Data Integration (stat08014)

Luciana Dalla Valle, University of Plymouth, UK *

Abstract

This article introduces some of the most popular techniques of data integration, that allow the combination of information coming from various sources. The illustration focuses, in particular, on the Bayesian generalized Heckman methodology and on the data calibration methodology based on vines and nonparametric Bayesian networks.

Keywords: *Administrative Data, Bayesian Networks, Copula, Data Integration, Heckmans Two-Step Method, Information Quality (InfoQ), Informative Priors, Official Statistics, Small and Medium-sized Enterprizes (SMEs), Vines.*

*email: luciana.dallavalle@plymouth.ac.uk

1 Introduction

Data integration is the process of combining heterogeneous data that originate from different sources, providing a unified view of these information [13]. The growing availability of data in every sector, including business, government and health care, and the increase in information quality standards, required the development of effective data integration methodologies, to obtain reliable and informative analyses. However, in most cases, the efficient aggregation of multiple datasets may be very complex, especially where data come in both structured and unstructured formats and need to be integrated from disparate sources stored in systems managed by different departments (see **stat06775**).

The existing literature on data integration ranges from traditional models based on regression and linear dependencies, to more complex models allowing to express non-linear dependencies and causal relationships. For an overview of the main integration techniques for survey data, see [14] and references therein.

An example of the application of traditional methodologies is provided by Foresti et al. [8], who used OLS to identify the determinants of sales growth, applying it to several integrated private databases. The authors agree that data aggregation from multiple information sources is key to decision-makers and that the matching of public with private databases is crucial for implementing new analyses that are functional to a new approach to business. An integrated databases was created to support analytic research requests by management and various decision makers, combining financial data with data from a range of official statistics providers, such as financial statements, EPO patents, foreign direct investment, ISO certificates, trade-marks, credit ratings and corporate group charts.

Other popular data integration techniques are discussed by Dong and Srivastava [7]. For example, record linkage merges records from different sources that are believed to belong to

the same entity (e.g. a person, household or business). In addition to record linkage, other methodologies to combine diverse information are described, such as schema mapping, data fusion and others.

Data integration techniques can be successfully used for missing data imputation, where estimates are created based on combining information from multiple data sources. These techniques allow us to use model developed on specific data sources to impute information for missing responses in other sources. For example, Di Zio et al. [6], Vicard and Scanu [17] and Chakraborty et al. [2] applied Bayesian Networks (BNs) <stat00258> to impute missing values and to integrate diverse data types, such as official statistics, historic data, survey data, management planning data and expert knowledge.

Bayesian methods (see **stat00207.pub2**, **stat03616**, **stat05825**) could be adopted to integrate data from different sources and to combine estimates across studies (see **stat04124**). The Bayesian approach assumes that prior knowledge about some parameters of interest is represented by a probability distribution (see **stat00243.pub2**, **stat00244**), which is updated after experimental data is observed. The combination of prior and experimental data generates the posterior distribution <stat07165>, upon which inference is based. Bayesian methods are often used in small area estimation, borrowing strengths from various data sources to obtain reliable estimates in subpopulations where the sample size is very small. In particular, benchmarking methods for small area estimates constrain a weighted average of the posterior means to equal prespecified estimates [16].

The remainder of this article illustrates some recent new developments in the literature of data integration.

2 A Bayesian Approach for Data Integration and Self-Selection Bias Modelling

The Bayesian generalized Heckman methodology [3] is a Bayesian approach which allows us to integrate productivity data of small and medium-sized enterprises (SMEs) with official statistics data and, at the same time, to correct self-selection bias (see **stat03338**, **stat03397**, **stat05290.pub2**).

Official statistics are fundamental sources of publicly available information that, unfortunately, is often neglected by SMEs. Bayesian methods allow us to combine official data, expressed as prior information, with other types of data, such as SMEs surveys. This approach is particularly useful when dealing with small samples.

When SMEs data is collected through a survey, one of the most common data quality issues is the high proportion of non-responses, that leads to self-selection bias. Self-selection bias originates when the a subgroup of units do not answer or do not fill in their questionnaires. The main issue with self-selection is that responders differ from non-responders and therefore estimating an effect from only the responders might confound the effect and the choice to respond (see **stat05107**).

Self-selection bias can be corrected applying Heckman's two-step method [9], which considers two equations tied together by a latent factor that allows the missing data associated with the non-responding subjects to be estimated. In particular, Heckman's method can be generalized using copulas <stat00943>, which allow us to assume different distributions for the marginals and to express various dependence structures (see **stat08012**, **stat08013**). This approach overcomes the restrictive assumptions of the original Heckman's approach improving its flexibility and facilitating its adoption to model various dependence structures in the data.

