

# **Integrated statistical systems: an approach to preserve coherence between a set of surveys based on the use of probabilistic expert systems**

Marco Ballin<sup>1</sup>, Stefano De Francisci<sup>1</sup>, Mauro Scanu<sup>1</sup>, Leonardo Tininini<sup>2</sup>,  
Paola Vicard<sup>3</sup>

<sup>1</sup>Istat, e-mail: {ballin, defranci, scanu}@[istat.it](mailto:istat.it)

<sup>2</sup>CNR, e-mail: leonardo.tininini@iasi.cnr.it

<sup>3</sup>Università Roma Tre, e-mail: vicard@[uniroma3.it](mailto:uniroma3.it)

## **Abstract**

The aim of this paper is to illustrate:

- 1) the use of graphical models known as Probabilistic Expert Systems (PES) for the description of the interrelationships of the phenomena observed in different surveys;
- 2) the use of the updating system of a PES as an automatic tool for establishing coherence between sample surveys in an integrated statistical system;
- 3) a system for the navigation of PES updated data to both assess the validity of some assumptions and perform some advanced analyses on both actual and updated data.

The motivating example is given by the integrated statistical system for agriculture, which is under construction. Other possible areas of use of PES are also decision making and policy evaluations.

**Keywords:** graphical models, dissemination of statistics, policy evaluation

## **1. Introduction**

When results of a sample survey are disseminated, it would be mandatory that the figures are consistent with the others of the same survey and with the ones of other surveys (on similar or overlapping topics). In the first case, internal coherence can be defined as the situation where all the figures of a survey can be produced marginalizing any disseminated table. In the second case, external coherence represents the situation where the figures of a variable studied in two or more different surveys (with the same reference population and time) are the same. The objective of this paper is to show how the dependence relationship among the variables of interest and the survey design is an important aspect to be considered in order to fulfill coherence properties. This notion leads to the so called PES based estimators (Ballin et al, 2005), which seem to be more efficient than usual estimators.

## **2. Motivating example**

In official statistics it is usual that some variables are observed in different surveys. Sometimes this fact has been seen as a problem of redundancy or a lack of efficiency in planning the questionnaires, because it increases the response burden. In the case of

structural agricultural surveys, the overlapping among questionnaires is planned. Such overlapping concerns the main aspects of farms (size, legal form, and so on) and it is used to update the sample frame and to ensure consistency between the surveys.

The construction of an integrated statistical system for the whole agricultural sector is the real world problem that motivated this research. The prototype of this system will consist mainly of three different statistical sources: 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.

A first problem is to ensure consistency on the common variables X, i.e. identification of a common statistical distribution for the common variables. This has effects on all the statistical tables that can be disseminated, even those that do not include X as explanatory variables.

Furthermore, it could be desirable 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.

### 3. Probabilistic expert systems

PES stands for Probabilistic Expert System (also known as Bayesian network, see Cowell et al, 1999) and it is a graphical representation (by means of a directed acyclic graph, DAG) of the joint distribution function consisting of nodes (the variables of interest) and directed edges (representing probabilistic dependence between pairs of nodes). Each node is assigned with the probability of the corresponding variable given its parents (when a node has no parents, it is assigned with its marginal probability distribution). In this set up, absence of directed edges corresponds to (marginal or conditional) independence between pairs of variables.

Figure 1: Four examples of PESs for three variables  $Y_1$ ,  $Y_2$  and  $Y_3$ .

