doi.org/10.26398/IJAS.0036-002

# MULTIVARIATE TECHNIQUES FOR ANALYZING AND PRESENTING OFFICIAL STATISTICS INDICATORS

**Ron S. Kenett**
*The KPA Group and the Samuel Neaman Institute, Technion, Israel*

***Abstract:*** *The production of official statistics is experiencing significant challenges including the handling of massive data sets, the application of computer intensive methods and the integration of data from different sources. Official statistics indicators provide a multivariate perspective, both in form and in content. This perspective requires an implementation of multivariate techniques for data analysis and presentation of findings. The information quality framework is a methodological approach that has been applied to many domain areas including the production of official statistics. Bayesian networks are graphical models that permit decision makers to evaluate alternative scenarios using official statistics. The article presents the information quality framework and discusses a Bayesian network application to Eurostat data. It begins with a background on current official statistics evolutionary changes and concludes with a discussion section that maps some of the challenges of official statistics.*

***Keywords:*** *Official statistics, information quality, indicators, multivariate analysis, Bayesian networks.*

## 1. Background

Official statistics need to be used to be useful. Quoting from Forbes and Brown, 2012: "An issue that can lead to misconception is that many of the concepts used in official statistics often have specific meanings which are based on, but not identical to, their everyday usage meaning. All staff producing statistics must understand that … their work translate the real world into models that interpret reality and make it measurable for statistical purposes. The first step … is to define the issue or question(s) that statistical information is needed to inform. That is, to define the objectives for the framework, and then work through those to create its structure and definitions. An important element … is understanding the relationship between the issues and questions to be informed and the definitions themselves." The challenge posed by this quote is a transformation of official statistics from a producer of numbers to a generator of information. This perspective significantly expands the traditional role of official statistics. To fulfill this role, several education programs provide qualified training to producers and users of official statistics. For example, the mission of the European Master in Official Statistics (EMOS) is to enhance the ability of students to understand and analyze European official data at different levels: quality, production process, dissemination, and analysis in a national, European and international context, see EMOS, 2021. As mentioned, a key task of modern official statistics is the generation of information. A general framework for designing and assessing information quality is proposed in Kenett and Shmueli, 2014, 2016a. Kenett and Shmueli, 2016b, provide an example where administrative data, collected for operational purposes, is combined with survey-based data to enhance the information quality of official statistics. The information quality framework consists of eight dimensions and requires an explicit determination of the goals of the analysis and a clarification of the available data, the methods of analysis used and the related utility function. Of impact on information quality of official statistics is the Generic Statistical Business Process Model (GSBPM) which describes statistical production in a general and process-oriented way. It is used both within and between national statistical offices as a common basis for work with statistics production to ensure quality, efficiency, standardization, and process-orientation and is used for all types of surveys, see GSBPM, 2021.

In general, modern statistics, machine learning, data science and in general, data analytics, are having a ubiquitous impact on industry, governments, business, and services (Kenett et al, 2022, 2023a). For an example of how these impact official statistics see Barcaroli, 2017, and Bhandari et al, 2022. For a general treatment of data science and the role of data scientists see Kenett and Redman, 2019. The next section is a high-level introduction to indicators, such as those published by national bureaus of statistics of national statistics organizations.

## 2. Indicators

Indicators come from the Latin word "indicator" that means "who or what indicates". They represent direct and indirect data driven measures. "… an indicator is not simple crude statistical information but represents a measure organically connected to a conceptual model aimed at describing different aspects of reality" (Maggino, 2018a). The construction of indicators involves the process of synthesizing indicators through aggregative–compensative and non-aggregative approaches. These methods apply a synthesis of units with reference to one or more indicators aiming at aggregating individual values at a microlevel. This synthesis allows a comparison of macro units with references of interest. In addition to these "numerical" approaches it is common to use graphical instruments such as dashboards (Maggino, 2018b). In any case, indicators need to be validated. This is sometimes called construct validation. We expand on this when, in the net section, we introduce the information quality dimensions.

