in a source are constant over time, then the administrative data that is published on this basis will be correct.

Measurement errors come to light if there are inconsistencies between the variables from the combined sources. There are inconsistencies, for example, in the characteristics of the units if:

- Variables from two sources describe the same event, but have a different outcome at the individual level. An example is that one source says that a person's gender is male, while another source says this same person's gender is female.
- There is a logical relationship between variables which, when used, produce an incorrect outcome. An example is that a person's annual wages are not equal to the wages that the person earned in the 12 months of the year in question.
- Levels and the associated flows do not reconcile: the level at reference time t plus the mutations between t and t+1 must produce the level at t+1. There is an inconsistency if the level of the population on 1 January 2007 plus all births and immigration less all deaths and emigration does not produce the level of the population on 1 January 2008.
- There is an impossible transition from one situation to another. For example, a transition from married to unmarried. This is impossible: there can only be a transition from married to divorced or widowed.
- There is an improbable concurrence of situations. For example, that a person has two full-time jobs at the same time for a longer period, or the person has a full-time job and receives full unemployment benefits at the same time.

Consistency is a quality aspect that can be viewed separately from bias or reliability. A statistical outcome gains in quality if it is consistent or made consistent with other statistical outcomes, so that a coherent description of subjects can be provided.

A special case of consistency is longitudinal consistency. By this, we mean that the data over a period of time is consistent or made consistent such that the scope of the mutations or changes is properly estimated for the populations or subpopulations. Longitudinal consistency is harmed, for example, by administrative delays. An example of this is that marriages of foreign nationals (residing in the Netherlands) concluded in their home country are sometimes registered up to two years after the marriage. If this administrative delay is not processed, the mutations in the marital status are underestimated for this category.

To correct measurement errors at a conceptual level, *harmonisation* is used. To correct measurement errors at data level, *adjustment for measurement errors* is employed.

3.2 Harmonisation

Research begins with the question of which concepts must be measured. These are initially defined conceptually. Examples can be derived from the concepts from the Bureau of Standards (Statistics Netherlands, 2009). The concepts must then be operationalised. This operationalisation consists of writing down the exact derivation of the concept from the information present in the data source. These derivation rules boil down to stating the precise criteria to measure the concept. In surveys, the operationalisation is found in how the questions are phrased, while in administrative records, there is much less freedom of definition: the quality of the operationalisation depends on the possibilities offered by the information in the register. Harmonisation involves information from one or several sources being brought under a single common denominator.

Example 1. The concept of “Job” is defined as follows: “A labour position filled by an employed person. An employed person can have several jobs at the same time. In that case, there is a main job and a side job” (StatLine, 2009). This definition is about the relationship between the company that is offering the position and the person that holds this position. The second half of the definition addresses the possibility of having several jobs at once, but that is not important here. The definition is operationalised by searching in the appropriate administrative data for combinations of people (identified by an anonymised personal identification number, the ‘RIN number’) and companies that offer the positions (identified by a business identification number, the so-called BE_ID). Each record where such a combination occurs represents a job with a start and end date (Arts and Hoogteijling, 2002; Bakker & Arts, 2003).

In the Register of Insured Persons (*Verzekerdenadministratie*) of the Dutch social security benefits administration (*Uitvoeringsinstituut WerknemersVerzekeringen* – UWV), companies are registered under the so-called BVA number, which stands for the Industrial Insurance Board & Registration Number (*Bedrijfsvereniging & Aansluitnummer*), and the employees are registered in this under their citizen’s identification number (*Burgerservicenummer* – BSN). After the BSNs have been converted into RIN numbers for security reasons, the BVA codes are converted to BE_ID codes. In the FiBase, a file with data from the companies for statutory payroll tax, the companies are registered under a payroll tax number and the employees under a BSN. By also converting the payroll tax numbers into BE_ID codes, the data is harmonised and the jobs in these two sources have the same meaning. The same is done for information from the Employment and Wages Survey (*Enquête Werkgelegenheid en Lonen*), with the understanding that this survey is comprehensive for part of the companies, and it only involves a sample of the other part. For the sampled part, the sample is taken from the General Business Register (ABR) in which the business units are identified by their BE_ID.

Since 1 January 2006, the abovementioned registrations have, for the most part, been combined in the so-called Policy Administration, and it is sufficient that the above rules are applied to records from this administration.

Example 2. “Educational attainment” is defined as “The level according to the Standard Classification of Education (*Standaard onderwijsindeling* - SOI) of the highest educational programme successfully followed”. For the application of the SOI, a list of educational programmes has been developed where each distinct programme has been assigned an ‘educational program number’. The SOI can be derived from these educational program numbers. To operationalise the educational attainment using administrative data, different sources are combined, both registers and surveys. Each of these sources contains mainly unique information about the educational programmes that people have followed during their life course. Each of these sources also has their own system of codes that records what educational program was followed and possibly completed. By making a conversion table to convert each of these code systems into the SOI educational program numbers, this information is harmonised.

