Evaluating, monitoring and documenting the effects of editing and imputation in ISTAT surveys

Orietta Luzi

ISTAT - Italian Statistical Office

Editing and Imputation in Official Statistics

Editing and Imputation (E&I) consists of an integrated set of actions aiming at

- obtaining complete and coherent data with respect to the specific survey quality needs
- providing information on collected data quality and error sources

Evaluating E&I in Official Statistics

General problem

Whatever action we perform on statistical survey data to make them acceptable with respect to the survey specific quality needs,we are conscious that:

- statistical properties of originally collected data are modified
- parameter's estimates are affected by non sampling errors, editing and imputation mechanisms (either random or sistematic)

Evaluating E&I in Official Statistics

Different evaluation needs

1.Before E&I: Evaluating the quality of E&I

Verifying the capability of editing/imputation methods of correctly identifying errors/recovering true data (e.g. for selecting the "best" approach to a survey/data problem)

2.During E&I: Evaluating the effects on of E&I

- –Measuring the modifications on both original distributions and relations due to E&I for tuning purposes
- –Assessing the effects on final estimates for estimation purposes

3.Documenting and monitoring E&I

- –Documenting the main characteristics and the overall effects of E&I processes for comparative evaluations over time or across similar surveys

Evaluating E&I processes: main past experiences at Istat

- Evaluating the quality of E&I methods: *the EUREDIT Project*
- Documenting and monitoring E&I processes: *the SIDI system*

Evaluating the quality of E&I methods
The Euredit Project

The **Euredit Project** (*EU Fifth Framework Research Program*) established a general framework for the comparative evaluation of E&I in terms of:

- **Experimental** approach
- **Evaluation** approach
  - *Evaluation criteria*
  - *Evaluation measures*

The EUREDIT Project

Experimental approach: simulation

1. A set of "true data" is artificially contaminated by using pre-defined (either MAR or MCAR) error mechanisms
2. Competitive E&I methods are evaluated by comparing true and edited/imputed data

Evaluation criteria

- Preservation of elementary data
- Preservation of distributions
- Preservation of aggregates
- Preservation of relations

Documenting E&I processes

The Information System for Survey Documentation (SIDI)

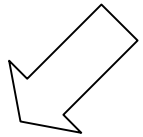
- SIDI is an information system devoted to support the survey managers in the following activities:
  - to monitor the production process
  - to analyse production processes over time
  - to evaluate effects of changes in the production process
- SIDI has a high degree of standardisation of both metadata and quantitative indicators
  - to allow the users to compare different surveys
  - to select surveys on the basis of several selection criteria

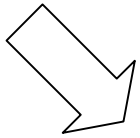
SIDI quality indicators

A set of **standard indicators** has been defined for each phase of the survey process

- same indicators regardless the survey typology (direct, administrative, mixed)
- standard formulae have been defined for each indicator

SIDI quality indicators





Metadata:

1. The survey information content such as statistical units and observed phenomena
2. The planning of the survey
3. The survey operations and the related quality control actions
4. On-line documentation: quality reports, papers, documents and questionnaire

Quality indicators:

1. Frame
2. Data collection
3. Data entry
4. Editing and imputation
5. Timeliness and punctuality
6. Costs

Accuracy

Information on E&I from SIDI

- SIDI standard indicators on E&I provide information on the overall impact of the specific E&I process adopted → quality of originally collected data
- SIDI metadata on E&I provide information on main characteristics of survey's E&I procedure

The implementation of the SIDI system

- SIDI manages more than 150 surveys
- The implementation of SIDI standard indicators is demanding task
  - additional response burden for survey managers
  - sometimes, survey procedures need to be renewed in order to compute indicators

  therefore, it was important to properly support the survey managers

Supporting Istat survey managers in producing SIDI standard indicators

- Especially designed training courses
- A net of quality pilots has been created (up to now 50 quality pilots have been trained for the most relevant Istat surveys)
- Developing generalised software:
  - to help the survey managers to calculate indicators
  - to avoid errors in calculations
  - to standardise the procedure
  - to speed up the procedure

Supporting Istat survey managers in evaluating and documenting E&I

Basic assumptions:

- The evaluation and documentation activities should be integrated in the production activity
- In order to evaluate E&I both qualitative (metadata) and quantitative information (indicators) is needed → *survey quality profile*

