Science of Computer Programming 74 (2009) 934–988

Contents lists available at ScienceDirect

Science of Computer Programming

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

Quantifying IT forecast quality

J.L. Eveleens *, C. Verhoef

VU University Amsterdam, Department of Computer Science, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 27 August 2008
Received in revised form 3 September 2009
Accepted 8 September 2009
Available online 12 September 2009

Keywords:
Forecast
Estimation
Cone of uncertainty
Reference cone
Forecast-to-actual ratio
Estimating Quality Factor
IT portfolio management

ABSTRACT

In this article, we show how to quantify the quality of IT forecasts. First, we analyze two metrics previously proposed to analyze IT forecast data—Boehm's cone of uncertainty and DeMarco's Estimating Quality Factor. We show theoretical problems with the cone of uncertainty (for example, that the conical shape of Boehm's cone is not caused by improved estimation, but can also be found when estimation accuracy decreases), and generalize it as a family of distributions that predict IT forecasts on the basis of expected accuracy and predictive bias. With these, we support decision making by providing critical information on IT forecasting quality to IT governors. We illustrate that plotting forecast-to-actual ratios against a predicted distribution reveals potential biases, for instance political, involved with IT forecasting. We illustrate our approach by applying it to four real-world organizations (1824 projects, 12 287 forecasts, 1059+ million Euro). We show that the distribution of forecast to actual ratios vary between organizations in at least three dimensions: in accuracy of estimation, in the tendency of forecasts to converge to the actual over the life of the project, and in systematic bias toward over- and underestimation. Moreover, we illustrate how to use the information to enrich forecast information for decision making. Finally, we point out that systematic biases, if not accounted for, make meaningless often-quoted rates of project success. We survey benchmarks related to forecasting and propose new benchmarks based on our extensive data.

The estimator's charter is not to state what developers should do,
but rather to provide a reasonable projection of what they will do.
–T. DeMarco [14]

1. Introduction

In large organizations, each year hundreds of IT projects are proposed. Due to limited time and budget, only a selection of these proposals receives funding. Decision makers must decide which of these projects are preferable given some qualitative and quantitative criteria, such as costs, durations and benefits of the projects. Namely, these criteria determine the value the proposals are expected to generate for the organization. An important part of the information IT governors have to support this decision making are forecasts. Therefore, forecasting quality is crucial. The forecasts also serve as budget and time constraints in many occasions.

Given the criticality of forecasting quality, it is surprising that almost no routine audits are being carried out on the quality of forecasting information itself. In this article, we propose how to quantify IT forecasting quality, so that IT governors know how good forecasts are and what bias they can expect. Often, IT executives and estimators have different interests in the forecasts. This causes these forecast to be used by them both for political reasons. This conflict of interest between the

* Corresponding author. Tel.: +31 0205987782.
E-mail addresses: laurenz@few.vu.nl (J.L. Eveleens), x@cs.vu.nl (C. Verhoef).

0167-6423/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.scico.2009.09.005

executives and estimators is not always clearly visible to the former, but can have significant impact on the quality of the forecasts. Therefore, quantifying the IT forecasting quality and the bias provides crucial information for executives to help steer their organization. We discuss how to visualize and quantify the accuracy of forecasts. This allows executives to fund those projects that minimize the risk of unexpected and unwanted underruns or overruns actuals.

Initial forecasts are expected to be inaccurate up to some point, yet these forecasts are extremely valuable in substantiating the go/kill decisions. The quality of the forecasts found in business cases is crucial to support the decision to fund or postpone a project. These early forecasts are significantly influenced by uncertainties, still they have substantial impact on the governing of the IT portfolio. Therefore, IT governors are helped by any tool aiding them in dealing with these forecasting problems. In this article, we show how to quantify the accuracy of early forecasts to enhance forecasting information to help executives with making decisions.

In some organizations, more detailed forecasts are made as projects progress to monitor them. These forecasts are expected to be more accurate than the initial forecasts as more information about the project is known and less time remains. If the quality improves over time, this certainly helps in more accurate monitoring, and gives the IT governor the tools to change plans if the expectations are not in alignment with the goals. IT governors often assume that forecasting information is void of politics and of good enough quality to steer. Intuitively, forecasts made later in a project are of higher quality. But is this true? And how do we quantify this? We will answer such questions in this article.

Even though the quality of forecasts is utterly important, we found that companies usually have no idea about their actual quality. Often, forecasts are assumed to be accurate but this is not assessed. For single projects, deviations between forecast and actual are expected due to unforeseen circumstances. It is commonly assumed that these effects will cancel each other out at the portfolio level, thus allowing decisions to be made on the aggregates. However, if all forecasts are overly optimistic or pessimistic the effects will not cancel each other when aggregated to the portfolio level. On the contrary, the impact of the deviations is amplified, degrading the quality to unacceptable levels. This results in steering on arbitrary numbers instead of data that presumably models reality in the future. In turn, this leads to gross under- or overfunding and thus to missed opportunities, since capital is wasted.

Therefore, it is critical to determine the quality of forecasts to assess their accuracy. There are two well-known tools dealing with this assessment. We discuss the merits and limitations of both these tools in this article. One tool, the Estimating Quality Factor (EQF) [14], is used to quantify the quality of forecasts of the estimators. As DeMarco stated: "The estimator's charter is not to state what developers *should* do, but rather to provide a reasonable projection of what they *will* do." DeMarco developed this tool to quantify the deviation between the projection of the estimators and the actual. The other tool is based on the so-called *cone of uncertainty* by Boehm [6]. This figure depicts deviations between forecasts and actuals by plotting forecast to actual ratios for different phases of a project. The plot shows a symmetric conical shape indicating a decrease in the deviations as a project progresses.

In this article, we illustrate how to generalize the cone of uncertainty to quantify certain quality aspects of IT forecasting. We show that a predefined referential cone visually assists in evaluating the differences between forecast and actual. For instance, a plot of forecast to actual ratios drawn with a reference cone can show forecasts are made of the minimum value instead of true predictions of the actual. Boehm was the first to describe the conical effect that was later on confirmed by others. However, we also found other shapes including wildly different ones up to the case where no conical shape whatsoever emerges. It turns out that depending on the bias of the forecasts, mainly the political undercurrent, different shapes emerge. In Boehm's case, the goal was to forecast the actual as quickly and accurately as possible. However, if the goal of the forecaster is, for instance, to lure one into a positive decision, they can provide for consistently low forecasts. This leads to a different shape than forecasts without political bias, as in Boehm's case.

Especially, when one never assesses IT forecasting quality, the bias of the forecasts can lead to extreme situations. We found in one case, forecasts up to 100 times the actual value where IT governors up to the highest level assumed the forecasts to be accurate. Needless to say that this is an entirely unwanted situation. In this article, we propose a method to reveal the bias by making the deviations of the forecast from the actual transparent.

Case studies. In this article, we analyze the IT forecasting quality of four large organizations. One organization is a vendor of commercial software, the second is a large multinational company, the third is a large multinational financial service provider and the last company is a telecommunications organization. The data of all organizations in total consist of 1824 projects with total actual costs of 1059+ million Euro and 12 287 IT forecasts. Compared with the 24 [8] or 25 [7] forecasts used to support Boehm's cone of uncertainty, the 20 data points used for DeMarco's EQF [14] and the 106 projects analyzed by Little [49] for both topics, our case studies form a sizable addition to the research done in the literature. We apply our approach to each organization to quantify the quality of their IT forecasts. We find that one organization grossly overstated its actuals. The second organization systematically underestimated them. The third organization was forecasting without bias and sustained high IT forecasting quality. The fourth organization has both forecasts without a bias but also forecasts that are too good to be true.

Benchmarks. Once it is known how to quantify the quality of IT forecasts, the question arises how it compares with its peers. For these comparisons, it is important that the same methods are used to assess the quality of IT forecasts and the benchmarks to be valid. In this article, we evaluate a number of benchmarks found in the literature related to IT forecasting. For instance, the often-quoted figures reported by Standish group [29–31] on project success are examined and turn out to be meaningless for benchmarking as we will argue in this article and elsewhere [18]. We propose ways to adequately compare

the IT forecasting quality with other organizations and describe valid benchmarks that are found in the literature and in our case studies.

Related work. IT forecasting has been an important topic for numerous years. Many methods have been developed to facilitate software cost estimation. COCOMO [6] and SLIM [61], for instance, are well-known methods. In an article of Briand et al. [9], an overview of the numerous authors is given, for instance [34,39,65], that have quantified the quality of software cost estimation methods including the forecasts made with such methods. The survey article [9] notes that many of these studies only have small data sets making it difficult to draw generalizable conclusions. An exception to this rule is the just referred overview article, where a total of 206 software projects are analyzed. In this article, we also analyzed a large sample of data regarding 1824 IT projects with 12 287 IT forecasts.

Note that the purpose of the authors surveyed in the article of Briand et al. [9] is to compare software cost estimation methods. The tools used to quantify the quality of IT forecasting are not subject to scrutiny, as in our article, and merely facilitate a way to make the comparison. In this article, the purpose of the quantification is to assist in governing IT. We illustrate what information IT governors are able to extract using our approach to improve decision making based on the forecasts. Therefore, this article describes more extensively how to quantify IT forecasting quality and challenges known results in this area.

For comparing software cost estimation methods, authors seldom use the EQF [14]. Only a small number of articles [47–49,64] refer to the EQF. We will argue that a popular method, the *magnitude of relative error* or MRE [13], has a severe drawback when compared with the EQF. Therefore, in this article, we elaborate on the EQF and its uses.

Although various authors [5,6,27,33,57,71] state that *the politics of forecasting* are highly influential for the forecasting quality, none have ventured to visualize their effects. In this article, we will show how this is done using our reference cone, which is a generalization of the well-known cone of uncertainty [6]. The cone of uncertainty is cited by many [12,38,43,50,52,67,71]. However, it is rarely assessed by others, whereas the results are based on one score of data points. The only exception we are aware of is the work done by Little [48]. In our article, we assess the cone of uncertainty as well and extend the analyses done by Little [48], by using his data in one of our case studies.

Finally, numerous benchmarks are given in the literature [4,5,14,29–32,47,52,57] related to the quality of forecasting. Yet none take into account the quantified effects of potential political or other biases. Therefore, many of these benchmarks are meaningless as it is unclear whether they are biased or how they are influenced (by politics). In this article, we illustrate how to incorporate the political nature of IT forecasting, which makes true comparisons possible.

Organization of this article. The remainder of this article is organized as follows. Section 2 introduces terminology that we use throughout the article. We also define what constitutes a high quality forecast. In Section 3, we elaborate on related work. In particular, we address various views on the cone of uncertainty that we found in the literature. Based on these views, we normalized the different notions used in those articles. We did this to unravel the nature of forecasting in order to carry out controlled simulation experiments. Our experiments either confirm or refute the various statements that were made in the literature. For instance, the intuition that forecasts improve because of improved accuracy of the estimation methods is refuted. Our simulations show that forecasts improve even with deteriorating accuracy of the estimation methods.

Section 4 discusses the tools that we use to quantify the quality of IT forecasts. The first tool, the Estimating Quality Factor or EQF, shows how to quantify the accuracy of the estimators. The second tool, the plot of the forecast-to-actual ratios against a reference cone, illustrates how the bias of forecasts is made visible. The tools combined provide the necessary information to quantify the quality of the IT forecasting practice inside an organization.

Apart from the simulation experiments, we also carry out four extensive case studies based on real-world data and present them in Section 5. We display several plots of forecast-to-actual ratios against a reference cone with different forecasting patterns, and provide the accompanying EQFs.

In Section 6, we describe how to use the analyses proposed in this article to enhance forecast information for decision making. We discuss three methods that provide additional information about the uncertainty of newly made forecasts.

Section 7 challenges the credibility of benchmarks related to forecasting in the literature. We review benchmarking information regarding EQFs and we challenge their credibility. We provide new benchmarks based on the four presented case studies: best, worst and mid-case benchmarks. Moreover, we show that the famous figures reported by Standish on project success are highly susceptible to political or otherwise biased forecasting. Therefore, these Standish figures are meaningless and we propose the use of modified definitions to recompute their benchmark figures.

On request of practitioners, in Section 8, we provide an overview of the lessons learned in this article. This section is especially focused on people who want to use our results. Moreover, it provides guidelines for practitioners to implement the proposed methodology and describes what information the methods will yield. It is possible to read this section without reading the previous sections. Finally, in Section 9, we conclude.

2. Terminology

In 2006, there was a lively discussion in IEEE Software [26] between Little, McConnell, Gryphon and Kruchten about the cone of uncertainty in general and Little's article [49] on this subject in particular. Since we not only found results similar to Little's, but also other results, we analyzed their discussion, which is cited and misunderstood by others [75]. We learned

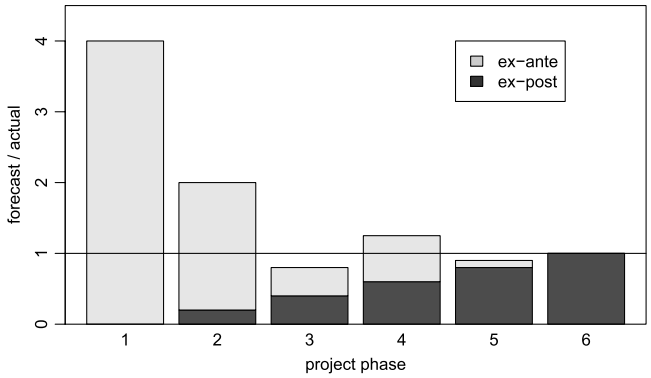

Fig. 1. Example forecasts.

that their terminology was rather subtle and that different authors meant different things by the same term. We feel this is not surprising as we ourselves had difficulty in defining the notions to clearly express the issues at stake in our article and in their discussion. To appreciate both our work and to be able to truly compare the different viewpoints put forth in their discussion, we introduce a standardized terminology. We transpose various views to this terminology in order to assess their validity. In this section, we also discuss and define a number of other terms that we will use throughout the article.

Forecast. We define a forecast as the total of the ex-post and the ex-ante part. The ex-post part is the part of the forecast that is already known—it is what has been done thus far. In many situations, one is able to measure this. The ex-ante part is a prediction of what lies in the future. For example, a running IT project has already burned cost and cost that are still to be made. The burned cost are the ex-post part and the future cost form the ex-ante part; their sum is the forecast of the total IT cost. In other words, a forecast (or forecasting) of a certain entity is a prediction of that entity.

In this article, the entities that we will analyze and focus on are the cost, and duration and functionality of IT projects, but our proposed approach applies to any kind of indicator with positive value. Almost any IT indicator satisfies the positive value restriction. For instance, the methods we will propose are also applicable when analyzing forecasts of effort or staff level. However, it is not possible to apply our approach directly to an indicator that is able to assume positive and negative values, such as the Net Present Value. Of course, it would be very interesting to use similar methods on the latter, since the most compelling reason to fund an IT-investment is its added value. But one must keep in mind that this value is determined by the cost, duration and benefits of a project. Therefore, the quality of such forecasts should be analyzed in order to assess the value a project will generate.

Both ex-post and ex-ante are phrases directly taken from Latin. Ex means 'from', post means 'after' and ante 'before'. In the Oxford dictionary [70], ex-post is defined as: based on actual results rather than predictions. Ex-ante is defined as: based on predictions rather than actual results. Although these terms are infrequently used in common English, we propose them as they precisely describe the necessary notions to appreciate our work and the related IEEE Software discussion. Both ex-post and ex-ante are also used in a book on forecasting [3], although there they have a slightly different interpretation.

Actual. We define an actual of a certain entity to be the final realization of that entity. For instance, if a project costs 1 million in expenses the actual is 1 million. In this article, we will make two assumptions about the actual. First, we assume the actual is objectively measurable and thus that manipulation of the final realization is not possible. In practice, this is not always the case. For instance, if a project runs out of budget it may happen that hours used for this project are booked on another project that has excess budget. In Section 5, we show that the methods we propose can also give an indication whether large scale manipulation is likely to have occurred in the data or not.

Second, we assume that IT governors want estimators to provide a prediction of the final realization. It is important to realize that, for the remainder of this article, we will use this final realization as reference to assess the quality of forecasts made. However, this need not always be the case. For instance, it is also possible that executives demand not a forecast of the final realization, but a conservative forecast, say the predicted final realization plus 20%. If these are the forecasts that are expected of the estimators, the actual must be set accordingly. Otherwise, the quality of the forecasts will not be fairly assessed. The methods we describe are easily adapted if another reference point is preferred.

Although it is possible to take other reference points at wish, we feel that the final realization is an adequate choice. If the estimators are able to accurately predict the final realization, executives can use the forecasts to derive any other reference point to their liking. For instance, if an IT governor wants a conservative forecast to determine the budget, the governor can use the forecast of the final realization and add a percentage to serve as budget. If the IT governor wants to determine budgets less conservatively, only the percentage needs to be changed. The estimators remain forecasting the final realization and are judged based on this reference point. Therefore, we have chosen the final realization as point of reference.

To illustrate our definitions, we provide in Fig. 1 six forecasts of some project indicator. Each vertical bar in the figure represents one of these forecasts. They consist of an ex-post and ex-ante part. For the first bar, the ex-post part is zero as nothing has been done at this point. As the project progresses, the ex-post part grows as work is done. The remainder of

Table 1
Translation of different views in terms of this article.

Author	Forecast	Ex-post	Ex-ante	Actual
Boehm	Estimate	–	–	Actual
McConnell	Estimate	–	–	Actual
Little	Estimate	–	Remainder	Actual
Gryphon	–	–	Estimate	–
Kruchten	Absolute overall uncertainty	–	–	–

work still to be done (= actual − ex-post) is predicted at each phase and forms the ex-ante part. The ex-ante part changes each time as less work remains. For the last bar, the ex-ante part is zero as all the work has been done. Together they make the forecast, which is the prediction of the total.

In Table 1, we transpose the various notions in the literature to the definitions we just proposed. The '–' indicates that the corresponding author did not define a term similar to the terms we defined. As our table shows, the notion of an estimate is used by different authors in different ways. For instance, Boehm, McConnell and Little discuss estimate in the sense of our forecast, but Gryphon uses estimate where he actually means the ex-ante part. Kruchten avoids the word estimate all together and talks of absolute overall uncertainty. Also, among the authors Little is the only one to label the ex-ante part of a forecast, but he did not label the ex-post part.

Ex-ante. In this article, we will assess the quality of the forecast, that is ex-post + ex-ante. However, an in-depth analysis of the ex-post and ex-ante portions individually is very insightful as well. A common re-estimation made is only of the ex-ante portion, for instance, "when will we be done". Or, requests for additional funds are predictions of just the ex-ante portion. The quality of these predictions are relevant to an IT governor to allow adequate monitoring of the project and to assign additional funds, if necessary. Therefore, analyzing the different parts of a forecast separately on top of the combined analyses we propose in this article, further enriches the assessment and knowledge of the quality of the forecasts made. It is recommended to undertake such analyses when these common re-estimations are made.

Note that it is possible to adapt our tools to perform analyses on solely the ex-ante part. If one is able to derive the actual remainder of work that is predicted by the ex-ante part, it is possible to draw the forecast-to-actual plot based on these data alone. That is, the forecast is replaced by the ex-ante part. This is, for instance, illustrated in an article by Little [49]. Moreover, it is possible to quantify the quality of these predictions with the EQF. While the tools remain similar, they then only analyze the ex-ante part.

Forecast quality. At this point, we define when we consider a forecast of a final realization to be better than another forecast. Let e be a forecast of actual a and f a forecast of actual b ($a, b > 0$). Forecast e is better than forecast f when

$$\frac{|e-a|}{a} < \frac{|f-b|}{b}.$$

In fact, we say that forecast e is better than f if the relative distance from the actual of e is smaller than that of f.

We also define when a project is better forecasted than another project. Let $e_1, e_2, \ldots, e_n$ be several forecasts made for a project k with actual a. Assume that there is some function $G_k = g(e_1, e_2, \ldots, e_n, a)$ that quantifies the quality of the forecasts made for that project. We assume that this function assigns a higher value to a higher quality of forecasts. For instance, it is possible to use functions such as the inverse of the median of the individual deviations to the actual, the inverse of the deviation of the initial forecast or some other function. Later on, we will use the EQF to instantiate function G_k. We define project k to be better forecasted than project l when $G_k > G_l$.

Since we will aggregate forecasts also to the portfolio level, we define when we consider a collection of projects, with their forecasting quality quantified by function G_k, to be better forecasted than other collections. Let E be a collection of p projects with their forecasting quality quantified by function G_k and F be a similar collection of q projects. We define $E = \{G_i : i = 1, \ldots, p\}$. The definition of F is similar. We define collection E to be better than collection F when the median of E is larger than the median of F.

Although other measures are possible, we used the median value as it divides the collection E in two. DeMarco [14, p. 14] advises to use the median value exactly for this reason. Additionally, in other publications, e.g. [69], we found the median value to be used to compare the quality of forecasts between organizations. Therefore, we find it more insightful than considering, for instance, the average. However, we will argue later on that it is best to not only use the median value, but also account for the variation of the forecasting quality.

Bias. Another term we will frequently use in this article is *bias*. In the Oxford dictionary [70], bias is defined as: a systematic distortion of a statistical result due to a factor not allowed for in its derivation. In our context, the statistical result is the forecast that is systematically distorted. In many articles, for instance [14,19,28,35,46,74], reasons for biases are given and discussed. For example, Lederer et al. [46] describe how different participants have different goals in IT cost forecasting and can thereby introduce political biases.

Another possible reason for a bias in forecasts are errors in the forecasting process, as described by Fairley [19]. For instance, suppose a forecasting tool is used that has certain parameters. In such a case, a bias can be introduced if the

parameters are not adequately set. However, it is also possible that estimators systematically feed the wrong data into the model, which introduces distortion.

Yet another reason is given in a book by DeMarco [14, p. 12], where he shows that a bias can be introduced if the ego of the estimator is involved in the project. As an example, he describes a simple experiment in which participants are asked to forecast their own performances on a trivial task. The same participants also predict the performance of others for the same task. The results show that a bias is introduced when the estimators predict their own performance, which is supported by others [28,35].

Weinberg et al. [74] discuss an experiment in which two seperate groups are given identical requirements, apart from the focus of their work. One group is given the task to create an efficient program using the least CPU time possible, while the other group should make a similar program in the least amount of time. Both are required to make a forecast of the time needed to complete. Weinberg et al. find that the latter group is more conservative in their forecast. They argue that meeting forecasts of an objective independent of those that are focused on, becomes less important. For instance, if the objective is to minimize the cost of the project, other objectives such as user satisfaction or minimal CPU time required by the program, become less stressed.

