Statistica Neerlandica (2012) Vol. 66, nr. 1, pp. 41–63
doi:10.1111/j.1467-9574.2011.00508.x

Topics of statistical theory for register-based statistics and data integration†

Li-Chun Zhang*

Statistics Norway, Kongensgt. 6 Pb 8131 Dep, N-0033 Oslo, Norway

Official statistics production based on a combination of data sources, including sample survey, census and administrative registers, is becoming more and more common. Reduction of response burden, gains of production cost efficiency as well as potentials for detailed spatial-demographic and longitudinal statistics are some of the major advantages associated with the use of integrated statistical data. Data integration has always been an essential feature associated with the use of administrative register data. But survey and census data should also be integrated, so as to widen their scope and improve the quality. There are many new and difficult challenges here that are beyond the traditional topics of survey sampling and data integration. In this article we consider statistical theory for data integration on a conceptual level. In particular, we present a two-phase life-cycle model for integrated statistical microdata, which provides a framework for the various potential error sources, and outline some concepts and topics for quality assessment beyond the ideal of error-free data. A shared understanding of these issues will hopefully help us to collocate and coordinate efforts in future research and development.

Keywords and Phrases: Combination of sources, data life cycle, representation, measurement, validity, equivalence, record linkage, statistical matching, micro integration, micro calibration.

1 Introduction

1.1 A remark on the development of register-based statistics

For some decades now, alongside survey sampling and population census administrative registers have been an important data source for official statistics. Not only do they provide frames and valuable auxiliary information for sample surveys and censuses, systems of inter-linked statistical registers (i.e. registers for statistical uses) have been developed on the basis of various available administrative registers to produce a wide range of purely register-based statistics (e.g. Statistics Denmark, 1995; Statistics Finland, 2004; Wallgren and Wallgren, 2007). A summary of the development of some of the key statistical registers in the Nordic countries is given in UNECE (2007, table on page 5), the earliest of which already came into existence

*lcz@ssb.no

† Correction added on 10 August 2015, after the initial online publication. A duplicate article, with a DOI in the format 10.1111/stan.508 was deleted from EV on 30 July 2015. This DOI now aliases to: 10.1111/j.1467-9574.2011.00508.x in STAN 66:1.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.
Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

in the late 1960s and early 1970s. Statistics Denmark was the first to conduct a register-based census in 1981; the next census in 2011 will be register-based in all the Nordic countries, while the so-called virtual census (e.g. Schulte Nordholt, 2005), which combines data from registers and various sample surveys in a more 'traditional' way, will be implemented in a number of European countries.

The major advantages associated with the statistical use of administrative registers, and data integration in general, include reduction of response burden, gains of long-term production cost efficiency, and potentials for detailed spatial-demographic and longitudinal statistics. But, although the trend towards more extensive use of administrative data has long been recognized (e.g. Brackstone, 1987), there is also a noticeable lack of *statistical* theories for assessing the uncertainty of register-based statistics (e.g. Holt, 2007).

It is tempting to reflect on the historical development of survey sampling by way of comparison. The *representative method* (Kiær, 1897) was presented by N. Kiær at the ISI meeting in 1895. Although he was unable to defend the approach on theoretical grounds, the practice continued to evolve thereafter. In 1924, the ISI formed a committee to investigate the 'application of the representative method' (Jensen, 1924). Its report stated: 'When ISI discussed the matter twenty-two years ago, it was the question of the recognition of the method in principle that claimed most interest. Now it is otherwise. I think I may venture to say that nowadays there is hardly one statistician, who in principle will contest the legitimacy of the representative method. Nevertheless, I believe that the representative method is capable of being used to a much greater extent than now is the case'. Indeed, the theoretical breakthroughs did not arrive until some 30–40 years after Kiær's initial presentation. Today, the contributions by Bowley (1926) and, in particular, Neyman (1934) are generally taken as the starting point of the theoretical development of the so-called design-based approach to survey sampling.

When viewed in this historical mirror, register-based statistics currently appear to be pretty much at a pre-Neyman stage in terms of their maturity. We believe that the key issue here, from a statistical methodological point of view, is the *conceptualization* and *measurement* of the *statistical accuracy* of register data which will enable us to apply rigorous statistical concepts such as bias, variance, efficiency and consistency, as we are able to do in other branches of the statistical science.

1.2 Life cycle of integrated statistical microdata

Administrative registers certainly do not provide perfect statistical data. For various reasons, the available values may differ from the ideal measures, and the measured objects may differ from the statistical units of interest. By convention these reasons are referred to as the error sources. Groves *et al.* (2004, Figure 2.5) provided a systematic outlook to the potential error sources for sample survey data, throughout their 'life cycle' starting from conception, to collection and processing, and then to the statistics produced. Bakker (2010) proposed an adaptation to register data,

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

based on 'the general idea that it is likely that the errors that normally emerge in surveys will also occur in registers'. Expanding and developing on these initiatives, we have arrived at a two-phase life-cycle model of integrated statistical microdata, which is presented in Figure 1. The rectangles describe the various states of data along the lines of 'measurement' and 'representation' respectively (Groves *et al.*, 2004). The broad arrows indicate the flow and/or processing between two successive states. A source of error is located between any two states, corresponding to an error type (the ovals). This two-phase model is more detailed and explicit than the Bakker's model (2010). The aim is to provide a framework that allows the disen-