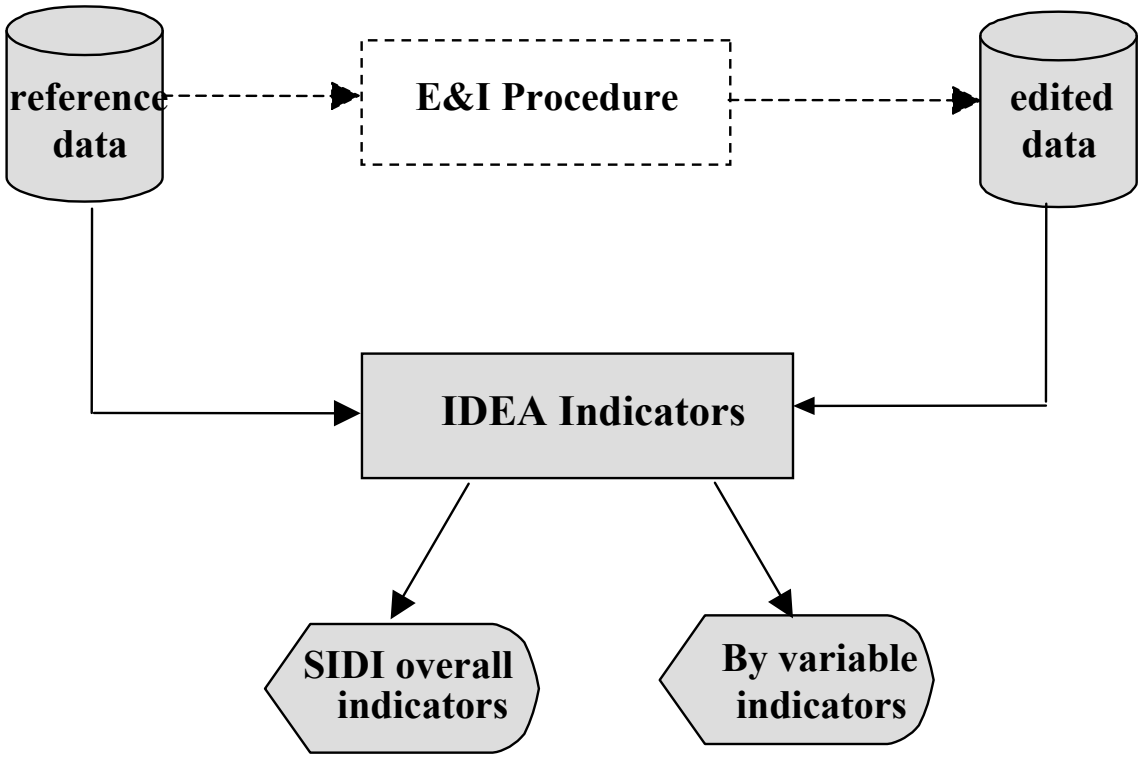
Supporting Istat survey managers in evaluating and documenting E&I: The IDEA software

Purpose

The IDEA software has been developed in order to:

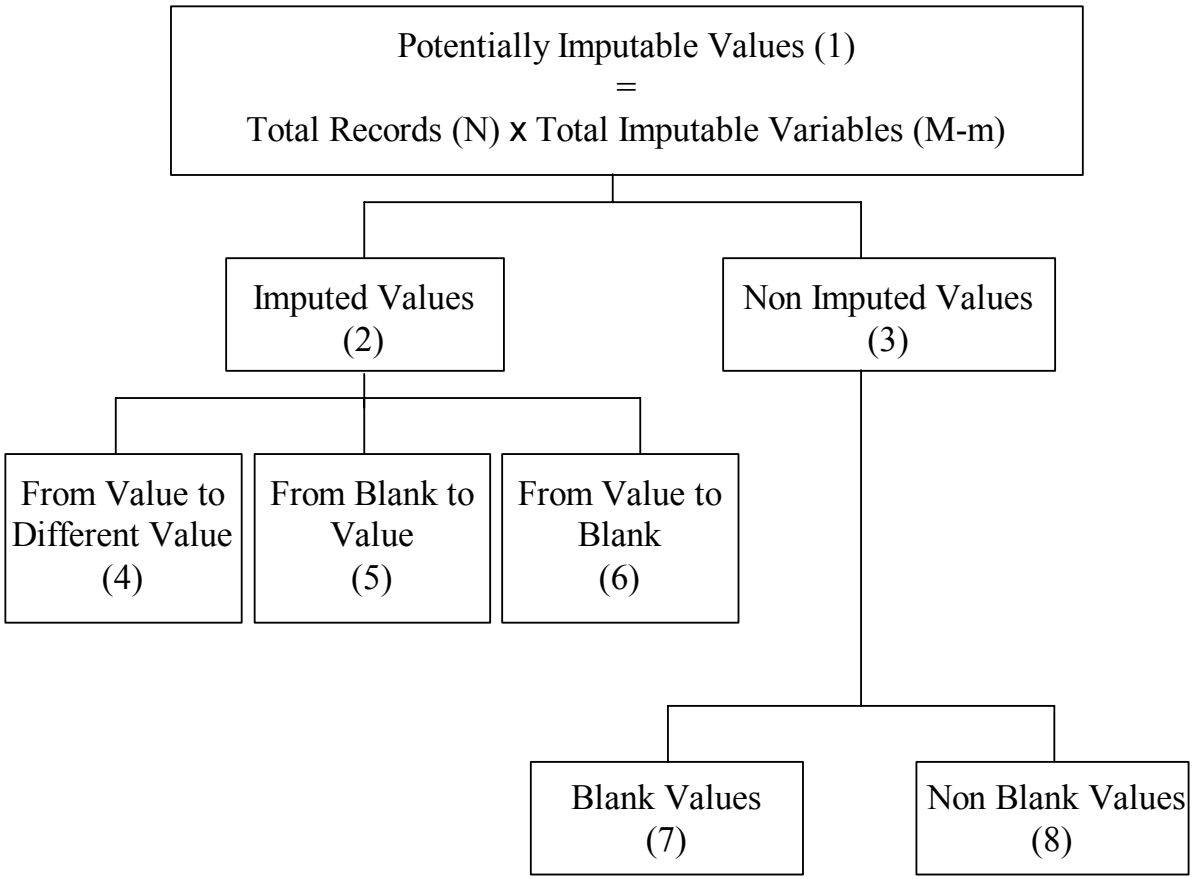
- provide survey managers with a standardized tool for computing the SIDI indicators for documentation purposes
- provide survey managers with a standardized tool for computing statistical measures for evaluation purposes at detailed variable level
- disseminate basic knowledge about the importance of evaluating E&I before, during and after data processing

