ESSnet KOMUSO

Quality in Multisource Statistics

http://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics_en

Specific Grant Agreement No 1 (SGA-1)

Framework Partnership Agreement Number 07112.2015.003-2015.226

Specific Grant Agreement Number 07112.2015.015-2015.705

Work Package 2

Quality measures and indicators of frames for social statistics

Version 2017-04-21

Prepared by: Li-Chun Zhang (SSB, Norway)

With contributions from Johan Fosen (SSB, Norway), Ildikó Szűcs (KSH, Hungary),
Christoph Waldner (STAT, Austria), Giovanna Brancato (Istat, Italy),
Ellen O'Connor (CSO, Ireland), Danuté Krapavickaité (LS, Lithuania),
and Regin Reinert, Peter Stoltze (DST, Denmark)

ESSnet co-ordinator: Niels Ploug (DST, Denmark), email npl@dst.dk, telephone +45 3917 3951

Table of Contents

Quality measures and indicators of frames for social statistics 1
1. Introduction 3
2. Frame and Frame errors 6
2.1 Traditional definition of frame 6
2.2 Enhanced population dataset (EPD) 7
2.2.1 Progressive frame data 8
2.2.2 Multisource frame data and alignment of multiple frame units 8
2.3 Frame Errors 10
2.3.1 10
2.3.2 Alignment error and unit error 11
2.3.3 Contact information error 12
2.4 Quantifying frame errors 12
2.5 Some comparisons between frames for business and social statistics 16
3. Quality assessment: items, approaches and methods 20
3.1 Specification 20
3.2 Assessment approaches 23
3.3 Methods 26
4. Applications 34
4.1 On-going survey 34
4.2 Modelling 43
4.3 Hybrid approach 48
4.4 Diagnostics 51
A diagnostic for late registration of emigration (PM1u) 53
A diagnostic for non-deregistration of emigration (CM1o) 53
Results 61
Appendix A. Istat Standard Survey Outcome Chart and Quality Indicators 64
References 66

1. Introduction

The current report presents the frame accuracy assessment framework developed in Work Package 2 (WP2) of *Komuso* (ESSnet on quality of multisource statistics) within the first Specific Grant Agreement, and illustrates its applications using real data in 6 countries. Relevant content from previous Deliverables in 2016 has been incorporated. The presentation is self-contained; it is arranged in a logical order and not strictly the order in which the various tasks were completed.

The aim of WP2 is to develop a theoretical framework for the assessment of frame quality, where the frame is constructed based on *multiple* input sources. The purpose is to provide the methodological basis for a quality guideline to the frames for social statistics. Our scope and perspective in WP2 are further clarified as follows.

To start with, the perspective of WP2 needs to be delineated from that of WP3, which aims to produce quality measures of multisource statistical outputs. While the quality of *any* statistical output necessarily depends on the quality of the relevant frame, it is beyond the scope of WP2 to spell out the consequences of frame errors on the quality of these outputs.

- For example, various population totals based on a frame of persons may be used to calibrate the sample weights in the Labour Force Survey (LFS). The frame over- and under-coverage errors will then affect e.g. the LFS Employment total estimator. While it will clearly be of interest to assess the extent to which the frame errors affect the final LFS estimators, in WP2 we must confine ourselves to measure *only* the coverage errors of the frame itself, but not e.g. the mean squared error of the Employment total estimator that depends on it.
- As another example, suppose an Employer/Employee Register (EER) is used to produce Employment statistics, then the present framework from WP2 is applicable insofar as the target population of employed persons is enumerated directly from the EER. However, suppose now that the EER is combined with a Population Register (PR), where the target population of persons between 16 and 74 is derived from the PR, and the EER provides only attribute data in order to classify whether a person is employed or not. Then it is possible to consider the quality of the resulting Employment statistics under the framework developed in WP3, and restrict frame accuracy assessment to the PR. Or, one may prefer to consider (employed, not employed) as an additional domain classification variable to the PR obtained from the EER, and assess the accuracy of Employment Statistics as that of the corresponding domain classification of the 'enhanced' population dataset under the present framework from WP2.

In short, ***we shall treat the frame as a multisource statistical product itself, and assess its quality as such***. This secures the relevance of the present framework to register-based census-like statistics, which is a key concern for WP2 of Komuso. The accuracy items, the assessment approaches and the associated methods reviewed and developed in this report are all directly applicable to register-based population statistics, as will be illustrated in the applications to be described later. It should be notice that, on the one hand, certain situations will be treated in WP3, where statistical outputs are affected by frame coverage errors; on the other hand, the coming quality guideline to frames (planned for the next project period of Komuso) will address the uses of frame and the consequences of frame errors in statistical production, where it is natural to take up such matters.

Next, while it is possible to speak of various quality dimensions such as relevance, accuracy, timeliness and comparability (e.g. Eurostat, 2011; Colledge, 1995), our focus in WP2 is ***accuracy***. One reason for this is that quality indicators for some of the other dimensions are more readily devisable. For example, whether the preparation of a relevant frame is delayed for a statistical process in a given situation is easily recognised and quantified. Moreover, the importance and means to prevent this from happening is more appropriately the topics of the coming quality guideline. Another reason is that the effects of many other quality issues, including the input data and metadata quality, will be assimilated in the accuracy assessment framework we develop. For example, internal incoherence of frame data or incompatibility between the input sources (including definitional difference, potential record linkage errors, etc.) must necessarily manifest themselves as certain frame errors, the extent of which can then be captured in frame accuracy assessment. Moreover, a frame based on multiple sources by definition must aim to integrate *all* relevant frame data in a statistical system. For example, it clearly would have been ineffective and suboptimal not to combine an address register with a person register when both are available. Thus, from a multisource perspective to statistical production, the compatibility between these two registers should and could be assessed and quantified in terms of the accuracy of a ***multisource integrated frame***.

Finally, in order to prepare the necessary statistical scientific foundation for the coming quality guideline to frames, a framework for accuracy assessment needs to be methodological and rigorous in nature. While a fit-for-purpose interpretation of quality will be important and necessary for the quality guideline, it is not - nor can it be - the ideal we strive for here. We shall therefore maintain the distinction between ***quality measure (QM)*** and ***quality indicator (QI)***, where only an estimate with its own quantified uncertainty (e.g. bias and variance) can provide a QM, whereas an estimate by itself can only be a QI. Of course, as we shall explain, given a parameter to be estimated, say, the total frame over-count, it may be difficult to obtain an acceptable quality measure due to the limitation of the resources and methods available. Quality indicators are thus both important and useful in practice. However, scientifically speaking, one cannot be content with indicators alone.

The rest of the report is organised in three chapters.

Chapter 2 covers the definitions of frame and frame errors. We extend the traditional definition of *sampling frame* to the situation where multiple sources of relevant frame data in a statistical system are combined and processed, to create an ***enhanced population dataset (EPD)*** as a multisource integrated frame for social statistics. The EPD does not only serve as a sampling frame but, more importantly, it will be directly used for delineating the target statistical populations in social statistics. The various existing frames and those that are currently under development at a number of European countries are all special cases of EPD. The formulation increases the relevance of the quality framework to register-based census-like statistics, where e.g. population and housing statistics are produced based on integrating data of different types of units. Consequently, we have found it necessary to extend the traditional classification of frame errors, in the context of frame as a multisource statistical product. The various frame errors will be motivated, explained and illustrated.

In Chapter 3 we propose *a list of 17 items* for frame quality assessment. Instead of laying down absolute thresholds of 'minimum quality', such as frame under-coverage error rate lower than 1%, we propose these as a *minimum* set of items for frame quality assessment. For each item one can obtain either quality measures or quality indicators, depending on the available data, resource and

methods. The 17 items are divided into two parts: A-list and B-list. The A-list contains items for which there exist established methods and past experiences for producing quality measures, whereas the B-list contains those that require further development and testing of readily applicable methods. We review and summarise the various approaches to frame quality assessment and the associated methods. The most readily applicable methods will be described in more details, including also some proposals for the B-list items. Notice that even though established quality measure methods exist for all the items in the A-list, some of them may be costly to implement, such as a coverage survey. As will be explained, we identified three approaches as the most promising for developing regular means of frame quality assessment at a lower cost, which are design-based methods that are applicable to on-going surveys, modelling and diagnostic methods based on multiple registers.

Finally, in Chapter 4, we present applications of the three aforementioned approaches to frame quality assessment. In addition, depending on the situation and the data that are available, we illustrate how they can be combined to yield what may be referred to as the hybrid approach. For each approach, the quality items to be assessed and the additional background and details of the associated methods will be described, and the results will be summarised and commented. A short appraisal of the approach will be given at the end of each relevant section or subsection.

2. Frame and Frame errors

2.1 Traditional definition of frame

We adopt the following as the traditional definition of frame:

Frame is any list, material or device that delimits, identifies, and allows access to the elements of the target (survey) population.

This is a simple summary of the ideas expounded in the two central references in the literature:

- Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. Wiley. (Chap. 3-5)
- Wright, T. and Tsao, H.J. (1983). A Frame on Frames: An Annotated Bibliography. In Tommy Wright, ed. *Statistical Methods and the Improvement of Data Quality.* Orlando, FL: Academic, 25-72.

Lessler and Kalsbeek (1992) list in addition a number of concepts the definition is meant to encompass. We summarise as follows.

1. The ***target (survey) population*** is finite and consists of ***population elements*** that are identifiable or accessible via the frame. The target population elements are also referred to as statistical units or observation units. The frame consists of ***frame units***. Some mechanism must exist for connecting the target population and the frame, i.e. it must be possible to *access* the population elements via the frame units. The ***survey (or study) population*** is the subset of target population, which is *accessible* via the frame. The discrepancy between the frame, the study population and the target population constitutes *frame errors*.
2. The frame units do not necessarily coincide with the population elements. More than one type of connection (or linkage) may exist between population elements and frame units. Non-existent, superfluous, or mistaken connections give rise to ***frame coverage errors***.
3. Frame data include ***classification variables*** for all the frame units. The target population can be partitioned into a set of ***domains*** according to the chosen classification variables, such as age, sex and region. Or, using classification variables, the target population may itself be defined as a domain of the population that is accessible using the entire frame, such as the permanent residents (i.e. target population) among all the persons that can be found in a population register. Domain classification errors are clearly a matter for frame quality assessment.
4. Access to population elements is needed in survey sampling, for which the frame must contain ***contact information*** of the frame units. Contact information errors are a concern for sampling frames. It may also be necessary to delineate the target population for estimation or analysis, with or without sampling. Frame quality for both sampling and estimation (or analysis) are relevant here. Notice that concerning sampling frame, one may distinguish between *direct* and *indirect* access to the target population. In the former case, such as when sampling persons from a population register, all the frame errors (to be described later) apply directly. However, in the latter case, the matter can quickly turn into one of estimation under imperfect frame(s), rather than assessing the quality of the frame *per se*. For example, when an address frame is used for sampling of households, the frame over-coverage error will need to be accounted for in the estimation directly, which in return will provide an assessment of the corresponding frame error as a by-product. As another example, the relevant domain classification variable of households may not exist in the address frame at all, in which case potential domain classification error is

more a matter of nonresponse or measurement error than errors of the frame. Such issues related to the uses of a sampling frame will not be discussed in this report.

Lessler and Kalsbeek (1992) identify the following problems related to frame errors:

- Missing population elements or under-coverage, which arises when some elements cannot be accessed from the available frame units.
- Inclusion of non-population (i.e. *erroneous*) elements, which occurs when some frame units correspond to elements outside of the target population, including when they cease to exist.
- Unrecognised multiplicity or duplicate listing, when two or more frame units refer to the same population elements, without this being recognised in the frame.
- Failure to account for clustering, when the inclusion probability of a population element is wrongly calculated for not taking into account the varying cluster sizes, which is more of an error in the use of frame than an error of the frame itself.
- Incorrect auxiliary information that provide classification variables.
- Inaccessible elements due to errors of contact information.

Below we extend the above traditional definition of frame and classification of frame errors in the context of multisource social statistics.

2.2 Enhanced population dataset (EPD)

In WP2 we are particularly concerned about frames constructed using data from administrative sources, possibly in combination with traditional fieldwork-based methods. As a frame based on multiple sources for social statistics one may envisage an ***enhanced population dataset (EPD)***, which consists of all the relevant available frame units including person, household, dwelling or address, etc. their classification variables and contact information. ***Central population register (CPR)*** that exists in a number of European countries are a special case of EPD. Linking patient and tax registers can yield an EPD; linking population and address registers another. The different EPDs differ in their content and quality. An EPD is said to be ***integrated*** if it provides a single source of available frame data for *all* social statistics. That is, if all the frame data are brought together in one place, regardless of how rich the content is or how accurate the data are. An integrated, rich and accurate EPD clearly represents the ideal state-to-be, as far as frames for social statistics are concerned.

Two particular error sources of an EPD are due to the ***progressive*** and ***multisource*** nature of the input frame data. For better appreciation of the relevant issues we draw comparisons to two other types of frame.

I) ***Frames based on census***, further exemplified by the area frame that is constructed in order to carry out the census, and the list frame that is the result of the census. In either case the data arises from a single source (or operation), and the frame will remain static until the next root-and-branch update. (Notice that in cases where a census-based frame is continuously updated after the census, i.e. by incorporate vital events, one would be talking about a multisource frame instead.)

II) ***Business Register (BR),*** which has as its main source the administration of *legal units,* but is typically supplemented with data on tax, employee benefits, commerce, etc. The frame data are collected from multiple sources, which are or can be updated more or less continuously, say, on a daily basis. Formally speaking, there are no essential conceptual differences between the BR and an

EPD. However, the two differ when it comes to the most prominent errors, as we will discuss later. It may also be noted that the deployment of BR is more common than EPD in Europe.

2.2.1 Progressive frame data

As mentioned above, the frames based on census are static and do not change until the next root-and-branch update. In contrast, both the BR and EPD can be updated on a daily basis, in principle as well as in practice. However, due to unavoidable delays and mistakes in the input administrative sources, both display *progressiveness* in the sense that the frame for a fixed reference time point, denoted by *t,* will differ depending on the time point the frame is constructed, denoted by *s* for $s \geq t$.

Administrative data are often generated by events, such as birth, death, change of address or civil status, etc. At least two time points are essential for each event that makes its way into the frame, or any statistical register: (1) the ***reference*** time point, i.e. when the event actually occurred, and (2) the ***registration*** time point, i.e. when the event was recorded in the administrative source. For events that are mistakenly reported or registered to start with, there may be additional time points when such mistakes are corrected. By delay in the administrative data, we mean the lag between the reference time point and the registration time point, provided which the frame data will be progressive in the sense defined above, despite the daily updating at the statistical office of the available administrative data.

Progressive frame data can affect all the frame errors to be classified. For instance, in survey sampling, there may be errors of contact information due to lack of updating, which creates difficulties for data collection. Moreover, the frame for sampling will generally differ from the frame for estimation, due to the time that elapses between the corresponding statistical processes, even though the statistical reference time point is fixed and the same on both occasions. The differences between the two frames can be articulated both in terms of coverage and domain classification errors, as we explain later on. Finally, because progressiveness is an important aspect of the input administrative data, we shall propose to assess the progressiveness of frame data directly, in addition to the various frame errors.

Remark. The progressiveness of frame data can have important consequences if they are explicitly acknowledged. For instance, the inclusion probability of the sample units can never be the sole basis for statistical inference in practice. For purely register-based statistics, an estimation perspective on the tabulated population quantities becomes necessary when it comes to quality assessment, even if explicit and reliable adjustments for the progressive data may be difficult or costly to achieve.

2.2.2 Multisource frame data and alignment of multiple frame units

An immediate consequence of multisource frame data in an EPD is the need of merging, i.e. record linkage, of two or more datasets. Unless there exists a unique identifier across the datasets, which allows for exact and perfect matching of the records, linkage errors may be unavoidable. See e.g. Harron et al. (2016) for a recent survey of data linkage methodology. While linkage error does not directly constitute a frame error itself, it can potentially cause all kinds of frame errors.

Next, an EPD can have multiple types of frame units. The “linkage” between any two sets of entities refers to the mapping between them. *A priori* one can envisage 4 possible linkages, namely, one-one, one-many, many-one, and many-many. Sirken and Levy (1974) refers to the number of “links” between an element and the frame as the *multiplicity* of the element. Two additional cases are one-

nil and nil-one, where the presence of one-nil "linkages" between frame units and population elements entails *erroneous* frame units, and that of nil-one "linkages" frame *under-coverage*.

- As an example of one-many linkages, consider an area (or cluster) sampling frame, where a frame unit is a building block that entails a number of population elements that are the households. The multiplicity of an element is 1 provided one-many linkage.
- As an example of many-one linkage let the frame units be parents and the population elements be children, where a child is surveyed if either parent is sampled. The multiplicity of a child depends on the sampling design. It is 2 for any child with two living parents, and if all the parents are sampled as separate individuals. Or, it is 2 only for any child with separated (or divorced) parents if the parents are sampled in households, i.e. household is the frame unit in this case.
- As an example of many-many linkage, suppose all the offspring of a sampled parent are to be included in the survey, and the parents are sampled in households.

As an extension of this standard point-of-view regarding frame units and population elements, consider the situation where *there are multiple types of frame units.* Suppose, for a register-based census, a **population dataset (PD)** is linked to the address register by a person's home address in the PD, and to the employee benefits register and higher education register by the personal identification number (PIN), and the postal address register by the name and postcode. The different types of frame units are person, home address, postal address, and business (or study) address. Consider a student who studies at a university located in a city different than the home address in the PD. The home and business addresses will differ, as well as possibly the postal and home addresses. All the frame address data are relevant for different purposes, ranging from household statistics, to statistics for municipality service planning, to statistics for national transport reengineering, etc. There arises thus the necessity to clarify the "linkage" between the different frame units themselves.

Table 1. Illustration of an alignment table

BU (person)	CU-1 (home address)	CU-2 (business address)	…	Telephone Nr.	…
Adam Smith	Smith-SO19xxx	-		123456 132415	
Eva Hanford	Smith-SO19xxx	Highfield-SO17xxx		324151	
Mark Smith	Smith-SO19xxx	London-WS5Dxxx		-	
Alan Smith	Smith-WC1Exxx	London-WC1Exxx		654312	
Sarah Sommers	Sommers-L17xxx	London-NQ6Axxx		-	
…	…		…	…	…

The "linkage" or association between the various frame units can be summarised using the alignment table (Zhang, 2012), as illustrated in Table 1. Here BU stands for *base unit,* and CU *composite unit.* Each CU may consist of one or several BUs, but it can never "cut across" a BU. While persons often are the BUs in social statistics, other possibilities do exist. For example, the EDAG (a big administrative dataset in Norway) contains all the payments related to contractual works as well as social and health benefits. The data are delivered to Statistics Norway monthly, and are initially organised as "reports" from different business and government agencies. For the purpose of harmonised Wage and Employment statistics, it is prudent to set the BU at a lower level than persons, because a person can have multiple jobs, which are the statistical units for which wage is

defined. Common examples of CUs in social statistics are family, household, dwelling, various addresses, school, employer, etc. The relationship between the different CUs, often non-nested, can be articulated via their relationships to the BUs.

Remark. Traditionally the concept of multiplicity has been used to characterise the access from frame units to population elements. However, as we have explained, the mapping between the different frame units is important as well. This is obvious when an EPD is used to define the target population directly. To understand the relevance and additional subtlety for sampling, consider the situation where persons are sampled in clusters. Let first the CPR address, i.e. a composite frame unit, be the cluster. Then, under-coverage may occur for a student who has an outdated home address in the CPR, but is actually living at a student home whose address does not appear in the CPR as private resident address. However, let now the CPR family, i.e. cohabiting parents and children, be the cluster, then under-coverage can be avoided for that same student, as long as one is instructed to follow up all the cluster members. This can be articulated in that the alignment between person (BU) and CPR family (CU-2) is *more reliable (or accurate)* than the alignment between person (BU) and CPR address (CU-1).

Remark. Many-many "linkages" between two types of composite frame units are 'decomposed' into two sets of many-one "linkages" between each type of CU and the BU. This is more practical than having to update the many-many "linkages" directly, especially because an EPD has to reconcile conflicting information from multiple input sources. For example, one may find different addresses for a person in the CPR, the TV license register, the patient register, etc. To make a selection of only one of them to keep in the EPD entails potential alignment errors. An alternative is to keep all of them under different names in an EPD, and to make a choice depending on the purpose when the data are actually used. In any case, maintenance and updating of alignment between different frame units should be treated as a regular and important statistical process related to the EPD.

Remark. There is no absolute distinction between a composite unit and a classification variable. For example, one may consider male and female to be two CUs, each having a many-one relationship with the base unit person. The distinction between a CU and a classification variable, or that between alignment and domain classification errors, is ultimately a choice based on ease and usage.

2.3 Frame Errors

Here we classify and explain five types of frame errors that emerge from the above:

- ***Coverage error*** due to missing, erroneous and duplicated frame units
- ***Domain classification error*** of frame units
- ***Alignment error*** between different types of frame units
- ***Unit error*** of composite frame units
- ***Contact information error*** of frame units

2.3.1 Coverage error and domain classification error

Frame *under-coverage* is the case if a population element is "missing" in the frame, i.e. if there exists no frame unit that provides access to that population element. Frame *over-coverage* is the case if it entails duplicated, non-existent or out-of-scope elements. It is common to distinguish between two main types of over-coverage: *duplicate listing* and *erroneous enumeration*. Duplicate listing refers to the case where a population element can be accessed via at least two frame units, without this being

recognised or planned as such in the frame. For example, two or more records in a patient register may refer to the same person, where each record is supposed to refer to a distinct person. A variation of the situation is when it is not feasible to determine the multiplicity of some population elements, referred to as the problem of unrecognised multiplicity, i.e. duplicate listing is known to be potentially the case for some population elements, without it being feasible to clarify or remedy the situation. Erroneous enumeration is the case when some frame units seemingly provide access to elements that are either non-existent or out-of-scope. Notice that it is possible to have duplicate listing of erroneous enumerations.