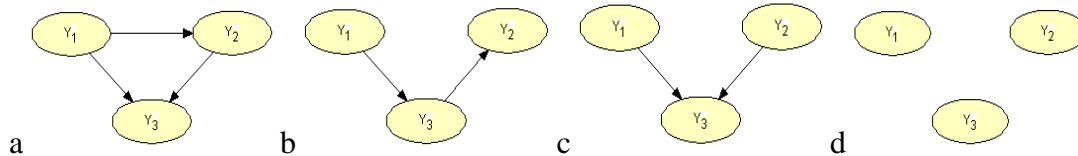


Figure 1 shows four examples of PESs for three variables  $Y_1$ ,  $Y_2$  and  $Y_3$ . Each figure corresponds to a particular model for the three variables, and describes the form of the joint probability distribution by a simple multiplication of the distributions attached to each node. For instance, the joint distribution for the PES in Figure 1.c is:

$$p(Y_1, Y_2, Y_3) = p(Y_1)p(Y_2)p(Y_3|Y_1, Y_2),$$

i.e. the joint distribution for the three variables is given by the product of the marginal distributions of  $Y_1$  and  $Y_2$  and the conditional distribution of  $Y_3$  given  $Y_1$  and  $Y_2$ . The four examples can be easily understood as the graphical representation of: a) the

saturated model; b) the model representing independence between  $Y_1$  and  $Y_2$  given  $Y_3$ ; c) the model representing independence between  $Y_1$  and  $Y_2$  but conditional dependence of  $Y_1$  and  $Y_2$  given  $Y_3$ ; d) the model representing independence between  $Y_1$ ,  $Y_2$  and  $Y_3$ .

PESs have been widely used in many applied settings: for instance in medicine, biostatistics, forensic identification, costumer segmentation and classification, computer troubleshooting, decision problems and so on (for more details on this topic, see Neapolitan, 2004). In the last years, PESs are also used in Official Statistics in contexts such as the description of the Census results (Getoor et al. 2001), or the treatment of missing values (Di Zio et al 2004, Di Zio et al 2005). A preliminary attempt to define PESs in the context of samples drawn according to complex survey designs is in Ballin et al (2005). In that paper, estimators based on the dependency structure between the variables of interest and the survey design are defined, taking respectively a model based and a model assisted perspective.

#### **4. PES as a tool for integration of two (or more) surveys**

PESs are particularly useful to integrate two (or more) surveys. In fact PESs are characterized by two positive aspects (Ballin et al, 2001):

- i) an easy-to-interpret, concise and, above all, informative way to represent a set of surveys with their dependence structure;
- ii) a method to update the information produced by a system of surveys, i.e. to pass information between surveys.

The first aspect points out that models avoiding the presence of unnecessary dependences are more appropriate for the definition of estimators. The second aspect is essentially based on the junction tree algorithm (Cowell et al., 1999). This algorithm, when one variable in a PES changes its marginal distribution, updates the distributions of the other variables.

The aim of this work is:

- 1) to propose a graphical representation (keeping some PES properties) in the case of integration of different sample surveys
- 2) to show how the junction tree works in this case.

##### **4.1. Graphical structures for the integration of different sample surveys**

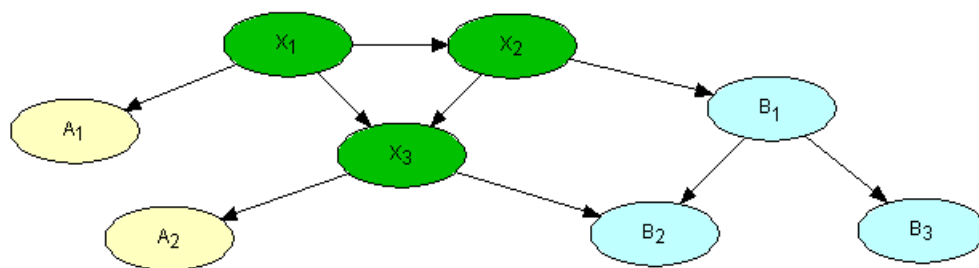
Ballin and Vicard (2001) introduced a graphical structure for the integration of two or more sample surveys with the requirement that each sample survey is represented graphically as a complete graph, i.e. its model is saturated. In this paper we will consider a more general situation.

Figure 2 shows an example where two surveys,  $A$  and  $B$ , collecting information on respectively  $A_1, A_2, X_1, X_2, X_3$  and  $B_1, B_2, B_3, X_1, X_2, X_3$ , are represented. This is a typical situation in many multipurpose surveys, where there is a core of common variables, i.e.  $X_1, X_2, X_3$ , and each survey investigates a particular topic.