In a book by McConnell [53], a distinction is made between a conscious and an unconscious bias. A conscious bias is a distortion that is introduced intentionally. For instance, if an estimator wants to present an IT project positively in order to get it approved, the estimator may underestimate the cost or duration of the project. Or, if the estimators want ample budget they can overestimate the cost to assure enough funds. An unconscious bias is a distortion that is unintentional. For instance, with a forecasting tool parameter settings can unintentionally be inadequately set.

In this article, we want to make the bias in the forecasts, both conscious and unconscious, transparent by plotting the deviations from the actual against a reference cone. In many occasions in this article we will speak of a bias caused by political reasons. However, the methods we will present are not restricted to this type of bias. How to avoid biases is discussed, for instance, in articles by Arkes [1] and Harvey [28].

Target and commitment. An important reason for political bias is the difference between forecast, target and commitment. A number of authors [2,36,44,49,53,73] have clearly described these notions. A target is a statement of a desirable final realization. A commitment is an agreement to meet a target. A forecast is the most likely outcome of the final realization. Jørgensen [36] notes that many people mix up these terms, even within the same project.

Armour [2] describes what the difference between the terms entails. Ideally, estimators give forecasts including their probability distribution. This will allow executives to set targets and commitments based on the probability of being able to meet them. For instance, suppose project costs are forecasted to be 1 million Euro with a 50% chance and 0.8 million with a 30% chance. If an IT governor is willing to take extra risk, a commitment of 0.8 million can be agreed upon with, for instance, the customer or the project team. If the extra risk is not worth it, the executive can commit to 1 million. With a forecast and its confidence interval or probability distribution, it is possible for IT governors to assess the risk they take by setting certain targets or commitments. In Section 6, we will show how to derive the confidence interval and probability distribution of the forecasts.

Project related terms. Above we stated that we will analyze forecasts of a value of interest, for instance, costs and durations of IT projects. However, we will not precisely define terms such as cost and duration, as this is not relevant for the analyses we perform. In fact, it does not even matter if inconsistent definitions of the value of interest are made for different projects within an organization, as long as the forecast of a project is based on the same definitions as the value of interest of that project. In our case studies, we did find consistent definitions to be used within the organizations, but not necessarily between the different organizations. However, for quantifying the IT forecast quality and comparing them between organizations, it is not necessary to define these terms precisely or adhere to equal definitions.

Also, we will not provide definitions of the start and end date of a project. If function G_k is independent of time, these definitions are not relevant similar to the definitions of the value of interest. If the function is dependent of time, these definition are still irrelevant within an organization as long as they are used consistently. They are, however, of importance when comparing forecast quality between organizations. Even though the Estimating Quality Factor is a function dependent on time, we will not give definitions of the start or end date of a project, as this is outside the scope of this article. Furthermore, the EQF metric is robust as small deviations in the definitions do not significantly affect the value. In our case studies, we found relatively small differences in the definitions between the organizations. These differences do not significantly influence the comparisons made in this article.

Now that we set the terminology and precisely defined the notions that we will use throughout this article, we commence with comparing the different viewpoints on the cone of uncertainty found in the literature.

3. Reviewing different cones

A well-known result in software engineering economics by Boehm is the so-called cone of uncertainty. This result is discussed in his book [6, p. 310–313], which describes the methods and procedures for software cost estimation. We recall this famous result in Fig. 2. It intends to illustrate the accuracy within which software cost forecasts are made as a function of the level of knowledge that is available. As described in a number of articles [8,52], the interpretation of the cone is that when one knows more about a software project, your forecasts will improve.

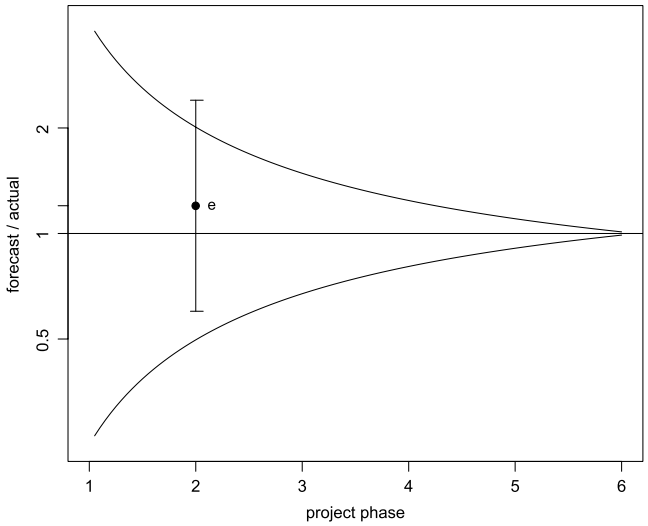

Fig. 2. Boehm's cone of uncertainty.

The horizontal axis in Fig. 2 represents time progression of the project in phases. The vertical axis compares the forecast with the actual by dividing the forecast by the actual value. This way, we see how much a forecast deviates from the actual at any given phase of the project. For instance, point *e* in Fig. 2 depicts a forecast made during the second phase of a project that turned out to be 1.2 times as high as the actual. The vertical axis is drawn on a logarithmic scale.

The vertical interval around *e*, depicted in Fig. 2, is known as a confidence interval. McConnell [52, p. 169] popularized the use of this interval for IT forecasts to give an indication of their uncertainty. Tockey [69, p.351–355] further explored the method by showing how to compute the uncertainty based on historical data. Others, for instance [75], have also used this confidence interval, which intends to contain the actual value in about 80% of the cases. We see that this is also the case in the example. Later on, we will discuss the usefulness and the limitations of the confidence interval.

Note that we will discuss forecast-to-actual ratios throughout this article. Therefore, we will refer to these ratios with the shorter term f/a ratio. This term is also used in other research areas [54]. When we plot f/a ratios, we refer to this figure as an f/a plot.

Different views. In a 2006 IEEE Software article [49], Little questions the implication of Boehm's cone of uncertainty [8] being that the uncertainty of the ex-ante part will diminish as the project progresses. In response to the article by Little, letters sent by Kruchten, McConnell and Gryphon [26] react to his findings. In this section, we address the issues raised by all four authors. Additionally, we will create different conical shapes using simulations and by doing so we provide insight in the discussion and clarify the different viewpoints. But first, we will summarize the various viewpoints of all authors involved.

- Boehm gives two implications of the cone of uncertainty in his book [6]. On the one hand, there is a need to be consistent in defining the objectives of the forecasts for the various parts of a software product. If the forecast of one part of a software system is based on a global design, it makes no sense to forecast another part of the same system based on a detailed design.

  On the other hand, the cone expresses that each forecast has some degree of uncertainty. Boehm argues that each forecast should include an indication of this degree of uncertainty. This is also underscored, for instance, by the articles of Cantor [10] and Laird [43].

  In later work [8], Boehm et al. add the implication that project uncertainties affect the accuracy of software cost forecasts. The more certain we are about the project, the more accurate we can forecast it.

  Although Boehm is right that the forecasting quality improves, this is only true when the goal of forecasting is void of bias. Forecasts improve as the project progresses as long as the goal of the forecasts is to predict the actual value as quickly and accurately as possible, which we will show in this section. Later on in Section 5, we will show that if there is a bias, forecasts do not necessarily improve. Recall that with forecasts we mean the total of the ex-post and ex-ante part.
- McConnell [52] endorses the viewpoint of Boehm that one can forecast a project more accurately as more knowledge becomes available. McConnell states in an article [26] and in his book [53, p.37] that Boehm's cone is however, only a best-case scenario and it is possible to do worse. McConnell describes that it is not possible to have consistently smaller bandwidths than the bandwidths given in Boehm's cone; a viewpoint that is being picked up by others [2]. According to McConnell, Boehm's cone therefore does not promise an improvement of the forecasts. In Section 5 and Section 6.2, we will show with our real-world cases that it is indeed possible to have larger bandwidths for the f/a ratios than Boehm's cone. However, we will also provide an example in which the range of the f/a ratios is consistently smaller than those of Boehm's cone, refuting McConnell's statement that Boehm's cone is a best-case scenario.

- Little describes that consecutive f/a ratios by definition converge to 1, but this does not mean the uncertainty of the ex-ante part decreases. He supports his claim by analyzing data from his company, which shows that the uncertainty of the ex-ante part is the same at each stage of the project. We corroborate with a simulation that Little is correct. The ex-ante part does not need to improve in order to obtain a conical shape.
- Kruchten asserts that Boehm's cone is about the "absolute overall uncertainty" or forecast in our terms and not about the uncertainty of the ex-ante part. Using a simulation, in this section we will show that Kruchten is right that Boehm's cone is about forecasts.
- Gryphon argues that Boehm's cone does not reduce by default, but it reduces because improved forecasting methods become available. In a report [76], the idea that the cone does not reduce by itself is even mentioned as a part of one of the ten most important ideas in software engineering. In this section, we will show that, in contrast to Gryphon's statement, improved forecasting methods themselves do not cause the reduction of the cone. It is even possible to acquire a conical shape with deteriorating forecasting methods as time progresses.

In the above discussion, it already appears that depending on the goals and conditions various outcomes are possible. Therefore, we continue to discuss what we call cone conditions.

3.1. Cone conditions

Boehm was one of the first to recognize a time-dependent converging effect of IT forecasts. However, the conditions under which this effect is present were neither systematically described nor investigated. We will reproduce conical shapes based on modest and reasonable assumptions by using simulation techniques, thereby investigating the phenomenon Boehm first observed. This implies that we need to make such assumptions not only explicit, but also executable. First, we will briefly enumerate all conditions and then we will discuss them in more detail. The conditions under which we are able to reproduce the conical shapes of Boehm's cone are as follows.

1. Completeness: We assume forecasts are made for an entire project.
2. Ex-post inclusion: We assume that each consecutive forecast incorporates the ex-post part and we assume this part is known with certainty.
3. Axis: The horizontal axis is a (relative) time axis. The vertical axis is the forecasted value divided by the actual value. This axis is drawn on a logarithmic scale.
4. Ex-post growth: The growth of the ex-post part at any time of the project is assumed to be determined by a function. We will use a constant function, which represents an evenly distributed growth of the ex-post part. However, if we look at labor, it is also possible to use, for example, a Rayleigh function [6,58].
5. Ex-ante accuracy: The accuracy of the ex-ante forecast is assumed to improve as the project progresses. Conventional wisdom indicates that we become better at estimating the ex-ante part as time progresses.
6. Symmetric ex-ante accuracy: We assume the accuracy of the ex-ante forecast to be the same in case of under- and overestimation.
7. Goal: The goal of the forecast is to predict without bias as quickly and accurately as possible the actual value of interest for the project.

Next, we elaborate on each summarized condition in more detail.

Completeness. The completeness condition indicates that we take into account all activities that relate to the value of interest. For instance, if we want to forecast the cost of a project, we also take into account all activities regarding the cost for creating the requirements and performing the business study.

One author [26] noted that iterative development of projects should be taken into account by analyzing each iteration separately. The first condition shows that this must only be done if the interest is in the value of each iteration. If the value of interest is the entire project, then taking into account all activities means that the forecast must include all iterations. As DeMarco [14] stated, the estimator should construct the forecast in such a way that the development method is reflected in it. However, in this article, we are not concerned how to construct a forecast but to quantify the quality of the resulting forecast. And from this perspective, the development method that is used is not relevant for the quantification as long as the forecasts are correctly compared with the actual of interest.

For example, suppose a project is executed in five iterations. At the start, a forecast is made for the cost of the entire project. After each iteration the forecast is then revised. In this case, both the initial forecast and the revisions are compared with the actual of the entire project. In our analysis of the f/a plot, this is viewed as one project with five consecutive forecasts.

Another example is a project that also takes five iterations. However, for this project forecasts are made at the start of each iteration for the cost of that single iteration. When analyzing the f/a plot, each of these forecasts is compared with the actual of the corresponding iteration. This means that when the project is finished, we have five different projects with an initial forecast that can be analyzed with an f/a plot. The difference between the examples is the value of interest of the forecasts.

Ex-post inclusion. The ex-post inclusion condition assumes that at any given moment during the project, information is available on what has been done so far. This represents an increase in knowledge. As the project progresses, we know more of the project and we know better what we have actually done so far. For instance, an ongoing project is spending money. Ex-post inclusion implies that we know at all times how much was already spent with certainty.

Axis. The axis condition explains the axes we use. In our simulation, we will use percentage of completion of the total duration of the project as opposed to the phases that are used by Boehm. Both are a representation of time. The phases used by Boehm do not need to be evenly spaced as percentage of completion does. The conical shape of the figure is not influenced however, by which one of the two time axes is used. It will merely be stretched or compressed in certain places.

The reason we opted for percentage of completion instead of the phases of a project is that our choice allows for better comparisons of forecasts. Let us elaborate on why this is so. We have extensive data, as described in another article [17], on how long phases of IT projects take as a percentage of the total. We found that the software development projects have considerable variation in the start and end of each phase, in terms of percentage of completion and have large overlaps between different phases. For instance, if we take into consideration forecasts made during the business study of different projects, this could mean for one project that it is made in the early stages of the project while no other phases have been started. In another project, it could mean the project is almost halfway and a number of other phases are already underway. The amount of knowledge available for both projects is thus completely different. The ex-post part of the first project is much smaller than that of the second project and the ex-ante part that still needs to be estimated is much larger than that of the second project.

By using percentage of completion, we make a comparison between forecasts that is more fair, as the forecasts are made when projects have had the same amount of time to work. However, this still does not mean that by working with percentage of completion the same amount of effort was carried out. In order to achieve this, we would have to normalize based on the effort. Since this information often is not readily available, in this article we consider only duration.

In the other article [17], we mapped phases to percentage of completion. Following that research, one can switch back and forth between the representation of time in phases and percentage of completion. We will do this later on to compare our results with those of Boehm's.

Ex-post growth. The ex-post growth condition means that the growth of the ex-post part is described by a function. We assume to know the ex-post part with certainty, and we use a function to determine the growth and thus the size of the ex-post part at any time during the project. Mathematically speaking, the function of the growth is the derivative of the size function of the ex-post part. Although the ex-post part is deterministic, the uncertainty remains in the ex-ante part of the forecast, as it is uncertain how the project will progress from that point onward. In a simplest case, we will assume a constant growth of the ex-post part. If the project is completed one-fifth, the ex-post part is one-fifth of the total as well. Another possibility is to assume a Rayleigh function, as effort of IT projects sometimes follow this distribution. In our investigations, we confine ourselves to using the constant growth function, but we also show that the same effects are found with a Rayleigh function.

Ex-ante accuracy. The ex-ante accuracy condition states that as the project progresses, the accuracy with which the ex-ante part is forecasted improves. Below, we will argue that this condition itself does not cause the conical shape of Boehm's cone. If the accuracy of the ex-ante part is constant or even decreases, we are also able to reproduce the conical shape.

Symmetric ex-ante accuracy. The symmetric ex-ante accuracy condition assumes that the ex-ante accuracy is symmetric on a logarithmic scale around the actual value. This means that for both under- and overestimation, the ex-ante accuracy is the same. If the maximum underestimation is twice as low, the maximum overestimation is twice as high. We assume this as Boehm's cone of uncertainty shows symmetry of the forecasts, also at the beginning where the ex-post part is zero. Therefore, it must apply to the estimation accuracy of the ex-ante part as well. Later on, we will illustrate more general assumptions for the ex-ante accuracy.

Goal. The goal condition is about the political undercurrent or other biases of forecasts. In this section, we assume that there is no bias, the forecast is meant to predict the actual as quickly and accurately as possible. As DeMarco showed in his book [14], forecasts mean different things to different people. For instance, project managers can perceive a forecast as a means to get enough budget, whereas their superiors can see it as the least amount of money necessary to perform the task. We will see such differences of perception in our real-world case studies.

3.2. Simulation

Based on the above assumptions on cone conditions, we carried out several simulations and it turned out that we were able to reproduce Boehm's cone of uncertainty. In fact, this provided us with a means to vary certain conditions and assumptions in order to test the various viewpoints of the different authors in their discussion about Boehm's and Little's cones.

It is possible to perform such simulations with any statistical package. In our case, we conducted the simulations with the statistical package R [62]. To give an idea on how much effort is invested in constructing such simulations, we provide below the R-code used to create Boehm's cone.

```
#Simulation of Boehm's cone of uncertainty
accuracy = 4
total = 100
n.of.draws = 1000
x = numeric(0)
ex.ante = numeric(0)
forecast = numeric(0)

for(i in 0:total) {
   ex.post = i  #assume constant growth function
   draws = (runif(n.of.draws) *
            (accuracy - 1/accuracy) + 1/accuracy)
   ex.ante = draws * (total - ex.post)
   forecast = c(forecast, (ex.post + ex.ante)/100)
   x = c(x, rep(i, n.of.draws))
}
plot(x, forecast, log = "y")
```

In this code snippet, at each percent of the project 1000 `forecasts` are made. Each of these `forecasts` in this simulation have a constant ex-ante `accuracy` of 4 as the project progresses. This means the `ex-ante` part is estimated to fall in the interval of 1/4th to 4 times its actual value. We use a factor of 4 since the forecasts at the beginning of the project in Boehm's cone use this estimation accuracy. As the ex-post part is zero at the beginning of the project, this factor applies to the estimation accuracy of the ex-ante part. The `ex-post` part is assumed to grow constant.

It is a common interpretation that the conical shape is caused by improved estimation accuracy of the ex-ante part, as time progresses. We will show that the conical shape is also reproducible in other situations. In particular, we analyzed three scenarios: the ex-ante accuracy increases, is constant or even *decreases* as time progresses. They all lead to Boehm's cone, albeit asymmetric around the actual value on a logarithmic scale as shown in Fig. 3. The asymmetry around the actual value of the cone of uncertainty is also observed by Laranjeira [45] and Cohn [12]. In fact, we will explain that Boehm's symmetric shape is theoretically reproducible, namely under very specific circumstances, but that these circumstances are highly unlikely to occur in practical situations.

To be more precise, we assumed the following:

- In the first plot of Fig. 3, we assumed the ex-ante accuracy to increase as time progresses. In this theoretical model, we assume the accuracy to increase linearly using the formula $accuracy = 4-2i/100$, where i is the percentage of completion of the project.
- In the second plot, we took the ex-ante accuracy to remain constant as time progresses. In this theoretical model, we assume the accuracy of the ex-ante part to remain 4 at all times. The above code snippet represents exactly this model.
- In the third plot of Fig. 3, the ex-ante accuracy *decreases* as time progresses. We used as theoretical model a linear decrease of the accuracy using the formula $accuracy = 4 + 2i/100$, where i is as mentioned before.

Of course, the above theoretical models are only meant to investigate whether any conical shape emerges. They are not necessarily real-world scenarios. What can be seen right away from the results of the simulation as visualized in Fig. 3, is that irrespective of the chosen accuracy scenario the conical shape is present. Whether the ex-ante accuracy increases, is kept constant, or decreases the forecast accuracy improves.

The clarification lies in the ex-post inclusion assumption, being that the ex-post part is included in the forecast. Let us explain: assume that one forecasts software cost. As the project progresses one's knowledge improves to know what has been spent, which is the ex-post part. This, one do not needs to predict. The remainder of the work to be done, which is the ex-ante part, is what one predicts. Even if the quality of this prediction decreases (to some extent), there is still convergence to the actual, since the part we know (ex-post) becomes larger and larger and the part we need to predict (ex-ante) becomes smaller and smaller. The effect of decreasing ex-ante accuracy is compensated by actual knowledge of the ex-post part.

3.2.0.1. Symmetry. We noted that Boehm's cone of uncertainty assumes symmetry on a logarithmic scale around *the actual value*. Our simulations do not reproduce this symmetry. The upper parts in the simulation, given the conditions, appear to be curving outwards with respect to the f/a ratio of 1 where Boehm's upper part was curving inwards. This asymmetry around the actual value is also caused by using the ex-post part available at each stage of the project. By adding the ex-post part to the ex-ante part, the symmetry of the forecasts around the actual value disappears.

Let us explain. For example, assume that we forecast the cost of a project. The actual cost of the project is $100. Assume that we have spent $50 so far. Thus, the ex-post part equals 50. We predict the ex-ante part with an accuracy of factor 2. This means we are able to estimate the ex-ante part in the range of 1/2 to 2 times the actual. Thus, we will estimate the ex-ante part, also being $50, somewhere in the range of 25 to 100. This range is symmetric on a logarithmic scale around 50.

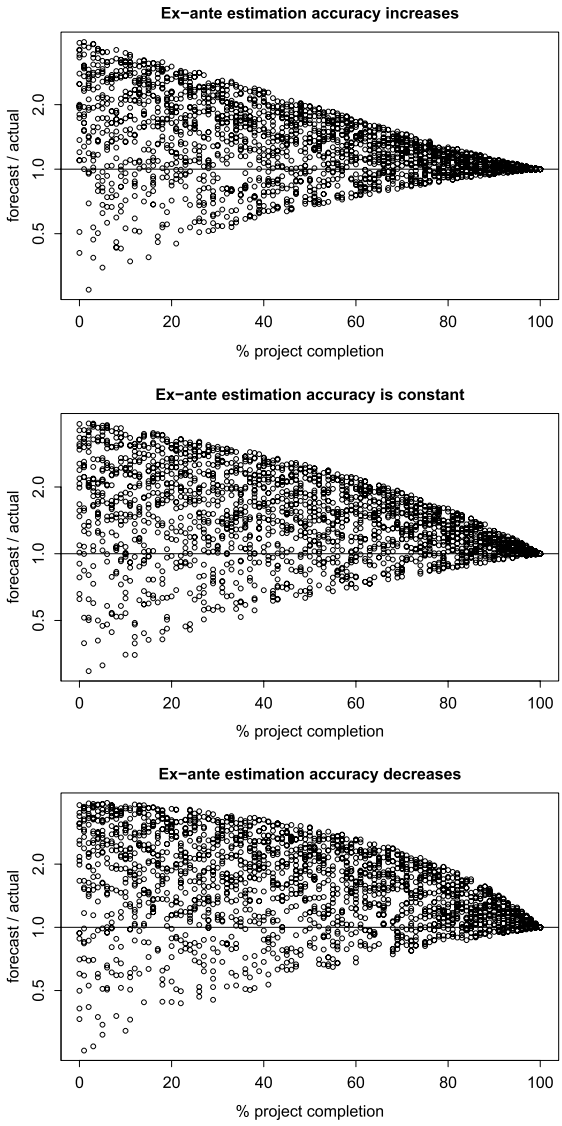

Fig. 3. Simulation of the cone of uncertainty with increasing, constant and decreasing ex-ante estimation accuracy.

