sciendo

Journal of Official Statistics, Vol. 38, No. 2, 2022, pp. 533–556, http://dx.doi.org/10.2478/JOS-2022-0025

Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes

Fabiana Rocci[1], *Roberta Varriale*[1], *and Orietta Luzi*[1]

Most National Statistical Institutes are progressively moving from traditional production models to new strategies based on the combined use of different sources of information, which can be both primary and secondary. In this article, we propose a framework for assessing the quality of multisource processes, such as statistical registers.

The final aim is to develop a tool supporting decisions about the process design and its monitoring, and to provide quality measures of the whole production. The starting point is the adaptation of the life-cycle paradigm, that results in a three-phases framework described in recent literature. An evolution of this model is proposed, focusing on the first two phases of the life-cycle, to better represent the source integration/combination phase, that can vary accordingly to the features of different types of processes.

The proposed enhancement would improve the existing quality framework to support the evaluation of different multisource processes. An application of the proposed framework to two Istat (Italian national statistical institute) registers in the economic area taken as case studies is presented. These experiences show the potentials of such tool in supporting National Statistical Institutes in assessing multisource statistical production processes.

Key words: Quality framework; multi-source processes; total survey error; statistical register.

1. Introduction

In recent years, the production of official statistics based on the combination of data from different sources has spread out in many National Statistical Institute (NSI), with the aim to reduce costs and response burden while delivering detailed and high-quality information on target populations and phenomena. In this view, new strategies in producing the required outputs need to be developed to move towards multisource processes exploiting as far as possible the integrated use of secondary data, possibly in combination with survey (primary) data. In this article, we refer to secondary data as data which *"are collected by others (i.e., not the NSI), used by an NSI for producing statistics and where the NSI has not defined the conceptual or process metadata"* (Memobust definition, Eurostat 2014), and in this context we focus on the use of administrative data, that are "data that is collected by sources external to statistical offices" (United Nations 2000).

Many new experiences have delivered important results, which can be considered at the basis of the modernization of NSIs. Along with these experiences, new production processes based on the integration of microdata from multiple sources are taking place,

[1] Istat, Via Cesare Balbo 16, 00184 Roma, Italy. Emails: rocci@istat.it, varriale@istat.it, luzi@istat.it

© Statistics Sweden

and new methodological issues are arising. A key issue to be considered relates to the need of a quality framework to assess the quality of these processes. In particular, nowadays NSI production processes are characterized by the intensive use of statistical registers: an appropriate quality framework should respond to the need for assessing the quality of the statistical register itself and its possible outputs. We refer to statistical register as a structured, regularly updated and authorized systematic collection of data and metadata (properties) at unit or event (object) level for a specific population carried out exclusively for statistics purposes.

The objects in the register are determined by definitions and classifications deriving from statistical criteria, and are unambiguously identified by a unique code (Wallgren and Wallgren 2014).

This article focuses on how to enhance the existing quality frameworks for a production process based on integrated sources, potentially both primary and secondary, in order to: (1) support the design of the production process, considering the possible different scenarios in terms of number and type of input sources; (2) monitor the process once it is put into production, identifying possible errors and ensuring that the design settings remain valid.

The starting point is considered to be the life-cycle paradigm (Groves and Lyberg 2010), according which the source of every potential error is identified in each phase of a given process. In the multisource context, the adaptation of the life-cycle paradigm is proposed by Zhang (2012) and applied by Statistics New Zeeland (2016), resulting in a two-phase life-cycle to represent a complete multisource statistical process. Afterwards, Reid et al. (2017) interprets this scheme as an application of the TSE, Total Survey Error (Biemer 2010) paradigm to the new realm of statistical production, which involves integrating and combining data from various sources. The authors start from Zhang's work and propose a three-phase framework where a phase concerning the final output is added. The three phases actually include: (1) a single source assessment, (2) an integrated data set assessment, and (3) an estimation and output assessment.

The analysis and application of the Zhang's two-phase framework to multisource processes developed by the Italian NSI (Istat) led to some considerations about the representation of all the actual process phases that are necessary to properly describe the process and, consequently, to identify the potential sources of errors.

Following Zhang's assertion that we should think to be in "*a pre-Neyman stage*" (Zhang 2012), more motivation arose to carry out more in-depth analyses, starting from the life-cycle concept, in order to investigate to which extent the existing frameworks could be possibly enhanced to catch all aspects and needs of representation and quality assessment for a multisource process.

The starting point is how to describe a process based on the integrated use of data sources, according to the available information on statistical units and variables. In this context, De Waal et al. (2020) describe some characteristics and the corresponding methodological issues of multisource statistics.

The aim of our work is to provide a more flexible tool to evaluate the quality of a multisource production process. Therefore, we propose a further evolution of the Zhang's two-phase life-cycle, to better describe the process of combining different data sources. In addition, to identify every potential error source we suggest an operational tool to connect the

steps of the production process to the phases of the quality evaluation framework. Finally, to enhance a vision embracing every type of statistical process, from the direct survey to new production strategies based on the (integrated) use of different types of data sources, we also propose to call it as Total Process Error (TPE). This would remind the need to carefully describe which type of data, modes of data collection and statistical outputs are involved in each phase of a statistical production process, to better take into account all its features.

In this article we do not discuss the third phase of the life-cycle introduced by Reid et al. (2017) concerning the evaluation of the process outputs: our aim is to propose an enhancement of the quality assessment throughout the multisource production process. This is expected to provide additional elements on the process quality, which are also useful in the evaluation of both the estimates and process outputs as introduced by Reid et al. (2017).

The article is organized as follows. Section 2 describes the current quality evaluation framework proposed by Zhang (2012) for processes using multiple sources. In Section 3, we introduce some considerations about this framework, highlighting the lack in representing and hence evaluating some steps of a multisource process. In Section 4, we describe the TPE framework, and in Section 5 we analyse the way to define it starting from Istat experience, with applications to two specific case studies in the area of economic statistical registers. Section 6 concludes the work with some reflections and future work.

2. The Current Quality Evaluation Framework for Processes Using Multisource Data

The reference literature of this article starts from the work by Zhang (2012), proposing a two-phase life-cycle model for integrated statistical microdata. The model provides a framework for the various potential error sources, and outlines some concepts and topics for quality assessment of multisource statistical processes. Zhang's framework can be interpreted as an extension of the Total Survey Error approach (TSE), since errors are linked to a life-cycle model. In 2016, Statistics New Zeeland applied and elaborated this framework describing the steps to assess the quality of an output or data set. Figure 1 and Figure 2

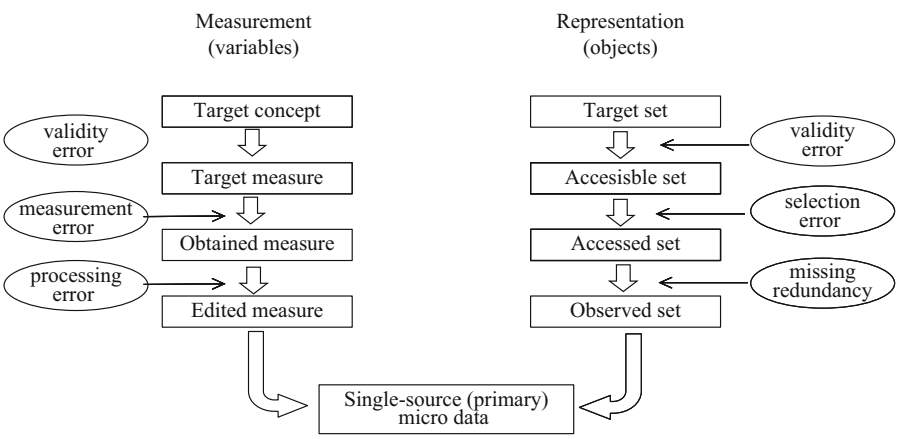

Fig. 1. Sources of error in phase one of Zhang's framework (Zhang 2012).