Integration of the two surveys essentially means coherence of information. Coherence can be obtained when the distributions of the common variables in two surveys are the same. This rarely happens in two surveys performed in distinct times (e.g.  $A$  before  $B$ ). The junction tree algorithm can be applied to update  $X_1, X_2, X_3$ , in  $A$  forcing them to have the same distribution estimated in  $B$ .

Note that the integration network in Figure 2 is *not* a *real* PES, unless  $(A_1, A_2)$  and  $(B_1, B_2, B_3)$  are independent given  $(X_1, X_2, X_3)$ . In general the integration network is obtained overlapping the PESs of different surveys with the requirement that the common variables  $(X_1, X_2, X_3)$  in Figure 2) form a complete subgraph and separate the sets of variables observed distinctly  $(A_1, A_2)$  and  $(B_1, B_2, B_3)$  in Figure 2). The integration network does not represent any more the statistical relationship between all the variables (due to the absence of information on the joint distribution of those variables that were never jointly observed, as  $A_1$  and  $B_2$ ). Nevertheless it is still precious for our purposes. In fact, the graphical structure of an integration network is still a directed acyclic graph that preserves the possibility to update the distributions of the different survey variables when new information becomes available.

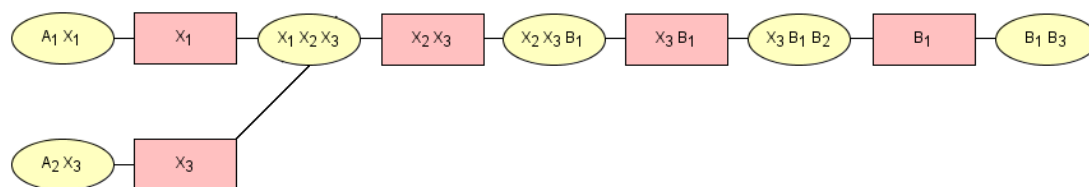
Figure 2. Integration network for surveys A and B. Green nodes represent the common variables



#### 4.2. How the junction tree works

The algorithms for propagating new information in an integration network (as well as in a DAG) are based on a graphical structure named junction tree. The junction tree is a *hypergraph* whose nodes, named *hypernodes*, are complete subsets of variables (named *cliques*). Furthermore, a junction tree should fulfill the *running intersection property* that is: for any two cliques  $C_i$  and  $C_j$  in the set of cliques and any clique  $C'$  on the unique path between them in the junction tree,  $C_i \cap C_j \subset C'$ . Rules for obtaining a junction tree from a DAG are described in Cowell et al. (1999).

For instance, the junction tree of the graph in Figure 2 is:



When new evidence about a node is available, it is propagated through the network by means of the distributions of the separators (rectangular shaped nodes) shown in the junction tree. For example, considering a generic clique  $C$ , its distribution is updated according to information carried by separator  $S$  as follows (see Cowell et al, 1999):

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

### **4.3. Propagation of information when samples are drawn according to complex survey designs**

Let us distinguish between two types of nodes in the integration network.

The first group of nodes corresponds to those that are observed in only one survey (as  $A_1$  and  $A_2$  in  $A$  and  $B_1$ ,  $B_2$ , and  $B_3$  in  $B$ )

We suggest that every distribution to be attached to these nodes in the integration network is estimated from the corresponding survey, according to the survey weights observed in those survey, using a Horvitz-Thompson estimator (or a Hajék estimator, if the weights do not sum to the population size).

The second group of nodes corresponds to those nodes that are observed in more than one survey.

There are two possibilities:

- a) substitute new information on the common variables  $X_1$ ,  $X_2$ ,  $X_3$  from one survey to the other (e.g. from  $A$  to  $B$ , because  $A$  is more recent, or more accurate, etc).
- b) estimate the joint distribution of  $X_1$ ,  $X_2$ , and  $X_3$  using  $A$  and  $B$  together, as a unique sample.