Underlying data flow



SIDI standard indicators

Indicators (rates, distributional statistics) based on the following elements



Statistical measures at variable level

Evaluation criteria

The EUREDIT evaluation criteria have been adopted

- *Preservation of elementary data*
- *Preservation of distributions and aggregates*
- *Preservation of (marginal and joint) relations*

Evaluation measures

An initial set of descriptive measures for quantifying differences among distributions and relations in the compared data sets was used for starting populating the system

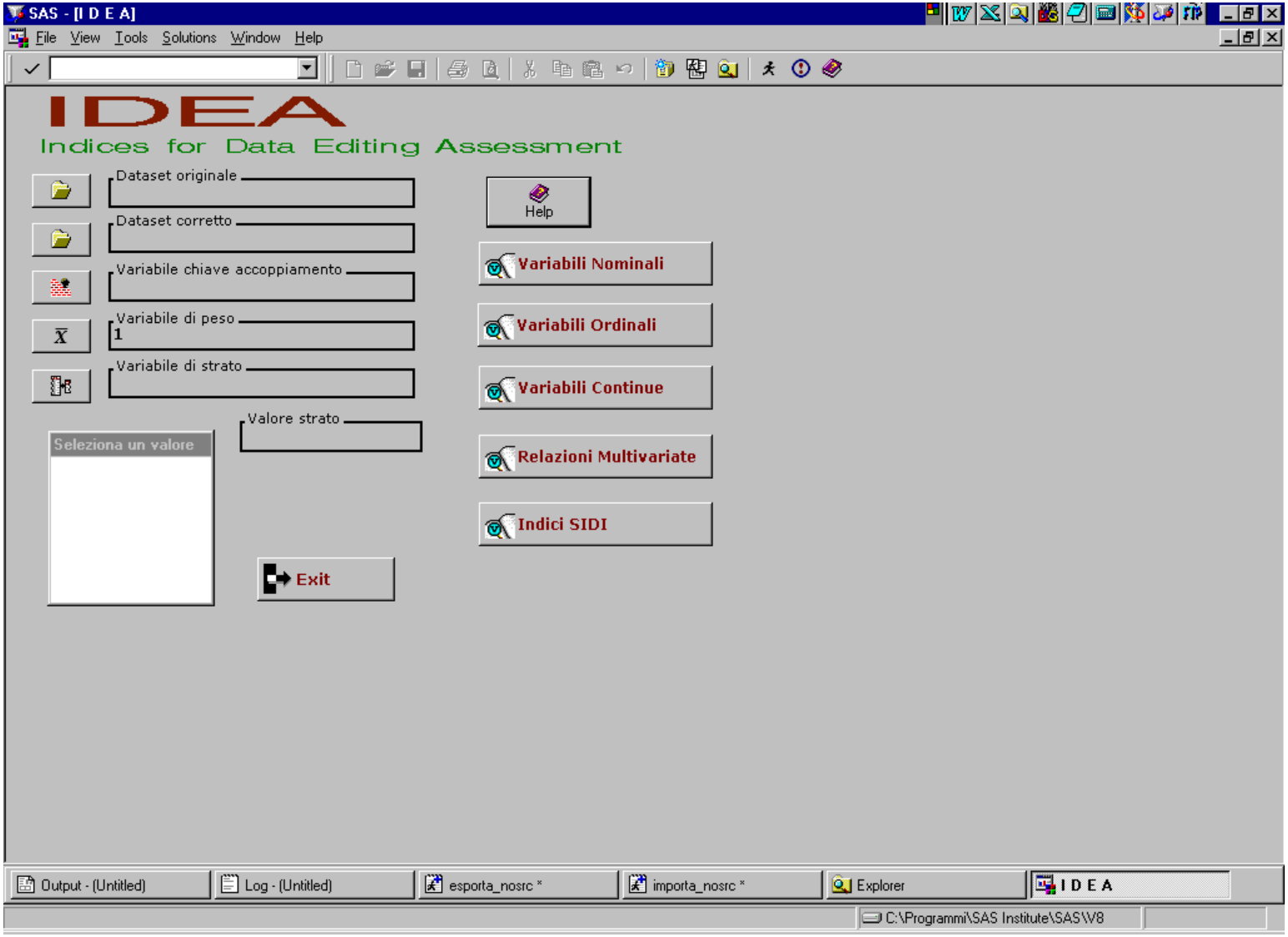
Statistical measures at variable level

- Separate evaluations are planned for different types of variables (nominal or ordinal variables, continuous)
- Indicators can be computed either at the end of the overall E&I process or after specific E&I sub-phases: in the latter case they provide information on the effects of the specific sub-phase on give subsets of variables, allowing for possible tuning
- Weighted indicators can be computed
- Indicators can be computed on the subset of values changed by the E&I process/method
- Indicators can be computed among different domains