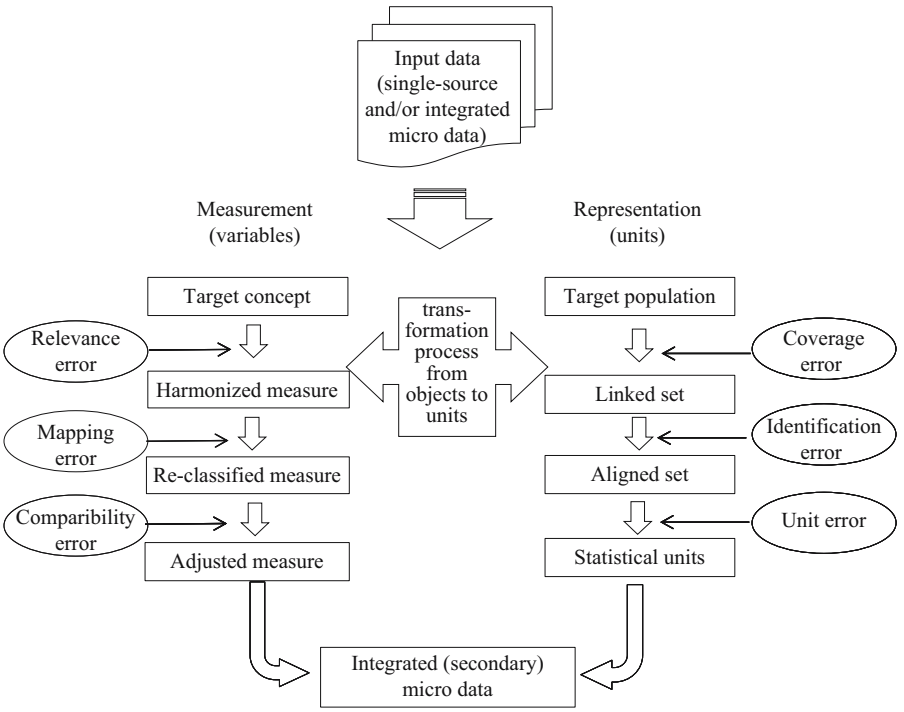

Fig. 2. Sources of error in phase two of Zhang's framework (Zhang 2012).

contain a graphical representation of the two-phase life-cycle diagram (Zhang 2012; Statistics New Zeeland 2016). The first phase, dealing with each single source, categorizes errors arising with respect to the original source's target objects and variables, in order to give a quality measure of the source itself. The second phase focuses on errors arising when data from several sources are combined or data from a single source are used to produce a statistical output, in order to assess the quality of the transformation process which is necessary to "adapt" the data from their original purpose to the statistical one. The input of this phase is the transformation from phase-one *object* and *variables* to phase-two *units* and *variables*, respectively, and accordingly the reference point corresponds to the statistical population and to the statistical concepts to be measured. Nevertheless, the input to the second phase is a data set represented by a single source and/or integrated micro data. The output is an (integrated) data set of micro data.

A third phase including the elaboration of the final statistical output was added by Reid et al. (2017). The latter model tries to comprehend an overall production process in the context of multisource data represented by administrative and survey data, where the integrated data set of micro data in Zhang's framework can play different roles according to the type of statistical process. Indeed, the integrated data set can be considered as the final output, complete for every observation and properly designed to achieve the statistical purpose, as in the case of the statistical registers. Otherwise, it can be an intermediate output: starting from it, statistical aggregates can be achieved (e.g., based on appropriate estimation methods).

The reflections by Reid et al. (2017) are quite straightforward and clear about the additional third phase. Furthermore, the authors address two interesting issues: the need of a more suited "statistical thinking" of the entire quality framework for the processes based on the use of multiple data sources, and the importance of a quality framework as a way to determine the strengths and limitations that different strategies (use of secondary data only, use of secondary data combined with direct survey data, use of survey data only) may have on the quality of a statistical output.

3. Considerations About the Existing Quality Framework for Processes Using Multisource Data

In many applications, the starting point to represent a statistical process based on the use of multisource data has been to reproduce what has been already defined for single direct surveys.

In general, researchers agree that major changes in the statistical production process call for a tailoring of the current approaches in terms of: (1) design, (2) implementation and (3) quality measurement and assessment.

In our experience, the existing two-phase quality framework described in Section 2 fails in representing (and evaluating) some steps of multisource production processes where important decisions about the design have to be taken. In particular, according to the sources' characteristics, the integration strategy (from phase-one to phase-two) can be straightforward or sometimes can be chosen among several alternatives that should be properly evaluated.

In order to understand whether and to which extent the life-cycle of a multisource statistical production process is appropriate and well represented, it is necessary to describe all the steps of the process itself, starting from the characteristics of the available data to the features of the statistical outputs. In this respect, De Waal et al. (2020) list eight basic situations of multisource processes, providing methodological guidelines to face every situation. In this section, selected situations are presented, to represent different scenarios of data integration. In presence of a complex picture of available data, several decisions need to be assessed during the process design and, at every release of the statistical process, the individual data sources may still be checked, but mainly to guarantee consistent quality each time a new version of the source is delivered to the statistical office.

For every multisource process, the analysis of the available data and information about the "statistical context" and of the required outputs is necessary. The statistical context is usually characterized by the target requirements such as the target population of N units composed by h strata, and the target statistical variables of interest Y_j $(j = 1, \ldots, P)$. Figure 3 represents a general scheme describing the statistical information context for a generic multisource statistical process, together with the target parameters θ_{hj} for strata h and variable j $(h = 1,\ldots, H;\ j = 1,..,P)$.

In a multisource context, the available sources S_q $(q = 1,\ldots,Q)$ providing information on the target statistical units and variables need to be analysed to assess their potential information. In particular, the following aspects are addressed:

- to verify that S_q $(q = 1,\ldots,Q)$ contains information about the phenomena under study;
- let $S_q(Y_j)$ $(q = 1,\ldots,Q;\ j = 1,\ldots, P)$ be the information on Y_j collected by source S_q, it is necessary to analyse the variable population coverage and whether its content is harmonized with the statistical definition of the target variable.

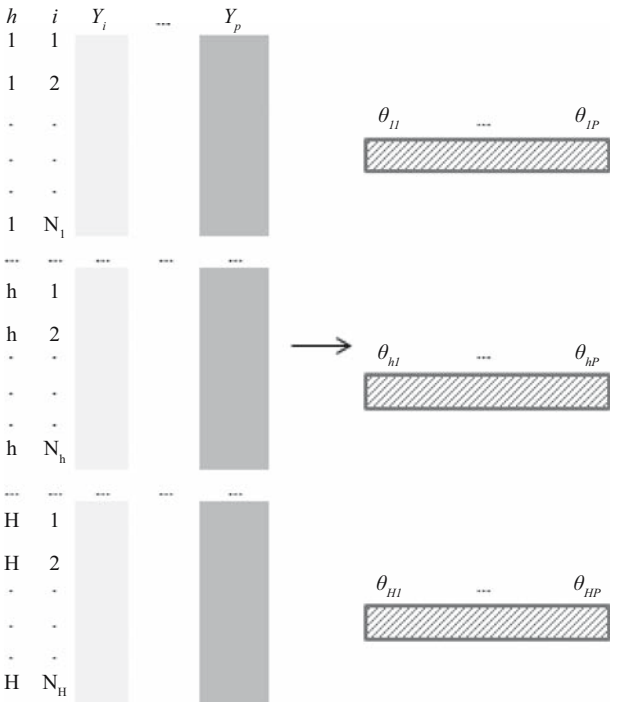

Fig. 3. Information context for a generic multisource statistical process.

