rivista di statistica ufficiale

n. 1
2018

In this issue:

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 1
2018

In this issue:

Istat

rivista di statistica ufficiale
n. 1/2018

ISSN 1828-1982

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Editorial Preface

The present issue of the *Rivista di statistica ufficiale* offers a reprint of an article written by Mauro Scanu and published in its past N. 3/2004. This both in order to make it easily accessible online, and to take the opportunity of showing again its scientific contents which are still extremely topical and of great interest for official statistics.

The paper deals, indeed, with Bayesian networks and their applicability in some critical issues related to survey data production and analysis.

The first part of the article focusses on the meaning and the concept of Bayesian networks, describing all their possible applications in the fields of decision-making, discovering causal relationships, risk prediction and assessment, data mining, reliability analysis, etc.

In its following sections, the article highlights the relevance of Bayesian networks for official statistics.

The second original article proposed in this issue is produced by Alessandra Fasano, Giulia Nisi and Ludovica Rossotti. By means of a complete literature review, this study provides an interesting analysis of coworking and investigates the drivers that are contributing to the rise and development of this new job method based on the use of shared working spaces.

The paper goes deep into several aspects of interest, such as the different workforce generations, their attitudes and behaviour in terms of work organisation, showing an overview of the increasingly worldwide spreading of coworking, with a specific focus on the Italian scenario.

The contents also stress the importance of this field of research, which proved to be strategic for enhancing the production of official statistics, especially with regard to the most innovative and current aspects of the labour market.

The present issue closes with the scientific paper by a group of experts representing a virtuous synergy between several research institutes of different countries: Simone Navarra, Marisa Cappella, Lars Age Johansson, László Pelikan Friedrich Heuser, Luisa Frova and Francesco Grippo.

The paper focusses on the decision tables for mortality coding and a web-based system developed by the Italian National Institute of Statistics - Istat within the framework of an international collaboration.

By means of this application, representatives from different countries can now collaborate in a coordinated and harmonised way, and commit to a common effort aimed at the simultaneous maintenance and update of the decision tables used for the underlying cause-of-death selection.

These tables provide criteria for the correct application of the selection rules of the *International Classification of Diseases*, published by World Health Organization – WHO.

The future development envisages the inclusion of these tables in order to implement a tool for their systematic management.

This represents a step towards the standardisation of multiple cause rules, which will result also in better multiple cause data and will be available for innovative research purposes.

Nadia Mignolli

Coordinator of the Editorial board

An overview of methods in official statistics based on Bayesian networks

Mauro Scanu [1]

Abstract

Bayesian networks are a graphical formalisation of a joint multivariate distribution. They are efficiently exploited in many different applied settings. In these last years, some applications in official statistics have been defined. This paper illustrates at first the concept of Bayesian networks, and then focusses on applications in official statistics.

Keywords: graphical models, imputation of missing items, complex survey designs.

1 Italian National Institute of Statistics - Istat (scanu@istat.it).
The views and opinions expressed are those of the author and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

1. Introduction

"*...Bayesian networks are complex diagrams that organize the body of knowledge in any given area by mapping out cause-and-effect relationships among key variables and encoding them with numbers that represent the extent to which one variable is likely to affect another...*" The previous quotation is from the Los Angeles Times (Helm, 1996). In that article Bill Gates and other researchers at *Microsoft* explain how the usual computers were deaf, dumb, blind and clueless, and how "Bayesian stuff" could be used in order to make computers more interactive with human beings. In the following example, *Microsoft* applications with Bayesian networks are briefly reviewed.

Example - The first Bayesian network application in *Microsoft* programmes is the so called paperclip (or *Office assistant*, see Figure 1), firstly programmed by Horvitz, a researcher at *Microsoft*. The annoying features of the paperclip may suggest the reader to immediately stop understanding and using Bayesian networks! However, as stated in the following quotation from a newspaper article (The Economist, 2001), the original tool has been modified: "*...The paperclip in question, as even casual users of Microsoft's Office software will be aware, is a cheery character who pops up on the screen to offer advice on writing a letter or formatting a spreadsheet. That was the idea, anyway. But many people regard the paperclip as annoyingly over-enthusiastic, since it appears without warning and gets in the way. To be fair, that is not Dr Horvitz's fault. Originally, he programmed the paperclip to use Bayesian decision-making techniques both to determine when to pop up, and to decide what advice to offer….The paperclip's problem is that the algorithm (sequence of programming steps) that determined when it should appear was deemed too cautious. To make the feature more prominent, a cruder non-Bayesian algorithm was substituted in the final product, so the paperclip would pop up more often….*".

Figure 1: The paperclip (*Office Assistant*) implemented in *Microsoft Office*

This first attempt has been followed by many Bayesian networks based tools more respectful of Bayesian network theory (see for instance the following web page: http://www.microsoft.com/research/default.aspx). They include the selection of the items in the sometimes long context lists, user modelling and intelligent user interfaces (not only the already discussed *Office Assistant* implemented in *MS Office*, but models, theory and systems implemented in *Priorities*), diagnostics, trouble shooting and sensor fusion. All these tools use the *Windows*-based application for Bayesian belief network (Belief network is a synonym of Bayesian network) construction and inference called *Microsoft Belief Networks* (MSBN, see Kadie *et al*, 2001), available free for non-commercial purposes (http://research.microsoft.com/adapt/MSBNx/).

All the applications described in the previous example deal with the "decision making" problem. This is not the only problem that Bayesian networks tackle. Among the others, Bayesian networks have been proved to be useful for discovering causal relationships, prediction, assessment of risk, evolution in a simulated world, data mining, reliability analysis. The application fields are the most diverse, from biology (analysis of gene expression data) to medicine (diagnostics), psychology (cognitive psychology), artificial intelligence, speech recognition and weather forecasting (for a complete overview of Bayesian networks applications see Neapolitan, 2004, Chapter 12). The use of Bayesian networks in all these fields is justified by the interaction between an easily manageable set of multivariate statistical models and the existence of fast and efficient statistical algorithms for their estimation and use. This aspect is the motivation of a profitable use also in many different official

statistics problems. Applications in official statistics are yet in their infancy. Preliminary results date to the beginning of this century (Getoor *et al*, 2001a; Sebastiani et al, 2001b, Thibaudeau *et al*, 2002). The topics of imputation of missing items and of the multivariate structure of estimators in finite survey sampling has been studied to a certain level of detail in a number of papers, and show that the models offered by Bayesian networks in official statistics are an extremely promising tool.

This paper is organised as follows. At first (Section 2) Bayesian networks are defined and some theoretical aspects are highlighted. Note that this paper does not aim at aim at giving a complete and mathematically exhaustive explanation of Bayesian networks: just those elements that will be of interest in the applications to official statistics are described at a certain level of detail, leaving the rest to the relevant literature. This section is based on many references (mainly Cowell et al, 1999, and Neapolitan, 2004; but also Charniak, 1991, and the web page on Bayesian networks managed by Kevin P Murphy: http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html). Sections 2.1, 2.2 and 2.3 are mainly based on Neapolitan (2004). Bayesian networks applications in official statistics (Section 3) include the treatment of missing items (Section 3.1, based on the results in Di Zio *et al*, 2003, 2004a-c, 2005) and the use in sampling from finite populations (Section 3.2 based on the results in Ballin et al. 2005a-e). At the end of each of these two last sections, the role of Bayesian networks and the advantages in their use are highlighted in separate comments. Section 3.3 describes some other Bayesian networks applications. Finally, possible future developments are discussed in Section 4.

2. Bayesian Networks

Usually dependence relationship between variables are modelled with specific functions of their parameters, as in the generalised linear models or in the loglinear models. Bayesian networks are different. They are a class of models based on 2 elements:

i. the presence or absence of (any kind of) probabilistic relationship between the variables, and

ii. the possibility to represent these probabilistic relationships graphically in such a way that it is possible to associate the joint probability distribution to the graphical representation in a non-ambiguous way.

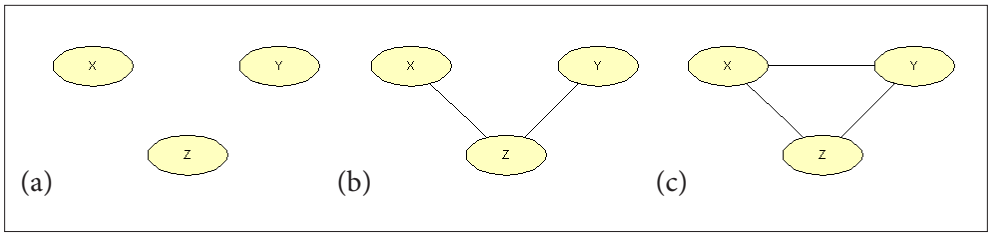
The first requirement makes this class quite general, and not dependent on specific functional definition of the dependence relationship of the variables. The second one is restrictive (for instance just some, but not all, of the loglinear models are Bayesian networks, see Warning 2 in Section 2.1).

In the following Bayesian networks are defined formally starting from the concept of Conditional Independence Graph (CIG) and Directed Acyclic Graph (DAG) as in Whittaker (1990). Finally the Bayesian network characteristics are shown with the help of some simplifying examples.

In general, a graphical representation of a multivariate variable $(X_1,...,X_K)$ is composed of a set of nodes V, each node representing one of the K variables, and a set of edges connecting pairs of nodes, E.

Conditional Independence Graphs (CIG) - A CIG is a graphical representation of the multivariate variable V composed of the pair (V,E), such that the edges in the set E are undirected and a pair of nodes is not connected by an undirected edge if and only if the two nodes are independent given all the other variables. Examples of CIG for three variables are in Figure 2.

Figure 2: Three CIGs for three variables *X*, *Y*, and *Z*



CIG (a) represents the situation of independence of the three variables, CIG (b) that *X* and *Y* are independent given *Z*, and CIG (c) that no conditional independencies characterize the three variables.

As a matter of fact, a CIG illustrates important features of the variables in *V*, in particular their dependence relationship. However, the joint probability distribution of *V* cannot be represented graphically, hence it is not yet useful for operative purposes.

Directed Acyclic Graphs (DAG) – In order to be operative, DAGs are appropriate. A DAG is a pair (*V,E*) of nodes and edges. Differently from CIGs, a DAG uses directed edges, henceforth arrows, for connecting pairs of nodes. The following elements characterize a DAG:

1. if there is an arrow from X to *Y* or from *Y* to *X*, *X* and *Y* are called *adjacent*
2. if there is an arrow from *X* to Y, *X* is a *parent* of *Y* and *Y* is a *child* of *X*;
3. the set of arrows connecting two nodes *X* and *Y* is called a *path*;
4. if there is a path from X to *Y*, *X* is an *ancestor* of *Y* and *Y* is a *descendent* of *X*;
5. if there is not a path from *X* to *Y*, *Y* is a *nondescendent* of *X*

The DAG has not associated any particular probabilistic feature of the variables in *V*, yet. One possibility that allows the operative use of the graphical representation linking the DAG with the probabilistic features of the variables is offered by the so called Markov condition.

> *Markov condition*. Let *P* be the joint probability distribution of the random variables represented by the nodes in *V*, and let the

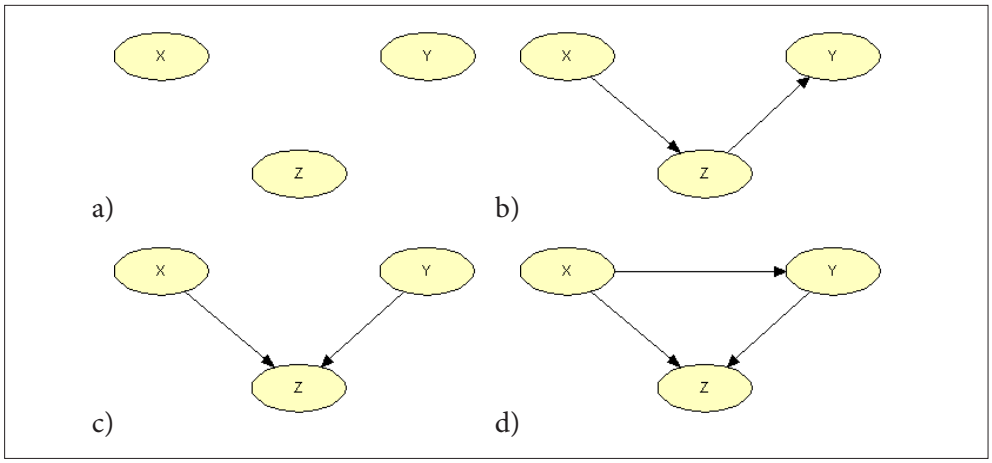
pair $G=(V,E)$ be a DAG associated to V. Then, the pair (P,G) satisfies the Markov condition if, for each variable X in V, X is independent of all its nondescendants given all its parents.

The pair (P,G) is a Bayesian network when it satisfies the Markov condition. Hence, the Bayesian network is an operative graphical representation of the joint probability distribution of the nodes in V. It is enough to associate each node X_j, $j=1,...K$, with the conditional distribution of X_j given its parents $pa(X_j)$ (when $pa(X_j)$ is the empty set, i.e. X_j is a root of the network, this conditional distribution is simply the marginal distribution of X_j). Then, the joint distribution function of the variables V is given by (chain rule):

$$P(X_1,...,X_k) = \prod_{j=1}^{K} P(X_j \mid \boldsymbol{p}\ (X_j)) \qquad (1)$$

Note that each multivariate variable can be factorised in the product of conditional distributions, but not all these decompositions correspond to a Bayesian network of the set of variables. As already said at the beginning of this section, a key issue is represented by the fact that the decomposition should be able to represent graphically the probabilistic relationship among the variables and describe it in a non-ambiguous way. Sometimes, this is not possible. For this reason, Bayesian networks are just a subclass of all the possible multivariate models: the Bayesian networks are the set of models for which it is possible to represent graphically the probabilistic relationship among the variables according to the Markov condition. Section 2.1 shows what this means in the case of three variables X, Y and Z.

Figure 3: Four possible Bayesian network structures for three variables



2.1 Meaning of different structures

Figure 3 shows some DAGs for three variables. According to the chain rule (1), these networks have the following interpretation (a thorough introduction on the concept of conditional independence is in Dawid, 1979).

Dag a) It has associated the following factorisation of the joint probability distribution: $P(X,Y,Z)=P(X)P(Y)P(Z)$. This case corresponds to the model of independence of the variables X, Y, Z.

Dag b) The joint probability distribution is $P(X,Y,Z)=P(X)P(Z|X)P(Y|Z)$. This is the case of conditional independence of X and Y given Z.

Dag c) The joint probability distribution is $P(X,Y,Z)=P(X)P(Y)P(Z|X,Y)$. This case corresponds to marginal independence of X and Y (just marginalize the joint probability with respect to Z) but conditional dependence of X and **Y** given Z. Note that it would not be possible to have at the same time X and Y marginal independent and conditional independent given Z in this network, unless (Z,X) is independent of Y or (Z,Y) is independent of X (as in the extreme case of DAG a) of complete independence; see also the following Warning 1).

Dag d) The joint probability distribution is factorised as $P(X,Y,Z)=P(X)P(Y|X)P(Z|X,Y)$. This is the *complete* model: all the dependencies between the variables are present. This model is also called *clique*.

When more than three variables are available, the possible dependence relationships are combination of the ones previously described. Two warnings are in order.

Warning 1 - As a matter of fact, the previous representations are not unique. In fact, for each CIG there can possibly be more than one Bayesian network, or better, given the same joint multivariate distribution P, more than one DAG. For instance, in Figure 2 conditional independence between X and Y given Z can be expressed uniquely by the CIG (b). On the contrary, different DAGs representing the situation of conditional independence of X and Y given Z can be defined via a suitable redirection of the arrows. These are shown in Figure 4. Their justification lies on the fact that, when X and Y are independent given Z, their joint probability distribution can be equivalently factorised as:

$$P(X,Y,Z)=P(X)P(Z|X)P(Y|Z)=P(Y)P(Z|Y)P(X|Z)=P(Z)P(X|Z)P(Y|Z).$$

Figure 4: Three equivalent Bayesian networks when *X* and *Y* are independent given *Z*

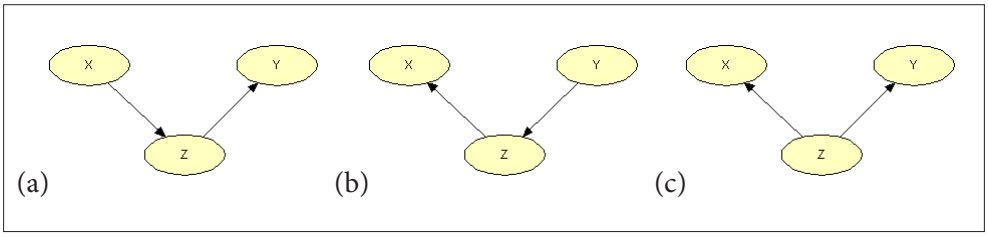


Figure 4 does not include the graph with the edges X -> Z and Y -> Z, i.e. Figure 3 c). In fact, this network has a complete different meaning. In order for Figure 3 c) to be consistent with the model represented by the equivalent networks of Figure 4, it is necessary to include an additional arrow linking X and Y. In other words, it is necessary to resort to a more complicated network than necessary (the clique, i.e. Figure 3 d). Hence, particular caution should be posed on the redirection of the arrows of a Bayesian network. The rules for arrows redirection and the definition of equivalent Bayesian networks are in Verma *et al* (1990).

