Topics of statistical theory for register-based statistics

Zhang, Li-Chun
Statistics Norway
Kongensgt. 6, Pb 8131 Dep
N-0033 Oslo, Norway
E-mail: lcz@ssb.no

1. Introduction

For some decades now administrative registers have been an important data source for official statistics alongside survey sampling and population census. 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 ones of which came already into existence in the late 1960s and early 1970s. Statistics Denmark was the first to conduct a register-based census in 1981, and the next census in 2011 will be completely register-based in all Nordic countries. Whereas 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 include reduction of response burden, long-term cost efficiency, as well as potentials for detailed spatial-demographic and longitudinal statistics. The trend towards more extensive use of administrative data has long been recognized by statisticians around the world (e.g. Brackstone, 1987). But also being noticed is a clear lack of *statistical* theories for assessing the uncertainty of register-based statistics (e.g. Holt, 2007).

It is tempting to reflect over the historical development of survey sampling for comparison. The *representative method* was presented by Kiær at the ISI meeting in 1895. Despite he was unable to defend the approach on a theoretical ground, the practice kept evolving afterwards. In 1924, the ISI formed a committee to investigate the feasibility of the representative method. The reporter Jensen 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". The theoretical breakthroughs, indeed, did not come until some 30 - 40 years after Kiær's initial presentation. Today, the contributions by Bowley 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.

Register-based statistics, when viewed in this mirror of history, appear currently 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* in register data which will enable us to apply rigorous statistical concepts such as bias, variance, efficiency and consistency, as one is able to do in other branches of the statistical science.

Administrative registers certainly do not provide perfect statistical data. For various reasons, the available, or observed, value (or values) may differ from the ideal measure (or measures) that is of statistical interest for the target units. 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 adaption to the administrative data. Expanding and developing on these initiatives, we have arrived at a chart of two-phase life cycle of integrated statistical micro data, which is now given in Figure 1. The squared boxes describe the different stages of data production. A source of error is located between any two stages, corresponding to an accuracy aspect (in the ovals). The word error is used here with an eye on the final state of the data because, between the data in two stages, it is sometimes more natural to speak of a deviation (or gap) rather than one of them being true and the other erroneous.

The entire life cycle of statistical micro data is generally divided into two phases. The first phase concerns the data from each single source, and the second phase concerns then the possible integration of data from different sources. There are mainly two inter-connected reasons for this distinction of two phases, 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 some 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 works that are carried out at the statistical office in order to make the data fit for statistical purposes. More generally, however, one may speak of secondary usage of any data, as long as they have been collected for a different purpose primarily, including the reuse of sample survey data for other purposes than the statistics that they have initially been collected for. In this way, also the life cycle of sample data will be extended beyond the first phase.

The second reason is the fact that, often, one can only make statistical uses out of an administrative register after it has been combined with data from other sources. A sample survey may suffice for a statistical purpose on its own because it is designed to be so. Auxiliary information can be incorporated to improve the efficiency or to reduce the various non-sampling errors, but even without them statistics can still be produced. The matter is different when it comes to administrative data. By definition an administrative register is established for purposes other than statistical. Integration is necessary in order to assimilate it into the statistical system at all. For instance, to produce statistics of the (highest) attained education level, one might make use of, say, a register of examination results. However, this register is organized on the basis of individual exam results, where the same person may have multiple entries, so that it is not possible to produce education statistics of persons without *further* processing. A two-phase distinction is natural here because one can not expect the register owner to do the job for the statistical agency. Moreover, a register of examination results will not contain information on everyone in the population. So integration with the Central Population Register will be necessary. Indeed, to control the quality, one would probably take into account, say, the register of school enrollment and any other information that may be relevant and helpful. While data integration is an essential feature associated with the use of administrative registers, it is unnecessary to restrict data integration to these alone. Also survey and census data may and should be integrated so as to improve their scope and/or quality. Indeed, any potential data sources for that matter.

It should be noticed that often the word single-source needs to be taken relatively, for the purpose of a suitable 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, i.e. 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 input data.

In what follows we describe the two phases of statistical micro data in more details, respectively, in Section 2 and 3. Some discussions on recent and possible future developments are provided in Section 4.