The forecast that we make at this time will range from $25 + 50 = 75$ to $100 + 50 = 150$. This means the accuracy of the forecast ranges from $75/100 = 3/4$ to $150/100 = 3/2$ times the actual value. This range however, is not symmetric around the actual value of 100 on both the absolute and logarithmic scale. Thus, by adding the ex-post part to the ex-ante part, we will in general not find symmetry around the actual value.

However, this does not imply that symmetry cannot be present. In two cases, is it possible to find symmetry around the actual value. The first possibility is when the ex-post part is unknown and is estimated with the same accuracy as the ex-ante part. In this case, we can find symmetry, however, this is not in accordance with Boehm's assumptions. Boehm explicitly assumed that there is an increase in knowledge, which implies the ex-post part is known to some extent.

Another theoretical possibility to achieve symmetry of the forecasts around the actual value is when we assume the ex-ante accuracy to be asymmetric and improve in a very specific way, which we will calculate later on. However, that specific way is unrealistic to occur in real-world cases.

It is also possible to find symmetry around other values than the actual value. For example, in the above calculations the values would be symmetric on a logarithmic scale around the value $100 \cdot \sqrt{9/8} \approx 106$. Namely, $150/(100 \cdot \sqrt{9/8}) \approx 1.41$ and $(100 \cdot \sqrt{9/8})/75 \approx 1.41$. In Section 4.2.1, we will explain this in more detail. Other possible symmetry is discussed in Section 6.3, where we will discuss the work done by Little [48]. In that article, it is shown that the ratios of the ex-ante part to the actual remainder of the work in his case study behave in a lognormal way, which is symmetric on a log scale.

At first glance it may appear nonintuitive that there is no symmetry around the actual value. Why is it possible to forecast something five times as high, yet not five times as low? But there is a good explanation. Namely, since we use the ex-post

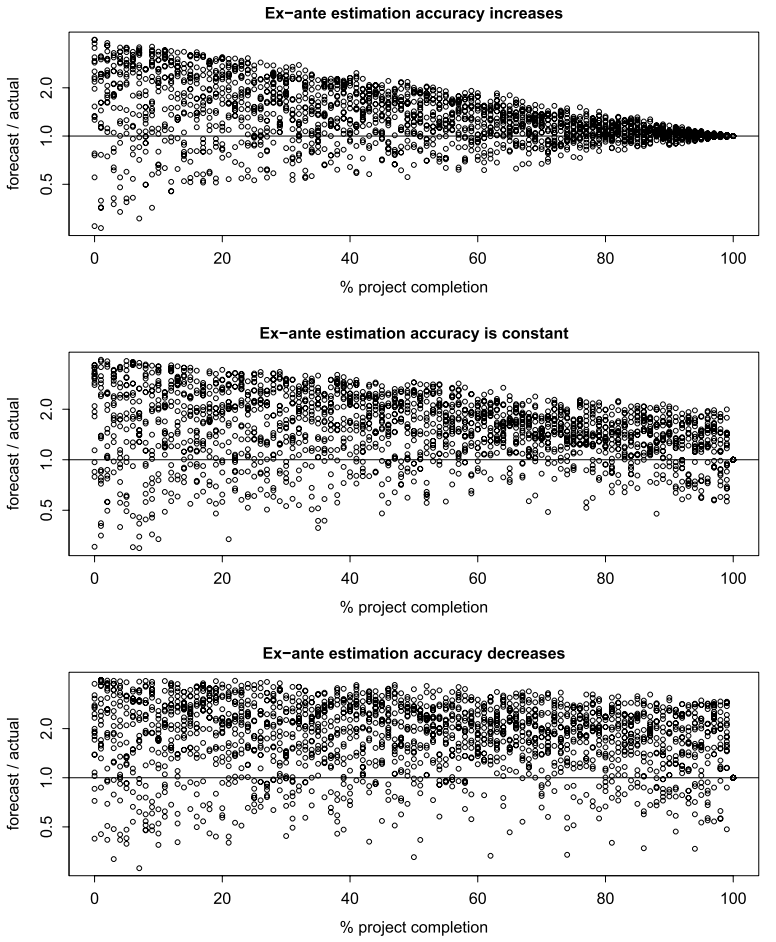

Fig. 4. Simulation of the cones of uncertainty with increasing, constant and decreasing estimation accuracy of the ex-ante part, assuming a uniform function for the ex-post part growth and *not* knowing the ex-post part exactly thus far.

part the lower limit is bounded. If the project is halfway, we are still able to forecast five times as high, but not five times as low. By then, the ex-post part may already be half of the actual and thus our maximum lowest forecast is twice as low. This bound makes it easier to forecast higher than lower. Therefore, the asymmetry of the cone of uncertainty around the actual value is quite reasonable.

Changing the conditions. In the code snippet, we assumed the growth of the ex-post part to be determined by a linear function. We wanted to verify whether using this particular function would be the cause of the conical shape. Another realistic assumption of the effort over time is, for instance, the Rayleigh function as found by Putnam [58,60]. We changed the simulation to use the Rayleigh function with a peak of the effort at 60% of the project's progress [6, p. 93], resembling a more detailed effort–time function. Using this assumption, we found the conical shapes to persist in the three scenarios: increasing, constant, and decreasing estimation accuracy of the ex-ante part. We do not depict the results as they are similar to those in Fig. 3. The figures show that the conical shapes are also found when another function is used instead of the linear function.

So far, we assumed to know the ex-post part precisely. In most organizations, much of this information is administrated well and easily obtainable. However, there are organizations in which this is not the case. Or, even if the information is available, it is not used in making the forecast. In these cases, we need to estimate the ex-post part as well. To investigate how estimating the ex-post part affects the cone of uncertainty, we assume that we are able to estimate the ex-post part twice as accurately as the ex-ante part. Even though the information is not known exactly, the estimator must have an idea of what has been done. Therefore, we assume that this knowledge allows the estimator to better estimate what has been done than what still needs to be done. In Section 5, we will also see an example in which this is not the case. There, the ex-post part is predicted with the same accuracy as the ex-ante part.

Using the assumption where we predict the ex-post part twice as accurate as the ex-ante part, the simulation results are shown in Fig. 4. As the plots show, even in these theoretical scenarios the conical shapes persist, although the deviations to the actual value remain relatively larger during the project since the ex-post part needs to be predicted, as well.

Summary. Under reasonable and modest assumptions, we are able to reproduce the shape found by Boehm, albeit that his symmetric shape around the actual value is neither naturally reproducible nor has a reasonable explanation. By challenging some of the assumptions, we still find this shape. An important finding is that even if we do not improve the accuracy of the ex-ante part, we will still converge to the actual as time passes. The simulations thus support the findings raised by Little in his article [49]. Kruchten was right in his response that the cone is about the forecast and not the accuracy of the ex-ante part. The simulations also illustrated that the improved estimation methods are not the reason for the accuracy of the forecasts to improve, refuting Gryphon's statement. Improved estimation methods, however, will make the cone converge faster.

4. Quality of forecasts

In the introduction, we argued that the quality of forecasting is important. As decisions are supported by forecasts, it is crucial that these predictions are as accurate as possible. Surprisingly, we also noted that often forecasting quality itself is not assessed by organizations. In this section, we want to address how to quantify the quality of IT forecasts. We will discuss tools that together make it possible to determine the quality. These tools are the EQF, introduced by DeMarco and the f/a ratios, which we just investigated, plotted against a reference cone.

First, we discuss DeMarco's EQF [14]. We will show that the EQF is a forecasting quality metric that measures the distance between forecasts and their actual value. This is applicable to each individual forecast, but also to all consecutive forecasts made for a value of interest of a single project. So, it is possible to assess the quality of individual forecasts, but more importantly, the quality of the process of IT forecasting. We will analyze the EQF and its variation with a box plot. We will argue that this is a useful tool for the management to determine forecasting quality. With the quality determined by the EQF, comparisons of the quality of forecasts can, for instance, be made between projects, portfolios or estimation methods.

Second, we discuss the f/a ratios plotted against a reference cone. In the previous section, we found that different conical shapes appear depending on simulation conditions we impose on the forecasts. In this section, we will propose a reference cone. With this reference cone, we are able to compare the forecasts of an organization to the standard of quality that the organization desires. This way, the management is able to assess whether forecasts comply with the conditions, such as the political nature of the forecasts. It allows executives to determine whether the forecasts are what they expect them to be, and take appropriate actions if they are not.

Together, these tools quantify IT forecasting quality and help the management to assess it. The EQF shows the quality of forecasts and allows for comparisons. The f/a plot gives insight in the bias of forecasts and the quality of forecasts. Therefore, the EQF and f/a plot plus our reference cone provide complete insight in the quality of the forecasts.

4.1. Estimating Quality Factor

The EQF was defined by DeMarco in his book, *Controlling software projects* [14]. The book describes in detail various aspects of the IT forecasting process. In this respect, DeMarco gives a definition of a forecasting metric, the EQF, which depicts the quality of forecasts made during a project. He defines the EQF [14, p. 146] by dividing the area under the actual value by the area of the difference between the forecast and the actual. In an article by Verhoef [72], the EQF is defined in terms that are more mathematical. We reiterate and correct the definition given there. Suppose a is the actual value ($a > 0$), t_a the time the actual is known and $e(t)$ the value of the forecast at time t ($0 \leq t \leq t_a$) in the project. Then, the EQF is represented mathematically by:

$$\begin{aligned} \text{EQF} &= \frac{\text{Area under actual value}}{\text{Area between forecast and actual value}} \\ &= \frac{\int_0^{t_a} a \, \mathrm{d}t}{\int_0^{t_a} |a - e(t)| \, \mathrm{d}t} \\ &= \frac{\int_0^{t_a} 1 \, \mathrm{d}t}{\int_0^{t_a} |1 - e(t)/a| \, \mathrm{d}t}. \end{aligned} \tag{1}$$

The assumption is that $e(t)$ is known for the range $[0, t_a]$. That is, we know at all times during the project what the value of the most recent forecast is. In practice, this is not always the case. For instance, often the first forecast of a project is not made right at the start of a project, but made after the start, at time $0 < x < t_a$. In such circumstances, a possible solution is to assume that the first forecast made at time x is actually made at the start of the project. Mathematically, this means we assume that $e(t)$ on range $[0, x)$ equals $e(x)$.

Fig. 5 depicts an example calculation of the EQF for a single project. For this project, in total 4 forecasts were made: at $t = 0$, $t = 0.2$, $t = 0.5$ and $t = 0.65$. For each forecast, we calculate the area between the forecast and the actual value. Of the first, we find the difference to be $1.5 - 1 = 0.5$. The duration of this forecast is $0.2 - 0 = 0.2$. Thus, the area between the first forecast and the actual of this project is $0.5 \cdot 0.2 = 0.1$. We repeat the calculation for the other areas and find the sum of the areas to be $0.1 + 0.06 + 0.045 + 0.07 = 0.275$. As the area under the actual is 1, we find the EQF value $= 1/0.275 = 3.6$.

The figure also contains two lines. In fact, these lines are of a reference cone with a predefined quality expressed in a desired EQF of 3.5. In the next section, we will explain the reference cone in detail. For the moment, it suffices to say that

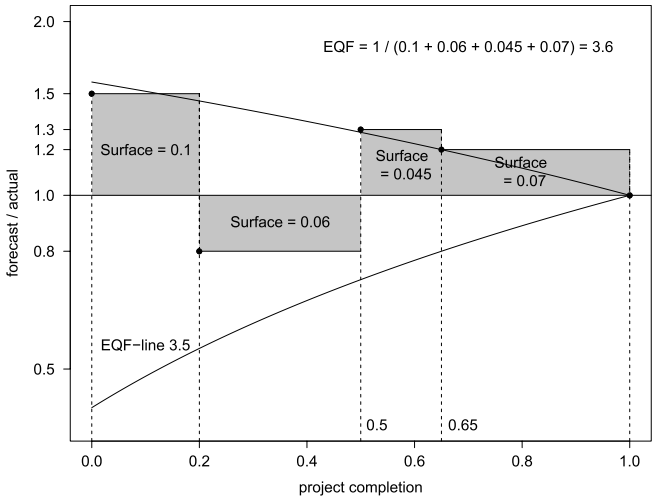

Fig. 5. An example EQF calculation of a single project with lines of a reference cone.

the surface between the upper line and the actual value (horizontal line at 1) computes to an EQF value of 3.5, a little less than our just found 3.6. The same applies to the surface between the lower line and the actual value.

A higher EQF means a better forecast. However, there is a difference between over- and underestimation. In case of at least one overestimate, the range of possible EQF values is zero to infinity. If forecasts are solely underestimations, the range of possible EQF values is between one and infinity. Since we cannot forecast a value smaller than zero, the surface between the underestimated forecasts and their actuals is at worst one, which results in a theoretical EQF value of one. In fact, in case of solely underestimations it is possible to constraint the range of possible EQF values even further given certain assumptions. For instance, if we assume that the ex-post part is known with certainty, determined by a constant growth function and used in the forecasts, the minimum EQF value for underestimations becomes two. Of course, such bounds hinge on the assumptions made, that need not hold in reality. However, the lower bound of one for solely underestimation holds for every situation. These bounds imply that it is in general easier to achieve better EQF values in case of systematic underestimation than in case of overestimation.

The EQF complies to the definitions, we imposed in Section 2. Recall that there we defined a forecast e to be better than forecast f, when

$$\frac{|e-a|}{a} < \frac{|f-b|}{b}$$

with a and b the corresponding actuals. Thus, as the proportional distance between the forecast and its actual becomes smaller, the forecast improves. The EQF complies as it also assigns a higher value in case of smaller deviations between the forecast and its actual. We also assumed some function $G_k = g(e_1, e_2, \ldots, e_n, a)$ that quantifies the quality of several forecasts made for a project k. This function must assign a higher value to a higher quality of forecasts. The EQF complies with these criteria and is thus suitable as function G_k.

MRE. The EQF is closely related to another well-known forecasting metric: the MRE or Magnitude of Relative Error. We found this metric is used more often than the EQF in the literature [13,34,39,40,53,55,56,68]. The MRE was introduced in 1986 in a textbook by Conte et al. [13]. This book defines the MRE as follows: suppose a is the actual value and f the forecast. Then the magnitude of relative error, or MRE, is as follows

$$\text{MRE} = \frac{|1-f/a|}{1}.$$

The EQF and MRE are related to each other. In fact, when only a single forecast is made for a project's actual value of interest, then EQF $= 1/\text{MRE}$ and the average EQF is the mean MRE (MMRE). In contrast with the EQF, the MRE is relatively simple to interpret. An MRE of 0.2 means the forecast deviates 20% from the actual value. The interpretation of the EQF is slightly less obvious. An EQF of 5 means the forecasts made for a single project have an average time-weighted deviation of $1/5 = 20\%$ to the actual value.

The real difference between the two becomes apparent when we take into account multiple forecasts for a single project. In this case, the MRE is not defined. The MRE assumes that only one forecast per project is made. To obtain the forecasting quality of a project with the MRE, it is possible, for instance, to take the average MRE of all forecasts. However, the drawback of taking the average is that the moment the forecast is made, is not taken into account. For instance, assume two forecasts are made for a project, one at the beginning and one at the end. Taking the average means we pretend the forecast made at the end is made during halfway of the project. This allows to boost forecasting quality measured per MRE by making a 'forecast' when the project is almost completed.

The EQF on the other hand does not take the average, but a weighted average of the forecasts made for a single project. The forecasts are weighted with the duration of the forecast. So, we cannot influence the EQF by making a new forecast at the end of the project as the influence of this forecast will be minimal (and it should be minimal). Therefore, the EQF is in general a more reliable and accurate measure for forecasting quality than the MRE.

Note that in the example given above, we discuss adding a forecast at the end of a project to already existing forecasts. With the example, we do not mean making the *initial* forecast nearly at the end of the project. Namely, in this case estimators are able to significantly influence not only the MRE, but also the EQF. In the mathematical definitions of the EQF, we assumed that $e(t)$ on range $[0, x)$ equals $e(x)$. The uncertainty of a forecast diminishes as the project progresses due to an increase in knowledge. Therefore, if estimators postpone making the initial forecast of a project, they acquire additional knowledge that helps making a more accurate initial forecast and improve the EQF value.

This is why we argued in Section 2 that for comparisons between organizations based on time-dependent quality measures, such as the EQF, it is important to use similar definitions of the start and end date of a project. For instance, if in one organization the project starts at the beginning of the requirements phase, and another at the beginning of the design phase, the knowledge available for forecasting is completely different. Such large differences in knowledge can significantly influence the comparison of EQF values. In our case studies in Section 5, we find the definitions used to be relatively similar allowing for fair comparisons.

EQF variants. We showed how to calculate the EQF. In addition, variations of this notion exist. For instance, DeMarco defined another time-weighted EQF [14, p. 147], which takes into account that deviations of the forecast from the actual in the beginning of a project are more important than in the latter part of a project. As we have seen in the previous section, the ex-ante part decreases as the project progresses. This makes accurate predictions of the ex-ante part in the beginning more important than in the end. Therefore, taking into account the time of the forecast makes sense.

DeMarco noted that it does not matter much how one calculate the EQF, as long as they are consistently used. We feel, however, that this alternative time-weighted EQF does have a significant advantage. It will create a larger incentive to focus on making forecasts as quickly and accurately as possible. This in turn will help make better decisions. However, in the remainder of the article, we will use the EQF as defined in Formula (1), since this version is the most well known, used in practice, and reported in publications.

EQF and f/a plot. The EQF is in fact a summary of the information that is captured within an f/a plot. Although information of the quality of the forecasts is encoded in the f/a plot, it is difficult to quantify its quality by looking at the conical shape. In Fig. 5, we depicted an example calculation of an EQF value for a single project. This figure used the same axes as in Section 3 and is an f/a plot, as well. Since the figure contains only a single project, it is possible to calculate the EQF. If we are to add more projects, it becomes hard to distinguish which forecasts belong to which project. This makes calculating the EQF values by looking at the f/a plot cumbersome. Therefore, summarizing the EQFs separately, is a useful addition to the f/a plot.

We note that the EQF is an addition to the f/a plot and is less powerful when used on its own. Since the EQF is a summary of the available information, it does not contain all aspects of the data that are present in the f/a plots. One aspect of the information that the EQF does not show is the potential bias of forecasts. That is, it does not distinguish between forecasts that are systematically lower or higher than the actual value. Thus, an EQF value does not show whether an organization has the tendency to underestimate or overestimate. However, this is valuable information. Namely, the EQF indicates whether the quality of the forecasts is good or bad. Yet, if all forecasts are lower than the actual value, it is possible to improve the quality by removing the bias.

Another aspect the EQF does not show is time. An EQF value does not show if the forecasts made in the beginning of a project are better than those made at the end of a project. For instance, if the forecasting method used in the middle of a project is worse than the one used in the beginning, we are not able to assess this with the EQF. These aspects are present in the f/a ratios plotted against a reference cone.

Comparing EQF values. With the EQF, we can compare projects with each other, but also aggregated forecasts on a portfolio level. Recall that we defined in Section 2, $E = \{G_i : i = 1, \ldots, p\}$ to be a collection of p projects with their forecasting quality quantified by function G_k. If we use the EQF as function G_k, we are able to compare the median value of different collections with each other and make comparisons on a portfolio level based on the EQF values. Below, we give some examples of potential comparisons.

- We can compare the IT forecasting quality of different projects. This allows the management to investigate whether certain projects or types of projects are better forecasted than other projects or types of projects.
- We can make periodic comparisons of the forecasting quality, for instance monthly. It is possible to analyze the quality of the forecasts made for the projects that are completed in a particular month. If we do this each month, we obtain an impression whether the quality of the forecasts improves, is constant, decreases over time, or displays other time-dependent (seasonal) effects.
- We can make a comparison between the forecasting quality of different portfolios. By assessing the collection of EQFs per portfolio, it is possible to decide which one is in overall better at forecasting. This way, the portfolios with the best forecasting quality can be rewarded.

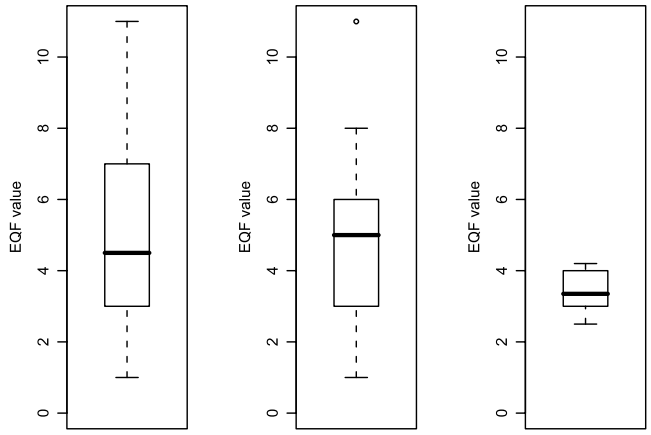

Fig. 6. Three box plots of constructed example portfolios with ten projects each. The median EQF values from left to right are 4.5, 5 and 3.35.

- We can compare new estimation methods with existing estimation methods. This allows the management to assess whether new methods cause the quality of the forecasts to improve. This is, for instance, done using the MRE in a number of articles [13,34,40,55].
- We can compare between organizations. Different organizations can benchmark their own IT forecasting quality with other organizations. This is possible within industries or even with organizations from other industries. However, this does require organizations to compute the EQF exactly the same way. Thus, a description must accompany the EQF values on how the numbers were calculated.

All the above comparisons are possible within an organization, but also between different organizations. Comparisons within a single organization allow for more detailed EQF calculations. An example of such a detailed calculation is to weigh EQF values of different projects by the size of the project. This way, larger projects have more impact on the outcome of the analysis. Such comparisons are more difficult to compute between different organizations as consensus is then needed on weighing the projects. It is, however, in both cases possible to perform the comparisons on a monthly or quarterly basis to check the development of the forecasting quality.

EQF variance. In a number of publications [14,47–49,72], we found that some of the comparisons were done solely based on single values without considering their spread. Articles by Lister [47] and Verhoef [72] state that projects with EQF values in the order of 10 can be considered good. DeMarco [14, p. 157] reports an *average* EQF value for a group of projects. Articles by Little [48,49] use the *median* value of a group of projects to compare the forecasting quality with forecasting quality of other industries. In this article, we advocate using the median value for comparisons and not the mean value. Namely, the distribution of the EQF values does not need to be symmetrical. Since the median value is not significantly influenced by large EQF values, it is, therefore, more robust than the mean value.

Already, the median is more useful than the mean value. However, besides considering the median value, it is most useful to take into account the spread or variation of the quality of the forecasts. If a portfolio has a good median EQF, but also a large variation, it means that forecasts are not consistent. It is easier to make decisions based on consistent forecasts than on highly volatile ones.

