ON THE USE OF EMPIRICAL OR ARTIFICIAL PROJECT DATA

By Mario Vanhoucke

ABSTRACT

This paper gives a brief overview of the artificial and empirical project data generated and collected by the researchers from the Operations Research and Scheduling (OR&S) group from Ghent University in Belgium. Artificial data are generated by project network generators under a strict design to control both the network structure and the resource constraints, while the empirical project data are collected over a time horizon of multiple years, using a standardized collection and classification method. All data are publicly available on the OR&S website (www.projectmanagement.ugent.be/research/data) and can be used anywhere for academic purposes. More detailed information on the network and resource parameters used to generate the artificial data and the classification process for the collection of empirical data is available in a paper published in the Journal of Modern Project Management (Vanhoucke et al., 2016).

1 INTRODUCTION

We live in a world where the words big data have become the buzzwords to refer to the increasing importance of the availability of data and the powerful methodologies to analyse these huge amounts of data for improved decision making. The presence of data has always been important for decision making in most management disciplines, and therefore also for improving the decision making process in project management and control. In the last decades, researchers have proposed various ways to generate artificial project data that is freely available at various places on the internet. While the generation of artificial data has some advantages compared to the use of real project data, researchers often had no choice but to generate artificial data due to the lack of publicly available empirical project data.

During the past years, the OR&S group has presented a standard method for collecting empirical project data, based on discussions with academics and professionals from the project management field. This novel classification and collection method has been integrated with the experience of generating artificial data, and has resulted in an overview article on the use of data in project management. More precisely, in the recent article entitled “An overview of project data for integrated project management and control" (Vanhoucke et al., 2016), an overview is given on the use of artificial and empirical project data for academic research purposes. The article shows how the available project data is used in the research field known as “Dynamic Scheduling" (Uyttewaal, 2005; Vanhoucke, 2012) or “Integrated Project Management and Control" (Vanhoucke, 2014) which is used to refer to the integration between baseline scheduling, schedule risk analysis and project monitoring and control. The article gives an overall overview of the wide variety of project data that are available and are used in various research publications. It shows how the combination of artificial data and empirical data leads to improved knowledge on and deeper insights into the structure and characteristics of projects useful for academic research and professional use. The current manuscript gives a brief overview of this study and gives references to various interesting sources for obtained project data.

In the next sections of this paper, an overview is given on the project data that are available in the Project Management literature. Section 2 gives a brief overview on the carefully controlled generation process of artificial project data using network and resource parameters as it is done in the literature. Section 3 briefly reviews the classification methods used for collecting empirical project data. In section 4, references are given to the publicly available websites and a discussion of the advantages and disadvantages of the databases is given. Section 5 mentions the freely available material that can be downloaded to support future project management research using the available data.

Ghent University, Tweekerkenstraat 2, 9000 Gent (Belgium)
Vlerick Business School, Reep 1, 9000 Gent (Belgium)
UCL School of Management, University College London, Gower Street, London WC1E 6BT (United Kingdom)

2 ARTIFICIAL DATA

The widespread use of artificial data for research purposes lies not only in the ease of the generation process, but also in the ability to generate the data with parameters relevant for the research study. The major aim of academic research is to develop new methodologies and test their performance on a wide range of problem instances in search for drivers of good or bad performance. Rather than presenting a methodology that can solve a problem for a specific project, the contribution of the research often lies in showing why the new methodology performs well in some cases, but fails to compete with alternative methodologies in other cases. This search for drivers that determine the performance of the new methodologies is crucial for academic research to provide insights into the characteristics of the newly presented ideas. These findings stimulate further developments and add additional improvement steps that can be tested and validated in future research studies. For these reasons, researchers generate the data according to the specific needs of the research study, and often do not feel the need to use real company-specific project data.

