2015-11

Official Statistics Data Integration for Enhanced Information Quality

Dalla Valle, Luciana

10.1002/qre.1859
Quality and Reliability Engineering International
Wiley

Official Statistics Data Integration for Enhanced Information Quality

Luciana Dalla Valle[1] (Plymouth University)
and
Ron S. Kenett (KPA Ltd., University of Turin and NYU Poly)

[1] Corresponding author: luciana.dallavalle@plymouth.ac.uk

ABSTRACT

This work is about integrated analysis of data collected as official statistics with administrative data from operational systems in order to increase the quality of information. *Information quality*, or *InfoQ*, is "the potential of a data set to achieve a specific goal by using a given empirical analysis method". *InfoQ* is based on the identification of four interacting components: the analysis goal, the data, the data analysis, and the utility, and it is assessed through eight dimensions: data resolution, data structure, data integration, temporal relevance, generalizability, chronology of data and goal, construct operationalization and communication. The paper illustrates, through case studies, a novel strategy to increase *InfoQ* based on the integration of official statistics with administrative data using copulas and Bayesian Networks. Official statistics are extraordinary sources of information. However, because of temporal relevance and chronology of data and goals, these fundamental sources of information are often not properly leveraged resulting in a poor level of *InfoQ* in the use of official statistics. This leads to low valued statistical analyses and to the lack of sufficiently informative results. By improving temporal relevance and chronology of data and goals, the use of Bayesian Networks allow us to calibrate official with administrative data, thus strengthening the quality of the information derived from official surveys, and, overall, enhancing *InfoQ*. We show, with examples, how to design and implement such a calibration strategy.

KEYWORDS: Information Quality (InfoQ), Data Integration, Bayesian Networks, Administrative Data, Official Statistics.

1) INTRODUCTION

The concept of *Information quality*, or *InfoQ*, defined by Kenett and Shmueli[1], focuses on assessing the utility of a particular dataset for achieving a given analysis goal by employing statistical analysis or data mining. This concept is more broad and articulated than data and analysis quality. *InfoQ* is based on the identification of four interacting components: the analysis goal, the data, the data analysis, and the utility, and it is assessed through eight dimensions: data resolution, data structure, data integration, temporal relevance, generalizability, chronology of data and goal, construct operationalization and communication. Empirical research should aim at maximizing *InfoQ*, in order to increase the value of statistical analyses, from a methodological as well as from a practical point of view. Ignoring *InfoQ* might lead to results containing no useful information, and therefore to a waste of time, effort and other resources.

The aim of this paper is to introduce a novel methodology designed to enhance *InfoQ* by using copulas and Bayesian Networks to integrate official statistics data with administrative or organizational data. Official statistics are produced by a variety of organizations including central bureaus of statistics, regulatory health care agencies, educational systems, and national banks. They represent important and rich sources of information about many aspects of the citizens' life including health, education, public and private services, as well as about the economic climate, the financial situation and the environment. Official statistics can be exploited not only by public institutions, but also by firms and organizations, in order to compare their performance against their competitors, measure the satisfaction of their customers, explore new markets and identify the most profitable locations to establish new subsidiaries. And these are only a few examples. The integration of official statistics data with other data sets can provide decision makers with high quality information, which we measure through the *InfoQ* concept. A related topic handled in the official statistics literature is small area estimation with external benchmarks[2]. Other papers dealing with this issue include Di Zio et al.[3] and Vicard and Scanu[4]. Here, our concern is with generating *InfoQ* and not only the control of estimation errors of various statistics.

However, due to the lack of effective data integration methodologies, the use of official statistics by decision makers is still rather limited. The recent growth in the number of available data sources and the increase in data quality standards motivates the use of innovative methods to aggregate results obtained from official statistics and from specific datasets in order to obtain reliable and informative analyses.

An example of data integration is provided by Foresti et al.[5], who state that the matching of public with private databases is crucial for implementing new analyses that are functional to a new approach to business. The authors describe the integrated data base maintained by Intesa Sanpaolo Bank in Italy for supporting analytic research requests by management and various decision makers. The bank uses regression models applied to internal data integrated with data from a range of official statistics providers such as:

- Financial statements (CEBI)
- EPO patents (Thomson Scientific)
- Foreign direct investment (Reprint)
- ISO certificates (Accredia)

- Trade-marks (UIBM, OIHM, USPTO, WIPO)
- Credit ratings (CEBI, Intesa Sanpaolo)
- Corporate Group charts (Intesa Sanpaolo)

A more sophisticated approach to data integration is illustrated by Dalla Valle[6], where data from surveys of companies in the north of Italy is combined with official data from the Italian stock exchange so as to calibrate the survey data. The methodology used in data calibration is based on copulas and non-parametric Bayesian networks.

Copulas are defined as multivariate distribution functions with uniform marginals, and are used to model the dependency structure of high-dependent multidimensional datasets. Vine copulas (or vines), in particular, are extremely flexible in high-dimensional cases, allowing the specification of various types of non-linear dependencies. Results from the application of vines are then used to determine the causal effects in non-parametric Bayesian networks (NPBNs). NPBNs require no assumption on the distributions of the marginals, unlike parametric Bayesian belief nets, and allow a straightforward interpretation of the casualties, thanks to their directed structure. Conditionalization of NPBNs can be easily used for calibration, since the graphical representation of these models permits an easy and clear illustration of the flow of influence among variables. For more on Bayesian networks applications to survey data analysis see Kenett and Salini[7,8], Salini and Kenett[9]. For more on Bayesian networks applications in general see Penny and Reale[10], Balin et al.[11], Kenett[12].

The integration of the different sources of information proposed here is performed using Bayesian networks, allowing us the maximization of *InfoQ*.

The remainder of this paper is organized as follows: in Section 2 we introduce NPBNs; in Section 3 we describe the concept of *InfoQ*; Section 4 illustrates the characteristics of the proposed novel methodology; Section 5 is devoted to the application of our methodology to case studies; finally concluding remarks are given in Section 6.

2) INTRODUCTION TO NPBNS

2.1 Copulas and Vines

Copulas, introduced by Sklar in 1959[13], have become very popular in finance, and have been applied to a wide variety of fields, like biology, medicine, social sciences and sampling theory. They allow calculating the joint multivariate distribution from the marginal distributions, incorporating their dependence structure. The main advantage of copulas is their flexibility, with the marginals described by any type of distribution and with various classes of copulas they can accommodate complex dependence patterns and departures from normality.

More formally, given d continuous random variables $X_1, \dots, X_d$, Sklar's theorem states that any joint multivariate distribution $F(x_1, \dots, x_d)$ of a random vector $X = (X_1, \dots, X_d)$ can be

uniquely determined as a copula *C* of its univariate marginals $F_1(x_1), \dots, F_d(x_d)$, via the expression

$$F(x_1, \dots, x_d) = C[F_1(x_1), \dots, F_d(x_d)].$$

Figure 1 depicts a bivariate normal copula density (left panel) and contour plot (right panel) with Gaussian marginals and correlation 0.5.

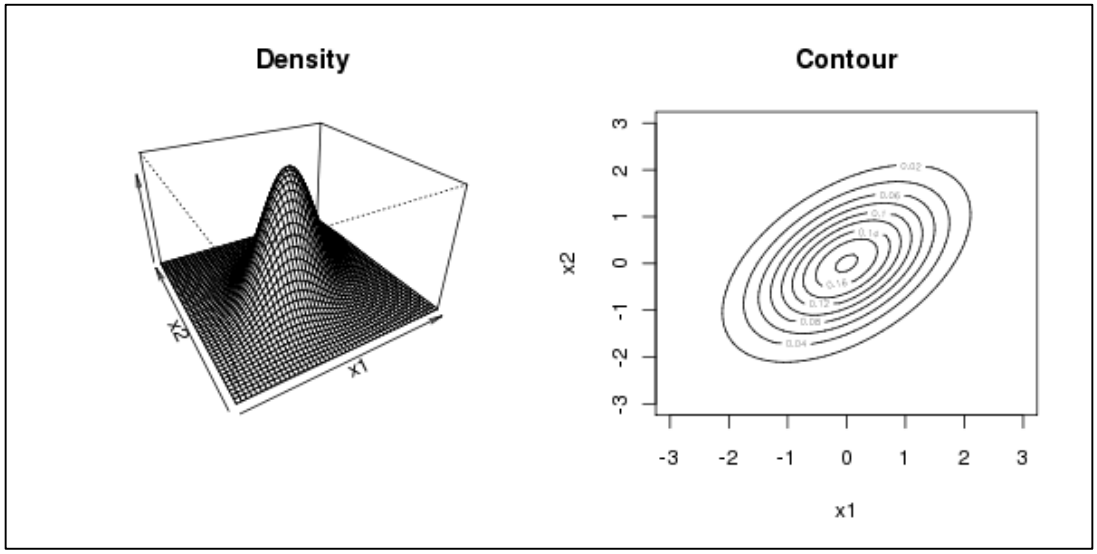

Figure 1: Bivariate normal copula.

However, while in the bivariate case copulas can be effectively used for dependence modeling, in the multivariate case copulas' flexibility is greatly reduced with the choice of families limited to the elliptical copulas (Normal and Student's t). For an overview of the main types of copulas and their characteristics see Joe[14] or Nelsen[15]. Recently pair copula constructions and their graphical representation, called vines, have been introduced by Aas et al.[16] to overcome the lack of flexibility of copulas in high-dimensional cases.
Vines are a flexible class of multivariate copulas based on the decomposition of a multivariate copula using bivariate (conditional) copulas as building blocks[17, 18]. A vine *V(d)* on *d* variables is a nested set of trees (connected acyclic graphs) $T_1, \dots, T_{d-1}$, where the variables are represented by nodes linked by edges, each associated with a specific bivariate copula. The edges of tree T_j are the nodes of tree T_{j+1}, for $j = 1, \dots d-1$.

In a vine, if two edges of a tree T_j share a common node, they are represented in tree T_{j+1} by nodes joined by an edge. Figure 2 illustrates a vine on *d*=3 variables, where the trivariate copula is decomposed into the three bivariate copulas c_{12}, c_{23} and $c_{1,3|2}$, depicted as edges linking the nodes.

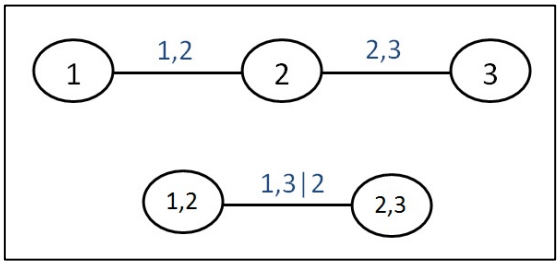

Figure 2: Trivariate vine.

Clearly, the higher the number of variables in a dataset the higher the number of trees in a vine, as seen in Figure 3 showing a vine on 4 variables. For more details about copulas and vines estimation, see Dalla Valle[6].

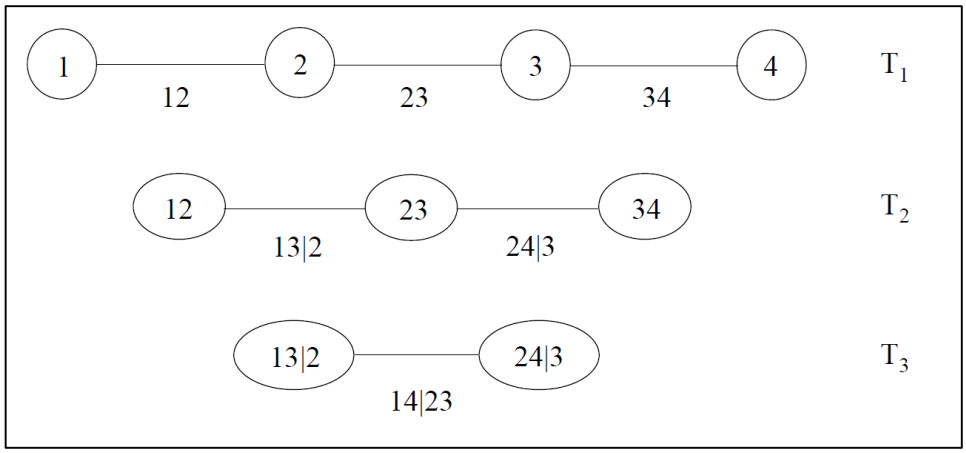

Figure 3: Vine on 4 variables.