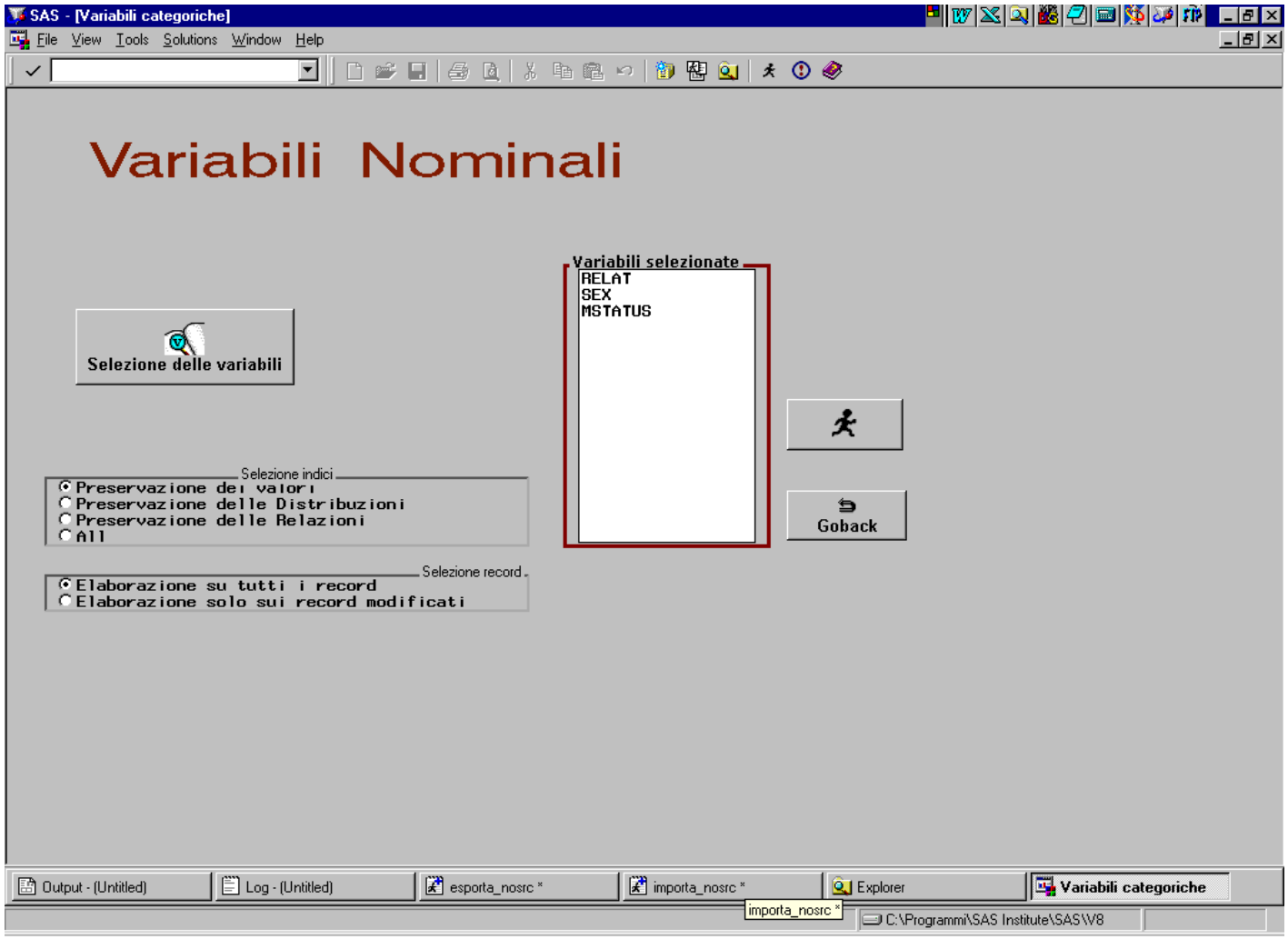
Other possible use of measures at variable level

- Together with SIDI indicators, to document the impact of E&I/the accuracy of original data at variable level
- *Evaluating the quality of E&I* when a set of "true" data is available, to perform
  - comparative evaluations of competitive methods
  - evaluation of the effectiveness of a single method

