The framework for quality assessment in the Istat Integrated System of Registers: An application to the estimation of the Attained Level of Education[1]

Romina Filippini[2], Sara Giavante[2], Gaia Rocchetti[2]

[2] *Italian National Institute of Statistics (Istat), Italy*

Abstract

The potentiality of using new data sources has led National Statistical Institutes to reorganize their production system towards a more structured use of administrative sources, able to provide detailed information while reducing costs and response burden. Exploiting administrative and survey data, the Italian production system of statistics has moved towards a register-based statistics production system, built upon an 'Integrated System of Statistical Registers' (ISSR) composed of base registers and satellite registers.

In this context of modernization, to improve process efficiency and monitor the accuracy of results, the Italian National Institute of Statistics (Istat) introduced a shared framework for the quality assessment and documentation of the production process, able to capture the characteristics and specificities of the new paradigm of statistical production.

One of the main Italian base registers is the Base Register of Individuals (BRI), a comprehensive statistical register storing data gathered from different administrative sources. Core variables like place and date of birth, gender, citizenship are associated to each unit. Moreover, a variable denoting people usually resident in Italy is attached. This subset of data is the basis of the new Italian census that is as much as possible register based. According to this idea, the Attained Level of Education (ALE) is a variable for which a prediction in the register for resident population is obtained through a model based on the integrated use of administrative and survey information.

This document describes the application of the quality framework to the estimation process of ALE in BRI 2022. Specifically, the metadata model is applied for a structured description of the entire production process and a system of quality indicators is computed to monitor each process step during its implementation. The application highlights the importance of the framework from various perspectives. Besides providing crucial information about the quality of the process and the resulting output, the computation of relevant indicators allowed to monitor quality aspects during the production phase and consequently to immediately recognize suspicious data, facilitating the timely implementation of appropriate corrections. Additionally, the initial implementation of the framework's metadata model resulted in a revision of the sequence of some process steps, leading to an enhancement in overall efficiency. Finally, the structured description of the process through the metadata models included in the framework ensures an understanding of the process even by non-experts and guarantees the reproducibility of the process in subsequent years.

Keywords: quality framework, quality indicators, Integrated System of Statistical Registers, Attained Level of Education

[1] This paper resumes the outcomes of a work jointly carried out by the authors, however Section 1 is attributable to Sara Giavante, Section 2 and Appendix are attributable to Gaia Rocchetti, Section 3 and Section 5 are attributable to Romina Filippini, Section 4 is attributable to Romina Filippini and Gaia Rocchetti.

1. The quality framework in the Integrated System of Statistical Registers

The National Statistical Institutes have long started a reorganization of the processes aimed at implementing an Integrated System of Statistical Registers. The construction of the Italian Integrated System of Statistical Registers (ISSR), divided into base registers and satellite registers, depending on the information obtained and the nature of the statistical units identified, is based on the extensive use of administrative data combined with other types of data sources (Istat, 2016). The aim is to maximize the potential of administrative data and ensure efficient integration. This approach enhances data quality, reduces costs, and responds more flexibly to information needs, thereby enhancing the production of official statistics and providing more accurate, timely, and efficient data.

This paradigm shift has necessitated a reflection on the methods by which to evaluate the quality of a system that concentrates and integrates data from a plurality of sources. In this regard, at Istat, a framework has been developed capable of capturing the salient characteristics of this registers system (Istat, 2023). The framework presents itself as a model that provides a set of quality indicators for monitoring and evaluating the products and production processes of the ISSR registers. For its construction, reference was made to models developed within the Unece framework, namely the Generic Statistical Business Process Model (GSBPM, Unece, 2019) and the Generic Statistical Information Model (GSIM, Unece, 2020), with the purpose of analysing, mapping, and documenting the processes of the ISSR.

The framework allows for systematizing the appropriate structural and referential metadata and quality indicators useful for monitoring and evaluating each step that characterizes the registry construction process. The set of identified metadata and quality indicators has been organized into a template consisting of documentation models, one for each sub-process, which guide the mapping of the process into its sub-processes and its description in terms of inputs, outputs, statistical methods, software used, etc.