An example of indicators is provided by the United Nations Sustainable Development Goals (SDG). This initiative aims at reaching 17 goals that are defined in a list of 169 SDG Targets. Progress towards these Targets is tracked by 232 Indicators. Official statistics indicators are used in a large variety of types and applications. The most popular and politically important indicators are macro-economic statistics, such as GDP, Current Account Balance, Public Deficit, Consumer Price Index, Productivity etc. These indicators typically come from administrative accounts. In areas not covered by macro-economic accounts, such as social statistics, environment, transport, agriculture, education, etc., indicators are mostly collected through surveys (Kenett and Salini, 2012, Eurostat, 2023). Macro-economic indicators often link to scientific theory derived from economic sciences. Non macro-economic indicators typically do not rely on established theory, which raises issues with their interpretation. In general, the aim of an indicator is to provide supporting evidence to decision makers. The quality of indicators is derived from their ability to provide answers to questions posed by decision makers, with the required accuracy, timeliness, consistency, etc.

An important aspect of indicators is that they can provide a multidimensional perspective. This requires proper multivariate display and data analysis (see Kenett and Maggino, 2021). In general, the challenge of transforming numbers to information is significant. In particular, the ability to evaluate alternative scenarios based on official statistics requires methods to analysis counterfactual thought experiments. The next sections introduce a framework for planning and assessing information quality using official statistics followed by an example of a multivariate analysis using Bayesian networks to assess alternative scenarios.

## 3. Information quality

Information quality (InfoQ) is defined as "the potential of a data set to achieve a specific (scientific or practical) goal by using a given empirical analysis method" (Kenett and Shmueli 2014). InfoQ is determined by the data (X), the data analysis method (f) and the analysis goal (g), as well as by the relationships between them. Utility is measured using specific metric(s) (U). Setting a study goals is typically an iterative process (see Kenett et al, 2023b). By examining each of these components, and their relationships, we can learn about the contribution of a given study as a source of knowledge and insight. A mathematical formulation of information quality is: InfoQ = U(f(X|g)). The components of InfoQ have been mapped to eight dimensions that represent a deconstruction of the concept. Here, we present the eight InfoQ dimensions and provide some guiding questions that can be used in planning, designing and evaluating reports based on official statistics.

i) Data Resolution

Data resolution refers to the measurement scale and aggregation level of the data. The data's measurement scale should be carefully evaluated in terms of its suitability to the goal, the analysis methods used, and the required resolution of the utility U. Questions one could ask to figure out the strength of this dimension include:

- Is the data scale used aligned with the stated goal of the study?
- How reliable and precise are the data sources and data-collection instruments used in the study?
- Is the data analysis suitable for the data aggregation level?

A low rating of data resolution is indicative of low trust in the usefulness of the study's findings. An example of data resolution is provided by Google's ability to predict the prevalence of flu based on the type and extent of Internet search queries. These predictions match quite well the official figures published by the Centers for Disease Control and Prevention (CDC). The point is that Google's tracking has only a day's delay, compared to the week or more it takes for the CDC to assemble a picture based on reports from doctors' clinics. Google is faster because it is tracking the outbreak by finding a correlation between what people search for online and whether they have flu symptoms, see Kenett and Shmueli, 2016a.

ii) 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. Questions to ask to figure out the strength of this dimension include:

- Is the type of data used aligned with the stated goal of the study?
- Are data-integrity details (corrupted/missing values) described and handled appropriately?
- Are the analysis methods suitable for the data structure?

A low rating of data structure reflects poor data coverage in terms of the project goals. For example, using a cross-sectional analysis method to analyze a time series warrants special attention when the goal is parameter inference, but is of less concern if the goal is forecasting future values.

iii) Data Integration

With the variety of data sources and data types available today, studies often integrate data from multiple sources and/or types to create new knowledge regarding the goal at hand (Dalla Valle and Kenett, 2015). Such integration can increase InfoQ, but, it can also reduce InfoQ. Data integration is particularly vulnerable to creation of privacy breaches. Questions to ask to figure out the strength of this dimension include:

- If the data integrated from multiple sources, what is the credibility of each source?

- How is the integration performed? Are there linkage issues that lead to dropping crucial information?
- Does the data integration add value in terms of the stated goal?
- Does the data integration cause privacy or confidentiality exposure concerns?

