THE STATISTICAL REGISTER FOR PUBLIC ADMINISTRATIONS: THE REFERENCE FRAMEWORK AND SOME METHODOLOGICAL ASPECTS[1]

Roberta Varriale, Nevio Albo, Cecilia Casagrande, Valeria Olivieri

Abstract. During the last decade, the Italian National Institute of Statistics has been engaged in a modernization program involving the use of statistical registers integrated into a single logical environment, the Italian Integrated System of Statistical Registers (ISSR), for supporting the consistency of statistical production processes and improving the quality of information for users. One object of the ISSR is the satellite statistical REgister for Public Administrations (REPA) that contains information on structural and economic variables on a subset of the Italian Public Administrations (PA). This subset includes specific sub-populations covered by the base business register related to the PA. Therefore, REPA extends, for each of those units, structural information coming from the base register with some economic variables obtained as the result of integration of data coming from administrative and survey sources. In this paper we describe some methodological aspects of the design and implementation of the production process, together with the structural metadata and the proposal of a structural variable for the functional classification of the statistical units.

1. Introduction

During the last decade, the Italian National Statistics Institute (Istat) has been engaged in a modernisation programme involving the use of statistical registers integrated in a single logical environment, the Italian Integrated System of Statistical Registers (ISSR) (Luzi *et al.*, 2019). ISSR comprises a series of statistical registers (basic, thematic and extended) that centralise and integrate data from administrative sources, statistical surveys carried out by the institute and new and emerging sources for the ongoing production of official statistics. The ISSR aims to ensure uniform management of the different themes (social, environmental, economic statistics, etc.) and conceptual, statistical and physical integration between the statistical units that make it up. One of the objects of the ISSR is the extended statistical REgister for Public Administrations (REPA), which includes the subset of Italian Public

Administrations in the so-called "S13 list" produced by Istat (https://www.istat.it/it/archivio/190748). In this paper we consider the units in the S13 list together with some structural information as the base Register of REPA and we will call it "Register S13" (RS13). Therefore, REPA extends the RS13 with some economic variables obtained as a result of the integration and elaboration of data coming from administrative and survey sources. The production process of the statistical Register for Public Administrations includes 3 objects: (i) the information base created by the integration of sources; (ii) the REPA itself; (iii) Frame PA.

The structure of the paper is as follows. Section 2 describes the production process of REPA and Frame PA for the whole reference population in the S13 list of public institutions, and section 3 focuses on the sub-population for Territorial Governments. Section 4 presents the metadata of REPA. Section 5 describes the new structural classification of institutions. Section 6 contains concluding remarks.

2. REPA and Frame PA

In the Section, we describe the production process of REPA and Frame PA for the whole reference population in the S13 list.

2.1. The units and the variables

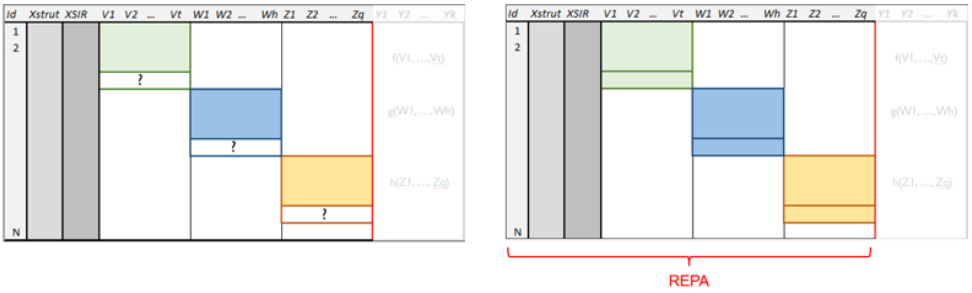
REPA is an extended Register of the base RS13, including all institutional units that are part of the general government sector and whose accounts contribute to the compilation of the consolidated profit and loss account of the General Government of Italy. RS13 is compiled according to ESA 2010, as defined by Regulation (EU) of the European Parliament and of the Council no. 549/2013, and on the basis of the interpretations of the ESA, provided in the Manual on Government Deficit and Debt published by Eurostat (Eurostat, 2019). According to the Regulation, RS13 is divided into 3 institutional sub-sectors: Central government (excluding social security funds), Local government (excluding social security funds) and Social security funds. It is possible to classify each sub-sector into different institutional typologies, some of which already populate the prototype version of REPA for Territorial Government. REPA contains a subset of the structural variables from RS13 and ISSR, a variable related to the proposed new structural classification of institutions, and a set of micro-data of an economic nature for each type of public institution. The structural variables are: identifiers and register variables, territorial variables, stratification variables, activity status, date of inclusion and possible exclusion from sector S13, transformation events. One of the ISSR variable is *Employees*. Frame PA is derived from REPA by aggregating REPA data into homogeneous variables, processable for the entire reference population and referred

