UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
Neuchâtel, 18 September 2018)

Selective editing in the Integrated Business Statistics System (SINTESI)

Luzi O., Manzari A., Pichiorri T., **Rocci F.**, Rosati S., Varriale R.,

Istat

SINTESI - Integrated Business Statistics System

A new infrastructure to support the whole statistical production system in the business area

Main purposes:

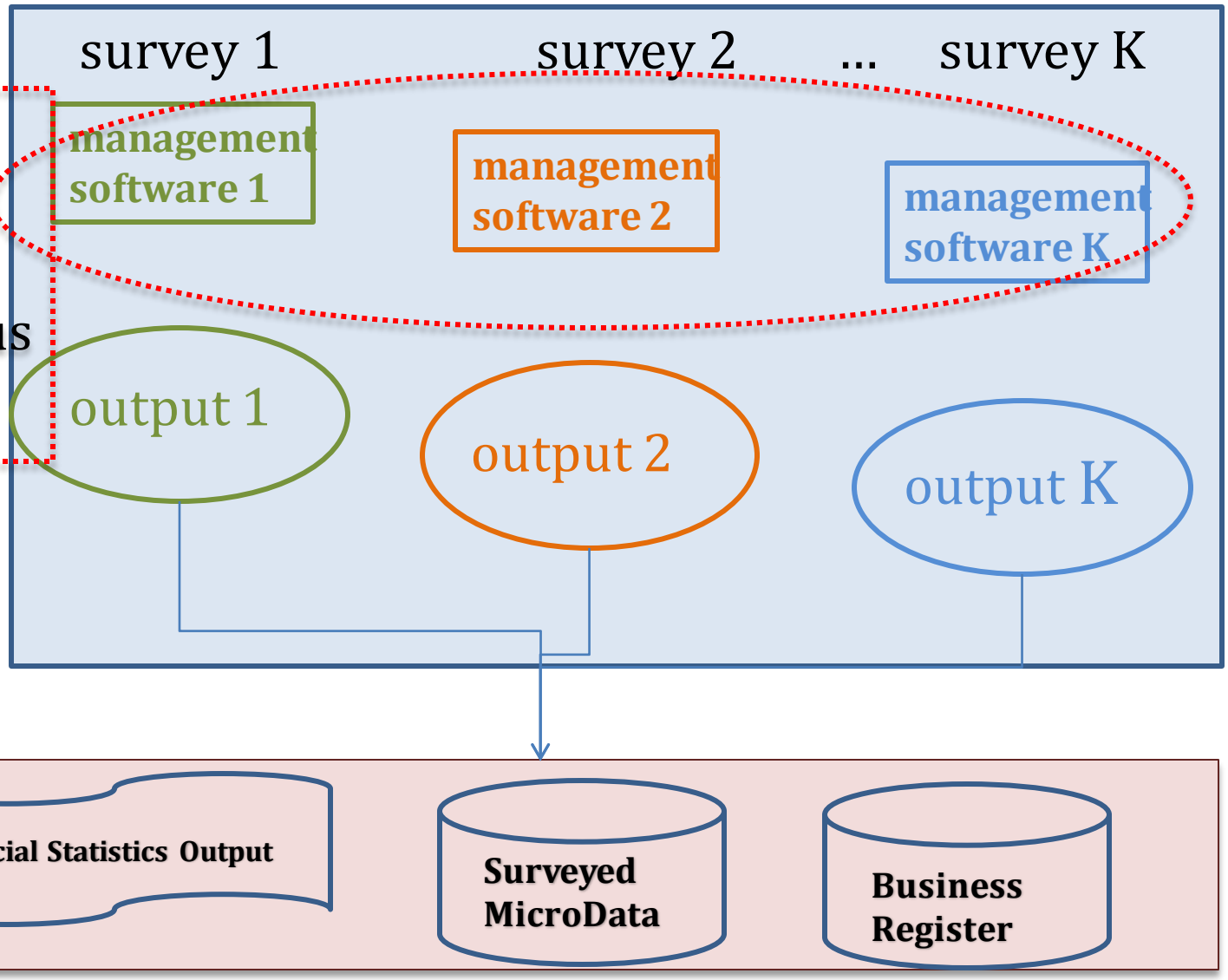
a. to simplify and to standardize the production processes

b. the elimination of redundancies and the improvement of the overall coherence of the statistical processes and statistics