A low rating on data integration is indicative of missed potential in data analysis.

iv) Temporal Relevance

The process of deriving knowledge from data can be placed on a timeline that includes the periods of data collection, data analysis, and usage of results as well as the temporal gaps between these three stages. Such gaps can be due to the employment of independent contractors or internal organizational poor coordination. The different durations and gaps can each affect InfoQ. The data-collection duration can increase or decrease InfoQ, depending on the study goal, for example studying longitudinal effects versus a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal. Questions to ask to figure out the strength of this dimension include:

- Considering the data collection, data analysis and deployment stages, are any of them time-sensitive?
- Does the time gap between data collection and analysis cause any concern?
- Is the time gap between the data collection and analysis and the intended use of the model (e.g., in terms of policy recommendations) of any concern?

A low rating on temporal relevance indicates an analysis with low relevance to decision makers due to data collected in a different contextual condition. This can happen in economic studies with policy implications that are based on old data.

v) 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. Questions to ask to figure out the strength of this dimension include:

- If the stated goal is predictive, are all the predictor variables expected to be available at the time of prediction?
- If the stated goal is causal, do the causal variables precede the effects?
- In a causal study, are there issues of reverse-causation?

A low rating on chronology of data and goal can be indicative of low relevance of a specific data analysis due to misaligned timing. A customer-satisfaction survey, that was designed to be used as input to the annual budget planning cycle, becomes irrelevant if its results are communicated after the annual budget is finalized (Kenett and Salini, 2012).

vi) Generalizability

The utility of $f(X|g)$ is dependent on the ability to generalize $f$ to the appropriate target population. Two types of generalizability are considered: 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 $f$ parameters estimated from one population, to predict individual observations in other populations. Determining the level of generalizability requires careful characterization of $g$. Generalizability is related to the concepts of reproducibility, repeatability, and replicability. Reproducibility represents insights that are replicable (but not necessarily identical), while repeatability is about achieving the same results in a repeated experiment. Replicability is used most often in genome wide association studies where a follow up experiment is conducted to identify a subset of genes as active, after following a large study investigating thousands of genes (Kenett and Shmueli, 2015). Repeatability relates to data quality and analysis quality, while reproducibility relates to InfoQ. Questions to ask to figure out the strength of this dimension include:

- Is the stated goal statistical or scientific generalizability?
- For statistical generalizability in the case of inference, does the paper answer the question "What population does the sample represent?"
- For generalizability in the case of a stated predictive goal (predicting the values of new observations; forecasting future values), are the results generalizable to the data to be predicted?

For more on Generalizability see Kenett and Shmueli, 2016b.

vii) 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. Action operationalizing results refers to three questions originally posed by Edwards Deming (Kenett and Redman, 2019):

- What do you want to accomplish?
- By what method will you accomplish it?
- How will you know when you have accomplished it?

Questions to ask to figure out the strength of construct operationalization include:

- Are the measured variables themselves of interest to the study goal, or is their underlying construct of interest?
- What are the justifications for the choice of variables?

Questions to ask to figure out the strength of operationalizing results include:

- Who can be affected (positively or negatively) by the research findings?
- What can he or she do about it?
- Who else?

A low rating on operationalization indicates that the study might have academic value but has little practical impact.

viii) Communication

Effective communication of the analysis and its utility directly impacts InfoQ. There are plenty of examples where the miscommunication of valid results has led to problematic outcomes. For a study of how to make more understandable National Assessment of Educational Progress (NAEP) and state test score reporting scales and reports, see Hambleton, 2002. Questions that a reviewer should ask to figure out the strength of this dimension include:

- Is the exposition of the goal, data and analysis clear?
- Is the exposition level appropriate for the readership of this report?

A low rating on communication indicates that poor communication might cover the true value of the analysis and, thereby, reduce the value of the information provided by the analysis.

Following this review of the information quality framework we now introduce Bayesian networks with an example. For more examples of applications of information quality to official statistics see Kenett and Shmueli, 2016a, 2016b.