The main advantage of copulas and vines is therefore the ability to model complex dependence patterns of variables in a flexible way. To integrate different sources of information we introduce non parametric Bayesian networks (NPBNs), which use vines to evaluate the dependence structure and determine the descriptive causal relationship among variables. We will use NPBNs to perform calibration via conditionalization.

2.2 Non-parametric Bayesian Networks

NPBNs originate from probabilistic graphical models, which represent multivariate densities via a combination of a qualitative graph structure that encodes independencies and local quantitative parameters. Bayesian networks (BNs) are directed acyclic graphs (DAGs) whose nodes represent variables and the edges represent causal relationships between the variables. These variables are associated to conditional probability functions that, together with the DAG, are able to provide a compact representation of high-dimensional distributions. For an introduction and for more details about the definitions and main results see, for example, Cowell[19], Jensen[20, 21], or Pearl[22]; for the use of BNs for problem solving and model building see Fenton and Neil[23].
These models have been applied in official statistics data analysis by Penny and Reale[10] who used graphical models to identify relevant components in a saturated structural VAR model of the quarterly gross domestic product that aggregates a large number of economic

time series. More recently, Vicard and Scanu[4] also applied Bayesian networks to official statistics, showing that the use of post-stratification allows integration and missing data imputation. For general applications of Bayesian networks see Kenett[12].

However the main classes of BNs are discrete, normal or discrete-normal, where discrete BNs are limited to small-sized datasets, and normal BNs are limited by the joint normality assumption. For this reason, researchers proposed alternative methodologies. Elidan[24] points out the need for a synergy between the copula framework and the field of machine learning. Kurowica and Cooke[25] and Hanea et al.[26] introduced continuous NPBNs, using copulas to realize rank correlations in directed acyclic graphs. Their approach is based on nonparametric statistical inference and elicited expert knowledge to understand the dependencies among the variables, and uses conditionalization for diagnosis and prediction.

The direct predecessors of a node, corresponding to a variable, are called *parents*, while the direct descendants of a node are called *children*. The DAG of a NPBN induces a non-unique ordering, and stipulates that each variable is conditionally independent of all predecessors in the ordering given its direct predecessors. The conditional independence statements encoded in the graph permits writing the joint density of d variables as

$$f(x_1, \dots, x_d) = \prod_{i=1}^{d} f_{x_i|Pa_i}(x_i|Pa_i)$$

where $f_{x_j|Pa_j}$, with $i = 1, \dots d$, is the conditional probability function associated to node i, that corresponds to variable X_i, and Pa_i is the set of all i's parents.

Hence, the nodes are associated with continuous invertible distributions, while each edge is represented by a (conditional) rank correlation. For more details, see Kurowica and Cooke[25]. Figure 4 shows a BN on four variables, where conditional rank correlations are listed on the edges.

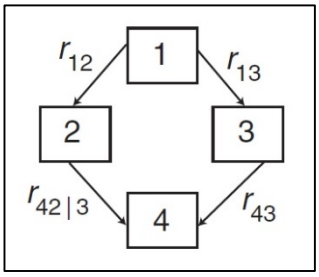

Figure 4. A BN on four variables with conditional rank correlations assigned to edges.

These assignments uniquely determine the joint distribution and allow us to factorize it. For each term of the factorization a vine is built, whose (conditional) rank correlations exactly correspond to those of the NPBN[26].

In the methodology proposed here we consider the (conditional) rank correlations and the marginal distributions needed to completely specify the NPBN as retrieved from data, thus requiring learning and validation. Learning a NPBN from a dataset considers one-dimensional marginal distributions directly from data, assuming that the joint distribution is

modelled by a normal copula. However, it is necessary to test the validity of the model assumptions. Validation involves two steps: validating that the joint normal copula adequately represents the multivariate data, and validating the NPBN as an adequate model of the saturated graph. The former statistical test indicates whether the joint normal copula adequately represents the original data. It compares the dependence structure of the original data with the dependence structure of the normal data. The latter statistical test indicates whether the current model is a good representation of normal data. It compares the dependence structure of the model with the dependence structure of the normal data. A suitable measure of multivariate dependence is the determinant of the rank correlation matrix. Statistical tests for validation of the NPBN are based on this determinant. For more details about NPBN learning and validation, see Hanea and Harrington[27].

3) INTRODUCTION TO INFOQ

As mentioned in the introduction, *InfoQ* is defined in Kenett and Shmueli[1] as the potential of a particular dataset to achieve a particular goal using a given empirical analysis method. The concept is derived mathematically by considering the utility of applying an analytic technology (*f*) to a data resource (*X*) for a given purpose (*g*). Formally the concept of *InfoQ* is defined as:

$$InfoQ(f, X, g) = U(f(X \mid g))$$

InfoQ is determined by four components: *g* (goal definition), *X* (data), *f* (analysis), and *U* (utility measure) as well as by the relationships between *X*, *f*, *g* and *U*. We expand below on these four components and, later, on eight dimensions used to deconstruct *InfoQ* for assessment and planning purposes.

Analysis Goal (g): Data analysis is used for variety of purposes. Three general classes of goals are causal explanation, prediction, and description. Causal explanation includes questions such as "Which factors cause the outcome?" Prediction goals include forecasting future values of a time series and predicting the output value for new observations given a set of input variables. Descriptive goals include quantifying and testing for population effects using data summaries, graphical visualizations, statistical models, and statistical tests.

Data (X): The term "data" includes any type of data to which empirical analysis can be applied. Data can arise from different collection tools: surveys, laboratory tests, field and computer experiments, simulations, web searches, observational studies and more. "Data" can be univariate or multivariate (one or more variables) and of any size (from a single observation in case studies to many observations). It can also contain semantic, unstructured information in the form of text or images with or without a dynamic time dimension. Data is the foundation of any application of empirical analysis.

Data Analysis Method (f): The term data analysis refers to statistical analysis and data mining. This includes statistical models and methods (parametric, semi-parametric, and non-parametric), data mining algorithms, and graphical methods. Operations research methods, such as simplex optimization, where problems are modelled and parametrized, fall into this category as well.

Utility (U): The extent to which the analysis goal is achieved is typically measured by some performance measure. We call this measure "utility". For example, in studies with a predictive goal a popular performance measure is predictive accuracy. In descriptive studies, common utility measures are goodness-of-fit measures. In explanatory models, statistical power and strength-of-fit measures are common utility measures.

To assess or design *InfoQ*, Kenett and Shmueli[1] propose eight dimensions used to deconstruct *InfoQ*. These are: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Chronology of Data and Goal, Generalizability, Operationalization and Communication. We proceed with a brief description of these dimensions.

1) Data Resolution. Data resolution refers to the measurement scale and aggregation level of the data. The measurement scale of the data should be carefully evaluated in terms of its suitability to the goal, the analysis methods used, and the required resolution of the utility *U*. A low rating on data resolution can be indicative of low trust in the usefulness of the study's findings.

2) Data Structure. Data structure relates to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, network data) as well as unstructured, non-numerical data (e.g., text, text with hyperlinks, audio, video, and semantic data). The *InfoQ* level of a certain data type depends on the goal at hand. A low rating on data structure can be indicative of poor data coverage in terms of the project goals.

3) Data Integration. With the variety of data source and data types available today, studies sometimes integrate data from multiple sources and/or types to create new knowledge regarding the goal at hand. Such integration can increase *InfoQ*, but in other cases it can reduce *InfoQ*, e.g., by creating privacy breaches. A low rating on data integration can be indicative of missed potential in data analysis.

4) Temporal Relevance. The process of deriving knowledge from data can be put on a time line that includes the data collection, data analysis, and results' usage periods as well as the temporal gaps between these three stages. The different durations and gaps can each affect *InfoQ*. The data collection duration can increase or decrease *InfoQ*, depending on the study goal, e.g., studying longitudinal effects vs. a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal. A low rating on temporal relevance can be indicative of an analysis with low relevance to decision makers due to data collected in a different contextual condition.

5) Chronology of Data and Goal. The choice of variables to collect, the temporal relationship between them, and their meaning in the context of the goal at hand affects *InfoQ*. A low rating on chronology of data and goal can be indicative of low relevance of a specific data analysis due to misaligned timing.

6) Generalizability. The utility of $f(X|g)$ is dependent on the ability to generalize f to the appropriate population. Two types of generalizability are statistical generalizability and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target

population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern/model *f* to other populations, or applying *f* estimated from one population to predict individual observations in other populations.

7) Operationalization. Two types of operationalization are considered: construct operationalization and action operationalization of the analysis results. Constructs are abstractions that describe a phenomenon of theoretical interest. Measurable data are an operationalization of underlying constructs. The relationship between the underlying construct and its operationalization can vary, and its level relative to the goal is another important aspect of *InfoQ*. The role of construct operationalization is dependent on the goal, and especially on whether the goal is explanatory, predictive, or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as the data are assumed to measure the construct, the variable is considered adequate. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variable is critical. A low rating on operationalization indicates that the study might have academic value but, in fact, has no practical impact.

8) Communication. Effective communication of the analysis and its utility directly impacts *InfoQ*. There are plenty of examples where miscommunication of valid results has led to problematic outcomes. A low rating on communication can be indicative that poor communication might cover the true value of the analysis and, thereby, dump the value of the information provided by the analysis.

4) DATA INTEGRATION METHODOLOGY

The methodology proposed in this paper aims at enhancing *InfoQ* through the integration, via NPBNs, of official statistics data with administrative or organizational data.

Official statistics are statistics published by government agencies or other public bodies such as central banks, national and international organizations and associations. Official statistics information covers different subject areas, such as economics, finance, demographics, health and society. Official statistics data quality is monitored by statistical organizations to produce relevant, objective and accurate statistics aiming at keeping users well informed and assisting good policy and decision-making. For this reason the data production process is carefully monitored by statistical agencies and some official statistics are revised after they have been published. Official Statistics data are usually survey-based, they arise from a scientific design and are professionally estimated by statistical agencies. An example of official statistics used for data calibration is provided in Dalla Valle[6] where the official statistics data is given by the Italian national stock exchange and the organizational data is a survey conducted by a trade association who surveyed its member.

Administrative and organizational data are important and useful resources that contain information which arises via the operation of a transaction, registration or as a record of service

delivery. They cover a wide range of areas, such as education, labour market, health, business and demographics, and they relate specifically to the administration of a system or process and are not primarily generated as research resources. Such data is routinely collected by firms, financial or educational institutions, consulting or marketing agencies and arise from records of a variety of delivered services and registered processes. Administrative data are often stored as electronic records that relate to individuals and/or organisations and summarise certain information for statistical purposes. This data is timely collected with high resolution; but often with data quality issues. For a complete overview on administrative data, see Jones and Elias[28].

Our proposed methodology leads to the maximization of *InfoQ* via data integration of official and administrative information thus enhancing temporal relevance and chronology of data and goal. The idea is in the same spirit of external benchmarking used in small area estimation[2]. In small area estimation benchmarking robustifies the inference by forcing the model-based predictors to agree with a design-based estimator. Similarly, our methodology is based on qualitative data calibration performed via conditioning on graphical models, where official statistics estimates are updated to agree with more timely administrative data estimates.

