KEYWORDS project data dynamic scheduling Monte Carlo simulation baseline scheduling schedule risk analysis project control

AN OVERVIEW OF PROJECT DATA FOR INTEGRATED PROJECT MANAGEMENT AND CONTROL

Mario Vanhoucke [1, 2, 3]
mario.vanhoucke@ugent.be

José Coelho [4, 5]
jose.coelho@uab.pt

Jordy Batselier [1]

[1] Faculty of Economics and Business Administration, Belgium

[2] Technology and Operations Management Area, Belgium

[3] UCL School of Management, University College London, United Kingdom

[4] Universidade Aberta, Portugal

[5] INESC – Technology and Science, Portugal

ABSTRACT

In this paper, an overview is given of the project data instances available in the literature to carry out academic research in the field of integrated project management and control. This research field aims at integrating static planning methods and risk analyses with dynamic project control methodologies using the state-of-the-art knowledge from literature and the best practices from the professional project management discipline. Various subtopics of this challenging discipline have been investigated from different angles, each time using project data available in literature, obtained from project data generators or based on a sample of empirical case studies. This paper gives an overall overview of the wide variety of project data that are available and are used in various research publications. It will be shown 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. While the artificial data can be best used to test novel ideas under a strict design in a controlled academic environment, empirical data can serve as the necessary validation step to translate the academic research results into practical ideas, aiming at narrowing the bridge between the theoretical knowledge and practical relevance. A summary of the available project data discussed in this paper can be downloaded from http://www.projectmanagement.ugent.be/research/data.

1. Introduction

In this paper, an overview is given of the data used to test research hypotheses on integrated project management and control. The discipline is also referred to as dynamic scheduling *(Uyttewaal, 2005; Vanhoucke, 2012)* and refers to the integration of three components of managing and controlling projects, known as baseline scheduling, schedule risk analysis and project control *(Vanhoucke, 2014).*

Each of the three disciplines has received attention from various areas of the academic community, which has resulted in a *(huge)* amount of published studies. In this paper, the focus is restricted to the development of quantitative optimization and simulation models for the three disciplines, for which a brief summary is given along the following lines. Firstly, the research on the construction of a project baseline schedule dates back to the '50s with the development of two methods, now known as the Program Evaluation and Review Technique *(PERT)* and the Critical Path Method *(CPM).* Years later came the extension concerning the incorporation of renewable resources with a limited availability. Due to the huge amount of research papers that have been written on this challenging topic, it is almost impossible to give the most important references. Research overviews have appeared in literature and research handbooks have been published. A recent survey of various variants for resource-constrained project scheduling problem is given in Hartmann and Briskorn *(2010)* and a student handbook has been published by Vanhoucke *(2012).* Secondly, the basic principle and underlying methodology used for schedule risk analysis is given by Hulett *(1996)* and an overview of Schedule Risk Analysis *(SRA)* is provided by Williams *(1995).* The technique has recently been used in various studies, such as Vanhoucke *(2010c)* and Elshaer *(2013).* Finally, the research on project control to monitor the performance of projects in progress has received an increasing attention in the past decade, and has resulted in a classification of the project control literature written by Willems and Vanhoucke *(2015)* and a review of analytical models and decision support tools in project control by Hazır *(2015).* The research on these three dimensions all make use of data, sometimes restricted to simple artificial examples, but often based on a wide set of generated data, or even a *(small)* set of empirical project data.

In this paper, an overview will be given of the availability and lack of project data for each of the three disciplines mentioned earlier. A distinction will be made between artificial datasets in literature and empirical databases with data from real projects. When a lot of data are available, as is the case for the baseline scheduling discipline, an overview will be given to bring structure and provide clarity for future researchers. For the disciplines that fall short regarding the availability of artificial or empirical data, references are given to the limited sets that are available and suggestions are formulated to extend the size of the available data.

The outline of this paper is as follows. Section 2 gives a brief introduction of the data types and sources, and why and how the careful selection of project data is important for researchers and professionals active in the field of integrated project management and control. This section is followed by two main sections for two classes of data. Section 3 then gives an overview of the main efforts done to generate and collect artificial data for the three disciplines mentioned earlier, and includes references to network generators, datasets available in the literature as well as efforts to classify the data in predefined classes. Section 4 provides a summary of an empirical dataset of projects and the classification scheme used to validate the quality of real data. Finally, section 5 draws overall conclusions and highlights important avenues for future research. An appendix has been added that gives a short overview of the references to the formulas used to generate data.

2. Project data

Since the recent explosion of digital data, *(project)* managers can measure and know significantly more about their business, and directly translate that knowledge on project performance into improved decision making. The big data hype requires that data are readily available to everyone, allowing a careful and intensive analysis to better measure project progress and therefore allowing to manage the performance of projects more precisely than ever before. This analysis requires data-intensive analytical techniques and methodologies from operations research, computer science and artificial intelligence that add an intelligence layer to big data to tackle complex analytical calculations much faster than ever before. The development of new and testing of existing analytical methodologies are often in the hands of academics studying the three disciplines separately, or aiming at presenting an integrated approach. Such studies require the presence of project data, in huge numbers, to test novel data-intensive ideas on scheduling, risk and control.

The reality is that these project data are often not available in these huge numbers, or - when available - clearly lack the required structure for research. Researchers often must fall back on their own data that differ from study to study, with an unknown source and with little to no relevance for sharing with others. To overcome these problems, many efforts have been made in the past decades to present sets of structured and well-designed data that can be shared and used among researchers for comparing and benchmarking new ideas. The focus of the following sections lies on giving an overview of the design of the existing datasets and on describing how they have been collected or generated, aiming at providing a clear overview of the various sets currently available.

Since the specific needs and details for data might differ between academic researchers and professionals, the next sections elaborate on the various sources of project data in literature *(section 2.1)* and on the difference between static and dynamic data *(section 2.2)* for integrated project management and control.

2.1 Data source

The data source in academic literature can consist of notional data, artificial data generated according to a well-defined process, or carefully collected empirical project data. While notional data only serve illustrative purposes, the difference between artificial data and empirical data is often more important, as they can be used to serve different, sometimes complementary needs.

- Notional data: Notional data consist of one or a few example projects used to illustrate calculations and to present the general relevance of the research idea under study. A single example is often constructed in such a way that it ideally shows the contribution of the newly presented method, and therefore, it often lacks any structure or value to claim the generalization of the research results.
- Artificial data: 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 the problem under study, 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 and provides insights into the characteristics of the newly presented ideas to stimulate further developments and fine-tuning in future research. As an example, in the study of Vanhoucke (2010a), it has been shown that the Earned Schedule (ES) method - at that time a novel extension of the traditional Earned Value Management method (EVM, Fleming and Koppelman (2010)) to measure the time performance of projects - worked well for projects with a rather serial structure, but could not be used for projects with a more parallel structure. The insight has led to follow-up papers by other researchers to develop good alternatives for parallel-structured networks, such as the method presented by Elshaer (2013).
- 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 slightly or 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.

Due to the different purposes of the two last data types, it is crucial for researchers to take a well-considered and balanced view on the use of theoretical artificial project data and empirical real project data in their research endeavours. It is the personal belief of the authors that the first and main focus of academic research should lie on using artificial project data based on a controlled and full-factorial design. In doing so, the researchers have full control over all the project parameters in order to obtain and present general results that are applicable for a wide variety of projects. It allows them to show why their methodologies work and fail, and it enables them to identify future research avenues. Only afterwards, these general results can be translated into practical guidelines and rules of thumb that differ from project to project, company to company and sector to sector. Empirical data serve very well for that purpose, and the resulting case study research should be used as a tool for validation of academic results and for tightening the gap between the academic endeavours and practical relevance, rather than for presenting generalized results. Nevertheless, empirical studies can certainly provide an impetus for new academic research. After performing both the general academic study and the empirical validation study, consultants can take over and extend the interesting ideas to sector-specific tools and methodologies, which should be - although very relevant - kept outside the academic environment.

2.2 Data type

The integrated project management and control methodology requires planning methods to schedule project networks prior to the execution of the projects, as well as project control methods to dynamically monitor the performance of projects in progress. Therefore, the required data needed to test new methodologies should be split in both static and dynamic data. This distinction is shown in **Figure 1**, which displays the project life cycle and the three components of integrated project management and control.

Static project data refer to all data necessary to model all processes carried out prior to the project execution. Obviously, planning and scheduling project activities with or without the presence of limited renewable resources belong to this class of processes and require data for project activities, precedence relations and the activity network, including estimates for time and costs of activities and their need for renewable resources.

Dynamic project data refer to all types of data required to model the progress of the project. The project control phase requires tracking data to measure the progress of the project at periodic time periods. These data should be collected at periodic intervals during project progress to measure the performance of projects and to enable the project manager to forecast the final project duration and costs as well as to take actions when the project runs into trouble.

While the distinction between static and dynamic data for the baseline scheduling and project control phases is straightforward and unambiguous, the third component, known as schedule risk analysis, can be considered as both static and dynamic. This component clearly satisfies the condition of static data, since the analysis of the risk of a schedule is done based on Monte Carlo simulations prior to the start of the project and serves as input for the control phase. However, these Monte Carlo simulations require data that is used to reflect and imitate project progress, and can therefore also be considered as dynamic.

Since the required data for these simulations is similar to the data for the project control phase, this phase will be classified as dynamic, and its specific data requirements will be described in section 3.2.

3. Artificial data

It has been mentioned earlier that one of the main advantages of creating artificial data is that researchers have full control over the parameters during the generation process. Through the use of a careful design, a dataset can be constructed that incorporates a wide and diverse set of different project parameters to assure that new methodologies can be tested for various project settings. Ever since the publication of Elmaghraby and Herroelen *(1980)*, who draw attention to the need for project datasets that span the full range of problem complexity, network and resource parameters have been proposed to describe the characteristics of projects and generators have been developed to generate artificial static data with these parameters. These parameters and generators, as well as the best-known datasets, are described in section 3.1. Generating dynamic data has been less controlled and formalized, since the imitation of real project progress heavily depends on assumptions made about the uncertainty and unexpected events that pop up during progress. Nevertheless, recommendations on the use of statistical distributions have been formulated in literature and are the topic of section 3.2.

3.1 Static project data

This section reviews the static project data parameters that are used by artificial project data generators to obtain data on project networks and project resources. These data generators have been used in literature to generate benchmark sets that are now commonly used and shared between researchers to compare and benchmark results of their studies.

3.1.1 Data parameters

Network topology: A first class of static parameters is used to describe the network topology of the project. The topological structure is defined by the specific assembly of project activities and precedence relations between these activities, and can lead to various structures. This search to model and measure the structure of a project network has resulted in various network parameters for which a non-exhaustive overview is given along the following lines.