## 4. Bayesian networks

Eurostat, 2023, provides survey-based information on health indicators in 36 countries over the past 20 years. It consists of data on various aspects of people's health status, which enables the analysis of public health issues, demographic patterns, socio-economic trends, and disparities in health statuses. Data on the following aspects are available:

- healthy life years

- self-perceived health and well-being
- functional and activity limitations
- self-reported chronic morbidity
- injuries from accidents
- absence from work due to health problems

Kenett and Salini (2009) showed how Bayesian networks can be used to analyze such survey data and enable the assessment of alternative scenarios. We present here this capability in the context of Eurostat data. This approach has been implemented in a wide range of application domains such as socio-ecological system resilience, see Cai et al (2018) and Adams et al (2022) and education surveys (Pietro et al, 2015). We begin by introducing Bayesian networks.

Bayesian networks (BN) apply a graphical model structure known as a directed acyclic graph (DAG) that is popular in Statistics, Machine Learning and Artificial Intelligence. BNs are both mathematically rigorous and intuitively understandable. They enable an effective representation and computation of a joint probability distribution over a set of random variables (Pearl, 1985, Kenett et al. 2022). The structure of a DAG is defined by two sets: the set of nodes and the set of directed edges. The nodes represent random variables and are drawn as circles labelled by the variable names. The edges represent links among the variables and are represented by arrows between nodes. An edge from node $Xi$ to node $Xj$ represents a relation between the corresponding variables. Thus, an arrow indicates that a value taken by variable $Xj$ depends on the value taken by variable $Xi$. This property is used to reduce, sometimes significantly, the number of parameters that are required to characterize the joint probability distribution (JPD) of the variables. This reduction provides an efficient way to compute posterior probabilities, given the evidence present in the data. In addition to the DAG structure, which is often considered as the "qualitative" part of the model, one needs to specify the "quantitative" parameters of the model. These parameters are described by applying the Markov property, where the conditional probability distribution (CPD) at each node depends only on its parents. For discrete random variables, this conditional probability is represented by a table, listing the local probability that a child node takes on each of the feasible values – for each combination of values of its parents. The joint distribution of a collection of variables is determined uniquely by these local conditional probability tables (CPT). The Eurostat case study presented here is based on discretized variables.

In learning the network structure, one can apply different network learning algorithms like the ones mentioned below in analyzing the Eurostat data. One can also manually include white lists of forced links imposed by expert opinion and black lists, of links that are not to be included in the network, even if the learning algorithm specifies it. In order to learn a BN that fully represents the joint probability distribution it represents, it is necessary to specify, for each node X, the probability distribution for X conditional upon X's parents. The distribution of X, conditional upon its parents, may have any form. Sometimes only constraints on a distribution are known. One can then use the principle of maximum entropy to determine a single distribution, i.e. the one with the greatest entropy given the constraints (Kenett and Salini, 2012). Often these conditional distributions include parameters which are unknown and must be estimated from data, for example using the maximum likelihood approach. When there are unobserved variables direct maximization of the likelihood (or of the posterior probability) is often complex. A classical approach to address this problem is the expectation-maximization (E-M) algorithm which alternates computing expected values of the unobserved variables conditional on observed data, with maximizing the complete likelihood assuming that previously computed expected values are correct. Under mild regularity conditions, this process converges to maximum likelihood (or maximum posterior) values of parameters (Heckerman, 1995).

Causal Bayesian networks are BNs where the effect of an intervention is defined by a ‘do’ operator that separates intervention from conditioning (Pearl, 2009). The basic idea is that an intervention breaks the influence of a confounder so that one can make a true causal assessment. The established counterfactual definitions of direct and indirect effects depend on the ability to manipulate mediators. A BN like graphical representation, based on local independence graphs and dynamic path analysis, can be used to provide an overview of dynamic relations. On the other hand, the econometric approach develops explicit models of outcomes, where the causes of effects are investigated and the mechanisms governing the choice of treatment are analyzed. In such investigations, counterfactuals are studied (Counterfactuals are possible outcomes in different hypothetical states of the world). In general, the study of causality involves: (a) defining interventions or counterfactuals,

(b) identifying causal models from idealized data of population distributions or empirical experiments and (c) identifying causal effects from actual data, where sampling variability is accounted for (Heckman, 2008). We focus here on a BN of 8 indicators from 36 countries with data from 2003 to 2022 (Figure 1). The range in the total number of surveys covered is 8724-9962. Overall, we have 73340 data points derived from the Eurostat surveys. We analyze this data with a BN.