The proposed methodology is structured in three phases:

Phase 1) Data structure modelling. This phase consists in conducting a multivariate data analysis of respectively the official statistics and administrative datasets, using graphical models such as vines and NPBNs. First, vines are employed to model the dependence structure among the variables, and then the results are used to construct the causal relationships in the NPBN. The model building phase for both the official statistics as well as administrative datasets allows us to establish and visualize the relationships among the variables and their reciprocal influences, identifying clusters of variables with common roles and single or groups of target variables driving the dependencies of the entire dataset. Moreover, this phase enables a comparison of the structures and causal relationships of both datasets, identifying the variables in common and those that are not, but may be incorporated in the subsequent phases of the methodology.

Phase 2) Identification of the calibration link. In the second phase a calibration link, in the form of common correlated variables, is identified between the official statistics and the administrative data. The calibration link is typically represented by a target variable or a group of target variables, common to both datasets and ruling their causal dependencies. The calibration link plays a key role in the entire dataset and its choice depends on the problem under study. Generally it is the most important variable (or group of variables) in the datasets in relation to the analysis goal, or it is a variable whose behavior is particularly important for that specific study. In this second phase the data analyst's experience may be fundamental in the identification of the calibration link.

Phase 3) Performing calibration. In the last phase the NPBNs of both datasets are conditioned on specific "target" variables in order to perform calibration, taking into account the causal relationship among all variables. Conditionalization is performed by "fixing" the

values of one or more target variables, setting them to be equal to the desired figures. Then, the effect of conditionalization on the remaining variables will be easily observed on the NPBN, which incorporates all the causal relationships. The conditioning variables are typically constituted by the calibration link, since understanding its behavior is the focus of the whole analysis. However, the conditioning variables may be different from the calibration link, when the analyst is more interested in the impact of other causal relationships. The calibration is performed when conditionalization of the target variables on one dataset forces the other dataset to agree with it and therefore additional information is brought to the whole analysis. In this phase, the effect of variables that are not common between the official and administrative data, can be observed and incorporated.

We demonstrate the application of these three phases using case studies from education and transportation.

5) THE CASE STUDIES

5.1 The Stella education case study

The *official statistics* dataset used in the first case study was collected by the Italian Stella association. Stella is an inter-university initiative aiming at cooperating and coordinating the activities of supervision, statistical analysis and evaluation of the graduate and post-graduate paths. The initiative includes universities from the north and the center of Italy. The Stella dataset contains information about the post-doctoral placement after 12 months of people who obtained a PhD in 2005, 2006 and 2007. After removing the missing data and selecting only currently employed individuals, we obtain a final dataset with 665 observations and 8 variables. The variables are:

- "yPhD": year of PhD completion
- "ybirth": year of birth
- "unistart": starting year of university degree
- "hweek": working hours per week
- "begsal": initial net monthly salary in euro
- "lastsal": last net monthly salary in euro
- "emp": number of employees
- "estgrow": estimate of net salary rise by 2011 in percent

1) Data structure modelling. We applied a vine copula to the Stella dataset to explore the dependence structure of the marginal distributions. The strongest dependencies are between "begsal" (the initial net monthly salary in euro) and "lastsal" (the last net monthly salary in euro). Moreover, the last salary is associated to the estimate of salary growth and to the number of employees of the company. We notice two groups of correlated variables: one group includes variables regarding the company (begsal, lastsal, estgrow, emp); the other group includes variable regarding the individual (yPhD, ybirth, unistart). The variable "hweek" (working hours per week) is only dependent on "lastsal" conditionally on "emp". The vine model helped to determine the conditional rank correlations, which are necessary to define the corresponding NPBN.

The NPBN represented in Figure 5 is the best network obtained by model validation, where the statistical tests performed on the determinant of the rank correlation matrix proves the validity of the NPBN. The Figure was created using the Uninet software (http://www.lighttwist.net/wp/uninet).
Each node of the NPBN is depicted as a rectangle, with the corresponding variable name on the top and mean ± standard deviation on the bottom. A histogram, or bar chart, of the distribution is plotted in each node.
Different colours of rectangle frames represent different roles of the variables. Here, variables related to the company are depicted in green, variables related to the individual are in yellow, and the calibration link variable (illustrated below) is in purple. This colour code is applied in Figures 5-10.

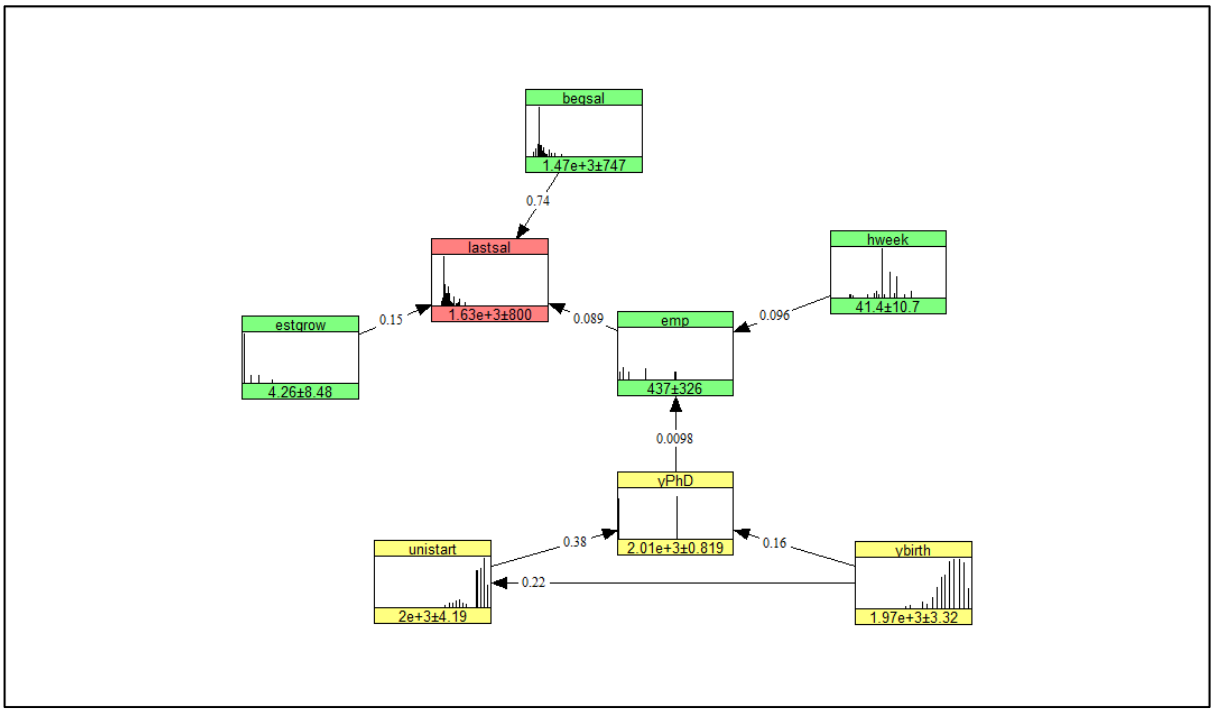

Figure 5. NPBN for the Sella dataset.

2) Identification of the calibration link. The calibration link used here is the "lastsal" variable. This decision was reached by discussion with education experts and is however subjective and context related.

3) Performing calibration. For calibration purposes, the Stella dataset is conditioned on a lower value of "lastsal" (which is reduced from 1630 to 1320), similar to the average salary value of the administrative dataset (the Graduates dataset) described below. In order to reach this lower value of salary, "begsal", "estgrow" and "emp" need to be decreased, as shown in Figure 6, respectively from an average value of 1470, 4.26 and 437, to an average value of 1310, 4.14 and 433.
In Figure 6 and in the following ones, the conditioning variable (the "lastsal" node in Figures 6, 8, 9) is depicted in grey, to denote conditionalization. In each rectangle, the old histogram

(before conditionalization) is represented in light grey, while the new histogram (after conditionalization) is in black. In Figure 7 the "begsal" node has been enlarged, to show the difference between the two histograms in more detail.
This step is essential to obtain a clear picture of the characteristics of graduate employees from official sources, and these results could be useful to make comparisons with the characteristics of employees graduated from different academic institutions.

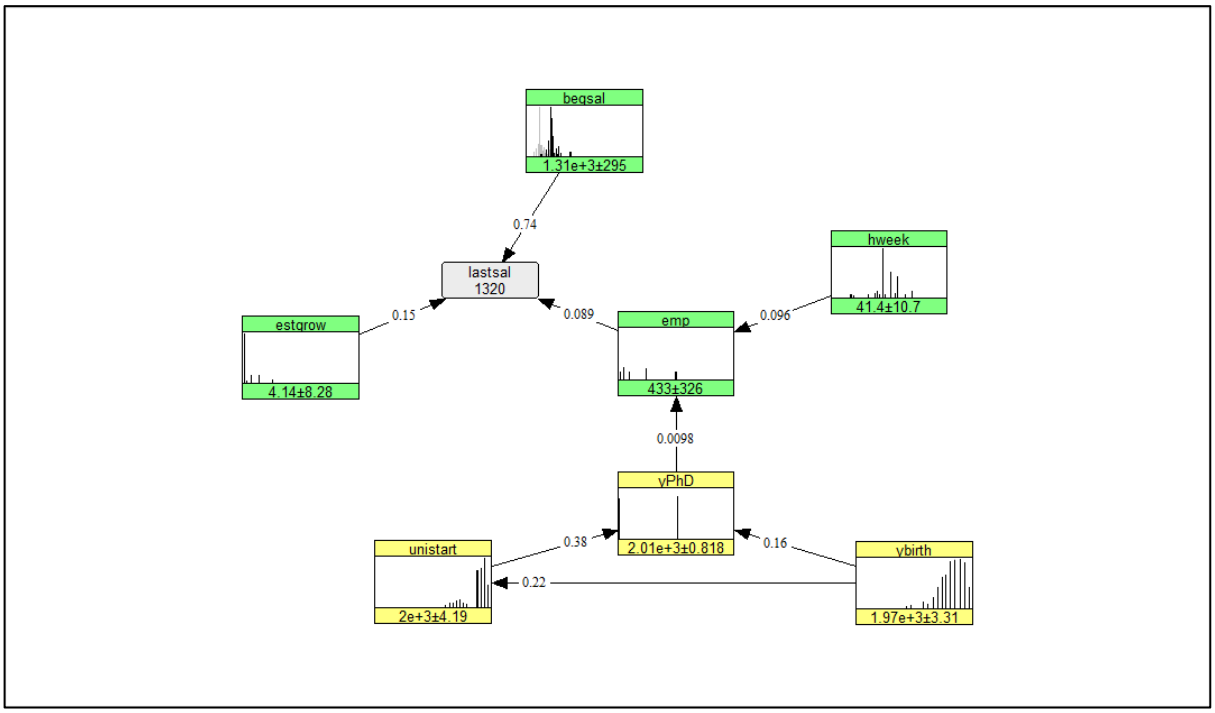

Figure 6: The Stella NPBN is conditionalized on a value of "lastsal" which is similar to the salary value of the Graduates dataset.

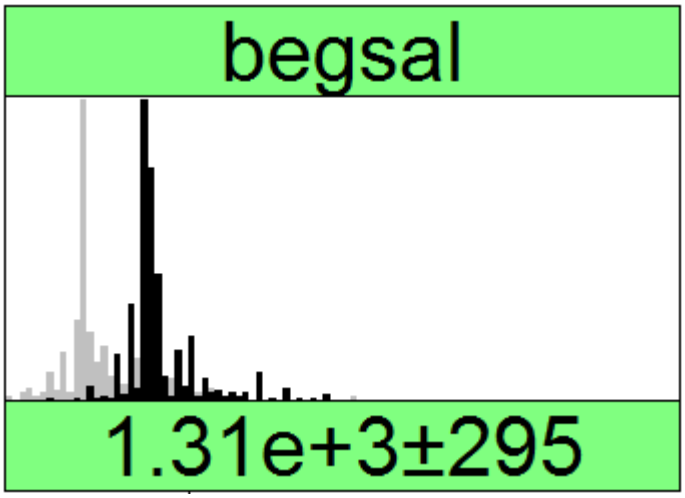

Figure 7: Enlarged "begsal" node of the conditionalized Stella NPBN.