Following De Waal et al. (2020), we go across several cases, to show how the methodology to integrate multisource data and to deliver the target estimates can be chosen among several strategies. The list of possible pictures varies according to the characteristics of the available data sources in terms of coverage of units and variables, existence of alternative sources for the same target variable, and so on. In the following, we present three extreme cases through graphical representations. In Figures 4, 5 and 6, the coloured areas correspond to the units covered by the available sources.

Case A: $\forall\ Y_j\ (j = 1, ..., P)\ \exists$ unique $S_q(Y_j)\ (j = 1,..., P;\ q = j)$, full units coverage.

Case B: $\forall\ Y_j\ (j = 1, ..., P_k)\ \exists$ unique $S_q(Y_j)\ (j = 1,..., p_k;\ q = j)$, full units coverage; $\forall\ Y_j$ $(j = p_{k+1}, ..., P)\ \exists$ unique $S_q(Y_j)\ (j = p_{k+1},..., P;\ q = j)$, no full units coverage.

Case C: $\forall\ Y_j\ (j = 1,...,P)\ \exists$ multiple $S_q(Y_j)$, that is, $(q = 1,..., Q;\ j = 1,...,P)$, full units coverage/not full units coverage.

So far, the two-phase life-cycle as proposed by Zhang clearly represent cases A and B. In case A, the errors may only derive from the transformation methods of the original source's objects and variables to the statistical target population and variables. The quality of both the multisource process and the derived target estimates depends on those methods. Nevertheless, in case B, the production of a complete statistical data set implies the use of microdata imputation methods: the introduced uncertainty is a component that needs to be taken into account when evaluating the quality of the entire process, apart from the transformation process. Case C represents situations where several available sources supply information about the same target statistical variables. In these cases, alternative strategies to achieve a statistical output can be feasible and we believe that the existing

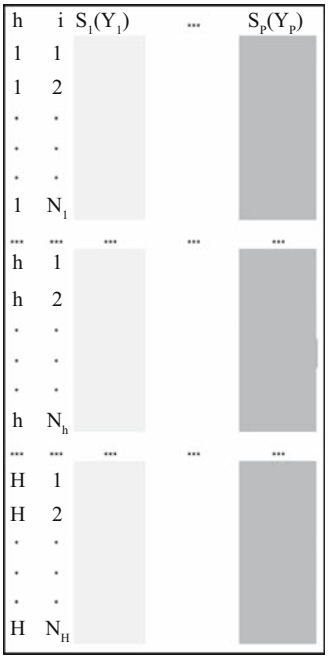

Fig. 4. Data sources non-overlapping in variables, absence of coverage issues (Case A).

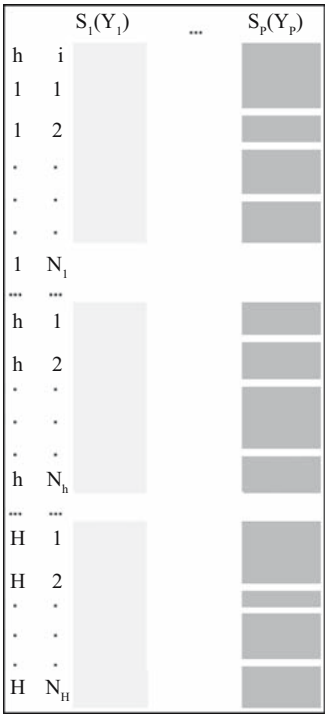

Fig. 5. Data sources non-overlapping in variables, presence of coverage issues for some variables (Case B).

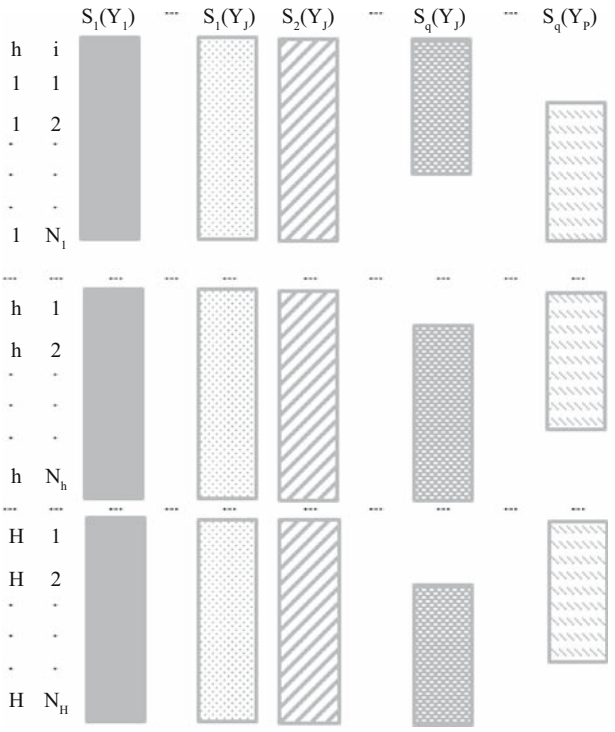

Fig. 6. Data sources overlapping in some variables, absence/presence of coverage issues (Case C)

quality framework lacks in precisely representing the more complex production processes. In this view, we propose a further specification of the Zhang model to create a more flexible total quality framework, that can be used both in the design phase or during the current process, once a specific strategy has been adopted.

4. A Proposal for the TPE Framework

Starting from the analysis on how to apply the framework proposed by Zhang (2012) and subsequently interpreted by Reid et al. (2017), a number of issues to be taken into account emerged.

First, it is important to observe how it is completely different to evaluate a multisource process in terms of: (1) a data set of statistical microdata, coming from the combination of different sources with full coverage for every target variable; (2) a statistical data set where the full coverage is obtained also through a micro imputation process; (3) final estimates of the target parameters, using different methodologies for each group of variables (and, in some cases, for each variable).

Following these considerations, some issues that needed to be further addressed were identified:

- there is a lack in literature of a well-defined vocabulary to better distinguish which kind of data, processes and outputs are involved in each phase. This is necessary in order to give a clear definition of the general framework of analysis,

- starting from the definition of statistical output provided by Reid et al. (2017), borrowed from the Organization for Economic Cooperation and Development (OECD) definition of a statistical product, there is a need, for each specific statistical process considered, to define and to distinguish different kinds of statistical outputs (e.g., full coverage statistical data set at microdata level such as a statistical register, estimates from the statistical data set, etc.) that can be obtained based on the use of multiple sources and to develop methods to ensure consistency among estimates. This is necessary in order to identify the most appropriate quality indicators in the different contexts, and
- the second phase of Zhang's framework should be further enhanced to trace every kind of actual assessment/integration/treatment phase of a process and better assess quality. The integrated micro data (cfr. Figure 2) can be obtained by using different integration strategies and treatments: as a consequence, in phase two it should be allowed to evaluate the effects of different alternative choices.

We propose a further evolution of the existing total quality framework, the TPE, which follows the life-cycle approach and takes into account every kind of statistical process.

The TPE is composed by two main phases, as in Zhang's framework, but we propose to split Zhang's second phase into two sub-phases to better identify the specific steps of the "transformation" process the original data have to go through, and to set up a system of indicators to evaluate each of them. It is worthwhile to note that an alternative modification to Zhang's proposal would be to include and describe different tasks/steps in carrying out the evaluations in phase 2. The reason why we decided to use sub-phases instead is because we want to stress that the process of quality assessment for a single source is quite different from a process of quality assessment for integrated data sources and, at the same time, we want to maintain Zhang's perspective of distinct quality assessment depending on the final purpose of each phase.

TPE framework can be represented as described in the following.