**Frequencies**

| Level | Count | Prob |
|---|---|---|
| People having a long-standing illness or health problem (%) | 9962 | 0.13583 |
| People having a long-standing illness or health problem (%) - Female | 8754 | 0.11936 |
| People having a long-standing illness or health problem (%) - Male | 8754 | 0.11936 |
| Self-perceived health is very good or good (%) - Female | 9244 | 0.12604 |
| Self-perceived health is very good or good (%) - Male | 9244 | 0.12604 |
| Self-perceived long-standing limitations (some or severe) in usual activities due to health problem (%) | 9934 | 0.13545 |
| Self-perceived long-standing limitations (some or severe) in usual activities due to health problem (%) - Female | 8724 | 0.11895 |
| Self-perceived long-standing limitations (some or severe) in usual activities due to health problem (%) - Male | 8724 | 0.11895 |
| Total | 73340 | 1.00000 |

N Missing 0
8 Levels

**Figure 1: The 8 indicators used in the Bayesian network analysis (JMP version 17.0)**

The original data has missing values and we carried out a multivariate imputation preprocessing step to handle this. Following that, all 8 indicators were discretized into 5 categories defined by equal width bins. An algorithmic BN structure analysis with Greedy Thick Thinning produced the DAG shown in Figure 3. An alternative Bayes search gave similar results.

The root of the DAG in Figure 2 consists of percentages of “Male having a long-standing illness or health problem” (top left). The bottom right variable in the DAG is “Female having a long-standing illness or health problem”. This is affected by the Male percentage directly and indirectly by “Female Self-perceived long-standing limitations (some or severe) in usual activities due to health problem”. The BN indicates that 36% of Males and 10% of Females are in the lowest category. In Figure 3 and Figure 4 we condition “Male having a long-standing illness or health problem” to be 100% in the lowest category and 100% in the highest category, respectively. This ability to study the impact of such conditioning is mirroring a mental process conducted informally by decision makers. This is sometimes labeled a “what if” analysis. BNs provide the means to conduct such an analysis in a systematic and reproducible way. Here, we show what would happen if 100% of the "Male having a long-standing illness or health problem" is in the lowest category or if 100% is in the highest category, respectively. These scenarios can represent specific initiatives designed to change the current situation shown in Figure 2.

With the conditioning in Figures 3 and 4, the percent of "Female having a long standing illness or health problem" in the lowest category, increases from 19% to 29%. On the other hand, the lowest category in "Female Self-perceived long-standing limitations (some or severe) in usual activities due to health problem" dropped from 81% to 8%.

These scenarios provide an estimate of the impact of focused interventions, based on past observed data. The example indicates what would be the impact on "Female Self-perceived long-standing limitations (some or severe) in usual activities due to health problem" of changes in the conditions of "Male having a long-standing illness or health problem".