Warning 2 - As already said, it is always possible to factorize a joint probability distribution, but it is not always possible to define a Bayesian network. An example is offered by loglinear models for categorical variables. It is easy to see that all the hierarchical loglinear models for three variables can be expressed as Bayesian networks but one: the one with the three way interaction set to zero. This loglinear model has a very peculiar aspect: the dependence relationship between the variables is not defined in terms of the joint probability distribution of all the variables, but by means of all the bivariate tables (distributions) of each couple of variables. In other words, it is true that each variable is connected with the others, although it is not the complete model (the saturated one). When factorizing the joint distribution of three variables X, Y, Z satisfying this model, the result is (no matter the order of the variables in the factorisation):

$$P(X,Y,Z)=P(X)P(Y|X)P(Z|X,Y).$$

Again, the result is the clique which is the appropriate factorisation for the saturated model (in other words, the factorisation of the joint distribution is the one associated to a more complicated dependence model). The appropriate representation of this loglinear model would actually involve three Bayesian networks, one for each minimal sufficient table for the model. Each model is a clique of respectively the pairs (X,Y), (X,Z), (Y,Z). As a matter of fact, there is not the possibility to describe this dependence relationship with a unique Bayesian network. Generally speaking, all those models that are defined via dependence relationship between subsets of variables in V and that cannot be expressed by an appropriate factorisation of the joint distribution function, do not have estimates of the parameters in closed form (e.g. the Iterative Proportional Fitting algorithm is used, calibrating successive estimates to the dependence relationship contained in each minimal sufficient table of the loglinear model). All these models are excluded by the set of models expressible as Bayesian networks.

Note that the previous problem does not apply to normal variables, i.e. multivariate normal variables can always be represented by Bayesian networks. This is due to the fact that multivariate normal variables are actually defined by the pairwise relationship of each couple of variables (subject to appropriate constraints on the variance matrix).

2.2 BN estimation

In the previous paragraph we have described a Bayesian network as a particular (graphical) model. Nothing "statistical" has been described. When just a sample of records where the variables V are observed is available, the Bayesian network should be estimated. There are many algorithms and methods for the estimation of a Bayesian network, some of them implemented in commercial or free software tools. A complete and updated reference is Neapolitan (2004). Here we review just the most important features on Bayesian network estimation.

The most important thing is that a Bayesian network is the pair (P,G), where P is the multivariate distribution of the variables V, and $G=(V,E)$ is the DAG. In this setting, only the set of nodes V is known in advance. The object of the inference is composed of two distinct elements:

1. the set of arrows E, or in other words the structure of a DAG
2. the conditional distribution of each node given its parents

In fact, the previous two elements define the Bayesian network and, by the chain rule (1), are able to define also the joint distribution of the variables V. It is worthwhile to mention three alternative approaches in estimating a Bayesian network.

The first one estimates at first the DAG structure, checking by appropriate independence and conditional independence tests whether undirected edges should be considered or not. Appropriate rules for the specification of the direction of the edges are defined in order to account for the relationship between variables (whether it is marginal or conditional independence). This estimation procedure of the structure is called *PC algorithm* (see Spirtes *et al*, 2000). Once the DAG structure is known, standard estimation methods (e.g. maximum likelihood estimation) can be applied in order to estimate the parameters of the conditional distribution of each node given its parents. This method is already implemented in commercial software tools, as Hugin (http://www.hugin.com). This approach is suitable when the data set is complete. Actually, some software tools allow to use this method also for incomplete data sets. In this last case, the PC algorithm is applied only on the subdata set of complete records, while the parameter estimation phase can be

performed on the overall data set. For instance, given the estimated structure, maximum likelihood estimation of the parameters can be performed with the EM algorithm.

The second approach is able to estimate with a unique procedure both the DAG structure and the parameters of the model given the structure via maximisation of the likelihood function (suitably penalised in order to avoid overspecification of the estimated model). This procedure has also been generalised to the case of partially observed data sets (Friedman, 1997). This approach is based on an extension of the *Expectation-Maximisation* (EM) algorithm for model selection problems that performs search for the best structure inside the EM procedure. Friedman proves the convergence properties of this algorithm, called *Model Selection EM*, and of one of its simplifications (in order to reduce the computational burden) *Alternating MS-EM*.

The third approach is just for incomplete data sets. It is a Bayesian approach developed by Sebastiani *et al* (2001a). This approach has the particular merit to highlight the different missingness mechanisms with the possibility to estimate the structure of a BN. Actually, the missingness mechanism can be considered as a set of additional dichotomous variables, showing whether each variables is actually observed or not. The multivariate structure of the variables of interest should take into account also their relationship with the indicators of missingness. This approach has not been implemented in any software tool, yet.

For a complete list of software codes and tools for using and estimating Bayesian networks and of their characteristics, see the webpage managed by Kevin P. Murphy (http://http.cs.berkeley.edu/~murphyk/Bayes/bnsoft.html) and the one of the gR project (graphical models in R: http://www.r-project.org/gR/).

2.3 Efficient use of the information in a BN

The Markov condition allows the identification of the relationship between a variable and its nondescendents. However it is still not clear the relation with all the other variables in V. The question is, given a variable X in V, which is the subsets of variables V' in V that makes X independent of all the

other variables in *V* given *V''*? *V'* is called the Markov blanket of *X*, henceforth *MB(X)*, and can be graphically determined in the Bayesian network structure via the following definition.

> *Markov blanket* – The Markov blanket *MB(X)* of a node *X* in *V* is composed by all the parents, children and parents of the children of *X*.

While it is evident the direct relationship of *X* with its parents and children, more attention should be given to its children's parents. The easiest example is offered by Network c) in Figure 3. In that case, *MB(X)* is composed by *Z* (its child) and *Y* (its child's parent). As already remarked, this network corresponds to considering marginal independence between *X* and *Y*, but conditional dependence of *X* and *Y* given *Z*. This last characteristic implies that *Y* should be included in *MB(X)* (*Z* alone is unable to make *X* independent of all the other variables given itself). Hence, in a multivariate setting the *MB(X)* is the subset of relevant variables for *X*: once *MB(X)* is known, all the other variables do not contain additional information on *X*.

3. Use in Official Statistics

Multivariate statistical models, as regression equations and loglinear models, are efficiently exploited in different official statistics problems: are Bayesian networks able to add something? The answer is yes, in many respects. First of all, Bayesian networks define models of interdependence between all the variables: variables relationship are easy to recognize. Secondly, this interdependence model allows a simplification of the joint distribution of the variables induced by the chain rule (1). Thirdly, each factor of the joint distribution can be easily estimated and used for operative purposes. Finally, when additional information is available (evidences, new distributions, additional records in the sample and so on) it can be easily used in order to update the joint distribution according to well established algorithms (see Cowell *et al*, 1999 and Cowell, 1998). All these elements suggest that some of the typical methodologies used up to now are just components of a larger family (see Ballin *et al*, 2005e for sampling and Di Zio *et al*, 2004a, for imputation). In the following a quick review of the use of Bayesian networks in official statistics is given. Note that most of the results have been obtained in the last 5 years. They should still be considered as research problems, and many issues have not yet been investigated. In the following, only categorical variables are studied. In fact, applications in this setting can be easily performed by means of the available software tools. The case of continuous variables still need to be further studied.

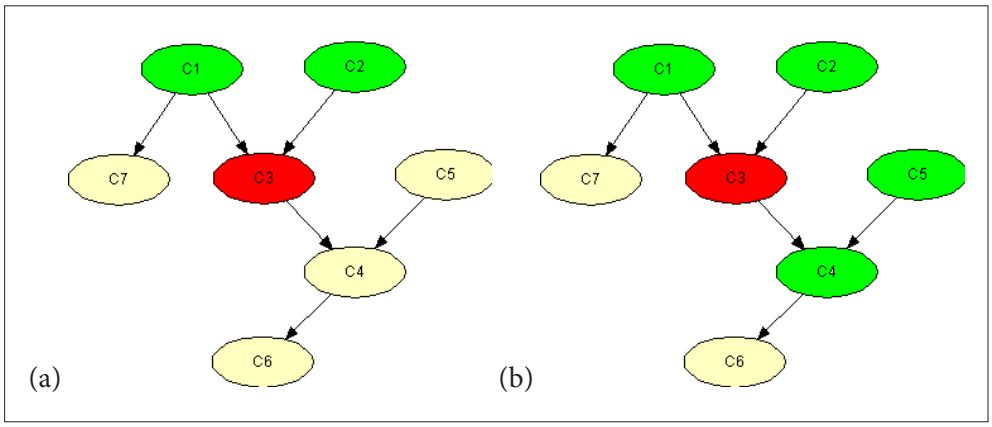
3.1 Imputation of missing items

This is maybe the most straightforward application of Bayesian networks, at least when missing data follow a Missing at Random mechanism (henceforth MAR; see Little *et al*, 1987, and references therein for a formal definition of MAR). Let $x=(x1_i,...,xk_i)$, $i=1,...n$, be a sample of n i.i.d. observations of the r.v. $X=(X_1,...,X_k)$, and assume that these records are just partially observed. Let $o(i)$ and $m(i)$ be subsets of $\{1,\ldots k\}$ such that $o(i) \grave{E}\ m(i) \cup \{1,\ldots,k\}$ and let $x_{o(i)}$ and $x_{m(i)}$ be respectively the observed and missing part of the record xi, i=1,…n. A usual practice for partially observed data set is imputation of missing values, i.e. generation of suitable values $\tilde{\mathbf{x}}_{m(i)}$ for the unobserved $x_{m(i)}$. Different imputation procedures have been defined. A "perfect" imputation

procedure would impute the missing part of the record with a random generation from the distribution of $X_{m(i)}$ given $X_{o(i)}$. This procedure can be considered as "perfect" because the new imputed data set would maintain the characteristic to be a random sample of n i.i.d. observations of X. Actually this procedure can be simplified in the sense that not all the conditional variables are necessary. A simplification that preserves the property to maintain the inferential characteristics of the imputed data set would consider a generation of imputed values from the distribution of $X_{m(i)}$ given $MB(X_{m(i)})$, where $MB(X_{m(i)})$ can possibly be a subset of $X_{o(i)}$. Hence, the identification of the Markov Blanket of $X_{m(i)}$ greatly simplifies the imputation procedure reducing the sets of conditionals and adapting the set of conditionals to the pattern of missing data in the record. For this reason, Bayesian networks are a useful tool for identifying which of the observed variables are necessary for imputation.

A preliminary formalisation of the use of the Bayesian network representation of the dependence relationship of the variables for imputation is in Thibaudeau *et al* (2002). In their paper, given a DAG structure, each missing variable is imputed drawing a value at random from its probability distribution given its parents. The imputation procedure starts from those nodes without parents. When all the missing items in these variables have been filled in, all the remaining variables whose parents are within the already imputed variables are imputed. When also these variables have been imputed, all the remaining variables whose parents are among those already imputed are imputed and so on.

Figure 5: Use of the dependence structure suggested by a Bayesian network. The alternative use of just the parents and Markov blanket of C3 is highlighted respectively in (a) and (b)



Their approach has been studied and generalised in some papers (Coppola *et al*, 2002a, 2002b, and Di Zio *et al* 2004a). In particular Di Zio et al (2004a) explains how logical constraints in terms of structural zeros can be easily considered in this setting. In fact, rules of compatibility between the observations on a unit can be defined as a fundamental aspect of the multivariate model for X. The possibility to specify Bayesian networks subject to logical rules, as the structural zeros, is a powerful approach that can be easily implemented during the Bayesian network estimation procedure. However, this approach actually does not exploit all the information in the data set: imputation of a missing variable is performed only by means of its parents, given an ordering among the variables (e.g. C_3 in Figure 5 (a) is imputed drawing randomly a value for its distribution given C_1 and C_2). Other papers (Di Zio *et al*, 2003, 2004b-c) have defined algorithms for the imputation of missing items with respect to the corresponding Markov blanket (e.g. C_3 is imputed conditioning on C_1, C_2, C_4 and C_5, see Figure 5 (b)). Manipulation of the Bayesian network in order to perform this operation is part of a software code in C++, described in Di Zio *et al* (2005).

An extension to the case of missing items in longitudinal surveys is in Righi (2005).

Comment: Bayesian networks appear as a device for exploiting most of the statistical information contained in the observed data set. Although the

use of random generation of imputations via conditional distributions is not new, Bayesian networks are a novel practice as far as the definition of the conditional variables is concerned. In a sense, the use of the Markov blanket of the unobserved variables makes the set of conditionals *adaptive* with respect to the pattern of missing values in each record. Adaptation is justified by the statistical relationship of the overall multivariate distribution. The variables not used as conditionals are independent of the missing variables given the conditional ones.

As a matter of fact, the multivariate distribution and the DAG structure should be estimated. The use of maximum likelihood estimators is particularly appropriate in this setting for their consistency. When the data set is large, the maximum likelihood estimate of the joint distribution function should be reasonably "near" to the true but unknown one. Hence, the imputed data set can be considered as "almost" generated by the true, and unknown, joint distribution function. Up to now, imputation by Bayesian networks has always been performed via estimation of the Bayesian network structure by the PC algorithm and, given the estimated structure, the conditional probability distributions are estimated via maximum likelihood. Other approaches in the estimation of Bayesian network structures for imputation are under study.

3.2 Estimation with completely observed samples drawn according to complex survey schemes

Also sampling methods from finite populations benefit of the multivariate relationship among the variables of interest (e.g. regression estimators). In general, special attention should be given to the sampling design. In fact, as stated in every modern textbook on sampling theory (e.g. Chambers *et al*, 2003), the sampling design is itself a variable and plays a very important role in the estimation process. Let $X_1,\ldots X_k$ be k variables of interest on a finite population of N units. Let a sample of n units be drawn from the population according to a complex survey scheme, with sample weights (defined by the pair design/estimator) w_i, i=1,…n. One of the most used estimators of the joint distribution function of the k variables is the ratio estimator:

$$\hat{F}(x_1,...,x_k) = \sum_{i=1}^{n} I_{x_1 \ldots x_k}(x_{1i},...,x_k) \frac{w_i}{\sum_{i=1}^{n} w_i} \qquad (2)$$

where $I(\ )$ is the indicator function, and $x_{1i},...x_{ki}$, i=1,…n, are the n observed records in the sample. The previous estimator can equivalently be rewritten via a Bayesian network model (preliminary results were obtained in Ballin *et al*, 2005a; advances are written in Ballin *et al* 2005b,c,d; and further extensions are in Ballin *et al*, 2005e). This new formalisation of estimator (2) is obtained via a new variable, S. This is the "design variable", with as many categories as the different inclusion probabilities, say $w_{(1)},\ldots,w_{(H)}$, and with marginal probability given by the fraction of the total weight of the units with the same sample weight:

$$P(S=h)=\frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}}$$

where n_h is the number of units with equal first inclusion probability $w_{(h)}$, h=1,…H. Given that S contains all the information on the sample design, conditioning on this variable produces estimators that are sample weights free. For instance, denoting with s_h the set of labels of the n_h units with weight $w_{(h)}$:

$$P(X_1=x_1 \mid S=h)=\frac{\sum_{i\in s_h} I_{x_1}(x_{1i})w_{(h)}}{n_h w_{(h)}}=\frac{\sum_{i\in s_h} I_{x_1}(x_{1i})}{n_h}$$

$$P(X_1=x_1 \mid X_2=x_2, S=h)=\frac{\sum_{i\in s_h} I_{x_1x_2}(x_{1i}x_{2i})w_{(h)}}{\sum_{i\in s_h} I_{x_2}(x_{2i})w_{(h)}}=\frac{\sum_{i\in s_h} I_{x_1x_2}(x_{1i}x_{2i})}{\sum_{i\in s_h} I_{x_2}(x_{2i})}$$

These definitions allow to rewrite (2) as the following:

$\hat{F}(x_1,...,x_k)=$

$$= \sum_{h=1}^{H} \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \sum_{i \in s_h} \frac{I_{x_1}(x_{1i})}{n_h} \sum_{i \in s_h} \frac{I_{x_1 x_2}(x_{1i} x_{2i})}{\sum_{i \in s_h} I_{x_1}(x_{1i})} \ldots \sum_{i \in s_h} \frac{I_{x_1 x_2 \ldots x_k}(x_{1i} x_{2i} \ldots x_k)}{\sum_{i \in s_h} I_{x_1 x_2 \ldots x_{k-1}}(x_{1i} x_{2i} \ldots x_{k-1i})}$$

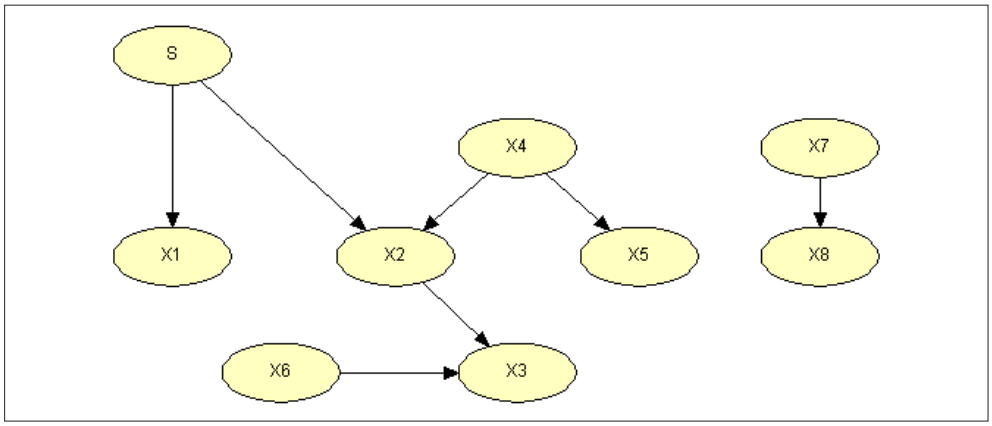
$$= \sum_{h=1}^{H} P(S=h) P(X_1 = x_1 \mid S=h) \prod_{j=2}^{k} P(X_j = x_j \mid X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, S=h) .$$

As a matter of fact, the usual ratio estimator $\hat{F}(x_1,...,x_k)$ implicitly assumes a particular model: the complete dependence model among ($S,X_1,\ldots,X_k$). In the Bayesian network terminology, the implicit model is the clique. If the dependency model for ($S,X_1,\ldots,X_k$) is simpler, the estimator (2) may result inefficient. An example taken from Ballin *et al* (2005e) is represented in Figure 6.

In order to define estimators that fulfil the dependence relationship between the variables and, at the same time, always use the sample weights, four different type of nodes have been defined.