4.1. Phase 1. Assessment of Single Data Sources with respect to Original Source Purposes

The first phase of a production process based on secondary sources consists in the quality assessment of each data source with respects to its original purposes. This phase is carried out separately for every source, that can cover different populations and is characterized by a peculiar structure and contents. This phase coincides with Zhang's phase 1. As a consequence, the potential error types are the ones reported in Figure 1. When errors are highlighted, some treatment can be possibly applied accordingly to the type of errors (UNECE 2019), consistently with original purposes.

4.2. Phase 2. Combination/Re-Use/Integration of Data Sources with respect to Target Statistical Purposes

4.2.1. Phase 2a. Assessment of Single Data Sources with respect to Target Statistical Purposes

Each source is separately evaluated to assess its quality with respect to the specific statistical targets (statistical units/variables). This phase provides useful elements to define

the data selection and integration strategy, for example, when multiple sources are available for the same target variables and/or sub-populations.

4.2.2. Phase 2b. Assessment of the Combined Data Sources with respect to Target Statistical Purposes

In this phase, the integrated data set is generated, and a further quality assessment is carried out. This phase partly corresponds to the Zhang's phase 2. Additional actions should be taken into account in order to allow the evaluation of the complete production process. Actually, the integrated data set is usually treated to solve possible statistical inconsistencies (e.g., outliers), or to impute partially or totally missing information (usually resulting from the sources incompleteness with respect to target variables and under-coverage with respect to target population, respectively), and so on.

For each phase and each potential error, specific indicators can be proposed for quality assessment. It is worthwhile to note that some types of error (and the corresponding quality indicators) may appear in more than one phase (e.g., coverage error).

Furthermore, we propose to enrich the TPE by an operative tool, represented by a cross-classification scheme (Table 1), describing the link between the N process steps of the entire production process and the two phases of the evaluation framework. This scheme is useful in order to show *where* the decisions on the production process are taken and, therefore, to support the process design and monitor the entire process once it is put into production. Furthermore, the scheme allows to adapt the TPE in a very flexible way to represent different production processes.

Based on the TPE, a system of indicators can be applied, that is meant to help as guidelines to identify potential sources of errors, to measure their effect on the output and to prevent them, in order to progressively improve the new production system and make it continuously evolve. An example of such thinking is illustrated in Lothian et al. (2019). The authors suggest to store the information about the system of indicators in the metadata of an "evolutionary schema", that is a schema for integrating and linking traditional and non-traditional data sets.

Table 1. Cross-classification scheme: production process steps versus TPE phases.

Process steps	Phase		
	1. Assessment of single data sources with respect to original source purposes	2. Combination/re-use/integration of data sources with respect to target statistical purposes	
		2a. Assessment of single data sources with respect to target statistical purposes	2b. Assessment of the combined data sources with respect to target statistical purposes
1	. . .	. . .	. . .
2	. . .	. . .	. . .
. . .	. . .	. . .	. . .
N	. . .	. . .	. . .

In the next section, we describe the application of the TPE to two case studies taken from Istat current experience, including the cross-classification scheme, to help understanding to which phase each process step has to be evaluated. The aim is to show the flexibility of TPE and its applicability to very different situations, that are common to NSIs, for designing and/or monitoring multisource processes. In particular, for the first case study a possible system of indicators is also shown. These indicators represent an example, as other measures could be defined in each phase of the quality assessment process.

5. Istat Experience, Case Studies

In order to describe the application of the TPE to Istat production processes, it is necessary to provide some preliminary information on the Istat organization.

In Istat, a new model of statistical production was launched in recent years and is still ongoing, with the aim of moving from a direct survey-based statistical system towards a new production paradigm based on a system of statistical registers. The new system is designed to be powered by multiple sources both secondary and primary, and organizes the information on the target phenomena available at microdata level through the integration of those sources. At present, the secondary data are represented by only administrative data (AD). An important feature of the system of statistical registers is that it involves different kinds of registers, defined according to the role they play in the statistical production system, and strongly connected through specific rules among both target (sub)populations and variables. Usually the main distinction is between *base* registers, that represent the statistical reference populations for all the statistical processes (individuals and economic units) and the *satellite* registers, that release additional variables usually representing specific phenomena (Wallgren and Wallgren 2014).

In this section, we describe the application of the TPE to two satellite statistical registers in the economic area taken as case studies: the Register for Structural Business Statistics (Frame SBS) and the Register for Public Administrations (Frame PA). The former has been developed in Istat in recent years, and is regularly used since 2016 for the yearly production of statistics under the European SBS regulation. In this case, TPE is used to *monitor* the process. The process of the register Frame PA is still under construction: in this case, TPE is used to guide the process *design*.

In this context, Istat has centralized some functions common to many statistical production processes, such as data collection, implementing an integrated and centralized system for the acquisition of AD owned by public and private Institutions to support all the ongoing production processes (Runci et al. 2016).

In particular, this centralized system represents the repository of AD sources acquired by Istat, as it stores all the available information related to the AD "objects". For each data source, the system provides a unique and stable identification number for all the statistical units, such as individuals, economic units, places, and so on.

Hence, the production process of a *base* register aims to identify the statistical units belonging to the entire target statistical population of the register itself. On the other hand, the production process of a *satellite* register aims to extend the information of either the entire target population of a *base* register, or a subset of it.

It is important to note that the identification of the statistical units at the centralized system does not imply any treatment of the external sources information. Given these premises, the TPE in Istat is characterized by some key elements:

- AD acquired through the Istat centralized system can be equated to AD acquired from external bodies in phase 1 of TPE framework,
- linkage problems may arise only at the stage of identification of statistical units, and
- target populations are assumed to be completely defined by *base* registers: issues related to the alignment of the statistical population of a *base* register with the actual target population may be assessed by the application of a TPE to the production process of *base* registers.

5.1. *The Statistical Register for Structural Business Statistics (Frame SBS) As a Case Study*

In this section, we describe the application of TPE to monitor the production process of the *satellite* register Frame SBS (Luzi and Monducci 2016), that represents a complete microdata set extending information of the Italian *base* Business Register (BR). BR represents the target population of Italian active enterprises, contains structural information on enterprises, such as economic activity, number of employees and turnover, and is produced annually according to the reference EU regulation. Frame SBS is built for the annual release of statistics on loss and accounts of Italian enterprises, is designed with respect to the international agreement on enterprises accountability, and covers industry, construction, distributive trades and services, broken down to a very detailed sectoral level.

The design and implementation of the register is the result of the joint work of methodologists, information technology experts and subject matter experts. It has to be highlighted that, as the latter contributed to the definition of the methodological architecture of the register, they are fully aware of the overall level of reliability of the register outputs at micro and aggregated level. At the moment, only domain estimates are disseminated to external users (e.g., to Eurostat), with associated indirect measures of their overall quality (e.g., overall coverage rate of the variable from the integrated AD, imputation rates, and so on.). Internal users mainly use register microdata as auxiliary information in data modelling, in combination with other registers and/or survey data, or for calibration purposes to improve business survey estimates.

Traditionally, in Italy SBS was estimated based on two direct annual surveys: the sample survey on Small and Medium Enterprises (SME), which involves annually about 100,000 enterprises with less than 99 persons employed, representing a population of about 4.3 million of units, and the total survey on Large Enterprises (LE), involving annually about 11,000 enterprises with 100 or more persons employed.

SBS variables are covered by a number of AD sources managed by bodies external to Istat. These sources can provide information on the enterprises' accounting variables at microdata level. Such AD sources are the Financial Statements (FS), the Sector Studies survey (SS), the Tax Return (TR) data collected through different forms (Unico, Irap).