One way to consider the variance is to create box plots of the EQFs as shown in Fig. 6. In such plots, the box depicts 50% of the data. The whiskers and the dots represent the other two quarters, with the dots representing potential outliers. The boldfaced bar in the middle of a box plot represents the median value of the data. An article by Kitchenham et al. [40] suggests to use box plots as well, to gain insight in the variation of the f/a ratios itself.

We constructed three example portfolios with ten projects each. The median EQFs of these portfolios are 4.5, 5, and 3.35, respectively. Based on this alone, we can conclude the middle portfolio to make the best forecasts. In Fig. 6, we depict the box plots of the example portfolios with which we have more information. For instance, the leftmost portfolio has a lower median than the middle portfolio, but also a higher chance of EQF values larger than 6. The rightmost portfolio has the lowest median, but also a smaller variance than the other portfolios. In this portfolio, the EQF will be most likely larger than 2.5 where in the other portfolios it can be as low as 1. None of these box plots are good or bad per se; it depends on the goals of the organization, which one is to be preferred. The purpose of this example was to illustrate that using box plots is more insightful than using just a single aggregated EQF.

In summary, by quantifying the forecasts through a box plot of the EQF forecasting metric, management has a tool to assess IT forecasting quality. The tool allows comparisons to be made and gives control to management over the forecasting process.

EQF revisited. Based on the above-discussed merits of the EQF, DeMarco [14] proposed to assess the estimator's performance based on the EQF. He stated that the success of the estimator must be defined as a function of the deviation of the forecast

to the actual, and of nothing else. The estimators will give a forecast that they feel is most accurate at a given time. And, the estimators will re-forecast when they gain high confidence that the forecast is an improvement of the previous one. Since the estimators are solely judged on the accuracy of the forecast, it is no longer in their interest to manipulate the forecast or change the timing of the forecast in order to take the politics involved into consideration.

We agree with DeMarco that the EQF should be used to judge the performance of the estimator. However, the EQF does introduce a side effect. By judging the estimators based on the EQF, their interest is to obtain a high EQF, which does not always align with real performance. The side effect of this is that the estimators benefit from making more forecasts. As soon as a discrepancy between the forecast and actual is found, the forecast will be altered by the estimators to improve convergence. Irrespective of the severity in discrepancy, the estimator benefits from taking it into account thus leading to an increasing number of forecasts per project.

Let us explain by an example how making more forecasts results in an improved EQF value. Suppose the actual, which we want to forecast, is 100 and 20 units of work are done each week. Suppose the initial forecast is 125 and if no new forecasts were to be made, the EQF value of this project would be 4.

During five weeks of the project, discrepancies between the forecast and actual are detected. An estimator can easily adapt the forecast for these differences, for instance, by using the ex-post part each week. This is done in the following manner. After one week, the ex-post part is 20 and the forecast of work done in this week is roughly $1/5 * 125 = 25$. We can 're-forecast' by replacing the ex-ante part of the first week of the initial forecast with the ex-post part. This creates a new forecast $20 + 4/5 * 125 = 120$. Repeating this procedure results in forecasts 115 after the second week, 110 after the third week and 105 after the fourth week. If we evaluate the forecasting accuracy by means of EQF for this project, we find an EQF value of 6.67. We note that our example assumes ideal circumstances. Only if the ex-post growth is adequately predicted, will this method yield increasing EQF values. Otherwise, this method can also cause decreasing EQF values.

This example shows that making more forecasts can result in a higher EQF value. By assessing the estimators solely on the EQF value, they are therefore inclined to make more forecasts. On the one hand, this is a preferable situation. Making additional forecasts in itself demonstrates the use of improved forecasting methods. Just the fact that the forecast is now using additional data from the ex-post portion is an improvement. Also, if the forecast takes into account velocity [12], such as measured in many agile projects, then that again is the use of an improved forecasting method. Moreover, when estimators adapt the forecasts frequently based on the most current information, IT governors are able to steer and monitor projects based on these up-to-date forecasts. This will allow executives to detect unwanted deviations from initial expectations as soon as possible.

On the other hand, making more forecasts can be undesired. Making a forecast or an adjusted forecast consumes time and money. In some cases, an estimator can be interested in adjusting a forecast, even though for the organization the change is insignificant and not worth the effort. Estimators have an incentive to adjust their forecasts regardless as they are judged on the resulting EQF value.

It is possible to take measures to reduce the benefit of making more forecasts. Most effective is to adapt the EQF calculation in such a way that it slightly penalizes for making additional forecasts. Assessing the estimator with such an adapted EQF, causes re-forecasting to only be beneficial when the discrepancy between the most recent forecast and the newly made forecast is significant. For small discrepancies, it will no longer be beneficial to change the forecast as this is penalized by the EQF calculation. Therefore, the incentive for estimators to make more forecasts is reduced.

Note that it is not advisable to restrict the estimators to a maximum amount of forecasts per project. By placing a restriction, the estimators have to consider the timing of re-forecasts made, while this was precisely an argument to advocate using the EQF. Therefore, such a restriction defies the purpose we try to achieve with the EQF.

In conclusion, we advise the organizations that want to introduce the EQF metric, to pay attention to a possible increase in the number of forecasts made. Adjusting forecasts for insignificant changes should be discouraged to prevent waste of effort on unnecessary forecasts.

4.2. f/a plots

In this section, we discuss f/a plots. In Section 3, we created different conical shapes in a plot of the f/a ratios. These shapes are determined by the conditions under which the forecasts are made. Thus, by looking at a plot of the f/a ratios, the shape gives us information about the assumptions under which the forecasts are made. This makes the f/a plot useful for management, as it will allow executives to see if the forecasts are made under the assumptions they expect them to be made.

Below, we elaborate on some of the conditions as described in Section 3.1 that cause different shapes of the f/a ratios and give us insight in how the forecasts were made.

- One of the conditions that influences the shape of an f/a plot is the goal condition. In the case of Boehm's cone of uncertainty, he assumed that the goal of the forecast is to predict as quickly and accurately as possible the actual value of interest of the project without bias. Indeed, the shape was centered around the actual value. However, if the goal of the forecasts is to give an optimistic prediction, we will find a shape that is for the most part below the actual value. Similarly, if the goal in general is to be conservative, we will find a shape that is for the most part above the actual value.

The goal is partly driven by the culture of an organization. If it is difficult to ask for more budget or time, the conservative approach will most likely be taken by the estimators. If it is difficult to get large amounts of budget or time, the progressive approach will be considered. Thus, the shape of the f/a ratios tells us what the goal is of the forecasts made.

- Another condition that impacts the shape of the f/a ratios is the ex-post inclusion condition. In case of Boehm's cone of uncertainty, he assumed that each consecutive forecast incorporates as much information of the ex-post part as possible. In the simulations in Section 3.2, we showed the impact on the shape of the cone if we need to estimate the ex-post part as well. In that case, the width of the conical shape is larger as the project progresses than if we know the ex-post part exactly. If such a shape is found, the management can investigate why this part needs to be estimated. Perhaps, the information is not available or it is not used. In either case, forecasts can be improved as the project progresses just by using the ex-post part.
- Yet another condition is the ex-ante accuracy condition. In Section 3.2, we showed that the conical shape persists irrespective of whether the ex-ante accuracy increases, is constant, or decreases. However, the rate at which the shapes converge to the actual is faster if the ex-ante accuracy increases. Therefore, the shape found in the f/a plot gives us an insight in the ex-ante accuracy of the estimation methods that are used.

Thus, we are able to derive information from the shape found in the f/a plot. This makes it a useful tool for the management to assess the quality of the IT forecasting process. The shape will tell us, among others, what the goal of the estimators is, whether all available information of the ex-post part is used or not and whether the estimation accuracy of the ex-ante part improves or not.

One may wonder why we use the f/a plot as the management tool and not a plot of merely the ex-ante part as advocated by Little [48]. There are two reasons for us to do this. The most important reason is that a plot of the ex-ante part gives less information than the plot of the f/a ratios. An in-depth analysis of the ex-ante part is very insightful, since common re-estimations made are of only this part. However, a plot of the ex-ante part leaves out any information of the ex-post part. Although it makes sense to use the ex-post part in forecasts, our case studies show this is not always done. The f/a plot does take the ex-post part into account and is, therefore, more useful as a management tool. However, note that a plot of the ex-ante part is a useful addition to the f/a plot.

The second reason to use the f/a plot is that a plot of the ex-ante part in general requires more data than the f/a plot. An f/a plot requires five values for each data point: the actual value, the forecasted value, the start date of the project, the end date of the project and the date the forecast is made. However, a plot of the remainder requires another value: the actual value of everything done up to and including the time the forecast is made, the ex-post part. This is needed to calculate the true ex-ante part (actual − ex-post) and the part of the forecast that is an estimation of the ex-ante part (forecast − ex-post). Note that there is an exception when the forecast in question is of duration. In this case, the missing variable is defined by the trivial equation (ex-post duration = date of forecast − start date).

4.2.1. The reference cone

In order to derive the information from the shape found in an f/a plot, we need to compare it with theoretical shapes based on certain assumptions. There are two possible approaches.

One approach is to use simulations to create different shapes by changing conditions using the trial and error method. We are then able to compare the shapes of the f/a ratios of the simulations with the shape of the actual f/a ratios and find those conditions that best resemble the data. However, this can be a difficult and time-consuming task. The number of assumptions we are able to change and the number of shapes we can create in this way are numerous.

Another approach is to restrict the comparisons by first defining a number of desired conditions the forecasts should comply with. Then, we compare the shape of the f/a ratios only with the shape of the simulation caused by the conditions we want. This way, we investigate whether the forecasts are made according to our expectations. If the shapes are dissimilar, we discuss the conditions with the estimators to find out which ones were violated.

In both approaches, we need to compare the shape of the data with a theoretical shape caused by the assumptions we impose. To ease comparison, we draw the theoretical shape together with the f/a ratios, so that we refer the data points to that shape. This reference shape or reference cone immediately allows us to spot deviations of the forecasts from the shape we would like it to have.

Below, we describe how to create such a reference cone. When more involved assumptions are made about the forecasts that we use below, the calculations will change, but the methodology will remain the same. As calculations can become involved, we recommend using computer algebra packages like Maple [51] to compute the results.

In our example case, we want the forecasts to abide by the conditions as formulated in Section 3.1. However, the ex-post growth and the ex-ante accuracy condition need to be further specified. We will assume the following to apply to the reference cone:

- Ex-post growth: The growth of the ex-post part is assumed to be described by a constant function.
- Ex-ante accuracy: The accuracy of the ex-ante part is assumed to remain constant as the project progresses.

With these conditions, we determine the shape of the reference cone. We find the shape by considering how a forecast is made under these conditions. We defined in Section 2 that a forecast consists of two parts: the ex-ante part and the ex-post part.

For the ex-post part, we incorporate as much information as possible and we know this information exactly. Since the ex-post part grows evenly during the project, each time unit x the same amount of work y gets done. The amount of work done at time x is thus described by $g(x) = y$. Since the total amount of work is 1 (100%), we find y by integrating $\int_0^1 g(x)\mathrm{d}x = 1$ which results in $y = 1$. We find the size of the ex-post part by solving the integral $p(t) = \int_0^t g(x)\mathrm{d}x = t$. Thus, at $t = 20\%$ project completion, $p(20\%) = 20\%$ of the work has been done.

The estimation accuracy of the ex-ante part remains the same as the project progresses. The ex-ante part at any time t of the project is of size actual − ex-post, with ex-post being $p(t) = t$. As the final value is always 1 in an f/a plot, we find the size of the ex-ante part to be $1 - p(t)$.

However, we do not know the ex-ante part exactly and have to estimate it. Assume that we are able to predict it with an estimation accuracy of c ($c \geq 1$). That is, the prediction of the ex-ante part lies within $1/c$ and c times the actual value. Thus, the lower bound of the ex-ante part is $(1 - p(t))/c$ and the upper bound is $c(1 - p(t))$. Now, the lower bound l and upper bound u of a forecast made at time t are mathematically defined by:

$$l(t) = p(t) + \frac{1 - p(t)}{c}$$
$$u(t) = p(t) + c \cdot (1 - p(t)).$$

These lines describe the conical shape of our reference cone. For each c we choose, we are able to plot the lines and create a reference cone. With these lines we are able to compare the data points to this reference cone, as we will do in Section 5. The factor c even allows us to obtain an impression of the accuracy of the forecasts. If most data points are within the reference cone, the quality of the forecasts will be near the factor c. If they are for the most part outside the reference cone, the quality of the forecasts is likely to be worse than the factor c.

In the mathematical description of the lines, we incorporated an estimation accuracy factor c. This factor determines the quality of the reference cone. In the previous subsection, we discussed another good candidate for this: the EQF. Therefore, we propose to use an EQF value to determine the quality of the reference cone. In order to do this, we need to calculate the area between the reference lines and the actual value. As we discussed in Section 3, the cone of uncertainty is not symmetric around the actual value. This means, given the same estimation accuracy factor c, the area between the lower reference line and the actual value is not the same as the area between the upper reference line and the actual value. Therefore, we assume that the estimation accuracy factor for the lower and upper bound will in general be different. To be more specific, we assume the estimation accuracy factor of the lower bound to be c_1 and for the upper bound to be c_2 with $c_1 \geq 1$ and $c_2 \geq 1$. This leads us to the following formulas:

$$l(t) = p(t) + \frac{1 - p(t)}{c_1} \tag{2}$$
$$u(t) = p(t) + c_2 \cdot (1 - p(t)). \tag{3}$$

Note that $p(t) = t$ given the assumption that the growth of the ex-post part is described by a constant function.

For these functions, we calculated the area underneath the lines by integrating the function over the duration. We already made such a calculation as visualized in Fig. 5. There we used 4 forecasts and now we use infinitely many in time. The calculations for the upper bound, of which we assume infinitely many forecasts are made that are systematic overestimations, are as follows:

$$\begin{aligned}
\text{upper bound:}\quad \mathrm{EQF}_u &= \frac{\text{surface actual}}{\text{surface upper} - \text{surface actual}} \\
&= \frac{1}{\int_0^1 t + c_2 \cdot (1 - t)\mathrm{d}t - 1} \\
&= \frac{1}{\left[\frac{1}{2}t^2 + c_2 \cdot \left(t - \frac{1}{2}t^2\right)\right]_0^1 - 1} \\
&= \frac{1}{\frac{1}{2} + c_2 \cdot \left(1 - \frac{1}{2}\right) - 1} \\
&= \frac{1}{\frac{1}{2}c_2 - \frac{1}{2}} \\
&= \frac{2}{c_2 - 1}.
\end{aligned}$$

Solving c_2 from this equation yields:

$$c_2 - 1 = \frac{2}{\mathrm{EQF}_u}$$
$$c_2 = 1 + \frac{2}{\mathrm{EQF}_u}.$$

By analogous calculations, we find that $1/c_1 = 1 - 2/\mathrm{EQF}_l$. Thus, we rewrite our previous formulas by replacing c_1 and c_2. This results in:

$$l(t) = t + \left(1 - \frac{2}{\mathrm{EQF}_l}\right) \cdot (1 - t) \tag{4}$$

$$u(t) = t + \left(1 + \frac{2}{\mathrm{EQF}_u}\right) \cdot (1 - t). \tag{5}$$

From $c_1 \geq 1$ follows $1 - 2/\mathrm{EQF}_l \geq 0 \rightarrow \mathrm{EQF}_l \geq 2$. For values $0 < \mathrm{EQF}_l < 2$, Formula (4) will not hold. Recall that in the previous section, we stated it is possible to further constrain the theoretically possible range of EQF values in case of systematic underestimation given certain assumptions. In this case, by assuming the ex-post part is used for creating a forecast and grows constant, the EQF is further bounded to a minimum value of 2 instead of 1. Therefore, in our model it is meaningless to draw lower limit lines with an EQF value between 0 and 2, as the underlying assumptions of the model are clearly not met in such a case. Therefore, in the next section, we will draw the reference line with a minimum EQF value of 2 for such cases. This also implies that in case of systematic underestimation, if EQF values lower than 2 are found, the ex-post part was not taken into account when creating the forecast or was not growing constant.

Since $c_2 \geq 1$, this means $2/\mathrm{EQF}_u + 1 \geq 1 \rightarrow \mathrm{EQF}_u > 0$. This is always the case, therefore Formula (5) holds for each EQF value. Note that Formulas (4) and (5) are not symmetric on a logarithmic scale. However, they are symmetric on an absolute scale.

Thus, we now have curves that describe the reference cone of which we also know the corresponding quality measured by a predefined EQF. This reference cone allows for comparison of the conditions under which the forecasts are made as well as the quality of the forecasts. In the rest of the article, we will use the following notation to indicate the reference cone we use. For instance, if we talk about reference cone(4.5, 8.5), we discuss a reference cone with lower bound formulated using Formula (4) and $\mathrm{EQF}_l = 4.5$ and an upper bound formulated with Formula (5) and $\mathrm{EQF}_u = 8.5$. We use reference cone(4.5), when a reference cone is drawn of which both lower and upper limit use the same EQF value of $\mathrm{EQF}_l = \mathrm{EQF}_u = 4.5$, in this case.

Although we will determine the quality of the reference cone using the EQF throughout this article, it is equal to using Formula (2) and (3) where we used c_1 and c_2. For some, it may be easier to determine the desirable values for c_1 and c_2 than to set a desired quality in terms of an EQF value. Our analyses can be performed with either; one can choose whichever method is most convenient.

Changing assumptions. In the above calculations, we assumed simple but reasonable conditions to apply to the forecasts. Below, we describe some possible extensions if more complex assumptions are made about the forecasts.

- We assumed the growth of the ex-post part to be determined by a constant function, thus leading to $p(t) = t$. In Section 3, we explained that similar conical shapes emerge when a Rayleigh function as the ex-post growth function is taken. If we assume a peak at 60% ($p = 0.6$) as described by Boehm [6, p. 93], the growth function $g(x)$ becomes more complex:

  $$g(x) = a \cdot \frac{x}{p^2} \cdot \mathrm{e}^{-\frac{x^2}{2 \cdot p^2}}$$

  with a being a scaling factor. The function indicates the amount of work that is done at time unit x. Since the total amount of work done must be 1 (or 100%), by solving the equation $\int_0^1 g(x)\mathrm{d}x = 1$ we conclude that

  $$a = \frac{1}{1 - \mathrm{e}^{\frac{-1}{2p^2}}}.$$

  Given $p = 0.6$, we obtain $a \approx 1.33$. With the growth function $g(x)$, we find the size of the ex-post part realized at time t by the integral

  $$p(t) = \int_0^t g(x)\mathrm{d}x = -a\mathrm{e}^{\frac{-t^2}{2 \cdot p^2}} + a.$$

  With this size of the ex-post part, we find the following formulas for the upper and lower bound of the reference cone:

  $$l(t) = 1 + \frac{1}{\mathrm{EQF}} \cdot (0.58 - 2.32\mathrm{e}^{-1.39t^2})$$

$$u(t) = 1 + \frac{1}{\text{EQF}} \cdot (-0.58 + 2.32e^{-1.39t^2}).$$

Note that although the same steps were undertaken as in the example with a constant growth function, the calculations involved with the Rayleigh growth function are more complex. We used the computer algebra environment Maple [51] to solve the computations.

- We assumed to include as much information as possible about the ex-post part and we assumed to know it with certainty. If this is not the case, we need to predict the ex-post part as well. In this case, we are able to predict this within $1/m$ to n times the actual value $p(t)$. That means the ex-post part becomes $p(t)/m$ for the lower bound and $n \cdot p(t)$ for the upper bound.

  However, translating the estimation accuracy parameters (m, n, c_1, c_2) in terms of EQF is not possible anymore without making some extra assumptions. For instance, we need to assume that we are able to predict the ex-post part say twice as accurate as the ex-ante part. That is, we assume $2m = c_1$ for the lower bound and $2n = c_2$ for the upper bound. With such assumptions, we can again translate the lines in terms of an EQF as we have shown above. Of course, one could also forgo translating the lines and simply choose appriopriate values for the estimation accuracy parameters (m, n, c_1, c_2).
- Recall that we stated symmetry around the actual value on a logarithmic scale of Boehm's cone of uncertainty is theoretically possible, if the ex-ante accuracy is asymmetric and improves in a very specific way. To be more precise, we obtain a symmetric reference cone on a logarithmic scale when the following equation holds for the constants c_1 and c_2 in Eqs. (2) and (3). For a given c_2, we need c_1 to be as follows

$$\frac{1}{c_1} = \frac{\frac{1}{p(t)+c_2(1-p(t))} - p(t)}{(1-p(t))}.$$

In fact, this ex-ante accuracy rewrites the formula for the lower bound. Using this ex-ante accuracy, we find the formulas for the upper and lower bound to be:

$$l(t) = \frac{1}{p(t) + c_2 \cdot (1 - p(t))}$$
$$u(t) = p(t) + c_2 \cdot (1 - p(t)).$$

For any given value c_2, we obtain a symmetric reference cone around the actual value on a logarithmic scale. However, in most cases, it does not resemble Boehm's cone as the upper bound is still curving outwards with respect to the f/a ratio of 1. The upper bound of Boehm's cone appears to be a hyperbolic function of the form $1/t$. Therefore, we approximate the shape of Boehm's cone by using for c_2 the following formula:

$$c_2 = \frac{1}{h \cdot t + i} + j.$$

Using nonlinear regression and assuming the phases Boehm uses have the same duration, we find that choosing $h = 2.409$, $i = 0.337$ and $j = 1.044$ results in a good approximation of Boehm's cone of uncertainty.

Although these formulas show it is theoretically possible to achieve symmetry around the actual value, it is unlikely to be found. First, the accuracy with which we predict the ex-ante part must be asymmetric. Second, given the accuracy c_2, c_1 must be exactly as we defined in the formula. It is unlikely there exists some estimation method that mimics the ex-ante part in such a precise manner. Therefore, the symmetry of Boehm's cone around the actual value will in general not be found in real-world cases.

- We also stated in Section 3.2 that symmetry on a lognormal scale is possible around other values than the actual value. We discussed an example of a project with an actual cost of \$100 that halfway spent \$50. Given a symmetric ex-ante accuracy of 2, we showed that a forecast would be between \$75 and \$150. This range is asymmetric on a lognormal scale around the actual value. However, it is symmetric around the value $100 \cdot \sqrt{9/8}$. We explain how to derive this result.