For sampling, estimation and dissemination the frame and study population that is accessible from the frame may be partitioned into *domains*, denoted by *d = 1, …, D*, according to some classification variables that are known for all the frame units. *Domain classification* is of critical importance to frame accuracy, just like coverage of the population as a whole. Domain misclassifications error is referred to as incorrect auxiliary information in Lessler & Kalsbeek (1992). We prefer to distinguish between domain classification variables and other auxiliary variables, although whether or not a variable is considered a domain classification variable may depend on the situation. For example, take the variable household size. When used as an auxiliary variable for sampling and estimation, the variable is important for the efficiency, but inference may still be valid despite errors in this measure. However, when it is used for domain classification, an error may result in a coverage error, either for the whole target population or the domain partition of it.

Notice that it is only unambiguous to speak of potential domain misclassification between frame units and population elements, or any two types of units, provided they are in a one-one relationship with each other. For example, a person may be located in City A according to the frame but actually lives in City B and therefore belongs to another domain. Notice that this is not a linkage error between frame units and population elements, since there is no question here that it is the same person in the frame and in the target population.

2.3.2 Alignment error and unit error

We introduce these as additional types of frame errors to the traditional classification. Despite the relevance of composite units in frames for social statistics, the alignment error and the unit error seem to have received only insufficient attention in the existing literature. We find it necessary to include them with regard to an EPD that is made possible from combining multiple data sources, e.g. with respect to register-based census-like statistics.

Alignment error between base units and a type of composite units is the case, if one or several base units are wrongly associated with a composite unit. As explained before, alignment between multiple frame units is important for an EPD in the context of multisource statistics. Alignment error is closely related to *unit error,* which is the case if there exist wrongly delineated statistical units. The former can be the cause of the latter. For example, two persons may wrongly form a household (as statistical unit) because of alignment errors between person and address. Nevertheless, one must distinguish unit error from alignment error, because sometimes it is necessary to *construct* a type of frame units based on the data one has, in which case unit errors may be unavoidable even if all the input data are error-free, as explained below.

Consider register-based household. Provided perfect dwelling registration in the CPR and perfect Dwelling Register (DR) and perfect linkage between the CPR and the DR, it is conceivable that one

may define a *dwelling household* to consist of all the persons registered at the same dwelling, and thereby obtain a perfect frame of dwelling households. Under the present framework, we do *not* consider such a dwelling household to be a constructed type of frame unit, precisely because it could be obtained directly from perfect input data. The perfection is simply another way of saying that there are no alignment errors. An example of constructed frame units in this setting is *living household*, which does not have to include everyone registered at the same dwelling, nor be limited to these. Living household is the ideal statistical unit for household income or expenditure statistics. Errors in a constructed living household is the case if two persons in different living households are placed in the same constructed living household, or if two persons in the same living households are placed in different constructed living households. An example may be a man and a woman both of 25 years old and living in the same apartment. They can be a couple or just flatmates: either way the constructed living household(s) may be mistaken, despite there are no alignment errors in this case. We refer to such errors as unit errors (Zhang, 2011). Alternatively, the situation may be conceptualised as *simultaneous* over- and under-coverage as discussed in Zhang (2012).

2.3.3 Contact information error

Contact information error is relevant when a frame is used for sampling, in order to collect data from or about the sampled population elements. Consider first address, which is an important piece of contact information for personal interview. As mentioned before, an address can also be regarded as a composite unit, in which case it is possible to treat this form of contact information error as an alignment error. Consider next telephone numbers as the contact information for telephone interview. Landline telephone number can possibly be treated as a “composite unit”, which can be connected to one or several persons as the base unit. On the other hand, it may be possible to treat mobile telephone number as a “base unit” in relation to person as a “composite unit”, i.e. a person may have more than one mobile phone. Similarly for email address, which is gaining increasing importance. Finally, the infrastructure of e-dialogue between the residents and the government is being implemented in many European countries, which can potentially become a more reliable and useful source of contact information. Despite formally it seems possible to classify contact information error as various forms of alignment error, we have chosen to retain it as a separate category because, apart from address, the contact information is only used to provide access to population elements, rather than serving as a frame unit or population element in its own right.

2.4 Quantifying frame errors

In this section we describe some basic means for quantifying the frame errors, both at the aggregated (macro) level and at the entity (micro) level. These provide the basis for the proposed quality measures and indicators later on.

2.4.1 Coverage and domain classification error

Table 2 below quantifies jointly under-coverage (missing frame units), over-coverage (erroneous frame units) and domain classification errors, in the case where the frame units and population elements are in a one-one relationship. The counts (N_{10}, N_{20}, ..., N_{H0}) are the numbers of erroneous units in the different frame domains, which may be considered as instances of one-nil linkage between the frame and the population. The counts (N_{01}, N_{02}, ..., N_{0H}) refer to the elements in the different population domains that are missing in the frame, which may be considered as instances of nil-one linkage between the frame and the population. The count N_{00} is 0 by definition in the present context, since nil-nil linkage does not exist.

At the macro level the domain classification errors can be quantified as $\sum_{k\neq h}^{H} N_{hk}$, the over-coverage error as $\sum_{h=1}^{H} N_{h0}$ and the undercoverage as $\sum_{h=1}^{H} N_{0h}$.

Table 2. Coverage and domain classification errors

Frame Domain Classification	Population Domain Classification				Erroneous Frame Unit
	1	2	…	H	
1	N_{11}	N_{12}		N_{1H}	N_{10}
2	N_{21}	N_{22}		N_{2H}	N_{20}
…					
H	N_{H1}	N_{H2}		N_{HH}	N_{H0}
Under-Coverage	N_{01}	N_{02}		N_{0H}	N_{00} (=0)

For example, let *h = 1, ..., H* denote the different household types such as (1) single-person household, (2) single parent with children, (3) couple without children, (4) couple with children, etc. The unit of Table 2 here must be person (i.e. base unit) -- it cannot be household because different households (i.e. composite units) cannot always be put in one-one correspondence. Then, N_{11} is the number of persons living in single-person households in the population that is correctly identified in the frame, and N_{21} is the number of persons living in single-person households in the population that is identified as living in households of single-parent with children in the frame, etc. Whereas $\sum_{k\neq 1}^{H} N_{1k} = N_{12} + \cdots + N_{1H}$ is the number of persons in the population who are wrongly identified in the frame as living in single-person households, etc.

Remark. The same table can be used to represent the Census and Census coverage survey (CCS) data. Then, N_{00} would represent the number of population elements that are missing in both Census and CCS, which in this case is assumed not to be equal to 0 and is the purpose of CCS.

Table 2*. Progressive errors of coverage and domain classification

Frame Domain Classification at t_1	Frame Domain Classification at t_2				Not in frame at t_2
	1	2	…	H	
1	N_{11}	N_{12}		N_{1H}	N_{10}
2	N_{21}	N_{22}		N_{2H}	N_{20}
…					
H	N_{H1}	N_{H2}		N_{HH}	N_{H0}
Not in frame at t_1	N_{01}	N_{02}		N_{0H}	N_{00}

Table 2 can be used to quantify the same types of errors due to progressive frame data. One only needs to modify the row and column headings as shown in Table 2*. Let the target population have reference time point *t.* Let $t \leq t_1 < t_2$, where t_1 and t_2 are two different measurement time points, aiming at the same target population. The count N_{00} is again 0 by definition. Notice that here all the discrepancies (with $i \neq j$) are relative between two frames that refer to the same target population, but *not* absolute between a frame and its target population, i.e. they are different to those in Table 2 both in values and interpretation.

Table 2 and 2* quantify the aggregated errors. On the unit level, multinomial probabilities can be used to describe the heterogeneity of the errors. Heterogeneity here means that the multinomial probabilities may vary for different frame units. (Otherwise, macro-level aggregated quantification would have sufficed.) Thus, for frame unit *i,* let

$$(p_{i1}, p_{i2}, ..., p_{iH}, p_{i0}) = E(\delta_{i1}, \delta_{i2}, ..., \delta_{iH}, \delta_{i0})$$

be the probability that it will contribute to the corresponding column in Table 2, where $\delta_{ih} = 1$ if the unit belongs to population domain *h*, and 0 otherwise, including the case of $h = 0$, i.e. erroneous inclusion frame units. Whereas, in the reverse direction, for population element *j*, let

$$(q_{1j}, q_{2j}, ..., q_{Hj}, q_{0j}) = E(\delta_{1j}, \delta_{2j}, ..., \delta_{Hj}, \delta_{0j})$$

be the probability that it contributes to the corresponding row in Table 2, where $\delta_{hj} = 1$ if the element belongs to frame domain *h*, and 0 otherwise, including the case of $h = 0$, i.e. missing in the frame.

2.4.2 Alignment error: Macro level

Perfectly accurate alignment between base units (BU) and composite units (CU) can be summarised as in Table 3 below, where N_{gh} BUs in domain-*g* are aligned with M_{gh} CUs in domain-*h*. Notice that since the base and coposite units are of different types, so are in general their respective domain classifications. For instance, one may have so-and-so many persons in a certain age-sex group (domain of BU) aligned with thus-and-so many households without children (domain of CU).

Table 3. Summary of alignment between BU and CU

	Composite Unit Classification				
Base Unit Classification	1	2	…	H	Base Unit Total
1	(N_{11}, M_{11})	(N_{12}, M_{12})		(N_{1H}, M_{1H})	N_{1+}
2	(N_{21}, M_{21})	(N_{22}, M_{22})		(N_{2H}, M_{2H})	N_{2+}
…					
G	(N_{G1}, M_{G1})	(N_{G2}, M_{G2})		(N_{GH}, M_{GH})	N_{G+}
Composite Unit Total	M_{+1}	M_{+2}		M_{+H}	N

It is of course possible and sometimes helpful to split Table 3 in two: one of N_{gh} and the other of M_{gh}. However, wherever ambiguity is not an issue, one can put the two in one (as here) to save space.

There is an asymmetry in Table 3 that is worth noting. It has fixed row sum (of N_{gh}) in row *g,* which is the number of BUs in domain-*g* that is fixed. Meanwhile, the column sum M_{+h} is not fixed, because a CU will be counted more than once provided it consists of BUs that belong to different domains.

In cases where the alignment between BU and CU is not perfectly accurate, Table 3 will not be enough because it represents only how the BUs are *actually* aligned with the CUs in the frame, but not how they *should be* aligned. Some variations are easily devised.

1) Let $N_{gh;0}$ be the *expected* number of type-*g* BUs that are aligned with type-*h* CUs, and $M_{gh;0}$ be the *expected* number of these type-*h* CUs. Denote the corresponding table by A_0.
2) Let $N_{gh;t}$ be the number of type-*g* BUs that are known (or considered) to be *correctly* aligned with type-*h* CUs, and $M_{gh;t}$ be the number of these type-*h* CUs. Denote the corresponding table by A_t.

Now, denote by A the actual table of alignment between BU and CU in the frame, then we have

$$E(A - A_t) = A_0 - A_t$$

It is possible to summarise the alignment between two different types of CUs in terms of the BUs. For example, in Table 3* below the relationship between CU-I and CU-II is summarised in terms of the

BUs, where N_{gh} stands for the number of BUs that belong to CU-I domain-*g* and CU-II domain-*h*. For example, a person (BU) may belong to a single-person family (CU-I domain) and a cohabitant without children household (CU-II domain), or a person may belong to a family with children (CU-I domain) and a single-person household (CU-II domain). Insofar as each BU must be associated with one and only CU of a given type, the table has fixed margins both in row and column. Similarly as above, one may distinguish between the observed and the expected (or true) values of Table 3*.

Table 3*. Summary of alignment between CUs in terms of BUs

Composite Unit-I Classification	Composite Unit-II Classification				Base Unit Total
	1	2	…	H	
1	N_{11}	N_{12}		N_{1H}	N_{1+}
2	N_{21}	N_{22}		N_{2H}	N_{2+}
…					
G	N_{G1}	N_{G2}		N_{GH}	N_{G+}
Base Unit Total	N_{+1}	N_{+2}		N_{+H}	N

As another example of macro level summary, in Table 3' the relationship between CU-I and CU-II is described in terms of the *parent units (PUs)*, where a parent unit is such that both types of units are nested in it. For example, let CU-I be (dwelling) household, and CU-II dwelling. A PU for both can be building, since both dwellings and households are nested in buildings, but a building can never divide a dwelling or a household. The number P_{gh} in Table 3' stands for the number of PUs that are associated *g* households and *h* dwellings. For instance, P_{11} is the number of buildings with one household and one dwelling, and P_{12} the number of buildings with one household and two dwellings – one of which is presumably not permanently occupied if all the alignments are correct, etc. The table neither has fixed row nor column margins, and the total P_{++} is in general not the total number of buildings. But it could have been if one had included a zero-count row and column in the table. Again, as with Table 3 previously, Table 3' can also be adapted to accommodate the observed table vs. the true table, or the expected table vs. the true table.

Table 3'. Summary of alignment between CUs in terms of PUs

Number of Composite Unit-I	Number of Composite Unit-II				Parent Unit Total
	1	2	…	H	
1	P_{11}	P_{12}		P_{1H}	P_{1+}
2	P_{21}	P_{22}		P_{2H}	P_{2+}
…					
G	P_{G1}	P_{G2}		P_{GH}	P_{G+}
Parent Unit Total	P_{+1}	P_{+2}		P_{+H}	P_{++}

2.4.3 Alignment error: Micro level

Heterogeneous alignment error on the micro level can be quantified in terms of the distribution of the allocation matrix of BUs to CUs (Zhang, 2011). Again, heterogeneity here means that the distribution of the allocation matrix may vary in different parts of the frame. To start with, the whole frame is partitioned into blocks of BUs, such that only the BUs in the same block may possibly be allocated to the same CU. For example, since persons at different street addresses cannot belong to the same (dwelling) household, each street address forms a block of BUs (i.e. persons).

Suppose there are *n* BUs in a given block, aligned with *m* CUs according to the frame. Let the corresponding *allocation matrix* ***A*** be an *nxn*-matrix of (0,1) indicators, whose element a_{ij} = 1 if the jth BU belongs to the ith CU, and 0 otherwise. Notice that there are *n-m* all-zero rows of the allocation matrix here. Moreover, for uniqueness of the allocation matrix, some scheme of row ordering will be necessary (Zhang, 2011).

As an example, consider the following three allocation matrices, which can be used to represent a block of 5 persons (BUs) grouped in two or three households (CUs):

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{A}^* = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Evidently, ***A*** and $\boldsymbol{A}^*$ can be used to represent the same situation, but only ***A*** is SUT. Next, suppose ***A*** is the allocation matrix according to the frame, whereas ***B*** is the reality, so that the first household consisting of the first three persons is correctly allocated, but not the last two persons. The misalignment of BUs and the constructed CUs here is the cause of frame unit (household) errors.

In this way, allocation matrix can be used to represent duplicated frame units, alignment errors between different types of frame units, as well as frame unit errors. The corresponding error at the micro level can be characterised by $P(\boldsymbol{B}|\boldsymbol{A})$, which is the within-block conditional distribution of the population allocation matrix ***B*** given the frame allocation matrix ***A***.

Notice that, conceptually, it is possible to represent all the BUs and CUs in the entire frame using a single allocation matrix. But this is hardly practical. Instead, potential misplacement of BUs in the different block can be envisaged as domain (i.e. block) classification error, thereby 'decomposing' the totality of frame alignment error into domain classification and within-domain allocation errors.

2.5 Some comparisons between frames for business and social statistics

For a better appreciation of the key quality issues of frames for social statistics, we compare them to those for business statistics. Table 4 summarises some prominent characteristics that differ between business and social statistics, and the generic types of frame errors they may affect. A more detailed elaboration of the key-word phrases and expressions in the table follows.

Units The term "business" has no unique or unambiguous definition. Still, roughly speaking, it can refer to an economic unit that is "engaged in the production of goods and services" (Colledge, 1995). The concept is much broader than colloquial usage for commercial or industrial activities, and can include government agencies, farms, non-profit organisations, etc. There is usually in operation a distinction between so-to-speak 'legally alive' and 'economically active'. While it may not seem to matter much in principle, since an 'economic inactive' unit can be assigned 0 measurements, it can affect the survey design and estimation efficiency if such units are not identified in the frame. The distinction between different types of business units is often based on characteristics such as legal ownership or liability, operational structure or relationship, kind-of-activity, location, etc. An important characteristic of the BR is that there does *not* exist a standard concept or practice of the

base unit (BU), from which one can build the different business units in use, such as enterprise, establishment, branch unit, etc.

In contrast, person is clearly the natural BU in frames for social statistics, when it comes to conducting household surveys. Notice that this does not imply that one necessarily has good quality of the frame unit person. For example, transient population arising from legally free movement or illegal migration may entail non-negligible frame coverage errors, despite the CPR may be of high accuracy for usual residents. In the countries that do not have a CPR, the frame quality of address, which is a composite unit, is often much better than that of person. The social statistics units that perhaps most resemble the business units in elusiveness are households, in particular the living household, at least insofar as frames based on administrative data sources are concerned.

Problems related to the various frame units in business and social statistic affect foremost the coverage, alignment error and unit errors. However, the lack of a natural definition of BUs in business statistics is an important reason that there currently does not exist a formal statistical theory of the alignment and unit errors in business statistics, as it is the case with households.

Table 4. Summary of some key characteristics and affected frame errors: business vs. social statistics

Characteristics	Business Statistics	Social Statistics	Frame Errors
Units	lack of standard base unit legal vs. operational kind-of-activity vs. location	person vs. household ⇔ BU vs. CU transient vs. permanent	coverage alignment unit
Contact	may vary for different statistics may be unknown or unclear	person member of household	contact information
Classification	NACE/SIC/etc. dictated by NA	natural interpretation no accounting framework	domain classification
Over time	merge/takeover/split (can be many-to-many) hard-to-trace revision of classification standard	merge/split of households (can be many-to-many) can be traced via BUs revision in limited extent	all types of frame errors
Measurement	size matters often continuous truncated/skewed/outlying	typically equal importance often categorical/nominal notable exception: income	output errors propagated from frame errors

Contact The right contact person may differ at the same business for different surveys. For instance, the person that has all the data for the Structural Business Survey may well be another to the one for the R&D survey. So even though it may not be difficult to get into contact with the business, the frame data of the contact person may be out-dated or incorrect. See Haraldsen *et al.* (2013).

In contrast, for household surveys, it is usually straightforward to identify the right person or persons to interview via any adult member of the household. But it may be more difficult to get into contact with the household to start with, whether by telephone or personal interview. For instance, in many countries the proportion of households for which at least one telephone number is available is

commonly below 95%. Notice that while non-contact is a common reason for nonresponse in household surveys, this is not necessarily due to the contact information error.

Classification The classification of industrial groups the business units belong to, such as NACE or SIC, is one of the biggest quality concerns for the BR. Of course, the problem cannot be entirely isolated from the unit and coverage problem. For instance, if an enterprise is wrongly divided into several local kind-of-activity units (LKAUs), then it is quite possible that the NACE classification of these LKAUs is mistaken as well. Even when the units are correctly delineated, assigning the NACE classification can be tricky. At the root of the matter is the "theoretical" nature of NACE classification, which to a large extent can be traced to the needs of National Account. Businesses simply do not form themselves in order to make the application of such statistical standards easier, and the lack of standard concept of BUs makes the problem harder to remedy.

In contrast, domain classification of person, household and other social statistics frame units often has a "natural" interpretation. For instance, a person may be classified according to sex, age, residence region, country of birth, etc. A household may be with or without children, with or without couple, with or without more than two generations, etc. Nor does one apply the same domain classification both to person (BU) and household (CU), except perhaps locality that necessarily is the same for all the persons in a given household. Therefore, domain classification errors of persons are 'simply' due to measurement errors in the relevant data, and those of households due to the (household) unit errors despite the data of the persons may be correct.

Over time The situation gets more complicated over time. To start with, it is difficult, in business as well as social statistics, to separate birth and death of the CUs (such as enterprises and households) from merge/takeover/split, although takeover may sound unnatural in social statistics. But it is possible to clarify the relationship between the households (or other CUs in social statistics) that exist at different time points, via the associated BUs (i.e. persons), even though it may be many-to-many to start with and difficult to summarise in a few words. Struijs and Willeboordse (1995, Fig. 4.1) represent the relationship between the businesses at two time points via the movements of the associated personals in much the same manner. However, since persons are *not* BUs in business statistics, such a representation nevertheless may convey little statistical data of interest, such as the flow of capital, the change of activity, the operational structure of the businesses, etc.

Another issue that seems to cause greater difficulty in business statistics is the revision of the classification standards themselves, such as NACE or unit types (e.g. the emergence of enterprise group). While some classification standards have changed over the time in social statistics, such as ethnicity, centrality, NUTS, etc., they have generated only limited repercussions compared to business statistics. For example, the increasing uptake of GIS in different data sources can be expected to lessen the pressure of geographical conglomeration due to changes in the NUTS standard. In contrast, any future changes of NACE will likely cause as much, possibly even more, trouble compared to the past.

Measurement Unlike the characteristics above that are related to representation (Groves et al. 2005; Zhang, 2012), measurement is not one that affects the frame quality *per se*. But it does affect the quality of the final statistical outputs via the propagation of the frame errors. In short, size matters in business statistics. Due to reasons such as truncation, skewed distribution and outliers in

measurement, relatively few units can often make a big difference to the results. In some ways, this can serve to reduce the impact of 'large' frame errors, such as the coverage or misclassification errors of small or micro business units, as long as one gets it right with the larger units in the BR. This is also an obvious reason why profiling is an important method for improving the frame quality of BR but not so for CPR.