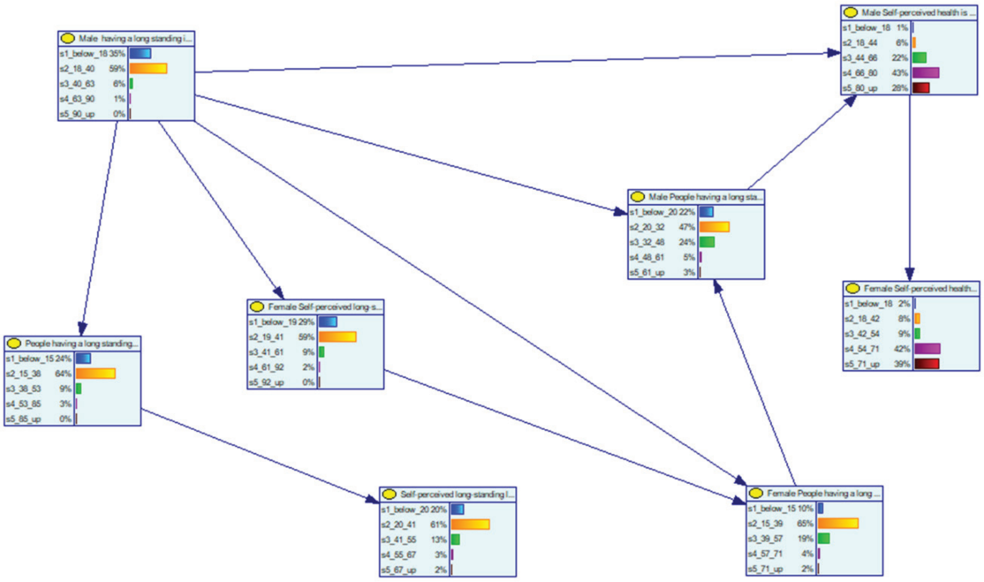


**Figure 2: Bayesian network analysis of the Eurostat data (GeNie version 2.0)**

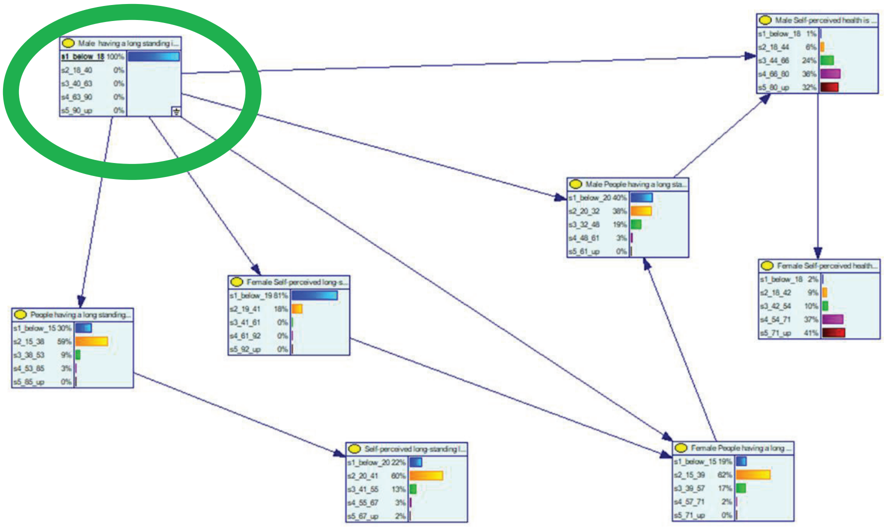


**Figure 3: Low level conditioned Bayesian network analysis of the Eurostat data (GeNie version 2.0)**

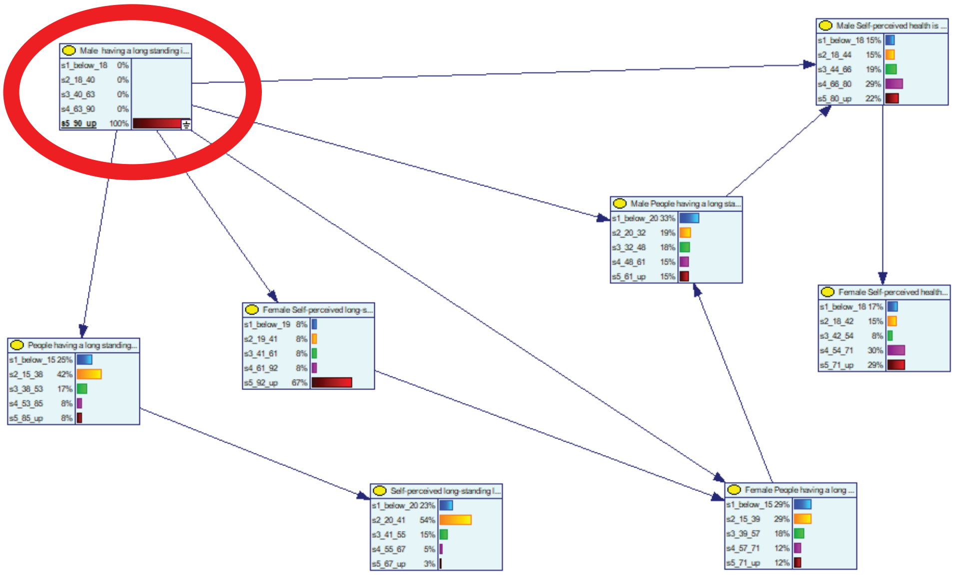


**Figure 4: High level conditioned Bayesian network analysis of the Eurostat data (GeNie version 2.0)**

## 5. Discussion

Official statistics is under strong evolutionary pressures. From a central and unique center of data production national statistics offices meet, on the one hand, alternative data providers and, on the other hand, changing expectations of users and customers. In this paper we touch on several aspects of this transformation and propose possible solutions. We list below some points that deserve more consideration in future work. We expand below on five such directions

i) A central challenge in data rich environments is data integration, the third InfoQ dimension. An example where this is needed in official statistics is in addressing survey mode effects. Surveys are typically conducted, simultaneously, on different platforms. Some surveys are conducted over the phone, some are face to face and some are based on omnibus panels. Integrating data from such sources is a much-needed competency. Dalla Valle and Kenett, 2015, propose a multivariate BN based method to calibrate such assembled data, in order to account for such mode effects.

ii) Official statistics indicators tend to be evaluated using univariate perspectives. This limits the quality of the information that can be provided. We provide an example of a multivariate analysis of survey data using Bayesian networks. Kenett and Salini, 2009, originally proposed it in the context of customer satisfaction surveys, but this also applies to official statistics.

iii) A similar transformation is occurring in the healthcare sector, see Bhandari and Kenett (2022). Both official statistics and healthcare services would benefit from coordinated initiatives, with mutual benchmarking targets.

iv) An area of research that deserves special attention is the development of methodologies for impact studies. The studies can be prospective (ex-ante) or retrospective (ex-post). They can combine observational data with randomized control interventions and case control analysis. Evaluations of ongoing interventions are called formative. Evaluations of past interventions are called summative. More knowledge is needed in conducting such studies.