2. Phase one

At phase one we are concerned with the data of a particular source on its own. The life cycle of sample survey data as charted by Groves *et al.* falls under this phase, except that we have excluded the process of post-survey 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 the line of representation first. 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 corresponds to what is known as the sampling error. The *observed set* may be referred to as the *net* sample that contains only the respondents, and the difference from 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. Notice that 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 distinctions are equally applicable to an administrative data source. The difference between the target and accessible sets may arise due to feasibility reasons. For instance, a job register may be intended to provide the basis for the administration of sick and child-care benefits. The target set should ideally contain all jobs that are 'substantial' enough to qualify for the benefits. To facilitate the reporting in practice, however, only jobs of certain regularity are instructed to be reported, hence accessible at all. Notice also that the 'units' of a job register like this are the various job-related events, such as hiring, dismissing, leave of absence, *etc.* Whereas labor-market statistics usually have persons as the statistical units. In order to make use of the job register, then, information by job events must be reorganized as information by persons, just like in the case with the examination register earlier. This is the reason that we choose to use the term *objects* for the units at phase one, to potentially distinguish these from the more familiar term *units* for statistical units at phase two. Next, to continue with the job register example, the accessed set would contain all the 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, while some of the reported events may be inadmissible due to confusion surrounding the administrative regulations. Finally, the validated set contains the objects that remain after a validation process by the register owner. Inadmissible objects may be detected at this stage, which nevertheless are attributed to the selection error. Still, some objects may be rejected at this stage due to *missing* information, while *redundancy* may remain or arise by mistakes. Although the difference between the validated and accessed sets may be small in many cases, the conceptual distinction between an object in a state as recorded and that after processed seems justified in general.

The line of measurement at phase one is a one-to-one mapping of the corresponding line of Figure 2.5 in Groves *et al.* (2004) for sample survey data. The corresponding terms there for the squared boxes are "Construct", "Measurement", "Response" and "Edited response". Construct is the ideal information that is sought about an object. It can be abstract and sometimes ambiguous, such as one's political orientation. Measurements are necessarily concrete and observable, such as the party one 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.

All these concepts can be applied analogously to administrative data. However, we have added "Registration" to "Response" for obvious reasons. Many administrative registers are based routine registration of events, and do not involve respondents as such. For the same reason we use the term *editing* in place of edited response, also because at this stage new variables/values may be derived from the ones that are actually observed. Finally, we have adopted the term *target concept* instead of *construct,* 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 will simply fail to serve the administrative purposes. Take the job register again. The routine for reporting may fail to capture some jobs that are as substantial as the ones that are actually being reported. For instance, suppose an employer is instructed to report any job that lasts more than two weeks and with a minimum of 20 hours in a week. Then a job that lasts for one week and two days will fail the criterion, despite it may well amount to more than 40 hours over the two weeks. On the other hand, what is 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 for describing or analyzing certain social-economic phenomenon. The idea of a theoretical construct for the measurement seems remote.

It is perhaps worthwhile to spend a little more time on the distinction between measurement and representation in relation to that between object and unit mentioned above. The term object has been introduced to contrast unit because so many administrative registers are initially organized by events. The examination register and job register have been mentioned. The Central Population Register consists of events such as birth, death, marriage, divorce, *etc.*; the VAT register is built on transactions, rather than by establishments or enterprises; and so on it goes. Unless the reporting/recording of events is fully automated, it may suffer from *delay* whichever the time point information is retrieved from the register for statistical purposes. Take the job register again. Suppose for illustration that a person is classified as "Employed" if he/she has an active job according to the job register, i.e. there is an event of hiring for a job in the register but no event of dismissing from the same job. Now consider the case when someone has been dismissed, but the event of dismissing somehow had not entered the register at the production time point. For the job register, the event of dismissing is an object along the line of representation, which belongs to the accessible set but not the accessed set, i.e. it is a selection error that probably has occurred by chance. However, if one had in fact further processed the job register and reorganized the information by persons, then this person should appear also in the accessed set, but with a misclassified employment status. In other words, one has now a measurement error instead of a selection error. As a statistical producer one needs to be prepared for both scenarios. It all depends on how far the register owner processes the data, and the structure of the data that are actually delivered to the statistical office. Still, in any case, the same selection error will most certainly cause a measurement error for any employment statistics at phase two of data integration in very much the same way. This is thus also an example for the necessity of the two-phase division in Figure 1, without which it seems difficult to handle the situation where an initial error along the line of representation eventually may be dealt with as a measurement error statistically.