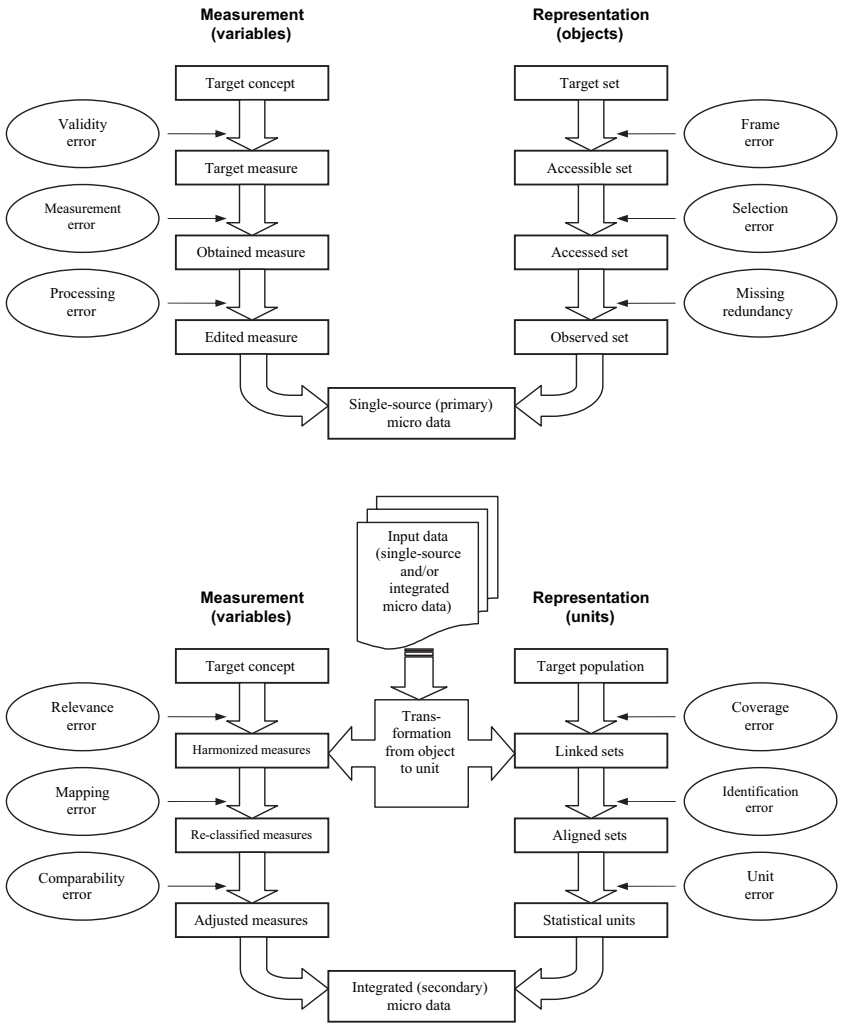

Fig. 1. Two-phase (primary and secondary) life cycle of integrated statistical micro data from a quality perspective. Data concept (square); Error Type (oval). Source of error indicated by narrow arrow; Input, flow and/or processing of data indicated by broad arrow.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

tanglement and clarification of the various types of potential errors, and their origin and detection in relation to the different production processes of integrated statistical microdata. It includes the life-cycle model of survey data (Groves *et al.*, 2004) as a special case.

Above all, we propose in general to divide the entire life cycle of integrated statistical microdata into two phases. The first phase concerns the data from each single source, the second phase concerns the possible integration of data from different sources, often involving necessary transformations of the initial input data. There are two main inter-connected reasons for this distinction, which are worth noting before we get into more details. The first one has to do with the fact that statistical usage of administrative data is *secondary* of nature, in contrast to *primary* usage of sample survey data that are designed and collected for certain defined statistical targets. Administrative registers are owned and, mostly, maintained by external register owners. The administrative data have already gone through a sequence of conception, collection and processing, albeit mainly for non-statistical purposes, before they are delivered to the statistical agency. These activities must not be confused with the work that is carried out at the statistical office in order to make the data fit for statistical purposes.

The second reason is the fact that using an administrative register for statistical purposes, or assimilating it into the statistical system at all, is often only feasible after it has been combined with data from other sources. A sample survey may suffice for a statistical purpose on its own because it was designed for this purpose. Auxiliary information can be incorporated to improve the efficiency or to reduce various non-sampling errors, but in principle it is possible to produce the statistics even without the auxiliary information. This is different in the case of administrative data. For instance, to produce statistics on the highest attained education level, one might make use of, say, a register of examination results. However, this register will not contain information on everyone in the population, so integration with the Central Population Register (CPR) will be necessary. Indeed, to control the quality, one would probably take into account, say, the register of school enrolment and any other relevant and helpful information. Note that, to integrate a register organized on the basis of individual exam results with the CPR, it must first be transformed it into a file organized on the basis of persons. A two-phase distinction is natural here because (i) one cannot always expect the register owner to carry out such necessary transformations for the statistical agency and (ii) ultimately the required transformations may differ depending on the statistical purpose, so that they cannot be set once and for all.

While data integration has always been an essential feature associated with the use of administrative registers, it should not be restricted to register data alone. Survey and census data may also be integrated – or any potential data sources for that matter – so as to extend their scope and improve the quality. Indeed, the secondary usage of sample survey and census data for purposes other than the statistics that they were initially collected for should be actively promoted. A two-phase

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

model emphasizes that there is and should be a life for these data beyond the first phase.

It should be noted that often the adjectives *single-source* and *primary* should be interpreted in *relative* terms, for the purpose of a conceptual distinction between the various input data from the perspective of second-phase data integration. For instance, data from the last census and concurrent surveys and registers may be combined to update population statistics of interest. In this context the last census may be regarded as a single data source on its own, that is, in juxtaposition with the survey and register data. However, the census data themselves may well have been produced by means of data integration in the first place, in which context they would be regarded as integrated statistical data based on multiple single-source (or primary) input data on that occasion.

1.3 Accuracy assessment

The assessment of errors basically starts out from the following set-up. On the one hand, there is the actual dataset, either at the end of the integration processes or in some intermediate stage. On the other hand, in concept there is an ideal (or preferred) dataset at the corresponding stage of processing, to be referred to as the *target* data. In many situations, though certainly not always, an *audit sample* of the target data is available. Moreover, the target dataset may be able to be linked to the actual dataset at the unit level, making it possible to obtain a probabilistic measure of the individual discrepancy between the two datasets, which then yields a basis for subsequent statistical inference. We shall refer to this as the unit-specific approach, since it is based on the observed discrepancy for specific units.

We notice immediately that the availability of an audit sample of perfect target data is often an idealization, since more often than not any obtained measures are inevitably contaminated by errors. In practice, then, one may opt for a tacit choice of the target data, as these are the best or preferred measures obtainable. Or, one might consider both datasets to be subject to errors. What characterizes the unit-specific approach is that it is not applicable if the two datasets are known to consist of entirely different units.

There are two shortcomings of the unit-specific approach. First, the required unit-level linkage may be impractical or difficult because no unique identifiers are available, or the linkage may even be prohibited by law. Second, unit-specific discrepancy measures the equality (or inequality) for a given unit. While this information is always useful for the *producer* of the data, it may not be relevant or may even be misleading for their *uses*. For example, consider a given set of units. Suppose that each unit is associated with two binary variables, which have the same sample mean, but without the values being equal for all the units. Since the two sets of variables have the same empirical distribution function, statistical inference should be identical based on either variable *alone*. Yet, the apparent unit-specific inequality may fail to reveal such statistical equivalence.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

Based on these considerations, we propose the concepts of statistical validity and equivalence between different microdatasets, as generalizations of the unit-specific equality concept. This will allow for alternative accuracy assessment approaches, both in situations where the unit-specific approach is difficult or prohibited in practice and where it is theoretically impossible.

The rest of the article is organized as follows. We elaborate the two phases of the life cycle of integrated statistical microdata in section 2. Some examples will be given of ongoing research projects and recent development of statistical theory for dealing with specific errors. In section 3 we consider some conceptual issues for accuracy assessment. We define the concepts of statistical validity and equivalence, discuss their implications in several related settings other than data integration, and outline some specific topics for research in data integration methodology. A few summarizing remarks on our general view of data integration will be given in section 4.