v) Finally, modern statistics offers an expanded range of analysis methods for inference and predictive analytics, see Kenett et al. 2022, 2023a, 2023b. National bureaus of statistics are typically not involved in analysis and focus on data production. There is however an iterative looping cycle between data collection and data analysis so that both activities cannot be disassociated. This emphasizes the role of national bureaus of statistics as educators of decision makers and the public at large. The more sophisticated the users and producers of official statistics, the better the information generation process.

The paper is designed to map current challenges of national statistics organizations and propose possible approaches to handle them. The information quality framework is presented as a way to address a wide-angle perspective of statistical analysis and Bayesian networks as a multivariate analysis approach that enables an assessment of alternative scenarios. These are only options and, undoubtedly, more such techniques will be offered in the future. The challenge of transforming producers of numbers to generators of information used by decision makers requires both methodological and practical advances.

## ACKNOWLEDGEMENT

The comments and suggestions of Dr. Smadar Lev from the Israel Central Bureau of Statistics, to an early version of this paper, are gratefully acknowledged.

## REFERENCES

Adams, K. J., Macleod, C. A., Metzger, M. J., Melville, N., Helliwell, R. C., Pritchard, J., and Glendell, M. (2022) Developing a Bayesian network model for understanding river catchment resilience under future change scenarios. EGUsphere, 2022, 1-35.

Barcaroli, G. (2017) Improving the quality of official statistics by using alternative data sources: the Istat experience, Big Data in Business and Industry ECAS – ENBIS summer course, September 9-10, "Conservatorio delle Orfane", Terra Murata - Isle of Procida – Italy.

Bhandari, D.R. and Kenett, R.S. (2022) Paradigm shift of statistical big data in healthcare: Management, analysis and future prospects, *Pravaha*, 28(1), 35–44. https://doi.org/10.3126/pravaha.v28i1.57969

Bhandari, D. R., Kenett, R.S., and Ravishanker, N. (2022) Data science for emerging application domains in Nepal. *Official Statistics for Nepal*: Issues and practices, 91, https://nepalindata.com/media/resources/items/20/bOFFICIAL_STATISTICS_OF_NEPAL_ISSUES_AND_PRACTICES.pdf#page=103

Cai, H., Lam, N. S., Zou, L. and Qiang, Y. (2018) Modeling the dynamics of community resilience to coastal hazards using a Bayesian network. *Annals of the American Association of Geographers*, 108(5), 1260-1279.