A first and simple parameter to measure the network topology is known as the Coefficient of Network Complexity *(CNC)*, originally defined by Pascoe *(1966)* as the number of arcs over the number of nodes for activity-on-the-arc[1] networks and later redefined by Davies *(1974)* and Kaimann *(1974, 1975)*. The measure has been adapted for activity-on-the-node problems by Davis *(1975)* as the number of direct arcs over the number of activities *(nodes)* and has been used in the network generator ProGen *(Kolisch et al., 1995)*. Some researchers have shown that the CNC fails to discriminate between easy and hard project networks and can therefore not serve as a good parameter for describing the impact of the network topology on the hardness of a project scheduling problem.

A second well-known parameter of the topological structure for activity-on-the-node networks is the Order Strength *(OS) (Mastor, 1970)*, defined as the number of precedence relations *(including the transitive[2] ones)* divided by the theoretical maximum number of precedence relations $\frac{n*(n-1)}{2}$, where n denotes the number of activities in the network. It is sometimes referred to as the density *(Kao and Queyranne, 1982)* or the restrictiveness *(Thesen, 1977)* and equals 1 minus the flexibility ratio *(Dar-El, 1973)*. Herroelen and De Reyck *(1999)* conclude that the OS, the density, the restrictiveness and the flexibility ratio constitute one and the same complexity measure. Schwindt *(1995)* and Demeulemeester et al. *(2003)* have used this parameter in the problem generators ProGen/Max and RanGen1, respectively.

Tavares et al. *(1999, 2002)* have presented several other parameters of network topology, which have been further developed by Vanhoucke et al. *(2008)* and implemented in the RanGen2 network generator. The first parameter I1 simply reflects the number of nondummy activities in

[1] In an activity-on-the-arc network, each arc represents a project activity and each node is used to denote a project event. This format is less used in integrated project management and control research, and hence, network topology parameters for this format are not discussed in this paper. In this paper, only network topology parameters for the activity-on-the-node networks will be discussed.

[2] When two direct or immediate precedence relations exist between activities (i, j) and activities (j, k), then there is also an implicit transitive relation between activities (i, k).

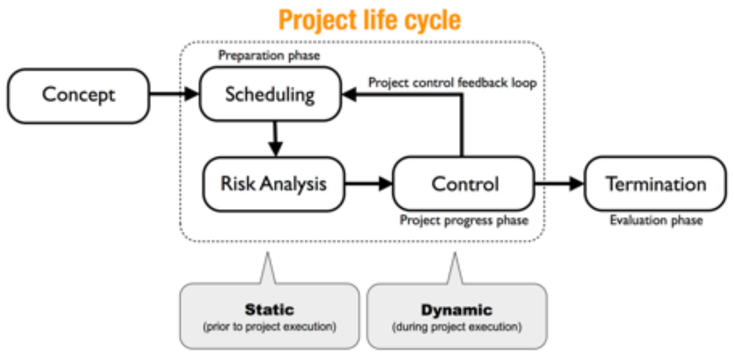


FIGURE 1. *The project life cycle with static and dynamic project phases*

the project. The other five parameters have originally been referred to as the I2 to I6 parameters and have been rescaled to lie between 0 and 1, inclusive, denoting the two extreme structures. Four of these parameters have been renamed to SP, AD, LA or TF *(see further for definitions)* to make them more intuitive and have been used for the generation of 4,100 instances used in a project control study of Vanhoucke *(2010a)*.

The first I2 parameter has been renamed to the Serial/Parallel *(SP)* parameter and measures the closeness of a network to a serial or parallel network. When SP = 0 then all activities are in parallel, and when SP = 1 then the project network is completely serial. Between these two extreme values, networks can be generated closer to either a serial or a parallel network. The SP parameter determines the number of serial activities in the network on the longest chain and can be considered as an easy-to-understand alternative for the OS.

The second I3 parameter, renamed to the Activity Distribution *(AD)*, measures how the project activities that do not belong to the longest chain are distributed in the network. When the longest chain defined by the SP parameter is considered as a number of serial activities that defines a number of levels in the project network, the AD parameter measures the width of each level along this longest chain. When AD = 0, all levels contain a similar number of activities, and hence, the number of activities is uniformly distributed over all levels. When AD = 1, there is one level with a maximal number of activities, and all other levels contain a single activity.

The third parameter, Length of Arcs *(LA)*, measures the length of each precedence relation *(i, j)* in the network as the difference between the level of the end activity j and the level of the start activity i. When LA equals 0, the network has many precedence relations between two activities on levels far from each other. Hence, the activity can be shifted further in the network. When LA equals 1, many precedence relations have a length of one, resulting in activities with immediate successors on the next level of the network, and thus little freedom to shift. This parameter is an alternative for the two parameters I4 *(length of short arcs)* and I5 *(length of long arcs)* of Vanhoucke et al. *(2008)*, which both measure the length of arcs in two different ways. To avoid confusion and overcome this close relation, LA is used and is equal to the I4 parameter, while the I5 parameter is no longer used.

The last parameter is the Topological Float *(TF)* that measures the topological float of a precedence relation as the number of levels each activity can shift without violating the maximal level of the network *(as defined by SP)*. Hence, TF = 0 when the network structure is 100% dense and no activities can be shifted within its structure with a given SP value. A network with TF = 1 is a network with a chain of activities defined by the value of the SP parameter *(these activities obviously have no topological float)*, while the remaining activities have a maximal topological float value.

Resource parameters: A second class of static parameters is used to describe the resource parameters of the projects. 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 Resource-Constrained Project Scheduling Problems *(RCPSPs)*. These parameters have been used to model and generate both renewable and nonrenewable resources. Renewable resources are available on a period-by-period basis, i.e. the available amount is renewed from period to period. Only the total resource use at every time instant is constrained. Typical examples are manpower, machines, tools, equipment or space, and this resource type is used in all RCPSP formulations discussed in section 3.1.3. Nonrenewable resources *(often referred to as consumable resources)* are available on a total project basis, with a limited consumption availability for the entire project. Typical examples are money, raw materials or energy. This resource type is less used in the academic literature and is only defined for one class of RCPSPs of section 3.1.3.

In order to describe and measure the relation between activities and resources, the number of renewable and nonrenewable resources must be specified, and it is common practice in the academic literature to set them to maximum four resource types, for both renewable and nonrenewable resources. Given the number of resource types available to execute the project, the resource requirements by project activities as well as the limited availability of the resources can be measured by various parameters.

The density of the resource requirements is used to describe whether an activity makes use of a particular resource or not, and is measured by the Resource Factor *(RF) (Pascoe, 1966)* or the Resource Use *(RU) (Demeulemeester et al., 2003)*. The RF simply calculates the average portion of resource types requested per activity, but the use of this resource parameter has been criticized in literature. Using the RF for generating project data is not always easy, since it is possible that no resource requirement will be generated for some activities while others use all the resources. Therefore, the RU has been proposed as an alternative and simply varies between zero and the number of resource types available and measures for each activity the number of resource types needed for its execution.

The connection between the resource requirements and the limited resource availability has resulted in parameters such as the Resource Strength *(RS) (Cooper, 1976)* and Resource Constrainedness *(RC) (Patterson, 1976)*. The RC is defined as the average resource requirement for all activities for a particular resource divided by the availability of that resource, and is therefore a simple and easy-to-understand measure to know how scarce the resource is. The RS is - although widely used in the academic literature - not so easy to understand and subject to debate among researchers. Its formula takes both the resource requirements of the project activities as well as the network structure into account, and is therefore criticized by De Reyck and Herroelen *(1996)* for being no pure resource parameter. Details are outside the scope of the current paper, and the reader is referred to Demeulemeester et al. *(2003)* for more information on the specific formulas and a detailed discussion on the advantages and disadvantages of this parameter.

The previous resource parameters have all been defined for renewable resources only. However, when project activities have multiple choices for activity durations and resource requirements *(this is referred to as the multi-mode case in section 3.1.3)*, both renewable and nonrenewable resources are used in the project data, and the RS must then be defined for both resource types. The RS for renewable resources with multiple activity modes has been defined by Kolisch et al. *(1995)* as a straightforward extension for the RS formula for single-mode activities, but Demeulemeester et al. *(2003)* criticize this redefinition and propose some adaptations since otherwise its use for project data generation can result in infeasible *(i.e. unsolvable)* activity/resource combinations. Furthermore, the RS for nonrenewable resources has been defined by Van Peteghem and Vanhoucke *(2014)* and is very similar to the RS for renewable resources. In order to bring structure to the different definitions of the previously discussed network topology and resource parameters, **Table 3** in appendix A is created to provide a summary and references to the exact formulas for both the network topology and resource parameters.

3.1.2 *Data generators*

Many of the network and resource parameters have been used to create project generators that automatically generate artificial static data using a range of values for these parameters. To the best of our knowledge, only one network generator is known to be strongly random *(Demeulemeester et al., 1993)*, which is a feature that expresses that the networks can be generated at random from the space of all feasible networks with a specified number of nodes and arcs. This feature is important to guarantee that all networks that can exist in practice can theoretically be generated by the network generator. Unfortunately, the generator makes use of the activity-on-the-arc format, which is less popular than its activity-on-the-node alternative format, and no other characteristics can be specified for describing the network topology. Therefore, this generator is not further discussed in this paper. All other data generators discussed in this section aim at generating activity-on-the-node project networks under a controlled design by predefining the topological structure of the network as discussed in section 3.1.1. These data generators have been used in research on the RCPSP *(see section 3.1.3)* and therefore also take the previously discussed network topology and resource parameters into account. An overview is given along the following lines.

ProGen is the network generator developed by Kolisch et al. *(1995)* and takes the CNC into account to measure the network topology as well as resource-related characteristics RF and RS. Schwindt *(1995)* extended ProGen to ProGen/Max, which can handle three different types of RCPSPs with minimal and maximal time lags, and relies on the OS instead of the CNC to measure network topology. Drexl et al. *(2000)* presented a project network generator ProGen/πx based on the project generator ProGen, incorporating numerous extensions of the classical RCPSP. Tavares *(1999)* has presented a new generator RiskNet based on the concept of the progressive level by using six topological parameters, referred to as I1 to I6 in section 3.1.1. Demeulemeester et al. *(2003)* have developed an activity-on-the-node network generator RanGen, which is able to generate a large amount of networks with a given value for the OS. Due to an efficient recursive search algorithm, RanGen is able to generate project networks with exact predefined values for different topological structure measures. Finally, Vanhoucke et al. *(2008)* have adapted RanGen to an alternative RanGen2 network generator taking the I1 to I6 into account *(with later some of them redefined to SP, AD, LA and TF)*. Both RanGen generators also consider all the resource parameters RS, RU, RF and RC.