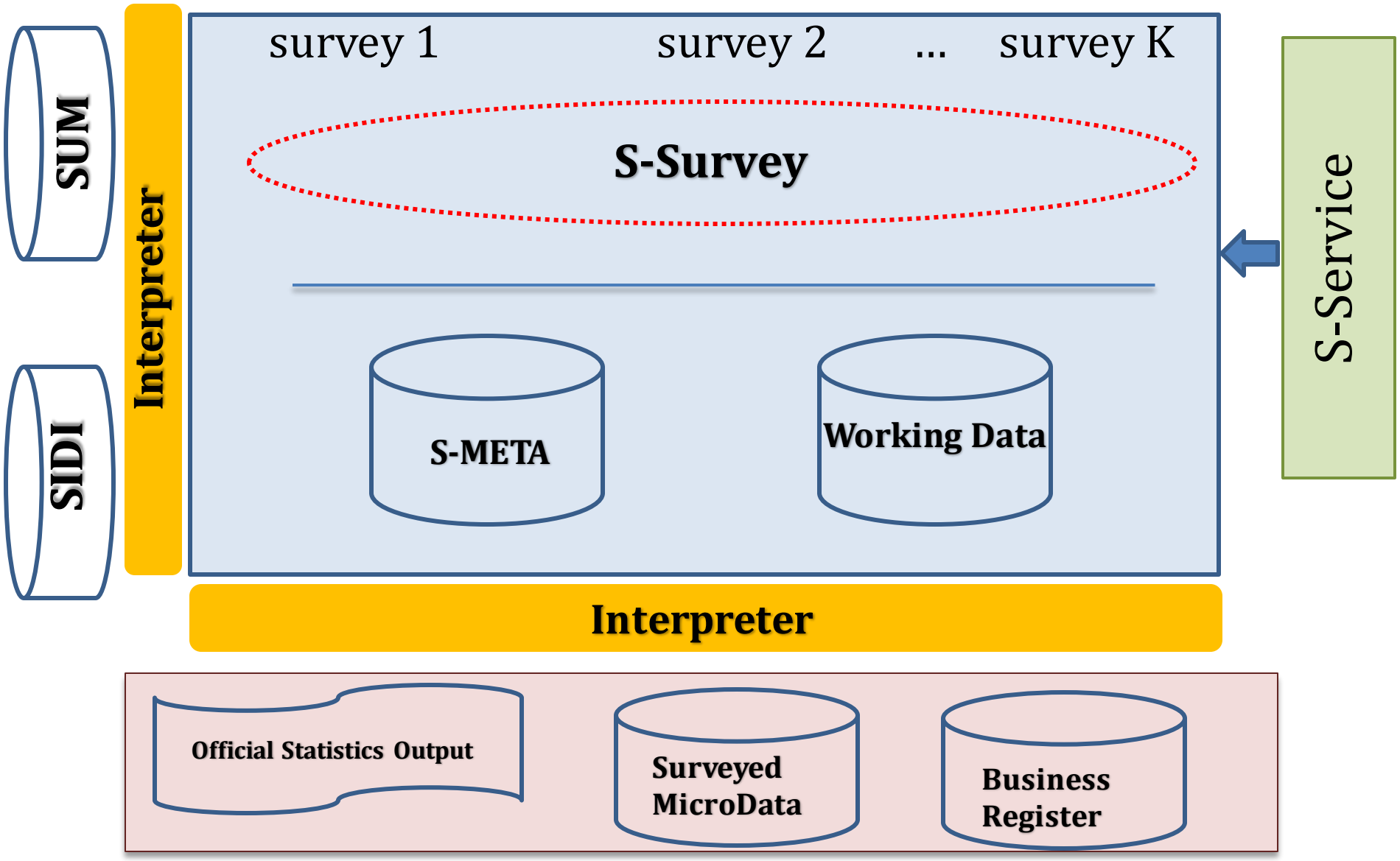
c. to reduce the costs associated with operational aspects of surveys, to increase efficiency and improve timeliness

Current situation…As-Is model

- ✓ High costs managment systems
- ✓ Heterogenous processes



...To-Be model


SUM
SIDI
Interpreter
survey 1
survey 2
...
survey K
S-Survey
S-META
Working Data
S-Service
Interpreter
Official Statistics Output
Surveyed MicroData
Business Register