- Type (a) nodes: these nodes admit S as a parent. In Figure 6, nodes X_1 and X_2 are type (a) nodes.
- Type (b) nodes: these nodes have at least a type (a) ancestor but S is not one of their parents. Node X_3 in Figure 6 is a type (b) node.
- Type (c) nodes: these are those nondescendants of type (a) and/or (b) nodes that do not admit S as a parent but that are (indirectly) linked to S. Figure 6 has two distinct groups of type (c) nodes: the first one is composed by the pair (X_4,X_5); the second one by X_6.
- Type (d) nodes: these are the nodes disconnected with S. In Figure 6, the couple (X_7,X_8) is a group of type (d) nodes.

Figure 6: Example of Bayesian networks for finite populations



The estimator of the joint distribution function will be of the following form:

$$\hat{F}(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) =$$

$$= \hat{F}(X_4, X_5)\hat{F}(X_6)\hat{F}(X_1, X_2 | X_4)\hat{F}(X_3 | X_2, X_6)\hat{F}(X_7, X_8)$$

where each component is estimated marginalizing their joint distribution with S with respect to S:

type (c)
$$\hat{F}(X_4, X_5) = \sum_{h=1}^{H} P(S = h)P(X_4, X_5 | S = h) \quad ;$$

$$\hat{F}(X_6) = \sum_{h=1}^{H} P(S = h)P(X_6 | S = h)$$

type (a)
$$\hat{F}(X_1, X_2 | X_4) = \sum_{h=1}^{H} P(S = h)P(X_1, X_2 | X_4, S = h)$$

type (b)
$$\hat{F}(X_3 | X_2 X_6) = \sum_{h=1}^{H} P(S = h)P(X_3 | X_2, X_6, S = h)$$

$$\hat{F}(X_7, X_8) \quad \sum_{1} (\quad) (\quad_7, \quad_8 | \quad)$$

type (d)

Note that type (b), (c) and (d) nodes may admit more than one subgroup (the two type (c) subgroups in Figure 6 are just an example). As shown in the previous example, each of these subgroups should be estimated distinctly. In general, if there are T, V and W distinct type (b), (c) and (d) nodes, with labels in the sets B_t, t=1,…T, C_v, v=1,…,V, D_w, w=1,…W, the general form of the Bayesian network (BN) based estimator is (Ballin $_{\text{et al}}$, 2005e):

$$\hat{F}(X_1,...,X_k) = \left[\prod_{v=1}^{V} \hat{F}(\mathbf{X}_{C_v})\right] \hat{F}\left(\mathbf{X}_A \middle| X_{C_v}, v = 1,...V\right) \left[\prod_{t=1}^{T} \hat{F}\left(\mathbf{X}_B \middle| \mathbf{X}_A, \mathbf{X}_{C_v}, v = 1,...V\right)\right]\left[\prod_{w=1}^{W} \hat{F}\left(\mathbf{X}_{D_w}\right)\right]$$

A Monte Carlo experiment in Ballin et al 2005(c) shows that the BN based estimators can be much more efficient than the usual ratio estimators. The key idea is that the use of estimators linear in the weights introduce implicitly dependence induced by marginalisation with respect to S. For this reason, each type of node and each subgroup should be estimated distinctly with respect to S. As a result, if the interest is just on a few marginal tables instead of the complete joint distribution of the variables of interest, this approach gives results which are internally consistent (see Ballin *et al*, 2005d), i.e. if two tables contain the same variable, its marginal distribution is always the same. Ballin *et al* (2005e, Proposition 1) define a list of necessary and sufficient conditions that ensure that the dependence model of the set of variables is respected (and hence the disseminated tables are consistent). Finally, Ballin *et al* (2005b) and (2005e) show that the usual calibration estimators (that in case of categorical variables are poststratification estimators) can be equivalently defined as updating procedures in a BN, and this ensures the possibility to enlarge the set of possible poststratification procedures.

Comment: As a matter of fact, it seems that survey weights may have an unpleasant effect on the usual estimators computed as linear functions of the weights: the introduction of dependencies that actually do not hold true. The introduction of a wrong dependence relationship makes the estimator

structure more complex, and consequently less efficient. BN based estimators are non linear in the weights but still make use of the weights, without the unpleasant introduction of spurious dependencies.

All the previous results are obtained given the BN structure. Estimation of a structure of a BN in a finite population setting is still an unsolved problem. The possible translation of the PC algorithm through changes of the test statistics in order to take into account the complexity of the survey design is discussed in Ballin *et al* (2005a).

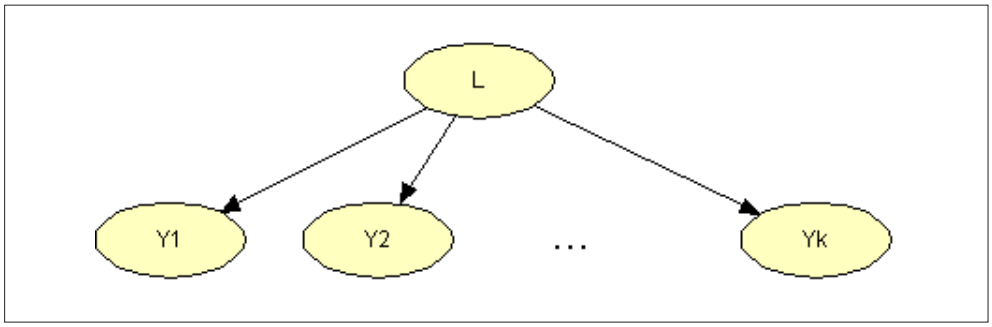
3.3 Other applications of Bayesian networks and possible extensions

Statistical Information Systems - One of the first Bayesian networks applications in official statistics is in Getoor et al. (2001a). They show how a very complex data base, as the one of the 1990 U.S. Census, can be easily and efficiently represented by a Bayesian network. The Bayesian network shows which contingency tables are necessary in order to describe the overall statistical information in the census, reducing the figures to store. They also show what numerical computations are necessary for any queries, i.e. how to join information from the different tables suggested by the Bayesian network. The example that the authors consider is relative to a data set which refers to just one kind of statistical unit. In general, the available tables may refer to different kinds of units: for instance some variables may refer to individuals, other to families, other to geographical (regions, counties,…) or institutional (hospitals, schools,…) entities. Getoor *et al* (2001b) show how to extend the concept of Bayesian network to the case of data referring to multiple kinds of units: they call this tool *Probabilistic Relational Model* (PRM). This tool is based on a knowledge representation language describe in Koller *et al* (1997). It seems particularly suitable for designing statistical information systems.

Record Linkage – When it is necessary to match records belonging to the same statistical unit in two data sets, but the record identifiers in the two data sets are subject to error, record linkage procedures are used (ISTAT, 2003, and references therein). Winkler (2002) shows which DAG structure is implicitly used for the naïve record linkage procedure. Assuming that the status of matched and unmatched pairs of records is represented by a (latent) variable

L, the DAG structure for the naïve record linkage procedure is represented in Figure 7. As a matter of fact, it corresponds to the so called conditional independence assumption.

Figure 7: Bayesian network for the naïve record linkage procedure, where *L* is the latent status of pair, and $Y_1,\ldots Y_k$ are the comparison of the two records in the pair with respect to the *k* matching variables



It is well discussed how the naïve record linkage procedure can lead to misleading results. It is important to investigate other approaches. For instance, Friedman (1997) shows how to estimate Bayesian networks in presence of latent variables. This approach can suggest alternative multivariate models able to link appropriately the record pairs.

Time series – Penny *et al* (2004) describe by means of BNs the multivariate dependence structures of time series. They apply this description to the quarterly gross national expenditure in New Zealand. Their objective is to identify which components of the gross national expenditure deserve to be improved in terms of timeliness.

4. Further developments

First of all, the applications discussed in Section 3 still need to be further explored and compared to the "traditional" ones. Nevertheless, it seems that Bayesian networks can be useful in many other different topics. Two of them appear particularly promising.

1. Integration of surveys – Following Ballin *et al* (2001), the different surveys can be designed as a *junction tree* (i.e. the tool used for the propagation of information in a Bayesian network, see Cowell, 1998, and Jensen, 1996). This network can perform as a tool for jointly analyzing variables only when strict model assumptions hold (this case corresponds to the statistical matching problem, see D'Orazio *et al*, 2005). Nevertheless it seems to be a formidable tool for updating survey results according to new information from archives or new surveys. In this case, it is necessary to understand the interaction between BN based estimators (Section 3.2) and calibration, poststratification, ratio raking (Harora *et al,* 1977a-b), and repeated weighting (Houbiers, 2003) estimators
2. Editing – The possibility to include logical rules in the estimation of the joint distribution of multiple variables, as well as to include the definition of "rare" events to be further investigated, suggest that editing procedures can be appropriately defined via Bayesian networks.

Acknowledgments

This is hopefully a comprehensive review article on the use of Bayesian networks in official statistics. I am indebted with all those who had the patience to introduce me to the different topics touched in this paper and with those I had the chance and fortune to discuss and work with: in alphabetical order Marco Ballin, Marco Di Zio, Orietta Luzi, Julia Mortera, Paola Vicard.

References

Ballin, M., M. Scanu, and P. Vicard. 2005a. "Bayesian Networks for finite populations". In *Atti del Convegno Metodi di Indagine e di Analisi per le Politiche Agricole (MIAPA)*: 95-106. Pisa, Italy, 21-22 October 2004.

Ballin, M., M. Scanu, and P. Vicard. 2005b. "Information propagation in finite survey sampling: Bayesian networks and poststratification". In Liseo, B., G.E. Montanari, e N. Torelli (*a cura di*). *Metodi Statistici per l'Integrazione di Dati da Fonti Diverse*. Milano: Franco Angeli.

Ballin, M., M. Scanu, e P. Vicard. 2005c. "Reti bayesiane per la costruzione di stimatori in popolazioni finite". In *Atti del Convegno Agristat, Verso un Nuovo Sistema di Statistiche Agricole*. Firenze 30-31 Maggio 2005.

Ballin, M., M. Scanu, and P. Vicard. 2005d. "Coherence of sample estimates for finite populations: some results based on Bayesian networks". In *Atti Convegno S.Co2005, Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*. Bressanone, Italy, 15-17 September 2005.

Ballin, M., M. Scanu, and P. Vicard. 2005e. "Bayesian networks and complex survey sampling from finite populations". In *FCSM Conference*. Arlington, Virginia, 14-17 November 2005.

Ballin, M., and P. Vicard. 2001. "A proposal for the use of graphical representation in official statistics". In *Atti del Convegno S.Co2001, Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*. Bressanone, Italy, 24-26 September 2001.

Chambers, R.L., and C.J. Skinner (*eds.*). 2003. *Analysis of Survey Data*. Chichester, UK: John Wiley & Sons.

Charniak, E. 1991. "Bayesian networks without tears". *AI Magazine*: 50-63.

Coppola, L., M. Di Zio, O. Luzi, A. Ponti, and M. Scanu. 2002. "Bayesian networks for imputation in official statistics: A case study". In *Proceedings of the Data Clean Conference*. Jyvaskyla, Finland, 29-31 May 2002.

Coppola L., M. Di Zio, O. Luzi, A. Ponti, and M. Scanu. 2002b. "On the use of Bayesian networks in official statistics". In *Atti della XLI Riunione Scientifica della Società Italiana di Statistica:* 237-240. Milano, Italy, 5-7 June 2002.

Cowell, R.G. 1998. "Introduction to inference for Bayesian networks". In Jordan, M.I. (*eds.*). *Learning in Graphical Models*. NATO ASI Series (Series D: Behavioural and Social Sciences), Volume 89. Dordrecht, The Netherlands: Springer.

Cowell, R.G., A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. 1999. *Probabilistic Networks and Expert Systems*. New York, NY, U.S.: Springer-Verlag.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching: Theory and Practice*. Chichester, UK: John Wiley & Sons.

Dawid, A.P. 1979. "Conditional independence in statistical theory". *Journal of the Royal Statistical Society*, Series B (*Methodological*), Volume 41, N. 1: 1–31.

Di Zio, M., M. Scanu, and P. Vicard. 2003. "Open problems and new perspectives for imputation using Bayesian Networks". In *Atti del Convegno* S.Co2003, *Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*: 170-175. Treviso, Italy, 4-6 September 2003.

Di Zio, M., M. Scanu, L. Coppola, O. Luzi, and A. Ponti. 2004a. "Bayesian networks for imputation". *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Volume 167, Issue 2: 309-322.

Di Zio, M., G. Sacco, M. Scanu, and P. Vicard. 2004b. "Some approaches in imputing missing items with Bayesian networks". In *Atti della XLII Riunione Scientifica della Società Italiana di Statistica*. Bari, Italy, 9-11 June 2004.

Di Zio, M., G. Sacco, M. Scanu, and P. Vicard. 2004c. "Multivariate techniques for imputation based on Bayesian networks". In Antoch, J. (*ed.*). *Proceedings Compstat 2004*: 928-934. 16th Symposium of IASC, Prague, 23-27 August 2004. Heidelberg, Germany: Physica-Verlag, Springer.

Di Zio, M., G. Sacco, M. Scanu, and P. Vicard. 2005. "Methodology and software for imputation of missing values by Bayesian networks". In Liseo, B., G.E. Montanari, e N. Torelli (*a cura di*). *Metodi Statistici per l'Integrazione di Dati da Fonti Diverse*. Milano: Franco Angeli.

Friedman, N. 1997. "Learning belief networks in the presence of missing values and hidden variables". In *Fourteenth International Conference on Machine Learning (ICML97)*.

Getoor, L., B. Taskar, and D. Koller. 2001a. "Selectivity estimation using probabilistic models". In *ACM-Sigmod*. Santa Barbara, CA, U.S., 21-24 May 2001.

Getoor, L., N. Friedman, D. Koller, and A. Pfeffer. 2001b. "Learning probabilistic relational models". In Dzeroski, S., and N. Lavrac (*eds.*). *Relational Data Mining*. Heidelberg, Germany: Springer-Verlag.

Harora, H.R., and G.J. Brackstone. 1977a. "An investigation of the properties of raking ratio estimators I: with simple random sampling". *Survey Methodology*, Volume 3: 62-83.

Harora, H.R., and G.J. Brackstone. 1977b. "An investigation of the properties of raking ratio estimators II: with cluster sampling". Survey Methodology, Volume 3: 232-243.

Helm, L. 1996. "Improbable inspiration". *Los Angeles Times*, October 28th, 1996.

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen, and V. Snijders. 2003. "Estimating consistent table sets: position paper on repeated weighting". *Discussion paper* N. 03005, CBS - Statistics Netherlands.

Istituto Nazionale di Statistica – Istat. 2003. "Metodi Statistici per il Record Linkage". *Collana Metodi e Norme*, N. 16. Roma: Istat.

Jensen, F.V. 1996. *An introduction to Bayesian Networks*. New York, NY, U.S.; Springer-Verlag.

Kadie, C.M., D. Hovel, and E. Horvitz. 2001. "MSBNx: A component-centric toolkit for modeling and inference with Bayesian networks". *Microsoft Research Technical Report*, MSR-TR-2001-67, July 2001.

Koller, D, A. Levy, and A. Pfeffer. 1997. "P-classic: a tractable probabilistic description logic". In *Proceedings of the Fourteenth Conference on Artificial Intelligence* (AIII-97): 390-397. Providence, Rhode Island, August 1997.

Little, R.J.A., and D.B. Rubin. 1987. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Chichester, UK: John Wiley & Sons.

Neapolitan, R.E. 2004. *Learning Bayesian Networks*. Upper Saddle River, NJ, U.S.: Prentice Hall.

Penny, R.N., and M. Reale. 2004. "Using graphical modelling in official statistics". *Quaderni di Statistica*, N. 6: 31-47.

Righi, P. 2005. "Trattamento delle mancate risposte nelle indagini longitudinali mediante modelli grafici ricorsivi". *Tesi di Dottorato in Metodi Statistici per l'Economia e l'Impresa*. Roma, Italia, Università degli Studi Roma Tre, XVI ciclo.

Sebastiani, P., and M. Ramoni. 2001a. "Bayesian Selection of Decomposable Models With Incomplete Data". *Journal of the American Statistical Association*, Volume 96, N. 456: 1375-1386.

Sebastiani, P., and M. Ramoni. 2001b. "On the use of Bayesian networks to analyse survey data". *Research in Official Statistics - ROS*, Volume 4, N. 1: 52-64.

Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, Prediction and Search. Second edition*. Cambridge, MA, U.S.: MITCogNet, MIT Press.

The Economist. 2001. "Son of paperclip". *The Economist, print edition*, March 24th 2001.

Thibaudeau, Y., and W.E. Winkler. 2002. "Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints". Technical report, RRS2002/9. U.S. Bureau of the Census.

Verma, T.S., and J. Pearl. 1990. "Equivalence and synthesis of causal models". In *Proceedings of the Sixth Conference on Uncertainty in AI*: 220-227. Cambridge, MA, U.S., 27-29 July 1990.

Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics*. Chichester, UK: John Wiley & Sons.

Winkler, W.E. 2002. "Methods for record linkage and Bayesian network". *Research Report*, Series n. 2002/5. U.S. Bureau of the Census.

Coworking: Evolution, Drivers and Spreading. A review for orienting suitable indicators for official statistics

Alessandra Fasano, Giulia Nisi and Ludovica Rossotti [1]

Abstract

This article provides an analysis of coworking as a new system of work organisation. Using a literature review, this article investigates the drivers that have led to the creation and development of this new work method in shared spaces. To this aim, the authors describe different workforce generations, their attitudes and behaviour in terms of work organisation. The study offers an overview of the current worldwide spreading of coworking with a specific focus on the Italian scenario.

Keywords: Coworking, Workspace, Work organisation, Community, Technological progress.

1 Alessandra Fasano (alessandra.fasano@unisalento.it); Giulia Nisi (giulia.nisi@icloud.com); Ludovica Rossotti (ludovica.rossotti@uniroma1.it).

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

1. Introduction

Coworking spaces appeared for the first time in 2005 in the United States and, since then, they have been continuously increasing, in both numbers and size. In 2017, there were approximately 13,800 coworking spaces worldwide (approximately 600 in Italy). These workplaces not only offer users shared workstations and services, but also serve as facilitators for networks and relationships, which are essential to address the current job market.