All previously mentioned network generators have been primarily built for generating networks for individual projects, but the generation of multiple project networks can be easily done to test planning, scheduling and control methodologies in a multi-project setting. However, the simple generation of multiple projects by the previously mentioned generators will ignore some specific settings of project portfolios that are not incorporated in the current single project data generators. To the best of our knowledge, there is only one multi-project network generator available in literature developed by Browning and Yassine *(2010)* that fully exploits the specific characteristics of the interaction between single projects by adding project portfolio parameters.

3.1.3 *Datasets*

The generators mentioned in the previous section have been used in various research studies to generate data for the specific research question of the study. However, some researchers have shared their data online, in order to enable other researchers to compare their results with previously obtained research results. Obviously, sharing project data with other researchers only makes sense when the data are used to study a well-known and widely investigated problem, in order to stimulate fair comparisons and evaluations of new research results with the state-of-the-art results that are currently available. The majority of data available for research focuses on the construction of a baseline schedule, and much less data are available for the schedule risk analysis and project control phases of dynamic scheduling. More precisely, data for constructing baseline scheduling focus on RCPSPs, in which the project activities have to be scheduled within the limited availability of resources. The generation of these datasets relies on both the network topology and the resource parameters described in section 3.1.1. Since the RCPSP is a problem that can be investigated under various extensions *(see e.g. Hartmann and Briskorn (2010))*, it is impossible and outside the scope of this paper to give a full overview of all available datasets in literature for all the possible extensions of the RCPSP. Therefore, a choice has been made to restrict the description to the RCPSP and two widely investigated extensions; one that incorporates activity cash flows and another that incorporates multiple modes for the project activities. Only one set is discussed regarding static project data used for schedule risk analysis and project control. This set only incorporates network topology parameters since no resources are taken into account. Obviously, since this set contains project networks used for research on the dynamic phases of dynamic scheduling *(see* **Figure 1***)*, these static project data should be extended with dynamic project data discussed in section 3.2. The description of the four classes of datasets is given along the following lines.

- **RCPSP:** The research on the well-known RCPSP aims at scheduling project activities within the limited availability of renewable resources so that the total project duration - often referred to as the project makespan - is minimized. Various exact algorithms have been developed to solve the problem to optimality, just like heuristic and meta-heuristic procedures to solve the problem to near optimality, and have resulted in a very competitive environment where new results are compared against other published results. The first dataset that has been used to test the ideas was the well-known Patterson set that is a collection of notional project examples from various papers in literature resulting in 110 unstructured projects. This set has long been the primary source for testing new procedures, until it was replaced by a bigger structured dataset once all 110 problems could easily be solved. This alternative set is known as the PSPLIB dataset (Kolisch and Sprecher, 1996) and is still used to benchmark new research results. Together with the new set, the authors proposed some criteria to set up a fair evaluation between different procedures, such as using a stop criterion of 5,000 schedules when population-based metaheuristics are used. All data have been generated by the ProGen generator, and researchers are stimulated to download the benchmark sets to evaluate their algorithms and to send their results to be added to the library. Up to today - almost 20 years after the introduction of four sets containing 30, 60, 90 and 120 activities for the projects - not all solutions currently found could be confirmed to be the optimal ones, despite the rapid increase of computer speed over the years, which makes the dataset still highly relevant for research purposes. An alternative set known as RG300 generated by RanGen has been proposed (Debels and Vanhoucke, 2007) and is available to researchers. This set contains projects with 300 activities and has been generated under a diverse structure of the network topology (using the OS) and resource scarceness (using the RC). Finally, a set known as RG30 (with 30 activities per project) has been constructed to compare the relation between the different network topology parameters and to show that some of the existing sets fall short on network topology diversity, as discussed in Vanhoucke et al. (2008).
- **MMRCPSP**: One of the best-known extensions of the traditional RCPSP concerns the inclusion of multiple modes for each project activity. This so-called Multi-Mode Resource-Constrained Project Scheduling Problem (MMRCPSP) assumes that each project activity can be executed in one of a set of predefined time/resource combinations (modes) where lower activity durations are linked to a higher renewable resource demand. A dataset containing projects with 50 and 100 activities has been put available by Boctor (1993), but the main dominant set has been the PSPLIB, since the library does not only offer single mode project data instances but also provides multi-mode instances under a controlled design. Due to the inherent complexity of the problem, the dataset is restricted to projects with 10 to maximum 30 activities. However, Van Peteghem and Vanhoucke (2014) have shown that the multi-mode set of PSPLIB falls short on some criteria and have presented three alternative sets. The main reasons are the limited range of the PSPLIB instances, both in terms of project structure as in number of modes per project activity. Moreover, not all instances of the PSPLIB can be solved as they contain infeasible mode combinations, while the three newly presented sets, known as MMLIB50, MMLIB100 and MMLIB+ are all feasible and a good algorithm should be able to find a near optimal or optimal solution for each instance.
- **RCPSPDC:** While the RCPSP and its extension to MMRCPSP aims at minimizing the project makespan (i.e. the total project duration), the extension to the well-known Resource-Constrained Project Scheduling Problem with Discounted Cash flows (RCP-SPDC) assumes costs for each activity and aims at maximizing the net present value of the project. Although much less investigated than the RCPSP, many research papers have been written on the problem, presenting exact and heuristic procedures for different payment models. Two datasets have been made available for this problem type, one with projects of 10, 20, 30, 40 and 50 activities (set DC1) that has been used to solve the problems to optimality (Vanhoucke et al., 2001) and a second one with 25, 50, 75 and 100 activities per project (set DC2) that has been used to solve the problem heuristically (Vanhoucke, 2010b).
- **EVM/SRA:** The previous datasets make use of both network topology and resource scarceness parameters, since they are mainly used for the development of various RCPSP algorithms. However, research on schedule risk analysis and project control seldom makes use of resource constraints, and the construction of a baseline schedule is often nothing more than an earliest start schedule using critical path calculations. Therefore, the construction of the static data consists of the generation of project networks with a controlled topological structure, but without the use of resource parameters. The 4,100 data instances of the dataset generated by Vanhoucke (2010a) (set MT) is the most complete set in terms of network topology, and has been generated by varying the SP parameter by nine settings (Set 1), and the AD, LA and TF parameters by four settings (Set 2 to 4, respectively). This has resulted in 900, 800, 1,200 and 1,200 instances, respectively, leading to 4,100 instances in total. Obviously, these static network data are then used in dynamic project control studies, using dynamic project data as discussed in section 3.2.
- **Summary table and critical remarks:** Table 1 gives an overview of the four classes of datasets. The table shows the values for the network topology and resource parameters for renewable and nonrenewable resources, and where applicable the number of modes for the MMRCPSP. The values used for generating the data are classified into three categories, displayed in the following format:
    - Class 1 (red cells). The values for the parameters that were set by the user as input values prior to the generation of the data are shown in the table and separated by a semi-colon in case multiple values are used.
    - Class 2 (green cells). The values for the parameters that have not been set by the user have been calculated afterwards using the definitions discussed in section 3.1 and shown in **Table 3**. For these parameters, the minimum and maximum values are calculated and displayed between brackets.
    - Class 3 (orange cells). A third class of values for the parameters is similar to class 1 (predefined by the user), but consists of values for which our calculations differ from the input values reported in the paper where the dataset has been proposed. These values are formatted as the values of class 2 (minimum and maximum value), but in an italic font to denote that the values should normally belong to class 1 (predefined by the user) but differ from the original paper values. These changes are summarized along the following lines.

- The CNC values are set to 1.50, 1.80 and 2.10 for the J30, J60, J90 and J120 instances of the single-mode PSPLIB, and to 1.50 (J10) and 1.80 (J12 to J30) for the multi-mode PSPLIB. However, the CNC calculations of the generator ProGen take dummy start and end activities into account, as well as the arcs that are connected to these dummy activities. In another paper by Vanhoucke et al. (2008), the CNC value is calculated as the number of direct arcs over the number of nodes, excluding all dummy activities and arcs connected to these dummies. Since this last definition is in line with the definition of the OS, that also excludes the presence of dummies, we have chosen to calculate CNC values according to this last definition (see **Table 3**), and hence, the values differ slightly from the originally reported values.
- Some of the instances of the multi-mode PSPLIB could never result in a feasible project schedule and have therefore been removed from the initial set. The two reasons for these infeasibilities are as follows:
    1. The number of generated instances for each set (J10 to J30) was set to 640. However, some of these generated instances have a total minimum nonrenewable resource demand (requested by all the activities) that exceeds the nonrenewable availability, and therefore, constructing a feasible project schedule is impossible. Removing these instances from the initial set has reduced the number of instances to on average 549 instances per set instead of the reported 640, which corresponds to what has been reported by Van Peteghem and Vanhoucke (2014).
    2. The number of modes for each activity has been set to exactly 3, but some of the activities have modes for which the renewable resource demand exceeds the resource availability. These modes have been removed from the set, which has resulted in an average number of modes lower than 3.
- The definitions of the RS, both for single-mode instances (renewable resources) and multi-mode instances (renewable and nonrenewable resources) differ among different sources in literature, and therefore, the following choices have been made:

1. The definition of the RS for the renewable resources for the single-mode PSPLIB, DC1 and MMLIB+ is the same as the definition used in literature, but a small adaptation has been made that results in some minor changes for the input values of these sets. More precisely, the definition of the RS has been extended with an extra condition that sets the RS value equal to 1 in extreme cases. The new formula is given in **Table 3**.
2. The definition of the RS for the renewable and nonrenewable resources for the multi-mode PSPLIB differs from the original definitions used by Kolisch et al. (1995) to avoid infeasible (i.e. unsolvable) activity/resource combinations, as mentioned earlier in section 3.1.1. More precisely, the definition of the RS for renewable resources is defined in Demeulemeester et al. (2003), while the definition for the RS of nonrenewable resources is taken from Van Peteghem and Vanhoucke (2014). Both definitions result in values for the RS parameter that differ from the originally reported values in the paper of the multi-mode PSPLIB instances.
3. The values for the OS for the DC1 set are not completely identical to the input values of the user since the ProGen/Max generator is not always able to generate project instances with the exact predefined OS values. Therefore, the values slightly differ from the values reported in the original paper.