For the Y_j $(j = 1,...,K)$ SBS target variable, the situation in terms of primary and secondary sources availability is represented in Figure 7. As introduced, the target

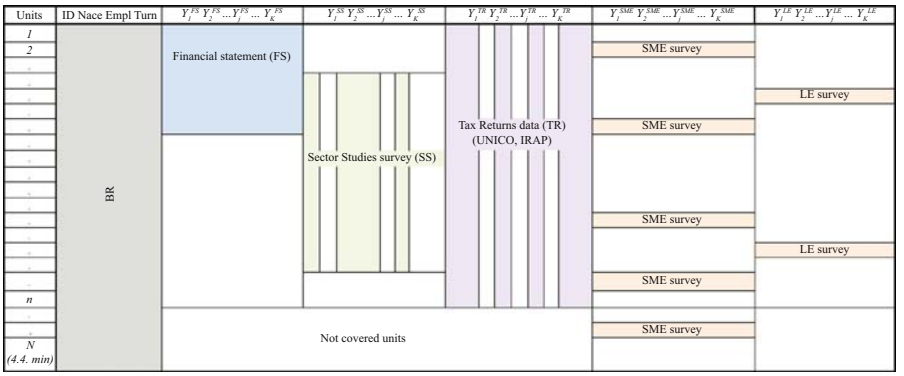

Fig. 7. SBS sources: administrative and survey data.

population is completely defined by the *base* register BR. Hence, two assumptions are made in Frame SBS production process: BR is equal to the target population and the AD sources have the same time stamp as the BR. Consequently, in this context no overcoverage is considered. For each source p ($p = FS, SS, TR, SME, LE$), the colored areas correspond to the covered enterprise sub-populations, identified based on the BR information. The variables Y_j^p ($j = 1,...,K$) represent the SBS variables covered by the p considered sources at microdata level. The variables Yj^{SME} and Y_j^{LE} relate to the SBS information collected by the annual sample survey on SME and LE.

At a first glance, the picture results as a chessboard, since some sources are overlapping but no one of them could cover the same set of variables neither the whole population. Many analyses were done, in order to assess how every source could be potentially used to achieve the more efficient result as possible in terms of coverage and accuracy of data. Furthermore, editing and imputation rules were applied to each source to ensure, when appropriate, internal consistency of each administrative object. The main issues that were observed during the preliminary analyses were:

- every AD source could guarantee a different degree of population coverage,
- administrative concepts of the different AD sources and the statistical concepts do not always agree, and
- difference of measurement on some of the variables present in different sources was registered.

Hence, two strategies for integrating the AD sources are possible:

Strategy A: for each statistical unit, after harmonizing the content of the variables Y_j^p ($j = 1,...,K$) into Y_j variables, integration of all available information coming from the AD sources is performed. Strategy A would maximize the overall quantity of information, but it does not guarantee internal consistency of each record.

Strategy B: a "priority" is assigned to every source (FS, SS and TR), based on its quality in terms of informative consistency with respect to the target phenomena, punctuality of the data supply, number of records, etc. After harmonizing the content of the variables, for each statistical unit of the register only one source is chosen, and the population coverage has different degrees. Strategy B would maximize the internal coherence of the data set, but provides a smaller quantity of information.

A constraint is that the final frame need to respect the internal consistency of information in each population unit, thus according to the adopted strategy different subsequent steps should have to be performed. In case of Strategy A, integrated data are treated/edited/imputed to ensure the internal consistency of each record. In case of Strategy B, subsequent treatment essentially consists in imputing missing information, as internal consistency for each record is already ensured in the original AD source. Strategy B has been chosen as it ensures the minimum amount of editing on the available data after the integration. This decision is peculiar of this production process and cannot be automatically extended to other processes without a proper evaluation.

Hence, the final design of the Frame SBS can be summarized as follows: different groups of variables have been identified, for which different production processes have been defined. These groups reflect different degrees of coverage of the AD sources and, therefore, different data quality levels. The sets of variables are:

- **Set of *BR* variables**: economic activity (Nace), Employment (Emp) and Turnover (Turn) of each enterprise,
- **Set of *core* variables:** The set of *core* variables Y_h $(h = 1, ..., H;\ H < K)$ that are the variables "highly" covered by the AD, so that the integration of different AD cover up to 95% of the target population for each variable. None of these variables is completely gathered by any data source, so that some partial and total unit non response is observed, and
- **Set of *components* variables:** The set of variables Y_j $(j = H + 1, ..., K;\ H < K)$ components of the *core* variables, which are not properly represented by AD.

Depending on the coverage and quality of administrative information, the component variables have been excluded from the Frame SBS process: the statistical register Frame SBS covers only the *core* variables Y_h $(h = 1, ..., H;\ H < K)$.

After deciding about the integration strategy and the set of *core* variables, the partial missing data on the integrated AD have been imputed, and, subsequently, totally missing units have been imputed to cover the total SBS target population. In general, imputation has been performed by using a combination of different methods, which have been applied to distinct groups of related variables, taking into account their distributional characteristics, their relationships with other variables, and exploiting all the available administrative information (Di Zio et al. 2016). It is worthwhile to note that measuring the additional uncertainty on register outputs due to imputation is still an issue under study at Istat (Di Zio et al. 2017; Alleva et al. 2021).

The output of this step is a frame (Figure 8) for the SBS target population defined by the Italian BR. It contains information on only the *core* variables Y_h $(h = 1, ...,\ H; H < K)$ at microdata level for *all* the units.

Summarizing, the process steps of the register Frame SBS are:

Step 1. Quality assessment on each AD source.

Step 2. Mapping of the coverage for every AD source for the whole system with respect to the K required variables (grouped in *core* and *component* variables) and the target population.

Step 3. Main decisions are taken about how to integrate AD sources.

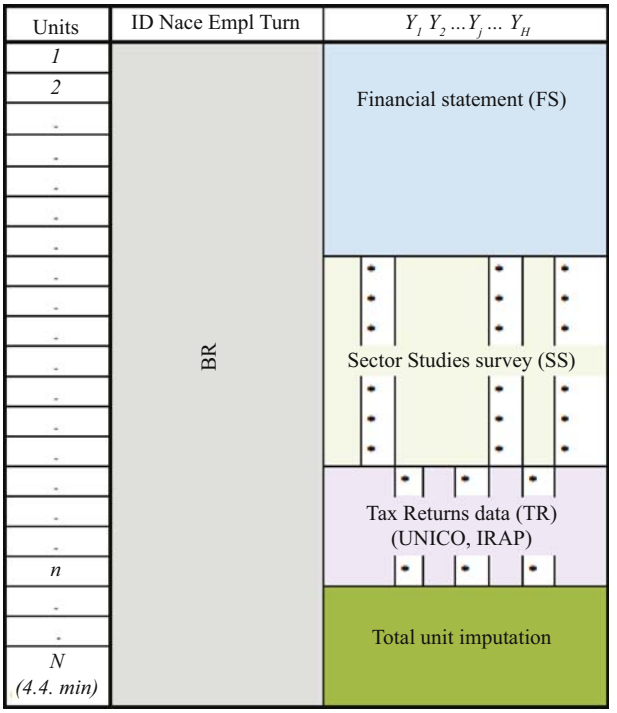

Fig. 8. Frame SBS statistical register: BR and H core variables.

Step 4. Imputation of the partial missing data on the integrated AD of the $H\ (H < K)$ *core* variables.

Step 5. Imputation of totally missing units of the $H\ (H < K)$ *core* variables to cover the total SBS target population.

The process steps have been cross-classified with the phases of the TPE framework, as shown in Table 2. In the table we used AD source instead of a generic data source as in Table 1 because, as already mentioned, the Istat system of statistical registers nowadays uses only AD as secondary data sources. As introduced, representing the process through the proposed scheme can help in understanding to which phase every process step has to be evaluated. In particular, TPE is used to monitor the quality of the entire production process that is repeatedly run, by identifying the source and phase of potential errors.