With few exceptions such as income statistics, the units in social statistics have typically more or less equal importance to the final outputs due to the categorical nature of the measurements. In fact, if anything, the frame errors can be 'amplified' because of the fact that frame errors in social statistics tend to occur for the population units with a more marginal status. For example, if under-coverage of the population is more likely to happen for persons without regular social and economic engagement with the society, then the resulting relative over-estimation of the labour force participation rate will be *greater* than the under-coverage rate itself. To illustrate, let N be the frame population size for labour force statistics, and δ the under-count of it, so that the population under-coverage rate is $\delta/(N+\delta)$. Next, let N_1 be the number of persons that are labour force active, where $N_1<N$. The observed labour force participation rate is N_1/N, while the true rate is $N_1/(N+\delta)$. Then, the relative over-estimation is $(N+\delta)/N - 1 = \delta/N > \delta/(N+\delta)$, i.e. greater than the under-coverage rate.

3. Quality assessment: items, approaches and methods

Building on the above, we specify a list of 17 items that measure frame accuracy. Each item may consist of one or several target parameters that need to be estimated. We then outline 5 approaches to quality assessment. Finally, we describe the *most readily applicable* methods of assessment.

3.1 Specification

In Table 5 we provide a description of the 17 items, followed by detailed specification. Here, CM is a shorthand for "coverage including domain classification measure", PM for "progressiveness measure", AM for "alignment measure", "UM" for "unit error measure", and IM for "(contact) information measure".

Table 5. List of frame accuracy measurement items

CM1. Total under- and over-coverage for the target population CM2. Total correct domain classification CM3. Domain-specific population under- and over-coverage CM4. Domain misclassification (i.e. cross-domain under- and over-coverage)
PM1 – PM4 Counterparts of CM1 – CM4, specifically due to delays in source data
AM1. Total of correctly aligned base units (i.e. persons typically) AM2. Domain totals of correctly aligned base units AM3. Distribution of correctly aligned base units by composite unit types AM4. Total of correctly aligned composite units (e.g. household, address, etc.) AM5. Domain totals of correctly aligned composite units
UM1. Total number of population composite units UM2. Domain total numbers of population composite units
IM1. Total of frame units of given type with (correct, invalid, missing) contact IM2. Domain totals of frame units with (correct, invalid, missing) contact

CM1 – CM4 for frame coverage and domain classification are specified based on the following table.

	Frame Domain				
Population Domain	1	2	…	H	Missing
1	N_{11}	N_{12}	…	N_{1H}	M_1
…			…		…
H	N_{H1}	N_{H2}	…	N_{HH}	M_H
Erroneous	R_1	R_2	…	R_H	

The in-scope frame total is $\sum_{i=1}^{H}\sum_{j=1}^{H} N_{ij}$ and the population total is $\sum_{i=1}^{H} M_i + \sum_{i=1}^{H}\sum_{j=1}^{H} N_{ij}$.

CM1u. Total under-coverage: $M = \sum_{i=1}^{H} M_i$

CM1o. Total over-coverage: $R = \sum_{i=1}^{H} R_i$

CM2. Total correct domain classification: $N_0 = \sum_{i=1}^{H} N_{ii}$

CM3u. Domain-specific population under-coverage: $\{M_i; i = 1,\ldots,H\}$

CM3o. Domain-specific population over-coverage: $\{R_j; j = 1,\ldots,H\}$

CM4. Domain misclassification: $\{N_{ij}; i \neq j, i = 1,\ldots,H,\ j = 1,\ldots,H\}$

We observe that the coverage errors in CM3 are for the whole population, hence may be referred to as *genuine* coverage errors, whereas the coverage errors in CM4 are between the domains and the misclassified frame units are still inside the population, hence may be referred to as *spurious* coverage errors. In principle, domain-specific coverage can be assessed using the same data and method as for population coverage, treating each domain as a separate population. However, there may not be enough data to produce reliable estimates in this way, as for population coverage.

PM1 – PM4 are counterparts to CM1 – CM4, specified for the following table.

Target population at time *t* Frame at t_1	Frame at t_2				Not in at *t* (according to frame at t_2)
	1	2	…	H	
1	N_{11}	N_{12}	…	N_{1H}	N_{10}
…			…		…
H	N_{H1}	N_{H2}	…	N_{HH}	N_{H0}
Not in at *t* (according to frame at t_1)	N_{01}	N_{02}	…	N_{0H}	

The frame for target population at time *t* constructed at time t_1 is being assessed at time t_2, where $t \le t_1 < t_2$. We assume progressive frame data for *t* converges by t_2, after which there will be no changes about the state-of-affairs at *t*. In practice, $t_2 - t$ can sometimes be many years.

PM1. Total under- and over-coverage: $M = \sum_{h=1}^{H} N_{0h}$ and $R = \sum_{h=1}^{H} N_{h0}$

PM2. Total correct domain classification: $N_0 = \sum_{i=1}^{H} N_{ii}$

PM3. Domain-specific population under- and over coverage: $\{\mathrm{N}_{0h}; \mathrm{h} = 1,...,H\}$ and $\{\mathrm{N}_{h0}; \mathrm{h} = 1,...,H\}$

PM4. Domain misclassification: $\{N_{ij}; i \neq j, i = 1,...,H, j = 1,...,H\}$

We observe that in order to assess PM1 - PM4, it is necessary to be able to distinguish in the frame *at least two dates*: one for the relevant demographic event (e.g. birth, death or change of status), one for the registration of that event. Assessment is then possible without additional data. When the convergence time point t_2 is known and feasible, in the sense that the observations of N_{ij} are available *in retrospect*, the measures PM1 - PM4 can be calculated almost surely. However, when t_2 is either unknown or infeasible, e.g. when some of the data sources are completely new, or when the frame at t_1 is assessed at some $t' < t_2$, more sophisticated methods are needed.

AM1 – AM5 for alignment between frame base units and any type of frame composite units are specified based on the following table.

Base Unit Classification (Total N)	Composite Unit Classification (Total M)				Base Unit Total
	1	2	…	H	
1	(N_{11}, M_{11})	(N_{12}, M_{12})	…	(N_{1H}, M_{1H})	N_1
2	(N_{21}, M_{21})	(N_{22}, M_{22})	…	(N_{2H}, M_{2H})	N_2
…			…		…
G	(N_{G1}, M_{G1})	(N_{G2}, M_{G2})	…	(N_{GH}, M_{GH})	N_G
Composite Unit Total	M_1	M_2	…	M_H	

When person is BU, typical examples of CU are address, household, family, building, etc. But the table above is generic and (BU, CU) can be defined for the situation at hand. For example, one may set buildings as the BU and spatial grids as the CU. In any case, according to the frame, N_{gh} type-*g* BUs are aligned with (or belong to) M_{gh} type-*h* CUs. Let $N_{gh;t}$ and $M_{gh;t}$ be the number of *correctly aligned* BUs and CUs, respectively, and $N_{gh;e}$ and $M_{gh;e}$ that of the *incorrectly aligned* units. We have

$N_{gh} = N_{gh;t} + N_{gh;e}$ and $M_h = M_{h;t} + M_{h:e}$

AM1. Total of correctly aligned base units: $N_t = \sum_{g=1}^{G}\sum_{h=1}^{H} N_{gh;t}$

AM2. Domain totals of correctly aligned base units: $\{N_{g;t} = \sum_{h=1}^{H} N_{gh;t}; g = 1,...,G\}$

AM3. Distribution of correctly aligned base units: $\{N_{gh;t}; g = 1,...,G, h = 1,...,H\}$

AM4. Total of correctly aligned composite units: $M_t = \sum_{h=1}^{H} M_{h;t}$

AM5. Domain totals of correctly aligned composite units: $\{M_{h;t}; h = 1,...,H\}$

Next, UM1 – UM2 are specified for any type of composite units subjected to unit errors.

	Frame CU Classification				
Population CU Classification	1	2	…	H	Missing
1	$M_{1;t}$	--	…	--	Z_1
2	--	$M_{2;t}$	…	…	Z_2
…		…	…		…
H	--	--	…	$M_{H;t}$	Z_H
Erroneous	$M_{1;e}$	$M_{2;e}$	…	$M_{H;e}$	

Here, h = 1, 2, …, H denotes some classification of the CU, such as dwelling household by the number of residents. An erroneous frame CU cannot become another true CU in the population. For example, if a dwelling household of two residents constitutes a unit error, then either these two person live in different addresses, or there are other persons at the address. In either case, this frame dwelling household is not a dwelling household in the population. The absence of spurious over- and under-coverage is a key difference between CUs and BUs.

UM1. Total number of population composite units: $N = \sum_{h=1}^{H} M_{h;t} + \sum_{h=1}^{H} Z_h$

UM2. Domain total no. population composite units: $\{N_h = M_{h;t} + Z_h; h = 1,...,H\}$

As explained in Section 2.3.3, it is often possible to represent the relationship between frame units and associated contact information data as alignment between suitably defined BU and CU. We have nevertheless retained contact information error separately, and phrased IM1 and IM2 accordingly.

	Contact Information			
Frame Unit Domain Classification	Correct	Invalid	Missing	Total
1	$N_{1;t}$	$N_{1;f}$	$N_{1;m}$	N_1
2	$N_{2;t}$	$N_{2;f}$	$N_{2;m}$	N_2
…	…	…	…	…
H	$N_{H;t}$	$N_{H;f}$	$N_{H;m}$	N_H
Total	N_{+t}	N_{+f}	N_{+m}	N

IM1: Total number of frame units with correct, invalid and missing contact information are, respectively, $N_{+t} = \sum_{h=1}^{H} N_{h;t}$, $N_{+f} = \sum_{h=1}^{H} N_{h;f}$ and $N_{+f} = \sum_{h=1}^{H} N_{h;m}$.

IM2: Domain totals of (correct, invalid, missing) contact information are (N_{ht}, N_{hf}, N_{hm}).

3.2 Assessment approaches

As it has been distinguished earlier, generally speaking, a *quality indicator (QI)* does *not* suffice as an end-point of quality assessment, despite it may provide highly suggestive evidence for potential problems or lack thereof. In contrast, a *quality measure (QM)* is necessarily an end-point of quality assessment, granted its own assumptions or on its premises. *No further quantification of the relevant error is needed, but only alternative ones.* Thus, for each item in Table 5, one may obtain either QMs or QIs for the corresponding target parameters. Consider for example CM2, total correct domain classification $N_0 = \sum_{i=1}^{H} N_{ii}$. It is possible to arrive at a specific guess of N_0, i.e. an estimate, in many different ways. We shall consider such an estimate a QM only if it is accompanied by its own uncertainty measure, such as bias and variance; otherwise it will be considered as a QI. In particular, there are situations where certain assumptions, which are necessary for the derivation of a QM, may be considered too stringent or unrealistic. One needs to be aware of such situations where the quality measures have apparent difficulties of their own. In this sense, it is possible to speak of QMs as more or less *reliable* or *robust.*

We summarise below 5 existing *quantitative* approaches for assessing frame errors. The first three of them are aimed at producing quality measures, the fifth one (diagnostics) is commonly used for producing quality indicators, and the fourth one (on-going surveys) is used for both. The elaboration below aims to provide an accessible general overview. Details of the most readily applicable methods will be presented in Section 3.3.

Coverage survey is commonly applied for assessing the under-coverage errors in census, in which case it is also known as the *Post Enumeration Survey (PES)*. See e.g. Nirel and Glickman (2009) for a review. Sometimes the result is used to adjust the direct census counts. This can be highly controversial, as in the US where the matter reaches the Supreme Court. In truth the operations involved in the census and coverage survey are so complex that a clear-cut statistical judgement is unlikely ever to be possible. The usefulness of the approach for providing QMs seems however beyond reasonable doubt.

The methodology, referred to as *Dual System Estimation (DSE)*, has its origins in the so-called *Capture-Recapture (CR)* methods developed in wild-life, social and medical applications, traditionally used for under-count adjustment. Imagine catching fish in a pond on two separate occasions (i.e. census and coverage survey), where one marks and identifies the fish (i.e. enumerated records) that happen to be caught on both occasions (i.e. the recaptures). Then, under a number of simplifying assumptions, including independent and constant-probability captures, it becomes possible to estimate the total number of fish in the pond (i.e. the target population), for which the captures on each occasion generally entail under-counts. The method can be generalised to more than two captures to allow for relaxation of the underlying assumptions. The capture probability can be allowed to differ across sub-populations.

Quality survey is also a sample survey based approach that can depend on the census, or the frame to be evaluated. For instance, sampling from the census enumerations is used for census over-coverage adjustment. The so-called *Reverse Record Check (RRC)* is used for assessing the census under-coverage errors in Canada. Five non-overlapping frames together provide a subset of the target population, the aim is to determine whether a selected person from it is missed in the census or not. The monthly Quality Assurance Survey at Statistics Canada employs dependent sampling from the BR to measure the accuracy of industrial classification. Van Delden et al. (2016) develop a model for assessing the misclassification of NACE and their effects using quality surveys. Similarly, one may consider the annual *Structural Business Survey (SBS)* partly as a periodic quality survey of the BR, with regard to the domain classification errors (e.g. NACE and no. employees), the unit errors and contact information errors.

Quality survey for CPR, typically in a census year, is a common approach in Scandinavia (e.g. Werner, 2014), where the CPR provides the sampling frame. Sometimes the survey is simply conducted by adding an extra module to the Labour Force Survey questionnaire, possibly with an enlarged sample size on the occasion. The purpose of such a quality survey is not the population coverage errors of the CPR, but the domain classification errors, the various alignment or unit errors such as that of dwelling and household, and the measurement errors of the census questionnaire returns.

Modelling Model-based assessment of frame coverage errors based on administrative sources has attracted growing interest in the recent year. We refer to the special issue of the Journal of Official Statistics (2015, vol. 31, issue 3) for several useful references in this regard. In principle, the modelling approach to coverage errors (or population size estimation) no longer require fieldwork census or coverage surveys, which is a useful and potentially powerful approach to frames based on administrative data sources. The traditional log-linear models for CR-data in 2 or more lists (e.g. Fienberg, 1972; IMGDMF, 1995a, 1995b) can be used to deal with under-coverage errors. But there is plenty of scope for a wide range of different models.

For example, an important methodological extension to the census CR methods concerns erroneous enumerations, i.e. over-counts. The reason is again the progressiveness of administrative data, where there is often little incentive for one to deregister from sources such as the Patient Register or the Electoral Register in a timely manner. Zhang (2015) studies CR-models that accommodate both over- and under-coverage register errors. The modelling approach no longer requires a census to start with, nor an independent coverage survey, provided there are more than 3 lists. But one of the lists must not have over-counts. Zhang and Dunne (2016) apply the *Trimmed Dual-System Estimation (TDSE)* in Ireland to assess the plausibility of population size estimates derived purely from administrative data sources, where both lists are allowed to have over- and under-counts.

Provided multiple datasets (or lists) that in union entail *only over-counts* of the target population, a potential alternative approach is *latent class (or entity) modelling* in combination with record linkage, also referred to as entity resolution or co-reference. See e.g. (Di Cecco et al. 2016; Stoerts, et al. 2015) Imagine K lists of records, where each record may or may not refer to a target population element (i.e. latent entity). The records in the same list that refer to the same entity represent duplicated enumerations; the records in the different lists that refer to the same entity can be conceived as the target for record linkage. The errors in the resulting population total are then due to de-duplication and record linkage errors, which are traditionally the topics of record linkage.

For two other examples of the modelling approach, Hedlin et al. (2006) develop a log-linear model to assess the time lag for the introduction of birth units in the BR, and Mancini and Toti (2014) a multilevel model of small area census population counts.

On-going surveys Business surveys typically have a feedback mechanism into the BR regarding the statistical units, their domain classification and contact information. The difference to the quality surveys concerns chiefly the purposefulness of the design, including the data collection protocol. Standardised documentation of survey outcome is especially important. In particular, Istat (Italy) collects and stores a set of classified outcome status (Appendix A) in the Quality Documentation System for all the surveys. We shall return to it in Section 3.3 when describing related assessment methods. Similar practice can be found at STAT (Austria), as described in Chapter 4 later.

Assessing frame accuracy based on data collected in on-going household surveys is more cost-effective than using designed coverage or quality surveys. However, there is a bigger challenge of non-sampling errors compared to business surveys. Legislation for data protection is potentially another problem in this context, which can limit effective cross-validation using multiple sources that are available. For such reasons the resulting quality measures may be less reliable than desired. Useful quality indicators can still be produced, as explained below under Diagnostics approach.

An important emerging issue is the integrated social survey design in connection with the census transformation programme (e.g. at Istat), whereby the traditional census will cease to exist, and the on-going social surveys should not only provide the so-called attribute statistics but also serve to maintain and strengthen the EPD for provision of basic population statistics.

Diagnostics We group under diagnostics a variety of somewhat informal approaches that can produce useful QIs and improve frame maintenance routines. Four examples are given below.

1. Net or gross discrepancy checks are commonly conducted in comparison to external sources such as census or surveys. See e.g. Myrskylä (1991) in relation with the first Finnish register-based census. As a more recent example, ONS (2013) compares various statistical population datasets to the census 2011 (adjusted) population count, to assess the net over- or under-counts. Notice that gross discrepancy checks are readily applicable 'internally' to the *same* source over time.

2. Sign-of-Life (SoL) is another highly common approach based on linking multiple datasets and rule-based micro integration, when it comes to quality assessment of both the EPD and BR. For instance, the tax authority sometimes maintains a list of persons who have not engaged with the public administration or services in the past over a given number of years, and have not responded to any contact attempts. These persons are likely erroneous enumerations if they are found in the EPD. See e.g. Wegfor (2015) for an example of assessing CPR over-coverage.

3. Quality indicator system (QIS) provides a systematic approach to entity-level QIs for statistical registers, including multisource frames based on administrative data. If one envisages a frame as a data matrix, then the QIS associates every entry of this matrix with a score, usually standardised to be between 0 and 1, as the corresponding QI. The approach was pioneered at Statistics Austria in connection with the register-based census 2011, where the scores are propagated from those of the

input datasets to the census statistical database by means of formal rules (e.g. Berka *et al.,* 2012). See e.g. Hendriks (2014) for an account of the QIS for the EPD in Norway.

4. *Indirect standard quality indicators from surveys* can provide useful information on the frame quality. For example, based on the classified survey outcome status at Istat (Appendix), indicators on the rate of over-coverage and the rate of units with errors in contact information are regularly computed based on the results of the data collection phase. The indicators are then aggregated at frame level, thus allowing for comparisons of quality among different frames and over time, assessment of quality after innovations or system changes. Indeed, improvements on these quality indicators reflect improvements in the quality of the frames. Similarly, when the frames are used for direct estimation exclusively or in conjunction with surveys, variable imputation rates could be considered as an indirect indicator of the data quality. The aggregated analyses of quality indicators from the surveys are not publicly disseminated, but are published in the intranet of Istat website, for internal assessment and monitoring purposes only.

Summary A summary overview of the assessment approaches is given below.

Assessment Approach	**Coverage & Domain Classification**		**Alignment and Unit**	**Contact**
	CM1-4 (A)	PM1-4 (B)	AM1-5, UM1-2 (B)	IM1-2 (A)
Coverage or Quality Survey	Using sample from the population: DSE and TDSE; Using sample from the frame: RRC, Census follow-up	---	Quality survey based on audit sample from the frame	As special case of multi-frame sampling
Modelling (only limited application, experience)	For coverage: log-linear models with 3+ lists, latent class (entity) models, etc. For domain classification: misclassification models, Structural Equation Models, etc.	Few existing examples of models for delays	Allocation error model	As special case of log-linear models
On-going Survey	Existing data collection protocol and quality indicators	---	Lack of standard data collection protocol	
Diagnostics	Net or gross discrepancy checks, Sign-of-Life, Quality Indicator System, etc.			

The classification of A or B-list (in parentheses) depends on whether established methods and experiences exist for producing the relevant measures. There are clear gaps in methodology regarding the newly introduced assessment items PM1-PM4, AM1-AM5 and UM1-UM2, which are all B-list items, and survey based estimation methods for domain-specific coverage errors, which requires larger samples when each domain is treated as a target population on its own. Moreover, mature methods based on coverage and quality surveys are costly to implement. Making better uses of the on-going survey and the model-based methods should be the focus of future development, in order to be able to assess the frame accuracy regularly and at a relatively lower cost.

3.3 Methods

Here we describe the *most readily applicable* assessment methods for the items in Table 5. Notice that it is often possible to apply the same method to several items in the list.

3.3.1 Dual system estimation (DSE) for under-coverage

Suppose two lists with x and n enumeration records, respectively. Let m be number of matched records between the two. The DSE estimator of the population size is

$$\hat{N} = n\frac{x}{m}$$

The DSE is a standard method for census under-count assessment -- see e.g. Hogan (1993), Renaud (2007), Nirel and Glickman (2009). It originates from the capture-recapture methods in Biological and Social Sciences. Wolter (1986) lists more than 10 model assumptions that may be required to motivate its validity. For instance, the assumption of independent enumeration of both lists has been a key concern for research and survey implementation.

However, the DSE can more easily be motivated by treating one of the lists as fixed, as demonstrated by the Reverse Record Check (RRC) approach at Statistics Canada. For census under-count adjustment, this sets up the DSE essentially as follows: a list of sub-population is constructed from the previous census and several administrative sources, yielding x; the census is the other list, yielding n, and m on matching the two lists. For this to be valid, one needs *three* assumptions:

- no over-coverage (erroneous or duplication) error in either list;
- no matching error between the two lists;
- uniform enumeration rate in *one* of the two lists, say, the one that yields n.

In RRC, one assumes uniform enumeration rate in the census; whereas in Post Enumeration Survey (PES), one may treat the census enumeration as fixed and assume uniform enumeration rate in the coverage survey. Notice that, in practice, one may apply the DSE within the different population strata, in order to relax the assumption constant enumeration rate across the whole population.