The previous discussion clearly illustrates the importance of structuring the vast amount of data used in the project management and scheduling literature. Because of the multiple and sometimes confusing definitions of some of the parameters previously discussed, **Table 3** gives an overview of all definitions used. **Table 1** is a summary of all the calculations, and the complete MS Excel file with all values for each individual instance as well as the datasets themselves can be downloaded from www.projectmanagement.ugent.be/research/data.

3.2 Dynamic project data

Dynamic project data are used to imitate project progress, in which deviations from the initial time and cost estimates for the activities result in projects finishing earlier than expected or with a certain delay, and with cost overruns or underruns. Unlike the static network and resource parameters that are used to generate static data under a wide range of various settings, the choice of generating dynamic data is more cumbersome as they should ideally reflect real-life scenarios. However, the real execution of projects is flavoured with unknown events and unexpected schedule deviations that cannot easily be captured by simple data parameters. Using unrealistic assumptions on real project costs or activity delays undoubtedly degrades the quality of the obtained research results, and hence, the credibility of their use for practical purposes. Therefore, project progress should be imitated in an experimental environment using statistical distributions that reflect the characteristics of real project progress. While the value of Monte Carlo simulations to imitate project progress has long been established *(Schonberger, 1981; Ragsdale, 1989; Williams, 1995; Kwak and Ingall, 2007)*, the choice of the right distribution to model activity duration uncertainty has been subject to a debate among researchers. Since the development of PERT, the beta distribution was assumed to be the best distribution to accurately represent the uncertainty present in the duration of activities of real-life projects. However, throughout the years, different authors have suggested alternatives, such as Kuhl et al. *(2007)* who proposed the use of the generalized beta distribution *(which has been used in the project control studies of Vanhoucke (2010c, 2011))*, but also the lognormal distribution *(Mohan et al., 2007)*, a mixture of beta and uniform distributions *(Hahn, 2008)* and the doubly truncated normal distribution *(Kotiah and Wallace, 1973)* have been applied to study stochastic activity networks.

The lack of realism is further strengthened by the presence of dependencies between unknown events that typify project progress. The simple use of statistical distributions

TABLE 1. *Details of four classes of artificial static datasets used in academic research*

		General					Network Topology							
		Dataset	Subsets	# Inst	Paper	Generator	# Act (or I1)	CNC	OS	SP (or I2)	AD (or I3)	LA (or I4)	I5	TF (or I6)
Resource-constrained Baseline Scheduling	RCPSP	Patterson		110	Patterson (1984)	Notional Data	[5-49]	[0.714-3.1]	[0.214-0.868]	[0.15-0.6]	[0.109-0.667]	[0-1]	[0-1]	[0-0.5]
		PSPLIB	J30	480	Kolisch and Sprecher (1996)	ProGen	30	*[1.4-2.067]*	[0.246-0.644]	[0.172-0.414]	[0.111-0.518]	[0-0.246]	[0.553-0.883]	[0.022-0.5]
			J60	480			60	*[1.45-2.083]*	[0.203-0.545]	[0.119-0.271]	[0.163-0.51]	[0-0.076]	[0.621-0.86]	[0.067-0.499]
			J90	480			90	*[1.467-2.089]*	[0.164-0.477]	[0.101-0.202]	[0.158-0.503]	[0-0.041]	[0.68-0.844]	[0.096-0.45]
			J120	600			120	*[1.475-2.092]*	[0.15-0.43]	[0.092-0.176]	[0.185-0.533]	[0.001-0.031]	[0.693-0.851]	[0.094-0.438]
		RG300		480	Debels and Vanhoucke (2007)	RanGen1	300	[10.657-18.047]	0.25; 0.50; 0.75	[0.013-0.06]	[0.061-0.345]	[0.029-0.11]	[0.505-0.915]	[0.043-0.407]
		RG30	Set 1	900	Vanhoucke et al. (2008)	RanGen2	30	[0.233-6.467]	[0.021-0.993]	0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9	[0.05-0.962]	[0-1]	[0-1]	[0-0.985]
			Set 2	180			30	[0.3-3.3]	[0.044-0.972]	0.20; 0.50; 0.80	0.15-0.85	[0-0.813]	[0-1]	[0.008-0.958]
			Set 3	240			30	[1.067-3.467]	[0.283-0.984]	0.20; 0.50; 0.80	[0.1-0.87]	0.15-0.85	[0.653-1]	[0-0.536]
			Set 4	240			30	[0.267-3.167]	[0.044-0.936]	0.20; 0.50; 0.80	[0.15-0.95]	[0-0.286]	0.15-0.85	[0.114-0.967]
			Set 5	240			30	[0.533-3.2]	[0.069-0.931]	0.20; 0.50; 0.80	[0.1-0.913]	[0-0.288]	[0.045-1]	0.15-0.85
	RCPSPDC	DC1	mv	1,800	Vanhoucke et al. (2001)	ProGen/Max	10; 20; 30; 40; 50	[0.7-2.3]	*[0.25-0.824]*	[0.222-0.846]	[0-0.917]	[0-0.8]	[0-1]	[0-0.864]
		DC2	npv25	180	Vanhoucke (2010b)	RanGen1	25	[1.6-3.08]	0.25; 0.50; 0.75	[0.083-0.417]	[0.063-0.667]	[0.04-0.434]	[0.222-0.941]	[0.032-0.588]
			npv50	180			50	[2.82-4.76]	0.25; 0.50; 0.75	[0.061-0.245]	[0.099-0.446]	[0.038-0.285]	[0.428-0.929]	[0.036-0.527]
			npv75	180			75	[3.733-6.4]	0.25; 0.50; 0.75	[0.041-0.189]	[0.042-0.443]	[0.048-0.225]	[0.445-0.922]	[0.045-0.461]
			npv100	180			100	[4.52-7.68]	0.25; 0.50; 0.75	[0.03-0.162]	[0.073-0.46]	[0.043-0.201]	[0.47-0.914]	[0.042-0.45]
	MMRCPSP	Boctor	boct50	120	Boctor (1993)	-	50	[1.38-1.9]	[0.793-0.927]	[0.388-0.612]	[0.346-0.571]	[0.043-0.355]	[0.841-0.969]	[0.04-0.172]
			boct100	240			100	[1.46-1.85]	[0.892-0.945]	[0.444-0.616]	[0.306-0.574]	[0.053-0.27]	[0.889-0.964]	[0.038-0.113]
		PSPLIB	J10	*536*	Kolisch and Sprecher (1996)	ProGen	10	*[1.2-1.2]*	[0.356-0.689]	[0.222-0.556]	[0.143-0.8]	[0-0.571]	[0.1-1]	[0-0.65]
			J12	*547*			12	*[1.667-1.667]*	[0.485-0.727]	[0.273-0.455]	[0-0.6]	[0-0.615]	[0.485-1]	[0-0.433]
			J14	*551*			14	*[1.643-1.643]*	[0.451-0.769]	[0.231-0.462]	[0.111-0.6]	[0-0.5]	[0.5-0.944]	[0-0.476]
			J16	*550*			16	*[1.688-1.688]*	[0.433-0.733]	[0.2-0.467]	[0.091-0.571]	[0-0.417]	[0.457-0.929]	[0-0.481]
			J18	*552*			18	*[1.667-1.667]*	[0.399-0.725]	[0.235-0.529]	[0-0.571]	[0-0.368]	[0.533-0.889]	[0-0.528]
			J20	*554*			20	*[1.7-1.7]*	[0.411-0.758]	[0.211-0.526]	[0.077-0.524]	[0-0.353]	[0.576-0.902]	[0.014-0.455]
			J30	*552*			30	*[1.733-1.733]*	[0.343-0.607]	[0.207-0.414]	[0.078-0.559]	[0-0.156]	[0.6-0.88]	[0.043-0.405]
		MMLIB	MMLIB50	540	Van Peteghem and Vanhoucke (2014)	RanGen1	50	[2.76-5.02]	0.25; 0.50; 0.75	[0.041-0.286]	[0.021-0.444]	[0.042-0.33]	[0.422-0.917]	[0.034-0.495]
			MMLIB100	540			100	[4.43-7.63]	0.25; 0.50; 0.75	[0.03-0.152]	[0.056-0.396]	[0.045-0.209]	[0.47-0.919]	[0.035-0.468]
			MMLIB+	3,240			50; 100	[2.62-7.63]	0.25; 0.50; 0.75	[0.03-0.286]	[0.029-0.475]	[0.024-0.33]	[0.418-0.932]	[0.028-0.509]
Schedule Risk Analysis Project Control	SRA/EVM	MT	Set 1	900	Vanhoucke (2010a)	RanGen2	30	[0.2-3.333]	[0.018-0.993]	0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9	[0.1-0.962]	[0-1]	[0-1]	[0-0.976]
			Set 2	800			30	[0.3-3.533]	[0.044-0.929]	0.20; 0.50	0.2; 0.4; 0.6; 0.8	[0-0.679]	[0-1]	[0.019-0.967]
			Set 3	1,200			30	[1.067-4.867]	[0.23-0.984]	0.20; 0.50; 0.80	[0.05-0.87]	0.2; 0.4; 0.6; 0.8	[0.631-1]	[0-0.658]
			Set 4	1,200			30	[0.467-3.467]	[0.076-0.917]	0.20; 0.50; 0.80	[0.05-0.957]	[0-0.394]	[0-1]	0.2; 0.4; 0.6; 0.8