A first set of suitable indicators are proposed by phase, subject (variables, objects and units), process step and error type. Both quantitative and qualitative measures are considered. In Table 3, quality indicators for the assessment of the first phase of the Frame SBS production process are suggested. Tables 4 and 5 contain a draft proposal of quality indicators for phases 2a and 2b according to the proposed TPE. It is worthwhile to note that indicators in Tables 3, 4 and 5 represent some example indicators for Frame SBS. It is also important to note that in Table 3 we use the term "object" to be consistent with Zhang's proposal. As explained, in Istat the identification of the statistical units in each data source is carried out by a centralized system for the acquisition of AD. This implies only the attribution of an identification number without any processing action on the external information. In Tables 3, 4 and 5, additional indicators can be added or some indicators

Table 2. Frame SBS: production process steps vs TPE phases.

Steps	Phase		
	1. Assessment of single AD with respect to administrative purposes	2. Combination/re-use/integration of AD with respect to target statistical purpose	
		2a. Assessment of single AD with respect to target statistical purposes	2b. Assessment of the combined AD with respect to target statistical purposes
1	Quality assessment of each AD source (*FS, SS, Unico, Irap*)		
2		Quality assessment of each AD source (*FS, SS, Unico, Irap*) in terms of SBS purposes	
3			Integration of AD sources (*FS, SS, Unico, Irap*)
4			Prediction/imputation of the missing values of the *core* variables for partially uncovered units
5			Prediction/imputation of the *core* variables for totally uncovered units

may be changed to respond to specific users' needs in other contexts, by respecting the correspondence with the quality framework phases. As a simple example, when probabilistic linkage is used to integrate data sources, it may be helpful to produce indicators assessing the quality of this procedure, such as counts or summary statistics for covariates by categories based on the probabilities associated with matches/non-matches between pairs of sources, as well as quantifiable results of any manual follow-ups done to verify the matches and non-matches. As another example, when TPE is applied to assess the quality of a *base* register, overcoverage errors should be accounted besides undercoverage errors.

5.2. *The Statistical Register for* Public Administration *(Frame PA) As a Case Study*

In the new Istat system of statistical registers, the *satellite* register of the Public Administration (Frame PA) is under construction. In this section, we describe the application of TPE to Frame PA, to illustrate as TPE is used in this case to support and guide the process design.

Frame PA aims at releasing economic variables on a subset of the Italian PA Institutions. This subset includes a specific sub-population covered by the *base* business register related to the PA, that we will name "Register S13" (RS13). Frame PA will

Table 3. Phase 1 quality indicators by subject, phase and error type. Case study Frame SBS.

Phase 1 indicators	**Objects. Accessible set -> Accessed set; Selection error**	
	Proportion of missing units with respect to financial statements (FS) theoretical/target population	[1-*No. units in the source/ Total No. units in the theoretical/target population in FS*] x 100
	Proportion of units of business register (BR) population in the source, by source *S*	[1-*No. units in the source/ Total No. units in BR*] x 100
	Adherence to reporting period, for FS	*No. units that do not adhere to the reporting period/Total No. units* x 100
	Qualitative indicators, by source *S*	*Changes in population coverage (Does coverage change over time?)* *Updating of reporting units (How are changes recorded and actioned? Is it proactive or reactive?)*
	Objects. Accessed set -> Observed set; Missing/redundancy error	
	Percentage of multiple records, by source *S*	*No. units in Source S with multiple identification code/No.of unique identification codes* x 100
	Qualitative indicators	*Detecting duplicate records (Describe how duplicate reporting units are identified)* *Methods of treating duplicate records (Describe how duplicate reporting units are handled)*
	Variables. Process step: Target measure -> Obtained measure; Type of error: Measurement error	
	Punctuality, by source *S*	*Date of receipt - Date agreed*
	Lagged time between reference period and receipt of data	*Date of receipt by Istat - Date of the end of the reference period over which the data provider reports*
	Qualitative indicators, by source *S*	*Changes in administrative forms*
	Variables. Obtained measure -> Edited measure; Processing error	
	Proportion of units failing edit checks, by source:	*No. units failing edit checks/ Total no. units checked* x 100
	Proportion of units with all implausible values, by source *S*	*No. units with all values implausible (missing or 0 or 1) / Total n.of units checked* x 100
	Proportion of units with all missing values, by source *S*	*No. units with all values missing/ Total no. units checked* x 100
	Proportion of edit rules failed at least once, by source *S*	*No. failed edit rules for source S/ Total no. edit rules for source S* x 100
	Proportion of imputed values, by source *S*	*Total no. imputed values in source S/ Total no. values in source S* x 100
	Composition of the proportion of imputed values, by source *S*	$\frac{Tot.\ no.\ values\ changed\ from\ a\ code\ to\ another\ code\ in\ source\ S}{Total\ no.\ imputed\ values\ in\ source\ S} \times 100$ $\frac{Tot.\ no.\ values\ changed\ from\ missing\ or\ zero\ to\ a\ code\ in\ source\ S}{Total\ no.\ imputed\ values\ in\ source\ S} \times 100$ $\frac{Tot.\ no.\ values\ changed\ from\ a\ code\ to\ zero\ in\ source\ S}{Total\ no.\ imputed\ values\ in\ source\ S} \times 100$

Table 4. Phase 2a quality indicators by subject, phase and error type. Case study Frame SBS.

	Units. Target population -> Observed set; Coverage error	
Phase 2a indicators	Proportion of SBS population units in source FS	*No. corporate enterprises of SBS population in source FS/ No. corporate enterprises of SBS population* x 100
	Proportion of SBS population units in sources SS, Unico, Irap	*No. units of SBS population in source S / No. units of SBS population* x 100
	Variables. Target concept -> Harmonized measures; Relevance error	
	Qualitative indicators, by source *S*	*Changes in definitions of all variables in each source and changes in definitions of Structural Business Statistics (SBS) variables (Does definitions change over time?) Conceptual scheme representing the reclassification of administrative concepts needed to produce the SBS variable definitions*
	Variables. Harmonized measures -> Re-classified measures; Mapping error	
	Quantitative indicators, by source *S*	*Comparison of each harmonized variable with SBS benchmark variable (histograms, univariate statistics, statistical tests, etc.), to be repeated when variable definitions change*
	Proportion of target variables which not require reclassification or mapping, by source *S*	*No. variables captured directly from source S / Tot. no. variables* x 100
	Proportion of target variables which can be derived through reclassification or mapping, by source *S*	*No. variables derived from source S after reclassification/ Tot. no. variables* x 100

extend, for each unit, structural information coming from the RS13 with some economic variables obtained as the result of integration of data coming from administrative and survey sources.

Frame PA includes different sub-populations. Nowadays, Istat is working on the subpopulation of local Authorities (municipalities, unions of municipalities, provinces, mountain communities, metropolitan cities). The first step to build Frame PA is to select statistical units from RS13, together with some structural information, such as address, number of employees (Empl.), and so on.

The main AD sources concerning the economic variables of these units are the Public Administration database (BDAP) and the information system on the payment and financial transactions of public bodies (SIOPE). BDAP records the accounting variables of balance sheets according to the financial statement management schemes; SIOPE is a system of digital collection of profits and payments made by treasurers and cashiers of all public administrations. Therefore, BDAP collects information on stocks, while SIOPE on flows.

The first four variables Y_1, Y_2, Y_3, and Y_4 we are treating relate to the revenues of the Institutions. BDAP collects all variables, while SIOPE collects only Y_4. The variables Y_j^{BDAP} $(j = 1,...,4)$ represent the variables covered by BDAP, while Y_4^{SIOPE} represents information on Y_4 from SIOPE. Since we are dealing with AD with reference to two years before, subject matter experts expect that the two sources provide the same information on Y_4.