Without losing generality, one may consider the x enumeration records as the targets for the other random list. The number of (valid) attempts is n, and the number of (re)captured records is m. As long as one can assume that all the x records have the *same* probability of being captured, denoted by π, m/x provides an unbiased estimator of π, and the DSE a method-of-moment estimator of the target population size N. As both n and m to vary, we have

$$\begin{aligned} V(\hat{N}) &\approx \frac{x^2}{E(m)^2}\left(V(n) - \frac{2E(n)}{E(m)}Cov(n,m) + \frac{E(n)^2}{E(m)^2}V(m)\right) \\ &= N(\frac{1}{\pi} - 1)\left(\frac{N}{x} - 1\right) \end{aligned}$$

Replacing N by the DSE xn/m and π by m/x, we obtain the variance estimator

$$\hat{V}(\hat{N}) = n(n-m)x(x-m)/m^3$$

Notice that this is the same variance estimate as when both lists are treated as random.

We refer to the Special Issue of Journal of Official Statistics (2015, vol. 31, issue 3) for many recent works on the topic, which serve as the point of departure for further exploration of the literature. For example, see Gerritse et al. (2015) for sensitivity analysis of the constant capture rate, Di Consiglio and Tuoto (2015) for the effects of linkage errors.

3.3.2 Trimmed dual system estimation (TDSE)

Suppose there are erroneous (i.e. out-of-scope) enumerations among the list with x records, but not in the other list with n records. Let r be the number of erroneous records, so that the number of in-scope records is $\tilde{x} = x - r$. The *ideal* DSE is then given by

$$\tilde{N} = n\frac{\tilde{x}}{m} = n\frac{x-r}{m}$$

i.e provided the other two assumptions of perfect matching and constant capture rate of the list with n records, respectively. Whereas the naive DSE, i.e. xn/m, is then an over-estimate. Zhang and Dunne (2016) propose the TDSE estimator, given by

$$\hat{N}_k = n\frac{x-k}{m-k_1}$$

where k is number of records *scored* (or *trimmed*) among the x records, of which k_1 records are among the m matched records between the two lists. This includes the naive DSE as the special case, i.e. $\hat{N}_0$ without any trimming at all.

It is shown that, as long as the scoring is more effective at picking out erroneous records than simple random sampling, we have

$$\hat{N}_k < \hat{N}_0$$

so that the TDSE adjusts the over-estimate $\hat{N}_0$ downwards. Meanwhile, as long as $r < k$, i.e. one does not trimmed more than the actual number of erroneous records, we have

$$\tilde{N} < \hat{N}_k$$

so that `over-trimming' is avoided. Finally, if all the r erroneous records are among the k scored ones, then the TDSE has the same expectation as the ideal DSE, i.e.

$$E(\hat{N}_k) = E(\tilde{N})$$

Zhang and Dunne (2016) provide three diagnostics for the stopping rule, i.e. when to stop the trimming. The approach is applied to the Irish administrative data, where there does not exist a central population register but a population dataset in the form of the Person Activity Register (PAR). It is demonstrated how to construct the method of scoring, and use the TDSE to analyse in details the potential over-coverage errors of the PAR, with respect to various sub-populations. For variance estimation, it is suggested to treat the convergent TDSE as if it were the ideal DSE, and use

$$\hat{V}(\hat{N}_k) = n(n-m+k_1)(x-k)(x-k-m+k_1)/(m-k_1)^3$$

Notice that, in the aforementioned application of the TDSE, the *n*-record list consists of the annual driving license dataset (DLD) of renewers and applicants. In other words, it is *not* a designed survey. Nonetheless the TDSE is applicable under the two assumptions: no matching error and uniform enumeration rate in the DLD. In Ireland, each license holder by law is required to renew the license every 10 years. Since the time point in life at which one obtains the driver license may be largely

independent of the subsequent events of life, the assumption of constant capture rate here essentially amounts to treating one's year-of-birth as a completely random event.

3.3.3 Frame-dependent sampling for over-coverage

Frame-dependent sampling for over-coverage and domain classification errors is a standard method of quality survey. The basic idea is straight-forward survey sampling: (i) select a sample of units from the frame, (ii) check how many of them are in the population and record the characteristics of the in-scope units, (iii) estimate the over-coverage or misclassification error and the associated sampling variance based on the sampling design. It seems helpful to incorporate the method in on-going surveys on a regular basis.

Consider an on-going survey that is *not* designed for frame assessment. Denote by s the selected sample, with associated design weights a_i for $i \in s$. Let L be the frame that is being assessed. Let U be the target population. Provided properly recorded survey status -- see Appendix for a template from Istat, so that one is able to identify the units in s (hence L) which are not in U. It is then possible to estimate the over-coverage error of L by

$$\hat{R} = \sum_{i \in s; i \notin U} a_i$$

and the associated sampling error of $\hat{R}$. Similarly, let L_h be a specific domain, and $s_h \subset L_h$ the domain subsample. An estimate of the domain-specific over-coverage error is

$$\hat{R}_h = \sum_{i \in s_h; i \notin U} a_i$$

Let U_h be the corresponding population domain. An estimate of the correctly classified frame units in L_h is given by

$$\hat{N}_{hh} = \sum_{i \in s_h \cap U_h} a_i$$

and the total of correctly classified frame units in L is given by

$$\hat{N}_0 = \sum_{i=1}^{H} \hat{N}_{ii}$$

We observe that, in practice, it may be difficult to assert the status of nonrespondents, i.e. category 15 in the Istat template. The estimates based on the remaining categories, as outlined above, are then lower-bound estimates of the corresponding targets. One can either consider these to be QIs or QMs with apparent limitations. Similarly, the direct sample proportions (as described in Appendix) can either be regarded as QIs or QMs that do not account for the sampling design effects.

The key to this approach is properly recorded survey status collected during the fieldwork. When the on-going survey is not aimed at frame assessment, the quality of such data may be poor without rigorous fieldwork management.

3.3.4 Log-linear models based on multiple lists

Log-linear capture-recapture models for multiple lists have often been used in Biological and Social Sciences for the estimation of `hidden' population size (e.g. Fienberg, 1972; IMGDMF, 1995a, 1995b). Applications to the general population are still under development -- see e.g. Griffin (2014) for an account of some recent developments in the US. We describe the basic log-linear modelling approach here, because the software implementation is readily available so that the practitioners can easily experiment with their own data.

Suppose one has K lists that enumerate the target population, all of which are subjected to *under-coverage but not over-coverage errors*. Let $Y_k = 1$ indicate that a population elements is enumerated in the *k*-th list, and $Y_k = 0$ otherwise. Assume $(Y_1,...,Y_K)$ follows a multinomial distribution, with parameters *N* and probability $\pi_{Y_1 \cdots Y_K}$ for *all* the population elements. Let $n_{Y_1 \cdots Y_K}$ be the observed cell counts, with expectation given by

$$\mu_{Y_1 \cdots Y_K} = E(n_{Y_1 \cdots Y_K}) = N\pi_{Y_1 \cdots Y_K}$$

The log-linear parameterisation of $\mu_{Y_1 \cdots Y_K}$ is given by

$$\log \mu_{Y_1 \cdots Y_K} = \lambda + \sum_{k=1}^{K} \lambda_{Y_k}^{(k)} + \sum_{k \neq j} \lambda_{Y_k Y_j}^{(kj)} + \cdots + \lambda_{Y_1 \cdots Y_K}^{(1 \cdots K)}$$

where the superscript clarifies the relevant lists for each log-linear parameter, i.e. / -term, and the subscript the values of the corresponding enumeration indicators.

The log-linear parameters are subjected to a set of constraints in order to avoid over-parameterisation -- see any textbook on log-linear models for details. When using standard software for fitting log-linear models, e.g. in R or SAS, one needs not to be concerned with such details, which are automatically handled by the software. However, the problem at hand implies that there is a structural zero cell, i.e. $n_{0 \cdots 0} = 0$, which is the number of elements that are missed by *all* the K lists. This can be accommodated by dropping, say, / . This results in a saturated model.

Further model reduction may be needed in order to estimate $\mu_{0 \cdots 0}$. For example, in the case with 2 lists, the largest identifiable model is [1][2], i.e. without the interaction term. Hence, one has to assume that the two enumeration indicators Y_1 and Y_2 are independent of each other, in order to apply Maximum Likelihood Estimation (MLE) under the log-linear model. This results in the DSE described earlier in the two-list case, and thus provides an alternative interpretation of the DSE under the log-linear modelling approach. More generally, with 3+ lists, having selected a suitable model, one obtains the MLE $\hat{\mu}_{0 \cdots 0}$ (Fienberg, 1972) and the corresponding population size estimate

$$\hat{N} = n_{obs} + \hat{\mu}_{0 \cdots 0}$$

where n_{obs} is the sum of the observed cell counts. Variance estimation requires an iterative procedure in general, since not all the MLEs can be given in closed forms. But it is available among the outputs of standard software.

As noted before, this log-linear modelling approach has been applied in many areas, and to hidden or hard-to-catch human populations, but not yet to the problem of estimating the size of the general population. A key concern is the sensitivity of the model assumptions. A related issue is the existence of alternative methods such as demographic accounting and the design-based DSE. Despite these alternatives are not unproblematic themselves, they do have a long tradition and seemed to have served the purpose well enough, making it difficult to adopt the pure model-based estimate.

In a situation without the census enumeration, though possibly with a coverage survey, a key difficulty of estimation based on administrative data is the presence of non-negligible over-coverage errors in the administrative enumerations. Di Cecco et al. (2016) extend the log-linear model to include a latent variable X, where $X=1$ if an enumeration record belong to the population and $X=0$ otherwise. See also Bryant and Graham (2015) for a Bayesian hierarchical modelling approach that treats the population size as a latent quantity, where the model no longer belongs to the class of log-linear models.

3.3.5 Two special models for 3 lists

Zhang (2015) consider the following setting of 3 lists in details: the first two lists A and B are both subjected to over- and under-coverage errors, and the third list S is only subjected to under-coverage error and has a constant capture rate. For example the list S can be the coverage survey, whereas the first two lists can be the census and the EPD, or two administrative registers. Of all the possible log-linear models of the list-population universe, only one is shown not to involve apparently *ad hoc* assumptions and can readily be extended to situations with more than lists A and B, which is given by

$$\operatorname{logit}\theta_{11} = \operatorname{logit}\theta_{10} + \operatorname{logit}\theta_{01}$$

where $\operatorname{logit}\theta = \log\theta - \log(1-\theta)$ and

$$\theta_{ab} = \Pr(I_{i\in U} = 0 \mid I_{i\in A} = a, I_{i\in B} = b)$$

Due to the closeness between logit and log functions for small q values, this is approximately equivalent to the assumption that

$$P(i \notin U \mid i \in A \cap B) = P(i \notin U \mid i \in A \backslash B) P(i \notin U \mid i \in B \backslash A)$$

It is argued that, as the over-coverage errors of both lists diminish, a more plausible assumption may be the following, i.e.

$$P(i \notin U \mid i \in A \cap B) = P(i \notin U \mid i \in A) P(i \notin U \mid i \in B)$$

which however does not belong to the class of hierarchical log-linear models. The main difference is that, under the second model, the over-coverage rate among the *matched* elements of both lists can be much lower that that of each list, i.e. most erroneous records are among the unmatched ones.

It is shown that closed-form method-of-moment estimator can be given by

$$\hat{\theta}_{10} = \frac{x_{01}}{n_{01}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{10}}{x_{10}}\right) \quad \text{and} \quad \hat{\theta}_{01} = \frac{x_{10}}{n_{10}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{01}}{x_{01}}\right)$$

for the first model, and

$$\hat{\theta}_{1+} = \frac{x_{+1}}{n_{+1}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{1+}}{x_{1+}}\right) \quad \text{and} \quad \hat{\theta}_{+1} = \frac{x_{1+}}{n_{1+}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{+1}}{x_{+1}}\right)$$

for the second model. Here, x_{ab} is the count of enumeration records with $I_{i\in A} = a$ and $I_{i\in B} = b$, and n_{ab} is corresponding count of those that can be matched to the third list S. For instance, x_{11} is the number of records in list A and B, and n_{11} that in list A, B and S. The estimate of the capture rate of list S is then given by

$$\hat{\pi} = \frac{n_{10}}{x_{10}(1-\hat{\theta}_{10})} = \frac{n_{01}}{x_{01}(1-\hat{\theta}_{01})}$$

under the first model and

$$\hat{\pi} = \frac{n_{1+}}{x_{1+}(1-\hat{\theta}_{1+})} = \frac{n_{+1}}{x_{+1}(1-\hat{\theta}_{+1})}$$

under the second model. The population size estimate, under either model, is given by

$$\hat{N} = x_{11}(1-\hat{\theta}_{11}) + x_{10}(1-\hat{\theta}_{10}) + x_{01}(1-\hat{\theta}_{01}) + \frac{n_{00}}{\hat{\pi}}$$

Variance estimation can be based on linearisation or bootstrap.

3.3.6 Domain misclassification error CM4

If some of the methods described above is applied to each domain separately, quality measure CM4 may require a very large sample. There are some other relevant methods that currently are under development. One of them is the approach for NACE misclassification (Van Delden et al., 2016). For social statistics, locality is naturally a domain classification of interest, where the number of locations may vary from small (e.g. region) to very large (e.g. municipality). Another relevant method is Dostál et al. (2016) for the German census. Clearly, reliable CM4 estimates can be used to construct relevant CM1-CM3 estimates and are therefore more demanding.

3.3.7 Coverage errors due to progressive frame data

Provided it is feasible to make assessment after the convergence time point t_2 *in retrospect*, direct cross-tabulation as described for the definition of PM1-PM4, yields all the desired figures almost surely, i.e. QMs with bias and variance both equal to zero, for the frame $L(t;t_1)$, i.e. frame for time point t that is constructed at t_1.

In case the difference between $t_2 - t$ is unknown, or too long for timely assessment, let t' be the *assessment* time point, where $t_1 < t' < t_2$. The figures obtained from direct cross-tabulation between $L(t;t_1)$ and $L(t;t')$ can be considered to yield a QI of the true error associated with the frame $L(t;t_1) = L(t;t+d)$, where $d = t_1 - t$. The observed error $e_t(d;t')$, i.e. obtained at t', is then most likely an underestimate of the true error $e_t(d) = e_t(d;t_2)$. More sophisticated methods would require modelling of the remaining future changes.

The observed error $e_t(d;t')$ is stochastic, now that $t' < t_2$. That is, by chance it is possible to observe a different number $e_t(d;t')$. To assess the variability of $e_t(d;t')$, it is natural to make use of data over time. Consider $L(t;t+d)$ for fixed d, but varying $t = T_1, T_2, ..., T_K$. Let $y_t = e_t(d;d')$ be the observed error associated with $L(t;t+d)$, which is assessed at time $t' = t + d'$, where $d < d'$. Under the simplifying assumption that the y_t's are IID observations of *different frames over time*, one may assess the corresponding expectation and variance accordingly, or indeed its distribution provided there are enough observations of y_t. The IID-assumption needs to be tested. The approach is currently being investigated, and we will report the results in the final deliverable of Kumoso.

3.3.8 Alignment, unit and contact information error

Design-based methods In principle design-based methods can be used to assess alignment, unit and contact information errors, in a manner similar to on-going surveys for coverage errors described above. The key difficulty may be the potential bias caused by nonresponse and measurement error. One may still consider the resulting estimates as QIs or QMs with apparent deficiencies.

Diagnostics By utilising multiple input sources to the EPD, diagnostic-based QIs can often be devised.

- For example, juxtaposing multiple sources of address from the population register, the postal address register, the TV license register, the higher education register, etc. it is possible to produce QIs of people with correct resident address, at-risk of having incorrect resident address, etc.
- Similarly, combining the population register and various Sign-of-Life data, e.g. tax returns, utility bills, etc. may be able to yield QIs for correct and incorrect address, and possibly over-coverage error in the next instance.
- Examination of the population register together with available contact information, such as address, telephone number, email address, etc. may be able to yield QIs for (correct, invalid, missing) contact information of various types.

4. Applications

Below is an overview of the applications to be presented, by target item and country.

List: Item	*Austria*	*Denmark*	*Hungary*	*Ireland*	*Italy*	*Lithuania*
A: CM1-CM3		CM1o	CM1o, CM3o	CM1-CM4	CM1o, CM2, CM3o	CM1-CM4
B: PM1-PM4		PM1u			PM1, PM3,PM4	
B: AM1-AM5	AM1-AM5					AM1-AM5
B: UM1-UM2						
A: IM1-IM2	IM1-IM2	IM1-IM2	IM1-IM2		IM1-IM2	

The following presentation is organised by assessment approach.

4.1 On-going survey

The approach is described in Section 3.3.3.

4.1.1 Italy: CM1o, CM2, CM3o, IM1-IM2

Both unweighted sample rates and design-weighted rates are calculated for the following items: CM1o, CM2, CM3o, IM1 and IM2, which are the ones feasible based on the available data. The results are not very different, so only the unweighted results are presented. As domains of interest, the five Italian geographical areas were considered: north-west, north-east, center, south, islands. In particular, an out-of-scope rate is calculated in addition to over-coverage rate (Appendix), i.e.

Out-of-scope rate = [(Out of scope)/(Resolved)]*100
Over-coverage rate = {[(Out of scope) + (1-α)*(Unresolved)]/(Total)} *100

The distinction is due to the presence of unresolved units for which it has not been possible to identify the eligibility status. It is customary for Istat surveys to set α=1, which yields a lower bound estimate of the true over-coverage rate, whilst the out-of-scope rate is an upper bound estimate. The units with invalid *or* partially missing contact information, which caused nonresponse in the survey, are used to calculate IM1 (overall) and IM2 (domain specific), given by

No-contact rate due to frame errors = [(No Contacts Due to Frame Errors)/(Resolved)]*100

Finally, some results for the dual system estimator (Section 3.3.1.) from the Post Enumeration Survey (PES) carried out for the 15° Census of population and households are also reported and commented. The PES was carried out during April-July 2012, and focused on items CM1 and CM3.

Background

Before 2011, the elementary sampling units of Istat samples for social surveys were drawn by the municipality staff, in the municipalities selected as primary sampling units, based on Istat instructions. Since then, the municipality data on population have been centralised at Istat. As preparatory operations for the 15th Population Census, municipalities were required to transmit to Istat, between January and mid-February 2011, their municipal population registers (Liste anagrafiche comunali - henceforth ***LAC***) detailing information on all registered residents as of December 31st 2010. Data were updated at the Census Day (October 9th 2011). The data is validated and integrated with other sources. The LAC contains the following frame units: institutional households, private households, persons. At present it has about 25 million households and 60 million persons.

From 2013 Istat carries out a yearly survey aimed at updating the LAC. In addition, every time a sample is selected from the LAC and field data collection is performed, standard survey outcome status (Appendix) is recorded for the units that have been contacted, and stored in Istat Quality Documentation System - SIQual and analysed (Brancato *et al.*, 2006).

Regarding the PES, in addition to the DSE for under-coverage, over-coverage has been assessed on the basis of the units in PES which were linked with more than one unit in Census (duplications) and the units in PES which were linked with units in Census in different places (misclassifications). The estimation in conducted in three steps: 1) estimate of the number of duplications from PES; 2) estimate of the number of misclassifications from the PES; 3) calibration of the estimates at the 1st step using the number of duplications observed at the census in order to improve the precision of the over-coverage estimates (the sample of the PES is not designed to obtain estimates of duplications and misclassifications). The described estimation process provides the weights for over-coverage by strata, weights that are afterwards applied to the Census individual records, and whose sum corresponds to the census total required in the dual system estimator.

Results

Data from several surveys that use LAC as sampling frame and for several years have been analysed. First, for a given year, results from different surveys are presented. Next, results for a chosen survey are given for several years to gauge the trend over time. Notice that, since the chosen survey is not based on a panel design, longitudinal results need to be interpreted with care.

Results on single year and multiple surveys (in percent)

Three surveys that selected the sample from LAC in 2014 were selected. They have similar sample sizes (approximately 25000 households), but different data collection modes.

Item	Survey 1	Survey 2	Survey 3
CM1o			
Out-of-scope rate	3,3	2,63	3,07
Over-Coverage Rate	3,3	2,54	2,84
CM2 (Total correct domain classification)			
North-west	96,62	93,23	89,65
North-east	96,69	94,51	93,42
Center	97,06	93,50	84,02
South	96,38	94,92	91,25
Islands	96,67	94,27	90,37
CM3o (out-of-scope rate)			
North-west	3,38	2,73	3,09
North-east	3,31	2,59	2,60
Center	2,94	2,39	3,01
South	3,62	2,67	3,27
Islands	3,33	2,83	3,64
IM1			
No-contact rate due to frame errors	1,15	2,44	2,72
IM2 (No-contact rate due to frame errors)			
North-west	1,71	3,00	3,38

North-east	0,93	2,87	2,73
Center	1,22	2,03	2,69
South	0,53	2,09	2,20
Islands	1,22	1,97	2,40

It is seen that the results are compatible when estimated from different surveys, thus increasing the confidence of the on-going survey approach. In particular, the out-of-scope and over-coverage rates are rather close to each other, in which case the inherent uncertainty due to the unresolved units is less worrisome in practice than what otherwise could have been.