2 Two-phase life cycle of integrated statistical microdata

2.1 Phase one

Phase one concerns the data of one single particular source. The life cycle of sample survey data as charted by Groves *et al.* falls under this phase, except that we have excluded the stage of postsurvey adjustment (such as weighting) along the line of representation. But the concepts of Groves *et al.* (2004, Figure 2.5) have been modified to accommodate data from the administrative sources as well.

Take first the line of representation. In survey sampling the *target set* would contain the units of the target statistical population, whereas the *accessible set* would correspond to the sampling frame. The difference between the two is known as *frame* error. Notice that in multiple-frame sampling, the accessible set would be more complicated to describe, but the concept still applies. Next, the *accessed set* would correspond to the *gross* sample, and the *selection* error (i.e. the difference between the accessible and the accessed sets) is what is traditionally known as the sampling error. The *observed set* may be referred to as the *net* sample that contains only the respondents, and the difference with the accessed set (i.e. the gross sample) is known as *unit* non-response (or *missing*). By convention we shall include the units of partial non-response (or item missing data) as part of the net sample and the 'holes' in the observation data matrix are regarded as a kind of imperfection that falls under the domain of measurement.

As a general note, the ovals of error types in Figure 1 point to the sources, not where they might be detected. For instance, some frame errors may only be detected after contact has been made with the gross sample and out-of-scope units identified between the gross and net samples.

The concepts are applicable to administrative data sources as well. The difference between the target and accessible sets may be caused by feasibility or other reasons. For instance, a job register should ideally contain information on, say, all

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

compensated contractual labour–market relationships. However, only the regular sector of the labour market reports to the authority and is thus accessible at all; the 'black-market' jobs are out of the frame. (Notice here that the 'units' of a job register are in fact the various job-related events, such as recruitment, dismissal, leave, etc., while labour-market statistics usually have persons as the statistical units. So in order to make use of the job register, information by job events must be reorganized as information by persons, just as in the case of the examination register earlier. This is the reason that we choose to use the term *objects* for the units in phase one, to potentially distinguish these from the more familiar term *units* for statistical units in phase two.) Next, the accessed set would contain all job-related events that are actually registered. Inevitably, some selection errors are associated with the reporting/recording process: some events will fail to be reported, or they may be reported with a delay, while some of the reported events may be inadmissible due to confusions surrounding the administrative regulations. Lastly, the observed set contains the objects that remain after a validation process by the register owner. Inadmissible objects may be detected at this stage, which nevertheless should be attributed to the selection error. But some objects may be rejected at this stage due to *missing* information, while *redundancy* may remain or arise by mistake. Although the difference between the observed and accessed sets may be small here, the conceptual distinction between an object in a state as recorded and that after processing seems justified in general.

Now let us take the line of measurement, which is a one-to-one mapping of the corresponding line in Figure 2.5 of Groves *et al.* (2004) for sample survey data. The corresponding terms there for the rectangles are 'Construct', 'Measurement', 'Response' and 'Edited response'. 'Construct' is the ideal information that is sought about an object. It may be abstract and sometimes ambiguous, such as political orientation. 'Measurements' are necessarily concrete and observable, such as the party voted for in the last election. The task is to design and choose the measurements to capture as closely as possible the construct (i.e. the ideal information). The overlap (or gap) between the two is then the *validity* (or invalidity). 'Response' is then the obtained measurement and 'Edited response' the corresponding value after editing and imputation. Here, imputation may refer to any changes made to the response, not just when a value is missing. The errors arising between the intended measurement and the obtained response are referred to as the *measurement* errors, whereas those between the response and edited response are the *processing* errors.

In Figure 1 these concepts have been modified to accommodate administrative data as well. Instead of 'Construct', we use the term *target concept*, partly to observe the parallel to 'target population' along the line of representation. Moreover, for administrative registers, the target information is seldom abstract or ambiguous along the line of measurement. Otherwise, it would simply fail to serve the administrative purposes. Take the job register again. What has to be reported *about* a job event is clearly defined, such as the social security number of the employee, the dates, etc., because the information is meant for administrative purposes, rather than

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

for describing or analysing certain social-economic phenomena. The idea of a theoretical construct for the measurement seems remote. Next, we use the term *target measure* instead of 'Measurement', and *obtained measure* instead of 'Response', and *edited measure* instead of 'Edited response'. In this way, the term measure designates the relevant values, whereas the term measurement designates the process of obtaining, or the collecting of, these values. We do not use the word 'Response', because many administrative registers are based on routine registration of events, and do not involve respondents as in the sense of sample surveys.

2.2 Phase two

Integration of data from multiple sources takes place in phase two. Here the *target population* and *target concept* are defined according to the statistics of interest, which usually differ from the respective primary target population and concept of each input data source. The target population is the set of statistical units that the statistics should cover, and the target concept is the information content of the statistics.

Now, very often a transformation from phase-one objects to phase-two units must be carried out as the first stage of processing in data integration, as a result of which information contained in the input data may be reorganized along the lines of representation and measurement. The term object was introduced in phase one to contrast with the term unit because many administrative registers are initially based on events. The examinations register and job register have been mentioned. The CPR consists of events such as birth, death, marriage, divorce, etc.; the VAT register is built on transactions, rather than establishments or enterprises, etc. In all these cases, the input data organized on the basis of primary objects must first be transformed to data organized on the basis of units suitable for integration.

It is important to realize that the transformation may affect the information flow along the lines of representation and measurement. Consider as an example the registration of a delayed job event. Suppose, for example, that because of the transformation a person is classified as 'Employed' if he/she has an active job according to the job register, that is, there is an event of recruitment but no event of dismissal from the same job. Suppose someone has been dismissed, but the dismissal event somehow had not entered the register at the moment of production (or data integration). For the job register, the event of dismissal is an object along the line of representation, which belongs to the accessible set but not to the accessed set, hence constituting a selection error. In phase two, after the transformation from event to person, this person would be included along the line of representation, but with a misclassified employment status. In other words, the phase-one selection error may cause an error with regard to the target measure (i.e. employment status) along the line of measurement in phase two. This is thus also an example for the necessity of the two-phase division in Figure 1, without which it seems difficult to clarify the situation where an initial error along the line of representation may nevertheless be dealt with statistically as an error along the line of measurement in phase two.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

Now take the line of representation. The units that result from the transformation process are not necessarily the target statistical units, unless these are readily available in all the input data. Generally speaking, the first thing to be accomplished is the *linked sets* (of units) across the relevant data sources. Consider the example where the statistical unit is the household. In most European countries there is no complete frame of households, in which case it is impossible to arrive at households directly after the transformation. Instead, they need to be created or validated at a later stage. One unit that can be used to link the relevant input data is the person, so various data sources containing information about family relationship, dwelling address, or even tax returns first need to be transformed to persons. Additional linkage may be possible, for example, using addresses to obtain housing information from the dwelling register. The potential differences between the target population and the set of statistical units covered by the linked data constitute the (over- and under-) *coverage* errors. The coverage error may occur through data linkage itself, if the linkage procedure is subject to error. But coverage errors may also be caused by the information associated with the linked units. For instance, the target population of household statistics should usually contain all people residing in private dwellings. Suppose the residence status comes from the CPR, whereas the private/institutional status of a residence comes from the dwelling register. In that case, any errors in either source or incompatibility between them may give rise to coverage errors of the target household population, despite the fact that households have not yet been created. Notice that the source of such coverage errors is not data linkage itself, but an earlier stage in phase one. However, in the statistical process it is possible to regard them as coverage errors, just as the coverage errors that do originate in the stage of data linkage. We shall come back to this later in connection with unit errors.