The framework comprises a general identification metadata template (General characteristics of a statistical register of the ISSR), which provides information about the Statistical Register of the ISSR and its sources, and includes a series of documentation templates for process and product description, containing metadata elements crucial for standard documentation, calculation, and interpretation of the indicators being described. For each documentation model, the template groups the informational objects into three macro-elements: Input, Subprocess, and Output.

In addition to the general identification template, the framework includes nine sub-process documentation models: (1) Check data availability, (2) Acquire data, (3) Conduct preliminary

evaluation, (4) Integrate data, (5) Classify and code, (6) Edit and Impute, (7) Derive new variables and units, (8) Calculate aggregates and (9) Validate outputs.

In this document, the quality framework is applied to the estimation process of the Attained Level of Education (ALE) in the Base Register of Individuals (BRI) for the year 2022. BRI is one of the Base Statistical Registers produced by Istat and it is the reference statistical register for Istat official statistical production concerning the population. BRI is built making an extensive use of administrative data, starting from demographic information and then making corrections based on signals derived both from administrative sources and social surveys.

After providing a general description of the context of ALE estimation carried out in Istat (section 2), the application of the framework to the ALE estimation process in 2022 is detailed. Specifically, section 3 illustrates the mapping of the process through the documentation template, while section 4 describes the computation of quality indicators. Finally, section 5 presents some concluding remarks.

2. The context of the Attained Level of Education

Since October 2018, the Permanent Census of Population and Housing has replaced the decade-long practice of conducting general census. The conventional method, based on a comprehensive field-based enumeration of individuals, families, and households, has been replaced by a strategy that heavily relies on integrating information stored in registers with data specifically collected through sample surveys. In this context, the BRI forms the foundation of the permanent census. The register variables, derived from administrative sources at the individual level, encompass essential demographic information such as gender, age, marital status, place and date of birth, and citizenship. However, to include other core variables, such as ALE, traditionally collected in the census but not completely covered by administrative sources, a sample survey referred to as Master Sample (MS) is conducted.

The MS allows the supplementation of register data in terms of coverage, quality and thematic information, including details related to education. Additionally, Istat exploits administrative information collected by the Ministry of Education, Universities, and Research (MIUR), which provides insights into students' ALE and course attendance. Istat processes and integrates MIUR data, leading to the creation of a database on Education and Qualification from administrative sources, referred to as BIT (“Base Integrata Titoli di studio”).

However, there are informational gaps in BIT data. Firstly, they pertain only to individuals entering a study program after 2011. Therefore, the 2011 Census is relied upon to fill this gap. The 2011 Census operations, whose reference date was October 2011, surveyed more than 59 million individuals, collecting data on educational attainment for persons aged 9 or older.

Furthermore, BIT data presents under-coverage issues since they only trace students enrolled in educational courses held in Italy, excluding qualification courses like Fine Arts, Drama, Dance and Music academic diplomas, as well as training and vocational careers managed by Italian Regions, which are not required to provide data to MIUR. The main consequence is a potential underestimation of ALE in the administrative source. Another critical issue concerning BIT data has to do with timeliness, since they are typically available with a delay of 1 or 2 years compared to the BRI reference time.

In this context, to produce ALE estimates for all the Italian resident population at reference time, Istat adopted a mass imputation approach based on the integration of different type of data (Di Zio et al., 2019). Specifically, to predict ALE referred to year 2022, Istat makes use of the following sources:

- BRI containing demographic information on the resident population in 2022 (BRI.22);
- Administrative data on ALE and school course attended in 2021 (BIT.21);
- 2011 Population Census (CENS11);
- Survey sampling data referred to October 2022 (MS.22);
- Auxiliary administrative information resulting from registration and cancellation forms for residence transfers from 2012 to 2021 (APR4.21).