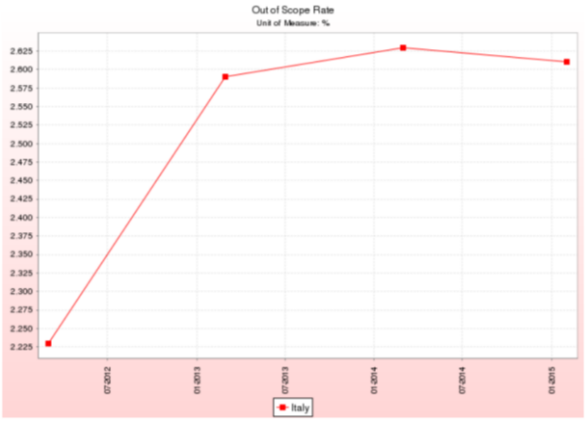
Results on single survey and multiple years

A social survey with the following characteristics was considered:

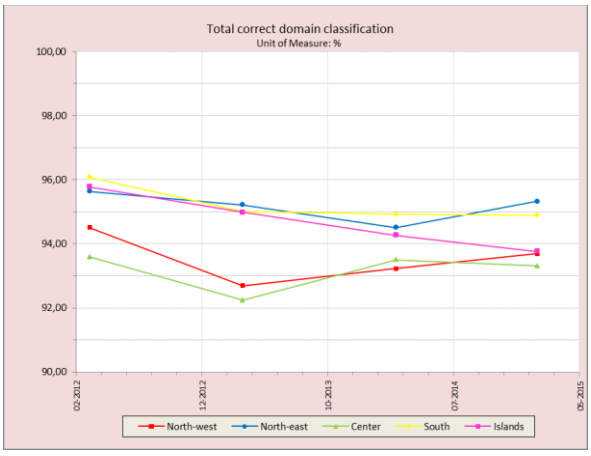
- Reporting and elementary sampling unit: household; sample size: around 25.000 units
- Data collection mode: Paper and Pencil Interviewing
- Sampling design: Partly single-stage and partly two-stage sampling with stratification of primary sampling unit (municipality). The PSUs are stratified according to the 20 Italian regions and by socio-demographic characteristics into metropolitan areas, metropolitan surrounding outskirts by inhabitants: <=2,000; 2,001-10,000; 10,001-50,000; >50,000.

The indicators relative to CM1o, CM2, CM3o, IM1 and IM2 were plotted for the four available survey editions, and the graphs are reported below.

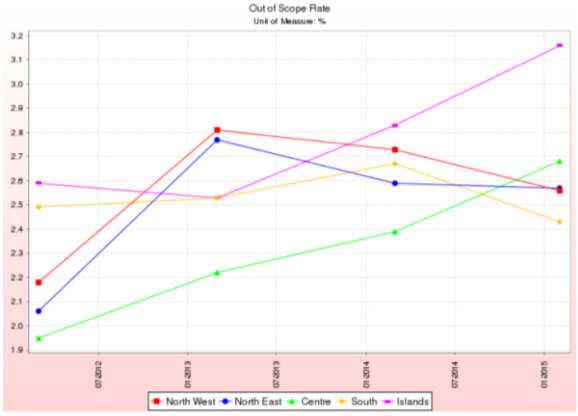
CM1o, Out of scope rate



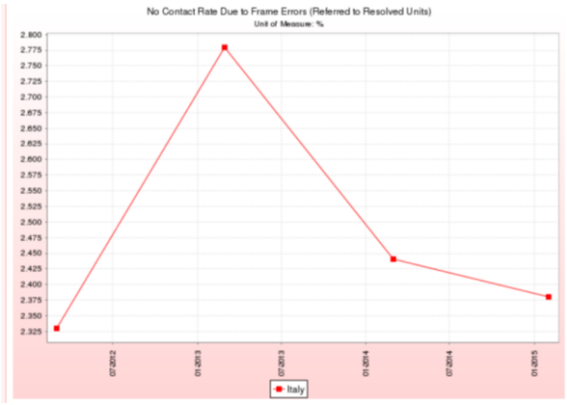
CM2, Correct domain classification rate



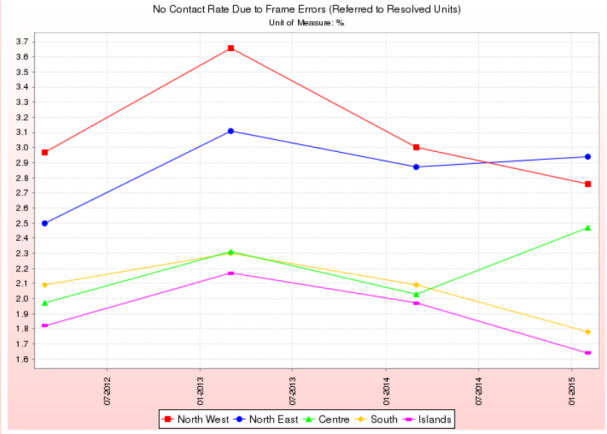
CM3o, Domain specific out of scope rate



IM1, No-contact rate due to frame errors



IM2, No-contact rate due to frame errors by domain



First of all, it should be noticed that changes over the years are rather small in absolute values. Still, in all the cases it appears that the error is the smallest in the first year 2012. A possible explanation is that in 2012 the frame still benefited from the recent update by the 2011 census data, while the routine successive updating of LAC was delayed to 2013. Since then, a set of activities aimed at improving the quality started to be regularly adopted, and this is reflected in the improvement that can be observed in the following two years, with some variability across the domains.

A comparison to the Post Enumeration Survey

The tables below are excerpts from the published census quality assessment (in Italian).

CM1, Estimate of the population and under- and over-coverage (PES, 2012)

N Census (a)	N estimated including over-coverage (b)	N estimated excluding over-coverage (c)	Gross under-coverage (d)	Over-coverage (e)	Net under-coverage (f)
59.132.045	60.002.997	59.774.142	870.952	228.855	642.097

CM1, Under-coverage and over-coverage rates (PES, 2012, in %)

Rate of gross under-coverage (h) = (d/b)*100	Over-coverage rate (i) = (e/b)*100	Under-coverage rate (h-i) = (f/b)*100
1,45	0,38	1,07

From this application CM1o=228.855 and the corresponding over-coverage rate=0,38%.

CM3, Domain specific population under- and over-coverage (PES, 2012, in %)

	CM3o	CM3u
North-west	0,48	1,30
North-east	0,32	0,76
Center	0,29	1,42
South	0,48	1,06
Islands	0,19	1,11

The PES provides the possibility to consider also other relevant domain, such as that by citizenship. The estimated under-coverage rate of foreigner population at the census is 11,07%. Over-coverage of the census 2011 is substantially lower than the estimated over-coverage of the LAC -- respectively 0,30% by the PES and about 2,22% from the given survey in early 2012. A potential issue with the routine updating of LAC is that, for administrative and economic conveniences, there is a tendency to inflate the list enumeration, i.e. not to delete persons that have moved away from the municipality.

4.1.2 Hungary: CM1o, CM3o, IM1-IM2

Design-weighted rates are calculated for the item CM1o and CM3o, and IM1-IM2. The domains are the 7 regions of Hungary. In particular, the out-of-scope rate among the unresolved units is estimated by the corresponding rate among the resolved ones, i.e.

$$\text{Over-coverage rate} = \{[(\text{Out of scope}) + (1-\alpha)*(\text{Unresolved})]/(\text{Total})\} *100$$

$$1-\alpha = (\text{Out of scope}) / (\text{Resolved})$$

The units with invalid contact information, which are unable to locate, are used to calculate IM1 (overall) and IM2 (domain specific), given by

$$\text{No-contact rate due to frame errors} = [(\text{No Contacts Due to Frame Errors})/(\text{Total})]*100$$

The approach is applied to the LFS for the Dwelling Register (as the sampling frame), and to the PIAAC for Population Register (as the corresponding sampling frame).

Background

The Dwelling Register is owned and maintained by HCSO, which is based on the 2011 census and is regularly updated by new and demolished dwellings. The institutions are excluded from the sampling frame for household statistics. The Population Register is obtained from the Central Office for Administrative and Electronic Public Services. The sampling frame for the PIAAC consists of adults of age 16 - 65, where age is defined with respect to the midpoint of data collection period (May – August). The register contains some persons who have emigrated from Hungary but have not been removed from the registry. The survey outcome status is not very different from the one shown in Appendix. The protocol is not standardised across the surveys, so there are some differences between the LFS (illustrated below) and PIAAC (omitted).

Code	Description	Classification of outcomes		
		Resolved	Scope	Frame error
10	Complete	YES	IN	NO
21	Unable to locate dwelling unit	NO	?	YES (contact)
22	Dwelling unit under construction/does not exist	YES	OUT	YES
23	Vacant dwelling unit	YES	OUT	YES
24	Address not a dwelling unit (e.g. institution)	YES	OUT	YES
31	Maximum number of calls reached	YES	IN	NO
32	Household has moved from this dwelling unit	Not applicable in first wave		
41	Refusal	YES	IN	NO
42/43	Contact person is unable to answer/Language problem	YES	IN	NO

Results

CM1o & IM1, Dwelling Register based on LFS (2014 – 2016, in %)

Item	2014	2015	2016
Resolved rate	98,85	98,89	98,93
Over-coverage rate (CM1o)	14,62	16,56	17,19
No contact rate due to frame errors IM1)	1,15	1,11	1,07

CM3o & IM2, Dwelling Register based on LFS (2016, in %)

Item	*Southern-Great-Plain*	*Southern-Transdanubia*	*Central-Transdanubia*	*Central-Hungary*	*Western-Transdanubia*	*Northern-Great-Plain*	*Northern-Hungary*
Resolved rate	99,22	99,42	99,53	98,37	99,13	98,60	99,36
Over-coverage rate (CM3o)	19,21	23,85	17,51	9,93	17,61	22,73	21,23
No contact rate due to frame errors (IM2)	0,78	0,58	0,47	1,63	0,87	1,40	0,64

CM1o & IM1, Population Register based on PIAAC (2016, in %)

Item	2016
Resolved rate	89,49
Over-coverage rate (CM1o)	9,80
No contact rate due to frame errors (IM1)	11,07

Clearly, there are considerable over-coverage errors in both registers. But the dwellings provide much more reliable contact data than the personal contact data in the Population Register.

4.1.3 Austria: AM1-AM5, IM1-IM2

Design-weighted rates are calculated for the following items: AM1-AM5. The frame is that for the register-based labour market statistics 2014 (RBLMS), covering the whole Austrian population in private household on 31 October. The domains are formed by region or demographic features. The on-going survey Micro-census (MC) of the 4th quarter of 2014 is used as an audit sample. For alignment error, the BU is persons in private households and the CU private households. The linkage between CUs in the RBLMS and MC is achieved via the household reference person in the RBLMS. The items AM1-AM5 are calculated as follows, with weighted totals.

$\boldsymbol{L}_{gh} := \{(x,Y) \in BU \times CU | x \in BU, Y \in CU, x \in Y, type_{BU}(x) = g, type_{CU}(Y) = h\}$

$N_{gh} := \#\{x \in BU | \exists Y \in CU\ s.t.\ (x,Y) \in \boldsymbol{L}_{gh}\}, N_g := \#\{x \in BU | type_{BU}(x) = g\}$

$N'_{g;t} := \#\{x \in BU \cap MC | \exists Y \in CU\ s.t.\ (x,Y)$ is correctly aligned in the MC and $type_{BU}(x) = g\}$

$N'_{gh;t} := \#\{x \in BU \cap MC | \exists Y \in CU\ s.t.\ (x,Y) \in \boldsymbol{L}_{gh}, (x,Y)$ is correctly aligned, i. e. $type_{BU}(x) = g, type_{CU}(Y) = h$ in the MC$\}$

AM1	Total correct aligned BU rate (with respect to the attributes $type_{BU}, type_{CU}$)	$R_t := \frac{N'_t}{\#BU \cap MC}$

AM2	Correctly aligned BU rate for the BU-domain g (with respect to the attributes $type_{BU}, type_{CU}$)	$R_{g;t} \coloneqq \frac{N'_{g;t}}{\#\boldsymbol{L}_g \cap MC}$
AM3	Distribution of correctly aligned BU rate (with respect to the attributes $type_{BU}, type_{CU}$)	$R_{gh;t} \coloneqq \frac{N'_{gh;t}}{\#\boldsymbol{L}_{gh} \cap MC}$

$M_{gh} \coloneqq \#\{Y \in CU | \exists x \in BU\ s.t.\ (x, Y) \in \boldsymbol{L}_{gh}\}, M_h \coloneqq \#\{Y \in CU | type_{CU}(Y) = h\}$

$M'_{h;t} \coloneqq \#\{Y \in CU \cap MC |\ Y \text{ is correctly aligned, i. e. } type_{CU}(Y) = h \text{ in the MC}\}$

AM4	Total correct aligned CU rate (with respect to the attributes $type_{CU}$)	$R_t \coloneqq \frac{M'_t}{\#CU \cap MC}$
AM5	Correct aligned CU rate for the domain (h) (with respect to the attributes $type_{CU}$)	$R_{h;t} \coloneqq \frac{M'_{h;t}}{\#\{Y \in CU | type_{CU}(Y) = h\} \cap MC}$

Example: $type_{BU} = \text{sex: } m, w;\ type_{CU} = \text{HH\_size: } 1,2,3+$

	1	2	3+
m (men)	(N_{m1}, M_{m1})	(N_{m2}, M_{m2})	(N_{m3}, M_{m3})
w (women)	(N_{w1}, M_{w1})	(N_{w2}, M_{w2})	(N_{w3}, M_{w3})

$N_{m2} \coloneqq$ number of men living in a 2 person HH according to the LMS

$N'_{m2;t} \coloneqq$ number of correct aligned [men living in a 2 person HH]

$N_t \coloneqq$ correct aligned rate of BU [with respect to sex and HH_size]

$N_{m2;t} \coloneqq$ correct aligned rate [of men living in a 2 person HH in the LMS]

$M_2 \coloneqq$ number of households with 2 inhabitants

$M'_{2;t} \coloneqq$ number of correct aligned households with 2 inhabitants

$M_t \coloneqq$ correct aligned rate of CU [with respect to HH_size]

$M_{2;t} \coloneqq$ correct aligned rate [of CU with 2 inhabitants]

Two types of contact data in the CPR are examined with regard to items IM1-IM2: 1) if an address from the CPR (source of RBLMS) is contactable, which is not the case if an interviewer reported “no access to the building”, “no private household at the address” or “address information not sufficient to find the building”; 2) if there is a telephone number available for a person in the CPR. The MC is used for the former, and the error is found to be negligible and the results omitted below. Three samples of the Holiday and Business trips survey, consisting of about 120.000 persons older than 14 years, are used to assess the availability of telephone numbers for persons in the CPR. For the actual survey only the part of the sample with available telephone number is used.

Background

Topics covered by the RBLMS are population, household and family characteristics. For this Statistics Austria distinguishes between 7 base registers and 8 comparison registers. The base registers (e.g. the CPR) are needed to provide all attributes of interest for the RBLMS. Additional supporting

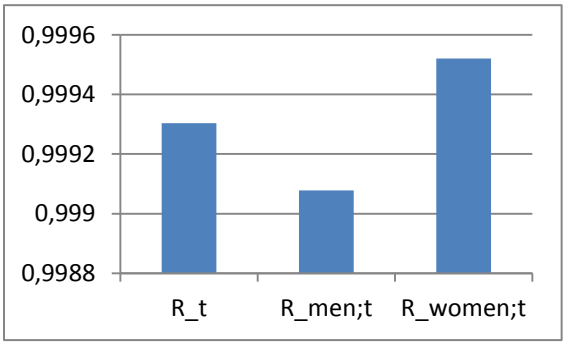
comparison registers are used to further improve the quality of the RBLMS, which gather additional information for cross-checks from more than 50 external data holders. The dwelling household is the adopted concept. Households in the RBLMS are generated by linking the CPR with the Buildings and Dwellings Register (BDR). These registers share the same numerical addresses for buildings, but not always consistent with each other on door numbers and therefore dwellings. The BDR is highly reliable at the building level. As far as dwellings are concerned, the linking of dwellings with people registered in the CPR is less successful due to some missing or wrong door numbers. In these remaining cases additional sources are used to generate households, e.g. relationships from e.g. the Central Social Security Register. The constructed CUs may suffer from alignment and unit errors.

The MC is a quarterly survey of about 22.000 private households (with about 47.000 in-scope persons). The LFS is integrated into the MC. The MC uses a rotating design, where in each quarter 20 percent of the sample units are refreshed and a household remains in the sample for 5 consecutive quarters. Private households (identified by addresses) with at least one person as main resident are included in the sampling frame, which is built using the CPR. All persons currently living at a selected address are interviewed for the MC, regardless of whether it is their main residence. In other words, a person may be interviewed at a different address in the MC than according to the RBLMS.

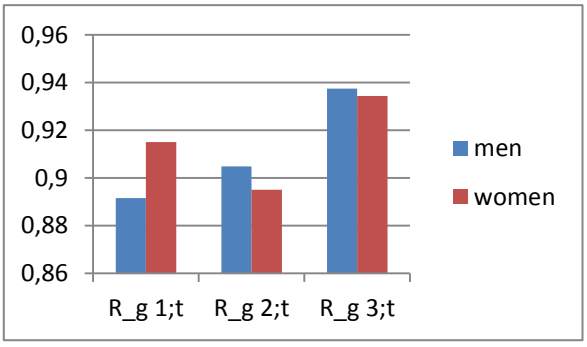
Results

In the results for AM1-AM5 below the BU is grouped by sex and the CU by (household) size 1, 2, 3+.

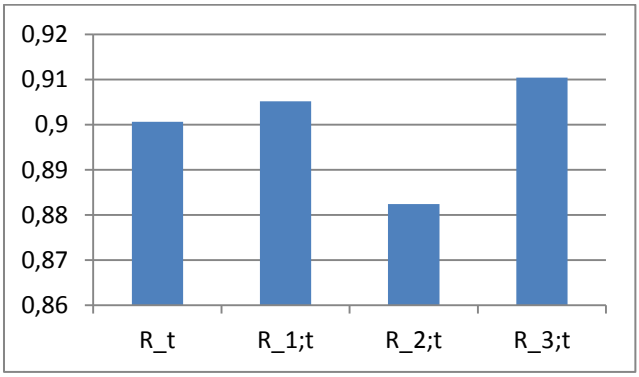
AM1 and AM2 (by sex) for RBLMS



AM1 and AM2 (by sex and size) for RBLMS



AM4 and AM5 (by size) for RBLMS



Overall, woman have a slightly lower alignment error rate than men (AM2), although in terms of the underlying distribution over the CU sizes (AM3) men have actually lower alignment error rates in 2 of the 3 size groups. Women living in single-person households have about 2% higher correctly aligned

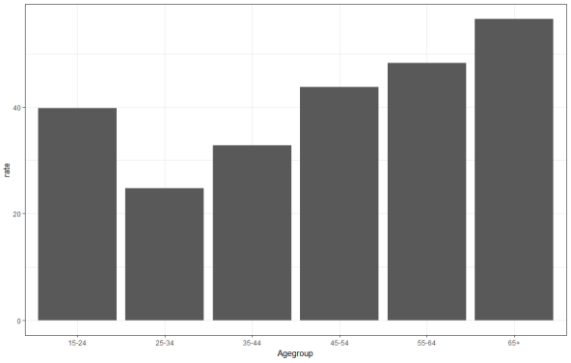
households than men living in single-person households. Regarding the items AM4 and AM5, as expected, increasing household size is associated with increasing alignment errors (AM5).

For contact data, the overall indicator for IM1 is 41.8% with contact telephone number. It is 42.2% for male and 41.3% for female. Breakdowns by NUTS-regions, age or income groups are given below.

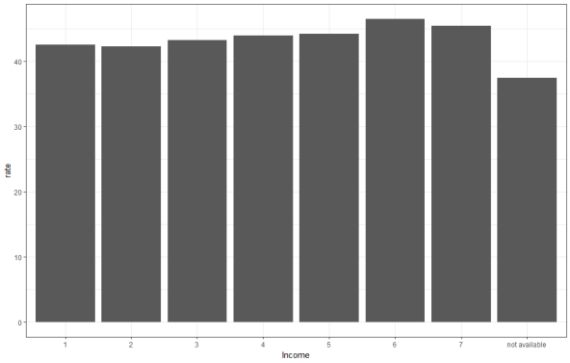
IM2, for CPR by NTS2-regions

AT11 - Burgenland	46.1%
AT12 – Lower Austria	45.8%
AT13 – Vienna	23.9%
AT21 – Carinthia	41.9%
AT22 - Styria	42.7%
AT31 – Upper Austria	52.4%
AT32 – Salzburg	47.2%
AT33 – Tyrol	47.5%
AT34 – Vorarlberg	43.6%

IM2, for CPR by age groups



IM2, for CPR by income groups



The availability of telephone numbers is very limited in urban areas, which is obvious when looking at Vienna with only about 24% of coverage with available telephone numbers. There is clearly a strong relation between the availability of a telephone number and age. Notice that persons in the youngest age group mostly still live with their parents and therefore in a household with a landline telephone. The availability increases very slightly with income. Not surprisingly, the availability is the lowest for the (last) group for whom income data is not available in the registers.

4.1.4 Appraisal

The approach based on frame-dependent on-going surveys is shown to be useful for assessing the quality of corresponding sampling frame, or other constructed frames to which micro-data linkage is essentially unproblematic. We notice the following in particular.

- Since the on-going surveys are not designed to estimate frame errors, the results are expected to be less accurate compared to the traditional coverage/quality survey approach, the latter being more powerful and informative but also more expensive. In any case, the approach based on on-going surveys should be helpful for providing timely monitoring of the frame quality over time and evaluating the effects of quality interventions on the frame.
- The approach is easy to implement, especially in surveys adopting computer assisted data collection modes, since the technique allows for automatic coding and storage of the necessary survey outcome status. As mentioned before, standardised protocol of survey outcome status

and enhanced fieldwork management can be expected to improve the comparability of the results across different surveys and over time.

- Intuitively one might feel more confident about calculating various error rates instead of error totals, e.g. due to the uncertainty surrounding the unresolved and, to some extent, out-of-scope units. To appropriately estimate the frame errors (either total or rate), a different weighting may be needed than the direct design weights, or the weighting for the survey outcome statistics of interest. The issue deserves more attention in future, in order to quantify the uncertainty of the estimates, and to transform them from QIs to QMs.

4.2 Modelling

The CSO Ireland applied the modelling approach to investigate coverage errors in Person Activity Register (PAR). Firstly, the DSE is used to generate population estimates (Sec. 3.3.1), which in turn provides an under-coverage estimate (CM1u) of the PAR. The results are further compared to the census results, which operates a census-night definition. Next, the TDSE (Sec. 3.3.2) is applied to explore the potential over-coverage errors of the PAR and the resulting population size estimate. This follows on from previous work done in the CSO (Zhang and Dunne, 2016).