3. Phase two

Integration of data from multiple sources occurs at phase two. The *target population* and *target concept* here are defined according to the statistics of interest, which are usually different 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.

Take first the line of representation concerning the units. As noted before, often, a *transformation* of information by objects to units is necessary. Moreover, the operation may affect both the line of representation and the line of measurement, because the information of the first-phase objects may be reorganized as measurements (i.e. variables) of the second-phase units, such as discussed in the examples of examination register and job register earlier. It is important to emphasize that the units that come out of this transformation are not necessarily the target statistical units. The aim is above all to obtain *linkable* units across the relevant data sources, as a preparation for *data linkage.* Consider the example of household. In most European countries there is no complete frame of households, in which case it is impossible to arrive at households as the units of linkage directly through the transformation stage. Instead, households need to be created or validated at a later stage based on all the relevant information from the different data sources. The unit at which level the data may be linked together is typically person in this case. Multiple data sources carrying information about family relationship, dwelling address, or even tax returns need first to be transformed to persons, which are then linked together at the stage of data linkage. Afterwards, 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.

Table 1. Alignment table illustrated.

Base Unit	Composite Unit Type-1	…	Composite Unit Type-D
1	1		1
2	1		2
3	2		2
…			

The task of *alignment* is to clarify all the relevant information for the creation (or validation) of statistical units in the linked data. Table 1 provides a generic representation of the results of alignment. 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, which may or may not be the units of data linkage. In any case, there is a many-to-one relationship between the base units and each type of composite units. But there may be no direct mapping between two types of composite units, as long as they can not be put into a hierarchical relationship to each other. For instance, in Table 1, the no. 1 composite unit of type-1 consists of base unit no. 1 and 2, whereas base unit no. 2 and 3 belong to the no. 2 composite unit of type-D, and so on. But there is no direct mapping between composite unit type-1 and type-D. Consider again the example of household. The base units are the persons. Family relationship between a number of persons can be summarized, say, in the form of a common and unique "family identity number" for these persons. Each family identity number then identifies a composite unit of a type that is different from a composite unit formed, say, by a "dwelling identity number". Moreover, imagine two parents and their adult son who live in a different dwelling with a cohabitant, and there will be no hierarchical relationship between the composite unit of dwelling and that of family. Of course, alignment as such is no simple task in many situations, and we shall refer to the potential errors as the *identification* errors.

Apart from possible mistakes in the alignment of the different type of units, difficulties may sometimes arise when it comes to 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 from each other. It is common practice in such cases to assign an industry code to the enterprise, which is in some sense judged to be the most important or 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 be problematic in many circumstances. Of course, the reason, as can be seen from the alignment table, is that a composite unit really admits only *partial classification*, as long as its base units may fall in different categories according to the same classification.

Having gathered and prepared all the relevant information after alignment, *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. For business statistics, one may expect to find the legal or tax units to be available to the statistical system. But these are not necessarily the statistical units of interest. The situation here, however, is less clear-cut and not as easy to describe in a few words, due to the different legislations and regulations in different countries. Two points are worth noting in particular. First, errors in a type of units may affect *all* the statistics that are based on these units, in which sense the unit errors are not restricted to any single statistical subject. For instance, the errors in households may affect not only the household statistics, but also demographic analysis, household income statistics, poverty mapping, *etc.* Next, the unit errors are conceptually different from linkage errors. For instance, suppose one is willing to define a "dwelling household" as the persons who share the same dwelling. Since the dwellings are composite units in relation to persons treated as the base units, it seems that the problem is formally the same as a many-to-one linkage problem between persons and dwellings. But this is true only if, say, the dwelling register is perfect, because otherwise the dwelling units can not be fixed themselves.

Take now the line of measurement concerning the variables. *Harmonization* is a kind of conceptual alignment of potentially multiple "proxy" measurements (i.e. from different sources) with regard to the target concept. It deals with the metadata, and nothing is actually done to the data themselves at this stage. If possible, one might arrive at a harmonized measurement that is closer to the target concept than any proxy measurement on its own. In any case, the extent of agreement between the target concept and the measurements is referred to as *relevance*. 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 covers also a many-to-one situation between the measurements and the target concept, i.e. in cases where there is not an explicitly formulated harmonized measurement.