3. Process documentation through the framework template

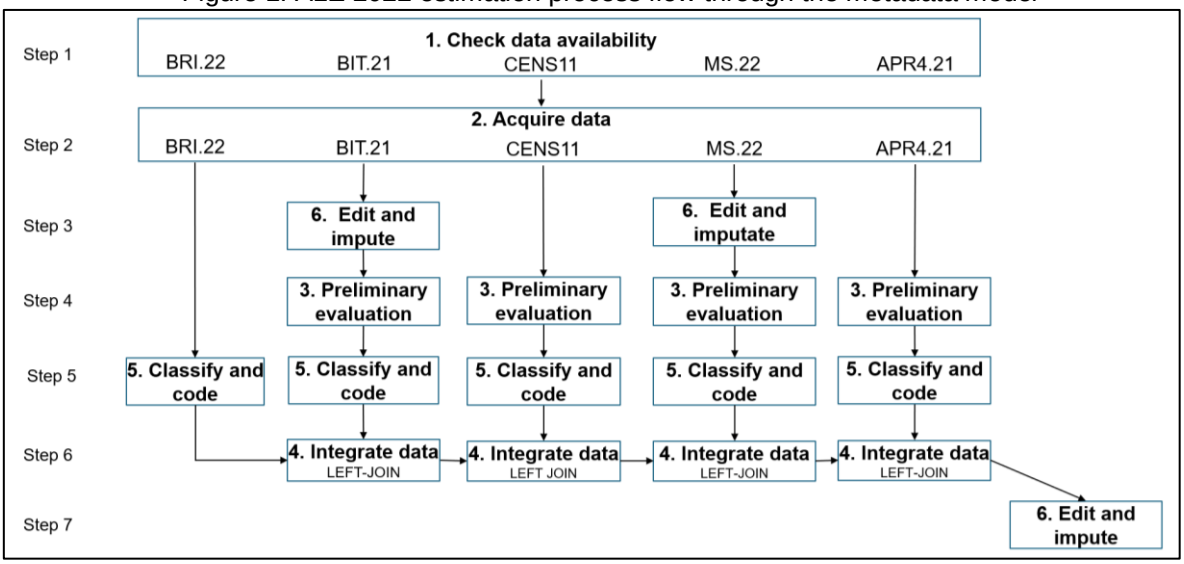
The process of ALE estimation starts from the construction of a dataset containing all pertinent information on the subject. Gathering the raw data and treating each primary source are essential prerequisites for this task. This involves identifying and correcting errors and inconsistencies to ensure the accuracy and reliability of each dataset, as well as imputing missing values, standardizing variables, and removing duplicate entries. The overall data cleaning process may be conducted on each individual data source, as well as on the integrated dataset, which includes the creation of new variables or aggregates that may enhance the predictive power of the model.

The structured depiction of the entire production process using the framework template significantly contributed to enhancing the process efficiency. Representing the flow of the process through its relevant steps allowed to highlight some critical points and prompted adjustments to the original procedure.

Figure 1 illustrates the updated process flow for ALE estimation in BRI.22, following recent adjustments. This process develops in 7 sequential steps and involves 6 distinct sub-processes, each applying specific treatments to the data, as described by the corresponding documentation template. Although the process is rather complex, providing a detailed

description of each sub-process (represented by a “node” in the process flow) through standardized templates, as outlined by the framework, improves the overall comprehensibility of the process. This helps making it accessible even to those who may not be directly involved in the ALE estimation process.

Figure 1: ALE 2022 estimation process flow through the metadata model



In the ALE estimation context, most actions are dedicated to constructing the dataset, which contains all the information significantly related to ALE. The initial two steps involve checks on the availability and collection of administrative and survey data required for the estimation. The corresponding documentation models are: (1) Check data availability and (2) Acquire data, each covering all five data sources exploited: BRI.22, BIT.21, CENS11, MS.22 and APR4.21 (see section 2). Steps 3, 4, and 5 focus on the pre-treatment of each individual data source, requiring actions related to (6) Edit and impute, (3) Conduct preliminary evaluation and (5) Classify and code. Step 6 involves the integration of information from the different sources on the reference population, (4) Integrate data. Lastly, in the final step, where the ultimate ALE estimation is carried out, editing and imputation is once again implemented.

For each “node” of the flow, the corresponding template must be filled-in with relevant information, describing the input data, the process or treatment required and the output data. Each output serves as the input for the subsequent step, following the process flow represented in figure 1.

As an example, Table 1 shows the model for the "Edit and impute" sub-process properly filled in for each step and each dataset where it appears in the process flow.