In Figure 8 the Stella NPBN is conditioned on a very low value of "lastsal", equal to 600. From the results we see that graduates with a low salary received also a reduced starting

salary, their estimated salary growth is low, and they are typically employed in smaller companies. On the contrary, graduates receiving a high salary generally started with a higher initial salary, their estimated salary growth is high and they are employed in larger companies (Figure 9).

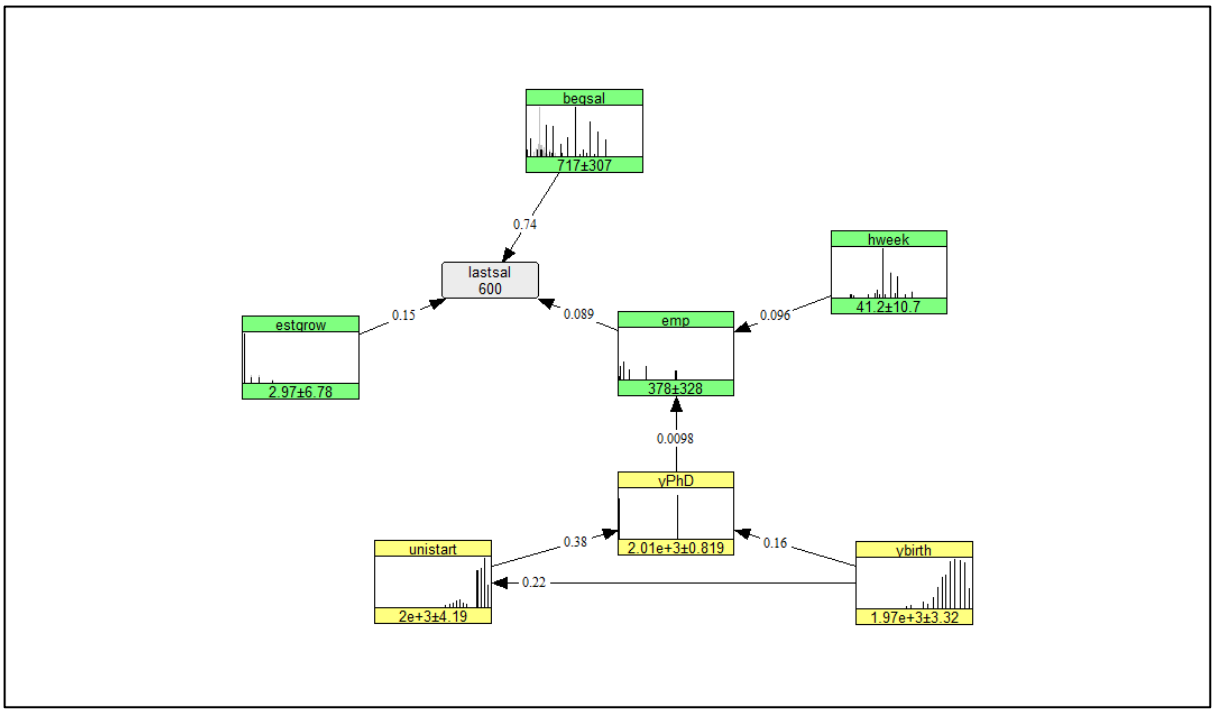

Figure 8: The Stella NPBN is conditionalized on a very low value of "lastsal".

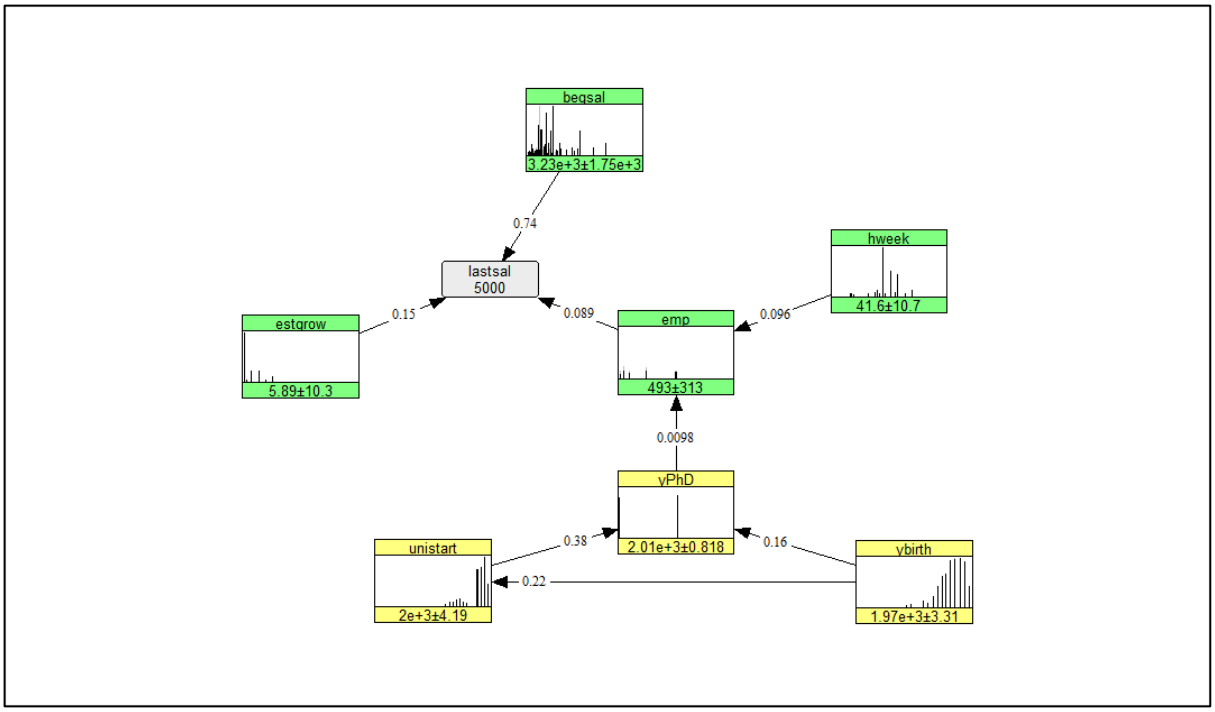

Figure 9: The Stella NPBN is conditionalized on a very high value of "lastsal".

Furthermore, the Stella dataset is conditioned on a low value of "begsal" and "emp" and for a high value of "yPhD" (respectively 1200, 253 and 2007). The variable "yPhD" is considered as a proxy of the starting year of employment, used in the Graduate dataset (Figure 10). In this case the number of employees and the starting year of employment mirror the characteristics of the Graduates dataset and suggest that the initial salary should also be decreased, to obtain a "lastsal" value comparable with the corresponding average value of the Graduate dataset. Therefore, official statistics data can be used to integrate missing information from administrative datasets, such as the "begsal" variable.

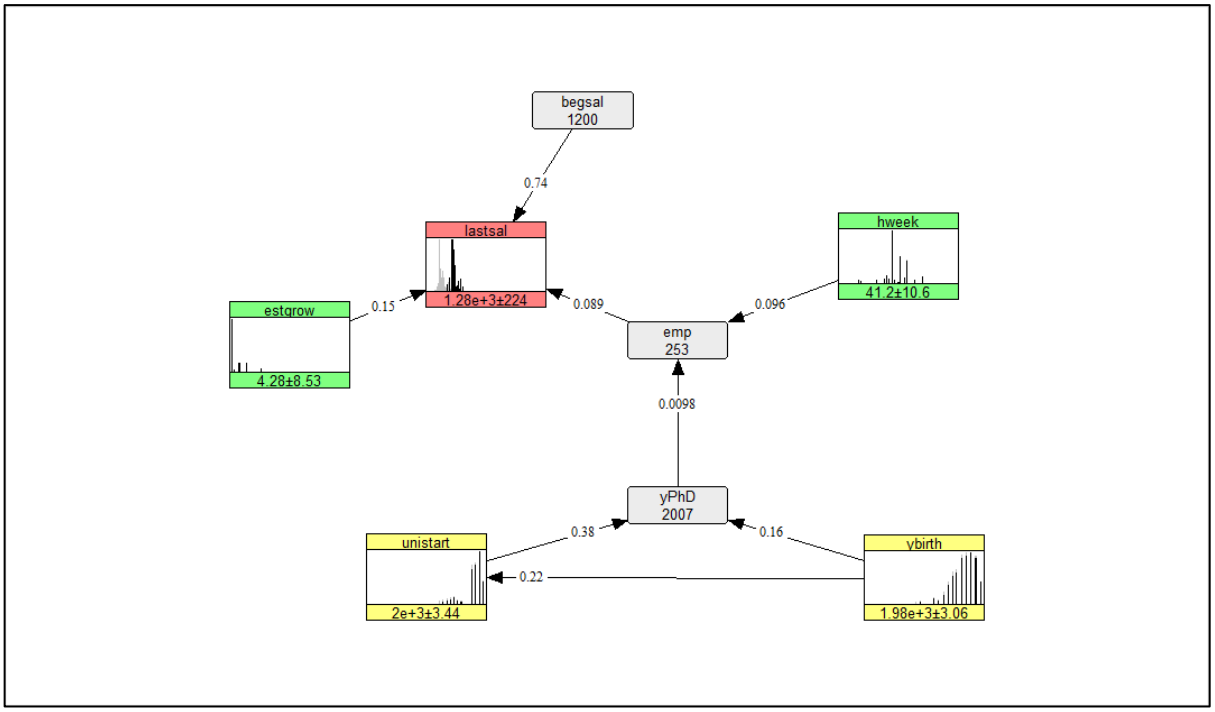

Figure 10: The Stella NPBN is conditionalized on a low value of "begsal" and "emp" and for a high value of "yPhD".

The *administrative dataset* of the Stella education case study contains information collected through an internal small survey conducted locally in few universities of Lombardy, in North of Italy, and in Rome. The dataset is called "Graduates". The sample survey on university graduates' vocational integration is based on interviews on graduates who attained the university degree in 2004. The survey aims at detecting graduates' employment conditions about four years after graduation. From the initial sample we only considered those individuals who are currently employed. After removing the missing values, we obtained a total number of observations of 52. The variables of this dataset are:

- "mdipl": diploma final mark
- "nemp": number of employees
- "msalary": monthly net salary in euros
- "ystjob": starting year of employment

1) Data structure modelling. We applied a vine copula to the Graduates dataset to explore the dependence structure of the marginals. Here the monthly salary is associated to all the remaining variables.

We then applied the NPBN to the Graduates data, and the network represented in Figure 11 is the best one obtained by model validation. We discuss below an interpretation of the colors produced by running the Uninet code. In black and white figures, this discussion should be skipped.
The colour code used for the Graduates dataset (Figures 11-13) is similar to that of the Stella dataset. Again, variables related to the company are depicted in green, variables related to the individual are in yellow, and the calibration link variable is in purple.