Dalla Valle, L. and Kenett, R.S. (2015) Official statistics data integration to enhance information quality, *Quality and Reliability Engineering International*, 31 (7), pp. 1281-1300.

EMOS (2021) https://ec.europa.eu/eurostat/cros/content/learning-outcomes-emos-programmes_en, retrieved June 21st, 2021.

Eurostat (2023) Health – Information on data, https://ec.europa.eu/eurostat/web/health/information-data#Health%20status

Forbes, S. and Brown, D. (2012) Conceptual thinking in national statistics offices, *Statistical Journal of the IAOS*, 28, p 89–98.

GSBPM (2021) https://ec.europa.eu/eurostat/cros/content/gsbpm-generic-statistical-business-process-model-theme_en (retrieved 21/6/2021).

Hambleton, R.K. (2002) How can we make NAEP and state test score reporting scales and reports more understandable? In *Assessment in Educational Reform*, edited by R.W. Lissitz and W.D. Schafer. 192–205. Boston, MA: Allyn & Bacon.

Heckerman, D. (1995) A tutorial on learning with Bayesian networks. Microsoft Research tech. report MSR-TR-95-06.

Heckman, J. (2008) *Econometric Causality*. International Statistical Review, 76, 1-27.

Kenett, R.S. and Salini, S. (2009) New frontiers: Bayesian networks give insight into survey-data analysis, *Quality Progress*, 42 (8), pp. 31-36.

Kenett, R.S. and Salini, S. (2012) *Modern Analysis of Customer Satisfaction Surveys: With Applications Using R*. Chichester, UK: John Wiley and Sons.

Kenett, R.S., and Shmueli, G. (2015). Clarifying the terminology that describes scientific reproducibility. *Nature methods*, 12(8), 699-699.

Kenett R.S. and Shmueli G. (2016a*) Information Quality: The Potential of Data and Analytics to Generate Knowledge*. John Wiley and Sons.

Kenett R.S., Shmueli G. (2016b) From quality to information quality in official statistics*. Journal of Official Statistics*, 32:1-22.

Kenett, R.S., and Redman, T. C. (2019) *The Real Work of Data Science: Turning Data into information, Better Decisions, and stronger Organizations.* John Wiley & Sons.

Kenett, R.S. and Maggino, F. (2021) Official Statistics Indicators: challenges in analysis and operationalization, *Journal of Official Statistics*, 37 (2), pp. 541-552.

Kenett, R.S., S. Zacks, S. and Gedeck P. (2022) *Modern Statistics: A Computer-Based Approach with Python*. Switzerland, AG: Springer Nature.

Kenett, R.S., S. Zacks, S. and Gedeck P. (2023a) *Industrial Statistics: A Computer-Based Approach with Python*. Switzerland, AG: Springer Nature.

Kenett R.S., Gotwalt C. and Poggi J.M. (2023b) An analytic journey in an industrial classification problem: How to use models to sharpen your questions. *Quality Reliability Engineering International*, 1-16. https://doi.org/10.1002/qre.3449

Maggino, F. (2018a) Indicators definition, Wiley *StatsRef: Statistics Reference Online*, DOI: 10.1002/9781118445112.stat08095, John Wiley and Sons,

Maggino, F. (2018b) Synthesis of indicators, *Wiley StatsRef: Statistics Reference Online*, DOI: 10.1002/9781118445112.stat08195, John Wiley and Sons.

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" (UCLA Technical Report CSD-850017). Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, 329–334

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, UK.

Pietro, L.D., Mugion, R.G., Musella, F., Renzi, M.F., Vicard, P. (2015). Reconciling internal and external performance in a holistic approach: a Bayesian network model in higher education, *Expert Systems with Applications*, 42(5), 2691–2702.