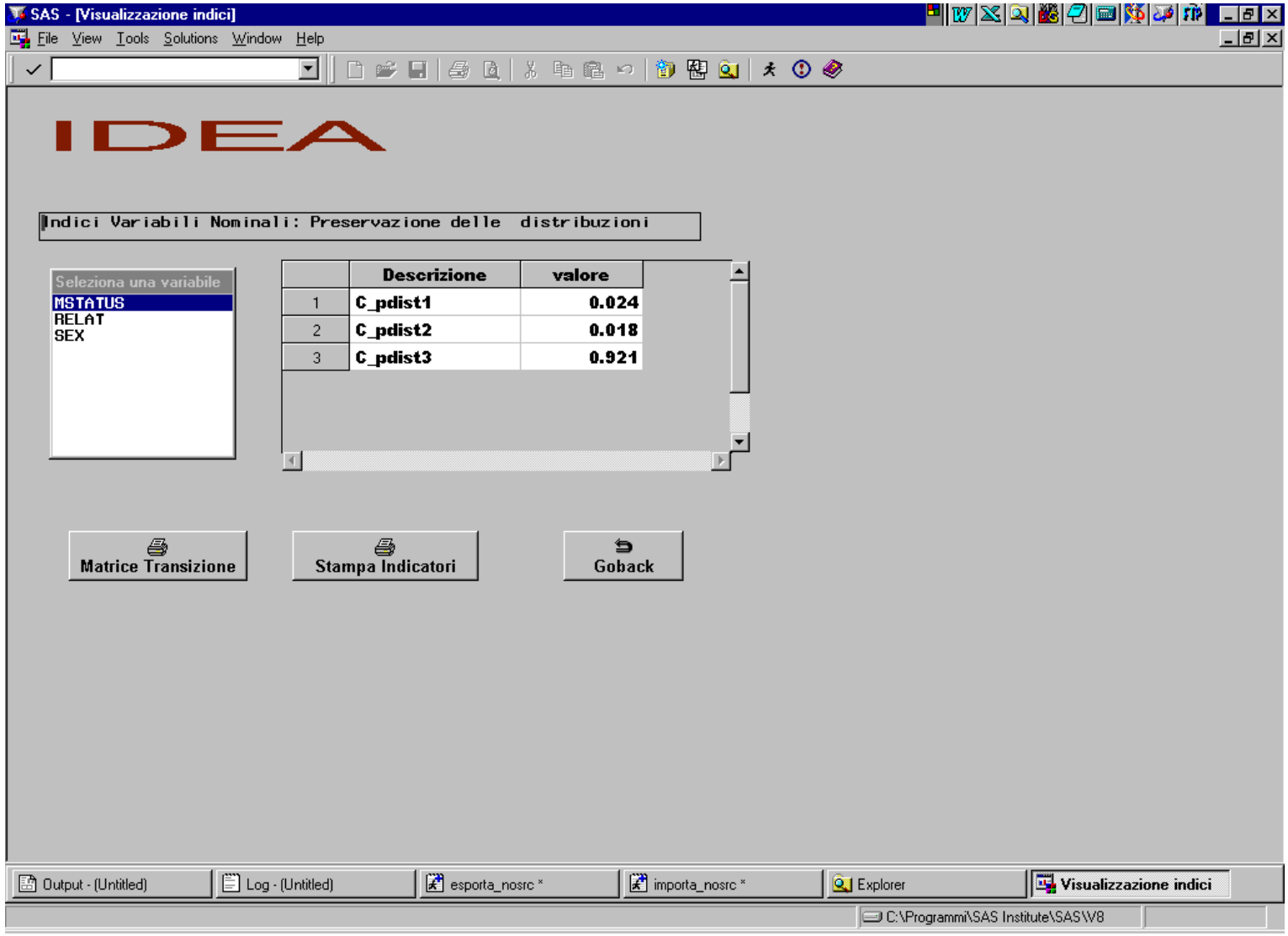
The IDEA software



The evaluation criteria



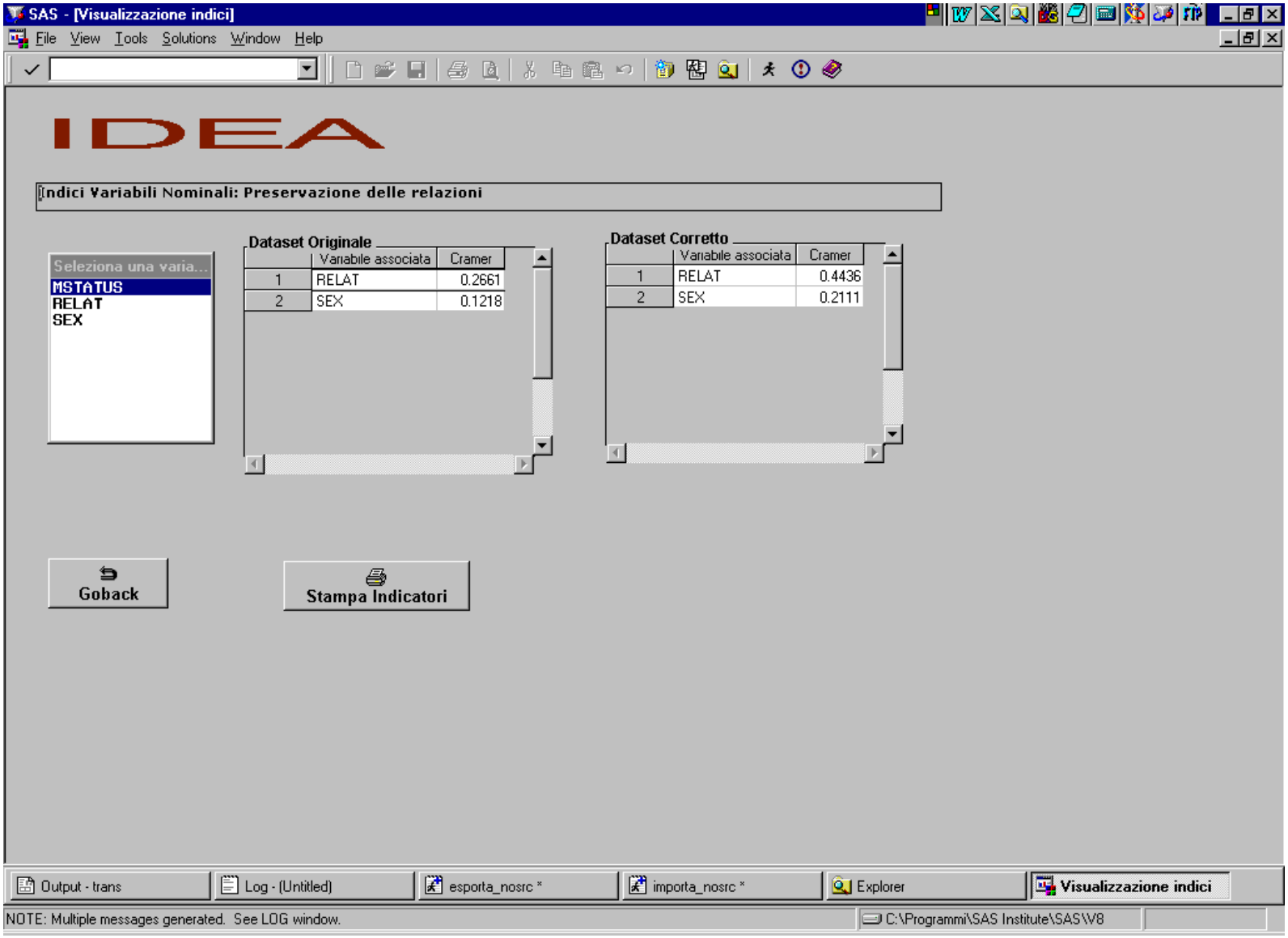