The first case is an answer to the following question: what is the new distribution of the variables  $B_1$ ,  $B_2$ , and  $B_3$  if the distribution for  $(X_1, X_2, X_3)$  is set to the one observed in  $A$ ? Hence, the propagation of information involves only the variables in the  $B$  survey.

The second case takes advantage of the information from both the surveys in order to get a new estimate of the joint distribution for the common variables  $(X_1, X_2, X_3)$ . This is not an easy task in the case of sample surveys collected by complex survey designs. A solution could be the calculation of new weights for the sample given by the union of  $A$  and  $B$ . This task can be performed following the “file concatenation approach” as in Rubin (1986) or, in case there is substantial knowledge on the design variables as in Torelli et al (2008). This new distribution will be included in the integration network, and the tables will be updated by means of the junction tree algorithm.

## **5. Navigating updated data**

Integration networks represent a very concise way to show the dependences between two or more surveys, as expressed by the variables in common. Nevertheless, in some cases it is important to analytically browse the data (often only small significant portions thereof) arising from the updating system of the corresponding junction tree, by assessing the validity of the updating process and possibly fine-tuning some of the assumptions made (e.g. the assumed order of reliability among the data sources to be integrated, which indirectly determines the structure of the junction tree). To this aim the data produced by the updating system explained in Section 4 can be used as a data source for the IstarMD system and later freely navigated by using its WebMD component.

IstarMD is a component of the Integrated Output Management System in Istat, an information system oriented towards the integration of part of the statistical data life cycle. It has been developed in order to integrate and manage the statistical data

supplied and validated by the statistical production areas in Istat to produce purposeful statistical outputs for end users.

The whole system can be considered as a multilevel and multiservice integration environment: it is supplied with flexible and multiple mechanisms for interchanging, sharing and integrating data. The system processes move through specific workflows and allow for:

- the design of semantic metadata layers for modelling statistical data marts of elementary data
- the extraction of both elementary and aggregated data from heterogeneous sources
- their transformation into multidimensional format
- the data loading into statistical data warehouses
- the management of metadata for Web navigation and aggregate data computation
- the dissemination of information to many different users, by means of different types of channels and technologies.

In this scenario, several different kinds of the integration network generated data could be easily browsed by using the IstarMD navigation system, e.g.:

- estimated data for a (possibly future) year starting from the data available for two past years;
- data obtained by the integration (and integration network-based updating) of interdependent variables coming from several, possibly partially incoherent, data sources;
- data obtained by artificially modifying the distributions of a few “pivot” variables to analyse how this impacts (through the integration network) the distributions of the other variables (this is a kind of “what-if” analysis).

The two main components of the IstarMD toolbox are the above mentioned WebMD (the component for multidimensional navigation and dissemination on the Web) and FoxtrotMD (the “administration” component for metadata management and aggregate data computation).

WebMD originates from the DaWinciMD dissemination system (Sindoni et al, 2006), initially developed to disseminate aggregate data from the 2001 Italian Population and Housing Census<sup>1</sup> and more recently used to disseminate, among the others, data from:

- the graduate education and employment Italian survey<sup>2</sup>;
- different surveys for setting up a system about “The framework for integrated territorial policies”<sup>3</sup>;
- the household budget survey of the Bosnia and Herzegovina Agency of Statistics<sup>4</sup>.

All data stored in IstarMD are based on the concepts of object, classification and basic table:

- an *object* corresponds to the application of an aggregation function to a certain collection of analysis units and typically takes its description from the first part of the corresponding statistical table’s title (more precisely the part before the “by ...” section of the title, i.e. the part independent from the classificatory

---

<sup>1</sup> <http://dawinci.istat.it/MD>

<sup>2</sup> <http://dip.istat.it/>; <http://lau.istat.it>

<sup>3</sup> <http://incipit.istat.it/>

<sup>4</sup> <http://hbsdw.istat.it/dawincibosnia>

details). Examples of objects are “Number of households”, “Resident population aged 6 and over”, “Average income”, etc.