The methodology was used to integrate a national-level survey with an official EU-level survey dataset. The national-level survey included information about new or significantly improved goods or services (product innovations) and new or significantly improved processes, logistics or distribution methods (process innovations), as well as about organizational and marketing innovation. The official EU-level survey provided a comparative assessment of the research and innovation performance of the EU Member States and the relative strengths and weaknesses of their research and innovation systems.

The application of the Bayesian generalized Heckman methodology allowed a proper evaluation of the productivity of SMEs by, on the one hand, removing self-selection bias and, on the other hand, integrating official statistics with other information.

<Figure 1 near here>

Figure 1 shows the histogram of the SMEs average turnover, obtained after data integration. The dashed line represents the true average value of turnover for the observed data, while the dotted line represents the average value of turnover predicted by the traditional OLS model. This result shows that the use of the OLS model in presence of self-selection underestimates the true value of the target variable. The Bayesian generalized Heckman model performs well and accurately predicts the true value of turnover.

3 A Data Integration Approach Based on Vines and Bayesian Networks

Data integration can be performed via calibration using vines and nonparametric BNs (NPBNs).

A vine is a multivariate copula, represented by a graphical model, that is constructed from a set of bivariate copulas, called pair-copulas. More specifically, the copula density is decomposed into a product of pair-copula densities. All these bivariate copulas may be selected completely freely as the resulting structure is guaranteed to be a valid copula. Hence, vines are highly flexible, and able to characterize a wide range of complex dependencies (see **stat08012**, **stat08013**).

Results from the application of vines are used to determine the causal effects in nonparametric NPBNs. The main advantages of the application of NPBNs are that they require no assumption on the distributions of the marginals, and allow a straightforward interpretation of the casualties, thanks to their directed structure. Conditionalization of NPBNs can be easily used for calibration, because the graphical representation of these models permits an easy and clear illustration of the flow of influence among variables [15], [1], [11].

This calibration approach allows to integrate different sources of information, including official statistics, organizational and administrative data [10], thus enhancing the overall information quality (InfoQ) of the study under consideration [12]. The 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.

<Figure 2 near here>

The data integration approach is structured in three phases, illustrated in Figure 2:

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.

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.

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.

This methodology was applied to integrate stock exchange data (official statistics) with a survey conducted by a trade association who surveyed its members (organizational data) [4]. Conditionalizing on the target variable represented by the companies' sales, the values of a specific financial dataset can be integrated with official statistics, for example to determine what characteristics a set of firms should have in order to perform similarly to the companies described in the official data source.

Other interesting applications of this methodology were in the areas of education and transportation [5]. The official statistics dataset used in the education case study contained information about the post-doctoral placement of PhD holders, while the administrative dataset contained information collected through an internal small survey on university graduates' employment conditions four years after graduation. The transportation case study used vehicle safety official information about the effect of car crashes on the human body and administrative information about vehicle crash tests collected by a car manufacturer company for marketing purposes. The application of the calibration methodology to the case study demonstrated that data integration facilitates the acquisition of 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 labour market. The transport safety case study enables manufacturers to compare specific vehicle safety data

with official data, identifying areas of improvements.

Furthermore, the data integration methodology improves several of the InfoQ dimensions, such as temporal relevance and chronology of data and goal, enhancing the overall level of InfoQ.

4 Related Articles

stat00207.pub2

stat00243.pub2

stat00244

stat00258

stat00943

stat03338

stat03397

stat03616

stat04124

stat05107

stat05290.pub2

stat05825

stat06775

stat07165

stat08012

stat08013

References

[1] Balin, M., Scanu, M. & Vicard, P. (2006) 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.

[2] Chakraborty, S., Mengersen, K., Fidge, C., Ma, L., & Lassen, D. (2015) Multifaceted modelling of complex business enterprises. *PloS one*, **10**, e0134052.

[3] Dalla Valle, L. (2016) The Use of Official Statistics in Self-Selection Bias Modeling. *Journal of Official Statistics*, **32**, 887–905.

[4] Dalla Valle, L. (2014) Official Statistics Data Integration Using Copulas. *Quality Technology and Quantitative Management*, **11**, 111–131.

[5] Dalla Valle, L. & Kenett, R. (2015) Official Statistics Data Integration for Enhanced Information Quality. *Quality and Reliability Engineering International*, **31**, 1281–1300.

[6] Di Zio, M., Sacco, G., Scanu, M. & Vicard, P. (2005) Multivariate techniques for imputation based on Bayesian networks. *Neural Network World*, **4**, 303–309.