In particular, coworking is characterised by some specific aspects, among them: community, openness, accessibility and self-sustenance.

To understand better this organisational mode, it is necessary to consider the drivers that have played a key-role in the technologic, social and economic scenarios. Among them: the Industrial Revolution (from Industry 1.0 to Industry 4.0) and the features of different workforce generations. Additionally, it is worth noting that other issues, related to the technological progress and to the constant requirement of reducing business expenses, have encouraged new working modalities, such as 'teleworking', 'hot desking' and 'smart working'.

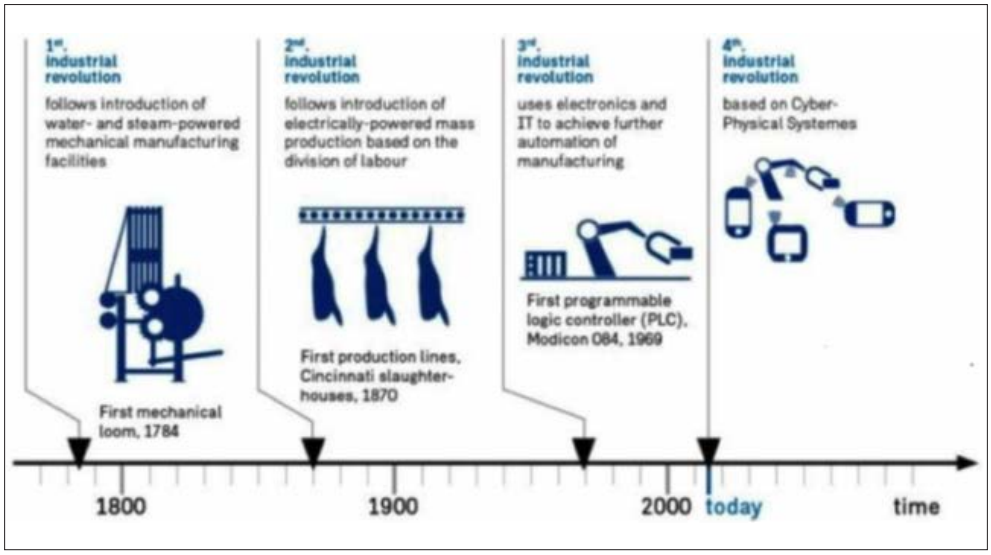
All these issues have led to an increasing demand of workspaces that could be more adequate for new work concepts. The following sections provide some significant examples of these workspaces; in particular, the spreading of coworking is considered both from a geographic and temporal point of view.

2. The drivers of change

In the last decades the world economy has been affected by significant changes, moving from a manufacturing economy to a digital economy mainly based on digital technologies (Figure 2.1) (Swann, 2017).

The First Industrial Revolution was characterised by the transition from hand production methods to machines, new chemical manufacturing and iron production processes, increasing use of steam power, development of machine tools and rise of factory system (Deane, 1971).

Figure 2.1 - From Industry 1.0 to Industry 4.0



Source: www.dfki.de

The Second Industrial Revolution used electric power to create mass production. This contributed to generate a wide range of employment opportunities for non-skilled workers, who became consumers thanks to low-cost products available on the market (Accornero, 1994; Mingione and Pugliese, 2010).

The Digital Age Revolution is the new productivity platform regarded, by the experts, as the Third Industrial Revolution (Murty, 2017); it uses electronics and information technology to automate production, resulting in a change of the traditional work process and in an outgrowing of the old organisational logics (Bonazzi, 2008; Catino, 2012). The Information revolution has caused a change of the labour market, resulting in a consequent decline of the employee number and an increase of skilled workers able to handle complex machinery. This trend occurred firstly in the Sixties, when computers began to be used for commercial purposes and, subsequently, in the Nineties when the 'World Wide Web' use spread rapidly (Berger *et al.*, 2014). Digital logic circuits and their derived technologies (including computer, digital cellular phone and the Internet) are crucial to this revolution.

The Digital Age Revolution caused (and is still causing) upheavals which are much deeper than those caused in the past by the technological revolutions (Frey and Osborn, 2015). The current evolution of this trend is leading to a Fourth Industrial Revolution powered by the Internet and Big Data with the ongoing development of cyber physical systems and smart factories (Schwab, 2017).

At the same time, the different Industrial Revolutions and the increasingly Globalisation have changed the skills of workers who have become more familiar with cognitive tasks and problem solving.

Workers have adapted their skills to new market demands associated to the introduction of new technologies (Ross, 2017).

As a matter of fact, the technological progress has caused two opposite effects in terms of employment: *i)* a decrease of workers as a direct consequence of the product process automation, *ii)* the creation of new professional skills and of production methods which require a high level of work flexibility (Frey and Osborn, 2015).

Furthermore, the technological progress has transformed both working and educational aspects, with a deep impact on the different generations and on their different approach to the labour market (Woolf, 2010). Indeed, Goldin and Katz (2007) defined the twentieth century history as 'the race between education and technology' with some differences among generations.

A 'generation' can be defined as the set of people born in a given period, living and growing up at the same historical moment and having a specific way of thinking, communicating and acting.

This also means that each generation has its own concept of work-life and a different approach to professional life.

Currently the workforce is made up of three categories (Figure 2.2): 'Baby Boomers', 'Generation X' and 'Generation Y'.

Figure 2.2 - The different composition of the workforce: comparison among the three generations

Topic	Boomers	Gen X	Gen Y
Work style	Get it done – no matter what it takes	Find fastest route to results; protocol secondary	Work to deadlines not schedules
Authority/ leadership	Respect for power and accomplishments	Rules are flexible; collaboration is important	Value autonomy; less inclined to pursue formal leadership positions
Communication	Formal and through structured network	Casual and direct; sometimes skeptical	Casual and direct; eager to please

Source: Chester, 2002, our elaborations

The 'Baby Boomers' were born between 1946 and 1964 (a period characterised by a significant increase of population, social security and economic prosperity). Typically, they have a permanent job and a low attitude

to technological tools. Their social identity is strictly associated to a specific work.

The 'Generation X', also known as 'Gen X' or 'Post Boomers', includes people born between 1965 and 1980. This generation typically lives in a context characterised by globalisation, work flexibility, mobility. Its social identity is associated not only to the work itself, but also to the satisfaction deriving from personal job and from private life.

The 'Generation Y', also known as 'Millennials' or 'Generation Next', includes people born between 1981 and 2000 (Cole *et al.*, 2002; Spiro, 2006). They typically live in a context strongly characterised by high technologic development, globalisation, mobility, job sharing and a widespread work flexibility (Howe, 2000). Their usual working day is characterised by an overlapping of work and life activities and by a low need of a permanent workplace.

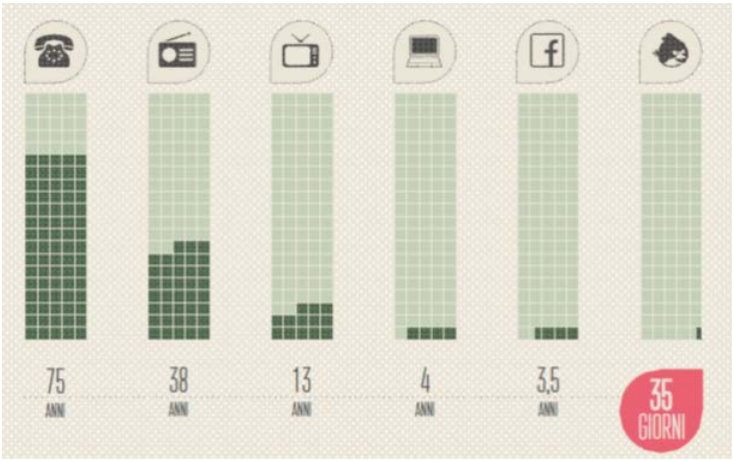
3. Some implications of technological progress

New inventions are spreading now much faster than in the past. Historically, technologies are adopted by a country during an average period of about 45 years and they usually are spread globally in an average period of 119 years.

The Internet has revolutionised this scenario: only 7 years are sufficient to reach every part of the world (Frey and Osborn, 2015).

The technology spreading time differs from country to country and it has been reducing thanks to the Internet (Figure 3.1). It is worth noting that digital technologies are able to put in connection people and ideas very quickly causing a substantial change in lifestyle and way of working.

Figure 3.1 - Technology spreading times to reach the target of 50 million users



Source: Frey and Osborn, 2015

Another important aspect is the number of tools necessary to work: these have been reducing thanks to the technological progress.

Since 1980, the typical desk setup has changed, because of the development of technology (Harvard Innovation Lab, 2015). In the past, a worker needed a desk full of several different tools (i.e.: stationery, fax, dictionaries, bulletin boards and calculator). Over the years, these tools became unnecessary. As matter of fact, icons on a computer desktop have substituted physical objects.

Nowadays a knowledge worker only needs a laptop, a smartphone and a pair of glasses on his desk (Figure 3.2).

Figure 3.2 - Evolution of the desk in the last 35 years

Source: Harvard Innovation Lab, 2015

Tools have been just limited to a computer, thus implying a no need of an office and a desk. In this way, workers can carry out tasks similarly anywhere and anytime.

The Internet is not the only factor that has led to a new attitude to work life and workplace.

The reasons that motivated this trend can be found in the following points: increasing amount of freelancers; work flexibility implying a high mobility requirement as well as a temporary workplace; faster and cheaper travelling than before; availability of 'cloud computing' access anywhere through a simple Internet connection to a smartphone or laptop (Levels, 2015).

4. New systems of work organisation

The new opportunities offered by technology and by the increase of flexibility generated new systems of work organisation, such as 'teleworking' and 'hot desking'. Thanks to these alternative work strategies, companies are now able to reduce costs related, for example, to the management of spaces.

> *"Telework is defined as a form of organising and/or performing work, using information technology, where work, which could also be performed at the employer's premises, is carried out away from those premises on a regular basis. The agreement concerns teleworkers with an employment contract and does not deal with self-employed telework"* (ETUC *et. al.*, 2006).

> *"Hot desking is an office organisation system, which involves multiple workers using a single physical workstation or surface during different time periods"* (Dubey, 2009).

The increasing need of a different way of working, not necessarily limited to a specific workplace, has led to the concept of 'smart working', intended as a flexible and fully autonomous working mode which is assessed not in terms of working time but through the obtained results. Moreover, there is neither a workplace nor work-related constraints.

The advantages related to these alternative forms of work have contributed to the spreading of coworking.

5. The spreading of coworking

The digital transformation has changed the economy and the technological innovation system. The actual economic context is characterised by a digital economy, based on digital computing technologies (Tapscott, 1995).

The digital economy permeates all aspects of society, including the economic landscape, the political decision-making process, the way people interact and the skills needed to get a good job. The emerging digital economy has the potential to generate new scientific research and breakthroughs, fuelling job opportunities, economic growth, and improving people life quality.

Nevertheless, the competitiveness of a country strongly depends on its ability to invest in Research and Development (R&S), in scientific and technological training and in the training of specialised professionals.

For this reason, in 2014, the European Union (EU) published 'Horizon 2020', the biggest EU Research and Innovation programme, where is claimed that investment in research and innovation is essential for the future of Europe.

In a frame of economic development and of the emergence of new technologies, two significant paradigms are establishing '*sharing economy*' and '*open innovation*'.

Sharing economy is an economic model in which individuals are able to borrow or rent assets owned by someone else. It is an alternative to the capitalistic system (Comito, 2016).

Open Innovation, also known as external or networked innovation, focusses on the scouting of new ideas, reducing risk, increasing speed and leveraging scarce resources.

These paradigms are indeed instruments to increase competitiveness.

To stimulate and accelerate these dynamics, in the last few years, more and more structures were developed that have become the preferred physical places in which all these concepts merge and find their utmost expression, thus contributing to the creation of an ecosystem of innovation. They are spaces with optimised sharing and collaboration among self-employed, small emerging companies (SMEs, spin-off or start-up), consolidated business

enterprises and representatives responsible for managing relationships. Among them: 'Science Park', 'Business incubation' and 'coworking space'.

"A Science Park (PST) is an organisation managed by specialised professionals, whose main aim is to increase the wealth of its community by promoting the culture of innovation and the competitiveness of its associated businesses and knowledge-based institutions. To enable these goals to be met, a Science Park stimulates and manages the flow of knowledge and technology amongst universities, R&D institutions, companies and markets; it facilitates the creation and growth of innovation-based companies through incubation and spin-off processes; and provides other value-added services together with high quality space and facilities" (IASP, 2002).

"Business incubation is a business support process that accelerates the successful development of start-up and fledgling companies by providing entrepreneurs with an array of targeted resources and services. These services are usually developed or orchestrated by incubator management and offered both in the business incubator and through its network of contacts. A business incubator's main goal is to produce successful firms that will leave the programme financially viable and freestanding. These incubator graduates have the potential to create jobs, revitalise neighbourhoods, commercialise new technologies, and strengthen local and national economies" (INBIA, 2007).

"Coworking spaces are created for the community and with the community in mind. It is not just a real estate business in which a physical space is rented: the role of the facilitator (or host, community leader, or any other title you want to use) is to enhance the connection and the interaction of coworkers to bring them value and to accelerate serendipity. It is a network, not just a place. It is not enough to put a bunch of people together in a room, you must work hard to create the right interactions that form a sense of community" (Valentino, 2013: 87).

The frequenters of these innovation spaces are various, such as digital nomads, freelancers or employees who work outside the company.

Start-ups are preferably set up in incubators or business accelerators and in PSTs, but in the early stages of their lives the lack of money often leads them to choose a solution like coworking.

Coworking spaces and, consequently, their related philosophy represent a bottom-up solution or a collective strategy for facing up to structural changes of the labour market. Furthermore, coworking represents a new modality of organising project-oriented work and largely freelance occupations as found in the cultural and creative industries (Merkel, 2015).

The number of coworking spaces and their variety will definitely continue to grow in the near future.

What does coworking mean?

> *"Coworking spaces are shared workplaces utilised by different sorts of knowledge professionals, mostly freelancers, working in various degrees of specialisation in the vast domain of the knowledge industry. Practically conceived as office-renting facilities where workers hire a desk and a wi-fi connection these are, more importantly, places where independent professionals live their daily routines side-by-side with professional peers, largely working in the same sector – a circumstance which has huge implications on the nature of their job, the relevance of social relations across their own professional networks and ultimately their existence as productive workers in the knowledge economy"* (Gandini, 2015: 125).

When was the term coined?

'Coworking', a term coined by Bernard De Koven in 1999 (Rief, Stiefel, and Weiss, 2016), was fundamentally different from traditional corporations, where work was under constant observation and assessment. The core concept of coworking is to work together as equals.

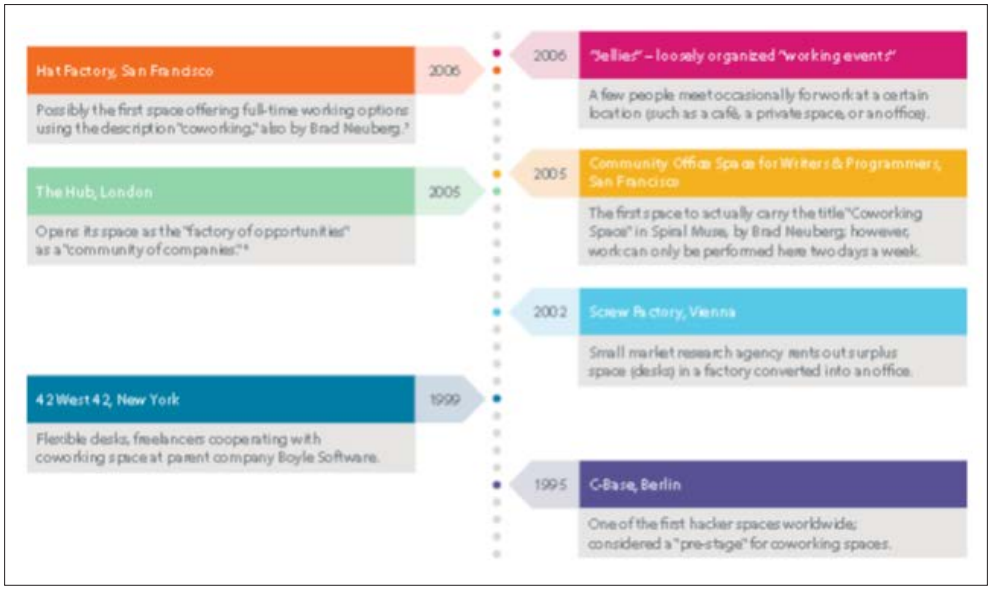
"When I coined the term coworking I was describing a phenomenon I called working together as equals […] I learned that the whole idea of working together as equals was a lot more revolutionary than I had naively assumed. For the most part, people don't work together as equals, especially not in the business world where they are graded and isolated, creating, for the majority of employees, an indelibly competitive relationship […] The environment created was also designed to allow coworkers to work together, as equals. But separately, each working on their own projects, pursuing their own, separate business interests. In this way, people were free to help each other without worrying about competitive pressures. And the result was productivity, community, and, surprisingly often, deeply shared fun" (De Koven, 2013: 45).

When did coworking become a real space?

In 1995, in Berlin a group of computer enthusiasts founded 'C-Base' (Figure 5.1). It is a hacker space considered like an early stage of a coworking space.

"Physical, community-oriented spaces where people with an interest in computers could gather to collaborate and work in an open-environment. While this model deviates from the coworking spaces we know of today, hacker spaces are viewed by some as setting the foundation for today's collaborative workspaces" (Enea, 2017:7).

Figure 5.1 - Coworking timeline



Source: Rief, Stiefel, and Weiss, 2016

In 1999, in the same year when De Koven introduced the notion of coworking, *42West24* sprung into the New York City scene. The space offers a pleasant work environment with flexible membership options for teams and individuals seeking a workspace, although the community concept was not emphasised.