…To-Be model – SINTESI -

Key Elements

- **Data Repository:** the logical environment of the generalized information sources - working data, business Register, auxiliary sources of information, raw survey data, historical survey data, etc.;
- **S-META:** a Metadata Repository containing all the metadata on survey units, variables, and statistical processes
- **S-SURVEY**: web application to manage the statistical elaboration for every phase of the production process
- **S-SERVICE**: catalog of Statistical Services implementing specific methodologies, properly defined for every phase of the S-Survey
- **INTERPRETER**: software component to handle the communication among different system components

As methodological Team

…To-Be model - E&I process in SINTESI -

- The first services to be implemented in 2018 relate to the data E&I process

- The Generic Statistical Data Editing Model (GSDEM) has been adopted as reference framework for designing an E&I process in SINTESI

- In GSDEM the E&I process is interpreted as a set of standardized, consistent statistical *functions* associated to some 'process steps':

  - *Editing systematic errors*
  - *Selective editing*
  - *Interactive editing*
  - *Automatic editing*
  - *Macro editing*

Selective editing

- Goal: methodological approach to identify influential errors in continuous data

  partitioning data into two sets

  - Critical set containing potentailly influential erros – Interactive Editing
  - Non critical set containing records supposed to be correct or with small erros – automatic Editing

- Key elements: units in the critical set are identified on the basis of a score function, which components are:

  - risk ~ probability of error occurence
  - influence ~ (expected) impact on estimates

SeleMix

- R software -

In SINTESI the **R** software **SeleMix** implementing a specific model-based selective editing approach has been chosen *Guarnera U., Buglielli T. (2013)*

Approach based on an explicit modeling *error mechanism* that contaminates the *true data via mixture model Di Zio M., Guranera U. (2013)*

- True data are thought of as *n* realizations from a random p-vector **Y** that, conditional on a set of *q* covariates **X**
- The intermittent nature of the error is modelled through a Bernoullian random variable
- Definition of the **score function** in terms of conditional distribution of the 'true' data given the observed data

The implementation strategy of selective editing in SINTESI - *via SeleMix* (1)

- Three steps to perform the selective editing should be explicitly defined by three function gathered by SeleMix into SINTESI:

  1. estimate model parameters (*ml.est function in Selemix*)
  2. predict the "true" values for the target variables (*ml.est/pred.y functions in Selemix*)
  3. identify influential errors (*sel.edit function in Selemix*)

- In step 1, the strategy to be implemented has to take into account the following issues:
  - Choose the target variable Y_1 or variables Y_1 and Y_2
  - identify any available auxiliary variables, e.g. X_1 and X_2, that are highly correlated with the target variables
  - choose the model: univariate (e.g. $Y_1 \mid X_1, X_2$ and $Y_2 \mid X_1, X_2$) or multivariate (e.g. $Y_1, Y_2 \mid X_1, X_2$)
  - define the estimation domains

The implementation strategy of selective editing in SINTESI - *via SeleMix* (2)

- Usually Step 2 is performed immediately after step1.

  Nevertheless, every process can set a different workflow from step 1 to step 2, depending on the survey process and its features. A preliminary data analysis is expected to give information about how to set the process for the specific survey under study.

- In step 3, the workflow of the process is has to take into account:

  - domain for selective editing application;
  - error threshold;
  - whether or not sampling weights are used;
  - whether or not known population totals are used.

How we are proceeding… (1)

- Short-term business statistics on business (STS) have been selected as starting application area of the new system:
  - they have relatively simple and sufficiently homogeneous methodological characteristics but intensive follow-up and interactive editing
  - their current information systems would gain in efficiency introducing product innovations of methodology and software generalization

- the *generic Statistical Data Editing (SDE) flow model for STS* proposed in GSDEM is taken as reference in the context of SINTESI.

