

INTEGRATED STATISTICAL SYSTEMS:
AN APPROACH TO PRESERVE COHERENCE BETWEEN
A SET OF SURVEYS BASED ON THE USE OF
PROBABILISTIC EXPERT SYSTEMS

Marco Ballin, Stefano De Francisci, [Mauro Scanu](#) (ISTAT)
Leonardo Tininini (CNR)
Paola Vicard (University Roma Tre)

NTTS, Bruxelles - 19 February 2009

OUTLINE

- 1 CONTAMINATION BETWEEN STATISTICS AND IT
- 2 BACKGROUND
- 3 COHERENCE AND POLICY EVALUATIONS
- 4 PROBABILISTIC EXPERT SYSTEMS
- 5 FROM PES TO INTEGRATION NETWORKS
- 6 COMPLEX SAMPLE SURVEYS
- 7 RESEARCH TOPICS

AIM OF THIS TALK

Statistics and IT can be usefully contaminated. The results can be effective also for Official statistics.

In the last years, joint research between Istat and academic institutions have distinctly addressed, among the others, these topics:

- 1 The use of probabilistic expert systems in official statistics (jointly with U. Roma Tre)
- 2 The production of integrated statistical information systems (jointly with CNR)

The objective of this talk is to provide the ideas behind a research project in Istat, that should integrate together the properties of probabilistic expert systems in the realm of integrated statistical information systems.

WHAT IS ALREADY AVAILABLE

- *Probabilistic expert systems* - A class of estimators defined by means of these models, that include HT estimator. Estimators are more efficient than HT, even if sometimes biased. Major motivation is the easy inclusion of an updating system that makes this class of estimators useful for poststratification.
- *Statistical information systems* - An integrated information system for the production of the statistical output (Istar) has been developed in order to integrate and manage the statistical data supplied and validated by the statistical production areas of Istat and produce purposeful statistical outputs for end users. The integrated system is based on the construction of several metadata layers. They cover not only the description, the design and the reference of the contents, but are also oriented towards the management of the navigation, the finding, the interchange and the semantics of the data.

MAIN MOTIVATION: COHERENCE BETWEEN ESTIMATES

Motivating example to tackle in the next future: an **integrated** statistical information system on agriculture. Its first version should contain results from

- FADN (Farm Accountancy Data Network Survey, focused on the economic performance of farms);
- FSS (Farm Structure Survey, focused on the structural aspects of farms as crop production, livestock, etc);
- sample frame (containing census results and data belonging to archives)

These data sources share many common variables (say X), especially on the farm structural characteristics.

COHERENCE BETWEEN ESTIMATES

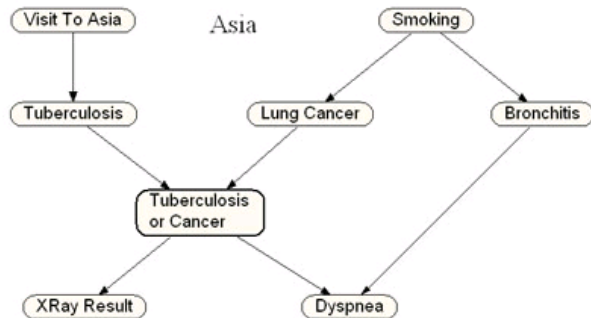
Once results from different surveys will be gathered in a SIS, some of the tasks to perform can be:

- 1 to ensure consistency on the common variables X ;
- 2 to simulate the distribution of some variables, under the control of a decision maker, in order to make clear the effects on the other variables (e.g. farmers' income, productivity, and so on), under the hypothesis of a fixed relationship between all the variables

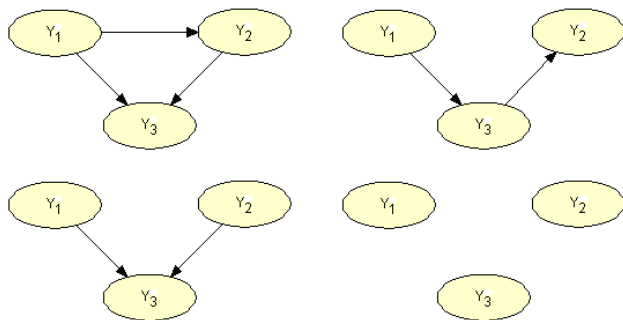
This talk proposes the use of statistical models (probabilistic expert systems) in order to fulfil the previous tasks