Table 5. Phase 2b quality indicators by subject, phase and error type. Case study Frame SBS.

	Units. Target population -> Linked sets; Coverage error	
Phase 2b indicato	Proportion of units of SBS population in the integrated data set (undercoverage). Also in longitudinal perspective.	*No. units of SBS population in the integrated data set/ No. units in the SBS population* x 100
	Proportion of units of SBS population in the integrated data set. Also in longitudinal perspective, by source *S*	*No. units of SBS population in the integrated data set from source S/ No. units in the SBS population* x 100
	Proportion of units of SBS population in the integrated data set with information present in only one source, by source *S*	*No. units of SBS population in only one source S/ No. units of SBS population in at least one source S* x 100
	Proportion of units of SBS population in the integrated data set with information present in more than one source	*No. units of SBS pop. in more than one source S/No. units of SBS population in at least in one source S* x 100
	Variables. Re-classified measures -> Adjusted measure; Comparability error	
	Proportion of units with influential values, by variable	*No. units with influential error/ Total no.of units* x 100
	Proportion of outliers, by variable	*No. units outliers/ Total no.of units* x 100
	Proportion of units with imputed values	*No. units with imputed values/ Total number of units* x 100
	Proportion of units failing at least one edit rule	*No. units failing edit checks/ Total no.of units checked* x 100
	Proportion of variable's values imputed, by variable	*No. units with imputed values for variable Y/Total number of unit* x 100
	Composition of the proportion of variable's values imputed, by variable	$\frac{\text{No. values of the variable Y changed from a code to a different code}}{\text{Total no. imputed values of variabel Y}} \times 100$ $\frac{\text{No. values of variable Y changed from missing or zero to a code}}{\text{Total no. imputed values of variabel Y}} \times 100$ $\frac{\text{No. values of variable Y changed from a zero to code}}{\text{Total no. imputed values of variabel Y}} \times 100$
	Impact of data editing and imputation on microdata, by variable	*Simple and quadratic distance between the pre-edited (Y) and post-edited (Y*) microdata of variable Y* $DL_1(Y_i, Y_i^*) = \sum_i N \lvert Y_i - Y_i^* \rvert$ */Total N. of units N* $DL_2(Y_i, Y_i^*) = \sqrt{\sum_i N (Y_i - Y_i^*)^2}$ */ Total N. of units* N_i
	Impact of data editing and imputation on distributions, by variable	*Kolmogorov-Smirnov distance on pre-edited and post-edited distributions* *Comparison of variable distributions (histograms, univariate statistics, etc.) pre- and post- editing and imputation*
	Impact of data editing and imputation on statistical relations between (set of) variables involved in the used models	*Pearson correlation index, Covariance matrix*
	Impact of data editing and imputation on statistical aggregates, by variable	*Tot. of the variable before editing and imputation /Overall total of the variable after editing and imputation* x 100

Variable Y_4 represents the total amount of revenues that each Institution receives during the year. The other variables represent, for each unit, specific components of the total Y_4. All the revenues are defined across 148 items. In the data, variable Y_4 is the only common information over the two AD sources; information on Y_j^{BDAP} $(j = 1,...,4)$, and Y_4^{SIOPE} is not necessarily present for each items; if Y_4^{BDAP} is present, Y_4^{SIOPE} should be present, and equal, and vice versa; if Y_4^{BDAP} is present, also Y_1^{BDAP}, Y_2^{BDAP} and Y_3^{BDAP} have to be present.

Figure 9 represents the theoretical scheme of the balance sheet including 148 items on the economic variables for each statistical unit; N is the total number of local Authorities, for each year. Symbol "X" used in Figure 9 represents the presence of information as an example.

From Figure 9, it is clear the complexity of the data structure underlying the register, which implies the difficulty of designing the register construction strategy. The use of TPE helps in splitting and describing the production strategy step by step.

As in Frame SBS, two strategies for integrating the AD were assessed.

Strategy A: for each statistical unit, integration of all available information coming from the two AD sources.

Strategy B: a "priority" is assigned to every AD source (BDAP and SIOPE), based on its quality. For each statistical unit of the register, only one source is chosen.

The choice of the integration strategy affects the subsequent steps that have to be performed to construct Frame PA. Following subject matter indications driven by the evaluation of AD quality, Strategy B was chosen. Since BDAP is evaluated to be the primary source of information and provides complete information on Y_4 for each statistical

Units	ID Address Empl	Item	Y_1^{BDAP}	Y_2^{BDAP}	Y_3^{BDAP}	Y_4^{BDAP}	Y_4^{SIOPE}
1	RS13	1					
1		2	X	X	X	X	
1		⋮					X
⋮		⋮	X	X	X	X	X
⋮		⋮					
⋮		⋮					
1		147					X
1		148	X	X	X	X	
2		1					
2		2					X
⋮		⋮					
⋮		⋮					X
⋮		⋮					
n		⋮	X	X	X	X	
		⋮					
⋮		⋮					
⋮		⋮	X	X	X	X	X
n		147					
n		148					
N		1	X	X	X	X	X
⋮		2					
⋮		⋮					
⋮		⋮					
⋮		147	X	X	X	X	
N		148					

Fig. 9. Frame PA AD sources: theoretical scheme of the balance sheet for the N statistical units (local Authorities).

unit, the subsequent treatment imputes data only when BDAP is totally missing. No action is done when information from BDAP and SIOPE do not correspond.

Figure 10 represents the situation in terms of source availability for the target variables. The target population is assumed to be completely identified by the RS13. Hence, two assumptions are made in Frame PA production process: RS13 is equal to the target population and the AD sources have the same time stamp as the RS13. For each source BDAP and SIOPE, the colored areas correspond to the covered units. For simplicity, Figure 10 does not report all variable names: in this application, information from each AD is present/absent with respect to all AD variables. $\text{BDAP}_{(\text{prev})}$ identifies information from BDAP source referred to the previous year rather than the one under analysis.

After integration, the imputation strategy for total missing units is under evaluation. The first step is to impute Y_4^{BDAP} directly using Y_4^{SIOPE}. Subsequently, Y_1^{BDAP}, Y_2^{BDAP} and Y_3^{BDAP} are imputed by using information from BDAP from the previous year, if present, otherwise from the same reference year. Median and nearest-neighbor donor with different strata combination are competing techniques.

The process steps of the register Frame PA are:

Step 1. Quality assessment of each candidate AD source: BDAP and SIOPE.

Step 2. Mapping of the coverage for every AD source with respect to the target statistical variables and the target statistical population (local Authorities).

Step 3. Main decisions are taken about how to integrate AD sources.

Step 4. Imputation of the total missing units with respect to BDAP source to cover the target statistical population, variable Y_4^{BDAP}.

Step 5. Imputation of the total missing units to cover the total RS13 target statistical population, variables Y_1^{BDAP}, Y_2^{BDAP} and Y_3^{BDAP}.