Categorical variables: preservation of distributions



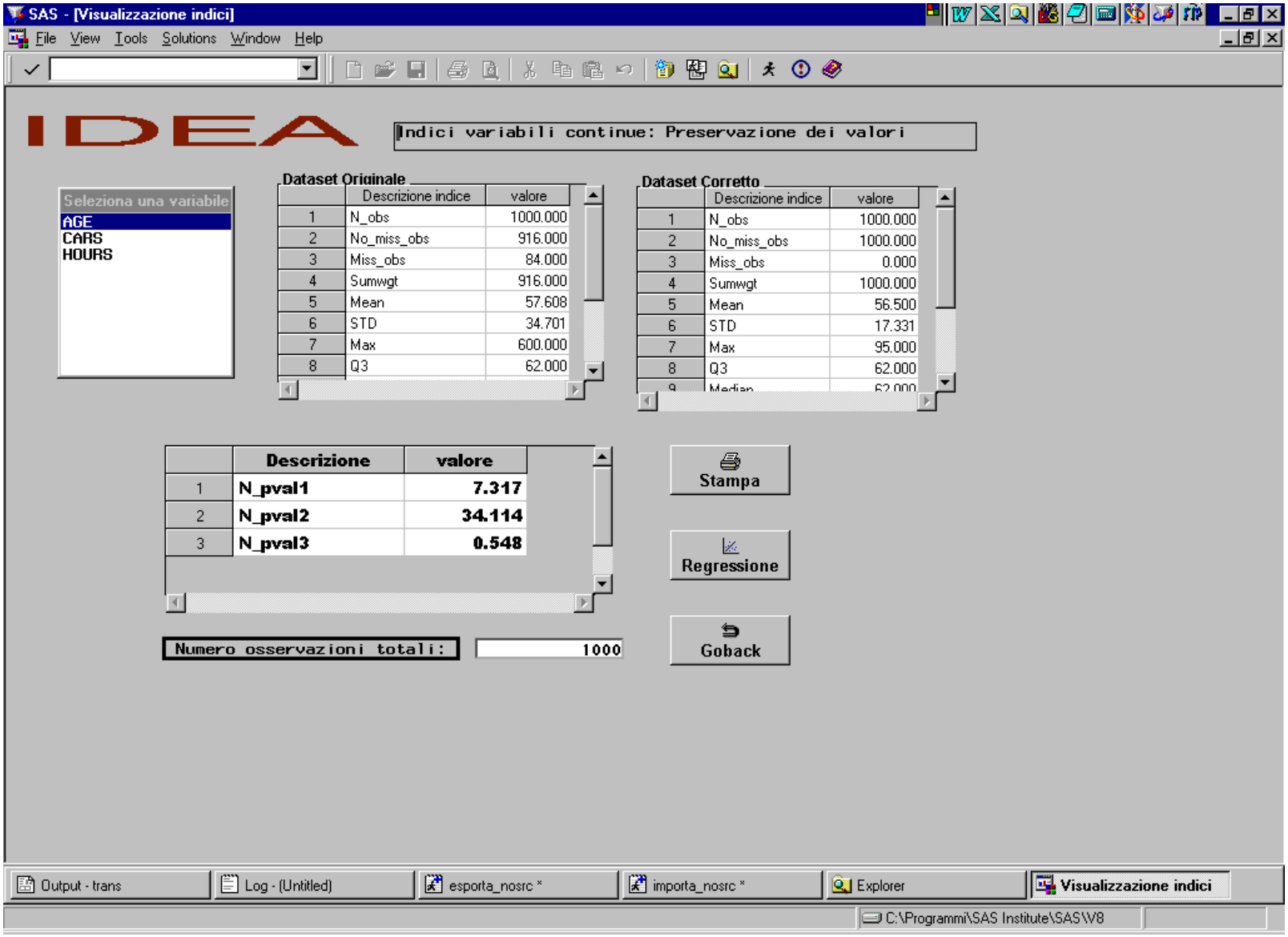
Categorical variables: Transition Matrixes

