European Journal of Operational Research 216 (2012) 386–396

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Modeling activity times by the Parkinson distribution with a lognormal core: Theory and validation

Dan Trietsch [a,*], Lilit Mazmanyan [a], Lilit Gevorgyan [a], Kenneth R. Baker [b]

[a] College of Engineering, American University of Armenia,Yerevan, Armenia
[b] Tuck School, Dartmouth College, Hanover, NH, United States

ARTICLE INFO

Article history:
Received 26 November 2010
Accepted 31 July 2011
Available online 10 August 2011

Keywords:
Project scheduling
Stochastic scheduling
PERT
Lognormal scheduling
Parkinson distribution
Linear association

ABSTRACT

Based on theoretical arguments and empirical evidence we advocate the use of the lognormal distribution as a model for activity times. However, raw data on activity times are often subject to rounding and to the Parkinson effect. We address those factors in our statistical tests by using a generalized version of the Parkinson distribution with random censoring of earliness, ultimately validating our model with field data from several sources. We also confirm that project activities exhibit stochastic dependence that can be modeled by linear association.

1. Introduction

Most published work on stochastic scheduling falls into one of two broad categories: (1) machine scheduling models and (2) project scheduling models. A key component of such models is a probability distribution for processing times or activity durations. Machine scheduling models tend to rely on the exponential distribution or the normal distribution. The exponential distribution often yields elegant results in problems that cannot be solved analytically for generic distributions (e.g., Bruno et al., 1981; Ku and Niu, 1986). The normal distribution is consistent with assuming that processing times are sums of numerous independent components of uncertainty so that the central limit theorem applies (e.g., Soroush and Fredendall, 1994). Project scheduling models, since the seminal work of Malcolm et al. (1959), have mostly relied on the beta distribution because of its flexibility and a claim that it is easy to estimate (Clark, 1962).

In this paper, we advocate the use of the *lognormal* distribution as a model for processing times and activity durations. We enumerate the various theoretical properties that support the use of the lognormal for both machine scheduling and project scheduling models, although our primary concern lies with the latter.

For the most part, the choice of a probability distribution for machine scheduling or project scheduling seems to be driven by convenience rather than empirical evidence. Efforts to validate assumptions about processing time distributions are scarce. For example, we have found no evidence in the literature that the beta distribution has *ever* been validated. Some progress has been made with data on surgery times (May et al., 2000; Strum et al., 2000), showing that the lognormal distribution provides the best fit by far. However, machine scheduling models have rarely considered the lognormal distribution (an exception being Robb and Silver, 1993). In this paper, we validate the use of the lognormal distribution as a model for activity times in several independent datasets obtained from project scheduling applications. By contrast, the beta distribution and the exponential distribution would fail in most of these cases.

Two practical issues arise in attempts to validate a particular probability distribution. One factor is the “Parkinson effect,” which is especially relevant in project scheduling: *reported* activity times may violate lognormality because earliness is hidden, not because the lognormal is a poor model. In other words, activities may finish earlier than estimated, or earlier than a given deadline, but there may be no incentive to report any outcome other than finishing on time. In such cases, the reported data contain a bias that obscures the underlying distribution. A second factor is that empirical data may be collected on a coarse time scale, leading to rounding of the actual times. However, rounding may cause false rejection of lognormality in standard tests, such as Shapiro–Wilk (Royston, 1993). These problems may explain why the lognormal has not been widely adopted for machine scheduling applications as well as project applications. In our validations, we recognize the Parkinson effect and the consequences of rounding. We introduce

[*] Corresponding author.
E-mail address: dan.trietsch@gmail.com (D. Trietsch).

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.07.054

a new version of the Parkinson distribution that helps diagnose whether the effect is present and makes possible accounting for it in simulation. In addition, we use statistical tests that account for the presence of ties occurring on a coarse time scale.

The vast majority of papers in both machine scheduling and project scheduling also rely on the assumption of statistical independence, but that is a very strong assumption. One serendipitous feature of the lognormal distribution is that it lends itself to use when statistical dependence is modeled by linear association (Baker and Trietsch, 2009a). In this paper, we also validate the linear association model for representing dependencies in empirical data, ultimately justifying linearly-associated lognormal processing times with different means but the same coefficient of variation.

Our results are relevant to both practitioners and theoreticians. The relevance to practitioners is direct: they can implement easier, more reliable stochastic estimates by our approach. The relevance to theoreticians is by informing stochastic scheduling models, such as the stochastic resource constrained project scheduling (SRCPS), which has attracted increasing attention over the last decade. Historically, SRCPS focused on minimizing the expected makespan under the earliest start policy, but that is not considered sufficient today (Demeulemeester and Herroelen, 2002). Most contemporary SRCPS models start with deterministic sequencing and include timing decisions that account for stochastic variation. The purpose is to obtain *proactive* schedules that hedge for variation (Herroelen and Leus, 2005). Hedging requires specifying safety time buffers. Models that study the tradeoff between minimizing the makespan (by reducing hedging) and achieving a stable or predictable schedule (by increasing hedging) are also known as *robust*. In both project and machine shop environments, however, we may expect deviations from plan during execution. *Reactive scheduling* models address the correct response (Aytug et al., 2005). The purpose of hedging is to reduce the expected cost during the reactive stage.