The table is continued here

General				Renewable Resource Parameters					Nonrenewable Resource Parameters			# Modes
Dataset	Subsets	# Inst	Paper	# RR	RF	RU	RS	RC	# NR	RF	RS	
Patterson		110	Patterson (1984)	[1-3]	[0.25-1]	[0.75-3]	[0-1]	[0.184-0.625]	n.a.	n.a.	n.a.	n.a.
PSPLIB	J30	480	Kolisch and Sprecher (1996)	4	0.25; 0.50; 0.75; 1.00	[1-4]	*[0.144-1]*	[0.102-0.607]	n.a.	n.a.	n.a.	n.a.
	J60	480		4	0.25; 0.50; 0.75; 1.00	[1-4]	*[0.141-1]*	[0.068-0.491]	n.a.	n.a.	n.a.	n.a.
	J90	480		4	0.25; 0.50; 0.75; 1.00	[1-4]	*[0.183-1]*	[0.048-0.449]	n.a.	n.a.	n.a.	n.a.
	J120	600		4	0.25; 0.50; 0.75; 1.00	[1-4]	*[0.089-0.521]*	[0.064-0.503]	n.a.	n.a.	n.a.	n.a.
RG300		480	Debels and Vanhoucke (2007)	4	[0.25-1]	1; 2; 3; 4	[0-0.42]	0.20; 0.40; 0.60; 0.80	n.a.	n.a.	n.a.	n.a.
RG30	Set 1	900	Vanhoucke et al. (2008)	4	[0.75-0.75]	3	[0.012-1]	0.4	n.a.	n.a.	n.a.	n.a.
	Set 2	180		4	[0.75-0.75]	3	[0.03-0.85]	0.4	n.a.	n.a.	n.a.	n.a.
	Set 3	240		4	[0.75-0.75]	3	[0.041-1]	0.4	n.a.	n.a.	n.a.	n.a.
	Set 4	240		4	[0.75-0.75]	3	[0.023-0.95]	0.4	n.a.	n.a.	n.a.	n.a.
	Set 5	240		4	[0.75-0.75]	3	[0.027-1]	0.4	n.a.	n.a.	n.a.	n.a.
DC1	mv	1,800	Vanhoucke et al. (2001)	4	0.25; 0.50; 0.75; 1.00	[1-4]	*[0-1]*	[0.261-0.825]	n.a.	n.a.	n.a.	n.a.
DC2	npv25	180	Vanhoucke (2010b)	4	[0.5-1]	2; 4	[0-1]	0.25; 0.50; 0.75	n.a.	n.a.	n.a.	n.a.
	npv50	180		4	[0.5-1]	2; 4	[0-0.765]	0.25; 0.50; 0.75	n.a.	n.a.	n.a.	n.a.
	npv75	180		4	[0.5-1]	2; 4	[0-0.522]	0.25; 0.50; 0.75	n.a.	n.a.	n.a.	n.a.
	npv100	180		4	[0.5-1]	2; 4	[0-0.367]	0.25; 0.50; 0.75	n.a.	n.a.	n.a.	n.a.
Boctor	boct50	120	Boctor (1993)	[1-4]	[0.667-1]	[1-3.68]	[0-0.25]	[0.443-0.822]	n.a.	n.a.	n.a.	[2.1-2.92]
	boct100	240		[1-4]	[0.669-1]	[1-3.57]	[0.021-0.182]	[0.464-0.765]	n.a.	n.a.	n.a.	[2.24-2.83]
PSPLIB	J10	*536*	Kolisch and Sprecher (1996)	2	0.50; 1.00	[2.2-4]	*[0-1]*	[0.098-0.684]	2	0.50; 1.00	*[0.098-0.684]*	*[1.9-3]*
	J12	*547*		2	0.50; 1.00	[2.25-4]	*[0-1]*	[0.092-0.512]	2	0.50; 1.00	*[0.092-0.512]*	*[1.833-3]*
	J14	*551*		2	0.50; 1.00	[2.357-4]	*[0-1]*	[0.081-0.494]	2	0.50; 1.00	*[0.081-0.494]*	*[1.929-3]*
	J16	*550*		2	0.50; 1.00	[2.438-4]	*[0-1]*	[0.076-0.409]	2	0.50; 1.00	*[0.076-0.409]*	*[2.25-3]*
	J18	*552*		2	0.50; 1.00	[2.389-4]	*[0-1]*	[0.067-0.353]	2	0.50; 1.00	*[0.067-0.353]*	*[2.278-3]*
	J20	*554*		2	0.50; 1.00	[2.45-4]	*[0-1]*	[0.069-0.315]	2	0.50; 1.00	*[0.069-0.315]*	*[2.35-3]*
	J30	*552*		2	0.50; 1.00	[2.5-4]	*[0.04-1]*	[0.046-0.26]	2	0.50; 1.00	*[0.046-0.26]*	*[2.867-3]*
MMLIB	MMLIB50	540	Van Peteghem and Vanhoucke (2014)	2	0.5; 1.00	[2.62-4]	0.25; 0.50; 0.75	[0.027-0.173]	2	0.5; 1.00	0.25; 0.50; 0.75	3
	MMLIB100	540		2	0.5; 1.00	[2.67-4]	0.25; 0.50; 0.75	[0.016-0.143]	2	0.5; 1.00	0.25; 0.50; 0.75	3
	MMLIB+	3,240		2; 4	1	[4-8]	*[0.193-1]*	[0.005-0.168]	2; 4	1	0.25; 0.50; 0.75	3; 6; 9
MT	Set 1	900	Vanhoucke (2010a)	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Set 2	800		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Set 3	1,200		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Set 4	1,200		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

to model deviations between real activity costs and durations and the baseline schedule estimates often ignores these dependencies, which make the resulting Monte Carlo simulation unrealistic or even irrelevant for practical use. A method to incorporate dependencies in project progress that is worth mentioning is the concept of the lognormal core presented by Trietsch et al. *(2012)*. These authors theoretically support a claim to use the Parkinson distribution and rely on a model using linear association to model statistical dependencies between activities. The proposed method that is used is an artificial simulation study by Colin and Vanhoucke *(2014)* and is validated by an empirical experiment on activity durations for project management simulation studies by Colin and Vanhoucke *(2015b)* using real projects from the database of Batselier and Vanhoucke *(2015a)*.

The risk of unrealistic assumptions and the resulting inability to imitate real project progress are probably the main reasons why the use of artificial data is often degraded to "too theoretical" and "overly unrealistic" by some researchers and many professionals. On workshops and conferences beyond the purely academic communities there is often an outspoken preference for empirical data to tighten the connection with reality. While the authors of this article recognize the importance of real-life data, they do not share this opinion completely as there are also some fundamental drawbacks on using empirical data, as will be discussed in section 4.

4. Empirical data

Since real project data come - by definition - from the execution of real projects, no distinction should be made between static and dynamic data. Every project should at least be planned with or without the presence of resources, and hence, the availability of static project network data is a minimum requirement for the execution and control of the project. Additionally, the dynamic project data should be carefully collected and analyzed and should therefore also be available. But the latter is not straightforward, and there clearly lies the danger in collecting real data. Often, the static and dynamic data are not readily available, or not well-structured, or for obvious reasons confidential and therefore forbidden to share. Moreover, projects are often monitored using the intuition of the project manager, with the help of unstructured data or data that are used as a starting point but that are never updated once the project enters the progress phase. Consequently, the data points are not always well updated or structured according to the needs of researchers, often resulting in a patchwork of data points that cannot be easily shared with other researchers and/or professionals due to the lack of structure and the unclear meaning. In a recent paper written by Batselier and Vanhoucke *(2015a)*, this problem has been recognized, and a classification system has been proposed to validate the quality and completeness of the data using a three colour system on three criteria, as will be discussed in section 4.1.

In section 4.2, the static data parameters of the previous section will be briefly discussed for the empirical dataset, as well as some newly defined dynamic data parameters. In section 4.3, the use of empirical data will be put in the right perspective by showing some advantages as well as inherent weaknesses of real data for research on integrated project management and control.

4.1 Data classification

Batselier and Vanhoucke *(2015a)* have presented a continuously extendible and publicly available set of empirical projects that outranks all existing empirical databases from the project management literature in both size and diversity. Moreover, the authors have also presented a so-called project card framework to ensure qualitative database extension regarding diversity and authenticity. A project card summarizes the specific details of a certain project and provides a tool for categorizing and evaluating these project data. It also includes some of the static parameters described in section 3.1.1 as well as dynamic parameters that will be discussed and summarized in section 4.2. Each project card is split up in three main parts, summarizing statistics for the three dimensions - baseline schedule, risk analysis and project control - of integrated project management and control.

Although the empirical projects are often much richer than the artificial data and include not only static data but also a wide variety of dynamic data parameters, not all data points were readily available and, even when they were, they were not always correct or were at least sensitive for interpretation. As an example, while the artificial project data all have perfect information about the resource use as described by the resource parameters presented earlier, some of the empirical projects simply did not make use of resource data, and therefore, these data could not be incorporated in the project card. But even for the projects that made use of resources, not all details of every little aspect of the resource use were completely based on the input of the project manager, and sometimes, assumptions had to be made. These two important differences between empirical data and artificial data have resulted in a classification system that measures both the completeness and the authenticity of the project data in order to validate the availability and realism of the empirical data.

4.1.1 Completeness

The completeness is measured as the extent to which each of the three dynamic scheduling dimensions 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. The baseline schedule dimension is said to be fully complete when all details for the project network have been included in the set, as well as data for resources and costs used by the project activities. As an example, projects that do not make use of resources are used for simple critical path calculations, and are not fully complete. The schedule risk analysis dimension is more complete when non-standard risk distribution profiles for activity durations were defined. The default distribution used is the triangular distribution with symmetrical tails to the left and right, but when these distributions have been replaced by other distributions, the data is said to be more complete. The project control dimension requires periodic data on real durations and costs in order to generate performance data using the EVM methodology. This data can be easily generated using Monte Carlo simulations described in Vanhoucke *(2010a)*, but when the tracking data was available and originated from user input instead of from simulations, the project is said to be more complete.

4.1.2 Authenticity

Next to an indication of whether the data are complete or not, the concept of authenticity was also introduced to indicate the source of the data and the degree of assumptions that had been made while entering the data. A distinction has been made between project authenticity that is used for the static data parameters and tracking authenticity that is relevant for the dynamic data parameters.

The project authenticity is said to be high when all static parameters, including activity, resource and *(baseline)* cost data were all obtained directly from the actual project owner. Full authenticity of data thus implies that the data collector did not make any personal assumptions regarding the relevant data types. It should be mentioned that it is perfectly possible that no data for resources is available *(resulting in a lower completeness value)* while still having a fully authentic project when no assumptions have been made for the remaining static data that were available.

The tracking authenticity is used to assess whether or not the dynamic data described in section 4.2 are authentic, and full tracking authenticity is achieved when the tracking data that were obtained from the project owner include actual activity start dates, durations and costs, without any modification or assumption made by the project collector. Both the project and tracking authenticities are evaluated according to the same color code-based approach as presented for the project completeness.

The concepts of completeness and authenticity could also be easily used for the artificial data described in the previous sections, but should lead to obvious results. Thanks to an artificial generation process using network topology and resource parameters, the static data could easily result in a 100% completeness, but due to its artificial nature the data would always be 0% authentic. The dynamic data is somewhat different. While the data for the schedule risk analysis dimension could vary from theoretical and artificial to inspired on real distributions, resulting in various values for completeness and authenticity, none of the previously described artificial datasets have project control data, and hence, have a zero completeness and authenticity score on this third dimension.

4.2 Data parameters

Table 2 gives an overview of the static and dynamic parameters for the empirical data, split up in nine subsets reflecting data for different sectors. The table has a similar structure as **Table 1**, but now includes not only static data *(network topology and resource parameters)* but also dynamic data parameters. Obviously, the table does not report values separated by a semi-colon *(class 1 in the paragraph "Summary table and critical remarks" of section 3.1.3)* since none of the networks have been generated. All data have been collected from real projects, and therefore, all values have been calculated afterwards upon availability. Appendix A provides a summary of the parameters with references to literature.