Table of MSTATUS by clr_MSTATUS						
	clr_MSTATUS					
MSTATUS	**1**	**2**	**3**	**4**	**5**	**Total**
.	12	58	13	3	14	**100**
0	0	1	1	0	0	**2**
1	187	3	0	0	0	**190**
2	0	461	0	0	0	**461**
3	0	0	54	0	0	**54**
4	0	4	0	44	0	**48**
5	0	2	1	0	123	**126**
6	0	4	0	0	2	**6**
7	0	2	0	0	1	**3**
8	1	1	0	1	2	**5**
9	0	5	0	0	0	**5**
Total	**200**	**541**	**69**	**48**	**142**	**1000**

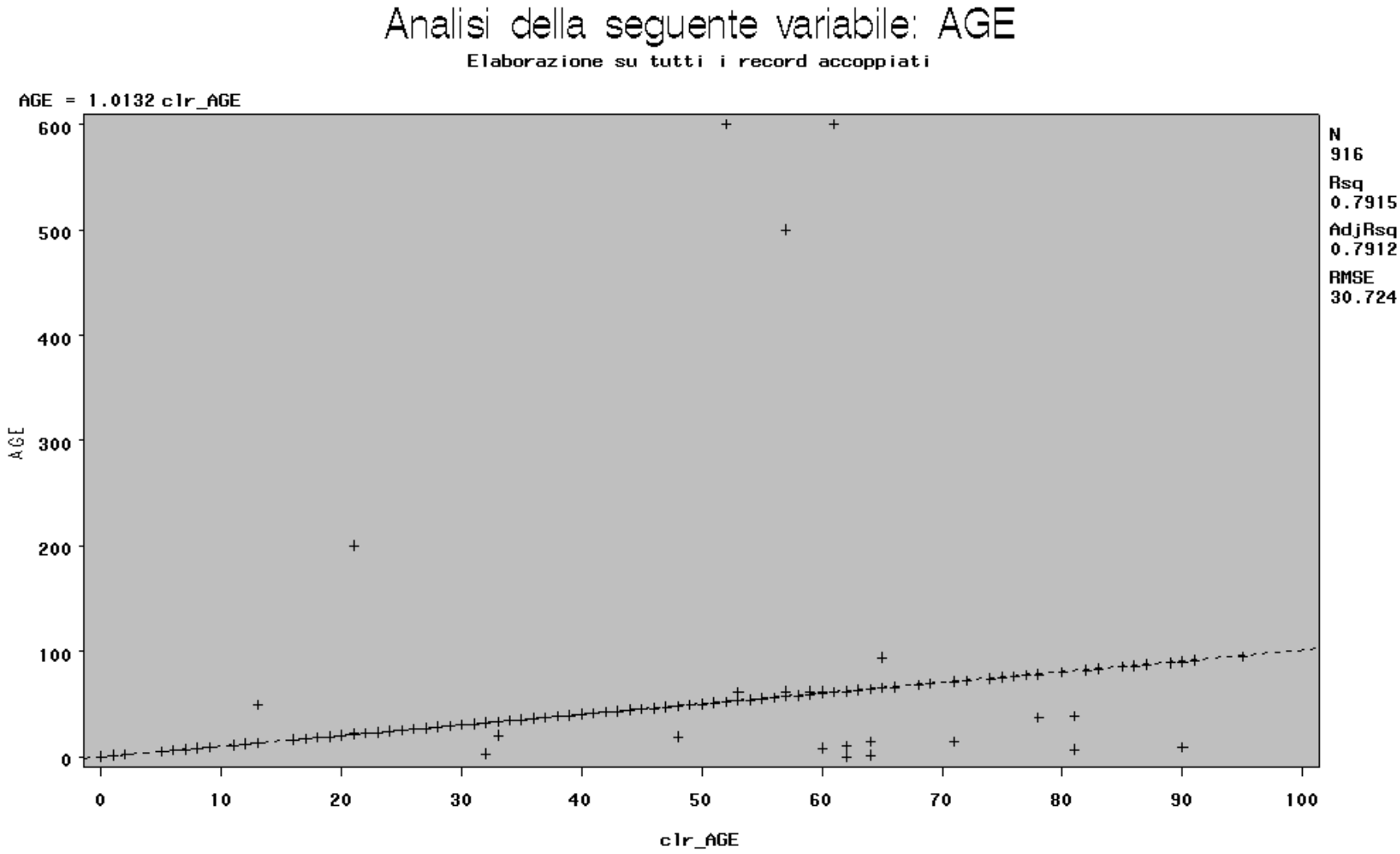
Categorical variables: preservation of relations