The DSE uses the PAR as list A, and the driving license renewal/applications dataset (DLD) as list B. The target population consists of persons over 18 years of age due to the use of the DLD. The TDSE used a series of trimmed PAR counts (X_T), which is obtained from trimming the initial X_0 by 'activity' i.e. sequential removal of the records associated with each input administrative source to the PAR.

Background

Ireland has a Unique Personal Identifier for all individuals i.e. a Personal Public Service Number (PPSN). Since January 1971 the number is issued automatically to everyone born in the Republic of Ireland and, since April 1979, to those who commenced or were in employment here. A PPSN is required for many interactions with government public services and can be applied for from the Client Identity Services section of the Department of Social Protection. Therefore, persons who were not included in the automatic issuing of PPSNs, such as home-makers and immigrants, would very likely have obtained a PPSN for some administrative activity over time.

The PAR has been produced based on births and deaths registers and a 'sign of life' basis i.e. use of the PPSN, using ten different administrative data sources (shown below) for the years from 2006 to 2014. It is available approximately one year after the reference year.

The administrative data sources (frames) used to compile the Enhanced Public Register PAR

Child Benefit	From Department of Social Protection Central Records System. One database with PPSN for child, one with PPSN for parent
P35 Activity	Employment or occupation pension activity from Revenue Commissioners
Income Tax	Self-employed income tax from Revenue Commissioners
Social Welfare	Unemployment activity Department of Social Protection
ECCE Child	Availing of the child pre-school scheme from the Department of Children and Youth Affairs
Post Primary	Registered in secondary school Department of Education and Skills
FETAC Activity	Further Education register (post-secondary) from Further Education and Training Awards Council pre-2012 Quality and Qualifications Ireland (QQI) post- 2012
HEA Activity	Registered at a University or college from the Higher Education Authority
State Pension	State pension activity from Department of Social Protection

Results

The population estimates are compared by sex and age (0 – 100). The differences between the estimators, DSE and TDSE, and the census are measured using root mean square deviation (RMSD). The population was divided into ten year age groups and the deviation of the DSE and TDSE from the census was calculated for the different age groups. Variance and confidence intervals for the estimators at each age/gender domain are also calculated. The formulae are given below.

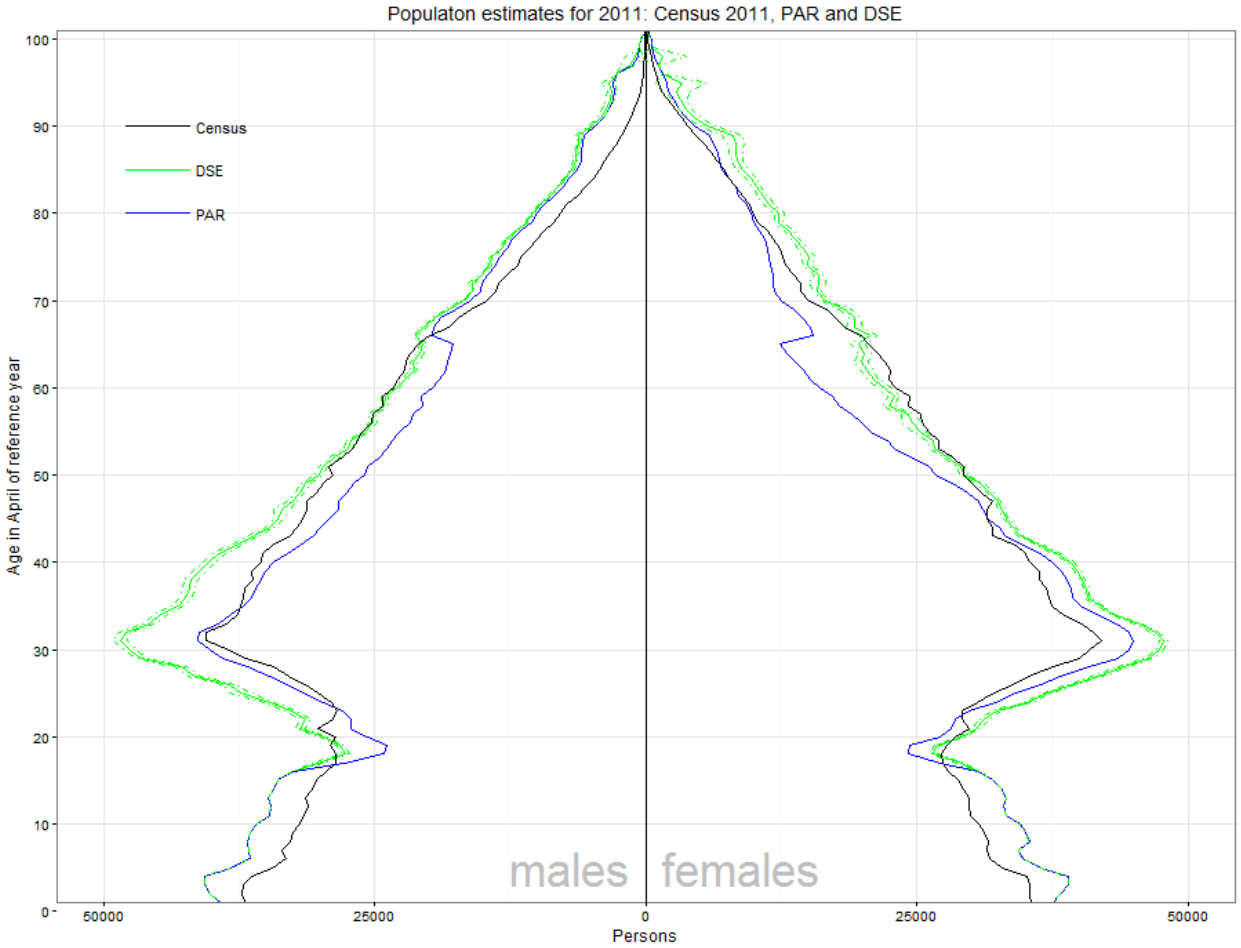
i =age, *c* = interval

$$RMSD = \sqrt{\sum_{i=i}^{i+c} \frac{(Census(2011)_i - Estimate_i)^2}{c}}$$

$$Var = \frac{X_0 * n * (X_0 \backslash n) * (n \backslash X_0)}{(X_0)^3}$$

$$95\% \text{ confidence interval } = \pm\sqrt{Var} * 1.96$$

The total population estimates by the PAR enumeration and the DSE is, respectively, 100.2% and 110.7 % compared to the census count. The age-sex distributions are plotted in the figure below.



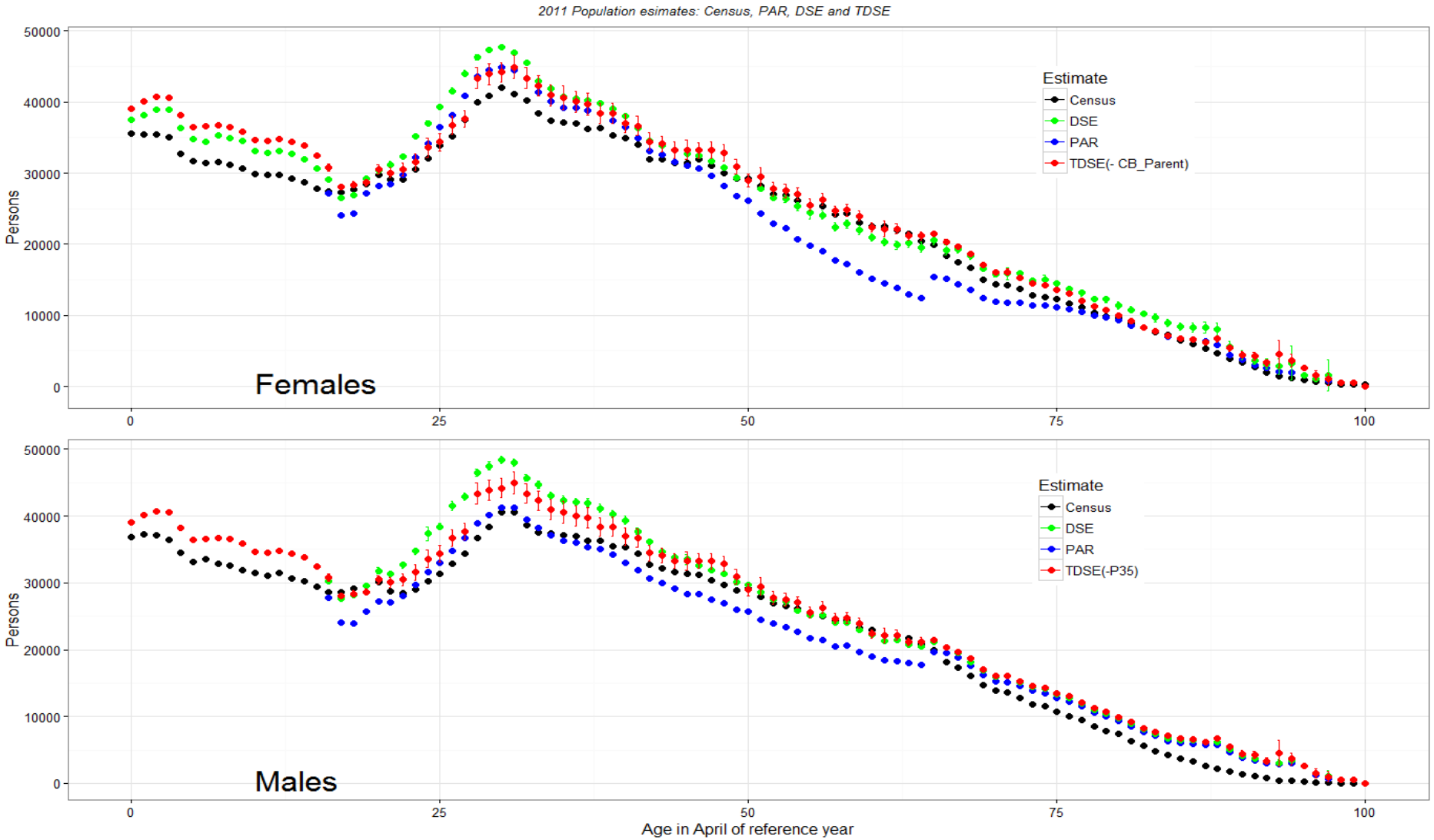
The census, personal activity register (PAR) and Dual system estimator (DSE) for 2011

Notice that currently the PAR over-counts children under 18, due to the nature of the data available for this age group. However, the DSE does not apply to this age group.

RMSD of PAR, DSE) and various TDSE (compared to 2011 census counts): O = over and U = under compared to census count. By age groups.

Female	**0 - 9**		**10 - 19**		**20 - 29**		**30 - 39**		**40 - 49**		**50 - 59**		**60 - 69**		**70 - 79**		**80 - 89**		**90 - 100**	
PAR	3339	O	2938	O	2527	O	2664	O	1330	U	5532	U	6141	U	1484	U	599	O	494	O
DSE	3339	O	2594	O	5073	O	4520	O	1747	O	1015	U	1543	U	2050	O	2224	O	1038	O
TDSE																				
-SWA	3339	O	2595	O	5064	O	4495	O	1679	O	2100	U	2543	U	2040	O	2218	O	1036	O
-P35	3339	O	2607	O	3527	O	3615	O	1512	O	1261	O	1774	O	3373	O	3873	O	2819	O
-Post Primary	3339	O	7164	O	5072	O	4521	O	1746	O	1014	U	1542	U	2050	O	2227	O	1038	O
-Income Tax	3339	O	2594	O	5088	O	4548	O	1765	O	964	U	1491	U	2079	O	2249	O	1035	O
- CB-Child	31035	U	12497	U	5073	O	4520	O	1747	O	1015	U	1543	U	2050	O	2224	O	1038	O
-CB-Parent	**3339**	**O**	**2613**	**O**	**4471**	**O**	**2357**	**O**	**803**	**O**	**1274**	**U**	**1582**	**U**	**2047**	**O**	**2223**	**O**	**1038**	**O**
-ECCE	3314	O	2594	O	5073	O	4520	O	1747	O	1015	U	1543	U	2050	O	2224	O	1038	O
-HEA	3339	O	2602	O	5065	O	4517	O	1747	O	1009	U	1535	U	2051	O	2222	O	1038	O
-FETAC	3339	O	2612	O	5026	O	4490	O	1732	O	1043	U	1548	U	2066	O	2222	O	1038	O
-SW Pension	3339	O	2594	O	5072	O	4518	O	1743	O	1026	U	1995	U	2672	U	2163	U	638	U
-SW UN	3339	O	2594	O	5073	O	4520	O	1747	O	1015	U	1543	U	2050	O	2224	O	1038	O

Male	**0 - 9**		**10 - 19**		**20 - 29**		**30 - 39**		**40 - 49**		**50 - 59**		**60 - 69**		**70 - 79**		**80 - 89**		**90 - 100**	
PAR	3522	O	3547	O	1849	O	862	U	2628	U	3467	U	2769	U	1956	O	2561	O	1814	O
DSE	3522	O	2733	O	6967	O	6174	O	2504	O	424	O	1502	O	2526	O	2976	O	2028	O
TDSE																				
-SWA	3522	O	2731	O	7000	O	6172	O	2361	O	647	U	1865	O	2517	O	2968	O	2027	O
-P35	**3522**	**O**	**2731**	**O**	**3661**	**O**	**3687**	**O**	**2178**	**O**	**866**	**O**	**1570**	**O**	**2697**	**O**	**3244**	**O**	**2359**	**O**
-Post Primary	3522	O	8003	O	6963	O	6173	O	2504	O	424	O	1502	O	2526	O	2976	O	2028	O
-Income Tax	3522	O	2733	O	7077	O	6210	O	3016	O	774	O	1459	O	2530	O	3011	O	2086	O
- CB-Child	32414	U	13371	U	6967	O	6174	O	2504	O	424	O	1502	O	2526	O	2976	O	2028	O
-CB-Parent	3522	O	2733	O	6979	O	5821	O	2254	O	386	O	1500	O	2525	O	2976	O	2028	O
-ECCE	3498	O	2733	O	6967	O	6174	O	2504	O	424	O	1502	O	2526	O	2976	O	2028	O
-HEA	3522	O	2750	O	6926	O	6188	O	2492	O	420	O	1499	O	2526	O	2976	O	2027	O
-FETAC	3522	O	2781	O	6933	O	6201	O	2564	O	428	O	1504	O	2528	O	2983	O	2028	O
-SW Pension	3522	O	2733	O	6966	O	6170	O	2494	O	417	O	752	U	850	U	470	U	143	U
-SW UN	3522	O	2733	O	6967	O	6174	O	2504	O	424	O	1502	O	2526	O	2976	O	2028	O



The PAR, DSE and TDSE most closely aligned with census counts for 2011: trimmed by 'Child Benefit – Parent' for females and trimmed by 'P35' for males.

The results for each 10-year age group are given for female and male separately. The PAR and the DSE provides population estimates above the census for males over age 65. The PAR records under count the over 65 females until age 80 and then returns numbers very close to the census figures. The DSE for females provides an estimate over the census count from age 65 to 90. The pension database is the main source of administrative data in the older population and its removal has the effect of drastically decreasing the size of X_T. However, trimming by pension activity actually provides a TDSE much closer to the census counts in the male population. The TDSE does result in a reduced estimate compared to the census counts for females in the age 60 – 100.

From age 20 to 50 in males and age 20 to 45 in females the DSE is markedly higher than the census and higher than can be accounted for by a census undercount e.g. an extra 6,310 males at age 32, +16% difference. While under-coverage in the census would account for some of this difference it unlikely accounts for all. It was hoped that a TDSE would provide better estimates in this age range by suitable trimming. Judging from the RMSD of the TDSEs, it seems that removal of the 'Child Benefit –parent' data produced estimates closer to the census estimate in the age 20 to 45 females with minimum effect on DSE in the other age groups. The removal of the 'P35' data provided estimates closer to the census estimate in the age 20 to 50 males with minimum effect on DSE in the other age groups. A closer comparison of these TDSEs and the other estimates are also given.

Appraisal

The development of social statistics from administrative data is in the experimental stage in Ireland. The census count is used as a standard although it is likely to have some under-coverage associated with it. The development of the administrative approach will on the one hand result in better frame quality and on other hand rely on improved estimation techniques like the TDSE.

Child Benefit in Ireland is a universal payment to parents of children and the main source of administrative data for the under 18s. It is not transformed by the modelling process used here because of the nature of the driver licence renewal database. Better quality data is needed for this section of the population. The 'Child Benefit –child' records used for the 2011 PAR consisted of records of children registered on the system. In the future it will be possible to use actual payments made rather than registration on the system for inclusion in the PAR and this change should provide an improved estimate for children.

Generally, it was found that trimming by administrative source has a smaller effect on X_T for females than males (indicating that more females are present on more than one list). It appears that the TDSE can provide better population estimates than the DSE for adults age 20 - 50. The strength of the TDSE approach can be seen in the removal of the 'P35' data which has a striking effect on the trimmed PAR, reducing X_0 by approximately half for the age 30 – 40 males, but producing a TDSE closer to the census estimate than by the untrimmed DSE. It has been noted before that TDSE can behave differently according to age and sex (Zhang and Dunne, 2016). The PAR trimmed by 'P35' provides better estimates for the age 20 – 45 females but yields larger errors in older age groups; trimming by the 'Child Benefit – parent' is aligned better with the census for females over all. Therefore, the best estimates in these age ranges were produced using different trimming according to sex.

There is potential to use administrative data for population estimates if the appropriate estimation methods can be determined. Making use of audit samples can be helpful in this respect.

4.3 Hybrid approach

Statistics Lithuania combine modelling with on-going survey data to assess the items CM1 – CM4. We refer to it as a hybrid approach because explicit logistic regression modelling is applied to some parts of the frame while design-based estimation is applied to some others. For domain specific assessment, the frame is divided into those with Lithuanian or foreign addresses.

The target population is individuals having personal ID number of Lithuanian Republic and older than 16 years. *Frame* consists of individuals in the Population Register (PR), who have personal ID number of Lithuanian Republic and is older than 16 on October 1st, 2016. Two important classification variables from the PR are *document type*: valid, non-valid; and *address type*: municipality, declared, foreign, and other, where "other" means errors such as house is destroyed, non-residential building, etc. The frame is merged with the databases of Labour Exchange (LE) and Social Security/Insurance institution (SS) on October 1st, 2016, in order to obtain Sign-of-Life (SoL) classification: SS only, LE only, no SS no LE, SS and LE. Cross-classification of these three variables gives rise to 32 cells below.

Document type	Sign-of-Life	Address type			
		Municipality	Declared	Foreign	Other
Valid	SS only	A1	B1	C1	D1
Valid	LE only	A2	B2	C2	D2
Valid	No SS no LE	A3	B3	C3	D3
Valid	SS and LE	A4	B4	C4	D4
Non-valid	SS only	U1	V1	Z1	Y1
Non-valid	LE only	U2	V2	Z2	Y2
Non-valid	No SS no LE	U3	V3	Z3	Y3
Non-valid	SS and LE	U4	V4	Z4	Y4

To obtain CM1-CM4, one needs to identify for each cell the number of individuals in or out of the target population. The latter may be further divided into those living abroad and those deceased. Under assumption that proven SoL, i.e. in LE and/or SS, means that the individual is living in Lithuania, a cell belongs to the target population, as long as it is on one of the 6 rows with proven SoL. For the remaining two rows, anyone except in cells A3 and B3 is considered to be living abroad, because the document type is non-valid to start with and there is no proven SoL. Finally, the individuals in A3 and B3 who do have valid document type can be divided into three groups: in-scope, abroad and deceased. Denote by I2 the number of individuals in B3 who live abroad and M2 those in B3 who are deceased; and I1 and M1 correspondingly for A3. Let the additional hats indicate their estimates. The 32 cells are thus mapped to the 3x2 table below for assessment of CM1-CM2.

Population	Frame		
	Living in Lithuania	Living abroad	Errors
In Lithuania (In-scope)	A1+B1+A2+B2+A4+B4+U1+V1+U2+V2+U4+V4 + (A3 - $\hat{\mathrm{I}}1$ - $\hat{\mathrm{M}}1$) + (B3 - $\hat{\mathrm{I}}2$ - $\hat{\mathrm{M}}2$)	C1+C2+C4 +Z1+Z2+Z4	D1+D2+D3+D4 +Y1+Y2+Y4
Abroad (Out)	$\hat{\mathrm{I}}1+\hat{\mathrm{I}}2$ + U3+V3	C3+Z3	Y3
Deceased (Out)	$\hat{\mathrm{M}}1+\hat{\mathrm{M}}2$		

The union of B1, B2, B3, B4 serves as sampling frame for social surveys at Statistics Lithuania. Using information on cause of non-response in the European health interview survey 2014, the number of individuals in B1+B2+B3+B4 who live abroad I2 is estimated using sample design weights:

$$\hat{\mathrm{I}}2 = \hat{t}_y = \sum_{k\in(B1+B2+\mathrm{B3}+B4)\cap s} d_k y_k \ ,$$

Here, *s* is a sample drawn from B1+B2+B3+B4, d_k the sample design weight, and $y_k = 1$ if individual did not respond to the survey due to living abroad, and $y_k = 0$ otherwise. Taking into account the SoL assumption these are all from B3. The number of people from B3 who passed away is estimated based on the corresponding cause of non-response in exactly the same way.