In 2002, the first shared workspaces appeared in Europe. In particular, in Vienna 'Screw Factory' was born. Considered as the mother of coworking, it is usually defined as a 'community centre for entrepreneurs' (Waber, 2014).

In 2005, in San Francisco 'Spiral Muse' represented the first working environment officially defined coworking. Neuberg the founder of this space wrote a blog article clearly describing the hallmark of coworking: the sense of community that it creates between users, thanks to the organisation of group activities that can encourage the sharing of ideas and experiences.

"Traditionally, society forces us to choose between working at home for ourselves or working at an office for a company. If we work at

a traditional 9 to 5 company job, we get community and structure, but lose freedom and the ability to control our own lives. If we work for ourselves at home, we gain independence but suffer loneliness and bad habits from not being surrounded by a work community. Coworking is a solution to this problem. In coworking, independent writers, programmers, and creators come together in community a few days a week.

Unlike a traditional office, in the Spiral Muse Coworking Group we begin the day with a short meditation and circle to set our personal and work intentions […]. Then, we work in the amazing Spiral Muse house, sitting at tables or relaxing on couches as we do our work. Even though each of us is doing separate work, perhaps programming or writing a novel, we can feel each-other presence, run ideas by the community. We take lunch as a group, and then later in the day have a 45-minute break, where we do a different healthy activity every day, such as guided yoga, meditation, a nice walk, or perhaps a bike ride in the sun" (Neuberg, 2005).

Starting from this period, the word 'coworking' became a commonly known word.

On January 2006 'CoworkingWiki', created by the co-founder of Hat Factory, debuted online with the following website description:

"What is coworking? The idea is simple: independent professionals and those with workplace flexibility work better together than they do alone. Coworking spaces are about community-building and sustainability. Participants agree to uphold the values set forth by the movement's founders, as well as interact and share with one another. We are about creating better places to work and as a result, a better way to work" (CoworkingWiki, 2006).

On February 2008, the New York Times published the first article on the theme of coworking.

> *"It seemed I could either have a job, which would give me structure and community or I could be freelance and have freedom and independence. Why couldn't I have both? As someone used to hacking out solutions, Mr. Neuberg took action. He created a world, coworking (eliminating the hyphen) and rented space in a building, starting a movement"* (Fost, 2008).

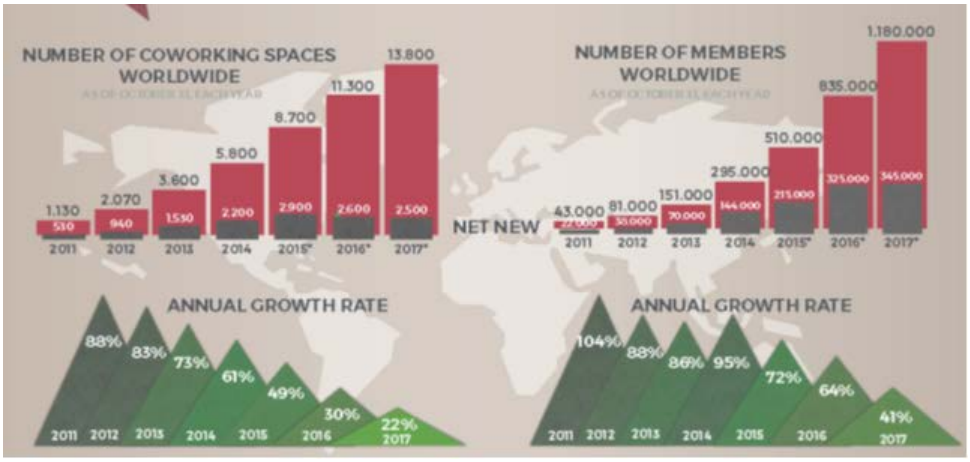
In 2010, 'Deskmag', the first digital magazine on coworking was published online and, in the same year, the first 'Coworking Global Meeting' was organised in Munich, involving 661 people coming from 24 different countries.

Since 2011, 'Global Coworking Unconference Conference' (GCUC), one of the most important conferences, has been periodically organised.

6. A statistical overview

The number of coworking spaces in the world has increased very fast: according to the 'Global Coworking Survey' the estimated number of coworkers in 2017 was more than one million with 13,800 spaces; the trend of these numbers has been continuously increasing (Figure 6.1).

Figure 6.1 - Number of coworking spaces and of members worldwide



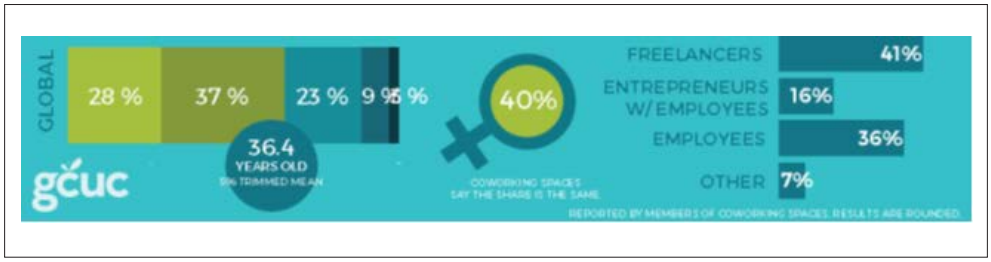
Source: Deskmag, 2017

Another important indicator is the user composition (Deskmag, 2017).

In the global context, coworking members are especially freelancers (41%) or employees (36%) (Figure 6.2).

Usually they are young people who are trying to undertake an entrepreneurial career as double job.

Figure 6.2 - The members of coworking spaces

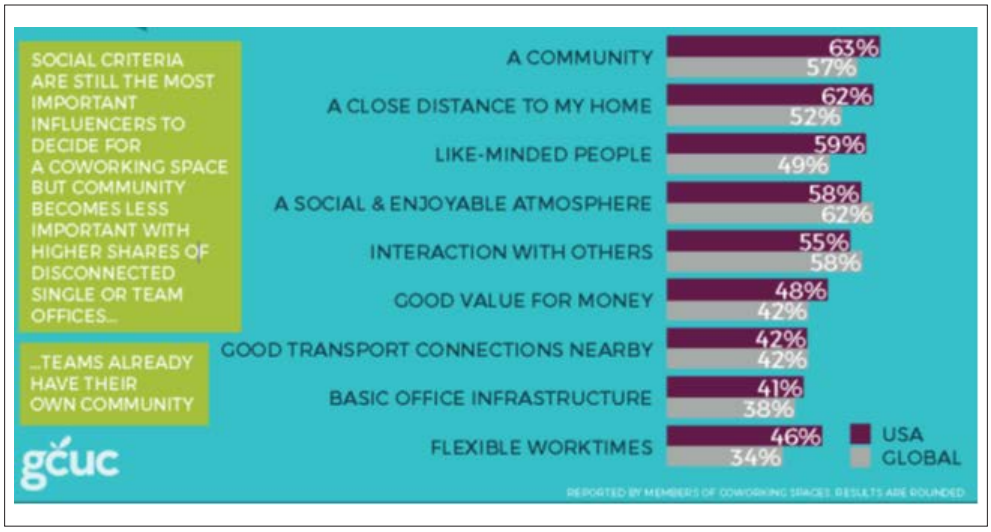


Source: Deskmag, 2017

The average number of members using coworking space has increased constantly: in 2012 the number of people involved was 38 and it doubled in just four years.

Among the motivations that lead users to choose a coworking space, the most frequent is the possibility to create a community (57%), while the lowest frequent is 'flexible worktimes' (34%) (Figure 6.3).

Figure 6.3 - Main reasons to choose a coworking space

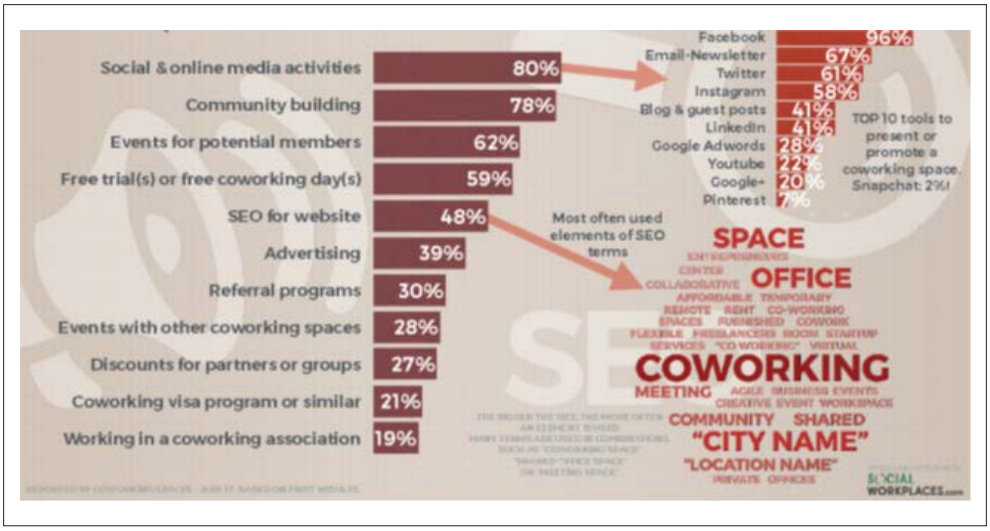


Source: Deskmag, 2017

The top tools to attract new members are the possibilities of promoting 'social & on line media activities' (80%) and 'community building' (78%).

The minority of coworkers (19%) stated 'Working in a coworking association' (Figure 6.4).

Figure 6.4 - Top tools to attract new members



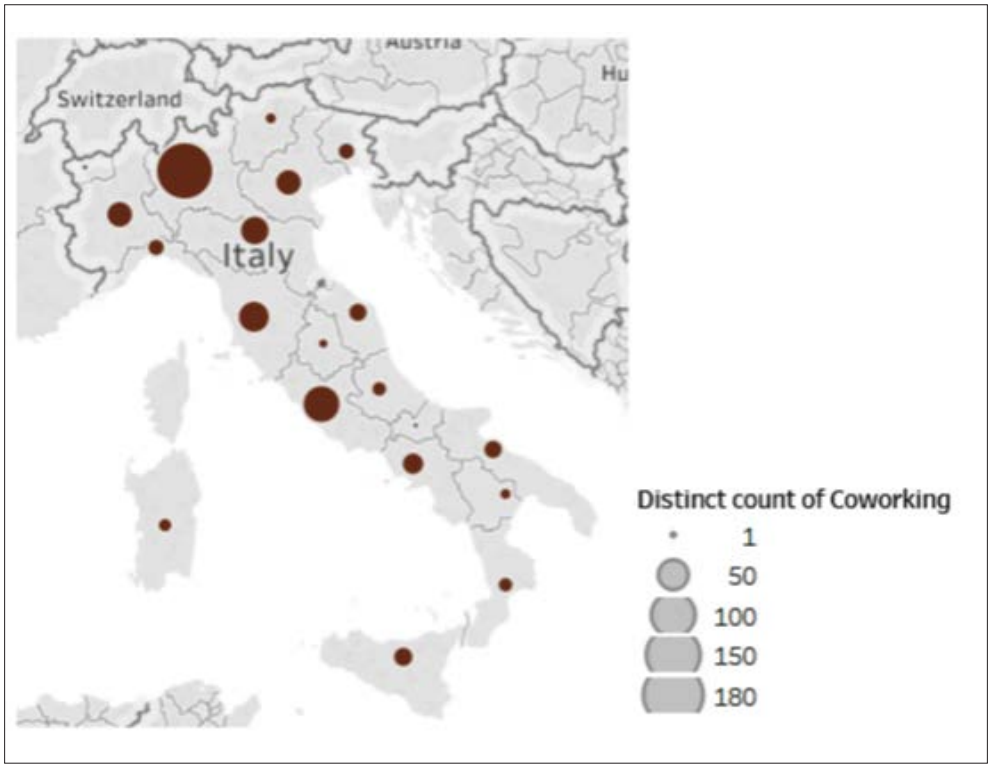
Source: Deskmag, 2017

7. Coworking in Italy

In the Italian context, coworking is rapidly evolving and its distribution is quite varied.

According to the available data (Enea, 2017), in January 2017 there were 588 coworking centres (423 in provincial capitals, 165 in other cities). The following list summarises the geographic distribution: 377 in Northern Italy, 161 in Central Italy and 90 in the South and islands (Figure 7.1).

Figure 7.1 - Distribution of coworking in the Italian regions in 2017



Source: Enea, 2017

In Italy, the first coworking idea took place in Lambrate (Milano) in 2008: 'Cowo' (coworking project) is the most popular coworking network in Italy, having offices in almost all the regions (Piemontese, 2016).

'The Hub' represents a network-connected coworking that has offices in Milano, Firenze and Roma as well as, in Southern Italy, in Bari, Catania and Siracusa.

The highest concentration of coworking spaces is in Northern Italy.

More in detail, in Milano we found besides 'Cowo' and 'The Hub' also 'Plan C', which has been designed for and by women; in Torino 'Cowo', 'Talent Garden', 'Toolbox'; in Alessandria 'Lab 121'; in Padova 'Talent Garden' and 'TalentLab'; in Modena 'Well_B_Lab' (a spin-off cooperative of the University of Modena and Reggio Emilia); in Firenze 'Multiverso'; in Roma 'Cowo', 'The Hub', '7h floor' and 'Let's Make'.

In Southern Italy, some local facilities are located in Napoli, Salerno, Bari, Catania, Siracusa and Cagliari.

A useful tool to find the nearest shared office is 'Coworkingfor', a search engine of coworking spaces.

The regional capacity to propagate and support this mode of work organisation also determines the spread of coworking. In fact, several Regions have promoted coworking supporting policies for young people and startup projects, such as: vouchers to rent coworking stations; funding to create coworking centres, incubators or business accelerators sometimes associated to urban regeneration; financial support for training activities; guidelines for coworking implementation.

The coworking organisational models can be top-down or bottom-up. In the first case, national or international companies or public administrations manage the organisation of coworking.

In the second case, small companies, start-ups and associations are in charge of the organisation.

8. Conclusions

Thanks to the development of the digital age and of the technological progress, as well as to the development of sharing economy and of open innovation, the 'knowledge worker' has no longer the need of a desk set: the new model of digital nomad is a worker who moves around the world making use of coworking spaces. The workplace becomes a shared space, where coworkers can build professional networks facing with today's labour market.

The analysis of the historical development of coworking arrangements highlights several aspects, as summarised in the following points.

- The importance of group activity as a fundamental element to create a sense of community and membership.
- The need of connection with other people as a key for promoting social networks.
- The creation of a sense of community to encourage exchange and contamination of ideas in different areas.
- The increase of heterogeneous teams composed by people that works in different contexts.
- The possibility to rent a desk in a shared workspace as an opportunity to reduce office costs.
- The opportunity to leave workers free of moving, thus also allowing the discovering of new places.
- The development of self-employment.
- The possibility of reusing architectural heritages originally built for different uses.

The occurrence of coworking has been spreading fast in the last decade and it is highly representative of new labour market trends. This solution matches flexibility needs, such as independence, innovation and cooperation and new necessities emerged in the recent past.

Finally, Coworking represents not only a new system of work organisation but also an answer to the isolation risk and to the need of work-life balancing.

For all this reasons, coworking represents a relevant field of study and analysis for official statistics, that should monitor its diffusion in terms both of spaces and typologies. In addition, it is worth deepening also the different aspects related to the users within urban contexts that are increasingly smart and oriented to the citizens' well-being.

References

Accornero, A. 1994. *Il mondo della produzione*. Bologna: il Mulino.

Bacigalupo, T., T. Sundsted, and D. Jones. 2009. *I'm outta here: how coworking is making the office obsolete*. New York, NY, U.S.: Harper Collins.

Beauregard, A., K. Basile, and E. Canonico. 2013. "Home is where the work is: a new study of homeworking in Acas – and beyond". *Research Paper*, N. 10/2013. London, UK: Acas Research and Evaluation Programme.

Berger, T., and C.B. Frey. 2014. "Industrial renewal in the 21st Century. Evidences from US cities". *Working Paper*. Oxford Martin School. Oxford, UK: University of Oxford.

Bonazzi, G. 2008. *Storia del pensiero organizzativo*. Milano: Franco Angeli.

Catino, M. 2012. *Capire le organizzazioni*. Bologna: il Mulino.

Chester, E. 2002. *Employing Generation Why?* Weatherford, Parker County, TX, U.S.: Tucker House Books.

Cole, G., R. Smith, and L. Lucas. 2002. "The debut of Generation Y in the American Workforce". *Journal of Business Administration Online*, Volume 1, N. 2.

Comito, V. 2016. *La sharing economy. Dai rischi incombenti alle opportunità possibili*. Roma: Ediesse.

De Koven, B. 2013. *The Coworking Connection*. Deep fun. https://www.deepfun.com/the-coworking-connection/.

Deane, P. 1971. *La prima rivoluzione industriale*. Bologna: il Mulino.

Deskmag. 2016. 2016 *Coworking Forecast*. http://www.deskmag.com/en/2016-forecast-global-coworking-survey-results.

Deskmag. 2017. *The 2017 Global Coworking Survey*. http://www.deskmag.com/en/background-of-the-2017-global-coworking-survey.

Dubey, N.B. 2009. *Office Management: Developing Skills for Smooth Functioning*. New Delhi: Global India Publications.

Enea. 2017. *Coworking…. Che? I nuovi volti nell'organizzazione del lavoro: un'indagine sul coworking in Italia*. Frascati: Laboratorio Tecnografico Enea.

European Trade Union Confederation – ETUC, Union of Industrial and Employers' Confederations of Europe – UNICE, Union Europeenne de l'Artisanat et des Petites et Moyennes Entreprises – UEAPME, and European Centre of Employers and Enterprises providing Public services - CEEP. 2006. *Implementation of the european framework agreement on telework report by the european social partners. Adopted by the Social Dialogue Committee on 28 June 2006*.

Fost, D. 2008. "They're Working on Their Own, Just Side by Side". *The New York Times*. February 20th, 2008.