Table 1: Model for the “Edit and impute” sub-process: step 3 (BIT.21 and MS.22) and step 7 (Integrated dataset)

Macro item	GSIM Object	Values		
		Step3: BIT.21	**Step 3: MS.22**	**Step 7: Integrated dataset**
Input	Transformable input	BIT.21 (Raw data)	MS.22 (Raw data)	BRI.22_integrated
	Parameters	Variable to be checked: ALE (check for not admissible combinations of modes)	Variable to be checked: ALE (check for not admissible combinations of modes)	Variable to be checked: ALE_INT (from step 6) - Covariates: demographic inf., school attendance - Auxiliary variable: flag "Inconsistent" (from step 3: BIT)
	Process support input	Validity and consistency edit rules for ALE and school attendance relationship	Validity and consistency edit rules for: (1) ALE and age, (2) ALE and school attendance relationship	- Consistency edit rules for ALE and age relationship - Conditions to identify "No-Change" individuals (0 prob. of achieving a new ed. level) - Estimation models for the Change subgroup of individuals
Process	Process function	Impute missing and check records against edits	Impute missing and correct erroneous ALE entries	Estimate ALE for all individuals with age≥9
	Process step	Edit and impute (5.4 in GSBPM)	Edit and impute (5.4 in GSBPM)	Edit and impute (5.4 in GSBPM)
	Process method	Deductive imputation	Deductive imputation	- Deductive imputation for "No-Change" and MS.22 individuals. - Model based imputation (Log-linear models) for others
	Rule	- If ALE=missing and school course =“Primary“ then ALE=2 "No ed." - If ALE is inconsistent with school course then "Inconsistent"=true	If ALE is missing or inconsistent, then impute ALE using administrative information. (Imputation rules available).	- Models are estimated within each Italian region - For individuals in BIT.21, where “Inconsistent"=false, school attendance is a covariate; otherwise not - For individuals not in BIT.21, ALE observed in MS.22 is the response variable
	Software	Oracle	Oracle	Sas
Output	Trasformed output	BIT.21 (corrected)	MS.22 (corrected)	BRI.22_integrated with complete ALE referred to 2022
	Quality indicators	..See section 4..	..See section 4..	..See section 4..

4. Quality indicators to monitor the process flow

Each documentation model outlining the production process incorporates the computation of quality indicators, underlying the importance of assessing data quality at each stage of the process. By incorporating quality assessment into each step, potential issues can be identified and addressed in a timely manner, enhancing the overall reliability and validity of the final estimate. Computation of quality indicators after the completion of each action specified in the metadata model, prior to proceeding with the subsequent step, emphasizes a proactive

approach to quality management and allows the identification and rectification of potential quality issues before they can propagate further in the process.

The standard quality framework offers a set of recommended indicators for assessing quality, corresponding to each sub-process. However, not all indicators may be relevant or applicable in every situation. It's crucial to consider the unique characteristics and requirements of each case when choosing which indicators to use. Moreover, some indicators presented in the standard version may not directly apply to the specific case. Therefore, they may either be declined or reinterpreted to ensure a more focused and targeted quality assessment process. This involves reviewing or adapting indicators to better suit the specific context, often necessitating modifications to their definitions to enhance relevance or utility in the given situation.

As an example, Table 2 shows how the quality indicators proposed for the “Edit and impute" sub-process have been interpreted and used in the context of ALE estimation.

Table 2: Quality indicators in the “Edit and impute” model: step 3 (BIT.21 and MS.22) and step 7 (Integrated dataset)

Indicators proposed in the framework template	**Indicators interpreted for the specific case of ALE estimation process**		
	Step3: BIT.21	**Step 3: MS.22**	**Step 7: Integrated dataset**
6.1. Records with at least one missing	6.1.BIT. Records with missing ALE	6.1.MS. Records with at least one missing on variables related to ALE	6.1.Int. Records with missing ALE (Not in MS)
6.2. Records failing at least one edit	6.2.BIT. Records with conflicting ALE and school course	6.2.MS.a. Records with conflicting ALE and age. 6.2.MS.b. Records with conflicting ALE and school course	6.2.Int.a. Records with conflicting ALE and age. 6.2.Int.b. Records with ALE from MS<ALE from BIT
6.3. Variables failing at least one edit	Not relevant	Not relevant	Not relevant
6.4. Item non-response rate per variable	Not relevant	6.4.MS.a. non response in ALE 6.4.MS.b. non response in other variables related to ALE	Not relevant
6.5. Imput. rate per var.	Not relevant	Not relevant	Not relevant
6.6. Modification rate per variable	Not relevant	6.6.MS Modification rate for ALE	6.6.Int. Records with estimated ALE different from ALE in BIT (referred to *t-1*)
6.7. Net imputation rate per variable	Not relevant	Not relevant	Not relevant
6.8. Cancellation rate	Not relevant	Not relevant	Not relevant
6.9. Weighted imp. rate (due to imputed values)	Not applicable	Not applicable	Non applicabile
6.10 Final item non-resp.rate per variable	Not relevant	Not relevant	Not relevant
6.11. Distance between variable distributions compared to other editions or sources	Not relevant	6.11.MS. Mean difference between pre- and post-correction ALE distributions	6.11.Int. Mean difference between estimated ALE distribution and weighted ALE distribution from MS