For people belonging to A3 additional SoL data do exist in Lithuania but are not accessible to this study. Instead, using the survey data, a logistic regression model is built for the *y*-variable:

$$P\left(y_k = 1 \middle| x_k\right) = \frac{e^{-(a+bx_k)}}{1+e^{-(a+bx_k)}}\ ,\ k \in \mathrm{B3}.$$

where *x* is a vector of auxiliary variables. But among the data available only the variable *age* is correlated with *y*. Under the assumption that the same model applies in A3, one obtains

$$\hat{p}_k = \frac{e^{-\left(\hat{a}+\hat{b}x_k\right)}}{1+e^{-\left(\hat{a}+\hat{b}x_k\right)}},\ \ k \in \mathrm{A3}.$$

Summing the estimated probabilities over A3 yields an estimator for I1: $\hat{\mathrm{I}}1 = \sum_{k\in\mathrm{A3}} \hat{p}_k$. It estimates the number of individuals in A3 who live abroad. Similarly for $\hat{\mathrm{M}}1$ – the number of deceased.

Results

Estimated cross-classification of frame by 3 variables

Document Type	Sign-of-Life	Municipality	Address type Declared	Foreign	Other
Valid	SS only	21266	1146209	2372	57
Valid	LF only	7974	115077	1146	19
Valid	No SS no LE	38547	1267490	257544	133
Valid	SS and LE	1265	22323	33	1
Non-valid	SS only	445	13079	139	5
Non-valid	LF only	189	1568	10	0
Non-valid	No SS no LE	2282	26155	40146	121
Non-valid	SS and LE	17	313	1	0

Population	Frame domain		
	Living in Lithuania	Living abroad	Errors
In Lithuania (In-scope)	2600857	3701	215
Abroad (Out)	54252	297690	121
Deceased (Out)	9090		

CM1-CM4, Frame domain (Lithuanian, Foreign) address

In-scope frame total	2956500
Population total	2956836
CM1u Total under-coverage	M=336
CM1o Total over-coverage	R=9090
CM2 Total correct domain classification	N_0=2956500
CM3u Domain-specific population under-coverage	{215; 121}
CM3o Domain-specific population over-coverage	{9090; 0}
CM4 Domain misclassification	{54252; 3701}

Appraisal

Although at the first sight the tables for CM look like simple cross-classification tables, they are not so easily obtained because not all the cell counts are observable directly, even in the sample. Some assumptions are needed without purposefully collected data for frame assessment. Sometimes the assumptions can take the form of a parametric model, such as the logistic regression model here. Access to additional databases like health service, pension, allowance and others could provide more Sign-of-Life data. On the one hand, frame cross-classification can be refined to have fewer or smaller uncertain cells, on the other hand, more powerful model can be built for the remaining uncertain units. Construction of the residence index for the Estonian population (Tiit and Maasing, 2016) provides a good example of the same hybrid approach. The index is calculated yearly based on 27 databases and the same index of the previous year, which takes value between 0 and 1 for every element of “an extended total population”. Together with suitable thresholds, this index can be used to classify the “extended

total population" into in- and out-scope target population elements. In the Estonian case the residence index allows one to estimate the resident population size and Population Register over-coverage due to persons who have left the country without registration. Notice that the construction is ultimately based on the rationality of decisions and comprehensive subject knowledge, insofar as any relevant survey data will suffer from non-identifiable non-sampling errors. Therefore, in practice, one aims to limit the effects of model misspecification to an acceptable level.

4.4 Diagnostics

4.4.1 Ireland: CM2, CM4

The CSO used a diagnostic approach to investigate frame errors CM2 and CM4. The domain of concern is defined by year-of-birth. The variable date-of-birth in the PAR (Section 4.2) is taken from the client record system of the Department of Social Protection, as it is considered most likely to be correct. The DLD (Section 4.2) is obtained from the national driving licence service (NDLS). In Ireland driving licences are renewed every ten years but more frequently for learners, commercial drivers and people over 70 years of age. The NDLS had over 2 million current driving licences on their database at the end of 2014. The NDLS database is not used to construct the PAR but the application/renewals in a year specific is used to generate a second list (NDLSr) in capture-recapture modelling (Section 4.2). The driver licence database also records the date-of-birth. A rules-based registration process has been introduced in Ireland to provide assurance about the information held by public agencies about an individual, called the Standard Authentication Framework Environment (SAFE). There are four levels (0-3) and the rules of establishing and authenticating identity increases with each level. In 2013 the SAFE2 level for identity registration was introduced in the NDLS for driving licence applications and renewals. This involves a face-to-face interview where the client is required to produce documentary evidence of identity, including photographic. SAFE2 provides 'substantial assurance' regarding identity. The CSO has investigated effects on potential domain classification errors, i.e. based on year-of-birth in the NDLSr, by comparing the error rate for 2012, before the introduction of SAFE2, and the error rate for 2014, one year after its introduction. For a diagnostic approach, error is the case provided a mismatch of year-of-birth in the PAR and NDLSr.

Results

Year	PAR	NDLS	Merge on PPSN		
	Total (over 18)	Total (over 18)	PAR ∩ NDLSr	PAR \ NDLSr	NDLSr \ PAR
2012	3,345,344	444,383	377,392	2,969,555	66,991
2014	3,081,838	311,842	252,738	2,829,466	59,104

Year	In-scope records	Variable: Year-of-birth			
		Correct/Match (CM2)		Error/Mismatch (CM4)	
2012	377,392	359,927	95.3 %	17,465	4.7 %
2014	252,738	248,538	98.3 %	4,200	1.7 %

The diagnostic results are presented above. The NDLSr recaptures 10 % and 8 % of the adult PAR in 2012 and 2014, respectively. These are the in-scope records where domain classification errors can be quantified. The total correct domain classification, CM2, is 95.3 % and 98.3 % for 2012 and 2014, respectively. The total domain misclassification, CM4, is 4.7 % and 1.7 %, respectively. It appears that the SAFE2 registration process has resulted in a 3% improvement in the accuracy of the variable year-of-birth since its introduction.

Appraisal

Investigating domain classification errors in this manner produces an indicator rather than a measure. However, it is a quick and straightforward method to assess frame quality. An automated process for comparing frames could be produced. The method can be up-scaled to assess more variables within the frames to produce overall indicators of frame quality. A minimum standard based on frame classification errors would be easy to understand for data providers and users. Domain classification error is a useful indicator to monitor frame quality over time and to monitor the impact of new initiatives, like SAFE2.

4.4.2 Denmark: PM1u, CM1o

DST applied diagnostics to assess items PM1u and CM1o. Delayed emigration registration constitutes the data for PM1u, where the target population is emigrants in a given year. Lack of Sign-of-Life (SoL) provides the data for potential non-deregistration (C1Mo), where the target population is the usual resident population in Denmark.

Background

Statistics Denmark relies fully on registers for producing population statistics. The centralized civil register in Denmark serves as a national register and was set up in April 1968. The Civil Registration System is regulated by act number 878 of 14 September 2009 with amendments in later years. The latest official Executive order is no. 5 of 9th of January 2013. A person needs to have the right to stay in the country before being a part of this register. This means that asylum seekers and illegal migrants are not part of the register. Today the main reference date for the population statistics in Denmark is the 1st of January for stock-data and the complete year (January 1st - December 31st) for the vital statistics. However, being completely register based, statistics can actually be produced for any given day during the year.

The Civil Registration System is used by the public administration in almost all cases which are related to individual persons. This means that there are many possibilities of detecting and correcting errors and deficiencies in the register. Thus a person cannot obtain work as an employee without giving information about his PIN-number (personal identification number) to the employer who is to withhold the employee's provisional tax and remit it to the tax authorities. This and several other reporting channels provide a relatively safe guarantee that all persons are actually registered in the Civil Registration System, and that the most frequently used information (addresses, family relations, etc.) are being correctly registered. For example, reporting to the register of increases and decreases in the population as a consequence of births and deaths are done by midwifes and doctors involved. This means that very reliable sources are used. Reports on decreases as a consequence of emigrations are however less

reliable, because it is the emigrant himself who is liable to notify the local population register about the emigration, and often there is little incentive to give this notification. The Civil Registration System, therefore, has persons registered who no longer live in the country. However, this situation is often rectified within a relatively short period when the municipality for different reasons notices that the person is no longer in the country. It could be that letters from the municipality to the individual are returned if the person cannot be found. Many students apply for housing benefits to the municipality. If the municipality then finds that someone else is already registered in the same dwelling, an investigation of a possible emigration is started. The Civil Registration System therefore has a very high quality. In short it can be said that the data quality of the Civil Registration System is very high, and that the problems which may arise are of little numerical importance.

A diagnostic for late registration of emigration (PM1u)

Around 17 percent of all emigrations in year X will be reported to the Civil Registration System later than February in year X+1. Producing direct register-based statistics on emigrations by time of *occurrence* would under-estimate the figures to a large extent, unless one is willing to wait till many years after the time of occurrence. Therefore the statistics on emigration (and immigration) is presented by the year of *registration* at the Civil Registration System and not year of occurrence. The following table provides a diagnostic on the effectiveness, when the statistics by year of registration is seen as an estimate of the number of migrations that takes place every year.

Year of occurrence and year of registration of emigration

		Year of occurrence							
		2008	2009	2010	2011	2012	2013	2014	2015
Year of registration	2008	43,490							
	2009	4,010	44,874						
	2010	1,173	3,812	45,882					
	2011	1,114	1,873	4,530	46,802				
	2012	346	697	1,127	4,091	47,988			
	2013	382	488	637	1,505	4,569	48,394		
	2014	257	423	590	972	1,639	5,491	49,218	
	2015	79	121	196	329	506	1,166	4,759	48,940

By publishing statistics by year of occurrence, there would have been 46,802 emigrations in 2011 noted in February 2012 (the date of publishing). By publishing the same statistics by year of registration a further (1,114+1,873+4,530+earlier figures) would have been added. This could be seen as an estimate for the emigrations that took place in 2011 but will be registered years later (4,091+1,505+972+329+later registrations).

A diagnostic for non-deregistration of emigration (CM1o)

It can be assumed that the population total does have a slight over-estimation in form of persons still registered in the register but having been away for several years. The diagnostic above shows the

observed late deregistration. A non-deregistration by definition cannot be observed directly. In this study, SoL activities are divided into three groups: demographic, health and socio-economic.

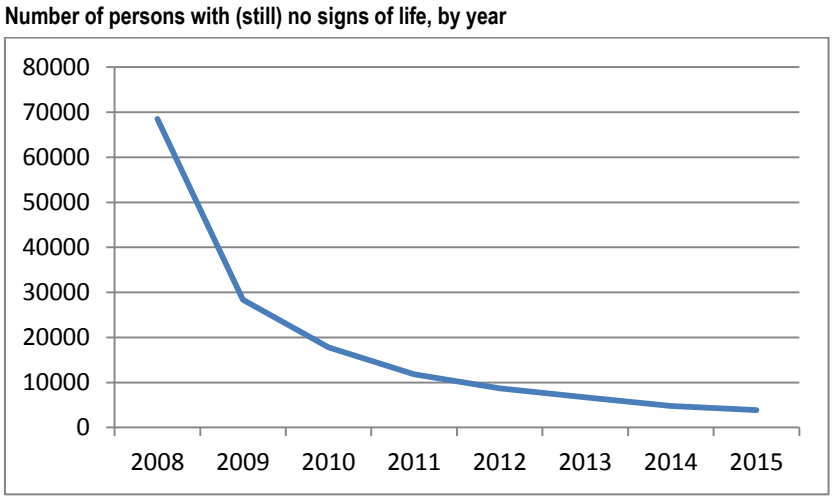
Socio-economic status
Self-employed
Assisting spouses
Top managers
Employees - upper level
Employees - medium level
Employees - basic level
Other employees
Employees - not specified
Unemployment
Subsidized employment without salary
Persons receiving holiday benefits
Guidance and activities upgrading skills
Unemployment benefit
Parental leave from unemployment
Maternity absence from unemployment
Sickness absence from unemployment
Cash benefit (passive)/cash benefit for Foreigners
Rehabilitation
Specially arranged scheme
Job clarification program
Enrolled in education

Demographics
Giving birth
Immigration
Change of marital status to married/divorced
Change of citizenship
Migration within the country

Health
General medical treatment
Specialist
Dentist/Dental hygienist
Chiropractor
Physiotherapist
Chiropodist
Psychologist
Other

The activities have been chosen since they require, to a large extent, that the person is in Denmark. The persons have been studied from 31st of December 2007 until 31st of December 2015, or until their death, their first emigration or disappearance during this period. A person can by the authorities be noted as disappeared and therefore be deregistered from the register. A person without any proven SoL during several successive years can be seen as a potential over-count.

Out of the registered Danish population on the 31st of December 2007 a total of 68,500 persons had no SoL activity in 2008. However, it can be noted that this group contains a large number of children who do not have an income of their own and who do not go to the doctors or dentists every year. Having two or more years with no proven SoL is therefore a stronger indication of over-coverage. Out of the population from the 31st of December 2007 a total of 28,355 does not have any of the above mentioned signs of life for the two years 2008 and 2009. The number of persons with three year of consecutive lack of signs of life is 17,765, and the number of persons with eight year of consecutive lack of signs of life is 3,892.



Appraisal

The SoL diagnostic is indicative but far from conclusive. In a sense, one might conclude that, for the reference time point October 31, 2007, the over-coverage (CM1o) is somewhere between 0 and 68500, since the number of (still) no SoL keeps decreasing every year. Moreover, a person may have left the country and then returned after some years, in which case the SoL can disappear for some years and then reappear afterwards. Therefore, even after many many years, there is still no assurance that the number of (still) no SoL is a valid estimate of total non-deregistration.

4.4.3 Italy: PM1u, PM3u, PM4

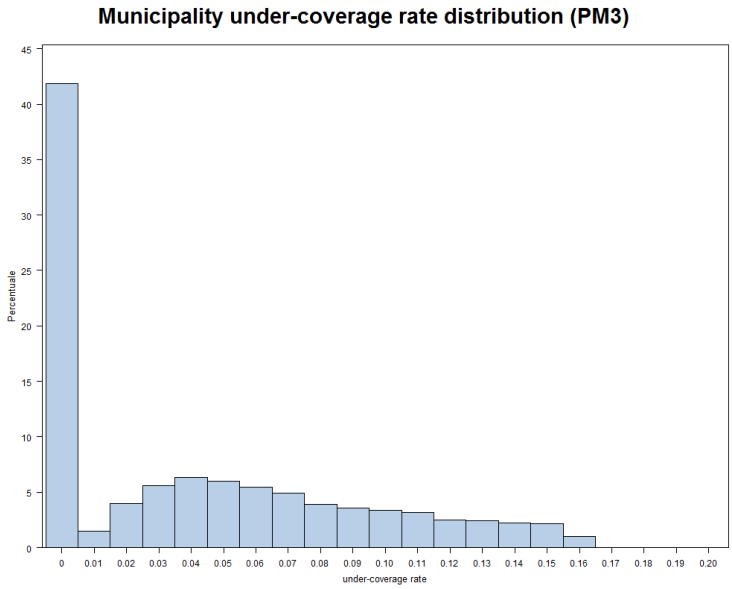
The diagnostic approach for progressiveness (Section 3.3.7) yields direct cross-tabulation between the frame for time t constructed at time t_1, i.e. $L(t;t_1)$, and that at time t_2 , i.e. $L(t;t_2)$. For this application categories are the Italian municipalities (H=8,048). The frame is LAC (Section 4.1.1). For this application, we have $t = t_1 = 1^{st}$ January 2014 and $t_2 = 1^{st}$ January 2015. However, for each entry (to LAC) event the date of the event and the date of its registration are available, whereas for events leading to an exit (from LAC), the corresponding record is simply deleted. As a consequence, only total and domain specific under-coverage at t_1 can be computed. With respect to domain misclassification, concerning the individuals moving from one municipality to another, the under-coverage in one place corresponds to over-coverage in another. Rates instead of totals for the corresponding items PM1u, PM3u, PM4 are computed.

Results

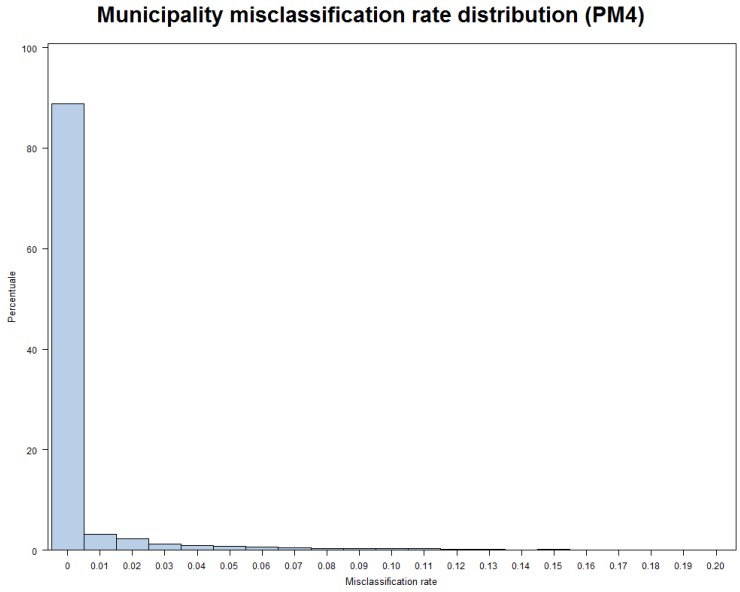
The total under-coverage rate, PM1u, is given by:

$$M = 100 * \sum_{h=1}^{H} N_{0h} / N^{t1} = 100 * 92627 / 60357255 = \mathbf{0,15}$$

The distribution of domain-specific under-coverage (PM3u) is given below. In about 42% of the municipalities, under-coverage due to progressiveness is null. For the remaining municipalities the under-coverage rate varies between 0,01% and 0,20%.



The distribution of domain-specific misclassification (PM4) is given below. The misclassification rate due to progressiveness is equal to zero in around 90% of the municipalities.



Appraisal

This application represents a preliminary analysis aimed at computing some progressiveness quality indicators. The approach proved to be easily applicable but allowed us to assess only a subset of the items PM1, PM3 and PM4, due to the lack of registration time point of exit events. Results on PM3s show higher municipality under-coverage rates than expected. The finding needs to be further investigated, e.g. by analysing the distribution of the time lag of the delayed entry events, which can provide hints on the most suitable time points for additional LAC collection within the year.

4.4.4 Lithuania: AM1-AM5

A simple diagnostic approach is applied to assess items AM1-AM5. The frame is the subset of the Population Register with declared address type. The BU is person, and the CU is the dwelling household. The 10 counties in Lithuania form the domains of BU, and urban/rural address the domains of CU. Rates are calculated in addition to totals as follows.

AM1. Correctly aligned base units rate: $R_t = \frac{N_t}{N}$, $N = \sum_{g=1}^{G}\sum_{h=1}^{H} N_{gh}$.

AM2. Correctly aligned domain base units rate: $R_{g;t} = \frac{N_{g;t}}{N_g}$, g=1,2,…,G.

AM3. Correctly aligned cell base units rate: $\left\{f_{gh;t} = N_{gh;t} / N_{gh}, g = 1...,G; h = 1,...,H\right\}$.

AM4. Correctly aligned composit units share: $C_t = \frac{M_t}{M}$, $M = \sum_{g=1}^{G}\sum_{h=1}^{H} M_{gh}$.

AM5. Correctly aligned domain composit units rate: $C_{h;t} = \frac{M_{h;t}}{M_h}$, h=1,2,…,H.

Declared addresses may be incomplete in the population register, if they lack of the house number. The problem is large in the rural area. Let U_{g2} be the individuals living in the rural area of county g. Put $U_{g2} = U_{g2;t} \cup U_{g2;e}$, consisting of $N_{g1;t}$ and $N_{g1;e}$ individuals with complete and incomplete addresses, respectively. The $N_{g2;t}$ individuals having house numbers are living at $M_{g2;t}$ addresses, for which the average household size is

$$r_{g2;t} = N_{g2;t} / M_{g2;t}\text{, } g\text{=1,2,…,}G.$$

For individuals belonging to $U_{g2;e}$ without house numbers, the number of addresses they are living in is unknown. Under the assumption that the distribution of household size, in the rural population of a county, is the same with or without house numbers, $M_{g2;e}$ in $U_{g2;e}$ can be estimated by

$$\hat{M}_{g2;e} = N_{g2;e} / r_{g2;t}\text{, } g\text{=1,2,…,}G.$$

The total number of addresses in the rural area of a county is $\hat{M}_{g2} = M_{g2;t} + \hat{M}_{g2;e}$. Similarly for the number of addresses in the urban area of a county $M_{g1;e}$, g=1,2,…,G. Overall, these yield

$$\hat{M}_1 = \sum_{g=1}^{G} \hat{M}_{g1}, \quad \hat{M}_2 = \sum_{g=1}^{G} \hat{M}_{g2}, \quad \hat{M} = \hat{M}_1 + \hat{M}_2.$$

Results

Distribution of BU (individual) and CU (address/dwelling houdehold)

County g	Urban Individual N_{g1}	Urban Adress $\hat{M}_{g1}$	Rural Individual N_{g2}	Rural Adress $\hat{M}_{g2}$	Total (individual)	N_g
1	89030	36607	65423	22832	154453	N_1
2	415906	172169	182473	61883	598379	N_2
3	243058	96557	100613	32145	343671	N_3
4	76294	30600	81551	27545	157845	N_4
5	143089	59785	101898	36019	244987	N_5
6	177850	75107	109495	39092	287345	N_6
7	44266	17571	65038	21306	109304	N_7
8	86564	34756	62591	19723	149155	N_8
9	80538	34245	65810	24699	146348	N_9
10	642311	254258	178522	59316	820833	N_{10}
Composite unit total		811656 $\hat{M}_1$		344560 $\hat{M}_2$	3012320 1156216	N $\hat{M}$

AM1-AM3, correctly aligned base units (individuals)

County g	Urban Correct $N_{g1;t}$	Urban Correct Rate $f_{g1;t}$	Urban Incorrect $N_{g1;e}$	Rural Correct $N_{g2;t}$	Rural Correct Rate $f_{g2;t}$	Rural Incorrect $N_{g2;e}$	Total correct $N_{g;t}$	Total Rate $R_{g;t}$
1	89030	1.000	0	37500	0.573	27923	126530	0.819
2	415809	1.000	97	135600	0.743	46873	551409	0.922
3	242884	0.999	174	74050	0.736	26563	316934	0.922
4	76255	0.999	39	51332	0.629	30219	127587	0.808
5	143079	1.000	10	67347	0.661	34551	210426	0.859
6	177785	1.000	65	88421	0.808	21074	266206	0.926
7	44256	1.000	10	43050	0.662	21988	87306	0.799
8	86559	1.000	5	48269	0.771	14322	134828	0.904
9	80514	1.000	24	39303	0.597	26507	119817	0.819
10	641667	0.999	644	118829	0.666	59693	760496	0.926
Total	1997838	0.999	1068	703701	0.694	309713	2701539	0.897

AM1. Total of correctly aligned base units N_t = 2701539, its rate is R_t =0.897.