Frey, C.B., M. Osborne, and Citi Research. 2015. *Technology at work. The future of innovation and employment*. Oxford Martin School, University of Oxford. Citi GPS: Global Perspectives & Solutions.

Gandini, A. 2015. "The rise of coworking spaces: A literature review". *Ephemera Journal*, Volume 15 (1): 193-205.

Goldin, C., and L.F. Katz. 2007. "The Race between Education and Technology: The Evolution of U.S. Educational Wage Differentials, 1890 to 2005". *NBER Working Paper Series*, Working Paper N. 12984. Cambridge, MA, U.S.: National Bureau of Economic Research – NBER.

Harvard Innovation Lab. 2015. *The Evolution of the Desk - 1980 to 2014*.

Howe, N., W. and Strauss. 2000. *Millennials Rising: The Next Great Generation*. New York, NY, U.S.: Random House.

International Association of Science Parks and Areas of Innovation – IASP. 2002. *Definitions. A glossary of some key terms and definitions from the industry of science and technology parks and areas of innovation*. Campanillas, Malaga, Spain: IASP.

International Business Innovation Association - InBIA. 2007. *A Practical Guide to Business Incubator Marketing*. Orlando, FL, U.S.: InBIA Publications.

Levels, P. 2015. "The future of digital nomads. How remote work will transform the world in the next 20 years". *Presentation at the Digital Nomad Conference*, Berlin, Germany, October 2015.

Merkel, J. 2015. "Coworking in the city". Ephemera Journal, Volume 15 (1): 121-139.

Mingione, E., and E. Pugliese. 2010. *Il lavoro*. Roma: Carocci editore.

Murty, P.S.R. 2017. "Digital Economy - B2C – Digitalization". *Paper statistics*. Social Science Research Network - SSRN.

Neuberg, B. 2005. Coworking: *Community for developers who work from home*. http://codinginparadise.org/weblog/2005/08/coworking-community-for-developers-who.html

Piemontese, N. 2016. *Coworking: la mappa degli spazi sul territorio italiano*. bianco lavoro: https://www.biancolavoro.it/.

Rief, S., K.P. Stiefel, and Agnes Weiss (Fraunhofer-IAO). 2016. *Harnessing the Potential of Coworking*. Holland, MI, U.S.: Haworth.

Ross, P.K, S. Ressia, and E.J. Sander. 2017. *Work in the 21st Century: How Do I Log On?* Bingley, UK: Emerald Publishing Limited.

Schwab, K. 2017. *La quarta rivoluzione industriale*. Milano: Franco Angeli.

Spiro, C. 2006. "Workplace transformation: Generation Y in the Workplace". *Defense AT&L Magazine*, November-December 2006: 16-19.

Stokes, K., E. Clarence, L. Anderson, and A. Rinne. 2014. *Making sense of the UK collaborative economy*. London, UK: Nesta.

Swann, T. 2017. "Information, cybernetics and the second industrial revolution". *Ephemera Journal*, Volume 17 (2): 457-465.

Tapscott, D. 1995. *The Digital Economy: Promise and Peril in the Age of Networked Intelligence*. New York, NY, U.S.: Mcgraw-Hill Education.

Valentino, R. 2013. *Coworkingprogress. Il futuro è arrivato*. Busto Arsizio: Nomos Edizioni.

Waber, B., J. Magnolfi, and G. Lindsay. 2014. "Workspaces That Move People". *Harward Business Rewiew*, Ottobre 2014 Issue.

Woolf, B.P. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Burlington, VT, U.S.: Morgan Kaufmann.

Decision tables for mortality coding: methods and tools for the management and documentation of changes

Simone Navarra [1], Marisa Cappella [1], Lars Age Johansson [2], László Pelikan [3], Friedrich Heuser [4], Luisa Frova [1], Francesco Grippo [1]

Abstract

Decision Table Editor (DTE) is a web-based system developed by Istat in the framework of an international collaboration (Iris Institute). By means of this application, experts from different countries can collaborate on the coordinated and simultaneous maintenance and update of the decision tables used for the underlying cause-of-death selection. These tables provide criteria for the correct application of the selection rules of the International Classification of Diseases (ICD10), published by World Health Organization (WHO) and periodically updated. They derive from those originally developed by the US National Centre for Health Statistics (NCHS) for the ACME software and represent a major tool for enhancing the international comparability of mortality statistics. One of the major achievements of the DTE is the improvement of transparency and documentation of changes introduced in the tables which have a direct impact on mortality statistics.

Keywords: Mortality coding, ICD10 updates, decision tables, Iris.

1 Italian National Institute of Statistics - Istat.

2 Swedish National Board of Health and Welfare.

3 Hungarian Central Statistical Office - KSH.

4 Federal Institute for Drugs and Medical Devices - BfArM (formerly, German Institute for Medical Documentation and Information - DIMDI).

The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

1. Introduction

Decision Table Editor (DTE) is a web-based application for the maintenance of the decision tables used for the selection of the underlying cause of death (UC). It has been developed by the Italian National Institute of Statistics - Istat in the framework of the collaboration with the Iris Institute which emerged from an international cooperation for the deployment, maintenance and development of the Iris software, an electronic system for automated coding of causes of death. The Institute is hosted at DIMDI (German Institute for Medical Documentation and Information) and the current cooperating partners are statistical institutions from France, Germany, Hungary, Italy, Sweden and United States (Iris Institute website www.iris-institute.org). Istat officially joined the group by means of an agreement with the DIMDI signed in 2012. Nevertheless the collaboration of Istat with the other European partners for the development of Iris software had begun two years earlier.

The decision tables are central to the function of Iris. The tables are primarily used by Iris software but they also constitute a support for manual coding and represent the knowledge base for the consistent and harmonised application of the international rules for the selection of the UC according to the provisions of the International Classification of Diseases and Related Problems, tenth revision (ICD10), published and revised by the World Health Organization (WHO, 2010). These tables make it possible to apply these rules by computer programmes and by coders with limited medical experience.

The knowledge database was first developed by the NCHS (US National Center for Health Statistics) for the ACME system (ACME tables). Successively, since 2011 it has been maintained by the Iris Institute for the inclusion of the annual WHO official updates of the ICD. Hence, the tables used by Iris differ by some extent from the NCHS ones (CDC, NCHS, 2016) for the inclusion of updates since 2010 on.

DTE is also accessible to the general public for downloading the decision tables in pdf format at the web-address www.iristables.istat.it.

2. Selection of the underlying cause and harmonised statistics[5]

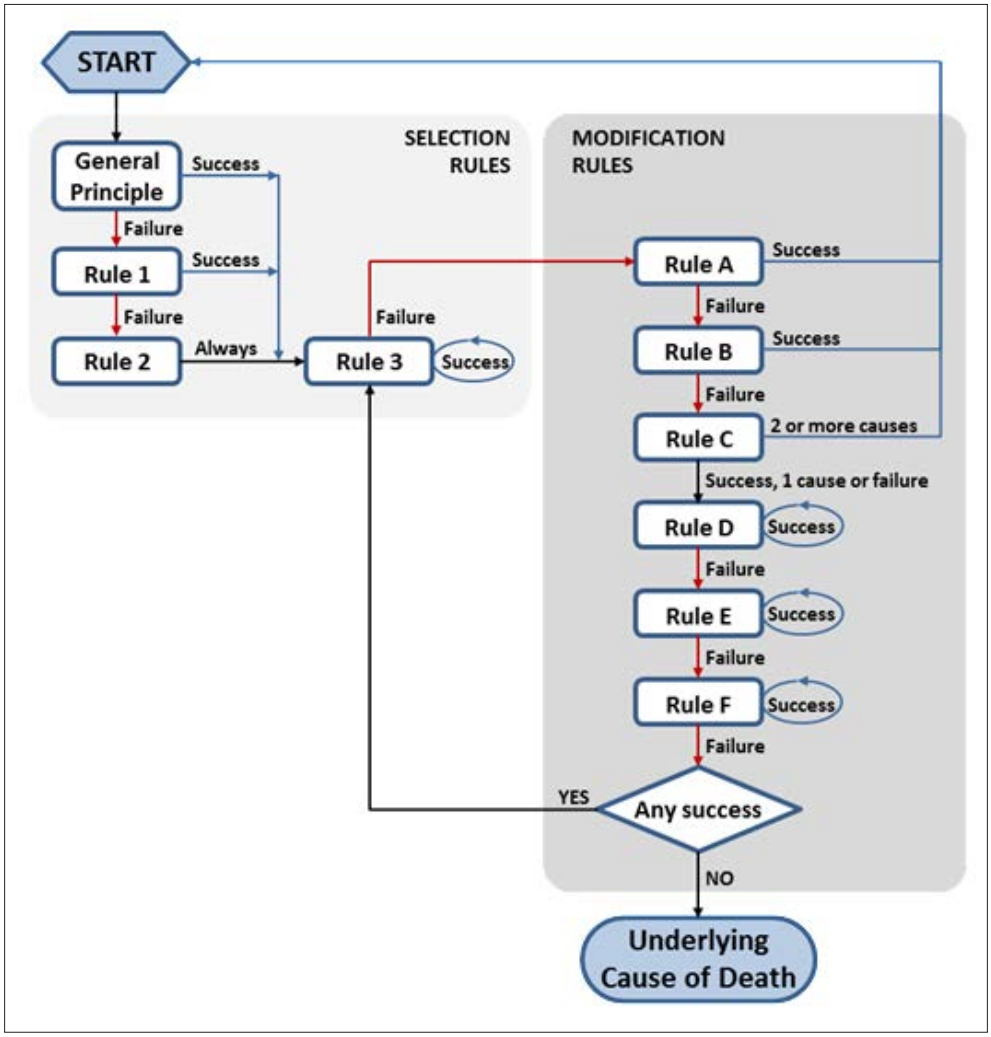
Comparison of mortality statistics are mostly based on the underlying cause of death. This is defined by the WHO (2010) as "(a) the disease or injury which initiated the train of morbid events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury". For each death, the UC is selected from an array of conditions reported by a physician on the death certificate through the application of the selection and modification rules contained in the appropriate revision of the ICD. Selection rules included in ICD are meant to be a systematic guidance for selecting the UC, thus ensuring comparability and uniformity in mortality statistics among different countries. Figure 1 represents a simplified schema of how the rules apply during the selection (WHO training tool). Although some details of the selection process are left out from this Figure, it clearly shows that the selection process can be seen as a complex algorithm with several decision nodes. The criteria for determining the success or failure of each node are described in specific instructions included in the volume 2 of the classification or by other provisions such as the inclusion/exclusion notes of the tabular list and the alphabetical index.

The procedures for selecting the UC imply two main steps. In the first, the selection is finalised to identify the antecedent originating cause which is the starting point of the sequence of events leading to death. This step primarily involves the application of General Principle or Rule 1 or 2. For the application of these rules, the sequence reported by the physician on the death certificate must be examined in order to evaluate its correctness. The classification provides instructions on sequences to be accepted and those to be rejected. After one of these rules, Rule 3 is applied, in order to evaluate if the cause identified in the previous steps can be considered an obvious consequence of another condition reported. Also in this case the ICD provides instructions for detecting obvious consequences. In the second step of the coding process, a modification of the selected cause is performed. This step is finalised to select a more informative condition if the first selected is an ill-defined (Rule A) or trivial affection (Rule B); to combine information reported

5 This description and furher parts of this paper refer to the 2010 edition of the ICD10. Although in 2016 the rule application schema and the name of the rules were deeply revised, the content of this paper remains still valid and applicable to the new framework of the ICD10 rules.

in different parts of the certificate (Rule C, linkage); to select a more specific condition (Rule D). This modification allows selecting a more informative condition for public health purposes.

Figure 1 - Coding rules and coding algorithm (a)



(a) This Figure is an adaptation of the flow chart included in the WHO training tool "ICD10 Interactive self-learning tool" available on the WHO website. It reflects ICD10 instructions until 2010. Although it leaves out some details of the selection process and does not contain special instructions such as surgery and procedures, it shows the complexity of the coding. For a complete and up-to-date information on this topic refer to ICD10 volume 2 and its updates on the following link http://www.who.int/classifications/icd/en/.

2.1 Automated coding and the Iris system

The international rules and instructions for the selection of the UC, leave space for interpretation, resulting in a certain degree of variability of the tabulated UC among coders (Harteloh et al., 2010) and, thereby, across countries. The interpretation derives from both the complexity of the algorithm and the criteria for decision making in each node. In order to face the problem, since the 1960s, US National Center for Health Statistics (NCHS) has been the major investor in the research and development of an automated mortality coding system and in 1968 developed the Mortality Medical Data System (MMDS) for the coding of both the UC and multiple causes on the death certificate (CDC website, about MMDS). MMDS consists of two main components: MICAR (Medical Information, Classification And Retrieval) and ACME (Automated Classification of Medical Entities). MICAR module assigns an ICD code to each condition reported generating the input for the ACME module which then, by using the set of logical decision tables, applies the international selection and modification rules, resulting in the selection of the tabulated UC.

A number of European countries implemented MMDS in the '90s of the last century. In some of them automated systems in languages other than English have been developed using the ACME decision table logic (Pavillon et al., 1999). France and Sweden in particular, started to cooperate on a common tool thanks to the experience of these countries in the use of automated coding. Successively, Germany joined the project and finally, in order to improve the international comparability of mortality statistics, Eurostat (the statistical office of the European Union) supported the development of Iris, a common, language-independent coding system that can be used for coding death certificates, written in any language, according to ICD coding rules and guidelines for the selection of the UC (Pavillon et al., 2007, Pavillon 2012). Version 4 of Iris uses MMDS components while version 5 contains a newly developed module, MUSE (Eckert, 2014).

3. Decision Tables for the selection of the underlying cause

The decision tables represent the knowledge base for the coding, both manual and automated, which allow taking decisions for every step of the coding algorithm represented in Figure 1. They are a formalisation of the instructions included in the volume 2 of the ICD10. This formalisation basically consists in the translation of the provisions of the Classification into relationships between pairs of ICD codes.

The tables were first developed by NCHS as part of ACME and are still released on official website as Part 2c of the Vital Statistics Instruction Manual series (CDC, NCHS, 2016). Nevertheless, when Iris was developed, some changes in the tables were needed in order to fit the specificity of the new software and also for including some official WHO updates. Despite these changes the Iris tables maintained the same structure as the NCHS ones. In Figure 2, an extract of Iris 2014 tables is shown (print version).

The Iris tables can be summarised as follows:

- valid codes table (corresponds to the NCHS tables A, B, C, G and H), includes the list of the ICD10 codes with the description of the properties of each code for mortality coding purposes. Certainly, not all the codes reported in the ICD are used for mortality coding and some of them are not used for the UC coding, but they can be used for multiple causes. Therefore, code validity, for both multiple and UC coding, is documented in the table as well as other flags informing on other characteristics such as: ill-defined condition activating rule A; trivial affection which activates Rule B; created code and, for these, the correspondence with the ICD10 valid codes used for data tabulation (NCHS Table G). Created codes are special codes not included in the ICD, used for capturing information contained in the diagnostic term, which is necessary during the coding process. In some cases, the regular ICD10 code is not sufficient for describing such detail indeed. As an example, the code A16.9 is used for coding both diagnostic expressions “tuberculosis” and “respiratory tuberculosis”. Nevertheless these two expressions can have a different behaviour during the UC selection. In order to distinguish between these two situations, the table includes the plain code A16.9 for coding “respiratory tuberculosis” while the

created code A16.90 is used for the term "tuberculosis" without other specifications.

- causal table (NCHS Table D), contains the accepted causal sequences and it is used for the application of General Principle, Rule 1 and Rule 2.
- modification table (NCHS Table E) lists modification relationships between codes. Various relationships can exist between two codes according to the reference rule. It represents the main guidance in application of Rule 3, and modification rules A, C and D.
  Both causal and modification tables contain ambivalent entries also indicated as "maybe" relationships. The maybes are generated by the fact that the ICD codes are used for coding broad groups of specific conditions while causal and modification relationships might involve only subsets of these. In these cases the UC selection depends on the analysis of the text reported by the physician and must be manually revised according to the explanation reported in the text next to the relationship involved. In NCHS tables the maybe explanations are included in a separate Table F. The maybe explanations are provided only for the modification table. As discussed previously, the created codes are used as well in these situations, with the advantage of allowing these cases to be automated processed.

In general, the causal and modification tables have a common structure and can be seen as a single component. Nevertheless, for practical reasons, they are generally presented as separate tables. Actually, the causal and modification tables are used in two separate moments of the coding process, first when applying the selection rules and second during the modification.

Figure 2 - Decision table structure

VALID CODES TABLE

Code	Ill-defined	Trivial	Created	Code conversion	Validity
A000	No	No	No		Valid for multiple and underlying
A001	No	No	No		Valid for multiple and underlying
…					
A169	No	No	No		Valid for multiple and underlying
A1690	No	No	Yes	A169	Valid for multiple cause only
…					
F03	No	No	No		Not to be used if underlying condition is known

CAUSAL TABLE

---E140-E141---			---E140-E141--- Continue		---E140-E141--- Continue
B252			K850	-K861	Y525
B263			K868	-K869	Y527
C250	-C259		M359		Y543
C788		M	O244		
D136	-D137		P350		---E142---
D350			Q871		
E050	-E69		Q900	-Q909	B252
…			…		…

MODIFICATION TABLE

---D739----

SMP	C261		
SMP	C788	M	Suba must be spleen
DS	C810-C969		
SMP	D139	M	Suba must be spleen
SMP	D377	M	Suba must be spleen
SMP	D730-D378		

Figure 2 shows the tables as they appear in the paper-based format, where the causal and modification tables are separate sets. On the other hand, Prospect 1 describes the variables of the tables as they were a single set and provides a short description of the variables included.