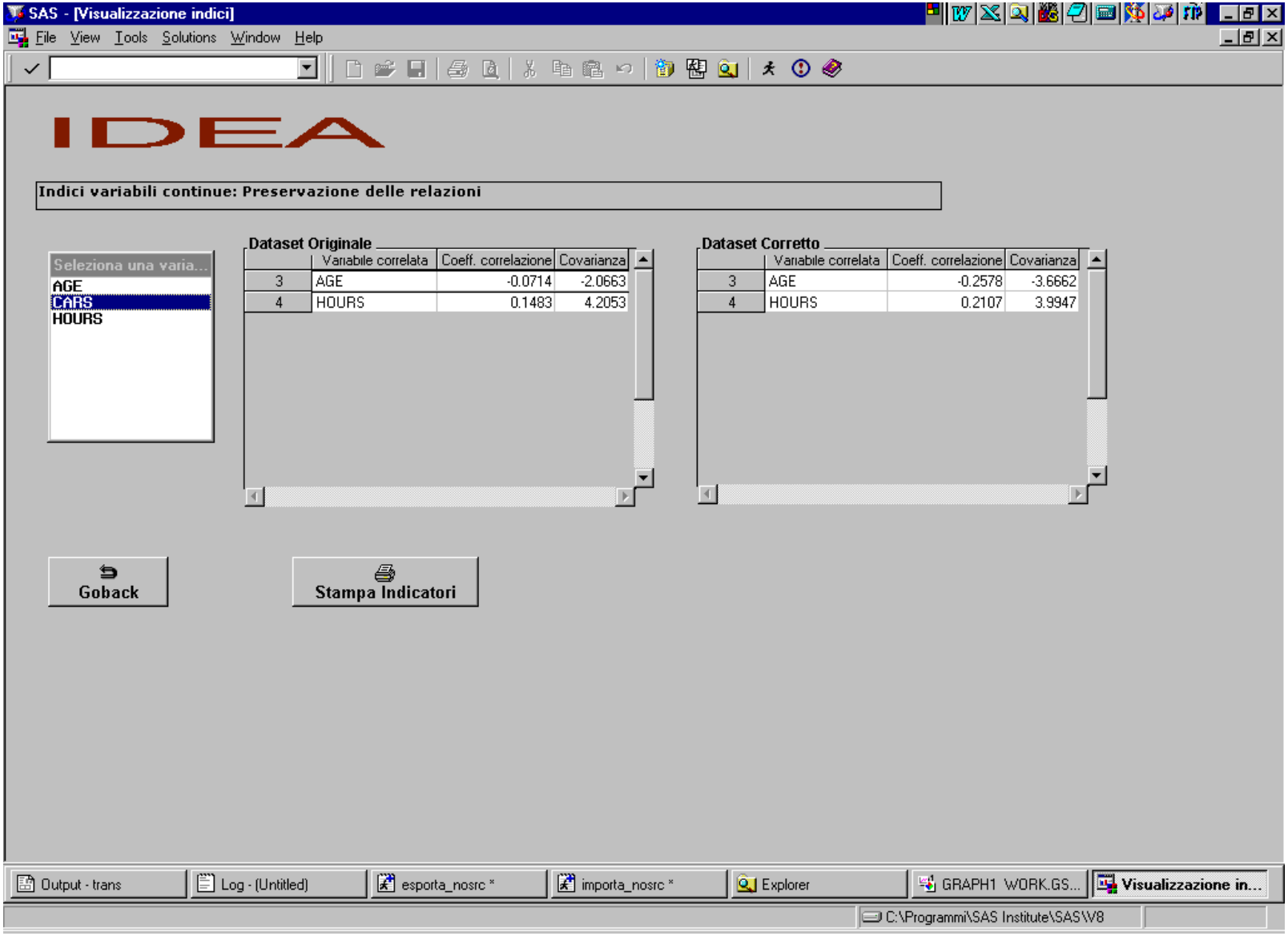
Continuous variables



Continuous variables: preservation of data



Continuous variables: preservation of relations



Main advantages when using IDEA

- Relating to SIDI indicators, additional burden on survey managers is eliminated
- Timeliness in producing SIDI indicators and updating the system is highly increased
- Open system: new indicators can be added in a very simple way
- Simple to use:
  - the most part of Istat surveys stores data in SAS archives
  - only raw and final data are required for all computations
  - neither technical skill nor additional programming effort is required to survey managers

Future work

- Adding new measures at macro (distributions, relations, aggregates) and micro level:
  - –Indicators from literature
  - –Indicators suggested by survey managers
- Identifying new standard measures for documentation purposes to be integrated in the SIDI information system