AM2. There are clear between-county differences in terms of correctly aligned base units $N_{g;t}$.

AM3. Distribution of correctly aligned base units consists in columns $N_{g1;t}$ and $N_{g2;t}$; their rates are presented in columns $f_{g1;t}$ and $f_{g2;t}$. The rates for rural area are much lower. The problem is negligible for the urban areas.

AM4. Total of correctly aligned CUs $M_t = M_{1;t} + M_{2;t}$ = 1050218, with rate C_t =0.908.

AM5. Domain totals of correctly aligned CUs are $M_{1;t}$ =811228 and $M_{2;t}$ =238990, i.e. the number of correctly aligned addresses in urban and rural area. Their rates are $C_{1;t}$ =0.9995 and $C_{2;t}$ =0.694, i.e. a much lower proportion of the addresses in rural area are aligned correctly.

AM4-AM5, Correctly and incorrectly aligned CUs (addresses)

County g	Urban		Rural		Total
	Correct $M_{g1;t}$	Incorrect $\hat{M}_{g1;e}$	Correct $M_{g2;t}$	Incorrect $\hat{M}_{g2;e}$	correct $M_{g;t}$
1	36607	0	13087	9745	49694
2	172129	40	45987	15896	218116
3	96488	69	23658	8487	120146
4	30584	16	17338	10207	47922
5	59781	4	23806	12213	83587
6	75080	27	31568	7524	106648
7	17567	4	14103	7203	31670
8	34754	2	15210	4513	49964
9	34235	10	14751	9948	48986
10	254003	255	39482	19834	293485
Total	811228 M_{1t}	428	238990 M_{2t}	105570	1050218 M_t

Appraisal

Relative measures such as rates and proportions can be more informative than totals for between-domain or group comparisons. The average dwelling household size based on the exercise above is 2.6, which is considerably higher than the real household size survey estimates, e.g. 2.18 based on the EU-SILC survey 2015. The assumption of equal household size distribution by county and urban/rural is thus too simplistic. More refined modelling is needed in order to better account for the heterogeneity.

4.4.5 Denmark: IM1-IM2

DST applied a diagnostic approach based on unweighted omnibus sample proportions to explore the correlations between contact data and survey nonresponse. There are two types of contact: digital post and telephone, where the latter is further distinguished into 7 categories depending on the reliability of the telephone number available. The domains are formed by region and demographics.

Background

Response rates in social surveys have declined over many years. Phone books do not have the same coverage as they used to. Instead commercial marketing agencies and opinion polls often use web based survey respondent panels as the new sampling frame. It is however not possible to draw representative samples from panels where the inclusion mechanism is not known. At DST representative samples are drawn from the population register. The only contact information that is included in the population

register is postal address. At DST survey questionnaires are collected via Digital Post and telephone. Web questionnaires may also be used.

Citizens in Denmark are automatically provided with digital post (DP). All communication to public services goes through this DP by email. In order to correspond by normal mail (postal letter), citizens actively have to unsubscribe DP. Many people do not check their digital post on a regular basis, but it is possible to set up your digital post, to send you messages by email and phone when you have received digital post. Usage of digital post is still in an evolving phase and it is assumed that users will adjust more and more towards a use where all digital posts are noted. Messages to survey respondents in the population register can be sent via DP. For those not registered with DP, the messages are rejected. Phone numbers are then extracted from public telephone bases. However, there is no guarantee that respondents receiving messages by DP will actually open their messages.

Omnibus surveys in Denmark are performed on a representative sample of the population in the age 15 to 74 years, where the sample is drawn from the population register. Initially and at the same time all respondents are contacted through DP and postal letter. The postal letters will typically arrive a couple of days after the digital post. The respondents who have not answered the questionnaire, either as response to the digital post or as response to the postal letter, within 10 days are contacted through telephone provided telephone contact information is available, on which occasion the respondent is either guided to a website for filling out the questionnaire, or the respondent can go through the interview directly on the phone.

When creating the sample, other persons in the household are also extracted from the population register. Phone numbers from other persons in the household are also looked up in phone number databases. Respondents are looked up in phone books by name and address. If there is a perfect match with name and address the match quality is categorized as 'A-quality'. There are 7 levels of match quality. If several numbers are found the levels of quality is used to prioritize the order in which respondents are tried to be contacted. The levels are shown in the table below.

Category	Description
A	Perfect match with name and address
A2	Number matches name and respondent's previous address from the population register
A3	Match on address and names of other persons in household
B	Match on address but not name – landline phone number
C1	Match between a very low frequent name and address has the same zip-code
C2	Match between a low frequent name and address has the same zip-code
C3	Match on address but not name – mobile phone number

Almost two thirds of the respondents can be matched with an A-category telephone number, while for about 20% no telephone number can be found. The remaining categories are small but dominated by A3 which is the phone number subscribed by other persons in the household. For the analysis below all numbers not in group A are treated as a single quality category.

Results

Contact data in the omnibus survey of the 4th quarter of 2016 are used to compile IM1 and IM2 for DP and telephone, respectively.

It is seen that the penetration of DP is fairly equal across the various domains. There are two groups where the level of DP penetration is clearly lower than in the rest of the population and that is amongst the elder population (above 60 years) and amongst single person households. The penetration amongst respondents not having telephone contact information is a bit lower than amongst those having telephone contact information. However, of the 20.1% persons in the sample who have no contact telephone numbers, 92.2% of them can be contacted through DP, leaving only 7.8% of 20.1% (i.e. 1.6%) of all the respondents with no DP or telephone number.

IM1-IM2, DP penetration in omnibus survey, 4th quarter 2016

		Sample	Percentage with DP
Month	October	1656	94,2%
	November	1657	94,1%
	December	1675	94,7%
Sex	Male	2454	94,0%
	Female	2534	94,7%
Age	16-19	339	98,8%
	20-29	855	98,6%
	30-39	805	97,1%
	40-49	936	96,0%
	50-59	927	95,5%
	60-70	837	87,6%
	70-74	289	79,2%
Size of household	1 person	1107	87,2%
	2 persons	2661	95,7%
	3 persons	821	97,9%
	4 persons or more	399	98,0%
Region	Capital area	1687	94,4%
	Central DK	1147	94,2%
	Northern DK	528	93,9%
	Zeeland	629	93,6%
	Southern DK	997	95,2%
Telephone	No telephone contact information	1005	92,2%
	Telephone contact information	3983	94,9%
Total		4988	94,3%

When it comes to the quality of telephone numbers, it is more likely to have no phone numbers for persons who do not have digital post. There are two groups with low level of DP penetration (above 60 years and single person households). The high quality level of phone number matches for elder persons

compensates a bit for this, but among single person household only 56% have phone numbers with quality category A (compared to 64% in the total sample). Hence single person households both have lower level of DP penetration and poorer quality of phone numbers. Finally, the quality of phone number matches in the capital region is lower than in the other regions of Denmark. It is likely related to the high level of single person households in the capital region.

IM1-IM2, quality of telephone numbers in omnibus survey, 4th quarter 2016

	Cateogry-A	Other than A	No phone number
October	64,7%	15,5%	19,7%
November	65,5%	15,6%	18,9%
December	62,1%	16,1%	21,7%
Male	64,3%	15,7%	20,0%
Female	64,0%	15,7%	20,3%
16-19	55,5%	27,4%	17,1%
20-29	42,9%	25,4%	31,7%
30-39	54,4%	18,8%	26,8%
40-49	69,2%	12,0%	18,8%
50-59	71,8%	11,5%	16,6%
60-70	77,9%	10,3%	11,8%
70-74	82,7%	6,6%	10,7%
1 person	55,8%	10,4%	33,8%
2 persons	66,3%	15,9%	17,9%
3 persons	67,1%	19,7%	13,2%
4 or more persons	66,7%	21,3%	12,0%
Capital area	58,3%	17,9%	23,8%
Central Denmark	65,6%	15,7%	18,7%
Northern Denmark	67,8%	15,2%	17,0%
Zeeland	68,4%	12,7%	18,9%
Southern Denmark	67,6%	14,3%	18,1%
Not Registered for DP	63,5%	8,9%	27,7%
Registered for DP	64,2%	16,1%	19,7%
Total	64,1%	15,74%	20,15%

Regarding the success rate of establishing survey contact, there are clear interaction effects between DP penetration and age, as well as between DP penetration and the quality/availability of telephone number. The same is also largely true regarding the interaction effects between quality/availability of telephone number and DP penetration, and between quality/availability of telephone number and age. Although the interactions between quality/availability of telephone number and age seem somewhat less pronounced than those between DP penetration and age. Marginally speaking, the quality/availability of telephone number is a stronger predictor for successful survey contact than DP penetration. Of course, there may well be confounding factors underlying both the quality/availability of telephone number and the success rate of establishing survey contact, so one should avoid drawing any

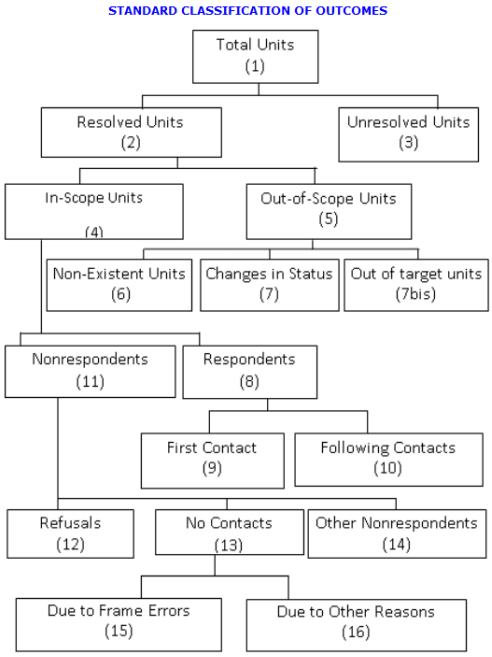
causal relationship between the two. Still, DP penetration almost doubles the contact rate among those without any telephone number, i.e. 38,1% vs. 19,2%. So it seems important in future to further enhance alternative digital means of contact.

Percentage of respondents with established contact by domain and contact information

	No DP	With DP	Category A	Other phone	No phone	Total
Sample size	282	4706	3198	785	1005	4988
October	58,3%	66,0%	76,3%	57,6%	36,4%	65,5%
November	58,2%	62,7%	71,5%	56,6%	36,0%	62,5%
December	59,1%	64,0%	75,3%	54,8%	37,4%	63,8%
Male	60,1%	64,7%	74,7%	55,7%	38,1%	64,4%
Female	56,7%	63,8%	74,0%	56,9%	35,2%	63,5%
16-19	50,0%	53,7%	61,7%	50,5%	32,8%	53,7%
20-29	8,3%	43,9%	58,9%	42,9%	22,9%	43,4%
30-39	34,8%	54,9%	65,3%	53,0%	32,9%	54,3%
40-49	51,4%	63,3%	69,8%	57,1%	40,9%	62,8%
50-59	50,0%	74,7%	79,7%	67,3%	51,3%	73,6%
60-70	66,3%	83,2%	85,3%	81,4%	53,5%	81,1%
70-74	75,0%	89,1%	92,5%	84,2%	38,7%	86,2%
1 person	51,4%	58,1%	72,7%	47,0%	35,0%	57,3%
2 persons	64,3%	68,3%	77,7%	60,3%	39,8%	68,1%
3 persons	82,4%	59,8%	68,4%	52,5%	30,6%	60,3%
4+ persons	50,0%	61,9%	68,8%	56,5%	31,3%	61,7%
Capital area	50,5%	60,4%	70,4%	56,3%	36,7%	59,9%
Central Denmark	68,7%	66,6%	75,1%	60,0%	42,8%	66,7%
Northern Denmark	65,6%	66,9%	78,2%	52,5%	34,4%	66,9%
Zeeland	62,5%	65,7%	74,9%	57,5%	37,0%	65,5%
Southern Denmark	52,1%	65,6%	76,9%	53,1%	30,0%	65,0%
No Digital post	58,5%		78,8%	36,0%	19,2%	58,5%
With Digital post		64,2%	74,1%	57,0%	38,1%	64,2%
Category A	78,8%	74,1%	74,4%			74,4%
Other phone	36,0%	57,0%		56,3%		56,3%
No phone	19,2%	38,1%			36,6%	36,6%
Total	58,5%	64,2%	74,4%	56,3%	36,6%	63,9%

Appendix A. Istat Standard Survey Outcome Chart and Quality Indicators

Istat currently collects and stores in the Quality Documentation System indicators for all the surveys Brancato, G. et al. (2004), and in particular the quality indicators for frames, are based on the following **Survey Outcome Chart (SOC)**:



STANDARD DEFINITIONS
Total Units (1): total number of units belonging to the survey of interest. For sample surveys, it corresponds to the number of sampling units
Resolved Units (2): units for which the eligibility status has been ascertained (vs unresolved)
In-Scope Units (4): units belonging to the target population
Out-of-Scope Units (5): unit not belonging to the target population, although included in the frame
Non-Existent Units (6): units that do not exist, although included in the frame due to errors or delays in updating the frame
Change in Status Units (7): units that changed their status becoming out of scope for the survey (e.g.: change of residence for a household, change of economic activity for an enterprise)
Out of Target Units (7bis): units that have never been in-scope, although included in the frame due to inclusion errors
No Contacts Due to Frame Errors (15): in scope units which have not been contacted due to errors or incomplete information in the frame (e.g.: wrong addresses)

The following **unweighted quality indicators** are routinely calculated:

Resolved Rate = [Resolved Units (2) / Total Units (1)] × 100

Overcoverage (Eurostat Quality Indicator) = [[Out-of-Scope Units (5) + (1-alpha)x Unresolved Units (3)] / Total Units (1)] × 100

(NB. Alpha= fraction of unresolved units that are considered eligible)

Frame Error Rate = [Out-of-Scope Units (5) + No Contacts Due to Frame Errors (15)) / Resolved Units (2)] × 100

Out-of-Scope Rate = [Out-of-Scope Units (5) / Resolved Units (2)] × 100

Non existent Rate = [Non-Existent Units (6) / Resolved Units (2)] × 100

Change in Status Rate = [Change in Status Units (7) / Resolved Units (2)] × 100

Out of Target Rate = [Out of Target Units (7bis) / Resolved Units (2)] × 100

No Contact Rate Due to Frame Errors (Referred to Resolved Units) = [No Contacts Due to Frame Errors (15) / Resolved Units (2)] × 100

COMPONENTS OF FRAME ERRORS RATE

Percentage of Out-of-Scope Units =[Out-of-Scope Units (5) / (Out-of-Scope Units (5) + No Contacts due to Frame Errors (15))]× 100

Percentage of No Contacts Due to Frame Errors Units = [No contacts Due to Frame Errors (15) / (Out-of-Scope Units (5) + No Contacts Due to Frame Errors (15))]× 100

COMPONENTS OF OUT-OF-SCOPE RATE

Percentage of Non-Existent Units = [Non-Existent Units (6) / (Non-Existent Units (6) + Change in Status Units (7) + Out of Target Units (7bis))]× 100

Percentage of Change in Status Units = [Change in Status Units(7) / (Non-Existent Units (6) + Change in Status Units(7) + Out of Target Units (7bis))]× 100

Percentage of Out of Target Units = [Out of Target Units (7bis) / (Non-Existent Units (6) + Change in Status Units (7) + Out of Target Units (7bis))]× 100

It is in principle possible to calculate the **design-weighted** counterparts, and the associated sampling errors, yielding **quality measure CM1** (over-coverage), and possibly **CM2** and **CM3** (genuine domain over-coverage) provided adequate domain sample sizes.

References

Berka, C., Humer, S., Moser, M., Lenk, M., Rechta, H. and Schwerer, E. (2012). Combination of evidence from multiple administrative data sources. *Statistica Neerlandica,* **66**, 18-33.

Brancato, G. et al. (2004). Standardising, Evaluating and Documenting Quality: the Implementation of ISTAT Information System for Survey Documentation – SIDI. Paper presented at the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, Germany, 24-26 May 2004.

Brancato, G. et al. (2006). Assessing Quality through the Collection and Analysis of Standard Quality Indicators: The ISTAT Experience, European Conference on Quality in Survey Statistics (Q2006), Cardiff, United Kingdom, 24-26 April 2006. http://www.statistics.gov.uk/events/q2006/downloads/T16_Signore.doc

Bryant, J.R. and Graham, P. (2015). A Bayesian approach to population estimation with administrative Data. *Journal of Official Statistics*, **31**, 475-487.

Colledge, M.J. (1995). Frames and Business Registers: An Overview. In Brenda G. Cox et al., ed. *Business Survey Methods*. Wiley, 21-48.

Di Cecco, D., Di Zio, M., Filipponi, D. and Rocchetti, I. (2016). Estimating population size from multisource data with coverage and unit errors. *In proceedings of* ICES-V 2016, Geneva.

Di Consiglio, L. and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, **31**, 415-429.

Dostál, L., Gabler, S., Ganninger, M. and Munnich, R. (2016). Frame correction modelling with applications to the German Register-Assisted Census 2011. *Scandevanian Journal of Statistics*, **43**, 904-920.

Eurostat (2011). European Statistics Code of Practice, available at
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-11-955/EN/KS-32-11-955-EN.PDF

Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, **59**, 409-439.

Gerritse, S.C., Van der Heijden, P.G.M. and Bakker, B. (2015). Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *Journal of Official Statistics*, **31**, 357-379.

Griffin, R.A. (2014). Potential uses of administrative records for triple system modeling for estimation of census coverage error in 2020. *Journal of Official Statistics*, **30**, 177-189.

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

Harron, K., Goldstein, H. and Dibben, C. (2016). *Methodological Developments in Data Linkage.* Wiley.

Hedlin, D., Fenton, T., McDonald, J.W., Pont, M. and Wang, S. (2006). Estimating the undercoverage of a sampling frame due to reporting delays. *Journal of Official Statistics*, **22**, 53-70.

Hendriks, C. (2014). *Improved input data quality from administrative sources through the use of quality indicators*. Paper to the Q-conference, Wien.

Hogan, H. (1993). The Post-Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, **88**. 1047-1060.

IWGDMF - International Working Group for Disease Monitoring and Forecasting. (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology*, **142**, 1047-1058.

IWGDMF - International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple-record systems estimation 2: Applications. *American Journal of Epidemiology*, **142**, 1059-1068.

Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. Wiley.

Mancini, L. and Toti, S. (2014). Dalla popolazione residente a quella abitualmente dimorante: modelli di previsione a confronto sui dati del censimento 2011. *ISTAT working papers, in Italian*.

Myrskylä, P. (1991). Census by questionnaire – Census by registers and administrative records: The experience of Finland. *Journal of Official Statistics,* **7**, 457-474.

Nirel, R. and Glickman, H. (2009). Sample surveys and censuses. *In Sample Surveys: Design, Methods and Applications, Vol 29A (eds. D. Pfeffermann and C.R. Rao)*, Chapter 21, pp. 539-565.

ONS - Office for National Statistics (2013). *Beyond 2011: Producing Population Estimates Using Administrative Data: In Practice*. ONS Internal Report, available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html

Renaud, A. (2007). Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results. *Survey Methodology*, **33**, 199-210.

Sirken, M.G. and Levy, P.S. (1974). Multiplicity estimation of proportions based on ratios of random variables *Journal of the American Statistical Association,* **69**, 68-73.

Snijkers, G., G. Haraldsen, L. Jones J. and D.K. Willimack (2013). *Designing and Conducting Business Surveys*. John Wiley & Sons, Inc.

Stoerts, R., Hall, R. and Fienberg, S. (2015). A Bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association, to appear*.

Struijs, P. And Willeboordse, A. (1995). Changes in Populations of Statistical Units. In Brenda G. Cox et al., ed. *Business Survey Methods*. Wiley, 65-84.

Tiit, E.-M. and Maasing, e. (2016). Residency index and its applications in censuses and population statistics. *Eesti statistika kvartalikri.* (*Quarterly Bulletin of Statistics Estonia*). 41-60.
http://www.stat.ee/publication-2016_quarterly-bulletin-of-statistics-estonia-3-16

Van Delden, A., Scholtus, S. and Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics, to appear*.

Werner, P. (2014). *Evalvering av Census 2011*. Statistics Sweden, internal report, in Swedish.

Wegfors, K. (2015). *Beskrivning av Registret*: *Registret över totalbefolkningen (RTB).* Statistics Sweden, in Swedish.

Wolter, K. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, **81**, 338-346.

Wright, T. and Tsao, H.J. (1983). A Frame on Frames: An Annotated Bibliography. In Tommy Wright, ed. *Statistical Methods and the Improvement of Data Quality.* Orlando, FL: Academic, 25-72.

Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics,* **27**, 415-432.

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

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, **31**, 381-396.

Zhang, L.-C. and Dunne, J. (2016). Trimmed Dual System Estimation. *In Capture-Recapture Methods in Social and Medical Sciences, eds. Dankmar Böhning, John Bunge and Peter van der Heijden.* CRC: Chapman & Hall, 239-259.