At the end of the ALE estimation process, 36 quality indicators are computed, pertaining to 4 different documentation models, carried out through 7 process steps (see Appendix). The indicators are computed using a uniform structure, presented as ratios between two quantities (Numerator/Denominator). This allows for the representation of the entire indicator system within a single table, facilitating comparisons. The table displaying quality indicators includes details such as the step and the corresponding sub-process where the indicator is computed, the dataset it applies to (which could be each individual source or the integrated dataset), a unique identification number, and the description of the indicator. Additionally, alongside the final resultant value, the table presents relevant numerical values - such as the numerator and denominator - from which the indicator derives. This setup offers a clear insight into the underlying data contributing to each indicator.

In subsequent years, the same structured table can be generated, allowing for the addition of columns referencing each year. This facilitates straightforward comparison of indicators over time. The complete table of indicators computed for the ALE 2022 estimation process is reported in the Appendix (Table A1).

5. Conclusions

The application of the framework for quality assessment in the ALE estimation context highlights the importance of integrating quality assessment into every phase of the variable production process. This approach not only improves data reliability and validity but also encourages a proactive approach to quality management, resulting in enhanced overall process efficiency and higher quality outcomes.

The availability of a standard template for documenting the overall process enables efficient organization of information, facilitates collaboration among team members, and enhances transparency in project management. On the other hand, the proposed framework promotes flexibility and adaptability in the quality assessment process, allowing expertise in selecting indicators that best align with the goals of the project, emphasizing the need for a thoughtful and context-sensitive approach to quality assessment.

The system of indicators describing a process can indeed be quite complex. Establishing a standardized and consistent method for their representation is essential for facilitating meaningful analysis and interpretation of the results. So, it becomes much easier to compare indicators across different datasets, time periods, or contexts. Moreover, monitoring these quality indicators over time provides valuable insights into the performance of the overall process flow, allowing to identify areas for improvement and implement necessary changes.

This approach leads to more efficient and effective processes, contributing to the continuous improvement of the system.

References

Di Zio, M., Filippini R., Rocchetti G. (2019). An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data. Rivista di Statistica Ufficiale, N. 2-3/2019.

Istat. (2016). Il Programma di Modernizzazione dell'Istat. https://www.istat.it/it/files/2010/12/Programma_modernizzazione_Istat2016.pdf.

Istat (2023). Monitoring and Evaluating the Quality of the Integrated System of Registers, Istat Working Papers N. 7/2023.

Unece (2019). Generic Statistical Business Process Model (GSBPM), Version 5.1, January 2019, United Nations Economic Commission for Europe. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1.

Unece (2020). Generic Statistical Information Model (GSIM): Communication Paper, Version 1.2, October 2020. https://statswiki.unece.org/display/gsim/GSIM+v1.2+Communication+Paper.