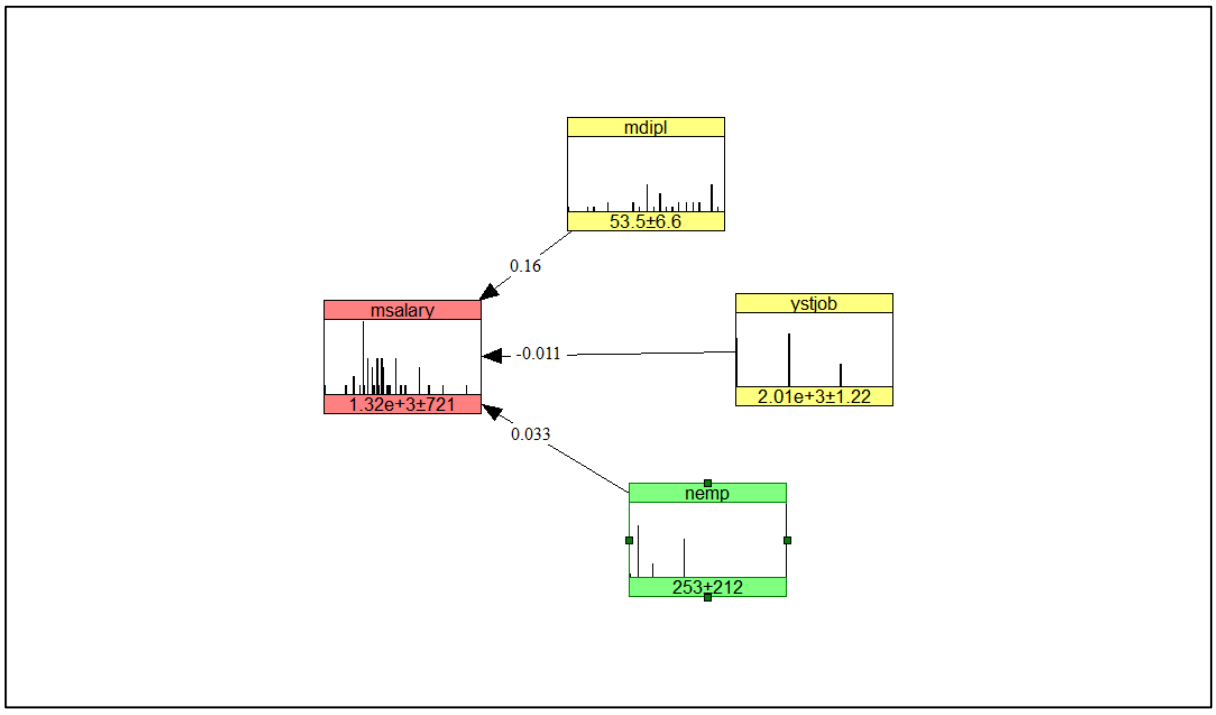

Figure 11. NPBN for the Graduates dataset.

2) Identification of the calibration link. The calibration link is the "msalary" variable, which is analogous to "lastsal" in the Stella official dataset.

3) Performing calibration. For calibration purposes, the Graduates dataset is conditioned on a high value of "msalary" (from 1320 to 1850), more similar to the average salary value of the Stella dataset (Figure 12). Graduates with a higher salary have better diploma final marks and they are employed in larger companies, as we see from the increased average values of "mdipl" and "nemp".

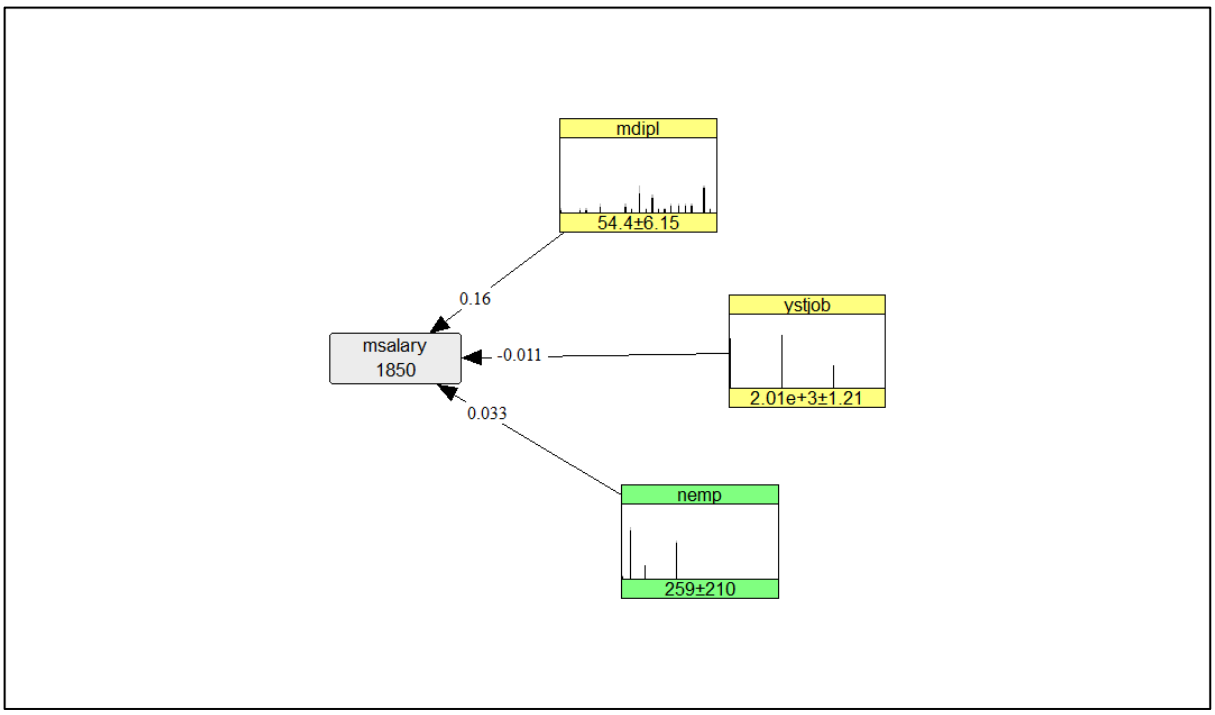

Figure 12: The Graduates NPBN is conditionalized on a high value of "msalary".

Finally, the Graduates dataset is conditioned on a high value of "mdipl" and "nemp" and for a low value of "ystjob". In this case the average salary increases to 1400, which is closer to the corresponding value of the Stella dataset (Figure 13). Therefore, through the results illustrated in Figure 13, a specific academic institution may identify which characteristics its graduates need to possess to receive a level of salary comparable to the official statistics salary.

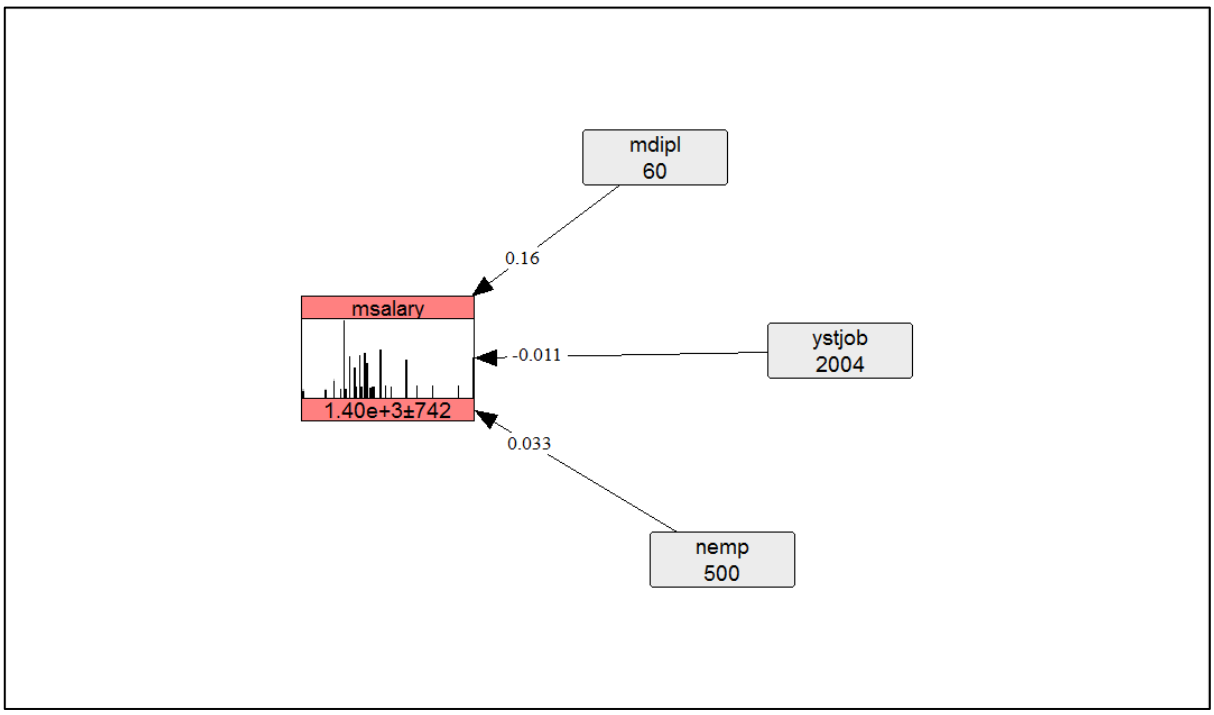

Figure 13: The Graduates NPBN is conditionalized for a high value of "mdipl" and "nemp" and for a low value of "ystjob".

We now describe the assessment of InfoQ for the data integration methodology applied to the Stella education case study, illustrating the InfoQ components and dimensions.

Stella education case study: Info Q components

g: Evaluating the working performance of graduates and understanding the influence on the salary of company-related variables and individual-related variables.
X: Combined survey data with official statistics on education.
f: Use of vines and Bayesian networks to model the dependence structure of the variables and to calculate the conditional rank correlations.
U: Assisting policy makers to monitor the relationship between education and the labor market, identifying trends and methods of improvement.

Stella education case study: The InfoQ dimensions

1) Data Resolution. Concerning the aggregation level, the data is collected at the individual level as in the Graduates as well as in the Stella dataset, aligning with the study goal. The Graduates dataset contains information collected through an internal small survey conducted locally in few Italian universities. Although the data collection should comply with a good standard of accuracy, detailed information about the data collection is not available for this dataset. Stella data are monitored by the authority of a consortium of large number of universities, guaranteeing the reliability and precision of the data produced. The Graduates data are not collected on a regular basis, while Stella produces periodic annual reports about education of a large number of Italian institutions. Considering the analysis goal of

regularly monitor the performance of the graduates in the labor market, there is still room for improvement for the level of *InfoQ* generated by this dimension, especially on the data resolution of the Graduates dataset.

2) Data Structure. The type and structure of education data of both datasets are perfectly aligned with the goal of understanding the influence on the graduates' salary of other information. Although the Stella data integrity is guaranteed by educational institution authorities, the dataset contained a small percentage of missing data, which were removed before implementing the methodology. The Graduates dataset contained a certain number of missing data (which were removed from the dataset), and no information describing the corruptness of the data is available. The level of *InfoQ* could be improved especially relatively to the Graduates dataset.

3) Data Integration. This methodology allows the integration of multiple sources of information, i.e. official statistics and survey data. The methodology performs integration through data calibration, incorporating the dependence structure of the variables using Vines and Bayesian Networks. Multiple datasets integration creates new knowledge regarding the goal of understating the influence of company and individual variables on the graduates' salary, enhancing *InfoQ*.

4) Temporal Relevance. The time gaps between the collection, analysis and deployment of this data should be of short duration and the time horizon from the first to the last phase should not exceed one year, to make the analysis effective for its goals. Stella data are updated annually and made available within few months, while for the Graduates data we do not have specific information about the time period between the data collection and deployment. Considering the goal of assisting policy makers to annually monitor the labor market and to allocate education resources, this dimension produces a reasonably good level of *InfoQ*, which could be increased by a more timely availability of the Graduates survey data. Moreover, the analysis of 2009 could be made relevant to recent years (e.g. 2014) by calibrating the data with dynamic variables, which would allow us to update the study information, enhancing *InfoQ*.

5) Chronology of Data and Goal. Vines allow to calculate associations among variables and to identify clusters of variables. Moreover, NPBNs allow predictive and diagnostic reasoning though the conditionalization of the output. Therefore the methodology is highly effective to reach the goal of identifying and understanding the causal structure between variables.

6) Generalizability. The diagnostic and predictive capabilities of Bayesian Networks provide generalizability to population subsets. The Graduates survey is generalized by calibration with the Stella dataset to a large population including a good number of universities. However, we could still improve *InfoQ* by calibrating the data with variables referring to other institutions, in order to make the study fully generalizable at a national level.

7) Operationalization. The methodology allows monitoring the performance on the labor market of graduates, describing the causal relationships between the salary and variables related to individuals and companies. Moreover, via conditionalization, it allows us to calibrate the results on education obtained by small surveys with the results obtained by official sources. Therefore, the outputs provided from the model are highly useful to policy makers. The use of a model with conditioning capabilities provides an effective tool to set up improvement goals and to detect weaknesses in the education system and in its relationship with industries.