Some proactive models do not require explicit distributional information, opting instead to allocate some arbitrary amount of safety time to the schedule in some predefined configuration. A practical heuristic for the allocation of safety time in projects is proposed by Pittman (1994). Goldratt (1997) promotes Pittman's heuristic (and other ingredients developed by Pittman) as the basis of Critical Chain scheduling. This heuristic is perhaps the best known approach to setting safety time buffers that does not require distributions. However, there is no field evidence that Critical Chain provides sufficient protection—our own data suggests it does not, because it lacks calibration. Emphatically, it is impossible to test such models without distributions. One of our empirical results that is important for theoreticians is that such testing must allow for a much higher coefficient of variation [*cv*] than is usually the case in the literature. Trietsch (2005) argues against addressing stochastic variation without stochastic analysis. In response, Ash and Pittman (2008) combine Pittman's heuristic with standard PERT distributions. Trietsch (2006) proposes a proactive timing approach to minimize total weighted flowtime costs, which also relies on explicit distributions (and allows stochastic dependence). Since the makespan is a flowtime, the model is more general than minimizing makespan, and it can also include a tardiness penalty. The model addresses the reactive stage cost indirectly, through the flowtime earliness and tardiness cost parameters. In contrast to earlier timing models, this approach makes no attempt to set explicit safety time buffers but instead sets planned release dates for each activity. Those release dates must satisfy a generalized newsvendor model. Bendavid and Golany (2009) solve that model by cross entropy. Baker and Trietsch (2009a) demonstrate that it is possible to find optimal release dates for any given sequence and any simulated sample in polynomial time. Dablaere et al. (2011) propose a very similar model for setting release dates and also use the newsvendor model and simulated samples. A related model maximizes net present value instead of minimizing weighted flowtime, and may also be addressed by setting release dates and optimizing them for a simulated sample (Wiesemann et al., 2010). Conceptually similar stochastic models involve setting due dates instead of release dates (e.g., Baker and Trietsch, 2009b).

In summary, the models we cite either rely on distributions or use them for testing, so identifying correct distributions and showing how to estimate their parameters is crucial. The fact that we validated the prevalence of stochastic dependence and the ability to model it by linear association is also important for future theoretical research on proactive models even if we ignore the Parkinson effect. If we do not ignore it, the introduction of the Parkinson distribution is a theoretical contribution of this paper. Another theoretical contribution is the introduction of a bootstrap simulation approach that facilitates scheduling new projects based on historical data. We also prove that the lognormal distribution can be used to represent ratios of actual time to estimated time even though they are not independent random variables.

Section 2 provides background information from the project scheduling literature and discusses published activity distribution models for PERT. Section 3 presents theoretical arguments for selecting the lognormal distribution as a model for activity time. Section 4 presents empirical support for that choice, based on data from several sources, including Trietsch et al. (2010)—which is an unpublished earlier version of this paper that we now use mainly for this purpose. In Section 5 we show how to account for the Parkinson effect and for ties in the data. Section 6 demonstrates that the stochastic dependence we encounter in our datasets can be modeled effectively by linear association. Section 7 contains our conclusion.

2. Activity time distributions

The basic tools used in project scheduling are two overlapping approaches introduced in the late 1950s: Critical Path Method (CPM) in Kelley (1961) and Program Evaluation and Review Technique (PERT) in Malcolm et al. (1959) and Clark (1962). Of the two, only PERT recognizes the probabilistic nature of activity times within a project. The deterministic assumption of CPM facilitated the developments of time/cost models (crashing) and of sequencing models that become necessary when resources are constrained (Demeulemeester and Herroelen, 2002). At the heart of the PERT method is a set of assumptions that facilitates a systematic and intuitively-appealing method for modeling stochastic behavior in projects. In this paper we address the most basic element of PERT: fitting a distribution to each individual activity time. According to Clark (1962), the beta distribution has the necessary flexibility, and a good way to estimate its parameters is by eliciting three values, an approach we call the *triplet method*. In particular, estimates are elicited for the minimal possible value (denoted *min*), the mode (*mode*), and the maximal possible value (*max*). These estimates are then used to define the mean μ, and the standard deviation, σ, using the formulas

$$\mu = (min + 4mode + max)/6, \tag{1}$$

$$\sigma = (max - min)/6. \tag{2}$$

Eq. (2) was selected arbitrarily, to resemble a truncated normal between ±2.96σ. Eq. (1) was then derived as an approximation for a beta distribution that matches Eq. (2) and the estimated *min*, *mode*, and *max* values. Clark (1962) states, "The author has no information concerning distributions of activity times, in particular, it is not suggested that the beta or any other distribution is appropriate." Furthermore, theoretically, with three exceptions—noted by Grubbs (1962) as part of a scathing critique of PERT—no beta distribution fits both estimators. Nonetheless, Eqs. (1) and (2) provide good