Appendix

Table A1: Quality indicators of the ALE estimation process in BRI 2022

Step: model	Dataset	Quality indicator		Num.	Denom.	Value
		ID	Description			
2: DatAcq.	BRI.22	2.1.BRI.a	Records uploaded compared to the previous year: Total	103,261,785	101,049,167	102.2%
2: DatAcq.	BRI.22	2.1.BRI.b	Records uploaded compared to the previous year : Resident	59,633,293	59,730,422	99,8%
2: DatAcq.	BRI.22	2.1.BRI.c	Records uploaded compared to the previous year: Resident with Age>=9	55,594,279	55,594,975	100.0%
2: DatAcq.	BIT.21	2.1.BIT.a	Records uploaded compared to the previous year: dataset on student enrollment	16,696,994	16,696,215	100.0%
2: DatAcq.	BIT.21	2.1.BIT.b	Records uploaded compared to the previous year: dataset on ALE	13,960,933	13,326,885	104.8%
2: DatAcq	MS.22	2.1.MS.a	Records uploaded compared to the previous year	2,562.600	4,816,285	53.2%
2: DatAcq.	MS.22	2.1.MS.b	Records uploaded with valid unit identification code compared to the previous year	2,546,758	4,803,240	53.0%
3: E&I	BIT.21	6.1.BIT	Records with missing ALE	2,550,362	16,511,294	15.4%
3: E&I	BIT.21	6.2.BIT	Records with conflicting ALE and school course	4,448	16,511,294	0.03%
3: E&I	MS.22	6.1.MS	Records with age MS≥9 and at least one missing on variables related to ALE	460,854	2,390,539	19.3%
3: E&I	MS.22	6.2.MS.a	Records with age MS≥9 and conflicting ALE and Age	229	2,390,539	0.0%
3: E&I	MS.22	6.2.MS.b	Records with age MS≥9 and conflicting ALE and school course	7,505	2,390,539	0.3%
3: E&I	MS.22	6.4.MS.a	Non response rate in ALE	111	2,390,539	0.0%
3: E&I	MS.22	6.4.MS.b	Non response rate in other variables related to ALE	460,744	2,390,539	19.3%
3: E&I	MS.22	6.6.MS	Modification rate for ALE	4,341	2,390,539	0,2%
3: E&I	MS.22	6.11.MS	Mean difference between pre- and post-correction ALE distributions	-	-	0.0%
4: Checks	CENS11	3.2.CEN	Records with duplicated unit identification code	72,509	59,433,363	0.1%
4: Checks	CENS11	3.3.CEN	Records with duplicated unit identification code and different ALE	28,054	59,433,363	0.05%
4: Checks	MS.22	3.2.MS	Records with duplicated unit identification code	4,208	2,395,568	0.2%
4: Checks	MS.22	3.3.MS	Records with duplicated unit identification code and different ALE	0	2,395,568	0.0%
6: Int.	CENS11	4.1.CEN	Missing or errors in linkage variable	381	54,962,619	0.0%

6: Int.	CENS11	4.2.CEN	Match rate	50.190.419	59,331,974	84.6%
6: Int.	CENS11	4.5.CEN	Hierarchical coverage: Records in BRI and CENS11 (not in MS/BIT)	35,925,845	54,962,619	64.6%
6: Int.	BIT.21	4.2.BIT	Match rate	14,724,269	16,696,742	88.2%
6: Int.	BIT.21	4.5.BIT	Hierarchical coverage: Records in BRI and BIT (not in MS)	14,097,350	54,962,619	25.7%
6: Int.	MS.22	4.1.MS	Missing or errors in linkage variable	15,764	54,962,619	0.03%
6: Int.	MS.22	4.2.MS	Match rate	2,381,189	2,544,241	93.6%
6: Int.	MS.22	4.5.MS	Hierarchical coverage: Records in BRI and MS	2,381,189	54,962,619	4.3%
6: Int.	APR4.21	4.2.APR4	Match rate	8,470,165	10,971,051	77.2%
6: Int.	APR4.21	4.5.APR4	Hierarchical coverage: Records in BRI and APR4 (not in MS/BIT/CENS11)	1,248,116	54,962,619	2.2%
7: E&I	BRI.22_int	6.1.Int	Records with missing ALE refferred to 2022 (Not in MS)	53,217,113	55,594,451	95.7%
7: E&I	BRI.22_int	6.2.Int.a	Records with conflicting ALE and Age	223	55,594,451	0.0%
7: E&I	BRI.22_int	6.2.Int.b	Records with ALE from MS<ALE from BIT.21	13,350	626,549	2.1%
7: E&I	BRI.22_int	6.6.Int.a	Records with estimated ALE>ALE in BIT (referred to t-1)	2,935,283	14,678,273	20.0%
7: E&I	BRI.22_int	6.6.Int.b	Records with estimated ALE 2022<ALE in BIT (referred to t-1)	24,633	14,678,273	0.2%
7: E&I	BRI.22_int	6.11.Int	Mean difference between estimated ALE distribution and weighted ALE distribution from MS	-	-	0.1%