PROBABILISTIC EXPERT SYSTEMS

- a statistical model
- nodes = variables, arrows = dependencies
- different uses in applied statistics



SOME CHARACTERISTICS



Some characteristics:

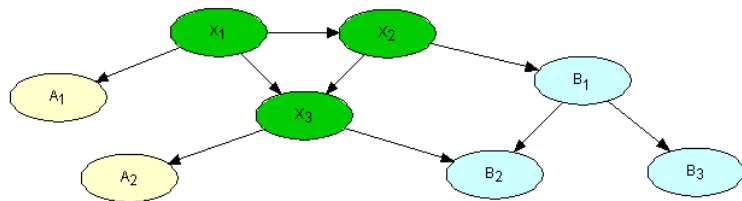
- 1 factorization lemma: $f(y_1, y_2, y_3) = \prod f(y_i | pa(y_i))$
- 2 updating system: once new information is available, it is disseminated by a simple algorithm based on the junction tree representation, and new tables are obtained

FROM PES TO INTEGRATION NETWORKS

Two groups of variables from two surveys

A: observe X_1, X_2, X_3, A_1, A_2

B: observe $X_1, X_2, X_3, B_1, B_2, B_3$

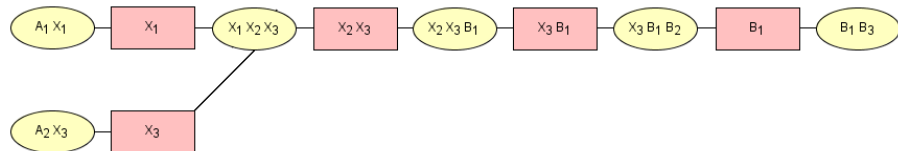


- Green nodes: common variables (constraint on PES: common structure, e.g. complete network)
- Yellow and light blue variables not connected!
- Hence: integration network cannot assume the role of the joint model between all the variables (it is a statistical matching problem)

UPDATING I

The updating system is very easy: it uses simple mathematical operators (products and ratios) on probabilities defined by a special network: the junction tree.

This is the junction tree of the previous example on A and B



UPDATING II

- 1 The yellow nodes represent the maximal subset of connected nodes in the integration network (cliques)
- 2 The pink rectangles are the common subsets of variables between adjacent cliques
- 3 (undirected) connections between cliques should fulfil the running intersection property
- 4 New information on a marginal distribution consists in repeatedly substituting the old joint distributions of the cliques:

$$p(i_c^*) = p(i_c) \frac{p(i_s^*)}{p(i_s)}$$

HOW TO ESTIMATE THE DISTRIBUTION IN A PES

The distributions of each node in a PES (or in an integration network) should be estimated. There can be different estimation approaches, according to the nature of the variables and to the objective to tackle

- ① if a variable is observed in only one survey, estimate its distribution from the corresponding survey
- ② if a variable is observed in more than one survey
 - ① substitute old/inaccurate information with information coming from the most updated/reliable survey
 - ② estimate a new distribution from all the surveys at hand

In either cases, update the integration network

WHAT SHOULD BE DONE I

- ① This approach relies on estimators defined on the statistical relationship between variables, expressed in terms of PES. Estimation of PES for complex sample surveys is a research topic. The first case to study: all the models are the complete model.
- ② It should be studied if it is useful to include nodes representing the different designs of the different sample surveys
- ③ Integrated statistical system can work with a selection of tables: This selection should be done carefully. The running intersection property of the junction tree should be fulfilled

WHAT SHOULD BE DONE II

- ④ The updating should be performed on one or more variables. In case of more than one variable, their joint distribution should be updated. Updating on more than one marginal distribution is not covered by the PES updating system (ratio raking to be included?)
- ⑤ The hierarchies of concepts (in particular objects and classifications) definable in IstarMD can help the designer (as well as the users) to distinguish and group the different kinds of data, e.g. if a table (or a fragment thereof) is based on an estimated forecast or derives from the modified distribution of a certain variable, etc

BIBLIOGRAPHY

Ballin, M., Scanu, M., Vicard, P. (2005) Model assisted approaches to complex survey sampling from finite populations using Bayesian networks. Working paper n. 54, Università degli Studi Roma Tre.

Sindoni, G., Tininini, L. (2008) Statistical Dissemination Systems and the Web. Handbook of Research on Public Information Technology, G.D. Garson and M. Khosrow-Pour (editors). Information Science Reference, 578-59

We wish to collaborate with researchers interested in the topic