- each object may have zero, one or more *classifications*, e.g. an object like “Resident population aged 6 and over” may be classified by “sex”, “five years age groups”, “marital status”, etc.
- the combination of an object with a certain number of classifications constitutes a *basic table* (or b-table for brevity), which is somehow an abstraction of a conventional published statistical table.

In order to represent an actually published table (and hence a collection of aggregate values) each b-table has to be *spatio-temporally instantiated*, i.e. be put in correspondence with one or more specific combinations of times and territorial areas. In other words, each b-table can have many different *spatio-temporal instantiations*, i.e. be associated with collections of aggregate values corresponding to different years (or months, or other time periods) and territories, possibly at different levels of detail. By structuring the tables to be published in terms of b-tables and spatio-temporal instantiations the dissemination designer can define very precisely which aggregate data can be accessed through the WebMD navigation interface and which cannot. Furthermore, the structuring of b-tables in terms of objects and classifications, as well as the territorial hierarchies defined in the spatio-temporal instantiation, enables the user to navigate the data by exploiting a multidimensional paradigm, very similar to the one made available by *data warehouse systems*, and based on the well known metaphor of the *data cube* (Kimball et al, 1998).

The concept of object in IstarMD basically corresponds to that of measure in a conventional data warehouse, although an object may also incorporate some *slicing* operations on the data cube. In order to guide the user in selecting the required cube, objects are organized into hierarchies, mainly based on generalization relationships, and the user can choose “generic” objects, i.e. those located in the higher levels of the hierarchy and hence corresponding to more general and abstract concepts. The system is able to combine the generic user choices and map them to the actual object-classification combinations specified by the metadata.

WebMD classifications basically correspond to specific dimension levels of the data cubes, although a classification’s structure can be more complex than usual flat dimension levels. Classifications can be shared by several objects (and hence data cubes), enabling a user to perform classification-based navigations: the user can select a combination of classifications and ask the system to show all available statistical aggregates (cubes) classified in that way, independently of the measure. As with objects, classifications are organized into hierarchies, to enable the user to express generic queries and consequently facilitate access to data.

The selection of a statistical table (cube) is based on the interdependent selection of the object and classifications of interest. Figure 3 shows the table selection page of WebMD, enabling the user to express the required table by selecting (without a predefined order and possibly only in part) the object, classifications, territory and year of interest.

Figure 3. The table selection page of WebMD

10 tables compatible with the choices already made (display the details)

Choices made
Object <any>
Classifications <unclassified data>
Year - <any>
Territorial partit. - <any>
Territory - <any>

Objects | Classifications | Territory | Year | Tables

Choose the **object** of interest

- Objects
  - Housing units
    - conventional dwellings
    - occupied housing units
  - Industry, trade and services
    - labour force
  - Population and households
    - households
    - couples
    - resident population

By performing some selection in one of the tabs (objects, classifications, territory and year) the user can then browse the information in the other four tab panels to analyze which data have been made available for publication. For instance after an object's selection he/she may view either all (and only) classifications that are available to classify it, or the territorial areas for which that object was elaborated, or also the list of statistical tables having that (or a more specific) object as component. Likewise, by selecting a specific year, the user can browse the other tab panels to view the tables available in that year, and correspondingly the objects and classifications constituting them, as well as the territories for which such tables are available in the selected year.

## 6. Feeding IstarMD with PES updated data

FoxtrotMD is IstarMD administration component specifically designed for metadata management and aggregate data computation. By means of FoxtrotMD the dissemination designer can define very precisely the statistical tables to show, as well as the rules to be applied on the source data in order to obtain the corresponding aggregate data. In the integration network context, source data will be not only the original source data, but also those updated and generated by the updating system of the corresponding junction tree.