Classification is obviously needed in order to derive the value of a harmonized measurement. But it also covers the situation where a re-classification is carried out. The variable "Occupation" (or "Job title") provides an example. Typically, there will be a whole range of various classifications of a certain type of occupation depending on the profession as well as the working place (i.e. corporate, institute, office, *etc.*). A re-classification of the input variable according to the statistical standard will then be necessary. However, this is not straightforward if the one set of categories do not have a well-defined mapping to the other. For instance, a "senior searcher" might be a "professor" or an "assistant professor", depending on the associated institute. Thus, errors are unavoidable, which will be referred to as the *mapping* error.

The remaining data processing is carried out at the stage of *adjustment,* which involves all the familiar editing and imputation activities. A key conceptual difference from editing at the first phase is the existence of inconsistency between data across the sources. Of course, errors in the input data may lead to inconsistency among the integrated data. For instance, a delay in the job register may cause an inconsistency with the social security information. Notice that such a situation does not necessarily imply a quality problem with the input data source. In the case of someone's losing a job, the event of dismissing 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 for benefits. Thus, the event may well be recorded at the social security earlier than at the job register, without the employer necessarily being negligent in any sense. In other words, inconsistency across data sources may be unavoidable even without any of them being deficient in quality. The reconciliation of inconsistency in multiple-sourced data on the micro level is often referred to as *micro integration,* which is rightfully a subject on its won. Ideally speaking, if each data source is error-free on its own, then all the remaining adjustments of data that are needed belong to the realm of micro integration. We refer to the potential errors as *compatibility* errors to focus on consistency instead of perfection in data.

4. Some recent developments

A shared understanding of the life cycle of integrated statistical data and the potential error sources can help us to collocate and coordinate the different research and development efforts. There are several ongoing joint European projects where the use of administrative data plays a prominent role, either in the frame of ESSnet projects (European Statistical System Network, http://essnet-portal.eu) or European Commission 7th Framework Programme (http://cordis.europa.eu/fp7). For instance, Work Package 4 "Improve the use of administrative sources" of the 7th-framework BLUE-ETS project (http://www.blue-ets.istat.it) aims to develop a framework for (administrative) input data quality under a broader perspective of data integration, in which context some of the work presented here was carried out. This is clearly related e.g. to Work Package 2 of the ESSnet project "Use of Administrative and Accounts Data for Business Statistics", which aims to develop concrete checklists for usefulness and quality of input administrative data, albeit with a strong *product* focus on business statistics. Whereas the entire ESSnet project "Data integration" is devoted to the theory and techniques for the production stage of data linkage (including both record linkage and statistical matching) and adjustment (with a focus on micro integration) at phase two in Figure 1.

Several research initiatives recently undertaken at Statistics Norway provide examples of developing statistical theories that deal with specific error sources, all of which are highly relevant for the forthcoming register-based census. The first case we would like to mention is the unit errors (2nd phase, Figure 1) in the household register. As noted earlier, the statistical unit household does not exist in any administrative source, 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 one to assess not only the uncertainty in the household statistics, but also the statistics that are 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 an *allocation matrix*. Each allocation matrix may contain several many-of-one mappings. The fact that the number of mappings are not fixed in advance is a key difference to a simple many-to-one linkage problem. To facilitate the practice, the allocation matrices are *blocked* (i.e. separated) from each other by address, which in the context of the household register may contain multiple dwellings that can not be identified due to shortcomings of the dwelling register. Each blocked allocation matrix have two realizations, one representing the true household allocations and the other 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, which then provide the basis for making statistical inference about the household register and the related statistics.

As a second case we take the register-based census employment status. The employment status was produced based on a number of the relevant administrative registers already in the last census. The statistical uncertainty of this variable can be estimated based on micro data that are linked to the Labor Force Survey (LFS) on the individual level. However, micro data linkage may not always be feasible or even permitted. Moreover, while equality between any two variables on the individual level implies equal statistics at aggregated levels, it is not a necessary condition for sufficient accuracy of one variable compared to the other at the aggregated levels. This is in fact an issue that may arise in many situations, for any type of errors along the line of measurement (2nd phase, Figure 1). A general theory for valid and equivalent statistical micro data is currently being developed. Particular methods for assessing the statistical accuracy of the register-based census employment status, which does not require micro data linkage with the LFS, are being studied as part of the ongoing ESSnet project Data Integration.