to as the "Frame PA variables": *Current revenues*, *Compensation of employees*, and *Purchases of goods and services*.

2.2. REPA and Frame PA production process

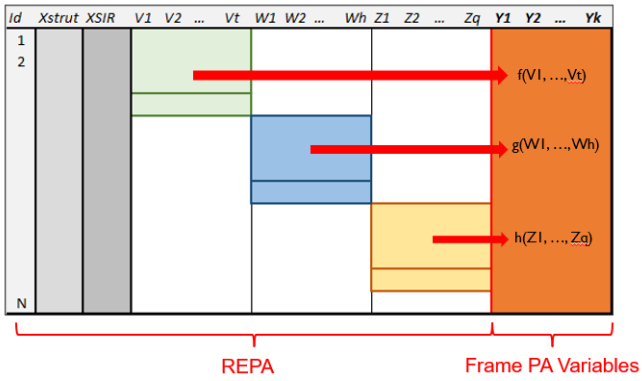
The REPA and Frame PA production process works in a differentiated way for different groups of institutions (sub-populations), defined through the classification of some structural variables (see Section 4.2). Its representation is in Figures 1 and 2, by using, as an example, three different sub-populations represented with different colours. *Id* is the identification code of each unit belonging to the RS13 and *Xstrut* are the structural variables coming from this Register together with the variable "New structural classification of institutions". *XSIR* represents the variables from the ISSR. Variables *V*, *W*, and *Z* are the variables of REPA for each sub-population, and *Y* are Frame PA variables that represent the final output of the process. Each sub-population has a different information structure (content of variables, sources, etc.). Figure 1.a shows how the output of data collection, harmonisation into statistical concepts and integration is differentiated by sub-population. This is followed by a phase of review, editing and imputation of total and partial non-response: the result is a complete dataset for each sub-population (Figure 1.b).

Figure 1 – *REPA production process: (a) Data collection, harmonization, and integration, (b) Data review, edit and imputation.*



As shown in Figure 2, this is followed by a process of transforming the original information into the output variables that will make up the Frame PA. This process is also differentiated according to the sub-populations.

Figure 2 – *REPA production process: Frame PA variables.*



REPA is still under development, the design and implementation of the Register is at a good stage for the sub-population of Territorial Governments (Varriale *et al.*, 2021). In the following section we describe in details its production process.

3. REPA and Frame PA for Territorial Government production process

In the Section, we describe the production process of REPA and Frame PA for the Territorial Governments (TG); we will refer to REPA TG as REPA.

3.1. The units and the variables

TG include regions and autonomous provinces and local governments i.e. municipalities, unions of municipalities, provinces, mountain communities and metropolitan cities. In 2019, the population of TG consists in 8749 units, representing the 84.3% of the total RS13 population.

The economic variables of REPA include accrual and cash values, for both revenues and expenditures. The accrual data for the revenue are the assessments (E1) while the cash data are the collections in accrual (E2) and the residual accounts (E3). For expenditures, the accrual data are the commitments (S1), the cash data are the payments on accrual (S2) and the residual accounts (S3). The information for both revenues and expenditures is organized into several hierarchical levels. By aggregating a selection of items, we obtain the economic variables of Frame PA (Guandalini *et al.*, 2022).

3.2. REPA and Frame PA production process: data sources and imputation

The primary source of information of REPA is the Public Administration Database (BDAP), providing all the information needed for REPA and Frame PA economic variables. BDAP will cover in the future also other types of institutions.