  We are interested in the value y around which the lower and upper bound are symmetric on a logarithmic scale. This means that we wish y divided by the lower bound is equal to the upper bound divided by y. That is, there is an equal factor between the lower bound and y and y and the upper bound. Mathematically, we need to solve the following equation

$$\frac{y}{t + 1/c_1(1-t)} = \frac{t + c_2(1-t)}{y}.$$

This leads to

$$y^2 = \frac{c_1 - c_2 \cdot c_1 - 1 + c_2}{c_1} \cdot t^2 + \frac{c_2 \cdot c_1 - 2c_2 + 1}{c_1} \cdot t + \frac{c_2}{c_1}.$$

Thus, given c_1 and c_2 we find the symmetry line around which the lower and upper bound are symmetric on a logarithmic scale. In our example, we had $c_1 = c_2 = 2$ and $t = 0.5$. Using these values in the formula leads to $y^2 = 9/8 \Rightarrow y = \sqrt{9/8}$. Thus, although it is unlikely the bounds are symmetrical around the actual value it is possible to find symmetry around other values.

4.2.2. Interpretation of the reference cone

We showed how to compute the lines of the reference cone. The reference cone allows for comparison with the pattern of the f/a ratios to see whether the ratios adhere to the conditions that the executives want. This is the main reason to use the reference cone and the reason it gives valuable insight in the bias present in your organization.

Moreover, by describing the reference cone in terms of EQF, we are also able to get an indication of the quantified quality of the f/a ratios. We stress that it is merely an indication of the quantification in EQF values and nothing more. For instance, one may incorrectly think that if all f/a ratios for a single project fall within the reference cone, the quality of the forecasts in terms of EQF must be better than that of the reference cone. However, this is not necessarily the case. Moreover, it is also possible for the projects to have a number of f/a ratios outside the reference cone and still have a better EQF value than the reference cone.

Let us explain this with examples. Recall Fig. 5 in Section 4.1, in which we showed an example of an EQF calculation for a single project with four forecasts. The figure contains a reference cone which is plotted using the assumptions described in this section and Formulas (4) and (5). The EQF value of these lines is taken to be 3.5. The example calculation for this project shows that the EQF value of the forecasts is 3.6, better that the reference cone. Yet, we find that not all the forecasts are contained in the reference cone. This example shows that it is possible to obtain a higher EQF value than the reference cone even if some forecasts are outside it.

For the second example, consider the same figure. Let us assume that we have a project that consists only of the initial forecast made in the figure. That is, the project consists of one forecast with an f/a ratio of 1.5. This means that the surface between the f/a ratio and the actual equates to 0.5 which leads to an EQF value of $1/0.5 = 2$. Thus, even though in this case the f/a ratio is contained in the reference cone, the quality in terms of EQF is worse than that of the reference cone.

Therefore, the reference cone merely gives an indication of the quantified quality of forecasts. In general, the quality of f/a ratios that are inside or close to the reference cone are comparable to the quality of this reference cone. However, it is not guaranteed and must not be viewed as such. This is also a reason to use a box plot of the EQFs, so that it is possible to view the quantified quality of the forecasting separately.

Summary. We showed that the shape of the f/a ratios plotted against a certain reference cone with prescribed EQF, reveals valuable board-level information by depicting and quantifying the quality of forecasts. An f/a plot is, therefore, an indispensable addition for decision makers. Depending on the forecasting conditions, different conical shapes appear. In order to derive the conditions that lead to the shape in an f/a plot, comparisons need to be made with reference cones. We showed how to infer such a reference cone to ease the comparison of various shapes. The reference cone shows how the data should behave for certain quality levels, expressed in an EQF value. This allows for quantitative and objective comparison of the data with the desired quality of forecasting. In the next section, we are going to apply our approach in a real-world context: we will analyze the forecasting quality and their bias of four large organizations.

5. Case studies

In this section, we apply the methods we proposed in this article to four real-world case studies. We use the EQF and the f/a plot with a reference cone, to gain insight in the bias and the quality of the forecasts in the organizations. The results show varying quality and conditions under which the forecasts were made in each of the case studies. In this section, we will compare the quality of the forecasts of each case study to one another. We will discuss benchmarking including the known related benchmarks in the literature, in the next section. First, we will briefly introduce the involved organizations and then we discuss them in detail.

Landmark Graphics Corporation. We obtained data from Todd Little of Landmark Graphics. It is the same data set that he used and reported in IEEE software [48]. Recall that his article initiated an extensive discussion in the same journal. Landmark Graphics is a vendor of commercial software that is used for oil and gas exploration and the production market. We are grateful to Todd Little for providing us with the data that consists of 121 software development projects executed in the period of 1999–2002. In total, Little provided us with 6245 forecasts that predict the duration of these 121 projects. In this section, we extend Little's analyses in the following sense. We show that the goal of this organization is to forecast the minimum value instead of the actual value. This causes most forecasts to be lower than the actual value.

Large multinational company X. The second organization is a large multinational company. Of 867 IT-enabled projects, of which at least 25% of the project costs consists of IT costs, we obtained forecasts and actuals of total project costs. The projects were all started and completed in 2005 or in 2006. In total, 3767 forecasts were made of these projects. In this organization, the forecasts turned out to be generally higher than the actual. Also, the EQF values are very poor when compared with the other case studies. After discussions with the organization, they corroborated to us that the goal of the forecasts was aimed at predicting the maximum value rather than the actual value. The process of budget approval was such that usually less funds are granted than proposed. Therefore, the forecasters tended to overestimate in order to obtain enough funds. Furthermore, this organization steered on Standish project success indicators, which, as we will elaborate on later in this article, induced overestimation as well.

Financial service provider Y. The third case study discusses a large multinational financial service provider. From this organization, we obtained data on 140 software development projects conducted in the period of 2004–2006. In total,

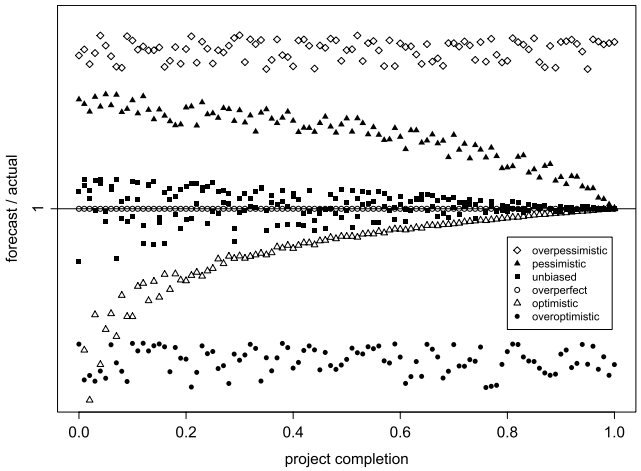

Fig. 7. Typical patterns in an f/a plot.

667 forecasts were made of the total cost of these projects. The quality of the forecasts of this organization in terms of EQF is high. Also, the f/a plot represents a conical shape as Boehm intended with the forecasts centered around the actual value. The plot indicates that the goal of the forecasts is to quickly and accurately predict the actual value.

Besides the cost forecasts, we also obtained 100 forecasts of the functionality of 83 software development projects. These projects were conducted in the period 2003–2005. As with the cost forecasts, the f/a ratios of functionality are similar to the conical shape found by Boehm. Therefore, the goal of the forecasts is to quickly and accurately predict the actual value. In this organization, all forecasts are checked by an independent metrics group (as advised by DeMarco in his book [14]). This aided in creating a corporate culture in which the forecasts only serve the purpose of accurately predicting the actual.

Telecommunications organization Z. The final case study analyzes data from a large organization in the telecommunications industry. We obtained data of 613 projects conducted in the period of 2002–2007. The data contained 1508 forecasts made for the total cost of the projects. The plots indicate that there was no bias in the forecasts. Yet, the quality of the forecasts in terms of EQF values is not high when compared with the other case studies. In this organization, projects were assessed via post calculations on their ability to generate value. Since less focus was put on the forecasts of the cost of the projects, project managers had no incentive to add a positive or negative bias. However, as a result they were not encouraged to improve the quality of forecasting of cost either.

5.1. Typical patterns

Before describing in detail the various case studies, we want to discuss a number of typical patterns in many f/a plots, among others those that we found in our presented case studies. Fig. 7 illustrates these typical patterns. Note that the size of the f/a ratios given in the figure is not relevant. Also, the figure depicts extreme situations, while in practice more variation is to be expected. For instance, an overpessimistic pattern, which in our typical example shows no underestimations, can in practice still contain such underestimations. Therefore, the illustration merely depicts the overall shape and location of the patterns.

We note that in real-world case studies, it is possible to find multiple patterns in one f/a plot. This indicates a heterogeneous set of f/a ratios. Even when a single pattern emerges, the data must be carefully examined for possible heterogeneity. In case of heterogeneity, different political or other biases can be present for the various projects. Finding characteristics that cause differences in the f/a ratios provides useful information for decision making when assessing different types of projects.

In the previous section, we discussed a number of conditions that impact the conical shape of the f/a ratios. In this section, we show the typical patterns that arise when varying the goal condition and ex-post inclusion condition. The ex-ante accuracy condition is not considered here as this does not alter the appearance of the typical patterns significantly, it merely changes the width of the conical shape. First, we will summarize the typical patterns depicted in Fig. 7. Then, we will give examples of situations in which such a pattern can occur. The typical patterns shown in the figure are summarized below.

- overpessimistic ($f/a \gg 1$). The forecasts are much larger than the actuals and no convergence to the actual takes place.
- pessimistic ($f/a \downarrow 1$). The forecasts are above the actual, but they do converge to the actual.
- unbiased ($f/a \rightarrow 1$). The forecasts converge to the actual and are both above and below the actual.
- overperfect ($f/a = 1$). The forecasts are equal or almost equal to the actual.
- optimistic ($f/a \uparrow 1$). The forecasts are below the actual value, but do converge to the actual.
- overoptimistic ($f/a \ll 1$). The forecasts are below the actual value and no convergence to the actual takes place.

overpessimistic. This pattern can be found when we consider an organization in which budget overruns are found to be negative. Suppose that the projects are considered successful when they are within budget and that budget is determined based on the initial forecasts. If the estimators are involved in the project, they need to make sure the initial forecast is high enough so that the actual value in the end is smaller than this prediction. This way, the project is considered a success. Thus, the goal of an estimator is to overstate the expected value of interest to create a safety margin to absorb potential problems in the project.

Also, suppose that the organization makes consecutive forecasts to monitor progression of the project. When these forecasts indicate budget will be left over, the excess amount is immediately reallocated to different projects. When it is very difficult to acquire additional funds for a project, estimators have no incentive making consecutive forecasts more accurate as time progresses as it is hard to regain funds that are immediately transferred to other projects. Therefore, no convergence to the actual takes place.

pessimistic. As the name implies, this pattern resembles a pessimistic pattern. The initial forecasts are set high to create a large safety margin for a project. However, when it is possible to obtain additional funds without much trouble, estimators are more inclined to reduce the safety margin as the project progresses. When the safety margin is still not enough, there is no problem in receiving more budget. Therefore, convergence to the actual will be present.

unbiased. Suppose an organization with an independent metrics group that is merely judged on the accuracy of its forecasts. In this situation, the estimators are interested in predicting the actual value without bias. Also, as they are not involved in any project, it is rather unlikely the ego of estimators can introduce a bias as DeMarco suggested in his book [14]. With an independent metrics group judged on forecasting quality, predictions will be adjusted as soon as discrepancies are found, making the f/a plot converge.

overperfect. Consider an organization that performs many projects based on a fixed price. In this case, the initial forecasts vary slightly from the actual, but after some political debate a forecast is given that is used as fixed price. Externally no uncertainty remains, but internally uncertainty remains on whether it is possible to meet the agreed price. However, often deviations from the actual to the fixed price will internally be minimized and variation is sought in the dimensions time and functionality. Sometimes, the actual may not even be administrated. In such cases, it is best to remove these forecasts from the analysis as these projects do not contain uncertainty on the value of interest. Such projects must be analyzed in other dimensions, where uncertainty does occur. This pattern is also an indication of possible data manipulation.

optimistic. This pattern boils down to what is called a salami tactic. The salami tactic indicates the initial forecasts only account for a small portion instead of immediately considering the complete picture. As the project progresses, the estimator will reveal that there is more to the project than was initially stated.

Another example is an organization that uses forecasts of the minimum value as targets. These forecasts serve as, often unattainable, targets that the project should meet. Once it is clear that the target cannot be met, the forecast is readjusted. This process repeats itself until the project is finished and the last forecast is actually met.

overoptimistic. In this case, the estimators are extremely optimistic. The knowledge that is gained during the project is not used in making improved forecasts. The administration of what has been done is either not correct or not used.

5.2. Landmark Graphics

Landmark Graphics is a vendor of commercial software that is used for oil and gas exploration and the production market. In the period of 1999–2002, 121 software development projects were undertaken for which Todd Little provided us with the data. The data, in total 6245 data points, consist of: the project numbers, the forecasted end dates; the date the forecast is made; the actual start of the project; and the actual end date of the project. These data allow us to calculate at what time during the project a forecast was made and the deviation between the forecast and the actual. We will also derive the EQF values for each project.

Little [49] analyzed nearly the same data of Landmark Graphics. He provided us with a couple of more projects than he analyzed in his article. His analyses provided a number of insights. Little found that the ex-ante accuracy remained constant during the course of the projects. Yet, he still found a conical-shaped figure when plotting the f/a ratios. In Section 3, we illustrated with simulations that this effect is caused by the actual use of the ex-post part in making the forecast.

We want to note that the f/a plot by Little [49] is slightly different from the one we present in this article. At Landmark Graphics, the forecasts of the projects are recorded weekly. Little used each week as a single data point, even if the forecast was not changed from the previous week. However, we chose to only take those forecasts into consideration that were actually changed, which left us with 923 forecasts. We did this for comparison reasons, as we only analyze newly made forecasts of the other case studies as well.

In Section 4.1, we showed that creating more forecasts can result in higher EQF values. As Little's data contain a forecast each week per project, we analyzed the data to check whether the EQF is influenced by the large amount of forecasts. In this case, the EQF was not affected by making a forecast each week, since the forecasts were simply reiterated instead of altered. Therefore, in this case the EQF is exactly the same when only the newly made forecasts are taken into account when compared with all the forecasts given. The benefit of analyzing the forecasts of each week as Little has done, is that the shape of the f/a ratios becomes more apparent.

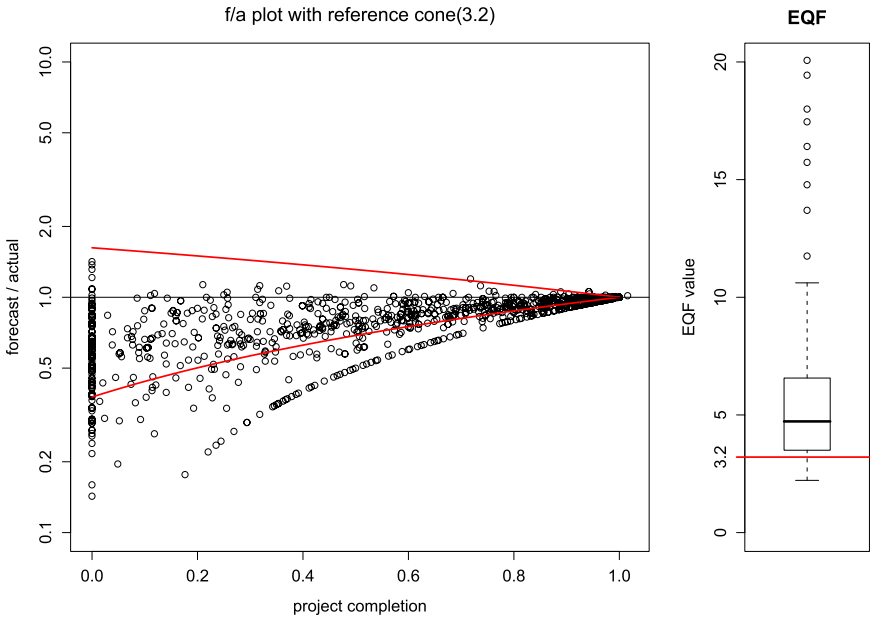

Fig. 8. f/a plot with reference cone(3.2) and EQF box plot of 121 projects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The analysis of Little also includes comparing the results from the analysis with benchmarks found in the literature. He reported that the EQF values and the ex-ante part estimates resemble lognormal distributions. We will discuss this topic in more detail in the next section, where we will show that we were unable to statistically reproduce his results. Little also compares the EQF values of the Landmark Graphics projects with benchmark EQF values reported in a book by DeMarco [14]. In his article, Little compares the median value found for the projects of Landmark Graphics with a reported mean value found by DeMarco. Since the median and mean value are quite dissimilar due to large outliers, this comparison cannot be used to assess the quality of the forecasts at Landmark Graphics. This is only a valid comparison if both are median or mean values. We will make a more fair comparison in Section 7.

We extend Little's analyses by applying the methods described in this article to his data. We depict the result in Fig. 8, which shows a plot of the distinct 923 f/a ratios plus a reference cone, and a box plot of the EQF values. Also, we have no reason to assume that the data set is heterogeneous.

Before we interpret Fig. 8, we note that the reference cone exactly follows the conditions that we used to derive Formulas (4) and (5) of the reference cone in Section 4.2.1. We now only need to choose an EQF value to draw the reference cone. If we choose a too high EQF value, the reference shape may be difficult to distinguish and it would be difficult to compare its shape with that of the f/a ratios. Therefore, we want an EQF value that most projects have attained. Boehm [6] describes that the lines in his cone of uncertainty represent 80% confidence limits. Therefore, we take our reference cone to resemble similar limits. That is, we want the reference cone to use that EQF value for which 80% of all projects have an EQF value higher than that limit. This leads to the 20% quantile of the EQF values. For Landmark Graphics, this resulted in an EQF value of 3.2.

The box plot of the EQFs in Fig. 8 has a median value of 4.7. The solid red line in the box plot corresponds to the 20% quantile with an EQF value of 3.2 that we used to draw the reference cone. The box plot shows that the EQF values are at least 2 and in many cases go up to 10. Also, a number of potential outliers are visible in the plot. In fact, it is quite common to find such high EQF values. If many forecasts are made, the chances are high for one to make at least a number of forecasts that are very accurate. Therefore, these potential outliers need not necessarily be stray values. It is, however, advisable to assess whether these points represent mere luck, accurate forecasting or manipulation of the data. In this case study, we found no reason to exclude them from the analysis. Later on, we will compare the EQF values to those found in the other case studies to assess whether we can consider the values of acceptable quality or not.

Now, we turn to the f/a plot itself. Our reference cone shows immediate room for improvement of the forecasting process at Landmark Graphics by removing the bias. The data indeed display a conical shape much like the shape of the reference cone. However, the data are shifted downward when compared with the reference cone and resembles the optimistic pattern described before. This is supported by a median f/a ratio of 0.85. There are quite a few data points that are lower than the reference cone and almost none that are higher.

This indicates that at least one of the conditions used for the reference cone does not apply to the data of Landmark Graphics. Since the data points are not centered around the actual value depicted by the horizontal line $f/a = 1$, the goal of the forecasts appears to be different from the goal we assumed for the reference cone. As the forecasts are in general lower than the actual value, it seems the forecasters try to predict the minimum value rather than the actual value.

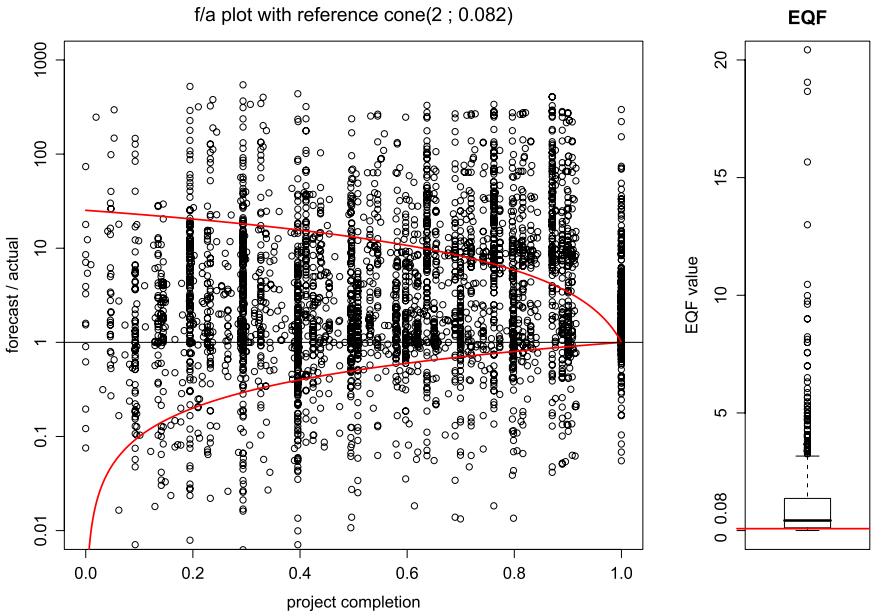

Fig. 9. f/a plot with reference cone(2,0.08) and EQF box plot of 867 projects.

Indeed, Little [49] confirms that the goal of the forecasts is different than the one we defined in the conditions. He describes that the corporate culture is such that the project teams consider the first possible end date of a project as the target. This means the goal is to predict the first possible moment the project can finish, thus a minimum instead of the actual value. This is not uncommon as DeMarco [15] described. He referred to the earliest possible date a project can finish as the nano-percent date. This nano-percent date is often used as target and causes the conical shape of the data to shift downward with respect to the reference cone.

Case summary. This first case study showed that plotting the f/a ratios together with a reference cone based on the 20% quantile of the EQF plus its box plot, enable decision makers to directly assess the quality and bias of forecasting within their organization. The reference cone reveals immediately that the forecasts are biased. The data resembles an optimistic pattern. It is possible to adjust the bias of the forecasts by rewarding estimators to stay within the reference cone. This will change the corporate culture so that the goal of a forecast becomes to quickly and accurately forecast the actual value and thereby improving the quality of the forecasts in this organization. However, since changing the corporate culture is time consuming and expensive, the organization may not gain much value from debiasing. Therefore, executives can decide not to debias the forecasts and to adhere to a different point of reference than the actual value, as we explained in Section 2. By choosing a different reference point, the EQF values will increase without having to change the estimation process. Naturally, executives then also have to account for the bias of the forecasts in their decisions. In Section 6, we show how to do this.

5.3. Large multinational company X

In the second case study, we assessed the quality of the forecasts of a large multinational company. This organization recorded data of all IT-enabled projects. These are projects that consist of at least 25% IT costs. In this case study, we investigated 867 projects that were undertaken in 2005 or 2006 with a maximum duration of one year. The data contain 3767 forecasts made of these projects and consists of: the project names; the actual start date of the project; the actual end date of the project; the date the forecast was made; the forecast of the cost; and the actual cost of the project.