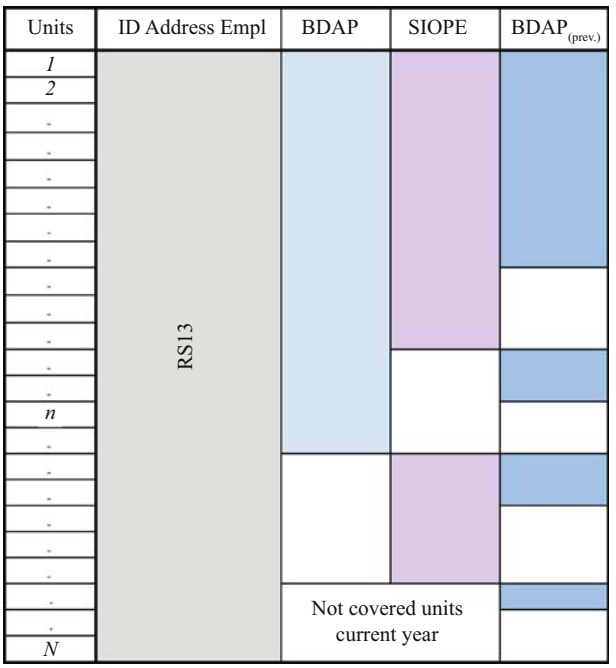

Fig. 10. Frame PA AD sources, by year.

Table 6. Frame PA: production process steps versus TPE phases.

Steps	Phase		
	1. Assessment of single AD with respect to administrative purposes	2. Combination/re-use/integration of AD with respect to target statistical purposes	
		2a. Assessment of single AD with respect to target statistical purposes	2b. Assessment of the combined AD with respect to target statistical purposes
1	Quality assessment of each candidate AD source *(BDAP, SIOPE)*		
2		Quality assessment of each AD source *(BDAP, SIOPE)* in terms of Frame PA purposes	
3			Integration of AD sources *(BDAP, SIOPE)*
4			Imputation of the total missing values of the variable Y_4^{BDAP}
5			Imputation of the total missing values of the variables Y_1^{BDAP}, Y_2^{BDAP} and Y_3^{BDAP}

It is worthwhile to note that in step 1 the term "candidate AD source" suggests a process that is still in development, where the final choice of which data sources to actually use has still to be made. At this stage, a lot of resources may be committed to thoroughly examine and compare the different candidate sources. Once the process will be set up for regular production, the choice of which data sources to use will be more or less fixed.

Table 6 cross-classifies Frame PA process steps with TPE phases. Also in this case, as in Table 2 we refer to AD source instead of a generic data source because, as already mentioned, the Istat system of statistical registers nowadays uses only this type of secondary data. For each phase and process steps, proper indicators are nowadays used to guide the design of the entire process. TPE is used as an instrument to identify potential source of errors and to measure their effect on the specific output of each phase. The final aim is to design the production process maximizing the quality of each step.

6. Conclusions and Future Work

In this article the TPE framework for the quality assessment of statistical processes using multiple data sources is proposed, starting from the scheme proposed by Zhang (2012), applied by Statistics New Zeeland (2016) and further developed by Reid et al. (2017). An in depth analysis of the existing frameworks in terms of life-cycle of a multisource process and the corresponding phases, where different types of errors can occur, has shown at this stage some additional gaps regarding how different decisions can be taken about combining data sources. In this article, we propose to split the second phase of the Zhang's

framework into two sub-phases, to better identify the different patterns the process can go through, taking into account all the features each data source present across time. An operational tool helps the application of TPE for different production processes, by cross-classifying the production process steps with respect to the TPE phases.

Since the processes using multiple data sources are not yet fully standardized, the TPE has been designed to be as flexible as possible, and to provide information useful to modify the processes according to possible changes in the input sources.

The identification of error sources in a multisource production process represents the basis for the systematic and continuous improvement of the quality of the entire process and its derived outputs, through the prevention/elimination (or at least the reduction) of such errors in the subsequent replications of the production process itself. The availability of quality indicators for different reference years will also allow the analysis of both data and process quality in a longitudinal perspective. In addition, based on the quality framework, a complete quality report could be developed for documentation and dissemination purposes.

This proposal represents an initial step of a more comprehensive project. First of all, there is a need of a well-defined vocabulary to describe which kind of data, processes and statistical outputs are involved in each phase. Furthermore, the establishment of an enhanced two-phase quality framework is expected to be the basis for further developments. As introduced, the next step will be to complete the framework including an "output validation phase", as already proposed in Reid et al. (2017).

Another important issue to be addressed is extending the TPE to multisource processes using also "new" types of secondary sources of data, such as big data, in combination with "traditional" secondary data (i.e., AD). Even if from a theoretical point of view TPE is useful to assess the quality of any multisource process, additional analyses are necessary in this area.

7. References

Alleva, G., P.D. Falorsi, F. Petrarca, F. and P. Righi. 2021. "Measuring the Accuracy of Aggregates Computed from a Statistical Register." *Journal of Official Statistics.* DOI: https://doi.org/10.2478/jos-2021-0021.

Biemer, P.P. 2010. Total Survey Error, Design, implementation and evaluation. *Public Opinion Quarterly*, 74 (5): 817–848. DOI: https://doi.org/10.1093/poq/nfq058.

De Waal, T., A. van Delden, and S. Scholtus. 2020. Multi-source Statistics: Basic Situations and Methods. *International Statistical Review* 88: 203–228. DOI: https://doi.org/10.1111/insr.12352.

Di Zio, M., L.C. Zhang, and T. de Waal. 2017. "Statistical methods for combining multiple sources of administrative and survey data." *The Survey Statistician* 76: 17–26. Available at: http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_July_20171.pdf (accessed March 2022).

Di Zio, M., U. Guarnera, and R. Varriale. 2016. "Estimation of the main variables of the economic account of small and medium enterprises based on administrative sources". *Rivista di Statistica Ufficiale*. N.1/2016. Available at: https://www.istat.it/it/files//2016/11/4_guarnera.pdf (accessed March 2022).

Eurostat. 2014. *Memobust Handbook on Methodology of Modern Business Statistics.* Available at: https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en (accessed March 2022).

Groves, R.M., and L.E. Lyberg. 2010. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74 (5): 849–879. DOI: https://doi.org/10.1093/poq/nfq065.

Lothian, J., A. Holmberg, and A. Seyb. 2019. An Evolutionary Schema for Using "it-is-what-it-is". *Journal of Official Statistics*, 35 (1): 137–165. DOI: https://doi.org/10.2478/jos-2019-0007.

Luzi, O., and R. Monducci. 2016. "The new statistical register Frame-SBS: overview and perspectives." *Rivista di Statistica Ufficiale*. N.1/2016. Available at: https://www.istat.it/it/files//2016/11/1_luzi.pdf (accessed March 2022).

Reid, G., F. Zabala, and A. Holmberg. 2017. "Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ." *Journal of Official Statistics*, 33(2): 477–511. DOI: http://dx.doi.org/10.1515/JOS-2017-0023.

Runci, M.C., G. Di Bella, and L. Galiè. 2016. *Il sistema di integrazione dei dati amministrativi in Istat.* Istat working paper, 18. Available at: https://www.istat.it/it/files//2016/11/IWP_18_20161.pdf (accessed March 2022).

Statistics New Zeeland. 2016. *Guide to Reporting on Administrative Data Quality*. Available at: https://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admindata-quality.aspx (Accessed March 2022).

UNECE. 2019. *Generic Statistical Data Editing Model.* Version 2.0, June 2019, Available at: https://statswiki.unece.org/display/sde/GSDEM (accessed March 2022).

United Nations. 2000. *Terminology on statistical metadata*. United Nations Statistical Commission and Economic Commission for Europe, Conference of European statisticians, statistical standards and studied, Geneva. 53. Available at: https://ec.europa.eu/eurostat/ramon/coded_files/UNECE_TERMINOLOGY_STAT_METADATA_2000_EN.pdf (accessed March 2022).

Wallgren, A., and B. Wallgren. 2014. *Register based statistics: Administrative data for statistical purposes*. John Wiley & Sons, Ltd.

Zhang, L.C. 2012. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* 66 (1): 41–63. DOI: https://doi.org/10.1111/j.1467-9574.2011.00508.x.

Received February 2020
Revised October 2020
Accepted August 2021