8) Communication. The graphical representations of vines and NPBNs are particularly effective to communication purposes as to a technical as well as non-technical audience. The visual display of a Bayesian Network makes it particularly appealing to decision makers who feel uneasy with mathematical models, producing a high *InfoQ* level.

Stella education case study	
InfoQ Dimension	***Score***
Data Resolution	3
Data Structure	3
Data Integration	5
Temporal Relevance	3
Chronology of Data and Goal	5
Generalizability	4
Operationalization	5
Communication	5
InfoQ Score	***33/40***

Table 1: InfoQ dimension scores for the Stella education case study.

Table 1 lists the scores of each InfoQ dimension for the data integration methodology applied to Stella education case-study. This assessment is subjective following consultation with experts and is based on the personal involvement of the authors in this project. As discussed above, the proposed methodology improves several infoQ dimensions, such as Data integration, Temporal relevance and Chronology of Data and Goal, enhancing the overall InfoQ level.

5.2 NHTSA transport safety case study

The *official statistics* data of the second case study is the Vehicle Safety dataset. The National Highway Traffic Safety Administration (NHTSA), under the U.S. Department of Transportation, was established by the Highway Safety Act of 1970, as the successor to the National Highway Safety Bureau, to carry out safety programs under the National Traffic and Motor Vehicle Safety Act of 1966 and the Highway Safety Act of 1966. NHTSA also carries out consumer programs established by the Motor Vehicle Information and Cost Savings Act of 1972 (website: www.nhtsa.gov). NHTSA is responsible for reducing deaths, injuries and economic losses resulting from motor vehicle crashes. This is accomplished by setting and enforcing safety performance standards for motor vehicles and motor vehicle equipment, and through grants to state and local governments to enable them to conduct effective local highway safety programs. NHTSA investigates safety defects in motor

vehicles, sets and enforces fuel economy standards, helps states and local communities to reduce the threat of drunk drivers, promotes the use of safety belts, child safety seats and air bags, investigates odometer fraud, establishes and enforces vehicle anti-theft regulations and provides consumer information on motor vehicle safety topics. NHTSA also conducts research on driver behaviour and traffic safety, to develop the most efficient and effective means of bringing about safety improvements. The dataset considered here contains information about the effect of car crashes on the human body and has 1241 observations and 14 variables, after removing the missing data, and it refers to cars manufacturers between the late eighties and the early nineties. The variables are:

- "HIC": Head Injury, based on the resultant acceleration pulse for the head centre of gravity
- "T1": Lower Boundary of the time interval over which the HIC was computed
- "T2": Upper Boundary of the time interval over which the HIC was computed
- "CLIP3M": Thorax Region Peak Acceleration, it is the maximum 3-millisecond 'clip' value of the chest resultant acceleration
- "LFEM": Left Femur Peak Load Measurement, indicates the maximum compression load for the left femur
- "RFEM": Right Femur Peak Load Measurement, indicates the maximum compression load for the right femur
- "CSI": Chest Severity Index
- "LBELT": Lap Belt Peak Load Measurement, indicates the maximum tension load on the lap belt
- "SBELT": Shoulder Belt Peak Load Measurement, indicates the maximum tension load on the shoulder belt
- "TTI": Thoracic Trauma Index, for a dummy it is computed from the maximum rib and lower spine peak accelerations
- "PELVG": the pelvis injury criterion, is the peak lateral acceleration on the pelvis
- "VC": Viscous Criterion
- "CMAX": Maximum Chest compression
- "NIJ": Neck Injury Criterion.

1) Data structure modelling. We applied a vine copula to the Vehicle Safety dataset to explore the dependence structure of the marginals. There are strong dependencies among the majority of the variables, e.g. between "CMAX" (Maximum Chest compression) and "CSI" (Chest Severity Index). The variable "CLIP3M" (Thorax Region Peak Acceleration) is only dependent on "CMAX" (Maximum Chest compression) conditionally on "LBELT" (Lap Belt Peak Load Measurement). The conditional rank correlations are determined by the vine model, and are used to define the corresponding NPBN. Figure 14 represents the NPBN for the Vehicle Safety data, and it is the best network obtained by model validation. Calibration link variables (described below) are depicted in yellow. This colour code is applied in Figures 14-16.

2) Identification of the calibration link. The calibration links are the "HIC", "CSI", "LFEM" and "RFEM" variables, which are analogous, respectively, to "Head_IC",

"Chest_decel", "L_Leg" and "R_Leg" in the administrative dataset (named Crash Test), which will be described below.

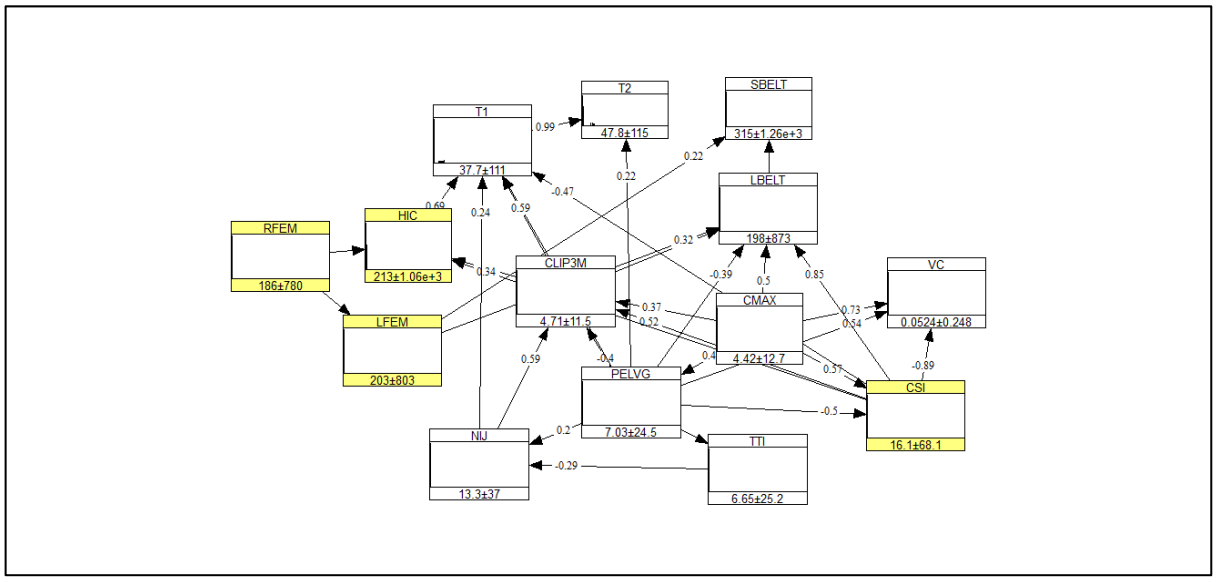

Figure 14. NPBN for the Vehicle Safety dataset.

3) Performing calibration. The Vehicle Safety dataset is conditioned on a low value of "RFEM" (reduced from 186 to 161), and a slightly higher value of "CLIP3M" (from 4.7 to 5). When changing the right femur and thorax region load values, "HIC" (head injury) decreases from 213 to 178, while "CSI" (chest severity index) increases from 16 to 25, becoming very similar to the corresponding values of the Crash Test dataset (Figure 15).

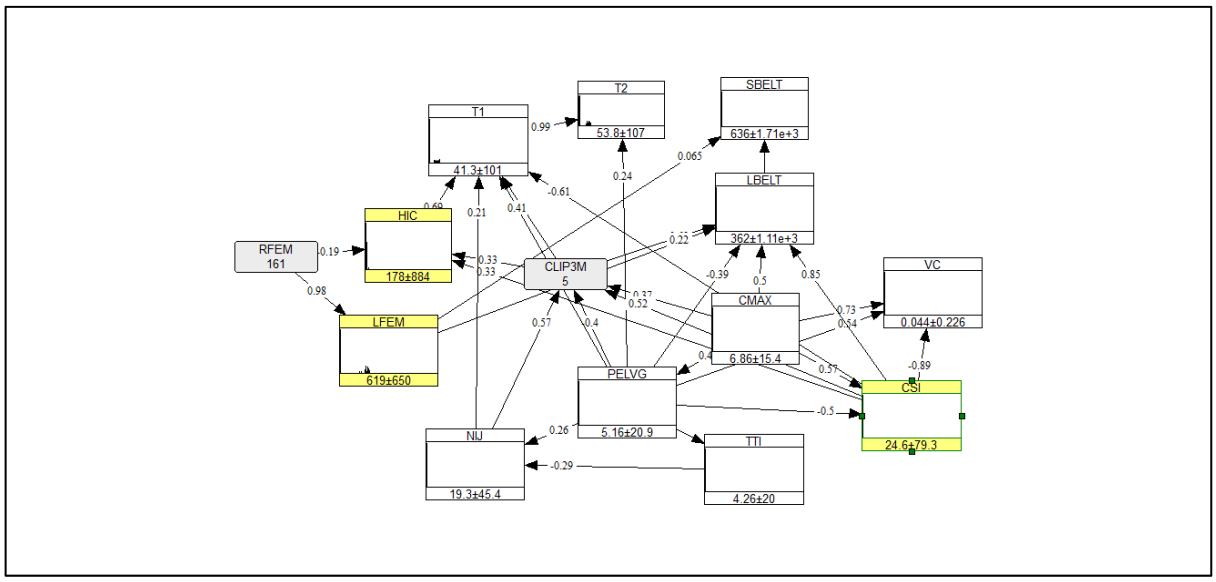

Figure 15: The Vehicle Safety NPBN is conditionalized on a low value of "RFEM" and high value of "CLIP3M" obtaining similar results to the Crash Test dataset.

Furthermore, the Vehicle Safety dataset is conditioned on a high value of "CSI" (from 16 to 22) and for a low value of "HIC" (from 213 to 158) (Figure 16). In this case, the left and right femur loads decrease (respectively from 283 to 152 and from 186 to 136), becoming closer to the corresponding average values of the Crash test dataset.

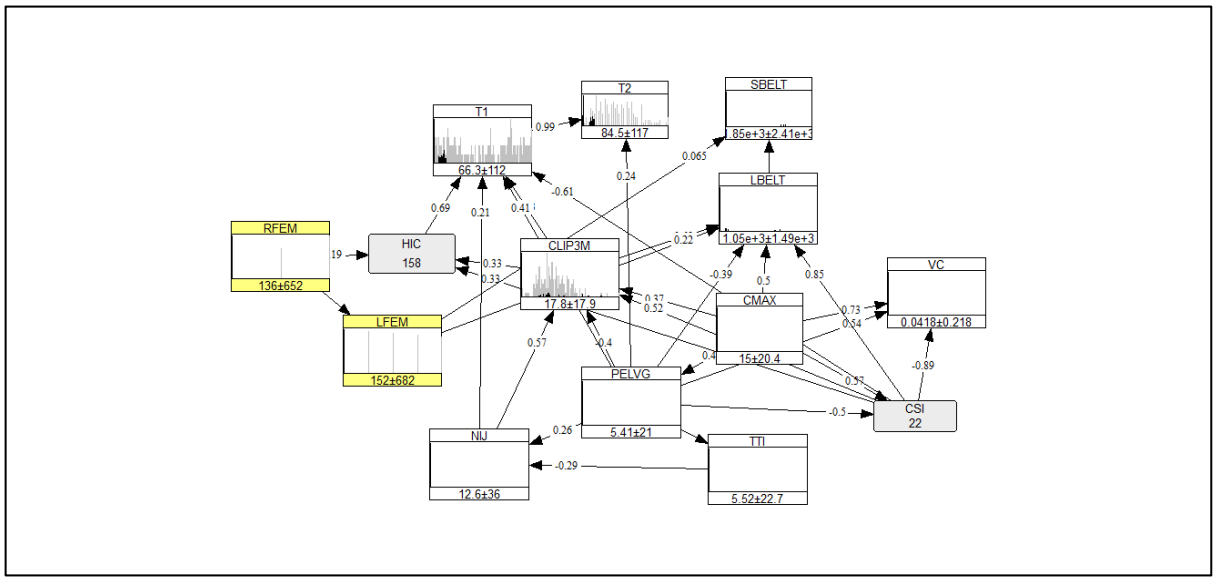

Figure 16: The Vehicle Safety NPBN is conditionalized on a high value of "CSI" and a low value of "HIC".