TABLE 2. *Details of nine classes of empirical datasets used in academic research (including static and dynamic parameters)*

			Static Project Data													
	General		Network Topology								Renewable Resource Parameters					
	Sector	# Inst	# Act (or I1)	CNC	OS	SP (or I2)	AD (or I3)	LA (or I4)	I5	TF (or I6)	RI	# RR	RF	RU	RS	RC
Empirical Database	Construction (civil)	12	73	1.22	0.57	0.43	0.48	0.18	0.86	0.31	0.71	7	0.48	2.04	0.30	0.54
	Construction (commercial building)	12	67	1.26	0.61	0.44	0.54	0.22	0.86	0.29	0.78	8	0.30	1.05	0.64	0.56
	Construction (industrial)	6	151	0.86	0.16	0.12	0.35	0.17	0.85	0.43	0.60	20	0.15	2.47	0.65	0.42
	Construction (institutional building)	5	127	0.86	0.36	0.37	0.64	0.06	0.96	0.56	0.87	-	-	-	-	-
	Construction (residential building)	9	45	1.21	0.62	0.49	0.52	0.21	0.89	0.25	0.83	6	0.47	1.53	0.53	0.56
	Education	2	123	0.71	0.06	0.14	0.49	0.00	1.00	0.60	0.86	2	0.75	1.25	0.41	0.52
	Event Management	3	40	1.40	0.49	0.31	0.45	0.10	0.81	0.14	0.58	6	0.37	1.28	0.49	0.63
	IT	5	84	1.10	0.68	0.53	0.56	0.15	0.92	0.16	0.72	5	0.39	1.30	0.53	0.70
	Production	2	26	0.77	0.49	0.48	0.82	0.00	0.93	0.47	0.81	-	-	-	-	-

The table
is continued
here

	General	Nonrenewable Resource Parameters			Time Data	Cost Data				SRA Distributions			
	Sector	# NR	RF	RS	PD [days]	Fixed Cost [€]	Variable Cost [€]	Resource Cost [€]	BAC [€]	Symmetrical [%]	Skewed [%]	No Risk [%]	Non-standard [%]
Empirical Database	Construction (civil)	2	0.27	1.00	445	435,253,639	186,632	69,664,583	505,104,854	96.5	2.5	0	1.0
	Construction (commercial building)	-	-	-	253	815,489	196,901	169,046	1,181,436	87.5	0.5	2.6	9.4
	Construction (industrial)	-	-	-	389	26,236,424	17,662,574	12,069,522	55,968,520	69.1	26.7	4.2	0
	Construction (institutional building)	-	-	-	271	5,424,084	4,026,616	0	9,450,699	98.4	1.3	0.3	0
	Construction (residential building)	-	-	-	294	4,690,247	487,652	55,473	5,233,372	100	0	0	0
	Education	-	-	-	131	90,000	2,792	21,530	114,321	73.2	4.1	22.8	0
	Event Management	-	-	-	237	25,827	0	15,435	41,262	63.6	13.2	22.3	0.8
	IT	-	-	-	139	33,780	933,332	104,183	1,071,296	99.8	0	0	0.2
	Production	-	-	-	323	533,024	0	0	533,024	100	0	0	0

The table
is continued
here

		Dynamic Project Data						
	General	Project Control						
	Sector	# Periods	Dev. Dur. [%]	Dev. Cost [%]	Avg. SPI	Avg. SPI(t)	Avg. CPI	Avg. p-factor
Empirical Database	Construction (civil)	101	0.1	10.5	0.95	0.96	1.19	0.93
	Construction (commercial building)	45	-5.6	-8.6	3.22	1.14	1.05	0.94
	Construction (industrial)	156	38.4	13.5	0.78	0.76	0.93	0.98
	Construction (institutional building)	52	24.4	10.8	0.87	0.77	0.94	0.93
	Construction (residential building)	107	10.8	9.4	0.89	0.88	0.91	0.99
	Education	13	13.5	93.9	0.78	0.78	0.52	0.84
	Event Management	-	-	-	-	-	-	-
	IT	28	18.8	2.9	0.88	0.88	0.98	0.98
	Production	60	23.5	-0.3	0.80	0.73	0.97	0.98

4.2.1 Static data

Most of the static parameters for the empirical data are identical to the artificial parameters and will not be repeated here. However, some new static parameters are shown in the table, which illustrates that empirical data are used for research in a different way than artificial data. These new parameters can be classified in two categories, as explained along the following lines.

First, a new parameter has been defined as the Regular/Irregular *(RI)* indicator, originally proposed by Batselier and Vanhoucke *(2015b)* and defined in a similar way as the Serial/Parallel *(SP)* parameter presented earlier. However, the RI parameter not only measures the structure of the network as is the case for the SP parameter, but also takes cost information and the timings of the project's baseline schedule into account. More precisely, the parameter reflects the cost accrue of a project from its start to its planned finish, and is used to provide a better indication of the expected accuracy of a certain dynamic control method using EVM. Just as a completely serial project has an SP of 1, a perfectly regular project - that is a project with a perfectly linear cost accrue - is characterized by an RI of 1. At the opposite end of the regularity spectrum, a maximally irregular project is represented by RI = 0 and occurs when the cost accrue is zero throughout the entire project life and suddenly jumps to the project budget at the project finish. Just like for the SP parameter, projects with different degrees of regularity are situated between these two extreme cases. Since none of the artificial projects contain cost data, the RI value has never been reported there. The main reason for the lack of cost data for artificial projects is that *(i)* costs are mostly irrelevant for scheduling projects *(as is the case for e.g. the RCPSP)* or *(ii)* - if relevant - they can be easily generated by generating random numbers. The latter has been done for solving the RPCPSDC using the DC1 and DC2 datasets. Moreover, Vanhoucke *(2010b)* has uploaded some of the generated numbers in cash flow files so that other researchers make use of the same cost data when comparing algorithms.

A second important difference is that the static data for the empirical projects also contains the parameters Planned Duration *(PD, expressed in working days)* and parameters for the project costs, displayed as Budget At Completion *(BAC)* in **Table 2** to be in line with the terminology of EVM. This total planned project cost is further split up into fixed activity costs *(€)*, variable activity costs *(€)*, and resource costs *(€)*. The reason why cost values are not available for artificial data has been discussed in the previous paragraph. The reason why values for the project durations are not available for artificial files lies in the fact that project durations are the result of the construction of a baseline schedule, and hence, is the outcome of a scheduling algorithm. This algorithm is developed by a researcher who makes use of the artificial data to test its quality hoping the results will outperform all previous algorithms on some criteria so that his/her hard work leads to a new academic publication. Consequently, the planned values for project durations are the output of the research, while they can be considered as input by the project manager who has put his real data available to the empirical database. The best *(i.e. the lowest)* found project durations for many of the artificial projects are often found by different researchers using different algorithms, and are referred to as best known solutions *(BKS)*. For some of the artificial datasets, the BKS are displayed in the MS Excel file previously mentioned.

4.2.2 Dynamic data

Dynamic data are the core of empirical project data since they reflect reality. As previously mentioned, the dynamic data can be split up into input distributions necessary to perform simulation studies for schedule risk analyses and data to monitor the progress of the project for project control. These two classes of dynamic data will be briefly discussed along the following lines.

- **Schedule Risk Analysis:** An SRA requires distributions of the activity durations in order to perform Monte Carlo simulations to measure the time/cost/resource sensitivity of project activities. These distributions are classified into four categories. The activities without uncertainty are assumed to be deterministic and are labelled with "No Risk". All others have distributions that can be symmetrical ("Symmetrical") or have a certain degree of skewness ("Skewed"). The symmetrical distributions are assumed to be triangular distributions defined by lower (a) and upper values (b) and the mode (m) expressed relatively to the baseline duration of the activity. As an example, using a standard symmetrical distribution with parameters (a, m, b) = (80%, 100%, 120%) for an activity with an estimated baseline duration of 10 days will have lower and upper values equal to 8 days and 12 days, respectively, and a mode equal to 10 days. The skewed distributions consist of left skewed distributions with (a, m, b) = (80%, 110%, 120%) and right skewed distributions (a, m, b) = (80%, 90%, 120%). All other distributions mentioned under the label "Non-standard" have another degree of skewness, or are even more advanced than using three point estimates that typify triangular distributions. Each activity of each project belongs to one of these classes, and a summary is given in the table that takes into account the size of the projects. More precisely, rather than simply reporting the average percentage of activities over all projects that belong to each class, a weighted average has been used taking the number of activities for each project into account.

  As an example, assume two projects, one project with 10 activities and all activities (100%) assigned to a symmetrical distribution, and another project with 100 activities with 50 activities (50%) assigned to a symmetrical distribution and the other 50 activities (50%) to a non-standard distribution. Instead of reporting the unweighed values $\frac{100\%+50\%}{2} = 75\%$ (Symmetrical) and $\frac{100\%*10+50\%*10}{110}$ (Non-standard) that largely ignore the size of each project, the table reports weighted values as $\frac{100\%*10+50\%*100}{110} = 55\%$ (Symmetrical) and $\frac{0\%*10+50\%*100}{110} = 45\%$ (Non-standard) to better reflect the degree of classification in each distribution class

relative to the size of the projects.

- **Project Control:** The project progress data used for project control consists of actual durations and costs for the project, known upon the project finish, as well as intermediate values collected during reporting periods when the project was in progress. Both the actual duration (in working days) and the actual cost (€ of the project belong to this first class and report real values for the time and cost aspect of the project. In the table, these values are reported as a percentage difference in comparison with their planned values reported by the static data parameters (with a positive number reflecting behind schedule/over budget and a negative number reflecting ahead of schedule/under budget). Intermediate data collected during the project life are collected at different time instances, and are represented by the number of reporting periods (# Periods) for which the performance of the project was measured. This number is equal to the total number of periods for which control data using EVM were gathered - rather than the average number - to give a good indication of the considerable amount of available data for project control. For projects without data of intermediate progress, this number is set to zero. These periodic measurements have resulted in various performance measures that indicate how the project is doing so far with respect to time and cost, and for which the average values (over all periods) are reported for the Schedule Performance Index ("Avg. SPI" and "Avg. SPI(t)"), the Cost Performance Index ("Avg. CPI"), and the p-factor that measures schedule adherence ("Avg. p-factor"). These measures all belong to the EVM methodology and a discussion of these metrics and their formulas is outside the scope of this paper. Interested readers are referred to Vanhoucke (2010a, 2014). Note that these values are calculated as the averages (for a certain sector) of the average SPI/SPI(t)/CPI/p-factor over all control periods (# Periods) of the projects in that sector, and not as the average of the final values at the project finish.