The purpose of *alignment* is to clarify all the relevant relationships between the different units in the linked data for the creation (or validation) of statistical units. Table 1 provides a generic representation of the *aligned sets* (of units). Each alignment table contains one type of *base units* and one or several types of *composite units*. The base units are the atomic building blocks of all the composite units in the same alignment table, in that there is potentially a many-to-one relationship between a number of base units and one composite unit, as for example in Table 1, where the no. 1 type-1 composite unit consists of base unit nos. 1 and 2, while base unit nos. 2 and 3 belong to the no. 2 type-D composite unit. There may also be no direct mapping between two types of composite units: for example, if the base units are persons, families are one type of composite units and dwellings another. Then, there may be no direct mapping between the two types of composite units, for example, in the case of a family of two parents and their adult son, where the son lives in a different dwelling with a cohabitant.

Since different data sources may contain conflicting information about the relationships between the base units and any type of composite units, the alignment table based on the linked datasets may differ from that derived from any of the input data sources alone. The dwelling information for students in Norway provides

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

Table 1. Alignment table (i.e. of the aligned sets of units)

Base unit	Composite unit type-1	...	Composite unit type-D
1	1		1
2	1		2
3	2		2
...			

an example. Every student has a registered dwelling in the CPR. But this may be their parents' home address, or their own actual dwelling at the place of study, as the registration rule is not strict in this respect. Suppose now that the CPR is linked to the official postal address register, and it turns out that a person (i.e. a base unit) is associated with one dwelling (i.e. a composite unit) in the CPR, and a different address in the address register; in this case it is the task of alignment to resolve the conflict. Subsequently, suppose there is a third school enrolment register, which shows that the person concerned is a student at a place different from the CPR address, but in accordance with the postal address. It would then seem justified to choose the dwelling at the postal address as the correct dwelling for that person. Clearly, alignment in itself may be no simple task in such situations, and we shall refer to errors in the alignment table regarding relationships between base units and composite units as *identification* errors.

Apart from possible mistakes in the alignment of the different type of units, difficulties may arise in the classification of the composite units. A typical example is the industry code of various business units. For instance, an enterprise (i.e. a composite unit) may consist of several legal units (i.e. base units), which have different industry codes. In such cases it is common practice to assign an industry code for the activity that is in some sense judged to be the most important or the most relevant. As a result of this, however, the totals aggregated from the enterprises by industry code may disagree with those aggregated from the legal units directly, which can raise problems in many circumstances. Of course, the reason, as shown in the alignment table, is that a composite unit really permits only *partial classification*, as long as its base units may fall in different categories according to the same classification. One may consider the problem here to be a special case of identification error, by regarding the industry as an additional type of composite unit. The practice of assigning a unique industry code to an enterprise then amounts to imposing a hierarchical relationship between two types of composite units (i.e. enterprise and industry) where such a relationship does not in fact exist.

Having clarified the alignment between the different types of units, *statistical units* may need to be created or validated. Again, the issue usually arises from the nature of secondary usage of administrative data, which are initially collected for different purposes. The statistical units of interest may simply not exist in any available data source, and need to be created by the statistician. We refer to the inevitable errors as the *unit* error. Household provides a typical example in social statistics. But the same problem exists in business statistics. Two points about the unit error are worth noting. First, errors in a type of unit may *propagate* to all statistics based on this unit.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

For instance, the errors in households affect not only household statistics, but also demographic analysis, household income statistics, expenditure, poverty mapping, etc. Second, unit errors are conceptually different from linkage errors. For instance, suppose a 'dwelling household' consists of the persons sharing a dwelling. As dwellings are composite units in relation to persons (base units), the problem seems to be formally the same as a many-to-one linkage problem between persons and dwellings. But this is only true if, say, the dwelling register is perfect, because otherwise the dwelling units themselves cannot be fixed.

There is an interesting interplay between coverage error and unit error with regard to composite units. Consider the following case. Suppose that persons A and B constitute a two-person household in the integrated data, and person C constitutes another single-person household. Suppose that in reality persons A and C belong to the same household, but person B constitutes a household on his own. Formulated as a coverage problem in terms of the household population, under-coverage of two households (i.e. household of persons A and C and that of person B) and over-coverage of two households (i.e. household of persons A and B and that of C) would occur simultaneously. Formulated as a unit-error problem, it could be said that on the one hand there is no coverage problem in terms of the base units (i.e. persons), but on the other hand, there are two inter-connected unit errors regarding the base units A, B and C. While both formulations may be correct in principle, the second one seems to be the natural approach under the proposed two-phase life-cycle model.

We now turn to the line of measurement concerning the variables. *Harmonized measures* are a kind of conceptual alignment of potentially multiple 'proxy' measures (i.e. from different sources) with regard to the target concept. These concern metadata and nothing is actually done to the data themselves at this stage. If possible, one might arrive at a standardised measure that is closer to the target concept than any proxy measure on its own. The variable 'Occupation' (or 'Job title') provides an example. Typically, there will be a whole range of various positions or titles of a certain type of occupation depending on profession as well as working place (i.e. corporate, institute, office, etc.), so that harmonization is needed in order to arrive at a common standard. The extent of disagreement between the target concept and the harmonized measures is referred to as *relevance error*. We use the term relevance instead of validity here, partly to avoid repetition, partly because relevance is a traditional term in register-based statistics. However, if a justification must be provided, it is that relevance also covers a many-to-one situation between the measures and the target concept, that is in cases where one refrains from explicitly formulating a standardized measure.

Re-classified measures are obtained by turning primary input-source measures into harmonized measures. However, this may not be straightforward if the set of input categories does not have a well-defined mapping with respect to the standard ones. Take for instance the occupation 'senior researcher', which may be a 'professor' or an 'assistant professor' depending on the institute concerned, where for illustration

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

we suppose that 'professor' and 'assistant professor' are part of the standard classification. But the required details might not be available from the institute, or they might be misunderstood or even neglected. Thus, errors in the re-classified measures are unavoidable in practice; they will be referred to as *mapping* errors.