The results illustrated in Figures 15 and 16 may be useful to compare vehicle official safety data with a specific vehicle manufacturer's data, identifying areas of improvements.

The *administrative dataset* of the NHTSA transport safety case study is a small sample of 176 observations about vehicle crash tests collected by a car manufacturer company for marketing purposes (Crash Test dataset). Vehicles containing dummies in the driver and front passenger seats were crashed into a test wall and information was collected recording how each crash affected the dummies. The injury variables describe the extent of head injuries, chest deceleration, and left and right femur load. The data file also contains information on the type and safety features of each vehicle. A brief description of the variables within the data is provided below.

- "Head_IC": Head injury criterion
- "Chest_decel": Chest deceleration
- "L_Leg": Left femur load
- "R_Leg": Right femur load
- "Doors": Number of car doors in the car
- "Year": Year of manufacture
- "Wt": Vehicle weitght in pounds

1) Data structure modelling. We applied a vine copula to the Crash Test dataset to explore the dependence structure of the marginals. Here the chest deceleration is associated to the head injury criterion and to the vehicle weight. The chest deceleration is associated to the left femur load only conditionally on the head injury criterion.

We then applied the NPBN to the Crash Test data. Figure 17 displays the best network obtained by model validation for the Crash test data. The colour code used for the Crash Test dataset is similar to that of the Vehicle Safety dataset. Calibration link variables (described below) are depicted in yellow, while variables relating to vehicles' features are depicted in green. This colour code is applied in Figures 17-19.

2) Identification of the calibration link. The calibration links are the "Head_IC", "Chest_decel", "L_Leg" and "R_Leg" variables, as explained above.

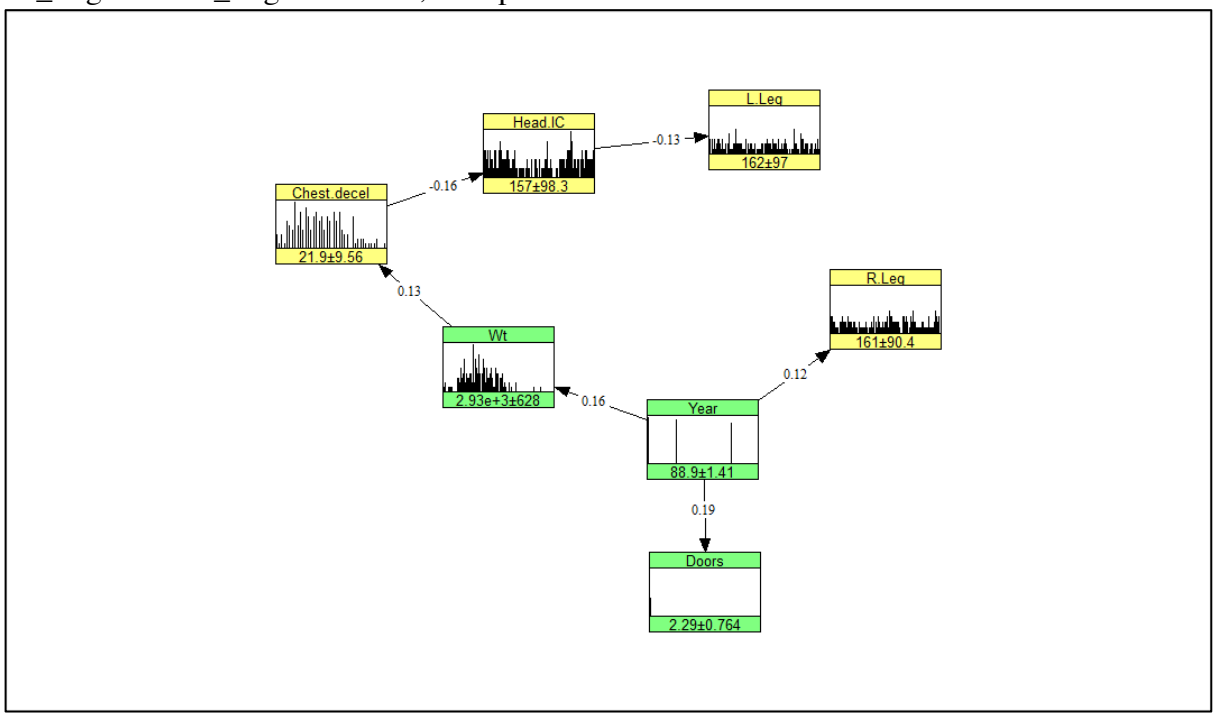

Figure 17. NPBN for the Crash Test dataset.

3) Performing calibration. The Crash Test dataset is conditioned on a high value of "Wt" (from 2930 to 5500) and "Year" (from 1988 to 1991). We notice from the changes in the injury variables that a recent and lighter vehicle causes less severe injuries to the head, but, at the same time, more severe injuries to other parts of the body (Figure 18).

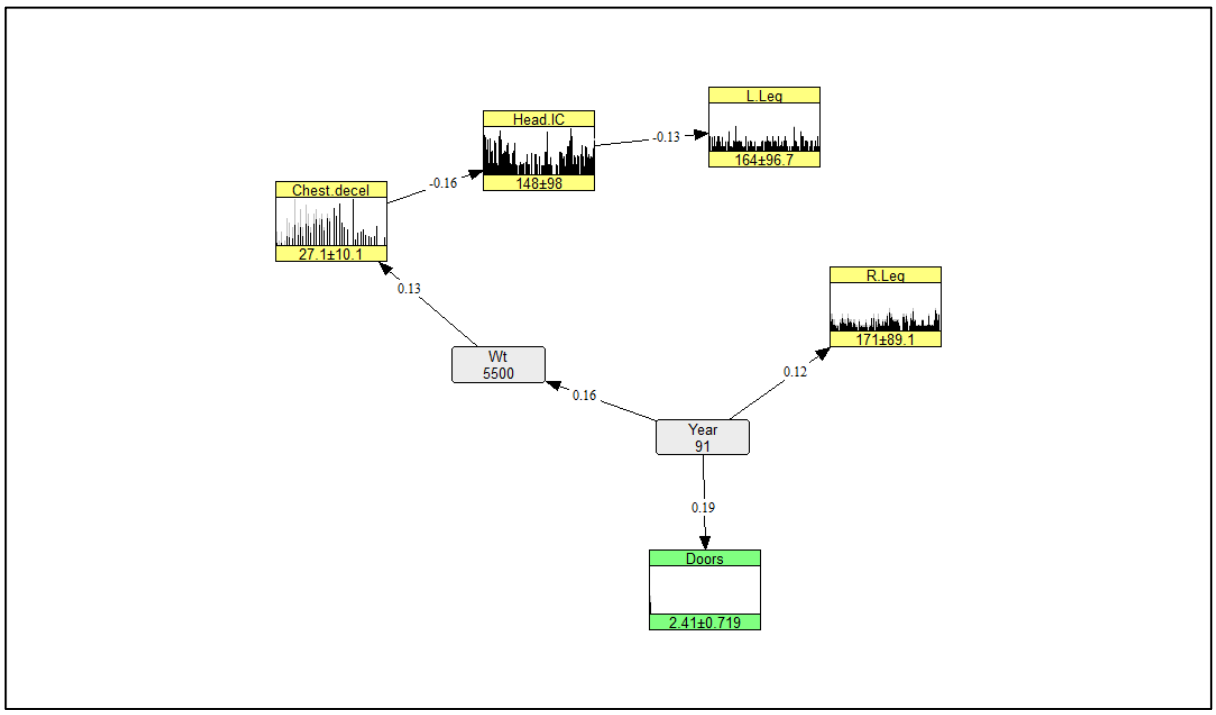

Figure 18: NPBN for the Crash Test dataset is conditioned on a high value of "Wt" and "Year".

Finally, the Crash Test dataset is conditioned on a low value of "Wt" (1600) and "Year" (1987) (Figure 19), obtaining opposite conclusions compared to the case depicted in Figure 18. Therefore, the calibration of the Vehicle Safety and the Crash Test datasets enables us to integrate information contained only in one of the two datasets (such as the car type variables), and thus to obtain a complete description of the relationship between vehicle features and potential injuries.

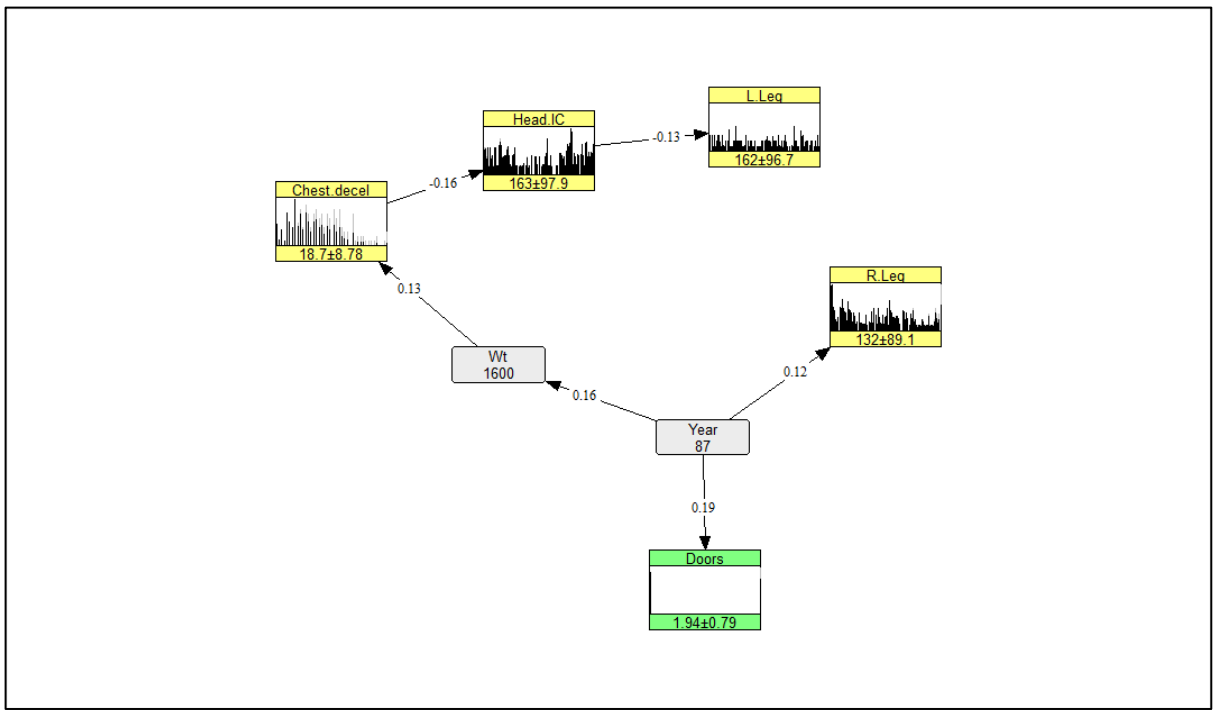

Figure 19: NPBN for the Crash Test dataset is conditioned on a low value of "Wt" and "Year".

We now evaluate InfoQ, through its components and dimensions, for the data integration methodology applied to the NHTSA transport safety case-study.