4.3 Evaluation

It has been previously mentioned that the use of empirical data is often favoured over the generation of artificial data due to their automatic reality check and their strong link with practical relevance. This practical realism is certainly one of the main advantages of using real project data, and it should be fully exploited for research purposes. Surprisingly, the real advantage of empirical data is that it can be used to create artificial data for simulation experiments and the like, by transforming the historical data into statistical distributions. As mentioned earlier, no problem for static, but a challenge for dynamic data. The method of Trietsch et al. *(2012)* and the application of this method by Colin and Vanhoucke *(2015b)* have been mentioned earlier in section 3.2 on using dynamic project data and are methods that enable researchers to transform empirical data points into statistical distributions for artificial project progress experiments. These methods show the relevance and importance of empirical data for research, since they connect the advantage of the realism of empirical data with the power of generating lots of artificial data in computer experiments.

The use of empirical project data is however 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. Indeed, the periodic dynamic data of projects are used as triggers for actions, often presented in performance indices and key performance indicators using EVM methods and the like. These triggers should be used in a careful way and should enable the project manager to take actions and spend time and money to solve problems, but should also refrain the project manager from taking actions when the project indicators report false problem warnings. The recent approaches of action tolerance limits *(Colin and Vanhoucke, 2014)*, statistical project control *(Colin and Vanhoucke, 2015a)*, artificial intelligence methods *(Wauters and Vanhoucke, 2014)* and decision support systems for project control *(Hazır, 2015)* all need the three components of the dynamic scheduling framework to a certain degree, and are set up to facilitate and/or improve the ability and quality of these corrective actions. The main purpose of many of the research studies is to contribute, directly or indirectly, to this challenging goal and to present methodologies to better control projects in progress and improve corrective actions, hereby assuming that these improved actions result in an increasing level of project success. The major and inherent weakness of empirical data lies in this fundamental and crucial research goal, since these empirical data include many of these corrective actions - often unknown. Since the empirical and periodic data points have been collected by project managers in charge of real projects, their ultimate *(and probably only)* reason why they have collected the data in the first place was not to share it with researchers but to support corrective actions. However, it is difficult, if not impossible, to distinguish between data with or without actions, since these two scenarios are never available in reality, and the actions are often unknown and only very vaguely described when asked to the manager. Often times, researchers end up with empirical data that include *(unknown)* corrective actions.

Despite this inherent weakness of empirical data, the advantage of using the historical data for artificial academic use offers a major contribution on top of solely using artificial data points. Therefore, the classification scheme presented in the project cards methodology of previous section is conjectured to be only a first step in the search for more and richer empirical data that can be used in academic research and transformed to controlled artificial data that better reflect reality in order to bring the newly developed methodologies even closer to the needs of professional project management.

5. Conclusions

In this paper, an overview is given on the use and generation of project data for integrated project management and control that focuses on the construction of a baseline schedule, the analysis of the schedule risk and the use of project control performance measures along project progress. It has been shown that the research endeavours of the past decades have collected and generated data from various sources, sometimes under a carefully controlled design, other times serving the specific needs of a research study but with no potential to be used elsewhere in literature.

A distinction has been made between static and dynamic project data, reflecting the way the data are used in the project life cycle. While the static data rely on network and resource parameters that can be varied over various values to generate project data under a full factorial design, the dynamic data are subject to choices that ideally should reflect real project progress. While the use of statistical distributions based on theoretical knowledge and historical data has been widely investigated in the literature, there is still no unified approach available for dynamic data generation.

Furthermore, a second distinction has been made between the generation of artificial project data and the collection of empirical project data, and it has been shown that both have value since they serve different purposes. While the artificial data allow an easy and controlled design of parameters to fully test new methodologies in literature, the drawback is that their parameters might not reflect real settings that occur in reality. This is particularly relevant and critical for the dynamic project data, and less relevant for static project data. The use of empirical data overcomes these drawbacks, but there still lacks a unified approach for the generation of artificial data, which remains subject to random choices for selecting distributions that imitate reality.

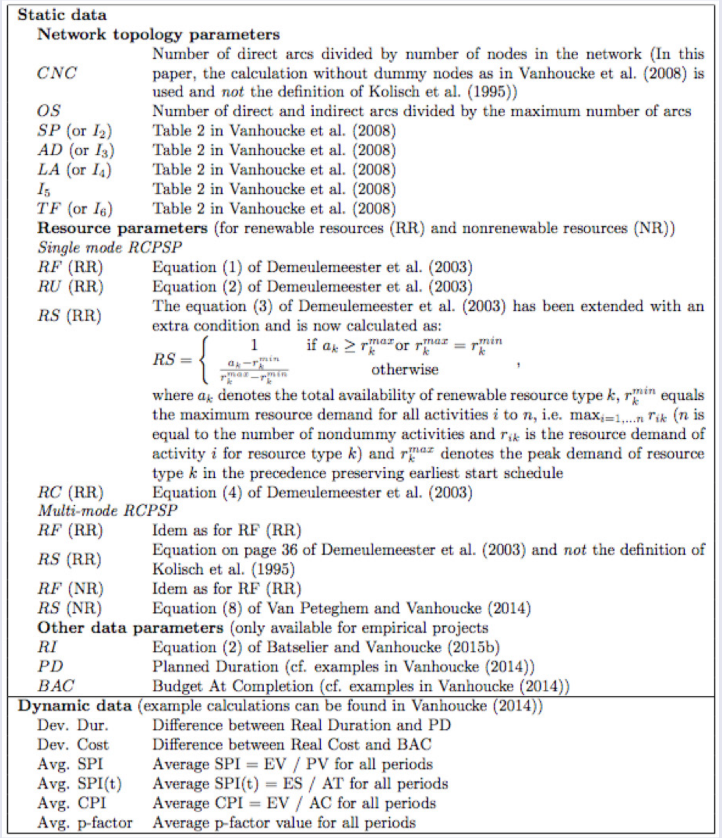
The previously mentioned shortcomings immediately define the future needs for project data generation to further enhance the research on integrated project management and control. A stronger synergy between empirical data and artificial data is necessary to increase the realism of research experiments. While this synergy is probably less critical for the static project data, it is undoubtedly crucial for the dynamic project data. While the generation of static project data under various settings for a set of parameters is likely to generate projects that also occur in practice, the imitation of project progress using artificial dynamic data is prone to errors or oversimplifications, and hence, to deviations from reality. A stronger link between empirical data and artificial data should reduce this risk by translating observations from reality into approaches for artificial data generation. While some efforts have been made in the past to propose mechanisms to derive dynamic data from empirical observations, it is believed that much more can and must be done in the future to formalize this artificial dynamic data generation process for use in an academic setting.

Acknowledgements

The support of the "Bijzonder Onderzoeksfonds" (BOF) for the project with contract number BOF12GOA021 is acknowledged and greatly appreciated.

APPENDIX A - LIST OF FORMULAS

This appendix (**Table 3**) gives a short and unambiguous overview of both the static data parameters used to measure the network topology and resource scarceness in section 3.1.1 and the dynamic data parameters used for the empirical data described in section 4.2. The table does not provide references to the original papers of the parameters, since these references have been mentioned throughout the text. Instead, it provides references to equations published in the academic literature so that the reader knows which formula has been used to calculate the network topology and resource parameters. This is particularly important for the static resource parameters, and more precisely, for the CNC and RS for renewable and nonrenewable resources used for the RCPSP and the MMRCPSP discussed in the paragraph "Summary table and critical remarks" of section 3.1.3 where it has been shown that alternative formulas are in use for the same parameter. These formulas have been used to calculate the values of the parameters shown in the summary **tables 1** and **2**. The values of all the parameters for the individual project files of each dataset are also available in an MS Excel file that can be downloaded from http://www.projectmanagement.ugent.be/research/data.

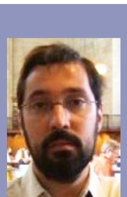


Static data	
Network topology parameters	
CNC	Number of direct arcs divided by number of nodes in the network (In this paper, the calculation without dummy nodes as in Vanhoucke et al. (2008) is used and *not* the definition of Kolisch et al. (1995))
OS	Number of direct and indirect arcs divided by the maximum number of arcs
SP (or I_2)	Table 2 in Vanhoucke et al. (2008)
AD (or I_3)	Table 2 in Vanhoucke et al. (2008)
LA (or I_4)	Table 2 in Vanhoucke et al. (2008)
I_5	Table 2 in Vanhoucke et al. (2008)
TF (or I_6)	Table 2 in Vanhoucke et al. (2008)
Resource parameters (for renewable resources (RR) and nonrenewable resources (NR))	
Single mode RCPSP	
RF (RR)	Equation (1) of Demeulemeester et al. (2003)
RU (RR)	Equation (2) of Demeulemeester et al. (2003)
RS (RR)	The equation (3) of Demeulemeester et al. (2003) has been extended with an extra condition and is now calculated as: $RS = \begin{cases} 1 & \text{if } a_k \geq r_k^{max} \text{or } r_k^{max} = r_k^{min} \\ \frac{a_k - r_k^{min}}{r_k^{max} - r_k^{min}} & \text{otherwise} \end{cases},$ where a_k denotes the total availability of renewable resource type k, r_k^{min} equals the maximum resource demand for all activities i to n, i.e. $\max_{i=1,\ldots,n} r_{ik}$ (n is equal to the number of nondummy activities and r_{ik} is the resource demand of activity i for resource type k) and r_k^{max} denotes the peak demand of resource type k in the precedence preserving earliest start schedule
RC (RR)	Equation (4) of Demeulemeester et al. (2003)
Multi-mode RCPSP	
RF (RR)	Idem as for RF (RR)
RS (RR)	Equation on page 36 of Demeulemeester et al. (2003) and *not* the definition of Kolisch et al. (1995)
RF (NR)	Idem as for RF (RR)
RS (NR)	Equation (8) of Van Peteghem and Vanhoucke (2014)
Other data parameters (only available for empirical projects	
RI	Equation (2) of Batselier and Vanhoucke (2015b)
PD	Planned Duration (cf. examples in Vanhoucke (2014))
BAC	Budget At Completion (cf. examples in Vanhoucke (2014))
Dynamic data (example calculations can be found in Vanhoucke (2014))	
Dev. Dur.	Difference between Real Duration and PD
Dev. Cost	Difference between Real Cost and BAC
Avg. SPI	Average SPI = EV / PV for all periods
Avg. SPI(t)	Average SPI(t) = ES / AT for all periods
Avg. CPI	Average CPI = EV / AC for all periods
Avg. p-factor	Average p-factor value for all periods

TABLE 3. *Summary of formulas for the static and dynamic parameters*

authors

Mario Vanhoucke is professor at Ghent University (Belgium), Vlerick Business School (Belgium, Russia, China) and University College London (UK) and guest professor in Peking University (China). He has a Master's degree in Business Engineering and a PhD in Operations Management. He teaches "Project Management", "Dynamic Project Planning", "Decision Making for Business", and "Business Statistics". He is an advisor for several PhD projects, has published more than 60 papers in international journals and is the author of four project management books published by Springer. He is a regular speaker on international conferences as an invited speaker or chairman and a reviewer of numerous articles submitted for publication in international academic journals. Mario Vanhoucke is a founding member and Director of the EVM Europe Association (www.evm-europe.eu) and partner at the company OR-AS (www.or-as.be). His project management research has received multiple awards including the 2008 International Project Management Association (IPMA) Research Award for his research project "Measuring Time – A Project Performance Simulation Study" which was received at the IPMA world congress held in Rome, Italy. He also received the "Notable Contributions to Management Accounting Literature Award" awarded by the American Accounting Association at their 2010 conference in Denver, Colorado.