With these data, we made an f/a plot and a reference cone, plus a box plot of the EQFs. The results are shown in Fig. 9. The reference cone plotted in the f/a plot is drawn using the same conditions as the reference cone used in the Landmark Graphics case study. Again, we chose the EQF value of the reference cone to be the 20% quantile, which of this organization is 0.08. However, given our conditions it is not possible in case of systematic underestimations to obtain an EQF value lower than 2. In fact, of underestimations an EQF value of 1 is the theoretical minimum. Therefore, the EQF value of 0.08 indicates that in this case study a lot of forecasts are overestimations. Since the reference cone given our conditions has a minimum bound of 2 for the lower limit, we have drawn the lower limit with this bound instead of the 20% quantile.

The plots in Fig. 9 provide a different picture than the Landmark Graphics data. It is difficult to immediately recognize a typical pattern in the f/a plot. Both the overpessimistic and overoptimistic pattern appear to be present as there is no convergence and there are both large under- and overestimations. Although the data appear to contain multiple patterns, we found no evidence of heterogeneity in the data. None of the project characteristics that were provided, were able to explain the different patterns found.

In the f/a plot, we see that most forecasts are plotted above the horizontal line where $f/a = 1$. This is confirmed by the median of the f/a ratios of 2.25. The EQF box plot corroborates that most forecasts are overestimations based on the low EQF values. The median EQF value of all projects is 0.43. In fact, 65% of the projects have an EQF value that is lower than the theoretical minimum in case of only underestimations of 1. Therefore, we find the pattern to be predominantly the overpessimistic pattern.

The EQF box plot depicts a large number of potential outliers. We find that the distribution of the EQF values has a heavy tail. Due to the large concentration of forecasts with low EQF quality, the EQF box plot shows the tail as outliers. However, similar to the Landmark Graphics case study, we did not find any reason to exclude these points from the data set.

Also, the forecasts do not converge to the actual as time progresses. The f/a plot does not resemble a conical shape at all, but resembles a pipe. In Section 3, we have seen a similar shape in Fig. 4, where we assumed that we were able to predict the ex-post part half as well as the ex-ante part and had deteriorating forecasting accuracy. However, in that case, we still had some convergence, which we do not find in the f/a plot of this case study. This implies that the ex-post part is most likely not used in the forecast.

We consulted with the organization to confirm our inferences from the plots. We did this initially without showing and discussing the outcomes of the analyses to confirm the results. We asked a number of project leaders on how the forecasts were created and asked executives on how the forecasts were used. These discussions confirmed our findings: forecasts were mainly determined by politics, seriously undermining the quality of the forecasts, while executives believed the forecasts to be accurate.

Of a given annual IT budget, not every project proposal obtained all the resources asked for. In order to fund as many projects as possible, projects usually received less funding than requested. The management demanded updated forecasts each month of every project so that, deviations from the forecast could be spotted early. If a project received more budget than needed, the excess funds were used for other projects as soon as possible. However, the corporate culture was such that the budget overruns were seen as negative and requesting a new budget was difficult. Namely, the organization adopted the Standish definition of project success, which means that the project can only be a success when it is within the budget. If a project was successful, it could even result in a bonus for the project members at the end of the year.

Therefore, project managers forecasted the cost higher than they expected the actual value to be. First, they knew they would get less than requested. Second, as reapplying for the budget was difficult, they wanted to be sure they had enough funds for the year even if things went wrong. Third, a project was considered successful when it stayed within the budget. Overstating budgets, thus increased the safety margin of success. Therefore, forecasters in this organization predicted much higher costs instead of the most probable actual cost.

Also, since reapplying for the budget was difficult, project managers refrained from lowering the forecasts using the ex-post part. If they have updated the forecast, it could result in excess funds being transferred to other projects. Project managers only allowed this to happen if they were absolutely certain that they did not need the money. In most cases, they did not significantly lower the forecasts so that there was no reason to reallocate budget from their project.

The reason that it was possible to grossly overestimate in this organization was because the management was unaware of the political bias or the quality of the forecasts. Instead, they beforehand assured us that all the forecasts were truly accurate. The management used the forecasts to monitor and govern the projects, decide upon the annual IT budget, and more. But this forecasting practice induced huge overestimating.

This in turn caused the projects not to get funded initially, as according to the forecasts there was not enough budget. However, many more projects could have been funded from the start of each year, since most projects have requested more than needed. As a result, opportunities were being missed by not allowing the projects to start as soon as possible.

Case summary. This case study emphasizes once more the need to quantitatively assess the bias and quality of the forecasting practice. To be able to make decisions based on forecasts, they must be void of politics and accurate to acceptable levels. Without knowing the forecasting quality, decisions are potentially based on highly inaccurate and politically poised data. Indeed, in this organization it resulted in missed opportunities. Plotting the f/a ratios with the reference cone and providing the EQF values resulted in revealing a biased and low quality forecasting practice.

5.4. Large financial service provider Y

In the third case study, we obtained data from a large multinational financial service provider. The data contained actuals and forecasts of both cost and functionality. We discuss each of them separately and also show the combination of the two.

Cost. We analyzed data from 140 software development projects from organization Y conducted in the period between 2004 and 2006. In total, 667 forecasts were made of the total cost of these projects. The data consist of: the project codes; the actual start date of the project; the actual end date of the project; the date the forecast was made; the forecast of the cost; and the actual cost of the project.

As before, we plot the f/a ratios against a reference cone and we summarize the EQF values in an accompanying box plot as shown in Fig. 10. The reference cone was plotted using the same conditions as in the other case studies. Again, we used the 20% quantile of the EQF values to instantiate the reference cone with; in this organization, this was 3.6. Also, we analyzed the data for possible heterogeneity. We found no reason to assume the data set is heterogeneous.

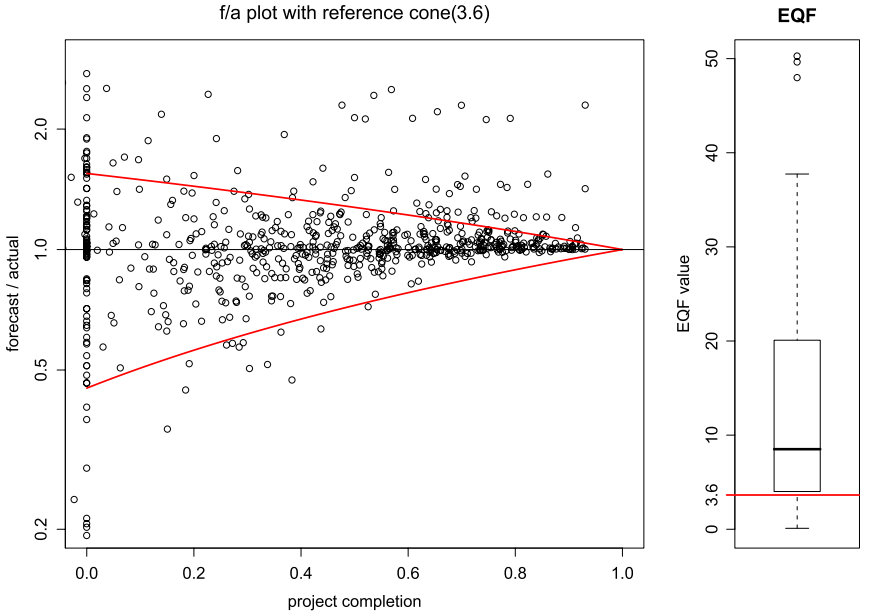

Fig. 10. f/a plot with reference cone(3.6) and EQF box plot of 140 projects.

The f/a ratios do form a conical shape and fall for the large part nicely within our reference cone. The data compare well with the unbiased pattern. This is supported by the median f/a ratio of 1.0. The forecasts converge to the actual value in the asymmetric way, as we would expect. This indicates that the conditions we used for the reference cone apply to the forecasts created by this organization. Moreover, the quality of the forecast is high with a median EQF value of 8.5.

To validate our findings, we discussed the results with the organization. Again, we did this initially without showing or discussing the outcomes of the analyses to confirm the results. We interviewed a number of project leaders on how they created their forecasts. Records were kept on what has been spent so far. Project leaders had full access to these data, wherein they used to regularly update their predictions on the remaining work and total cost. They used the data themselves to monitor the progress of their projects. So, in accordance with our expectations, the ex-post part was used to make forecasts in the organization.

In this organization, the forecasts were also used as budgets. However, all the forecasts were checked by an independent metrics group in the organization. This independent group used methods such as predictions based on the function points countings [16,23], to assess the validity of the forecasts made by the project leaders. If the difference between the forecasts by both parties was too large, budget was not granted until the discrepancies were resolved. Although the project leaders indicated to use a small safety margin in their forecasts, they were unable to increase this margin without proper argumentation. The check by the independent metrics group prevented them from grossly underestimating and overestimating. The goal of the forecasts was simply to predict the actual value as clearly seen from the plots in Fig. 10.

Functionality. The data set for the functionality from the same organization Y involves 83 software development projects in the period 2003–2005. The data contain 100 forecast, which were made for the functionality of the projects measured in function points [16,23]. The same data set is also used in an article by Kulk et al. [41], in which the difference between the forecast and actual was attributed to requirements creep. The article proposes methods to detect projects out of control using early warnings. In this article, we analyze the deviations themselves to quantify the quality of the forecasts made.

In Fig. 11, we plotted the f/a ratios of the function point forecasts with the reference cone using the same conditions as before, and an EQF box plot. The 20% quantile of the EQFs is 2.6. Again, we found no reasons to believe that the data set is heterogeneous.

The figure shows a similar situation as with the cost forecasts for the functionality f/a ratios. The ratios in the figure appear unbiased, which is supported by the median f/a ratio of 1.0. Also, with the exception of a number of outliers, the ratios converge to the actual value. The EQF quality of the projects with a median of 6.4 is relatively high. As with the cost forecasts, the f/a ratios for functionality follow the conditions of the reference cone.

To count the functionality of the projects, multiple experienced function point counters were used. None of them were involved in the execution of the project. Therefore, they had no incentive other than predicting the actual value of the number of function points.

Combined. Of the above projects, in total 55 software development projects contained forecasts and actuals of both cost and functionality. These projects entailed 231 cost forecasts and 69 functionality forecasts. Similar to the previous analyses, these subsets were both unbiased and converging to the actual value. With a median EQF of 9.0 for the cost and 5.0 for the functionality forecasts, the quality remains high.

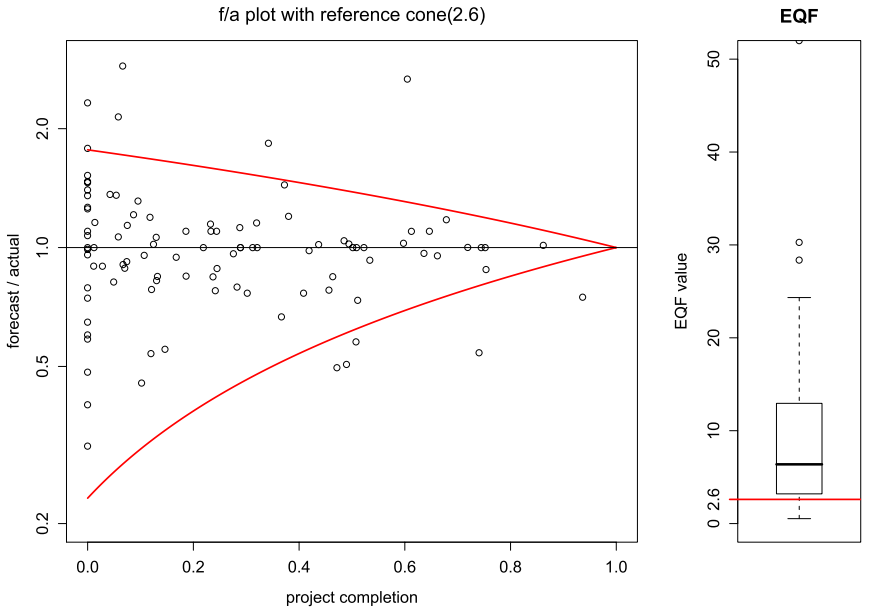

Fig. 11. f/a plot with reference cone(2.6) and EQF box plot of 83 projects.

Case summary. The f/a ratios in this organization compare well with our reference cone, as defined in Section 4.2.1 for both cost and functionality. Combined with relatively high EQF values, this shows the organization is able to make accurate forecasts that are aimed at predicting the actual value without bias. The case study provides evidence that having an independent metrics group is a proper method to obtain a good IT forecasting practice.

5.5. Large telecommunications organization Z

The fourth case study analyzes data from a large international telecommunications organization. This case study consists of 613 projects that entail 1508 forecasts of the cost of the projects. The data consist of: the project codes; the actual start date of the project; the actual end date of the project; the date the forecast was made; the forecast of the cost; and the approved cost of the project.

Note that we only obtained the approved cost and not the actual cost. In this organization, the actual cost of projects were aggregated to higher levels without storing information on the actual cost per project. Therefore, it was not possible to obtain the actuals of each project as they were no longer traceable in the aggregated numbers. In the analysis, we used the approved cost as though they were actuals to plot the figures and compute the EQF values. Although the approvals are approximations of the actual, we feel that the discrepancies do not influence the overall conclusion of this analysis. In this organization, the actuals were always lower than the approved budget. Of most projects, the budget was approved in several phases. New budget was only requested when the previously approved budget was spent. Therefore, deviations between the actual and the approved budget are maximally as large as the difference between total approved budget and the total approved budget minus the last approved additional budget. However, the discrepancies do affect the comparisons to public benchmarks that we will conduct in the subsequent sections. Therefore, we will not use this case study for that purpose.

In Fig. 12, we made a plot of the f/a ratios, the reference cone and the EQF values of the projects in the same manner as before. The 20% quantile of the EQF values used for the reference cone in this organization was 2.1. The reference cone was drawn under the same conditions as before.

The figure immediately shows something strange with the EQF box plot. Instead of the box, we only notice the median value 13.0 (the bold bar), the 25% quantile line and the 20% solid quantile line. The box does not appear in this box plot as the 75% quantile turns out to be infinite. This means that more than 25% of all the projects are able to perfectly predict the approved cost of the project. This resembles the overperfect pattern, which clearly shows in the f/a plot as well.

However, considering that we used the approved budgets instead of the actuals, this is not surprising. In this organization, the forecasts made were used to determine the budget. Therefore, many approvals were equal to the forecasts made. This is also the cause for many projects to have an infinite EQF value. There are a number of possible reasons why a project can have exactly the same approval as forecast:

- Projects could be fixed budget projects. That is, the maximally allowed cost of the project is agreed upon by the parties involved. In many such cases, the project will spend all the budget as focus is placed on the dimensions time and functionality.
- Some projects could have finished with money to spare, but this was not registered in the data set.

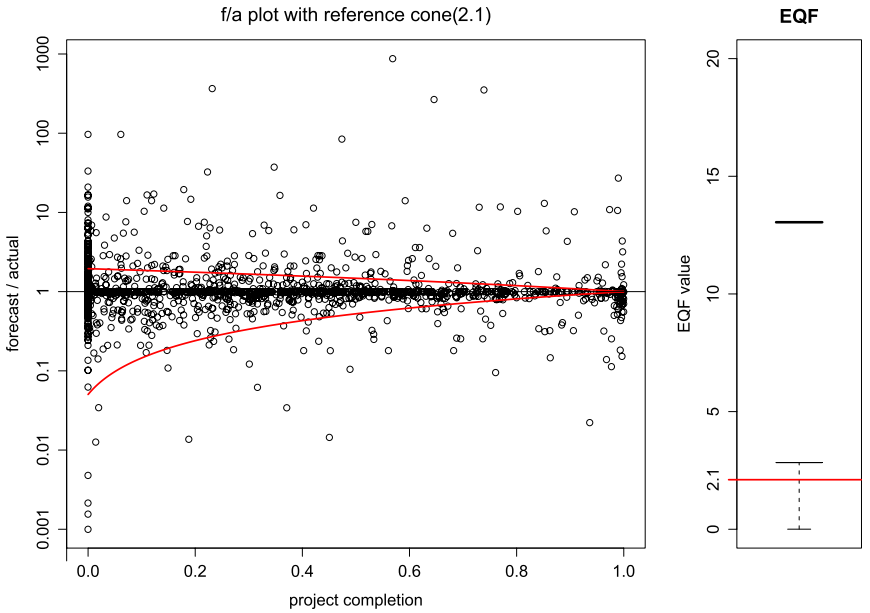

Fig. 12. f/a plot with reference cone(2.1) and EQF box plot of 613 projects.

- Some of the project data may have been manipulated. For instance, the budget leftovers of a certain project were used to account for hours of another project.

In case if the projects are fixed budget projects, it is best to eliminate them from the analysis. In such cases, the projects will minimize the deviation of the actual to the fixed price; most of the money given will be spent. Therefore, no information is gained on the quality of forecasting as there is little uncertainty remaining in a fixed budget agreement. Of such projects, we must analyze the forecasts of time and functionality to see how accurately the projects are predicted. As we were unable to distinguish which of these projects were fixed budget or not, we decided to remove all the projects with an infinite EQF value.

Recall that in the other case studies we also found potential outliers in the EQF box plot. In these cases, we did not exclude these values since they could have been achieved by accurate forecasting or by mere luck. In this case, the large amount of projects with infinite EQFs with respect to the entire data set make these reasons for high values unlikely in this situation. Therefore, in contrast to the other case studies, in this case we exclude some of the outliers, namely those with infinite EQF.

After discussing the initial results with the organization, we were asked to analyze the data for possible trends. As we were given 6 years of information, the conjecture was the data contained a trend. Therefore, we grouped the f/a ratios based on the year a project received its first approved budget. We analyzed them by plotting the f/a ratios and reference cones of each year. The analysis revealed different patterns for the years 2002–2004 and 2005–2007. In the years 2002–2004, we hardly found any convergence, whereas in the later years convergence was present. After consulting with the organization, we found that the forecasts were recorded differently starting from 2005. This indicated the initial data set contained heterogeneity. Therefore, we decided to remove the data from the projects before 2005 from the analysis as well.

The remaining data consist of the projects started in 2005 or later and with a finite EQF value. Of these data, we have no reason to assume that it contains remaining heterogeneity. The data consist of 307 projects and 971 forecasts. Again, we made a plot of the f/a ratios, a reference cone and a box plot of the EQF values in Fig. 13. The 20% quantile of the EQF values is 1.5. Since this value is lower than the minimum required of the lower bound of the reference cone, we plot this lower bound with the value 2 as we did before.

After removing the projects with infinite forecast quality, we obtained an EQF box plot as we have seen in the other case studies. The median EQF value of the remaining projects is lowered to the value 4.3. Also, the 20% quantile has reduced to 1.5. Although many overperfect forecasts are removed, still many f/a ratios in the figure are located on the horizontal axis. The f/a plot still shows the overperfect pattern. Again, this is not surprising considering the fact that we compared the forecasts to the approved budgets.

The 2005–2007 data resemble the unbiased pattern, as we have identified in organization Y as well. The overall pattern in the figure shows the f/a ratios converge to the approved values. Also, no clear bias is distinguishable from the plot. This is supported by the median of the f/a ratios of 1.0.

We corroborated the results with the organization. We found that in this organization project development followed a number of phases. In every phase, an approval was needed to proceed. For some phases, a forecast of the expected cost of that phase was required. Therefore, the forecasts made were partitioned in what had been done and what remained. So, it is likely that our assumptions of the reference cone concerning the ex-post part are satisfied.

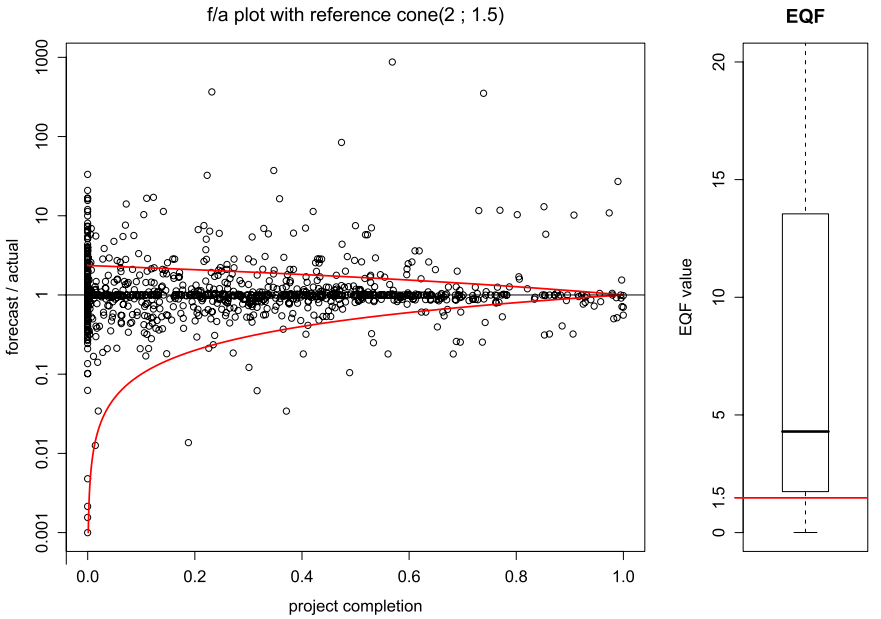

Fig. 13. f/a plot with reference cone$(2, 1.5)$ and EQF box plot of projects started after 2004 and without projects with an infinite EQF value.

Table 2
Case comparison based on EQF values and f/a patterns.

Organization	20% quantile	Median EQF	Average EQF	Patterns	Number of projects
X	0.08	0.43	1.6	overpessimistic	867
Z	1.5	4.3	85.9	unbiased	307
LGC	3.2	4.7	6.3	optimistic	121
Y functionality	2.6	6.4	9.9	unbiased	83
Y cost	3.6	8.5	36.9	unbiased	140

A review board assessed the forecasts and provided budget accordingly. However, the projects were not judged on the quality of the forecasts of the costs. A project was judged by its ability to generate business value and its time to market, a good idea in our viewpoint. This was achieved through post calculations to assess whether the targets set in the business case were met. So, primary focus was put on the added value for the business, then the cost per function point and finally on the plan accuracy. This enabled the forecasters to predict the actual cost of a project without the need to, subconsciously, introduce a bias. However, as the forecasts of economic key performance indicators like Net Present Value and cost per function point were considered more important than the plan accuracy, the quality of the latter was not the main focus.

Although the organization had no focus on the quality of the forecasts of cost, we note that it is very relevant. The forecasts of cost are needed to compute the Net Present Value. In order to obtain reliable Net Present Value data, the quality of cost forecasts is crucial and must not be ignored.