For the population of the regions and autonomous provinces, the BDAP source has no total non-response. Therefore, the variables REPA and Frame PA are obtained through a transcoding procedure. On the other hand, the population of local governments is characterised by total non-response. Before applying the transcoding procedure to obtain the REPA and Frame PA variables, a data imputation step is necessary. The imputation procedure uses an auxiliary information source, i.e. the Information System on the Operations of Public Bodies (SIOPE).

The entire REPA and Frame PA production process is repeated over time, using the same RS13 reference year and different provisions of both BDAP and SIOPE. It is important to note that BDAP and SIOPE are related to the same reference universe, the units belonging to the S13 list, but have a different periodicity in terms of availability of the data. For BDAP there are six different provisions relating to data referring to the same period, while SIOPE's provision is "continuous". In a generic year, at time *t* in the July month there is the first provisional supply of the BDAP's data referring to time *t-1*. The definitive provision usually takes place in May of the following year and the data are referred to time *t-2*. Just to give an example, let's consider the production process of REPA, reference year 2019: data from RS13 refer to 2019, the first BDAP provision for the same reference year is in July 2020, and the last BDAP provision is in May 2021.

BDAP provisions are characterised by a gradually decreasing non-response rate. Therefore, with regard to the 2019 data, the non-response rate for the total of local units decreases from 20.3% for the July 2020 BDAP provision to 7.4% and 3.6% for the following two provisions, but it remains high for mountain communities (from 58.3% to 52.3% and 44.4%) and for unions of municipalities (from 38.3% to 31.0% and 24.4%). The cases of total non-response of the BDAP source are distinguished according to the presence of information in the SIOPE source. The assumption underlying the imputation process is that the BDAP source is complete, i.e. if the BDAP source is available, no imputation is necessary.

The imputation method is described below. In the following notation, *V1-V3* denote both variables 1 to 3 of revenue (*E1,E2,E3*) and variables 1 to 3 of expenditure (*S1,S2,S3*) coming from BDAP. Variable *V4* is the result of the sum of *V2+V3* and *V4SIOPE* is variable *V4* coming from SIOPE, both for revenue and expenditure.

The first assumption underlying the imputation procedure is that for each unit *i* and for each item (148 for revenue and 22 for expenditure): $V4_i = V4SIOPE_i$, where $V4_i$ and $V4SIOPE_i$ represent the value of *V4* and *V4SIOPE* for unit *i*, respectively.

Therefore, the first step in the imputation process is to impute variable *V4* with the value of variable 4 from SIOPE: $V4_i^* = V4SIOPE_i$, where $V4_i^*$ is the imputed value of variable *V4* for each non-responding institution *i*. Subsequently, to impute the variables *V1* and *V2*, we use the median ratio between *V1* and *V4* - *r1(t)=V1(t)/V4(t)* - and *V2* and *V4* - *r2(t)=V2(t)/V4(t)* - calculated in each imputation stratum by using of all the information collected on the respondent units at time *t*. Therefore, the imputation of variables *V1* and *V2* is carried out by the relations:

- $V1_i^*(t) = MEDstr[r1(t)]\ V4_i^*(t)$
- $V2_i^*(t) = MEDstr[r2(t)]\ V4_i^*(t)$

where *MEDstr[r1(t)]* and *MEDstr[r2(t)]* are the median of *r1(t)* and *r2(t)*, calculated in appropriate imputation strata defined by the variables: Region, Institutional typology of entities, and aggregations of items in "Piano dei conti". Variable *V3* is then obtained through the relationship *V3=V4-V2*.

The choice of the imputation method was based on different types of evaluation. In particular, the two main working hypotheses for imputing missing values at time *t* were: (i) *cross method*: to use as auxiliary one the information collected in all responding institutions at time *t*; (ii) *longitudinal method*: to use as auxiliary information the longitudinal profile of the institution itself from the previous year, if available; and to use the information on all the responding institutions at time *t*, otherwise. First, Monte Carlo simulation studies were carried out on units with information from both BDAP and SIOPE to assess the impact of the imputation strategy on the estimates of the REPA variables at the aggregate level, i.e. by region, by title and by type of institution. Then, we analysed the longitudinal data of the institutions: the distribution of the items is subject to large variations from year to year, which invalidates the assumption underlying the longitudinal method of stability of the institutional profile. Furthermore, we evaluated the practical system management: the cross method is more responsive to any changes that occur on the respondents and is characterised by more simplicity for managing a control process. Finally, the goodness of fit of the chosen imputation strategy was confirmed by the analyses of subject matter experts of the final data.