A related topic that has been studied is a modeling approach to delays and mistakes in the job register. (Zhang and Fosen, 2010). Initially, these cause selection errors (1st phase, Figure 1) in representation. At the stage of data integration, however, the target population is given by the Central Population Register, and the phase-one selection errors manifest themselves as measurement errors (2nd phase, figure 1) instead, which amount to misclassifications of a binary employment status. Notice that, since the delays are being updated and mistakes corrected over time, the misclassification errors for a given statistical reference time point (say, the 1st of November in 2009) vary according to the time point of ‘measurement’, i.e. the time point at which information is retrieved from the job register. It turns out that it may take many years before the amount of updates and corrections drop to a negligible level. Thus, the misclassification errors necessarily generate ‘noises’ at the time point of production, which must be distinguished from the true information (or ‘signals’ about the labor market). In particular, under the assumption that misclassification is subjected to an unknown random mechanism, 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 register-based detailed statistics. Unlike the aforementioned approach to assessing the accuracy of the register-based census employment status, the methodology here is applicable to purely register-based data in the absence of survey data for comparisons.

To summarize, the national statistical institutes face the challenge of finding a way between budgetary constraints and ever increasing demand on statistical information. Efficient use of all data available is naturally an option that must be explored. Administrative register data have become a major data source for official statistics, carrying with it many new and difficult theoretical challenges. The 20th century has witnessed the birth and maturing of sample surveys. The 21st century will be the age of data integration.

REFERENCES (RÉFERENCES)

Bakker, B. (2910). *Micro-integration: State of the Art.* Paper for the joint UNECE/Eurostat Expert Group Meeting on Register-Based Census. The Hague, the Netherlands.

Brackstone, G.J. (1987). Issues in the Use of Administrative Records for Statistical Purposes. *Survey Methodology,* vol. 13, pp. 29 – 43.

Groves, R.M., Fowler Jr., F.J., Couper, M., Lepkowski, J.M., Singer, E. and Tourrangeau, R. (2007). *Survey Methodology*. New York: Wiley.

Holt, T. (2007). The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper. (With discussions). *The American Statistician,* vol. 61, pp. 1- 15.

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*, nr 97, 558-606.

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*, Vol. 22, pp. 25-37.

Statistics Denmark (1995). *Statistics on Persons in Denmark – A register-based Statistical System*. Luxembourg: Eurostat.

Statistics Finland (2004). *Use of Registers and Administrative Data Sources for Statistical Purposes – Best Practices in Statistics Finland.* Handbook 45, Statistics Finland, Helsinki, Finland.

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 Wallgren, B. (2007). *Register-based Statistics - Administrative Data for Statistical Purposes.* John Wiley & Sons, Ltd.