José Coelho is a researcher at the Technical University of Lisbon (Portugal) and Professor in the Universidade Aberta (Open University in Lisbon (Portugal)) in the Department of Sciences and Technology where he teaches courses on computer science such as "Programming", "Introduction to Artificial Intelligence", "Programming of Digital Artefacts", and "Modern Heuristics". His main research areas are on project scheduling, e-learning and digital art. He has published 6 papers in international journals, 10 papers in national journals and conferences, has produced 11 software packages, and developed 8 didactic and pedagogical tools. He is a regular reviewer for international journals and member of the coordinating council of the Department of Sciences and Technology, the commission of the Doctor's Degree in Digital Media Arts, and vice-coordinator of the Master Degree in Graphical and Audiovisual Expression

Jordi Batselier holds Master's degrees in Civil Engineering (2011) and Business Economics (2012) from Ghent University (Belgium). Since 2012 he is working as a PhD researcher at the Operations Research & Scheduling research group of the Faculty of Economics and Business Administration of Ghent University. His research interest lies in project management, more particularly, in performing project control by means of earned value management. His specific research actions are focused on the empirical evaluation and development of forecasting techniques for project duration and cost, on which he has published several papers in international journals.

references

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

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

Batselier, J. and Vanhoucke, M. *(2015b).* Project regularity: Development and evaluation of a new project characteristic. Submitted to Journal of Systems Science and Systems Engineering.

Boctor, F. *(1993).* Heuristics for scheduling projects with resource restrictions and several resource-duration modes. International Journal of Production Research, 31:2547-2558.

Browning, T. and Yassine, A. *(2010).* A random generator of resource-constrained multi-project network problems. Journal of Scheduling, 13:143-161.

Colin, J. and Vanhoucke, M. *(2014).* Setting tolerance limits for statistical project control using earned value management. Omega The International Journal of Management Science, 49:107-122.

Colin, J. and Vanhoucke, M. *(2015a).* Developing a framework for statistical process control approaches in project management. International Journal of Project Management, 33*(6)*:1289-1300.

Colin, J. and Vanhoucke, M. *(2015b).* Empirical perspective on activity durations for project management simulation studies. Journal of Construction Engineering and Management, 04015047, 1-13.

Cooper, D. *(1976).* Heuristics for Scheduling Resource-constrained Projects: An Experimental Investigation. Management Science, 22:1186-1194.

Dar-El, E. *(1973).* MALB - A heuristic technique for balancing large single-model assembly lines. IIE Transactions, 5:343-356.

Davies, E. *(1974).* An experimental investigation of resource allocation in multiactivity projects. Operational Research Quarterly, 24:587-591.

Davis, E. *(1975).* Project network summary measures constrained-resource scheduling. AIIE Transactions, 7:132-142.

De Reyck, B. and Herroelen, W. *(1996).* On the use of the complexity index as a measure of complexity in activity networks. European Journal of Operational Research, 91:347-366.

Debels, D. and Vanhoucke, M. *(2007).* A decomposition-based genetic algorithm for the resource-constrained project scheduling problems. Operations Research, 55:457-469.

Demeulemeester, E., Dodin, B., and Herroelen, W. *(1993).* A random activity network generator. Operations Research, 41:972-980.

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.

Drexl, A., Nissen, R., Patterson, J., and Salewski, F. *(2000).* ProGen/πx - an instance generator for resource-constrained project scheduling problems with partially renewable resources and further extensions. European Journal of Operational Research, 125:59-72.

Elmaghraby, S. and Herroelen, W. *(1980).* On the measurement of complexity in activity networks. European Journal of Operational Research, 5:223-234.

Elshaer, R. *(2013).* Impact of sensitivity information on the prediction of project's duration using earned schedule method. International Journal of Project Management, 31:579-588.

Fleming, Q. and Koppelman, J. *(2010).* Earned value project management. Project Management Institute, Newton Square, Pennsylvania, 4th edition.

Hahn, E. *(2008).* Mixture densities for project management activity times: A robust approach to PERT. European Journal of Operational Research, 188:450-459.

Hartmann, S. and Briskorn, D. *(2010).* A survey of variants and extensions of the resource-constrained project scheduling problem. European Journal of Operational Research, 207:1-15.

Hazır, Ö. *(2015).* A review of analytical models, approaches and decision support tools in project monitoring and control. International Journal of Project Management, 33*(4)*:808-815.

Herroelen, W. and De Reyck, B. *(1999).* Phase transitions in project scheduling. Journal of the Operational Research Society, 50:148-156.

Hulett, D. *(1996).* Schedule risk analysis simplified. Project Management Network, 10:23-30.

Kaimann, R. *(1974).* Coefficient of network complexity. Management Science, 21:172-177.

Kaimann, R. *(1975).* Coefficient of network complexity: Erratum. Management Science, 21:1211-1212.

Kao, E. and Queyranne, M. *(1982).* On dynamic programming methods for assembly line balancing. Operations Research, 30:375-390.

Kolisch, R. and Sprecher, A. *(1996).* PSPLIB - A project scheduling problem library. European Journal of Operational Research, 96:205-216.

Kolisch, R., Sprecher, A., and Drexl, A. *(1995).* Characterization and generation of a general class of resource-constrained project scheduling problems. Management Science, 41:1693-1703.

Kotiah, T. and Wallace, N. D. *(1973).* Another look at the pert assumptions. Management Science, 20*(1)*:44-49.

Kuhl, M. E., Lada, E. K., Steiger, N. M., Wagner, M. A., and Wilson, J. R. *(2007).* Introduction to modeling and generating probabilistic input processes for simulation. In Henderson, S., Biller, B., Hsieh, M., Shortle, J., Tew, J., and Barton, R., editors, Proceedings of the 2007 Winter Simulation Conference, pages 63-76. New Jersey: Institute of Electrical and Electronics Engineers.

Kwak, Y. H. and Ingall, L. *(2007).* Exploring monte carlo simulation applications for project management. Risk Management, 9*(1)*:44-57.

Mastor, A. *(1970).* An experimental and comparative evaluation of production line balancing techniques. Management Science, 16:728-746.

Mohan, S., Gopalakrishnan, M., Balasubramanian, H., and Chandrashekar, A. *(2007).* A lognormal approximation of activity duration in pert using two time estimates. Journal of the Operational Research Society, 58*(6)*:827-831.

Pascoe, T. *(1966).* Allocation of resources - CPM. Revue Française de Recherche Opérationnelle, 38:31-38.

Patterson, J. *(1976).* Project scheduling: The effects of problem structure on heuristic scheduling. Naval Research Logistics, 23:95-123.

Ragsdale, C. *(1989).* The current state of network simulation in project management theory and practice. Omega The International Journal of Management Science, 17*(1)*:21-25.

Schonberger, R. *(1981).* Why projects are "always" late: A rationale based on manual simulation of a PERT/CPM network. Interfaces, 11:65-70.

Schwindt, C. *(1995).* A new problem generator for different resource-constrained project scheduling problems with minimal and maximal time lags. WIOR-Report-449. Institut für Wirtschaftstheorie und Operations Research, University of Karlsruhe.

Tavares, L. *(1999).* Advanced models for project management. Kluwer Academic Publishers, Dordrecht, 1999.

Tavares, L., Ferreira, J., and Coelho, J. *(1999).* The risk of delay of a project in terms of the morphology of its network. European Journal of Operational Research, 119:510-537.

Tavares, L., Ferreira, J., and Coelho, J. *(2002).* A comparative morphologic analysis of benchmark sets of project networks. International Journal of Project Management, 20:475-485.

Thesen, A. *(1977).* Measures of the restrictiveness of project networks. Networks, 7:193-208.

Trietsch, D., Mazmanyan, L., Govergyan, L., and Baker, K. R. *(2012).* Modeling activity times by the Parkinson distribution with a lognormal core: Theory and validation. European Journal of Operational Research, 216:386-396.

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

Van Peteghem, V. and Vanhoucke, M. *(2014).* An experimental investigation of metaheuristics for the multi-mode resource-constrained project scheduling problem on new dataset instances. European Journal of Operational Research, 235:62-72.

Vanhoucke, M. *(2010a).* Measuring Time - Improving Project Performance using Earned Value Management, volume 136 of International Series in Operations Research and Management Science. Springer.

Vanhoucke, M. *(2010b).* A scatter search heuristic for maximising the net present value of a resource-constrained project with fixed activity cash flow. International Journal of Production Research, 48:1983-2001.

Vanhoucke, M. *(2010c).* Using activity sensitivity and network topology information to monitor project time performance. Omega The International Journal of Management Science, 38:359-370.

Vanhoucke, M. *(2011).* On the dynamic use of project performance and schedule risk information during project tracking. Omega The International Journal of Management Science, 39:416-426.

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., 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., Demeulemeester, E., and Herroelen, W. *(2001).* On maximizing the net present value of a project under renewable resource constraints. Management Science, 47:1113-1121.

Wauters, M. and Vanhoucke, M. *(2014).* Support vector machine regression for project control forecasting. Automation in Construction, 47:92-106.

Willems, L. and Vanhoucke, M. *(2015).* Classification of articles and journals on project control and earned value management. International Journal of Project Management, 33:1610-1634.

Williams, T. *(1995).* A classified bibliography of recent research relating to project risk management. European Journal of Operational Research, 85:18-38.