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, derives from the modified distribution of a certain variable, etc. By using FoxtrotMD the dissemination administrator can:

- manage the *objects* of interest for the statistical tables to be disseminated, in particular their descriptions in the two languages chosen for publication, the related statistical tables (i.e. tables defined using a given object), as well as the parent (hierarchical) relationships between them. As mentioned above, objects can indeed be arranged into a hierarchical structure based on generalization. In particular, this mechanism can be exploited to group some objects, making them, for example, children of the same *abstract object* "Population estimates". An IstarMD abstract object like "Population estimates" is mainly introduced to group concepts and facilitate the user access to data, although it does not correspond to a single actual statistical table.

- manage the *classifications* of interest for the statistical tables to be navigated, in particular their descriptions in the two languages chosen for publication, the corresponding modalities in both languages, the related statistical tables (i.e. tables defined using a given combination of classifications), as well as the parent relationships between them. Similar to objects, classifications can indeed be arranged into a hierarchical structure based on generalization and level of detail.
- manage the *statistical tables* to be disseminated, defined by the combination of an object with a certain number of classifications. Each table will have its own multi-language descriptions, object and classification components and possibly multiple *spatio-temporal instantiations*, i.e. combinations of territories and years for which data are available (and have to be disseminated). FoxtrotMD also enables the dissemination administrator to define the rules to extract and aggregate the data to be disseminated, starting from tables of microdata or from a table of data updated by the integration network.
- compute and store the *aggregate data* to be disseminated. By using the specified rules, the *ETL component* of FoxtrotMD can extract/aggregate the data and store them in the aggregate data table used during statistical table visualisation by WebMD. The aggregation process is automatically performed at all levels of the territorial partitioning hierarchy specified by the administrator.

## 7. Further problems and conclusions

Apart from the problem of integration, it is worth mentioning other possible uses of PES. For instance, it is possible to use these tools for the evaluation of policies. This can be performed by modifying some variables (under the control of the decision maker) and verifying the effect on the statistical tables of interest.

Additionally, PES can easily be extended as a decision tool, that includes also possible costs and utilities of an action.

These tools can be included in the PES. In this case, PES do not only preserve consistency among the information gathered between different statistical sources, but they constitute a valid tool that facilitates the communication with the users of official statistics.

## References

- Ballin, M., Vicard, P. (2001) A proposal for the use of graphical representation in official statistics, in *Proceedings of the conference SCO2001*, 24-26 September 2001, Bressanone/Brixen.
- 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.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*, Springer, Heidelberg.
- Di Zio, M., Scanu, M., Coppola, L., Luzi, O., Ponti, A. (2004) Bayesian Networks for Imputation, *Journal of the Royal Statistical Society, A*, **167**(2), 309-322.

- Di Zio, M., Sacco, G., Scanu, M., Vicard, P. (2005) Multivariate techniques for imputation based on Bayesian networks, *Neural network world*, **2005/4**, 303-309.
- Getoor, L., Taskar, B., Koller, D. (2001) Selectivity estimation using probabilistic models, *Proceedings of ACM-SIGMOD 2001 International Conference on Management of Data*, Santa Barbara, California, USA.
- Kimball, R., Reeves, L., Ross, M., Thornthwaite, W. (1998) *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*, Wiley, New York.
- Neapolitan, R. E. (2004) *Learning Bayesian Networks*, Prentice Hall, Upper Saddle River.
- Rizzo, F., De Francisci, S. (2007) An integration approach for the Statistical Information System of Istat using SDMX standards, *Meeting on the Management of Statistical Information Systems (MSIS 2007)*, Geneva, 8-10 May 2007
- Rubin, D. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, 4, 87-95.
- Sindoni, G., Tininini, L. (2006) Statistical warehousing on the Web: navigating troubled waters. In *Proceedings of the International Conference on Internet and Web Applications and Services (ICIW'06)*, IEEE Press, 2006.
- 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
- Torelli, N., Ballin, M., D'Orazio, M., Di Zio, M., Scanu, M., Corsetti, G. (2008) Statistical matching of two surveys with a non randomly selected common subset, *Proceedings of the ESSnet-ISAD workshop*, Vienna, 29-30 May 2008.