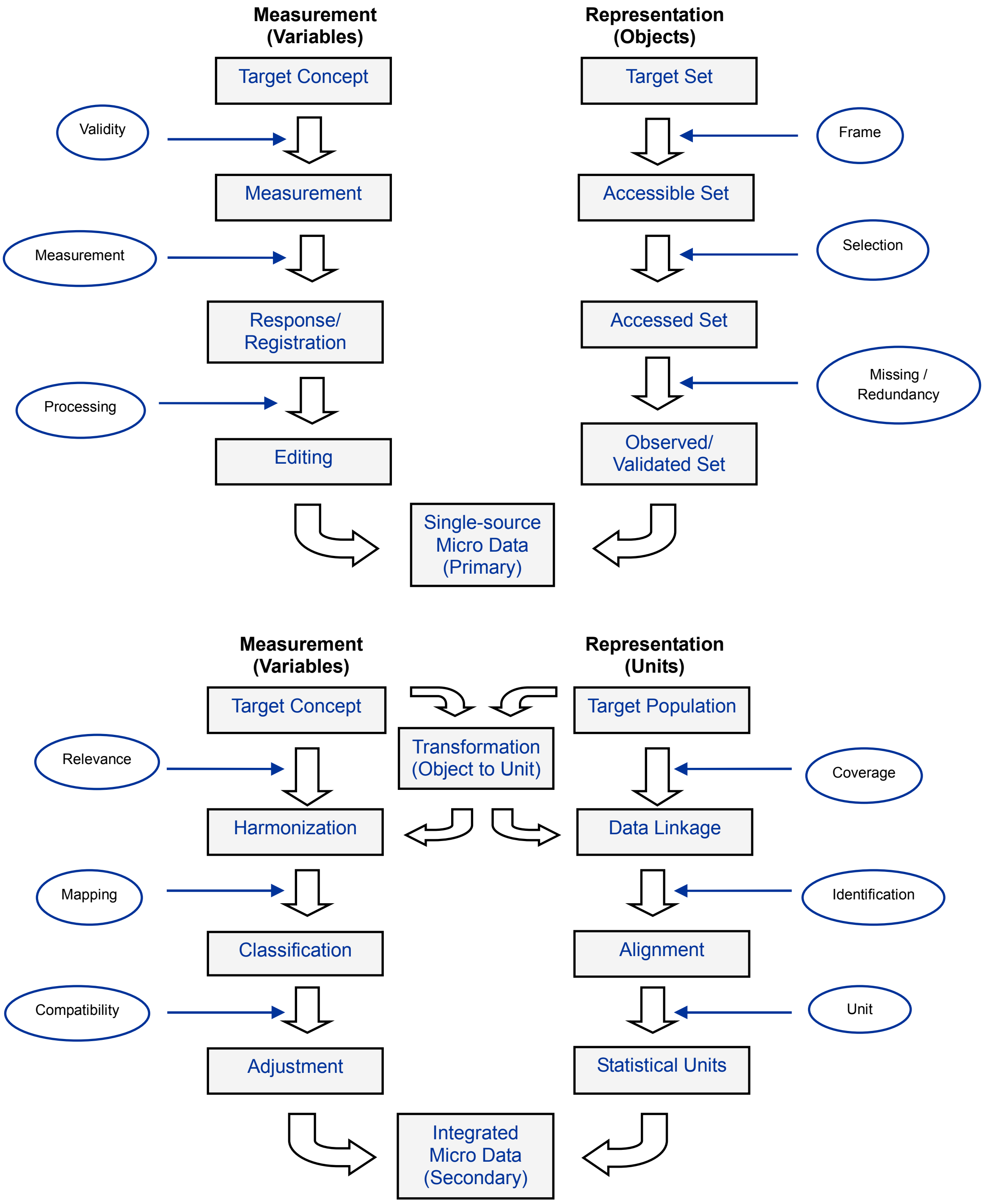
Zhang, L.-C. (2010). A Unit-error Theory for Register-based Household Statistics. To appear in *Journal of Official Statistics*.

Zhang, L.-C. and Fosen, J. (2010). Assessment of Uncertainty in Register-based Small Area Means of a Binary Variable. To appear in *Journal of Indian Society of Agricultural Statistics.*

RÉSUMÉ (ABSTRACT)

For some decades now administrative registers have been an important data source for official statistics alongside survey sampling and population census. Reduction of response burden, long-term cost efficiency as well as potentials for detailed spatial-demographic and longitudinal statistics are some of the major advantages associated with the use of administrative registers. However, administrative registers certainly do not provide perfect statistical data. Moreover, often, one can only make statistical uses out of an administrative register after it has been combined with data from other sources. At the present stage, there is still clearly a lack of statistical theories for assessing the quality of such register-based statistics. We believe that a key issue here is the conceptualization and measurement of the statistical accuracy in administrative register data, which will enable us to apply rigorous statistical concepts such as bias, variance, efficiency and consistency, e.g. as one is able to do when it comes to survey sampling. In this paper we present a general outlook to the various potential errors for multiple-source integrated statistical data, of which the administrative data are a large part. A shared understanding of these error sources shall hopefully help us to collocate and coordinate efforts in future research and development.

Figure 1. Two-phase (primary and secondary) life cycle of statistical data from a quality perspective. Stage of data production (square); Accuracy concept and source of error (oval).


Measurement (Variables)
Representation (Objects)
Target Concept
Target Set
Validity
Frame
Measurement
Accessible Set
Measurement
Selection
Response/ Registration
Accessed Set
Processing
Missing / Redundancy
Editing
Observed/ Validated Set
Single-source Micro Data (Primary)
Measurement (Variables)
Representation (Units)
Target Concept
Target Population
Transformation (Object to Unit)
Relevance
Coverage
Harmonization
Data Linkage
Mapping
Identification
Classification
Alignment
Compatibility
Unit
Adjustment
Statistical Units
Integrated Micro Data (Secondary)