Case summary. In this case study, we analyzed several years of forecasting data. We were able to use our approach to identify the differences in the quality of forecasting over the years. We addressed this heterogeneity in the data set, by removing f/a ratios from earlier years and removing projects with infinite EQF values. We found no bias in the f/a ratios, yet it is possible to improve the quality. Since the organization put more emphasis on the value a project generates than the cost, less attention was given to the quality of the forecasts of the cost. We agree that the value of a project is important, but do note that the value is influenced by the cost. Therefore, it is equally important to check and improve the quality of the forecasts of cost as well.

5.6. Case comparisons

After having analyzed the four case studies separately, we are able to compare the organizations with one another. First, we will summarize the main characteristics of the case studies in terms of EQF values and patterns found in the f/a plot for each organization in Table 2. Note that the number of projects used for this table is only 1518 instead of the mentioned total of 1824 in the introduction. This is because, we left out a number of projects in the analyses of organization Z. Recall that we verified all our analyses, both before and after showing the results, through discussions with the organizations.

To properly compare the organizations, we need a criteria to determine what is good and what is not. In Section 2, we defined a collection E of p projects to be better forecasted than a collection F when the median of E is larger than the median

of F. E is defined by $E = \{G_i : i = 1, \ldots, p\}$ and F is similar. In Section 4.1, we showed that the EQF is suitable as function G_i. Thus, the forecasts of one organization are better than those of others, if in case the median EQF value of all projects is higher. The table is sorted based on this criterion.

Concluding from the analyses, we find organization X to produce the worst forecasts out of the case studies. In this organization, the forecasts hardly have any relation to the actual values of the projects. Highly influenced by politics, the quality of the forecasts is very poor. The f/a ratios resemble an overpessimistic pattern, indicating that the estimators in general overestimate the value of interest.

Although organization Z does not have a particular bias, the quality of the forecasts is not as high as for Landmark Graphics and organization Y. Since the quality of the forecasts is not checked and it is not given particular attention, there is no incentive to improve the forecasting process.

Landmark Graphics has a reasonable quality when compared with the other case studies. For instance, the forecasts are considerably closer to the actual than in the case of organization X. However, the quality could still be improved when compared with the quality of the forecasts in organization Y. The main difference is the bias of the forecasts that is not present in organization Y.

Out of all the case studies, organization Y has the best quality of forecasts with a median EQF value of 6.4 for functionality and 8.5 for cost forecasts. The forecasts made in this organization are more accurate than those made in the other organizations. With the goal to forecast the actual value as quickly and accurately as possible, the forecasts provide the organization with usable predictions for their IT projects.

In Section 7, we will compare the case studies with known benchmarks from the literature. In those comparisons, we will not consider organization Z. The benchmarks from the literature have been calculated using actuals. In organization Z, we used approvals instead of actuals, since only aggregated actuals on more than one project were present. Therefore, we have refrained from using that data in further comparisons.

In our case studies, we found that, although estimators and IT governors assumed political influences to be present and most knew how forecasts were made, none were aware of the impact on the quality of IT forecast quality. In organization X, executives ensured us that the forecasts were accurate, yet the tools lead us to a different conclusion. Therefore, the tools give valuable insight in the quality of your IT forecasts by quantifying the quality and making biases transparent.

6. Enhancing forecast information

In the previous section, we showed that an f/a plot, our reference cone, and a box plot of the EQF values enable quantifying and assessing the quality of IT forecasts. The analyses provide IT governors with useful information about the conditions under which the forecasts are made. It allows executives to detect biases, assess improvements made in the forecasting process and make comparisons between, for instance, portfolios. Moreover, it enables using the quantified quality to enhance the available forecast information for decision making.

In this section, we will discuss three methods that provide such enhanced forecast information. The first method applies the information gained in the analyses of the previous section to newly made forecasts. We show that the quantified quality in terms of EQF and the bias of the organization enables making basic calculations to assess the uncertainty of new forecasts. For instance, it is possible to make statements such as: given the EQF and bias, it is likely that the project will cost roughly between 1 and 1.7 times the forecasted value. This type of information is easily adopted even if hardly any data are available.

The second method enhances the information further and is known as a confidence interval. With this interval, it is possible to make statements such as: the actual will be between 0.9 and 1.4 times the forecasted value with 80% certainty. McConnell [52] gives values to create intervals around the forecast so that in 80% of the cases the actual value will be contained in the interval. In this section, we show that the values given by McConnell are not applicable in general. In most cases, the results from applying these figures do not lead to an accurate description of the uncertainty surrounding a forecast. However, the method itself is very useful when organizations derive the ranges of the confidence intervals based on their own data. We will provide these ranges for our case studies.

The third method we discuss, is a generalization of the confidence interval. When enough f/a ratios are available, it is possible to determine the distribution of groups of f/a ratios. When such a distribution is known, it allows making statements such as: with 90% probability the project will cost at least 1 million Euro and with 15% it will cost 1.8 million Euro or more. We will discuss the theoretical distributions suggested in the literature. We will argue that in the literature there is hardly any evidence that the f/a ratios in general belong to a theoretical distribution. Therefore, it is better to use the historical data available to derive the empirical distribution.

All of these methods use historical data to make statements about future forecast uncertainty. The underlying assumption of the three methods is that the historic data provide reasonable projections for the future. It is assumed that no trend breaks take place and the past and present situation of the organization are similar. If this assumption does not hold, the information gained with the methods must be regarded with extra care.

6.1. Basic calculations

The first method to enhance forecast information is based on the quantified quality of the forecast, in case if only limited information is available. The method consists of basic calculations, which we illustrate using an example.

Consider an IT governor who needs to decide on a project proposal of an organization. The project is forecasted to cost 1 million Euro. Suppose that the quality of the forecasts based on historical projects is quantified using our proposed methods and found to have a median EQF value of 5. Moreover, the forecasts are biased and follow an optimistic pattern. That is, the forecasts are in general smaller than the actual value.

Assuming that the past and present situation in the organization are the same, this information enables the IT governor to assess the forecast of our example project proposal in the following manner. We wish to assess the deviation of the initial forecast, yet we only know the quantified quality in terms of EQF. However, recall that in Section 4.2.1 we determined the relationship, given certain assumptions, between the f/a ratios and the EQF with the reference cone. This relationship was defined using the following formulas.

$$l(t) = t + \left(1 - \frac{2}{\mathrm{EQF}_l}\right) \cdot (1 - t)$$

$$u(t) = t + \left(1 + \frac{2}{\mathrm{EQF}_u}\right) \cdot (1 - t).$$

In this case, we are interested in the deviation at $t = 0$. Since the forecasts in our example are in general smaller than the actual value, we are able to approximate the optimistic bias by only considering $l(0)$. Note that one of the assumptions of the formulas was that the forecasts are unbiased. Although this assumption does not hold in our example, considering $l(0)$ is the best approximation that we are able to make with the limited information available. In Section 7.2 we show how to correct for biases if more information is available.

Given our EQF value of 5, we find that $l(0) = 0 + \left(1 - \frac{2}{5}\right) \cdot (1 - 0) = 0.6$. Thus, half of the projects will have an f/a ratio between 0.6 and 1. Therefore, the project has roughly a 50% chance to cost between 1 and $1/0.6 \approx 1.7$ million Euro.

To give another example, suppose that the forecasts in the organization were unbiased. In this case, the actual can turn out to be higher or lower than the forecast. Then we also need to assess $u(0)$. We find that $u(0) = 0 + \left(1 + \frac{2}{5}\right) \cdot (1 - 0) = 1.4$. Therefore, the project would have roughly a 50% chance to cost between $1/1.4 \approx 0.7$ and $1/0.6 \approx 1.7$ million Euro.

Note that this method provides for very rough calculations. For instance, in this example, we apply the EQF values calculated for *all* the forecasts to a single *initial* forecast. We have seen that the initial forecast in general has a larger uncertainty than forecasts made later in the project. Although we make a correction for this using the reference cone, the 50% deviations applied here are only a rough prediction. Much better would be to apply the analyses of the previous section only to the initial forecasts. Using that information, these basic calculations become more accurate.

But more importantly, we apply the EQF that does not distinguish between an under- or overestimation. In the optimistic example, we used the bias to assume that all the forecasts are underestimations, while in reality some overestimations may also occur. Such overestimations would make the interval of [1, 1.7] too narrow and the chance of 50% of being in that interval too high. On the other hand, in the unbiased example, not knowing the direction potentially makes the interval of [0.7, 1.7], a too wide a range of values.

Still, this type of information will enhance the quality of the forecast information provided to the IT governor. It enables statements to be made regarding the uncertainty surrounding the forecast and gives a more realistic assessment of the true cost of the project.

Next, we will describe two additional methods that further enhance the forecast information. Note that these methods are advanced and require more data than merely the median EQF value. We advise an organization that has not performed any of our previous analyses described in Section 4, to be cautious in using these methods. If one has not performed these analyses or do not have the data, it is better to first make an adequate assumption of one's own bias and quality in terms of EQF. Then, it is possible to enrich decision making in the way we have described above. For more information to make an adequate assumption of your quality in terms of EQF, consider the values we found in our case studies in Section 5 and other EQF benchmark values, which we will describe in the next section.

6.2. The confidence interval

The second method we discuss to enhance the forecast information is well known in statistics and is known as a confidence interval. Boehm [6] argues that each forecast should include an indication of its degree of uncertainty. McConnell [52] proposes a method that gives a prediction of this uncertainty by creating an interval around a forecast using the cone of uncertainty. Tockey [69] explored this method further, while others [11] also discuss and advocate the use of such intervals.

First, we explain informally what a confidence interval is and how it can be used to enhanced forecast information for decision making. Note that both Kitchenham et al. [40] and Tockey [69, p.351–355] describe and informally explain how to create the intervals. However, these articles do not contain a formal description as we will give. We explain how to compute the confidence interval for any given data set. We illustrate this by calculating the intervals for our case studies from Section 5. Finally, we assess the values McConnell advocates, which he took from Boehm. To assess the applicability of these benchmark values, we applied them to our case studies. The results of this analysis show that the values given by McConnell are a poor indication of the uncertainty surrounding the forecasts in the case studies. It is better to derive organization-specific values of the confidence interval, in order to get a valuable prediction of the uncertainty.

Informal description. Recall forecast e in Fig. 2, in Section 3. During the time this forecast is made, the actual value is unknown. However, at that time it is valuable to obtain an impression of the uncertainty of the forecast and to obtain an idea on the likely range of values that contains the actual value. For this purpose, McConnell suggests to use the boundaries of Boehm's cone of uncertainty. From Boehm [6, p. 310], we know that these intend to represent 80% confidence limits. This means that 80% of the forecasts should be within these boundaries. By using them to create an interval around each forecast, it is assumed that 80% of those intervals created will contain the actual value. Later on in this section, we will explain in detail how to compute and apply the boundaries.

In Fig. 2, we applied the boundaries of the cone of uncertainty to forecast e, which resulted in the vertical solid line around the forecast. This interval is a confidence interval and in this case it actually does contain the actual value.

At the time a forecast is made and the interval of that forecast is constructed, we do not know whether the actual value is actually contained in the interval. However, the confidence interval provides useful information. When a decision needs to be made based on a forecast, the interval gives insight in the accuracy of the forecast. For instance, consider an example where a decision needs to be made for a project that is forecasted to cost 1 million. Suppose that the interval we created ranges from 0.7 to 2. Now, we are able to enrich our forecast information by adding that the interval has an 80% chance of containing the actual value. Thus, most likely the project will cost between 0.7 and 2 million. This tells us that we must not be surprised when the project will be twice as expensive as initially forecasted. Also, it allows for the best and worst case analyses that take into account the uncertainty of the forecast.

Formal description. In an article by Smithson [66], a confidence interval is defined as follows. Denote the actual value, of which the true value is unknown at the time, by θ. Assume that a confidence level of $100(1-\alpha)\%$ is given, where α lies between 0 and 1. Let N be the sample size of the data set. A two-sided confidence interval consists of an upper confidence limit U and a lower confidence limit L, such that under repeated random samples of size N, the interval between U and L contains θ's true value $100(1-\alpha)\%$ of the time. It is common to choose the upper and lower confidence limit in such a way that $100(1-\alpha/2)\%$ of the data is lower than the upper limit and $100(1-\alpha/2)\%$ of the data is higher than the lower limit. The underlying assumption of a confidence interval is that the samples are created in the same way. That is, the samples are assumed to be homogeneous data.

Since the forecasts are the samples, the assumption is that they are made with the same estimation method. If different estimation methods are used to create the forecasts, the confidence interval, therefore, does not apply. In that case, we need to split the forecasts into groups that are made with the same estimation method and apply the method to each group individually.

We stated that the interval depends on the confidence level $100(1-\alpha)\%$. But what choice of α makes a good confidence level? We provide two extreme situations to give an idea. Suppose we create for our example project a confidence interval with a confidence level of 1% and find it will cost between 0.99 and 1.01 million. Although the interval is quite narrow, it does not help much in making a decision. The only thing the interval shows is that the actual cost of the project most likely will not be within this interval.

On the other extreme, suppose we create a confidence interval with a confidence level of 99% that finds the cost of the project is between 0 and 5 million. Although we are quite confident that the actual cost of this project is within this range, this does not help us much either in decision making as the interval is quite wide. The range of possible values is simply too broad to adequately assess whether the project is worth undertaking.

So, to answer our question, ideally we want a narrow confidence interval with a high confidence level. An indication of the width of the interval is obtained by looking at the ratio of the upper confidence limit and the lower confidence limit: U/L. In case a confidence level of 80% is taken, this ratio is known as the $p90/p10$ ratio also described by Little [48]. If two confidence intervals have the same confidence level, we prefer the interval with the lowest U/L ratio.

6.2.1. Organization-specific intervals

With this formal definition of the confidence interval, we are able to compute organization-specific intervals based on the historical data and explain as follows. First, we determine the division of the forecasts that we are interested in. That is, we group the forecasts by determining the phases or the range of percentage of completion for which we want to know the confidence limits. For each group, we denote f_i to be the f/a ratio of forecast i contained within that group and $f = \{f_i\}$. Let x be the desired confidence level for the confidence interval with $0\% \leq x \leq 100\%$. Let r be the $(100\% - x)/2$ quantile of f and s the $100\% - (100\% - x)/2$ quantile of f. For instance, if we take a confidence interval with confidence level of $x = 80\%$, r is the 10% quantile and s the 90% quantile. We calculate the confidence limits by $U = 1/r$ and $L = 1/s$. Note that the upper confidence limit is determined by the lower quantile and the lower confidence limit by the upper quantile.

In accordance with the above explanation, we derived the confidence limits that correspond to 80% confidence intervals for the case studies based on their historical data. We computed the reciprocals of the 10% and 90% quantiles of the f/a ratios. These values form the lower and upper confidence limits and are summarized in Table 3. We chose to group the forecasts based on the percentage of completion. These particular groups have been made to ease the comparison with the values suggested by McConnell, which we will discuss below and explain in detail how we derived this specific grouping.

The values given in Table 3 provide us with useful information. Consider our example project proposal with an initial forecasted cost of 1 million Euro. And, let us assume that the initial forecast is made in the range of $0\% - 14.4\%$. The intervals are created in such a way that they contain the actual value in 80% of the cases. Thus, if the project is performed by Landmark

Table 3
Confidence intervals with confidence level 80% derived from historical data for three organizations. Organization Z was excluded as in that case study we analyzed approvals instead of actuals. These figures are not generally applicable to different data sets.

	LGC			X		
% of completion	L	U	U/L	U	L	U/L
0%–14.4%	1.11	3.30	3.0	0.06	7.82	130
14.4%–21.8%	1.06	2.21	2.1	0.05	7.77	155
21.8%–30.8%	1.11	3.21	2.9	0.04	2.85	71
30.8%–40.7%	1.04	2.74	2.6	0.07	4.55	65
40.7%–100%	1.00	1.35	1.4	0.03	1.77	59
	Y cost			Y functionality		
% of completion	L	U	U/L	U	L	U/L
0%–14.4%	0.59	1.89	3.0	0.67	1.76	2.6
14.4%–21.8%	0.73	1.58	2.2	0.95	1.56	1.6
21.8%–30.8%	0.76	1.49	2.0	0.89	1.28	1.4
30.8%–40.7%	0.83	1.31	1.6	0.63	1.15	1.8
40.7%–100%	0.80	1.07	1.3	0.91	1.75	1.9

Table 4
Confidence intervals with confidence level 80% derived from Boehm as used by McConnell [52, p. 169]. These figures are not generally applicable to different data sets.

Phase	L	U	U/L
Initial product concept	0.25	4.0	16
Approved product concept	0.50	2.0	4
Requirements specification	0.67	1.5	2.2
Product design specification	0.80	1.25	1.6
Detailed design specification	0.90	1.10	1.2

Graphics, it is likely that the project will cost between 1.11 and 3.3 million Euro. In case of organization X, it can cost between 0.06 and 7.82 million Euro. Organization Y could account for the cost to be between 0.59 and 1.89 million Euro. In each of the organizations, the interval enriches the available forecast information.

It is possible to carry out the procedure for any given subset or grouping of the forecasts. For instance, it is helpful to group only initial forecasts. This way, the values found are applicable to any initial forecast made to obtain an indication of the uncertainty. Note that it is also possible to apply this method to only ex-ante predictions. This is interesting when many re-estimates are made. The values for the confidence intervals are different for each subset, confidence level and data set.

6.2.2. Benchmark values

In his book [52], McConnell gives benchmark values for the confidence limits with a confidence level of 80%. McConnell uses the same values as given by Boehm [6]. The values are summarized in Table 4, which also shows the U/L ratio. Note that McConnell and Boehm opted to make groups based on the project phases instead of our grouping of the forecasts based on the percentage of completion.

But how predictive are the values proposed by McConnell when they are applied to forecasts of different data sets? As the values of McConnell are based on Boehm's cone of uncertainty, these values are dependent on the specific conditions that form the cone. For instance, the forecasts in Boehm's cone are assumed to have the goal to quickly and accurately predict the actual value. We found that this is not always the case. Therefore, we want to know whether the values given by McConnell and Boehm are applicable even if the conditions under which the forecasts are made, are different from those of Boehm's cone.

When we compare the values of McConnell in Table 4 with those of our case studies in Table 3, we find them to be quite different. Most of the U/L ratios of Landmark Graphics and organization Y are smaller than those of McConnell. Also, the calculated quantiles L and U vary between all case studies and those of McConnell. Already, the applicability of the values of McConnell to our case studies is questionable.

In the above comparison, we fixed the confidence level to 80% and computed the corresponding confidence interval bounds. To further assess the quality of the values given by McConnell, we fix the confidence limits by applying his values for L and U to our case studies. With these intervals we compute the confidence levels they obtain in our cases, which according to McConnell should be about 80%.

However, the case studies either do not have information on the phase in which the forecasts are made or use different phases than those used by McConnell. Therefore, we need to translate the phases used by McConnell to the percentage of completion to be able to apply the benchmark values.

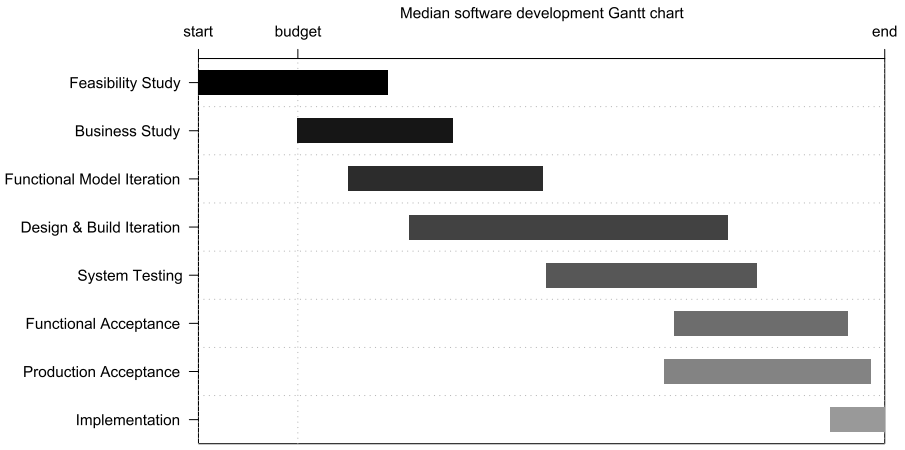

Fig. 14. An aggregated Gantt chart of median start and end times of different phases for 318 projects normalized for their duration (taken from an article by Eveleens et al. [17]).

Table 5
Translation between our Gantt chart and the McConnell phases.

McConnell phase	Organization phase(s)	% of completion
Initial product concept	Feasibility study	0%–14.4%
Approved product concept	Business study	14.4%–21.8%
Requirements specification	Functional model iteration	21.8%–30.8%
Product design specification	1/2 of Design & Build iteration	30.8%–40.7%
Detailed design specification	1/2 of Design & Build iteration, System testing Functional acceptance Product acceptance and implementation	40.7%–100%

Fortunately, in an article by Eveleens et al. [17], a translation was made between the phases and percentage of completion for organization Y. In that article, an aggregated Gantt chart is shown that gives the median start and end time of a phase normalized for the duration of 318 projects. For the sake of ease and completeness, we recall Fig. 14 from that article. However, the phases used in that organization are slightly different from those used by McConnell. Therefore, we mapped the phases of the organization to those used by McConnell in Table 5. Note that the translation presented in this table corresponds to the grouping of the forecasts we chose in Table 3.

In the translation to the percentage of completion, we used the median start dates of the Gantt chart to indicate the end of the previous phase and the beginning of a new phase. For instance, in the translation, the Feasibility Study ends with the median start of the Business study, and so on. As we have summarized in the table, the initial product concept of McConnell coincides with the feasibility study of the financial service provider Y. Given the 318 projects of which we have duration information this phase takes up to 14.4% of completion, and for very small projects this can drop to about 0%. The other phases are dealt with in a similar fashion. For those phases that do not coincide, we made pragmatic choices.

Using this translation, we applied the benchmark values given by McConnell to the forecasts of the case studies. We assumed that the forecasts in each individual case study were created using the same estimation method. We were unable to group forecasts based on the estimation method, simply because none of the case studies had necessary information on the methods used to create the forecasts. By assuming this, we were able to apply the confidence interval.

Although it is unlikely that the assumption will hold in all case studies, the analysis is not influenced by it. McConnell suggests applying the values to your forecasts without specifying anything about the estimation method used. Therefore, McConnell also makes the implicit assumption that the forecasts are made with the same estimation method, even though this will not always hold.