The process from the re-classified to the *adjusted measures* may involve all the familiar editing and imputation activities. A key conceptual difference from editing in the first phase is the existence of inconsistency across the data sources, which may not necessarily imply a quality problem with the input data source from the register owner's perspective. For instance, if someone loses a job, the event of dismissal should be reported by the employer to the job register, albeit within an allowed time lag, whereas the person him- or herself would report to the social security services, in order to receive a benefit. Thus, the event may well be recorded at the social security services earlier than in the job register, without the employer necessarily having been negligent in any sense. In other words, inconsistency across data sources may be unavoidable even without any of them being deficient from their respective producers' point of view. The micro-level reconciliation of inconsistency in multiple-sourced data is often referred to as *micro integration*, which is rightfully a subject on its own (Bakker, 2010). Even if each data source itself is error-free, and there are no mapping errors in the re-classified measures, adjustments may still be necessary. There is thus a source for *compatibility* errors between the re-classified and *adjusted measures*, by which we place the emphasis on the consistency of the various data, rather than their perfection.

2.3 Some recent developments

A shared understanding of the life cycle of integrated statistical data and potential error sources can help us to collocate and coordinate various research and development efforts. For instance, Work Package 4 'Improve the use of administrative sources' of the ongoing EU 7th-framework BLUE-ETS project (http://www.blue-ets.istat.it) aims to develop a framework for the input quality of administrative data, that is, quality evaluated at the end of the first phase and possibly for general statistical purposes. Work Package 2 of the ESSnet project 'Use of Administrative and Accounts Data for Business Statistics' (European Statistical System Network, http://essnet-portal.eu) aims to develop concrete checklists for usefulness and quality of input administrative data, albeit with a strong product focus on the actual business statistics that are to be produced at the end of the second phase. The ESSnet project 'Data Integration' is devoted to the theory and techniques for data linkage (including both record linkage and statistical matching) and micro integration (from harmonization of measures in concept to actual adjustments of data) in phase two in Figure 1.

As examples of statistical theory for dealing with specific errors, we would like to mention two research initiatives recently undertaken at Statistics Norway, both highly relevant for the forthcoming register-based census. The first case concerns the unit errors (second phase, Figure 1) in the household register. As noted earlier, the

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

statistical unit household does not exist in any administrative source in Norway, and must be constructed on the basis of the relevant primary-source units such as person, family and dwelling. A statistical theory for household unit errors has been developed (Zhang, 2010), which allows the assessment of not only the uncertainty in the household statistics, but also in the statistics based on household units. The approach can be applied to any composite units given the base units as described in Table 1. The mappings between the two types of units are formally represented by *allocation matrices,* each containing possibly several many-to-one mappings. The fact that the number of mappings is not fixed in advance is a key difference from a simple many-to-one linkage problem. To facilitate the practice, the allocation matrices are *blocked* (i.e. separated) from each other, by address in the context of the household register. An address may comprise several dwellings that cannot be identified because of the shortcomings of the dwelling register. Each blocked allocation matrix has two realizations, one representing the true household allocations and the other representing those according to the actual household register. The joint distribution of the pair of allocation matrices can be estimated based on an *audit* sample where both are observed. This distribution then provides the basis for statistical inference about the household register (see Table 4 in Zhang, 2010), and the related statistics.

The second case concerns a modelling approach to delays and mistakes in the job register. (Zhang and Fosen, 2011). Initially, these cause selection errors (first phase, Figure 1) in representation. At the stage of data integration, however, the target population is given by the CPR, and the phase-one selection errors cause misclassifications of a binary employment status. For statistical treatment, one may therefore re-formulate these errors as 'measurement errors', that is deviation of the adjusted measure from the harmonized measure (second phase, Figure 1). Notice that, as delays are made up for and mistakes corrected over time, the misclassification errors for a given statistical reference point in time (say, 1 November 2009) vary according to the moment of 'measurement', that is the moment at which information is retrieved from the job register. It turns out that it may take many years before numbers of delay updates and corrections drop to a negligible level. Thus, misclassification errors necessarily generate 'noise' at the moment of production, which must be distinguished from the true information (or 'signals' about the labour market). It is shown that the induced bias may be substantial and, hence, damaging for statistics at detailed levels. A sensitivity analysis approach has been developed, which can provide useful information for the dissemination of detailed register-based statistics. The methodology here is applicable to purely register-based data in the absence of survey data for comparison.

3 Accuracy assessment of integrated microdata

3.1 Traditional topics of data integration and related settings

The concept of target data applies to all three 'traditional' topics of data integration:

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

1. Record linkage starts with two separate sets of units that are at least partially overlapping, if not in fact identical. The result is a linked dataset with all units that were able to be matched between the two initial sets of units. The target dataset is then the matched dataset, in which all the identical units are correctly matched.
2. Statistical matching concerns the situation where the two datasets to be integrated have actually no, or only very few, identical units. Statistical matching *on the micro level* aims to create a single dataset that contains *both* sets of variables, either for the union of the two initial sets of units or only one of them. In any case, given the units to be included in the integrated dataset, the target dataset is the one that contains the true values of all the variables for these units.
3. Micro integration is applied in situations where variables from different sources may have values that are incompatible or inconsistent with each other at the unit level. Micro integration aims to generate a dataset in which all perceived incompatibility or inconsistency has been removed. The target dataset would then contain the *true* values of all the variables involved, but may contain different units than the integrated dataset, for instance if the integrated dataset is subject to linkage errors, or if it is a statistical register that suffers from under- or over-coverage.

The observation of the simple binary data example in section 1.3 can be generalized as follows. Suppose the integrated data (denoted by Z) are to be used together with some auxiliary data (denoted by X), instead of the target data (denoted by Y). For statistical inference, it is then important *how the joint distribution of* (Z, X) *compares to that of* (Y, X). The unit-specific approach, however, focuses on the joint distribution of (Z, Y). Of course, micro-level equality between Z and Y would guarantee valid inference. But micro-level equality is generally not the case in reality, while valid inference does not necessarily require micro-level equality, and comparisons between (Z, X) and (Y, X) in principle do not need to be unit-specific.

We notice that such a theoretical perspective may be useful in several other settings:

1. Indirect (or proxy) reporting may be opted for if direct reporting is not possible, for example available household members answering questions on behalf of absent household members. The indirectly collected data may be considered as surrogates for the target data that could have been collected directly.
2. An interview object may give different answers when asked the same question repeatedly, because of cognitive or social-cultural reasons. This is an example of unstable reporting. It is possible to postulate an underlying latent target variable, but it is also possible not to do so.
3. Some data may be collected by post or through the Internet, while others are collected by direct interviews. Data collected through a particular mode

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

may be seen as the golden standard and, hence, the target data. Or all the modes may be considered to be subject to measurement errors.

4. In the context of releasing public microdata, the target data may be available, but there may be concerns about unsecured sensitive information. Using synthetic data instead of the true data may avoid disclosure risks, provided the true statistical information is maintained (RUBIN, 1993).