- In the *generic SDE flow model for STS* influential errors need to be timely identified while optimizing the trade-off between costs (associated to the interactive review of data) and accuracy of output statistics

How we are proceeding… (2)

- The main steps:
  - To provide the SDE workflow of the current production process for each pilots survey
  - to extract their operating structure
  - to design a generalized *SDE flow model* for STS build the statistical services fine-tune the mode of application of each service to each specific STS

Survey on Turnover and Orders in the Industry

- general description -

- The *Survey on Turnover and Orders in the Industry* covers the mining and manufacturing economic activities (NACE Rev. 2 sections B and C)

- Two indices are released by means of the monthly survey :
  - the turnover index measures the trend over time of the sales of industrial companies
  - index about the orders

- The survey is based on a panel of enterprises, selected at the base year from the Italian Businesses Register (ASIA).
  This results in about 7.000 surveyed enterprises, whose data are elaborated according to their Kind of Activity Unit (KAU)

- Statistical results are released monthly, with only 30 days of delay with respect to the end of the reference month.

- The survey process runs continuously during the period of every month.

Survey on Turnover and Orders in the Industry
- E&I process -

- Few deterministic edit rules are run, based on a comparison with the longitudinal profile of the enterprise itself

- The main rule calculates the variation over the same month of the previous year, for which records with ±30% variation are considered to be anomalous.

- The interactive controls are based on an intense work of collecting information, both directly from the respondents both from other auxiliary information delivered by other survey approximately about the same period.

Survey on Turnover and Orders in the Industry
- The simulation -

- To test the performance of SeleMix approach, a first experiment has been implemented, based on data of the years 2017 and 2016

- A simulation is designed to run over every month of the year 2017, trying to model the same longitudinal relationship as the current process is based on

- To define a proper benchmark:
  - for each month we have:
    i. raw data : Y_{Rt}
    ii. validated data: Y_{Vt}
  - Where $Y_{Rt} \neq Y_{Vt}$ data are flagged as being erroneous for any reasons (flagged), for which further analysis should be done to understand what caused them
  - Selective editing model is estimated for the set of data originally corrected plus those records that were flagged

Survey on Turnover and Orders in the Industry
- The simulation -

- The final aim would be to assess which data the model identify as influential errors. The results are compared to the flagged observations

- The target variable is the monthly raw data on turnover to be related to the turnover of the same month of the previous year, as the longitudinal information represents commonly an auxiliary information for the STS editing rules

$$Y_{Rt} \sim Y_{Vt-12}$$

Table 1. Results of SeleMix for the Survey on Turnover and Orders in Industry: surveyed units and influential data per month – Year 2017

Month	Total n. obs	flagged (a)	Outlier and Influential (b)			Flagged and Outlier and Influential (a∩b)		
			tsel 0.003	tsel 0.004	tsel 0.005	tsel 0.003	tsel 0.004	tsel 0.005
1	6713	107	363	321	137	20	16	13
2	6729	84	398	252	214	21	16	15
3	6758	72	179	128	94	13	7	7
4	6740	67	394	328	260	10	7	6
5	6743	76	308	140	133	13	11	11
6	6748	82	300	214	181	15	12	10
7	6743	89	365	313	215	16	16	12
8	6623	100	604	494	357	19	19	18
9	6715	102	299	192	183	9	5	5
10	6685	111	316	173	144	11	6	6
11	6678	141	317	244	193	12	9	9
12	6655	186	484	268	144	25	16	12

Survey on Turnover and Orders in the Industry
- The simulation -
- First results-

- Additional analyses are needed to evaluate performance of the proposed selective editing method in the specific context to better define a strategy to release a generalized service.

- As an example, the following methodological issues need to be addressed:

- the impact of the influential errors on the final indices needs to be quantified;
- a comparison between the set of flagged data and the influential ones can suggest which errors mechanism the model identifies;
- any possible seasonal effects need to be assessed and adjusted.

Conclusion and next steps

- The AS-IS description of the STS statistical processes has still to be finalized, because some issues result difficult to be standardized.

- The first mapping of the data processing flow of the Survey on Turnover and Orders in the Industry revealed some key aspects in order to understand how to proceed.

- At present, the “transformation” of SeleMix algorithms for the Selective Editing in a Statistical Service SINTESI is an ongoing activity.

- Selective editing can allow to save time and costs for interactive editing, since only units that result to be both potentially erroneous and influential are revised.

- The first experiments and results on monthly data of the Survey on Turnover and Orders in the Industry should allow to understand which kind of errors are detected on the basis of the current procedures in respect of SeleMix approach.

Thank you