The major part of harmonisation consists of formulating a number of decision rules that indicate as precisely as possible how a certain concept is measured, given the available information in the data. To do this, knowledge is needed about the scientific and social meaning of the concept as well as knowledge of the information that is available in the source or sources.

3.3 Adjustment for other measurement errors

First, harmonisation is used to resolve inconsistencies. If differences still remain, an assessment is made as to the best source for each of the variables. In determining the best source, account should be taken of the administrative practice. In particular, if a variable is not very important for a reporter (such as the start and end date of a job in tax data), the quality is often doubtful. A variable in a source can be strong on some points and weak on others. For example: the annual wages of people working for the National Government in source A is not as good as the same piece of data in source B, but for the other sectors, the annual wages in source A are of better quality. By giving preference to a certain source (possibly under one or multiple conditions), the annual wages can be determined (Bakker, 2003).

If the quality of the sources is not sufficiently known, then a new variable can also be determined based on two or more sources. An example of this is taking the average value of two sources. In addition, decision rules can be formulated to force a relationship between different variables into being correct. Here, which information is adapted depends on the quality of the sources.

3.4 Applicability

Both harmonisation and the adjustment of measurement errors are carried out using decision rules. To properly formulate these decision rules, knowledge is needed of the actual situation in society and the content and quality of the information from the available sources.

Identifying measurement errors is not an easy task, nor is formulating decision rules. In some cases, there will be rules to correct incorrect records, but which will also convert a small number of correct records into incorrect ones. To determine the scope of these two occurrences, additional research is needed. A focused survey, for example, could be used for this purpose.

4. Conclusion

It is important to document which values are corrected and which decision rules were used for this purpose. This is necessary to make the process reproducible. Such documentation is also important for the users of the resulting micro files. In harmonisation, it is also important to properly define the concepts used, to determine the decision rules and to include these in the documentation.

5. References

Al, P. and Thijssen, J. (2003), Bespiegelingen over het waarom, de mogelijkheden en beperkingen van micro-integratie in de sociale statistieken, In: J. Nobel, S. Algera, M. Biemans and P. van der Laan (eds.), *Gedacht en gemeten*. Statistics Netherlands, Voorburg/Heerlen, pp. 112-122.

Arts, K., Bakker, B.F.M. and Lith, E. van (2000), Linking administrative registers and household surveys, In: P. Al and B.F.M. Bakker (eds.) *Re-engineering Social Statistics by micro-integration of different sources*. Theme issue Netherlands Official Statistics, vol. 15, Summer, pp. 16-22.

Arts, C.H. and Hoogteijling, E.M.J. (2002), Het Sociaal Statistisch Bestand 1998 en 1999. *Sociaal Economische Maandstatistiek* 2002 (12), 66-71.

Bakker, B.F.M. (2003) Hoe nieuw zijn nieuwe ideeën?, In: J. Nobel, S. Algera, M. Biemans & P. van der Laan (eds.), *Gedacht en gemeten*. Statistics Netherlands, Voorburg/Heerlen, pp. 123-132.

Bakker, B., & Arts, K. (2003), Dynamiek op de arbeidsmarkt; gegevens over stromen uit het Sociaal Statistisch Bestand. In: B.F.M. Bakker & L. Putman (eds.), *De virtuele volkstelling en SSB.* SISWO/CBS, Amsterdam, pp. 59-70.

Bakker, B.F.M., Linder, F. and Roon, D. van (2008), Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys. In: *IAOS Conference on Reshaping Official Statistics*. International Association of Official Statistics and the National Bureau of Statistics of China, Shanghai, 14-16 October.

Bethlehem J.G. (2007), *Weighting adjustment for non-response*. Methods Series document, Statistics Netherlands, The Hague [English translation from Dutch in 2011].

Groves, R.M., Fowler F.J. Jr., Couper, M.P. Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004), *Survey Methodology*. Wiley Interscience, New York.

Israëls, A., Pannekoek, P. and Schulte Nordholt, E. (2007), *Imputation*. Methods Series document, Statistics Netherlands, The Hague [English translation from Dutch in 2011].

Statistics Netherlands (2009), http://www.cbs.nl/nl-NL/menu/methoden/begrippen/default.htm. Statistics Netherlands, The Hague/Heerlen.

StatLine (2009), http://www.cbs.nl/nl-NL/menu/methoden/begrippen/default.htm?ConceptID=92. Statistics Netherlands, The Hague/Heerlen.

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Micro-integratie				
1.0	16-02-2009	First Dutch version	Bart Bakker	Piet Daas Eric Schulte Nordholt
English version: Micro integration				
1.0E	17-02-2011	First English version	Bart Bakker	