The causal and modification tables contain address and subaddress codes. The address is either a single 3-5 digit code or a span of codes enclosed in dashes (e.g. “---E142---” is a single code, “---E140-E141---” is an interval of codes). The subaddress is given under the address and may also consist of a single code or a span of codes. Note, for instance, that the span E050-E69 includes all the valid codes from the valid codes table from E050 to E69. In the modification table the following acronyms precede each subaddress indicating the relationship with the respective address and designating the applicable rule: DS, DSC, IDDC, SENMC, SENDC, LMP, LMC, LDP, LDC, SMP, SMC, SDC. In some cases an additional code is reported on the right of the subaddress (not shown in the Figure). This code, referred as recode,

identifies a code resulting from the combination of the tentative UC (address) and another code on the death certificate (subaddress). Table D contains just one type of relationship between address and subaddress so the acronym is not reported but it is understood as DUE. The symbol "M" is used in both table D and E to denote ambivalent (maybe) relationships. Reasons to these ambivalences are displayed next to the "M" and provide further guidance in the selection of the most appropriate UC. For some cases special attention is required when applying a rule. These entries are flagged with a symbol "#" (not shown in the Figure).

Prospect 1 - Variables of causal and modification tables and types of relationships (a)

Variable	Modality	Description	
Address	A000-Y98	Also referred as anchor code or simply code. It is the tentative UC resulting from the selection process. It can be represented as a single code or as a span of codes (address1-address2).	
Subaddress	A000-Y98	Also referred as subanchor code or subcode. It is another code present on the death certificate. It can be represented as a single code or as a span of codes (subaddress1-subaddress2).	
Relationship		Also referred as rule, is the type of relationship that links address and subaddress codes and indicate which ICD10 rule is applicable.	
	DUE	Due to	General Principle, Rule 1 and 2
	DS	Obvious consequence	Rule3
	DSC	Obvious consequence with combination	
	IDDC	Ill-defined, in due to position with combination	Rule A
	SENMC[b]	Senility, in mention position with combination	
	SENDC[b]	Senility, in due to position with combination	
	LMP	Linkage, in mention position with preference	Rule C
	LMC	Linkage, in mention position with combination	
	LDP	Linkage, in due to position with preference	
	LDC	Linkage, in due to position with combination	
	SMP	Specificity, in mention position with preference	Rule D
	SMC	Specificity, in mention position with combination	
	SDC	Specificity, in due to position with combination	
Recode	A000-Y98	Is the code resulting from a combination of the address and subaddress, when modification rules are applied for the relationships DSC, IDDC, SENMC (b), SENDC (b), LMC, LDC, SMC, SDC.	
Maybe flag	M	Indicates ambivalent relationships: entries with ambivalent relationships are flagged with the letter "M". Both causal and modification table contain ambivalent relationships, but explanation are provided only for those in modification tables.	
Maybe reason	Free text	Shows the reason for ambivalent relationship. Reading the reason, the coder can decide if the relationship expressed in the entry is applicable.	
Special note	#	For some cases special attention is required when applying a modification rule. These entries are flagged with a symbol "#". This field is also referred as "neocode".	

(a) As causal and modification table share the same structure, they can be considered as an unique body. The causal table contains a single relationship which is "DUE". All the other rules refer to the modification tables.
(b) From the 2016 edition of the tables the rules SENMC and SENDC have been deleted. For SENDC the IDDC rule has been used, the new rule IDMC has been created to substitute SENMC and for other uses as well.

The structure shown in Figure 2 refers to the compressed format where relationships between ICD codes are represented, when possible, as intervals of codes. This representation is necessary in order to make paper-based tables more readable to coders. In this compressed form the tables includes more than 94,000 rows (2014 version). However, the relationship between intervals of codes is a synthetic representation of all the relationships between single pairs of codes. When the relationships between intervals of codes are resolved into relationships between single pairs of codes, the number of relationships expressed in the table account for more than 31 million. The tables in which the intervals of codes are resolved are referred as normalised tables. Table 1 shows the comparison between the compressed and the normalised structure of the tables.

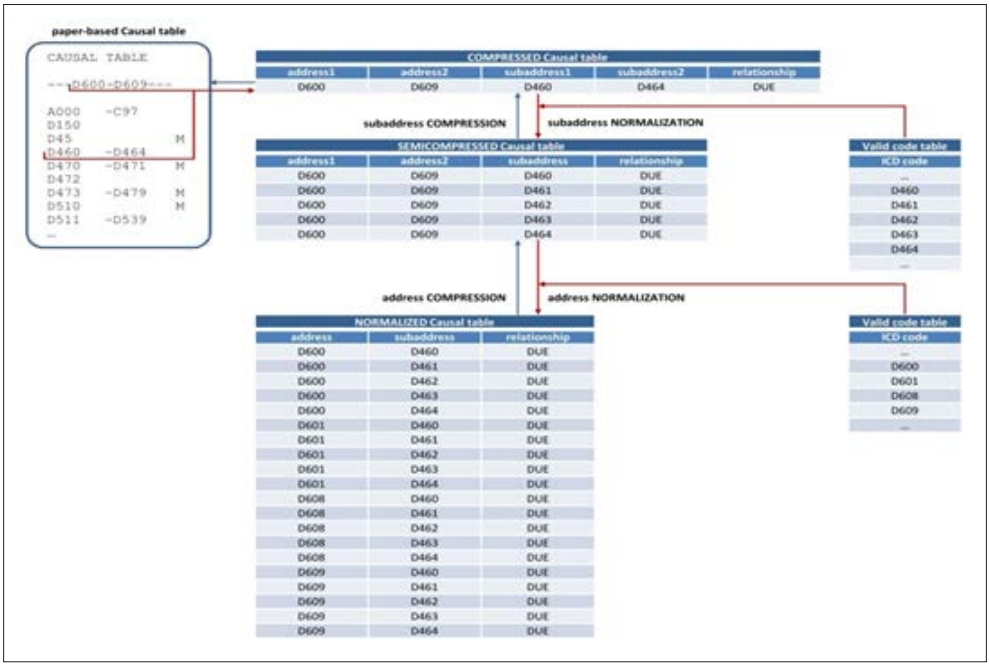
Table 1 - Compressed and normalised structure of the tables (2014 edition)

	Table D		Table E	
	Compressed	Normalised	Compressed	Normalised
Total rows	57,844	29,677,852	36,566	2,127,820
of which:				
rows with maybe	20,815	1,375,960	8,835	86,300
recode required (a)	-	-	14,496	98,756
other notes (a)	-	-	439	13,432

(a) Not applicable for causal table.

In Figure 3 an example of the normalisation procedure for a given row of the causal table is provided. To make this normalization, both address and subaddress intervals are resolved into single codes depending on the list of valid codes. The Figure shows how from a single row representing a “due to” relationship between two intervals of codes, 20 normalised rows are obtained: the product between 5 codes in the address interval (D600-D609) and 4 codes in the subaddress interval (D460-D464). Normalisation is a reversible process. Normalised tables can be compressed back to the non-normalised format through the compression procedure which is the inverse of normalisation.

Figure 3 - Normalisation and compression procedure



3.1 ICD updates and table editing

Maintenance of the decision tables is necessary for the up-to-date and the correct functioning of the Iris software. Actually, any change in the tables affects the result of Iris coding. Table maintenance consists in the annual revision in order to fulfill two needs:

- to correct errors such as incorrect or missing causal relationships or linkages;
- to apply the WHO official updates (WHO website, list of official updates). Actually, modifications in the ICD, its rules or their interpretations are implemented by editing the appropriate decision table.

It is convenient to remark that, although the editing of the tables is performed by cooperating partners of the Iris Institute, it strictly depends on decisions taken at international level and in particular it is performed, as much

as possible, according to the process of ICD updating. This process involves different organs within the net of the Collaboration Centers of the WHO for the Family of International Classifications (WHO-FIC). The official updates to the ICD10 are approved at an annual meeting by the Update and Reference Committee (URC) and published on the WHO website in the format shown in Figure 4. For the mortality application, a specific organ of the WHO-FIC called Mortality Reference Group (MRG) functions as a consulting body. The scope of the MRG is to improve the international comparability of mortality data by making decisions on coding issues, suggesting clarifications of coding instructions as well as other changes to the ICD10. This organ is also helped by a more practical group (Table Group) that recommends changes to the tables.

Figure 4 - WHO official ICD10 updates. Extract from the "Cumulative official updates to ICD10 of volume 2" available for download in pdf format on WHO official website

Instruction	Instruction manual entries	Source	Date approved	Major/ Minor update	Implementation date
Move location of sequelae of TB and add mention of chronic forms of hepatitis to section 4.2.2 of ICD-10 volume 2	**4.2.2 Accepted and rejected sequences for the selection of underlying cause of death for mortality statistics** … (a) Infectious diseases The following infectious diseases should not be accepted as due to any other disease or condition, except when reported as due to human immunodeficiency virus [HIV] disease, malignant impairing the immune system: • typhoid and paratyphoid fevers, other salmonella infections, shigellosis (A01-A03) • tuberculosis (A15-A19) • sequelae of tuberculosis (B90) The following infectious and parasitic diseases should not be accepted as due to any other disease or condition (not even HIV/AIDS, malignant neoplasms or immunosuppression): • cholera (A00) • botulism (A05.1) • plague, tularaemia, anthrax, brucellosis (A20-A23) • leptospirosis (A27) • tetanus, diphtheria, whooping cough, scarlet fever, meningococcal disease (A33-A39) • diseases due to Chlamydia psittaci (A70) • rickettsioses (A75-A79) • acute poliomyelitis (A80) • Creutzfeldt-Jakob disease (A81.0) …	MRG 1798	October 2011	Minor	January 2013

4. Decision Table Editor

4.1 Objectives

Since 2011 Iris Institute has updated and maintained decision tables taking into account the annual provisions of the WHO, even if not all the updates have been fully implemented. Updating process originally adopted was based on a spreadsheet structure. The major limits of this kind of tool were, first of all, a limited possibility to trace and retrieve changes introduced in the tables, especially for documenting the rationale of the changes. Second, it implied a significant manual intervention, increasing the chance of error. Certainly, the complexity of the tables shown above makes the table editing not a trivial task. For instance, the compressed format of tables D and E complicates data manipulation because, generally, the updating requires the disentanglement of many code intervals. Moreover the high interrelation existing among the relationships included in the tables implies that changes in one relationship can have impact on many others. Third, it did not allow the simultaneous work of different experts: updates could happen only in series but not in parallel. For all these reasons, it was essential to develop a reliable system for the annual table updates, as little dependent on direct manual intervention as possible.

To respond to the need of a continuous table updating, the Italian National Institute of Statistics - Istat, in the framework of the agreement with Iris Institute, developed the Decision Table Editor (DTE) web application. DTE is an online work platform conceived to allow international coding experts to cooperate in maintenance, production and distribution of the decision tables. DTE is therefore designed as a work and production environment rather than a mere instrument for table consultation, although data retrieval features are available for internal users.

In summary, the objectives of the DTE are:

- to handle simultaneous and coordinated access to the tool of experts from different countries for updating decision tables;
- to document the annual updates;
- to check for duplications and inconsistencies;
- to avoid manual intervention on the tables;

- to produce the decision tables used by both Iris system and manual coders;
- to store, retrieve and browse annual versions of the tables.

4.2 System overview

The system is a Java web-based application which allows managing the updating process of the decision tables. In particular:

- editing:
  - valid ICD codes;
  - decision tables;
- validation and production of annual tables;
- management of primary tables;
- browsing and downloading.

The management of the system functionalities is performed by means of a very user-friendly interface. The database of the application has been designed in Oracle and comprises of two main data groups:

- the first group is the data storage of the historicised decision tables;
- the second group is designed for recording the changes required by annual updates, and can be considered a data flow recording.

The storage group contains tables for valid codes, decision tables (both causal and modification are stored in the same table) and maybe reasons. The information of the decision tables is kept in a normalised form, as described in Figure 3, i.e. the relationships kept in the tables refer to pairs of codes and not to intervals. This way of storing information, although highly memory consuming (more than 31 million rows are needed), facilitates the updating procedures and makes data retrieval more flexible. Moreover, each record contains fields for both start and expiration year as well as a reference to the reason for the change (Id of the update giving rise to the starting or expiration), making possible to store and retrieve all annual versions of the tables and the origin of change (historicisation).

The data flow group contains the information that the system uses for making the changes to the tables according to each implementation year.

4.3 Collaborative, coordinated and controlled workflow

The first problem encountered in designing DTE, was the need for a definition of a rigorous workflow for the table updating. In this paragraph the flowchart of the updating workflow is described.

Different profiles are designed for different tasks and, in order to trace all the activities performed on DTE, access to system requires username and password and implies a three-tiered permission architecture.

The three internal user profiles are: Administrator, Supervisor and Editor.

The implementation of annual WHO updates, as well as correction of errors, consists in the addition of new rows and modification or deletion of existing rows from the tables of the previous year. Nevertheless, with the DTE, these modifications are not directly performed on the tables but are inputted in a specific encoding panel and successively applied to the tables by the system itself. Every change in the tables is maintained in order to track and retrieve different annual versions. This updating process is designed for releasing and storing a single annual version of the tables in the database. Changes can be made several times in a given year but only one final edition is kept.

The complete workflow is represented in Figure 5 and it is described below.

Phase 1. Data input: update definition and check

Editors are involved in phase 1 of the process. Their main task is to insert data derived from the agreed updates to be implemented in the year. DTE system is designed to manage simultaneous access of different editors. Furthermore, when one or more editors work on data entry, changes to database are univocally identified allowing to trace the source of any modification and the operator who made it. The detailed steps of this phase are the following:

- *Updating of valid codes table.* Definition of expiring date for expired codes, addition of new codes, modification of attributes (trivial, ill-defined, etc.). Each modification in the table must be documented in an appropriate field with reporting the rationale and source.
- *Update definition.* For each given year, the list of updates impacting on causal and modification table is defined with the description of the rationale and source. This task is reported in a specific input panel (Figure 6, upper part of the update input panel).
- *Encoding.* For each update, editors enter in the lower part of the input panel (Figure 6) the rows of causal and modification tables which must be deleted, added or updated according to the instructions reported in the upper part of the panel. During data typing the system performs online check of the input.
- *Check "within".* After the encoding is completed, a check is run in order to identify possible inconsistencies among encoding rows referring to the same update. In order to carry out this quality control, the system performs normalisation of the encoding (Figure 3). From this point onward all check procedures operate on normalised tables. When this check does not find errors, the normalised encoding rows are stored in a table called update table. Updates will be applied to the historicised causal and modification tables in a later stage.

Phase 2. Table production

This phase is coordinated by the supervisor, who runs the check and updating procedures. In the updating procedures the changes described in the normalised update table are applied by the system to the historicised tables. Before changes become effective, test tables are produced.

- *Check "between".* Update table produced in the previous phase comprise updates entered by different editors. This implies that updates may be incompatible with each other. In order to identify these errors, the supervisor runs the encoding check "between" procedure. This could produce errors that must be revised manually by the supervisor through the correction panel #1. The procedure cannot proceed further until these errors have been corrected.

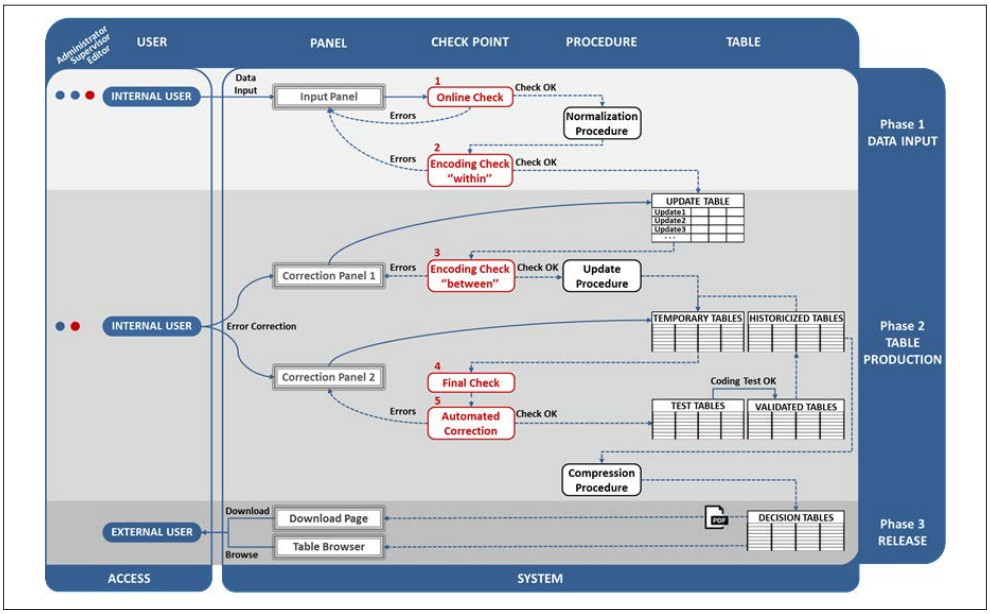
- *Update procedure.* This procedure compares the update table with the historicised tables, applies changes according to the actions specified by the editor and produces a temporary set of updated tables (temporary tables). These must undergo the validation steps described in the following bullets before making the changes effective and stable.
- *Final check and automated correction.* This procedure checks for inconsistencies in the temporary tables. Most errors require a deterministic correction and they are automatically corrected. Others are displayed and must be revised manually by the supervisor through the correction panel #2. The presence of errors stops further processing.
- *Coding test.* Once the updated tables are free of errors, the supervisor can download tables (test tables) for running coding tests that would show the impact of the updated tables on the data and identify possible errors occurred during the update;
- *Validation.* After analyzing the test results the supervisor validates the tables. The validation is a procedure that transfers the changes from the temporary validated tables to the historicised tables.

Phase 3. Release: browse and download

After validation, the tables become available for downloading and browsing. Finally, external users can download the decision tables in pdf format from the "Download" section of the site.

The table update activities are coordinated by a supervisor who controls the transition to the next steps of the process. In particular, the supervisor defines the timing for the termination of the encoding and the starting time for the table updating procedures. Morover, the supervisor can unlock, if necessary, the activities of editors on already implemented updates and re-run updating procedures for a given year.

Figure 5 - DTE workflow overview (a)



(a) Access to the editing part of the system is limited to internal users. Spots in the "ACCESS" area represent user profiles allowed in the different phases; red spots indicate user profiles mainly involved in the specific phases. Manual and automated procedures are represented by solid and dotted arrows respectively. Check points (in red) are ordered by occurrence.