The generation of artificial data has been the topic of academic research for decades, and has resulted in project data generators to design the look and feel of the generated data using network and resource metrics. The design of an artificial dataset is then mostly done with one of the data generators under a controlled design using predefined parameters for the topological structure of the project network and the scarceness of resources used by the activities. The OR&S group has been very active in this field, and has developed a random network generator known as RanGen1 (Demeulemeester et al., 2003) and further extended to RanGen2 (Vanhoucke et al., 2008) which makes use of various metrics for the network topology and resource scarceness, as briefly summarized along the following lines:

- **Network topology:** The topological structure is defined by the specific assembly of project activities and precedence relations between these activities. The way the activities are connected to each other using the precedence relations results in a project structure that varies between a complete serial network to a complete parallel network. Metrics such as the Order Strength (OS), the Serial/Parallel indicator (SP) and the Coefficient of Network Complexity (CNC) have been used in the academic literature and has resulted in various insights. As an example, in previous articles in the Measurable News, it has been shown that Earned Value/Schedule indicators are more reliable for forecasting when the topological structure of the project network, measured by the SP indicator, is more serial than parallel (Vanhoucke and Vandevoorde, 2007, 2008, 2009; Vanhoucke, 2010).

- **Resource scarceness:** Modelling the demand for resources by activities as well as the limited availability of the project resources has resulted in various resource parameters to model so-called resource-constrained project scheduling problems. These parameters have been used to model and generate both renewable resources and nonrenewable (or consumable) resources, and typically quantify the relation between the activities and resources using metrics such as the Resource Scarceness (RC), the Resource Strength (RS) and the Resource Use (RU).

For a complete overview of most commonly used network topology and resource scarceness metrics, their formulas and their use for the generation of artificial data, the reader is referred to the paper by Vanhoucke et al. (2016).

3 EMPIRICAL DATA

The major reason why empirical data must be used in research is to validate academic results for practical use, showing the relevance in a real-life setting that often differs either slightly or sometimes dramatically from the well-designed artificial data. As a professional, the availability of data allows testing ideas on company-specific data to fine-tune existing or new methodologies to the unique and specific aspects and settings of the company culture, personal wishes and particular needs of the project manager. Rather than providing insights into drivers for good or bad performance of the newly presented methodologies, the focus often lies on adapting and modifying the methodology in order to optimize its performance for a specific setting. Without real project data, this translation process from theory (academic research with artificial data) to practice (professional experience with empirical data) remains a theoretical exercise with little to no value.

Despite the fragmentary availability of empirical project data, little to no work has been proposed to share the data between professionals and researchers to carry out research, nor to propose a standardized way to further expand the current databases. As far as the author knows, much of the work presented in articles using empirical project data was done on real projects that could not be shared with other researchers due to confidentiality reasons, due to the lack of a clear structure in the data or maybe due to the absence of a general format that could be used for sharing. Therefore, only recently, the OR&S group has presented a construction and evaluation framework for a real-life project database in Batselier and Vanhoucke (2015) using a standardized approach for collecting data and a uniform format to present the data for further use and sharing between researchers.

While the artificial databases in literature focus solely on the design of the project network and resources constraints, the collection of empirical project data is often done from a totally different point of view, focusing more on the completeness of the data for integrated project management and control, and on the authenticity of the data that assesses their value in real-life.

- **Completeness:** The completeness is defined as the extent to which each of the three dynamic scheduling dimensions (baseline scheduling, schedule risk analysis and project control) was covered by the project data, and is expressed by a three-level color code which is based on the traffic light approach proposed by Anbari (2003). A green, yellow and orange color respectively indicates full, mediocre and rather poor completeness of data.

- **Authenticity:** The concept of authenticity is used to indicate the source of the data and the degree of assumptions that had been made while collecting the data. A distinction has been made between project authenticity that is used for the parameters defining the project network and resource constraints (as done for the artificial data) and the tracking authenticity that is relevant for the dynamic data parameters that define the project progress. The latter parameters use the well-known Earned Value Management parameters to measure the periodic performance of the projects in progress.