The entire REPA production process is scheduled in this way: different provisional data are available during the year and for each reference period of the data. Frame PA is released once in a year for the final data delivery (May *t+2*). The process works in synergy with other Istat production structures, in particular those dealing with economic statistics and National Accounts.

4. Metadata

In the Section, we present the metadata activity that aims to describe how the concepts used in REPA have been structured.

4.1. REPA: metadata activity on units of analysis

The structured description of the metadata is one of the outputs of the REPA and provides information on which are the concepts that define it, framing and organizing them in a standardized way. The metadata activity therefore aims to apply a model for a structured description of concepts and it was carried out according to the principles of the standard "Generic Statistical Information Model (GSIM)" (HLG-MOS, 2020). On the basis of GSIM, it was possible to provide the REPA with the fundamental concepts for its documentation. Subsequently, we proceeded with the definition and description of the units type of the Register, the variables involved and the classifications correlated to the categorical variables.

Metadata activity, in addition to providing a set of organized and documented concepts, promotes the integration and the sharing between the REPA and the other ISSR Registers, especially with the base RS13. As introduced in Section 1, being REPA an "extended" Register, it has the specific purpose of extending the information of the base RS13, providing specific information that is not contained therein.

The elementary unit of analysis of the REPA is based on the more general concept of “economic unit”. This concept, defined in the framework of other Registers, describes an "entity that carries out economic activity of production, consumption or exchange". However, the REPA Register is just referred to "Institutional Units belonging to the institutional sector of Public Administrations" disseminated through the S13 list that is redefined every year and related to an annual reference period.

4.2. REPA: metadata activity on variables and classifications

Categorical variables are those variables that allow the reference population to be "partitioned" on the basis of specific characteristics. All those variables that have an *enumerated value domain* (according to GSIM concept), with defined modalities, belong to this type of variables. These variables are always associated with a classification that organizes their modalities in a structured way. In the case of REPA, the main categorical variables are: the Italian statistical classification of economic activities (Italian version of the Nace, *Nomenclature statistique des activités économiques dans la Communauté européenne*), the institutional typology and the legal form. These variables are inherited from the base RS13 . Some of them, together with the type of accounting of the institution and the economic operations

of the institution itself, have been used to determine the different sub-populations on which the entire REPA construction process have been built (see Section 2.2).

The variables that extend the REPA information set with respect to the base RS13 are those numerical (with *described value domain* according to GSIM), i.e. those relating to economic aggregates of income and expenditure provided, in the case of the sub-population of Territorial Governments (TG), from administrative sources BDAP and SIOPE. Up to now, the REPA prototype has been implemented only for the sub-population of TG which, as already highlighted, includes the institutional typologies of regions and autonomous provinces and local governments. The numerical variables of REPA, unlike the categorical ones, have specific sources depending on the reference sub-population and therefore the treatment relating to the non-responses is designed taking into account the source that is used for the imputation (Figure 1).

Table 1 – *The data structure components of final output FRAME PA. Territorial Governments sub-population. IU: Institutional unit.*

IU	NACE	Institut. typology	Type of Accountab.	Economic operability	CR	CE	PGS
1	…	Municip.	Financial	Current/ No current	Num.var.	Num.var.	Num.var.
2	…	Municip.	Financial	Current/ No current	Num.var.	Num.var.	Num.var.
3	…	Municip.	Financial	Current/ No current	Num.var.	Num.var.	Num.var.
…	…	…	…	…	…	…	…
i	…	Province	Financial	Current	Num.var.	Num.var.	Num.var.
…	…	…	…	…	…	…	…
j	…	Region	Financial	Current	Num.var.	Num.var.	Num.var.
…	…	…	…	…	Num.var.	Num.var.	Num.var.
n	…	…	…	…	Num.var.	Num.var.	Num.var.