[7] Dong, X.L. & Srivastava, D. (2015) *Big Data Integration. Synthesis Lectures on Data Management*. Morgan Claypool Publishers.

[8] Foresti, G., Guelpa, F. & Trenti, S. (2012) Enterprises in a globalized context and public and private statistical setups. *SIS Scientific Meeting 2012*.

[9] Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, **47**, 153–161.

[10] Jones, P. & Elias, P. (2006) *Administrative data as a research resource: A selected audit*, London: National Data Strategy.

[11] Kenett, R.S. (2016) On Generating High InfoQ with Bayesian Networks. *Quality Technology and Quantitative Management*, *13*, 309–332.

[12] Kenett, R.S. & Shmueli, G. (2014) On information quality. *The Journal of the Royal Statistical Society - Series A*, **177**, 3–38.

[13] Lenzerini, M. (2002) Data Integration: a theoretical perspective. Proceedings of the 21st ACM-SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. pp. 243–246.

[14] Lohr, S.L. & Raghunathan, T.E. (2017) Combining Survey Data with Other Data Sources. *Statistical Science*, **32**, 293–312.

[15] Penny, R.N., Reale, M. (2004) Using Graphical Modelling in Official Statistics. Quaderni di Statistica, **6**, 31–48.

[16] Pfeffermann, D. (2013) New Important Developments in Small Area Estimation. *Statistical Science*, **28**, 40–68.

[17] Vicard, P. & Scanu, M. (2012) Applications of Bayesian Networks in Official Statistics, in *Advanced Statistical Methods for the Analysis of Large Data-Sets*, A. Di Ciaccio, M. Coli & J. M. Angulo Ibanez, eds., Springer, pp. 113–123.

5 Figures

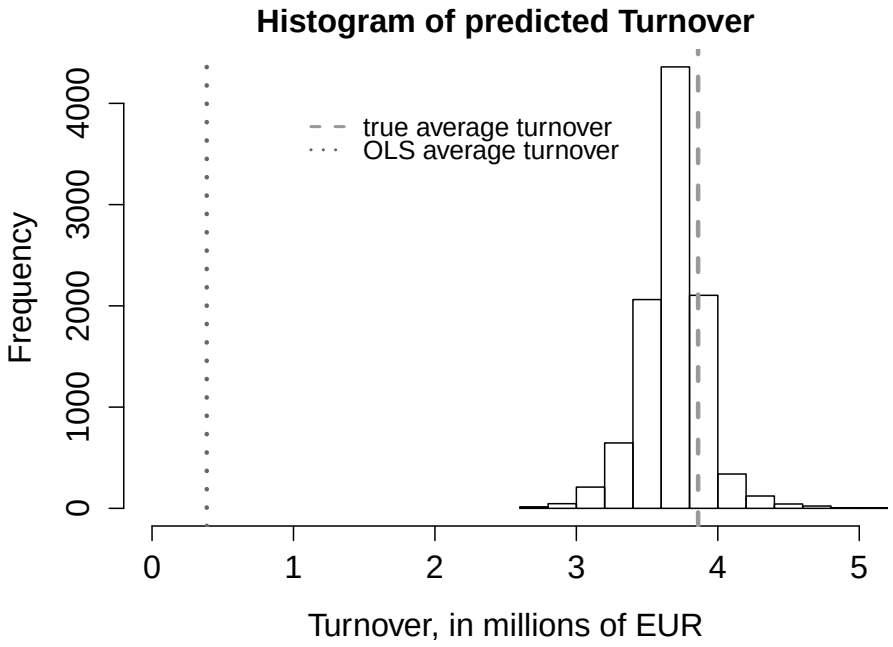

Figure 1: Histogram of Turnover predicted via the Bayesian generalized Heckman approach. The plot compares the new average turnover estimate with the biased OLS estimates.

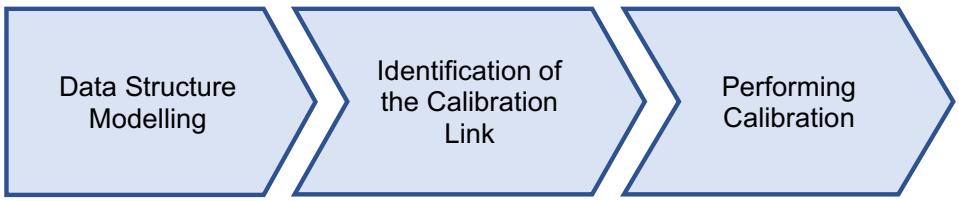

Figure 2: Graphical representation of the data integration methodology.