4 PROS AND CONS

It is tempting to favour the use of empirical data over the use of artificial using arguments that focus on the realism of empirical data and on the restrictions inherent to generated artificial data. This argument is often stated by professionals who correctly argue that research should support the real needs of project managers, and not vice versa. While most researchers agree with the latter statement, it should be recognized that the use of artificial data has some important advantages compared to empirical data. Figure 1 shows the main advantages and disadvantages of artificial and empirical data, and a brief discussion is given along the following lines.

One of the main advantages of artificial data is the ability to generate the data according to the specific needs of the research study. The use of artificial data is crucial for researchers to provide insights into the project drivers that determine the quality and accuracy of project schedules, risk metrics and control methodologies. It is therefore the personal belief of the author that the first and main focus of academic research should lie on using artificial project data based on a so-called controlled and full-factorial design. Thanks to the controlled generation process, the researchers have full control over all the project (network and resource) parameters in order to obtain and present general results that are applicable for a wide variety of projects. Through the use of simulated computer experiments, new relations between the generated project drivers and the computer output can be found and can provide insights into the behaviour of scheduling and control techniques.

However, it is sometimes extremely hard and a waste-of-time to convince professionals about the relevance and advantages of artificial project data. The straightforward advantage of using empirical data is that they better reflect the real project management world compared to the artificial data, and therefore it is often concluded that empirical data should be favoured above artificial data use. However, the use of empirical project data is not without danger. It must not be forgotten that the ultimate goal of research on integrated project management and control is to improve the decision-making process during project progress. Research focuses on evaluating current methods and presenting novel techniques for project control that should be used as triggers for corrective actions to bring projects in danger back on track. These triggers should be used in a careful way and should enable the project manager to take only actions when really necessary. Therefore, the control methodologies should only provide warning signals when the project has a high probability of running out of control, and no warning signals for every little change in the project with a low impact on the project objectives. The main purpose of many of the research studies is to contribute, directly or indirectly, to this challenging goal of presenting methodologies to better control projects in progress and improve corrective actions. The major and inherent weakness of empirical data lies in the fact that these empirical data include many of these corrective actions.

Without an explicit distinction between data and actions, computer experiments cannot test the accuracy and quality of the warning signals of control methods, and can therefore not provide any insights into the relation between project drivers and the performance of these methodologies.

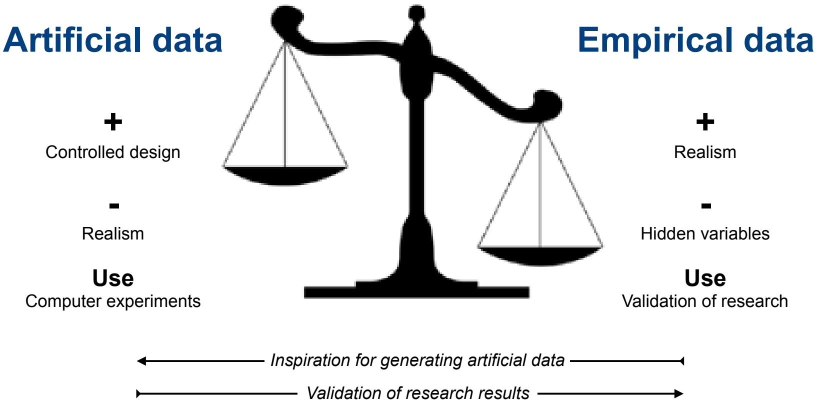

Figure 1: Artificial and empirical project data

Despite the major advantage of using artificial data in a research setting, researchers should also consider the use of empirical real project data in their research endeavours. The use of an empirical set during research increases the likelihood to carry out relevant research and reduces the risk of generating artificial data that have little to no practical value. Moreover, new research results obtained from the experiments on the artificial datasets should be validated on empirical projects to assess their realism and their potential use in practice, and should ideally lead to practical guidelines and insights relevant for real projects. Hence, the main advantage of empirical data for research lies in the validation of existing research results rather than in the creation of new research insights.