Table 1 shows the data structure components of the final output FRAME PA, for the TG institutional units (IU). As introduced, the informative detail of the numerical variables within the FRAME PA output (derived from REPA) consists of three variables, one relating to income variables and two relating to expenditure variables: *Current revenues (CR)*, *Compensation of employees (CE)* and *Purchase of goods and services (PGS)*. These variables derive from the aggregation of available income and expenditures items, at a greater level of detail. The choice of these three final variables allows a comparability of the meanings of the economic variables between units classified on the basis of the variable "type of accountability" (financial or economic-patrimonial).

5. The new structural classification of a legal-functional and territorial type

The design and implementation of the REPA and Frame PA includes the definition of a new structural classification of the statistical units to support both the treatment processes and the analysis of the economic variables present in the extended register.

The need to define a new structural variable arose in the light of the descriptive limitations that distinguish the other structural stratification variables borrowed from the ISSR in the REPA, which are normally entrusted with the task of classifying all the units of the S13 list for any control activity, publication and interpretation of the related data. Those variables are: *institutional typology*, *Nace*, *territory* (municipality, province, region) of PAs. The first and most complex of these variables, the institutional typology variable, classifies the units of the S13 liston the basis of their legal, functional and territorial characteristics, but in an incomplete and incoherent manner that does not ensure that a consistent part of PAs are assigned to homogeneous sub-populations in terms of all the main characteristics just mentioned. On the other hand, Nace and the sub-variables on the location of units have the limitation of being based on a single classification criterion, namely functional or territorial, which alone is not capable of distinguishing PAs into sufficiently homogeneous sub-populations.

Therefore, another structural stratification variable was created with the aim of building an effective tool for describing the units, i.e. capable of distinguishing the largest number of PAs according to coherence and relevance, organising them into homogeneous subpopulations.

In terms of coherence, main general methodological criteria underlying the statistical classification activity have been applied for the new variable, with greater rigour than is expected for the structural variables institutional typology and Nace. These criteria are:

- completeness of the classes (few units in "other" class)
- mutually exclusive classes (same level of generality between classes)
- no underpopulated classes (< 10 units).

In terms of both relevance and coherence, it was envisaged that the new variable would classify the units on the basis of the same characteristics (mentioned above) of the individual structural variables of the institutional typology, Nace and localisation, synthesising and/or integrating them through unique and more significant dimensions. These dimensions are associated with the main characteristics of the units:

- for legal characteristics → functional autonomy level; governance model (association or institution);

- for functional characteristics → nature of activity (administrative functions, operating services, final services); sector of activity (health care, protected area management, business loans, etc.);
- for territorial characteristics → territorial level of activity (national, regional, local).

The resulting new variable produced classes of units that were completely homogeneous from a methodological and structural point of view in 92.5% of cases, compared with 37.5% of the classes belonging to the main structural stratification variable *institutional typology*.

This greater descriptive effectiveness of the new structural variable of a juridical-functional and territorial type manifests itself in a particular way with reference to the set of 894 units (8.6% of the total S13 population in 2019) whose primary source of information is the Istat statistical survey RIDDCUE (Collection of Information, Data and Documents necessary for the Classification of Economic Units in the institutional sectors established by the European System of Accounts 2010) (https://www.istat.it/it/archivio/219736), and which at the same time cannot be effectively distinguished by the structural variable of the institutional typology.

Unlike the other sources of information for the economic variables of the extended register (both from administrative archives and the statistical surveys), the source constituted by the RIDDCUE survey does not identify a specific and homogeneous sub-population of the base RS13. This is due to the specific purpose of this annual survey, which is to collect economic information on a heterogeneous set of units to determine their affiliation to one of the institutional sectors defined by ESA 2010, including sector S13. Of the units whose primary source is the RIDDCUE survey, the majority (894 units) cannot be broken down into homogeneous sub-populations from a legal and/or functional and/or territorial point of view using the institutional typology variable, because it reserves only generic, residual and underpopulated classes for this set of units. This heterogeneous set of institutions represents the second largest group of units of REPA after the sub-population of TG and the one with the highest total non-response rate.