In all these situations, it may be argued that statistical accuracy is not primarily a question of whether the data to be used are equal to the target data on the individual level. Indeed, in the case of releasing synthetic microdata, one would rather avoid such equality if this is justified. The purpose is to use the data one has (or prefers) *in place of* the target data for certain needs.

3.2 Statistical validity and equivalence

Let Y and Z be two arbitrary random variables (or sets of variables). Let X be a set of auxiliary variables. Denote the joint distribution function of Y and X by $f(Y=y, X=x)$, and that of Z and X by $f(Z=y, X=x)$. Notice that, using a somewhat compact notation, by y we also denote the outcome of Z. For instance, Y may be the target data and Z may be the integrated data and X may be the set of relevant auxiliary variables. We have

$$f(X=x, Y=y)=f(X=x)f(Y=y \mid X=x) \quad \text{and}$$
$$f(X=x, Z=y)=f(X=x)f(Z=y \mid X=x)$$

where $f(X=x)$ denotes the marginal distribution function of X, and $f(Y=y \mid X=x)$ and $f(Z=y \mid X=x)$ the conditional distribution functions of Y and Z given X, respectively. We assume independence between the observations. This is the most usual case in practice. Otherwise, we may assume that the distributions are defined for independent clusters (of observations).

If $f(Y=y, X=x)=f(Z=y, X=x)$, then $(Z,\ X)$ has the same joint distribution as $(Y,\ X)$. Since the marginal distribution $f(X=x)$ is the same, this is equivalent to $f(Y=y \mid X=x)=f(Z=y \mid X=x)$. The joint data $(Z,\ X)$ should then provide valid inference if these are used in place of $(Y,\ X)$, so we may say that Z is *valid* for Y with respect to X, or *vice versa*.

For instance, let Z be the proxy interview data, and let Y be the direct interview data. Then, the indirect interviews are statistically valid for direct interviews, provided $f(Y=y \mid X=x)=f(Z=y \mid X=x)$, where X may contain the variables used in weighting. The two interview outcomes do not need to be identical, should they be administered on the same individual.

In this way statistical validity is defined with respect to the distribution functions. It does not presume that the observations of Z and Y have to be available for the same units, or the same sample size. Let $s(Y,\ X)$ denote the units for which $(Y,\ X)$ is available, and let $s(Z,\ X)$ denote the units for which $(Z,\ X)$ is available. It is possible to investigate whether $f(Y=y \mid X=x)=f(Z=y \mid X=x)$ no matter whether $s(Y,\ X)$

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

is the same as $s(Z, X)$ or not. In other words, the concept of validity allows for comparisons that are not unit-specific. For instance, to study the validity of proxy interview data, one may compare these with the direct interview data that are obtained for other persons.

Still, $s(Z, X)$ may not yield *identical* inference as that based on $s(Y, X)$, even if Z is valid for Y with respect to X. For that we need a stronger condition. Let

$$p(X=x, Y=y; s(y,x)) = p(X=x; s(y,x))p(Y=y \mid X=x; s(y,x))$$

be the empirical distribution function of (Y, X) in a *given* set $s(y, x)$, and let

$$p(X=x, Z=y; s(z,x)) = p(X=x; s(z,x))p(Z=y \mid X=x; s(z,x))$$

be that of (Z, X) given $s(z, x)$. The notation emphasizes that, even if $f(X=x)$ is the same, $p(X=x; s(y,x))$ may differ from $p(X=x; s(z,x))$. We say that the dataset $\{(z,x); s(z,x)\}$ is *empirically equivalent* to the dataset $\{(y,x); s(y,x)\}$ if $p(X=x, Y=y; s(y;x)) = p(X=x, Z=y; s(z,x))$.

Empirical equivalence ensures identical statistical information in the two given datasets. Equality on the individual level is not necessary given empirical equivalence. Moreover, empirical equivalence can be considered as a non-parametric version of the sufficiency principle in the parametric setting of statistical inference. According to the sufficiency principle (e.g. Birnbaum, 1962), all relevant information in the data is summarized in the sufficient statistics of the data, whose distribution function is proportionally related to that of the whole dataset by a constant that does not depend on the model parameters. Thus, in a parametric setting, a dataset may be said to be *statistically equivalent* to another dataset, provided identical sufficient statistics. However, in contrast to empirical equivalence, statistical equivalence as such is model-dependent.

3.3. Public microdata and disclosure control in terms of validity and equivalence

Not surprisingly, ideas that are similar to the concept of validity and equivalence can be found in the literature on data integration and other related topics. For instance, so-called matching noise (Paass, 1985) can be regarded as an expression for deviation from validity. Here we would like to discuss the relations to public microdata subjected to disclosure control in particular.

Microdata quality for analytical purposes is of concern with regard to public microdata subject to disclosure control. In particular, the following approaches can be distinguished:

1. From the census file, samples of anonymised records (SARs, http://www.ccsr.ac.uk/sars) can be drawn according to $p(X=x, Y=y; U)$, where U denotes the finite census population. By construction, the SARs approach yields valid datasets with regard to the known distribution $p(X=x, Y=y; U)$ under simple random sampling. But validity is also achievable under complex

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

sampling designs, either unconditionally or conditionally on X, after appropriate weighting adjustments.

2. In the case of a survey sample $s(y,\ x)$, synthetic datasets can be generated according to $p(X=x,\ Y=y;s(y,x))$, or some estimated distribution $f(X=x,\ Y=y)$ on the basis of $\{(y,x);s(y,x)\}$, without conditioning on the available x values, that is, $\{x;s(y,x)\}$. The key difference between this micro simulation approach and the SARs approach is that the true joint distribution is unknown and must be estimated. In addition, when sampling from an estimated distribution under a statistical model, instead of the observed empirical distribution, the generated data may not correspond to any real units and are hence said to be 'imaginary' (Fienberg, Makov and Steele 1998). In any case, the micro simulation approach may yield synthetic data that are valid *in expectation*, provided the distribution from which the data are generated is an unbiased estimator of the underlying target distribution.
3. Fienberg *et al.* (1998) proposed to condition on $\{x;s(y,x)\}$ and randomly generate synthetic data Z according to the empirical distribution of $p(Y=y\mid X=x;s(y,x))$, or $f(Y=y\mid X=x)$ that is estimated under a suitable parametric model. These are referred to as 'pseudo' microdata in contrast to the micro simulation approach. The data obtained in this way may be *conditionally* valid in expectation given $\{x;s(y,x)\}$, provided the conditional distribution of Y given X is an unbiased estimator of the target conditional distribution.
4. Earlier, Rubin (1993) outlined a similar approach under the Bayesian multiple imputation framework, albeit without the emphasis on conditioning of $\{x;s(y,x)\}$.

In brief, all the aforementioned approaches may lead to valid, but never equivalent microdata. Notice also that, in all the cases, units with unique, or very low frequency, of $(x,\ y)$ value may still appear in the synthetic data, and in general posterior disclosure control is necessary. The distinction between validity and equivalence raises the question whether public microdata should not aim directly at equivalence. Two alternative approaches seem to be worth considering.