While research aims at presenting general results, and professionals aim at obtaining specific added value for their business, both should aim at tightening the gap between the theory and practice to support better decisions for projects. Research should certainly be translated into practical relevance and not only stay in the papers written by academics, but empirical evidence from professionals should also provide an impetus for new academic research. When academics and professionals work in perfect harmony, researchers can focus on general studies with some additional empirical validation, after which professionals can take over in order to extend these new methodologies to sector-specific software tools and techniques, which should be - although very relevant - kept outside the academic environment.

5 CONCLUSION

This article is a summary of the more complete article published in the Journal of Modern Project Management and gives an overview of the use of artificial and empirical project data used by and shared between researchers and professionals. All project data can be downloaded from www.projectmanagement.ugent.be/research/data. This website contains a full overview of the data generated and collected by the OR&S group which can be summarized as follows:

- The previously mentioned paper “An overview of project data for integrated project management and control" can be freely downloaded from this website. The paper contains a full overview of all metrics used for the generation of artificial data and the collection of empirical data. The website also contains a download link to the RanGen generator and references to other generation software tools.
- An MS Excel overview of the data generation process, including references to the data generator and the values for the network and resource metrics, is available for each single project instance of each dataset.
- All data can be downloaded from the website, and currently consists of 10 artificial sets (with thousands of projects) and one empirical set with 100 projects in January 2016 (and growing). Details on the format used to represent the data in Computer Les

is available, and a converter to transform the empirical sets to MS Excel is available for download.

6 ACKNOWLEDGEMENTS

We acknowledge the support provided by the “Bijzonder Onderzoeksfonds" (BOF) for the project with contract number BOF12GOA021, by the “Fonds Wetenschappelijk Onderzoek" (FWO) for the project with contract number G015711N and by the National Bank of Belgium (NBB).

References

Anbari, F. (2003). Earned value project management method and extensions. Project Management Journal, 34(4):12-23.

Batselier, J. and Vanhoucke, M. (2015). Construction and evaluation framework for a real-life project database. International Journal of Project Management, 33:697-710.

Demeulemeester, E., Vanhoucke, M., and Herroelen, W. (2003). RanGen: A random network generator for activity-on-the-node networks. Journal of Scheduling, 6:17-38.

Uyttewaal, E. (2005). Dynamic Scheduling With Microsoft Office Project 2003: The book by and for professionals. Co-published with International Institute for Learning, Inc.

Vanhoucke, M. (2010). Measuring time: An earned value performance management study. The Measurable News, 1:10-14.

Vanhoucke, M. (2012). Project Management with Dynamic Scheduling: Baseline Scheduling, Risk Analysis and Project Control, volume XVIII. Springer.

Vanhoucke, M. (2014). Integrated Project Management and Control: First comes the theory, then the practice. Management for Professionals. Springer.

Vanhoucke, M., Coelho, J., and Batselier, J. (2016). An overview of project data for integrated project management and control. Journal of Modern Project Management, 3(2):6-21.

Vanhoucke, M., Coelho, J., Debels, D., Maenhout, B., and Tavares, L. (2008). An evaluation of the adequacy of project network generators with systematically sampled networks. European Journal of Operational Research, 187:511-524.

Vanhoucke, M. and Vandevoorde, S. (2007). Measuring the accuracy of earned value/earned schedule forecasting predictors. The Measurable News, Winter:26-30.

Vanhoucke, M. and Vandevoorde, S. (2008). Earned value forecast accuracy and activity criticality. The Measurable News, Summer:13-16.

Vanhoucke, M. and Vandevoorde, S. (2009). Forecasting a project's duration under various topological structures. The Measurable News, Spring:26-30.