4.4 Table editing

Editors enter updates by translating text instructions into relationships between ICD codes. To do this, DTE provides an input panel where editors can document single updates by specifying a unique identifying name, textual recommendation, source and implementation date. The input panel is also equipped with an encoding panel where editors can specify the relationships between codes to be added, deleted or modified in the tables for the implementation year. Therefore, besides the transformation of text into relationships between codes, the editor must specify which actions should be performed for each specified row, namely addition, deletion or modifying.

In Figure 6 the general structure of the input panel is shown and a practical example of manual encoding is also provided. Referring to Figure 4, WHO recommends the implementation in 2013 of an update to the volume 2 of the ICD10. The instruction is:

"The following infectious diseases should not be accepted as due to any other disease or condition, except when reported as due to human immunodeficiency virus [HIV] disease, malignant impairing the immune system:

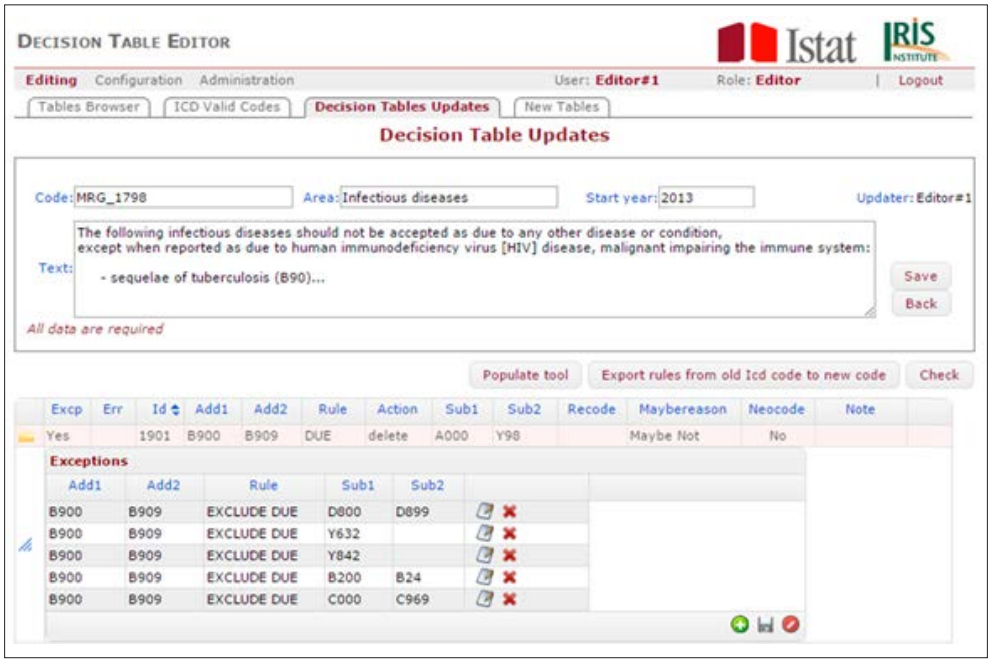
• sequelae of tuberculosis (B90)…"

This instruction includes a statement and the related exception. The editor has to manually encode both of them row by row.

The encoding of the statement *"sequelae of tuberculosis should not be accepted as due to any other disease or condition…"* implies a "delete" action as the address *"should not be accepted as due to"* the subaddress so the relationship must be deleted from the tables. In case of affirmative statement (e.g. *"can be due to"*), action field would be set to "add".

The exception to the previous statement *"…except when reported as due to human immunodeficiency virus [HIV] disease, malignant impairing the immune system"* is encoded as well and rule is automatically set according to the rule entered in the related encoding.

Figure 6 - Update input panel



4.5 Table browser and encoding features

Table browser

To allow retrieval of table data, the DTE system is equipped with a table browser utility. The table browser allows retrieving data from historicised tables which are stored in a normalised structure and returns data in a compressed form. The search can be performed with very flexible criteria such as: year of edition, codesets[6], type of relationship, maybes or recodes. The upper part of the table browser panel allows specifying all the criteria for the search.

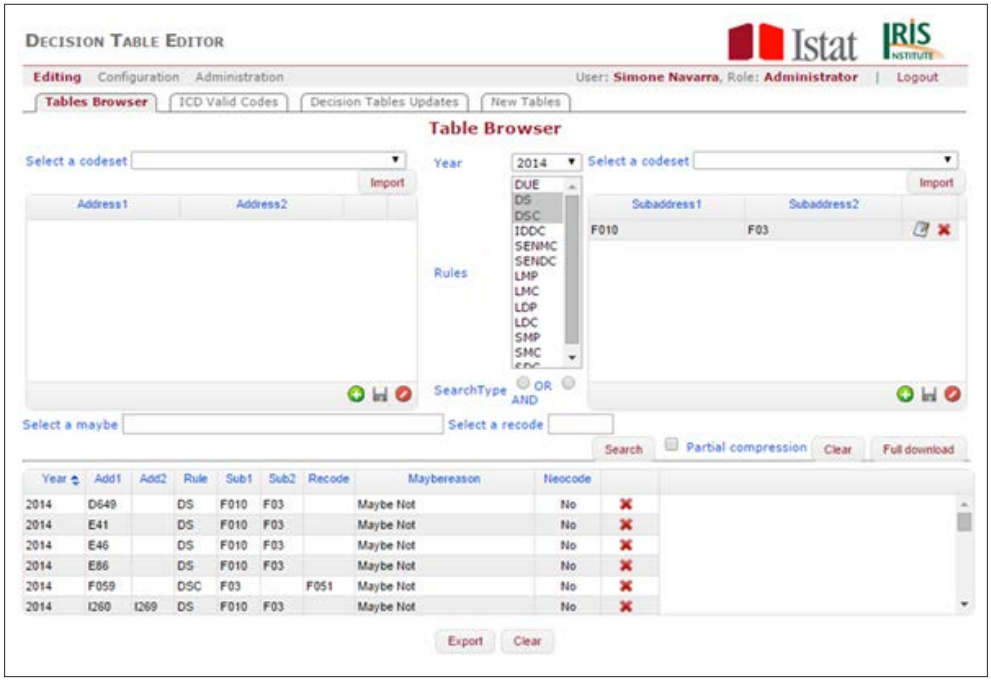
The search results are restituted in the bottom part of the panel in different formats also specified in the criteria panel:

- partial compression (only subaddress is compressed into intervals);
- double compression (both subaddress and address are compressed);
- exported in csv format.

A screenshot of table browser is presented in Figure 7.

6 A codeset is a collection of non-consecutive ICD10 codes which refer to a specific broad group. For example the codeset “dementia” groups codes from different ICD chapters such as F01-F09, G30.

Figure 7 - Table browser (a)



(a) The Figure shows a search on 2014 tables of all conditions that can be considered obvious consequence (rules DS and DSC) of dementia F01-F03. Only a part of the results retrieved are shown in the Figure.

Populate tool

Manual encoding is the simplest way to enter encoding rows. Nevertheless it does not take into account the information of the relationships contained in the actual tables (for instance, whether or not relationships specified in the updates already exist in the tables). In this sense it is a blind update. To avoid this problem a tool is designed for retrieving and modifying existing rows from the tables. As a support to manual encoding, the input panel provides the editors with a populate tool. This instrument is especially useful when large sets of relationships need to be handled at the same time avoiding time-consuming manual data entry.

For example, it may be necessary to modify (according to the WHO update) all the relationships involving a given code or pair of codes. The populate tool allows searching for all these relationships in the existing tables

(the last updated version) and uses them to populate the encoding panel where they can be manually edited. In other cases it may be required to link groups of address and subaddress codes through a given relationship in all possible combinations. The populate tool allows to calculate all these combinations sparing the user the effort to type them one by one in the encoding panel.

The populate tool shares most of the functions with table browser but differs from it in the following features:

- table reference year cannot be selected but it refers to the last available;
- search results are exported to the encoding panel.

An additional tool provided, called rule export, is used when it is necessary to create all the relationship for a code, for instance when an update creates a new code. In these cases, by means of the rule export tool, it is possible to attribute to the new code all the relationships of another code (both in address and subaddress). These are successively exported in the encoding panel for revision.

4.6 Quality control and validation

During the process, many check points have been designed in order to ensure the quality of the information produced. When checks are run, the errors discovered are distinguished into:

- *hard errors.* They must necessarily be corrected by the operator otherwise the process cannot go to the successive step;
- *soft errors.* They are displayed to the operator but they can be either corrected or accepted;
- *automatically corrected errors.* They are not displayed and they are automatically corrected because the correction is univocal.

Online check

The online check verifies the formal correctness of each row entered in the panel, independently from other rows. It is performed during data input by procedures embedded in the encoding panel. The following aspects of data consistency are checked:

- *code validity:* according to the year of implementation of the update;
- *consistency of the spans.* When a span is reported, the second code in the span must be a successor of the previous;
- *applicability.* Some modalities of the relationship variable (see Prospect 1) can be applied to a restricted set of address codes:
  - IDDC can be used only if the address contains exclusively ill-defined codes, whose list is specified in the valid codes table;
  - SENDC and SENMC can be used only if the address contains exclusively senility codes, whose list is specified in the valid codes table[7];
  - LMC, LMP, LDC, LDP cannot be used for ill-defined and senility codes;
- *recode.* The recode must be specified only for relationships requiring combination (DSC, IDDC, SENMC, SENDC, LMC, LDC, SMC).

When an error occurs during data input, prompt messages are triggered.

Encoding check "within" and "between"

The encoding check is a procedure for verifying the consistency of each encoding row inputted in the database with the others. Therefore it takes into account the overall data input, not the single row. The "within" check is run by the editors and examines the consistency of rows referring to a single encoding panel. The "between" check is run by the supervisor and verifies the consistency of the overall encoding for a given year. Inconsistencies checked in these steps are:

- *contradictory actions (hard error).* Two or more rows contain the same address, subaddress and relationship but action is opposite (i.e. add and delete the same relationship);
- *duplication.* If there is a duplicated combination of address, subaddress and relationship the presence of a maybe reason or not defines whether

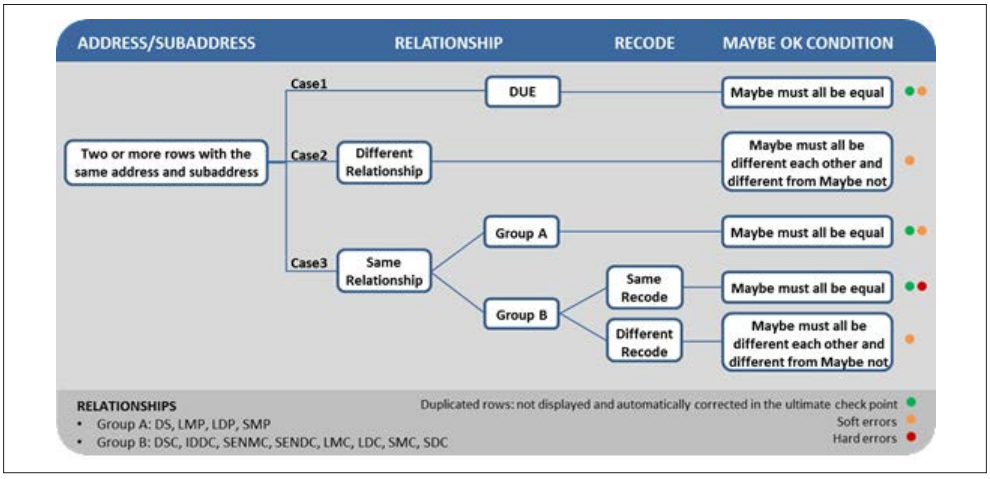
7 From the 2016 edition of the tables the rules SENMC and SENDC have been deleted. For SENDC the IDDC rule has been used, the new rule IDMC has been created to substitute SENMC and for other uses as well.

the error is a soft error, a hard error or an automatically corrected duplication (Figure 8). Rows containing the same address, subaddress, relationship, maybe and recode are considered duplications. They are not shown as errors because they will be automatically corrected in the ultimate check point;

- *maybe reason specification (soft error).* Two or more rows contain different modification relationships and no maybes are specified (more details in Figure 8).

If errors are present, details are displayed in a separate window. Further, inconsistent rows are highlighted in the encoding panel so that the editors can correct them. The encoding check (within) procedure can be repeated several times until all inconsistencies are removed. The encoding of checked updates is closed and the updates are directed to the supervisor (Phase 2) so that editors can no longer modify data unless the supervisor considers necessary to unlock and return them to Phase 1.

Figure 8 - Maybe reason specification check



Final check and automated correction

Changes introduced by the update procedure may be a source of new inconsistencies between rows in the updated tables. Therefore, a set of checks is performed for the following aspects:

- *applicability (hard error).* Described above;
- *symmetry.* This refers to the presence of two rows containing the same relationship where the address and subaddress are interchanged.

  Example:

row #	address	subaddress	relationship
1	A	B	DUE
2	B	A	DUE

  Relationships can be divided into symmetric (DUE and LMC; can display symmetry), and non-symmetric (all the others):

  - IDDC, SENMC, SENDC, LDP, LDC, SMP, SMC and SDC relationships must not display symmetry (hard error);
  - Presence of symmetry for DS, LMP, IDDC and SMC relationships (soft error);
  - LMC relationship must display symmetry (missing rows are automatically inserted);
- *modification relationship (soft error).* A pair of address and subaddress cannot have more than one type of modification relationship (all relationships except for DUE are modification relationships);
- *duplication.* Described above;
- *maybe reason specification (soft error).* Described above.

If errors are present, details are displayed in a separate window and errors are manually revised by the supervisor.

In the very last automated correction the following aspects are checked and errors automatically corrected:

- *duplication (hard error).* Simple combinations of address, subaddress and relationship as well as duplications deriving from previous checks are deleted (only one row is kept);
- *reflexive due to relationship (hard error).* Every code must be linked to itself by DUE relationship. Missing rows with DUE relationship are automatically added;

- *implicit due to relationship (hard error).* A pair of address and subaddress linked by DS or DSC relationship must be linked by DUE relationship as well. Missing rows with DUE relationship are automatically added.

5. Conclusions and future steps

The current version of DTE includes decision tables for the selection of the UC. However, an additional set of tables is designed for a preliminary step of the coding.

The UC selection is only a part of the overall coding process indeed, and in a previous step an ICD code is assigned to all the conditions reported on the death certificate. This task, referred as multiple cause coding, is critical for the successive step of the selection. During the multiple cause coding other information should be taken into account because conditions can get different ICD codes according to variables such as:

- age and gender of decedent;
- interval between onset of diseases and death, when reported;
- manner of death;
- presence and positioning of other conditions on the certificate;
- pregnancy status.

In analogy with the UC selection tables, a set of multiple cause coding tables exist and have been developed as documentation of Iris.

The future development of DTE envisages the inclusion of these tables in order to provide a tool for their systematic management.

This is a step toward the standardisation of multiple cause rules which will result also in better multiple cause data, that will be available for innovative research purposes.

Acknowledgments

DTE has been developed by the Italian National Institute of Statistics - Istat as one of the activities for the collaboration within the Iris Institute, established through an agreement with the German Institute of Medical Documentation and Information (DIMDI), and signed by the two institutions in 2012.

Besides the authors, the development of DTE has been possible thanks to the advice of the other members of the core group of the Iris Institute, namely Robert Anderson, Isabelle Bonellie, Olaf Eckert, Xavier Lavin, Gérard Pavillon and Stefanie Weber.

The software has been developed for Istat by the contractor Top Network S.p.A. In particular, the activities were performed as follows: Angela Ciocci, coordination of the development activities; Claudio Carotenuto, development of the interface; Sara Fiorini, development of database and procedures.

References

Centre for Disease Control and Prevention – CDC, National Center for Health Statistics – NCHS. *About the Mortality Medical Data System*. http://www.cdc.gov/nchs/nvss/mmds/about_mmds.htm.

Centre for Disease Control and Prevention – CDC, National Centre for Health Statistics – NCHS. National Vital Statistics System. 2016. I*nstruction manuals. ICD-10 ACME Decision Tables for Classifying Underlying Causes of Death*.

Eckert, O. 2014. "Improvement of Mortality Statistics by MUSE". Presentation at the *European Conference on Quality in Official Statistics – Q2014*. Vienna, 2-5 June 2014.

Harteloh, P., K. de Bruin, and J. Kardaun. 2010. "The reliability of cause-of-death coding in The Netherlands". *European Journal of Epidemiology*, Volume 25: 531-538.

Iris Institute. *Official website*. www.iris-institute.org.

Pavillon, G., M. Coleman, L.A. Johansson, E. Jougla, and J. Kardaun J. 1998. "Final report on automated coding in member states". Eurostat Working Papers, Population and social conditions. Luxembourg: Office for Official Publications of the European Communities.

Pavillon, G., L.A. Johansson, D. Glenn, S. Weber, B. Witting, and S. Notzon. 2007. "Iris: A Language Independent Coding System For Mortality Data". In WHO – Family of International Classifications Network – FIC. *Annual Meeting.* Trieste, Italy, 28 October – 3 November 2007.

Pavillon, G., and L.A. Johansson. 2012. "The Iris International Coding System Of Causes Of Death". In WHO – Family of International Classifications Network – FIC. *Annual Meeting*. Brasilia, Brazil, 13-19 October 2012.

World Health Organization - WHO. *ICD10 Interactive Self-Learning Tool*. http://apps.who.int/classifications/apps/icd/icd10training/

World Health Organization - WHO. *International Statistical Classification of Diseases and Related Health Problems. 10th Revision, 2010 edition*. Geneva, Switzerland: WHO.

World Health Organization - WHO. *List of Official ICD-10 Updates*. https://www.who.int/standards/classifications/classification-of-diseases/list-of-official-icd-10-updates.

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici e ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti per il perseguimento degli obiettivi della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca Istat". Nel 1999 la collana viene affidata a un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna a essere editore in proprio della pubblicazione.

1B012016003000000

ISSN 1828-1982

€ 15,00