1. Aim for full empirical equivalence first, followed by disclosure control. Notice that, for contingency tables, empirical equivalence implies equality between two tables, so that the approach simply amounts to disclosure control of the target contingency table. Notice also that the loss of information through disclosure control is uncontrolled in theory, despite the fact that the result may be acceptable from both an inferential and a public relations point of view.
2. Generate equivalent synthetic data aimed at *coarsened* information, where disclosure control is embedded in the coarsening of the statistical information. Contingency tables subject to minimal sufficient marginals (e.g. Fienberg *et al.*, 1998) is an example of this approach. The table to be

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

released is statistically equivalent to the target table under the assumption of a log-linear model up to the corresponding interaction structure, without being equivalent to the target table in general. Another example is multivariate datasets subject to given mean and covariance matrix, which is equivalent to the target dataset under the assumption of a linear regression model with constant variance and normally distributed residuals. The loss of information is clearly defined under this approach, which may satisfy users with specific and limited needs.

3.4 Validity versus accuracy

It is important to realize that statistical validity is not the same thing as accuracy or efficiency at aggregated levels.

As an example, consider the annual register-based employment status in Norway. The annual Employment Register (ER) is a statistical register derived from a number of administrative registers including the Employer/Employee register, the Wage sum and Tax registers, the Self-employment register, etc. Among other things, it contains the employed/not-employed status, referred to as the ER status, for the whole target population at a given reference time point for each calendar year. The target employment status follows the ILO definition, and is available for the respondents in the quarterly Labour Force Surveys (LFS). Disregarding the potential measurement errors in the LFS, we shall tacitly consider the LFS status as the target measure. The micro integration process that generates the ER status is designed accordingly. Nevertheless, perfect agreement between the ER and LFS status is impossible for a number of reasons, such as the definitional bias in the administrative sources, inconsistency on the individual level between the underlying administrative data sources, intrinsic random variations in the registration processes, etc.

One motivation for the ER is to produce employment statistics at a low level of aggregation. Consider first the ER-based employment rate. We shall tacitly disregard all the random variations in the ER, assuming that these are negligible at the level of interest. Still, the ER employment rate is subject to a definitional bias. Next, consider the direct LFS estimator based on the data within the domain of interest. We shall tacitly disregard all the non-sampling errors and assume that the direct LFS estimator is unbiased. But it is subject to a sampling variation. We thus face a trade-off between the bias of the ER employment rate and the variance of the corresponding LFS estimator. It is now conceivable that, while the balance may be in favour of the LFS estimator at a more aggregated level, such as the national total, the ER employment rates may become more accurate (say, in terms of the MSE) than the corresponding LFS estimators at a more detailed level. Fosen and Zhang (2011) show that this is the case at the municipality level.

In short, even though the integrated statistical data are not fully valid in terms of microdata quality, they may outperform the available target data when it comes to producing statistics.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

3.5 On an alternative approach to record linkage

According to the tradition of Fellegi and Sunter (1969), record linkage attempts to maximize the probability of correct linkages between the units of two datasets. Often, however, the subsequent analysis proceeds as if the linked dataset were free of linkage errors. Chambers (2008) formulated a framework that can be used to modify regression analysis based on the probabilistically linked data. The approach is based on modelling the mapping between the linked data and the target data. A tacit assumption is the absence of non-linkable units. That is, the two datasets are the same size and the units are always assigned a one-to-one mapping as the result of record linkage. In this case, the mapping between the dependent variables in the linked dataset and those in the target data can be represented by a permutation matrix, that is, a matrix obtainable by a (row) permutation of the corresponding identity matrix of the same dimension. Random variations of the linked dataset, where the randomness is caused by the nature of *probabilistic* record linkage, can then be incorporated into the analysis in an appropriate manner.

Some extensions to this approach have been developed more recently (Kim and Chambers, 2010a,b). One is to allow for the situation where some of the independent variables of the subsequent regression analysis may be obtained from a different linkage procedure than that of the dependent variable. In this case, part of the true design matrix (i.e. of the linked independent variables) is not observed, and a mapping matrix for these variables is needed in addition to the one for the dependent variable. Another is a situation where a sample of the units (i.e. the first dataset) is linked to the population of units (i.e. the second dataset). The relationships between the variables under (hypothetical) full data linkage can be written down by augmenting the non-sampled units of the first dataset, leading to a corresponding partition of the full-data mapping matrix. The basic linkage error model can then be applied to the different parts of this matrix in an appropriate manner, as the two full datasets are still assumed to contain the same units, after which the analysis based on the actually linked dataset can be derived.

Now, consider an alternative approach to record linkage, where one relaxes the Fellegi–Sunter maxim of maximizing the probability of correct linkage. Instead, one may choose to maximize the number of linked records to start with. Of course, the initial links should be as plausible as possible. In other words, instead of maximizing the probability that an established link is correct, one might start by minimizing the probability of missing a correct link. This approach would typically result in an initial linked dataset that is the size of the smallest of the two original datasets.

In the next stage, one would consider adjusting the initial linked dataset to improve its *(statistical) validity*. This requires a somewhat different approach to auditing. Traditionally, the results of recording linkage are audited to provide an estimate of the probability of correct links. It rarely produces the probability of missing links, although this would certainly be desirable from a theoretical point of view. Otherwise, one is bound to assume that the probabilistic linkage process

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

is 'non-informative' for the subsequent analysis. Just imagine that, as a hypothetical example, one achieves perfect linkage among a subset of the units and no links otherwise. Then the linkage process is informative or not, depending on whether the linked units are a random subset of the target dataset or not. However, nothing is guaranteed by the fact that the linkage process is perfect according to the Fellegi–Sunter maxim.

For the alternative approach, one would expect two measures from auditing. The first one, based on the correct links identified in a random audit sample, concerns the target joint distribution of interest. The second one, based on both the correct and the incorrect links in the audit sample, concerns the actual joint distribution based on the initial linked dataset. The initial linked dataset is valid if the two distributions agree with each other. Otherwise, there is room for adjustment. Notice that this kind of evaluation does not require a unit-specific approach. The comparison is between two distributions, rather than for discrepancies at the unit-level.

Obviously, some issues here need more research to make the process work. The audit sampling design is probably more efficient if the selection probability varies in a sensible way across the initial linked dataset. This also means that there may be an adjustment issue when evaluating the two joint distributions in the audit sample. When adjusting the initial linked dataset, should one treat the units differently according to their probabilities of being correct links, or should one simply disregard this and concentrate exclusively on the validity of the joint distribution of the linked data? Is validity (on expectation) achievable in general? Provided it is, for inference based on the linked dataset, one still needs to take into account the fact that the validity is estimated and not given. This involves the potential trade-off between using the linked dataset, which is associated with uncertainty surrounding its validity, and just the correct links identified in the audit sample, which presumably has a much smaller sample size.