Table 2 shows the main groups of the RS13 units according to their primary sources of information and their classifiability by institutional typology.

The importance of the new classification variable is manifold.

For a subset of PAs only the new legal-functional and territorial structural variable has the characteristics of isolating, autonomously or by crossing with other structural variables, homogeneous sub-populations. This characteristic is potentially important for the design of the imputation process of total non-response to economic data. In fact, the results of the simulations for TG have shown that imputation using a cross method based on the definition of homogeneous strata for imputation

guarantees greater stability of estimates over time and therefore less distortion of the data than a longitudinal method.

Table 2 – *Main groups of units by primary source of economic data and classifiability according to institutional typology (i.t.), year 2019.*

Groups	#	%	Primary source	Classifiability by i.t.
Territorial governments with specific source	8749	84,3	BDAP	Yes
PAs without specific source	894	8,6	RIDDCUE survey	No
Other PAs with specific source	405	3,9	General Government accounts Consolidated income statements of NHS Bodies Economic and patrimonial balance sheets of the Universities Final accounts of the Chamber of Commerce Collection of final balance sheets of Social Security Institutions (BICEP)	Yes
PAs without specific source	325	3,2	RIDDCUE survey	Yes

Furthermore, the introduction of the new structural variable in REPA and Frame PA makes it possible to deepen the description and analysis of the economic variables of these units, which would otherwise be associated with a composite and numerous subgroup of "other public administrations", thus compromising any data interpretation activity.

Finally, if the REPA and the Frame PA were to include in the future the population of public institutions of the Register Asia - Public Institutions, the new variable would ensure the same results in terms of data processing, dissemination and analysis also for other 2853 active institutional units from 2019 that at the moment are not considered in the RS13 framework. In fact, the Register Asia - Public Institutions, representing another object of the ISSR, includes more units than RS13. Those additional units would allow a better integration of all statistical registers on public units.. This macro-group of units would present similar problems concerning their breakdown into homogeneous sub-populations on the basis of the source of the economic data and of the available structural variables.

Despite the great benefits that this new classification can bring, more work is needed to validate this information.

6. Conclusion and future work

In this paper we have described some methodological aspects of the design and implementation of REPA production process, together with the structural metadata and the proposal of a new variable for the structural classification of statistical units.

Other work remains to complete the REPA, such as the design and implementation of the REPA for the other sub-populations, the completion of the metadata process and the further integration of the process with the other Istat production processes. Finally, a process for validating the quality of the register needs to be developed.

References

EUROSTAT 2019. Manual on Government Deficit and Debt IMPLEMENTATION OF ESA 2010 2019 edition, *Eurostat Manuals and guidelines, Economy and finance*. Publications Office of the European Union, available on-line: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-gq-19-007 (retrieved on 13/07/2023).

HLG-MOS 2020. The Generic Statistical Information model (GSIM v.1.2), available on line https://statswiki.unece.org/display/gsim (retrieved on 13/10/2023)

GUANDALINI A., PASSANTE D., VARRIALE R. 2022. The revenues of local governments in the statistical register for public administrations: inequality decomposition by sources, *RIEDS*, Vol. 76, No.3, pp. 17-28.

LUZI O., ALLEVA G., SCANNAPIECO M., FALORSI P.D. 2019. Building the Italian Integrated System of Statistical Registers: Methodological and Architectural Solutions. ESS Workshop on the use of administrative data and social statistics, Valencia,4-5June2019, available on-line:https://cros-legacy.ec.europa.eu/system/files/building-italia-integrated-system_istat_0.pdf (retrieved on 09/10/2023).

VARRIALE R., LORI M., MANTEGAZZA F. 2021. Stato di avanzamento dei lavori Focus: Frame PA enti territoriali. Nota tecnica Istat, Dicembre 2021.

Roberta VARRIALE, Sapienza University of Rome, roberta.varriale@uniroma1.it
Nevio ALBO, Istat, nalbo@istat.it
Cecilia CASAGRANDE, Istat, casagran@istat.it
Valeria OLIVIERI, Istat, vaolivie@istat.it