NHTSA transport safety case study: Info Q components

g: Assessing motor vehicles safety and evaluating the severities of injuries resulting from motor vehicle crashes.
X: Combined small sample of crash test data with motor vehicle safety official statistics data.
f: Use of vines and Bayesian networks to model the dependence structure of the variables and to calculate the conditional rank correlations.
U: Assisting manufacturers and policy makers to set safety performance standards for motor vehicles and motor vehicle equipment, aiming at improving the overall safety of vehicles, and therefore reducing deaths, injuries and economic losses resulting from motor vehicle crashes.

NHTSA transport safety case study: The InfoQ dimensions

1) Data Resolution. Concerning the aggregation level, the data are collected at crash test level, as in the Vehicle Safety as well as in the Crash Test dataset, aligning with the study goal. The Crash Test dataset contains information about vehicle crash tests collected by a car manufacturer company for marketing purposes. The data contain variables measuring injuries of actual crash tests, and are collected following good accuracy standards. Vehicle Safety data are monitored by the US National Highway Traffic Safety Administration (NHTSA), guaranteeing the reliability and precision of the data produced. The Crash test

dataset was produced for a specific study conducted internally by a car company and data are not collected on a regular basis, while the NHTSA produces periodic reports about transport safety of a large range of motor vehicles. Considering the analysis goal of regularly evaluating motor vehicle safety, there is still room for improvement for the level of *InfoQ* generated by this dimension, especially on the data resolution of the Crash Test dataset.

2) Data Structure. The type and structure of car safety data of both datasets are perfectly aligned with the goal of assessing motor vehicles safety and evaluating the severities of injuries resulting from motor vehicle crashes. Although the Vehicle Safety data integrity is guaranteed by the NHTSA, the dataset contained a small percentage of missing data, which were removed before implementing the methodology. The Crash Test dataset did not contain missing data and an exploratory data analysis revealed no corruptness in the data. The level of *InfoQ* for this dimension is therefore high.

3) Data Integration. This methodology allows the integration of multiple sources of information, i.e. official statistics and marketing data. The methodology performs integration though data calibration, incorporating the dependence structure of the variables using vines and Bayesian Networks. Multiple datasets integration creates new knowledge regarding the goal of assessing motor vehicles safety and evaluating the severities of injuries resulting from motor vehicle crashes, enhancing *InfoQ*.

4) Temporal Relevance. The time gaps between the collection, analysis and deployment of this data should be of short duration, to make the analysis effective for its goals. Vehicle Safety data are updated semestrarely and made available within few months, while for the Crash Test data we do not have specific information about the time period between the data collection and deployment. Considering the goal of assisting policy makers to set safety performance standards for motor vehicles and motor vehicle equipment, this dimension produces a reasonably good level of *InfoQ*, which could be increased by a more timely availability of the Crash Test data. Moreover, the analysis of the ’87-’92 cars could be easily made relevant to recent years (e.g. 2014) by updating the data including more recent vehicle models, thus enhancing *InfoQ*.

5) Chronology of Data and Goal. Vines allow to calculate associations among variables and to identify clusters of variables. Moreover, NPBNs allow predictive and diagnostic reasoning though the conditionalization of the output. Therefore the methodology is highly effective to reach the goal of identifying and understanding the causal structure between variables.

6) Generalizability. The diagnostic and predictive capabilities of Bayesian Networks provide generalizability to population subsets. The Crash Test survey is generalized by calibration with the Vehicle Safety dataset to a large range of motor vehicles population, enhancing *InfoQ*.

7) Operationalization. The methodology allows assessing motor vehicles safety and evaluating the severities of injuries resulting from motor vehicle crashes, describing the

causal relationships between injuries in various parts of the body, in case of an accident, and the type of vehicle. Moreover, via conditionalization, it allows us to calibrate the results on vehicle safety obtained by small datasets with the results obtained by official sources. Therefore, the outputs provided from the model are highly useful to policy makers, to set guidelines aiming at improving the overall safety of vehicles. The use of a model with conditioning capabilities provides an effective tool to set up improvement goals and to detect weaknesses in transport safety.

8) Communication. The graphical representations of vines and NPBNs are particularly effective to communication purposes as to a technical as well as non-technical audience. The visual display of a Bayesian Network makes it particularly appealing to decision makers who feel uneasy with mathematical models, producing a high *InfoQ* level.

NHTSA transport safety case study	
InfoQ Dimension	***Score***
Data Resolution	3
Data Structure	4
Data Integration	5
Temporal Relevance	3
Chronology of Data and Goal	5
Generalizability	5
Operationalization	5
Communication	5
InfoQ Score	35/40

Table 2: InfoQ dimension scores for the NHTSA transport safety case study.

Table 2 lists the scores for each InfoQ dimension of the data integration methodology applied to the NHTSA transport safety case study. Again, we need to point out that this assessment is subjective following consultation with experts and is based on the personal involvement of the authors in this project. The proposed methodology maximizes InfoQ, by increasing the scores of Data Integration, Temporal Relevance and Chronology of Data and Goal. The overall InfoQ of this second case study is higher than the InfoQ of the Stella case study, due to a higher score in Data Structure and Generalizability.

6) DISCUSSION AND CONCLUSION

Policy makers, analysts and managers, nowadays, have access to a number of administrative and organizational data such as various types of specific "ad hoc" surveys. However, the results obtained from the statistical analysis of these data, in the light of official statistics results, may be difficult to interpret, and sometimes the outcomes may seem even contradictory. Hence, due to the complexity of data integration, often official statistics are simply ignored, omitting a fundamental source of information, which has indeed the potential of enriching the statistical analysis and produce more valuable results in relation to the goals.

In this paper we introduced a novel methodology to maximize InfoQ. This approach is based on the integration of official statistics data with administrative or organizational data. The integration is performed in three steps.
The first step consists of *data structure modelling* of both datasets, using vine copulas to model the dependencies between the considered variables in a flexible way, allowing us to describe all types of relationships, such as asymmetric and tail dependence. The vine outcomes are then used to build the causal relationships of a NPBN, where the normality constraint of the traditional BN is relaxed, allowing us the use of various types of distributions. NPBNs not only allow for the inclusion of continuous variables in the model, but also of discrete variables, such as "yPhD" in the Stella dataset and "Years" and "Doors" in the Crash Test dataset. As illustrated by Hanea et al.[30], the NPBN approach may be successfully applied to high-dimensional discrete datasets, in contrast to traditional discrete BN models, that become intractable in high-dimensions.
The *identification of the calibration link* is the second step, where common correlated variables are identified in the two datasets. This step is subjective and context-related and the advice of experts is recommended.
Finally, the third step consists in *performing the calibration*, where the NPBNs constructed for the two datasets are conditioned on specific key variables, modifying the value of the calibration links and allowing us an easy comparison of the two datasets.

We applied the proposed data integration methodology to two case-studies: one on education and the other one on transport safety.
The data integration enabled us to acquire a lot of useful information in relation to the goals of the analyses. The education case-study allows decision makers to perform comparisons between the characteristics of graduates from a specific institution with official results, examining the relationship with the labor market. The transport safety case-study enables manufacturers to compare specific vehicle safety data with official data, identifying areas of improvements. Moreover, in both case-studies variables existing in only one of the two calibrated datasets are integrated, providing a more complete illustration of the considered problem.

The case-studies demonstrate that our methodology improves several of the InfoQ dimensions, such as Temporal Relevance and Chronology of Data and Goal, enhancing the overall level of InfoQ.

REFERENCES

1. Kenett, R.S. and Shmueli, G. On information quality. *The Journal of the Royal Statistical Society - Series A* 2014; 177: 3-38.

2. Pfeffermann, D. New Important Developments in Small Area Estimation, *Statistical Science* 2013; 28: 40-68.

3. Di Zio, M., Sacco, G., Scanu, M., Vicard, P. Multivariate techniques for imputation based on Bayesian networks. *Neural Network World* 2005; 4: 303–309.

4. Vicard, P. and Scanu, M. Applications of Bayesian Networks in Official Statistics. In: A. Di Ciaccio, M. Coli & J. M. Angulo Ibanez (Ed.). *Advanced Statistical Methods for the Analysis of Large Data-Sets*, Springer, 2012; 113-123.

5. Foresti, G., Guelpa, F. and Trenti, S. Enterprises in a globalized context and public and private statistical setups. *SIS Scientific Meeting 2012*.

6. Dalla Valle, L. Official Statistics Data Integration Using Copulas, *Quality Technology and Quantitative Management* 2014; 11: 111-131.

7. Kenett, R.S. and Salini S. New Frontiers: Bayesian networks give insight into survey-data analysis, *Quality Progress* 2009; 42: 31-36.

8. Kenett, R.S. and Salini, S. *Modern Analysis of Customer Satisfaction Surveys: with applications using R*, John Wiley and Sons, 2012.

9. Salini, S. and Kenett, R.S. Bayesian Networks of Customer Satisfaction Survey Data, *Journal of Applied Statistics* 2009; 36: 1177-1189.

10. Penny, R.N. & Reale, M. Using graphical modelling in official statistics, *Quaderni di Statistica* 2004; 6: 31-48.

11. Balin, M., Scanu, M. and Vicard, P. Paradata and Bayesian networks: a tool for monitoring and troubleshooting the data production process, *Working paper no. 66*, Dept of Economics, Universita degli Studi Roma Tre, Italy, 2006.

12. Kenett, R.S. On Generating High InfoQ with Bayesian Networks, *Quality Technology and Quantitative Management* 2015, in press.

13. Sklar, M. Fonctions de repartition a *n*-dimensions et leurs marges, *Publications de l'Institut de Statistique de l'Universite de Paris* 1959 ; 8: 229-231.

14. Joe, H. *Multivariate model and dependence concepts*, Monographs on Statistics an Applied Probability, 73, Chapman & Hall, London, 1997.

15. Nelsen, R. B. *An introduction to copulas*, Springer-Verlag, New York, 2006.

16. Aas, K., Czado, C., Frigessi, A. and Bakken, H. Pair-copula constructions of multiple dependence, *Insurance: Mathematics and Economics* 2009; 44: 182-198.

17. Bedford, T. and Cooke, R.M. Probability density decomposition for conditionally dependent random variables modeled by vines, *Annals of Mathematics and Artificial Intelligence* 2001; 32: 245-268.

18. Bedford, T. and Cooke, R.M. Vines - a new graphical model for dependent random variables, *Annals of Statistics* 2002; 30: 1031-1068.

19. Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Sciences, Springer, 1999.

20. Jensen, F. V. *An Introduction to Bayesian Networks*, London: Taylor and Francis, 1996.

21. Jensen, F. V. *Bayesian Networks and Decision Graphs*, Springer 2001.

22. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo: Morgan Kaufman, 1988.

23. Fenton, N and Neil, M. *Risk Assessment and Decision Analysis With Bayesian Networks*, CRC Press, Boca Ranton, USA, 2013.

24. Elidan, G. Copulas in Machine Learning. *Copulae in Mathematical and Quantitative Finance*. Proceedings of the Workshop Held in Cracow 2012. Springer, 2003; 39-60.

25. Kurowicka, D. and Cooke, R. *Distribution-Free Continuous Bayesian Belief Nets*, Proceedings Mathematical Methods in Reliability Conference, 2004.

26. Hanea, A., Kurowicka, D. and Cooke, R. Hybrid Method for Quantifying and Analyzing Bayesian Belief Nets, *Quality and Reliability Engineering International* 2006; 22: 613-729.

27. Hanea, A. and Harrington, W. Ordinal Data Mining for Fine Particles with Non Parametric Continuous Bayesian Belief Nets, *Information Processes Journal* 2009; 9: 280-286.

28. Jones, P. and Elias, P. *Administrative data as a research resource: A selected audit*, London: National Data Strategy, 2006.

29. Kenett, R.S. and Shmueli, G. From Quality to Information Quality in Official Statistics, *Journal of Official Statistics* 2015*, in press.*

30. Hanea, A.M., Kurowicka, D., Cooke, R.M. and Ababei, D.A. Mining and visualising ordinal data with non-parametric continuous BBNs, *Computational Statistics and Data Analysis* 2010, 54: 668-687.