3.6 Micro calibration: a method of data integration

Although all the Nordic countries have announced that the next census – in 2011 – will be register-based, they will nevertheless utilize information from sample surveys more or less directly. For instance, the Norwegian register-based census employment data are constructed in such a way that the national employment total would agree with the corresponding LFS estimate. The situation can formally be described as follows. It is possible to obtain a *purely* register-based census file. Suppose that for one of the census variables, a survey variable exists which can be regarded as providing the target measure. The question then arises whether the census variable can be adjusted on the micro level, by a minimum amount in some sense, so that the distribution of the adjusted census variable agrees with what can be inferred based on the observed target measure in the survey, possibly jointly with a chosen set of covariates. In other words, the census variable achieves statistical validity as a result of the adjustment. We shall refer to such a method as *micro calibration*. In the case

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

of the Norwegian census employment status, the aim is only the target mean. Had the entire distribution been 'matched', the census total would automatically have agreed with the target total. Moreover, by means of micro calibration, one can adjust a register-based census file to agree with all the relevant survey information, which is a new method of data integration.

There seems to be a connection to the literature of statistical matching and, in particular, the analysis of the so-called uncertainty space (e.g. Conti *et al.*, 2009). The information on different variables in the separate surveys is integrated into a single census file by means of micro calibration, just as in statistical matching, which constructs the joint information of multiple variables that are not jointly observed in either of the separate (usually, two) datasets that have no, or virtually no, overlapping units. Given that there are actually no observations of the target joint distribution, standard statistical matching procedures proceed under the conditional independence assumption (CIA). That is, conditional on the variables that are actually observable in both data sources, the rest of the variables are independent between the two sources. The CIA is ultimately unsatisfactory. Of course, it may be the case that, as more and more common variables become available, the CIA becomes less and less critical for the results. But this would also imply that less is gained through statistical matching. The study of uncertainty is therefore an interesting recent development in statistical matching. Essentially, it concerns the analysis of the so-called *uncertainty space*, which is the set of all the possible (generally not unique) distributions of the random variables that are compatible with the available information.

It would appear that the uncertainty space is the same when either statistical matching or micro calibration is applied given two separate survey datasets. The difference is the actual integrated dataset. In statistical matching, it is centred around the expected joint dataset under the CIA. In micro calibration, it would be close to the initial, say, register-based census file. In other words, the information in the two survey datasets is combined with the information in the initial census file in micro calibration, whereas in statistical matching they are combined only with the hypothetical CIA but no empirical information. These issues should provide interesting topics for research.

4 Summarising remarks

To summarise, official statistics production today faces the challenge of navigating a course between budgetary constraints on the one hand and the ever increasing demand for statistical information on the other. Naturally efficient use of all data available through data integration is an option that must be explored; administrative register data is a major, but not the only, aspect of this. Many new and difficult theoretical challenges provide interesting topics for further research. The 20th century witnessed the birth and maturing of sample surveys; the 21st century will be the age of data integration.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

Acknowledgements

I would like to thank the referees and the guest editors for their helpful comments and suggestions.

References

Bakker, B. (2010), Micro-integration: State of the art, in: *ESSnet on Data Integration, Draft Report of WP 1*, 55–78.

Birnbaum, A. (1962), On the foundations of statistical inference, *Journal of the American Statistical Association* **57**, 269–306.

Bowley, A. J. (1926), Measurement of the precision attained in sampling, Memorandum published by the Int. Stat. Inst., *Bulletin of the International Statistical Institute*, Suppl. to Vol. **XXII**, *Book 1,* 1–62.

Brackstone, G. J. (1987), Issues in the use of administrative records for statistical purposes, *Survey Methodology* **13**, 29–43.

Chambers, R. (2008), *Regression analysis of probability-linked data*, Official Statistics Research Series, vol. 4, Statistics New Zealand.

Conti, P. L., M. Di Zio, D. Marella and M. Scanu (2009), *Uncertainty analysis in statistical matching*, First Italian Conference on Survey Methodology (ITACOSM09), Siena.

Fellegi, I. P. and A. B. Sunter (1969), A theory for record linkage, *Journal of the American Statistical Association* **64**, 1183–1210.

Fienberg, S. E., Makov. and R. J. Steele (1998), Disclosure limitation control usingperturbation and related methods for categorical data, *Journal of Official Statistics* **14**, 485–502.

Fosen, J. and L.-C. Zhang (2011), *Quality assessment of register-based census employment status*, *Proceedings of the International Statistical Institute,* World Congress, Dublin.

Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer and R. Tourrangeau (2004), *Survey methodology*, Wiley, New York.

Holt, T. (2007), The official statistics Olympic challenge: wider, deeper, quicker, better, cheaper. (With discussions), *The American Statistician* **61**, 1–15.

Jensen, A. (1924), The report on the representative method in statistics, *Bulletin of the International Statistical Institute*, Vol. **XXII**, Book 1, 359–380.

Kim, G. and R. Chambers (2010a), *Regression analysis for longitudinally linked data*, Working Paper Series 22–10, University of Wollongong.

Kim, G. and R. Chambers (2010b), *Regression analysis under incomplete linkage*, Working Paper Series 17–09, University of Wollongong.

Kiær, N. (1897), *The representative method of statistical surveys (1976 English translation of the original Norwegian)*, Central Bureau of Statistics of Norway, Oslo.

Neyman, J. (1934), On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society* **97**, 558–606.

Paass, G. (1985), *Statistical record linkage methodology, state of the art and future prospects, Bulletin of the International Statistical Institute, proceedings of the 45th Session, LI, Book 2.*

Rubin, D. (1993), Discussion, statistical disclosure limitation, *Journal of Official Statistics* **9**, 461–468.

Schulte Nordholt, E. (2005), The Dutch Virtual Census 2001: a new approach by combining different sources, *Statistical Journal of the United Nations Economic Commission for Europe* **22**, 25–37.

Statistics Denmark (1995), *Statistics on persons in Denmark – a register-based statistical system*, Eurostat, Luxembourg.

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

© 2011 The Author. Statistica Neerlandica © 2011 VVS.

UNECE (2007), *Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics*, United Nations Publication, ISBN 978-92-1-116963-8.

WALLGREN, A. and B. WALLGREN (2007), *Register-based statistics – Administrative data for statistical purposes*, John Wiley and Sons, Chichester.

Zhang, L.-C. (2011), A unit-error theory for register-based household statistics, to appear in *Journal of Official Statistics* **14**, 415–432.

ZHANG, L.-C. and J. FOSEN (2011), Assessment of uncertainty in register-based small area means of a binary variable, to appear in *Journal of Indian Society of Agricultural Statistics*.

Received: 7 June 2011. Revised: 8 August 2011.

© 2011 The Author. Statistica Neerlandica © 2011 VVS.