Each forecast in the case studies is multiplied by the upper and lower confidence limit of McConnell given in Table 4. For instance, a forecast made for Landmark Graphics at 23% of completion belongs to the Requirements specification phase of McConnell and is multiplied by 1.5 for the upper limit and 0.67 for the lower limit. This creates the confidence interval around each forecast. For each such constructed interval, we checked whether the actual value was indeed contained in it. That is, we assessed the confidence level of the confidence interval. In Table 6, we enumerate these confidence levels for each case study. Note that according to Boehm and McConnell this should be around 80% in all cases, and not a lot more or less.

The results show that the confidence limits proposed by McConnell do not always lead to accurate results. For Landmark Graphics and organization X, the confidence intervals do not contain the actual value close to 80% of the times at all. In these organizations, only sometimes the confidence interval contains the actual value. Recall that a confidence interval with a too

Table 6
The confidence levels for the case studies with the confidence intervals proposed by McConnell applied to them.

% of completion	LGC	X	Y costs	Y functionality
0%–14.4%	97.2%	53.5%	96.0%	100.0%
14.4%–21.8%	85.3%	23.9%	94.1%	100.0%
21.8%–30.8%	43.6%	23.1%	86.0%	100.0%
30.8%–40.7%	26.9%	18.2%	82.8%	57.1%
40.7%–100%	59.6%	10.8%	68.4%	58.6%

Table 7
Different translations from project phases to percentage of completion.

	Original	Shift +5%	Shift +10%
Initial product concept	0%–14.4%	0%–19.4%	0%–24.4%
Approved product concept	14.4%–21.8%	19.4%–26.8%	24.4%–31.8%
Requirements specification	21.8%–30.8%	26.8%–35.8%	31.8%–40.8%
Product design specification	30.8%–40.7%	35.8%–45.7%	40.8%–50.7%
Detailed design specification	40.7%–100%	45.7%–100%	50.7%–100%

Table 8
Confidence levels of the confidence intervals given by McConnell applied to the Landmark Graphics case study with different translations.

	Original	Shift +5%	Shift +10%
Initial product concept	97.2%	97.6%	97.8%
Approved product concept	64.8%	69.1%	74.9%
Requirements specification	38.1%	43.1%	52.6%
Product design specification	27.1%	28.7%	33.3%
Detailed design specification	55.7%	58.7%	62.0%

low or high confidence level is not helpful at all. Therefore, in these cases, the intervals do not provide much information. Only for organization Y, the intervals are reasonable for both cost and functionality. In the beginning, the confidence level is larger than 80% and in the end it is lower than 80%, nonetheless the intervals contain the actual value reasonably often.

To assess how sensitive these results are to the particular translation we applied between the phases and percentage of completion, we carried out two more translations that are summarized in Table 7. The first translation shifts the phases by 5%. The first phase takes 5% longer and the last one 5% shorter. The second translation shifts the phases by 10%. Using these translations, we again calculated the confidence levels that correspond to the confidence limits given by McConnell and Boehm. We summarized these confidence levels of the data of Landmark Graphics, in Table 8. We omitted the values for organizations X and Y as they show similar variations as the Landmark Graphics data and lead to the same conclusion.

Although Table 8 shows quite some variation in the confidence levels when different translations are used, for each of these translations we still draw the same conclusion in our analysis. Namely, the benchmark values given by McConnell do not provide for intervals with an adequate confidence level in different organizational settings.

The results are explained by the conditions under which the forecasts are made in each case study. The conditions that apply to organization Y are in alignment with the assumptions underlying Boehm's cone of uncertainty. And indeed, applying the confidence intervals gives reasonable results. In the other case studies, the conditions do not match with Boehm's assumptions, and our analysis reveals that in that case confidence intervals are unreliable. But even with organization Y, the intervals are not optimal.

Instead of using the benchmark values suggested by McConnell, we recommend to derive the limits of the confidence interval based on the historical data of the organizations themselves. To illustrate the difference between the McConnell values and organization-specific values, we visualized the different bounds of the intervals in Fig. 15. The dashed lines in the figures are the confidence interval ranges given by Boehm. The solid lines are the confidence intervals with a confidence level of 80% based on the data of the organizations themselves, as show in Table 3.

The figure visualizes the difference between what McConnell advocates to use and what the organizations must actually use to obtain 80% confidence intervals. The organization-specific confidence intervals differ significantly from those suggested by McConnell. Even the ones for the reasonably fitting intervals for organization Y of both cost and functionality are quite different. These results indicate that it is most useful for organizations to derive the values of the confidence interval using their own data.

Summary. Confidence intervals provide useful information to the managers at the time decisions need to be made and allow to enhance forecasting information for decision making. For instance, with the intervals questions are answered such as: what is a likely range of values that the actual can attain? We showed how to compute these organization-specific ranges of the intervals. Moreover, we illustrated that the benchmark values proposed by McConnell have very limited use. They are

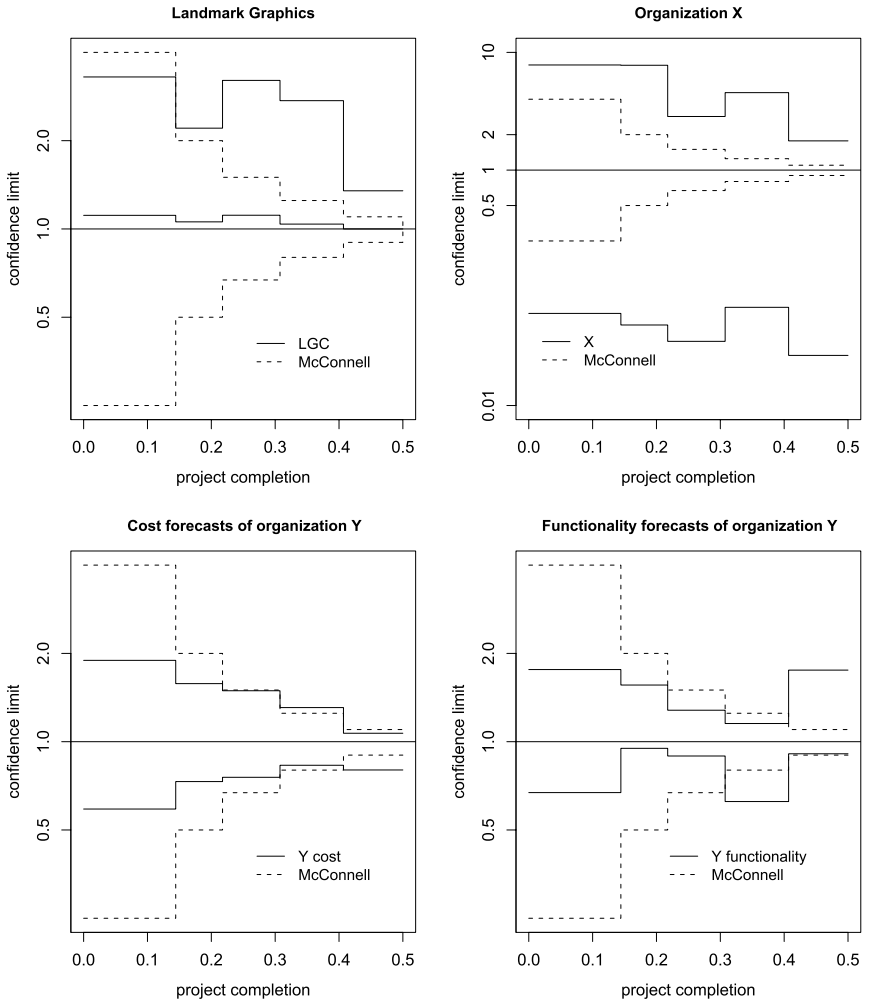

Fig. 15. Visualizing the differences between McConnell/Tockey confidence limits and organization-specific confidence limits.

only useful in organizations that make forecasts under similar circumstances as McConnell assumed. This implies that there is no bias, forecasting quality coincides, and phases fit perfectly. In other cases, the values of his intervals are too narrow or broad, since conditions differ from the assumptions made by Boehm. These benchmark values do not give an accurate indication of the uncertainty of the forecasts. Therefore, it is advisable for the organizations to compute the confidence limits based on their own homogeneous data.

6.3. Distribution of ratios

The third method that enriches forecast information is the distribution of the f/a ratios. This distribution is a generalization of the confidence interval. In that method, two quantiles of the entire f/a distribution are used. This method allows to assess any quantile that may be interesting. Moreover, this enables computing the historic chance of an actual being higher or lower than a given threshold.

To show how the distribution of the f/a ratios provides for valuable information, we will use the empirical distribution as an example. We use this distribution as it uses the available historic information and is generally applicable. In an article by Feller [20] and in a book by Fisz [21], a formal mathematical definition is given for the empirical distribution. Let $X_1, \ldots, X_N$ be mutually independant random variables of a cumulative distribution function $F(x)$. Let $X_1^*, \ldots, X_N^*$ be the variables rearranged in the ascending order of magnitude. The empirical distribution of the sample is the step function $S_N(x)$, defined by

$$S_N(x) = \begin{cases} 0 & \text{for } x < X_1^* \\ \frac{k}{N} & \text{for } X_k^* \leq x < X_{k+1}^* \\ 1 & \text{for } x \geq X_N^*. \end{cases}$$

This definition enables deriving the cumulative empirical distribution function of the f/a ratios. In Fig. 16, we show these functions of the initial forecasts of Landmark Graphics, organizations X and Y. The horizontal axis shows the value of the f/a

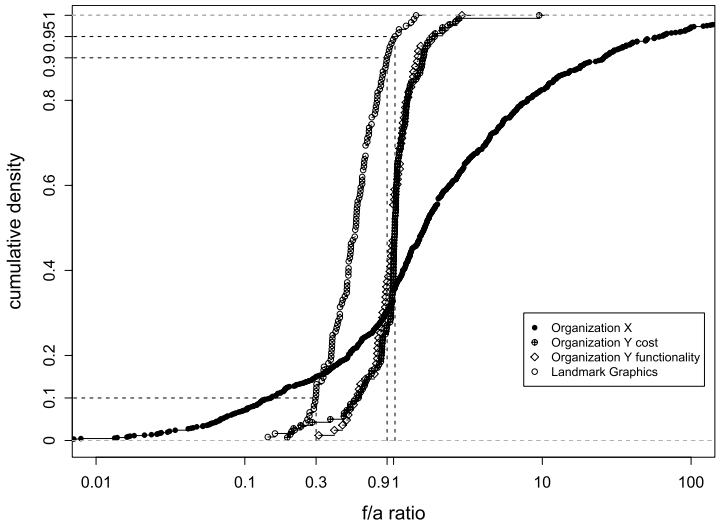

Fig. 16. Empirical cumulative density function of the initial forecasts of organization X, organization Y and Landmark Graphics. These distributions are not generally applicable to different data sets.

ratios. The vertical axis depicts the cumulative density. A point at f/a ratio 0.5 and cumulative density 0.4 means that 40% of the forecasts in the data had an f/a ratio of 0.5 or less. In a statistical package like R [62], these figures are easily created using the command `plot(ecdf(dataset))`.

Note that similar to the confidence interval, it is important to make an adequate grouping of the forecasts. In this figure, we chose to group the initial forecasts together, as we are interested in assessing the quality of these forecasts in our example.

We illustrate how this distribution function can enrich forecast information using an example. Consider again a project, in which an IT governor has to decide on a project proposal with forecasted cost of 1 million Euro. With the distribution of the f/a ratios, we are able to find the chance based on the historical data that the project will cost more. If the project is conducted by Landmark Graphics, we find that in 95% of the historical projects the initial f/a ratio is smaller or equal to 1. Thus, with a 95% probability the forecast will be smaller than the actual and the project will cost more than 1 million Euro. Moreover, we find there is a 10% chance the f/a ratio is smaller than 0.3. This indicates that with a probability of 10% the project will cost $1/0.3 \approx 3.33$ million Euro or more.

With this kind of information, an executive is able to assess the project proposal. This enables determining targets and commitments based on the risk the organization is willing to take, as we described in Section 2. For instance, a commitment could be made that the project should cost no more than 1 million Euro. However, the odds of achieving this commitment are slim. The IT governor could also commit to a project cost of 3.33 million Euro, which will have a fair chance of succeeding. Therefore, the additional information provided by the distribution of the f/a ratios is of great value.

6.3.1. Benchmark distributions

We showed that the distribution of the f/a ratios is valuable for decision making. However, the question remains what the distribution of the f/a ratios is. In our example, we have used the empirical distribution. The advantage of this distribution being that it is generally applicable to any data set, since it uses the actual historical data. In the literature, a number of alternative distributions are proposed. In an article by Pescio [56], distributions such as the beta, triangular and log–logistic distribution are suggested. Both Putnam [59] and Laranjeira [45] advocate the beta distribution. However, none of the authors provide any evidence that we are aware of that these distributions are an adequate approximation of the true distribution of the f/a ratios.

An exception to this statement is an article by Little [49]. In that article, Little analyzed the distribution of the ex-ante ratios. With ex-ante ratio we mean the forecast of the ex-ante part divided by the true ex-ante part, which is the actual minus the ex-post part, or mathematically: forecasted ex-ante/(actual − ex-post). He depicted two cumulative density plots: one of the initial f/a ratios and one of the ex-ante ratios of three other phases. Little compared the cumulative density plots with the one belonging to the lognormal distribution and found them to be quite similar. On the basis of these plots, Little concluded that the ex-ante ratios followed a lognormal distribution.

We and others [44] understood from Little's article that his finding of lognormality also applied to the f/a ratios. Through personal communication, Little assured us that the findings are only applicable to the ex-ante ratios. Still, for completeness sake, we want to check whether the f/a ratios are lognormal for two reasons. First, we are, in this article, interested in the distribution of the f/a ratios. And second, we wish to prevent further misinterpretations of Little's findings. Below, we will also statistically assess the lognormality of the ex-ante ratios. We are grateful to Little for that he provided us with his data to perform these analyses.

Table 9
Test results whether forecast/actual ratios of Landmark Graphics are lognormally distributed.

	Initial		Planning	
Test name	p-value	Outcome	p-value	Outcome
Shapiro–Wilk	0.546	Accept	0.000	Reject
Anderson–Darling	0.609	Accept	0.000	Reject
Cramer–Von Mises	0.520	Accept	0.000	Reject
Lilliefors	0.169	Accept	0.000	Reject
	Development		Stabilizing	
Test name	p-value	Outcome	p-value	Outcome
Shapiro–Wilk	0.000	Reject	0.008	Reject
Anderson–Darling	0.000	Reject	0.048	Reject
Cramer–Von Mises	0.000	Reject	0.075	Accept
Lilliefors	0.000	Reject	0.090	Accept

Table 10
Test results whether the initial forecast/actual ratios of organization X and organization Y are lognormally distributed.

	Organization X		Organization Y cost		Organization Y functionality	
Test name	p-value	Outcome	p-value	Outcome	p-value	Outcome
Shapiro–Wilk	0.000	Reject	0.000	Reject	0.022	Reject
Anderson–Darling	0.000	Reject	0.000	Reject	0.003	Reject
Cramer–Von Mises	0.000	Reject	0.000	Reject	0.002	Reject
Lilliefors	0.000	Reject	0.000	Reject	0.060	Accept

f/a ratios. The data provided to us by Little contains data of four phases. The phases used at Landmark Graphics were: the initial phase; the planning phase; the development phase; and the stabilizing phase. Each of the phases contains 121 f/a ratios and 121 ex-ante ratios. To statistically test the f/a ratios for lognormality, we took the log of the ratios and tested them for normality. To test the null hypothesis that the log of the data is normally distributed, we used four different tests. These tests are the Shapiro–Wilk test, the Anderson–Darling test, the Cramer–Von Mises test and the Lilliefors test. Each of these tests results in a p-value. This value indicates the likelihood that the null hypothesis is correct. For low p-values, we reject the null hypothesis. Statistically speaking, rejection means that the log of the data is not normally distributed, and thus the data are not lognormally distributed.

We applied the four tests to Little's f/a ratios of each of the four phases. The results of the tests are summarized in Table 9. As a threshold for accepting or rejecting the null hypothesis, we used $\alpha = 0.05$.

The table shows that we accept the null hypothesis that the initial f/a ratios are lognormally distributed. All the four tests indicate that the log of the data is normally distributed. The tests of the f/a ratios of the stabilizing phase give contradicting statements. The results are not conclusive, whether these ratios are statistically lognormal or not.

However, the f/a ratios made during planning and development are not lognormally distributed. The tests indicate that these f/a ratios do not adhere to a lognormal distribution. Therefore, as Little indicated, we cannot draw the conclusion that the f/a ratios are lognormally distributed. Note that removing a couple of projects does not significantly improve the p-values of these tests.

Moreover, we also tested the initial f/a ratios of the other organizations as depicted in Fig. 16, for lognormality. The p-values in Table 10 indicate that none of these data sets are statistically lognormal.

Ex-ante ratios. Since Little analyzed the ex-ante ratios, we also wanted to verify whether these are lognormally distributed. Again, we took the log of the ex-ante ratios and tested for normality with the same four tests. The results of the tests are summarized in Table 11. As a threshold for accepting or rejecting the null hypothesis, we used $\alpha = 0.05$ again.

The p-values in the table show that, except for the initial ex-ante ratios, which are equal to the initial f/a ratios, we reject the hypothesis, which means that the ex-ante ratios are statistically not lognormally distributed.

However, further analysis shows that the tests would not have lead to rejection when we remove 5 projects of the Planning phase, 10 of the Development phase and 3 of the Stabilizing phase. These outlying projects are relatively few projects, since they make up only 4%, 8% and 2% of all the projects. Therefore, from a practical perspective the data appear to be lognormal.

Yet, almost none of these projects overlap. If we were to remove these projects, it would mean removing 13% of all projects. Moreover, the projects to be removed are all the lowest f/a ratios within the data set. That is, the projects to be removed are not particularly random. Therefore, assuming the data are lognormal would be practically sensible, yet it does mean making the assumption certain low f/a ratios cannot be drawn from the approximate distribution that do occur in

Table 11
Test results whether ex-ante ratios are lognormally distributed.

	Initial		Planning	
Test name	*p*-value	Outcome	*p*-value	Outcome
Shapiro–Wilk	0.546	Accept	0.000	Reject
Anderson–Darling	0.609	Accept	0.000	Reject
Cramer–Von Mises	0.520	Accept	0.000	Reject
Lilliefors	0.169	Accept	0.000	Reject
	Development		Stabilizing	
Test name	*p*-value	Outcome	*p*-value	Outcome
Shapiro–Wilk	0.000	Reject	0.000	Reject
Anderson–Darling	0.000	Reject	0.000	Reject
Cramer–Von Mises	0.000	Reject	0.001	Reject
Lilliefors	0.000	Reject	0.004	Reject

practice. The evidence provided is not adequate to statistically verify lognormality for this data set, let alone that it provides evidence that it generally applies to other data. Therefore, others who wish to use such a distribution should be cautious to apply it to their data.

General theoretical distribution. Is it reasonable to assume that there exists a general theoretical distribution that applies to most f/a ratios? In the previous section, we showed that the values of the confidence intervals depend on the assumptions under which the forecasts are created. Since these intervals are a summary of the distribution of the f/a ratios, it is likely that these distributions also depend on the assumptions made. In Fig. 16, we plotted the cumulative density functions of our case studies. The figure shows, similar to the confidence intervals, that the distributions are quite different for each organization. The figure indicates none of the distributions are part of a common theoretical distribution.

If there is a theoretical or practical basis that the data should adhere to some distribution, from a practical perspective one could use it. For instance, in Little's case, one could assume the ex-ante ratios to behave lognormally. However, choosing a distribution of which it is uncertain it applies to the data, can result in no information or even misinformation to be gained. Since the distributions suggested in the literature lack the theoretical basis and statistical evidence to support their claim, we caution others to apply these distributions to their own data set. We advocate the use of the empirical distribution, if there is no such basis or evidence to support a choice of distribution. When enough data are present, one is able to use that empirical distribution on any data set to provide management information.

Summary. We showed that the distribution of the f/a ratios provides valuable information to IT governors. Such a distribution allows to determine the historical chance of an actual being higher or lower than a given threshold. In the literature, a number of distribution are proposed, yet we found that the theoretical basis and statistical evidence lacks to support these suggestions. From a practical perspective, one could apply them, but we caution to use these distributions and advise to use the empirical distribution.

7. Benchmarks

In this section, we discuss and review benchmarks found in the literature that are related to forecasting. We discuss the origins of the benchmarks and assess whether they should be used for comparisons. It turned out that some of them are inapplicable for benchmarking, as we will argue in this section.

First, we discuss a number of benchmarks found in the literature related to EQF values. One benchmark is an average EQF value of DeMarco [14]. We also discuss the median benchmark values given by Lister [47]. Although there is no data supporting the benchmark values of Lister, they appear to be an adequate indication of forecasting quality at the time of writing this article. The values given by Lister compare well with the values we find in our case studies.

Second, we address the benchmark figures reported by Standish Group on successful and challenged projects. In another article [18], we argue with the help of the tools proposed in this article that these figures are meaningless for benchmarking. For instance, we show that the definitions used by Standish do neither incorporate underruns nor the political nature or other biases of the forecasts. In this article, we introduce new definitions that do incorporate underruns and the bias of forecasts, to enable making proper comparisons.

Third, besides the Standish Group benchmarks on project success, we also discuss overrun benchmarks reported in the literature. Already, for many years a large variety in overruns have been reported. And although almost all the authors mention potential politics involved in their quantifications, none of them show how their figures are influenced by this recognized phenomenon. Consequently, their benchmarks provide an unclear picture and are useless for any meaningful benchmarking.

Table 12
Summary of benchmark EQF values found in the literature and our case studies.

Source	Median EQF	Average EQF	Number of projects
Organization X	0.43	1.6	867
DeMarco	–	3.8	Unknown
DeMarco–Little	1.9	4.2	20
Landmark Graphics 1999	4.7	6.3	121
Lister	4–9	–	–
Organization Y functionality	6.4	9.9	83
Landmark Graphics 2001 [48]	7.0	–	–
Landmark Graphics 2002 [48]	7.6	–	–
Landmark Graphics 2003 [48]	8.4	–	–
Organization Y cost	8.5	36.9	140
Kulk et al. [42]	9.4	–	221

7.1. EQF benchmarks

Despite the fact that using EQF values is a sound idea, information in the literature on the subject is sparse: over the three decades that this notion is published, we found only a handful of benchmarks. In this section, we will discuss them. One benchmark is one of the most widely known benchmarks for EQF values that stems from the creator of the EQF: Tom DeMarco. Another benchmark is given by Lister. In addition, Little provided recent EQF information. Finally, we found an EQF benchmark in an article by Kulk et al. [42]. Together with our cases studies, we will present these EQFs and propose them as new benchmark values. We encourage others to do the same, so that more EQF benchmarks become available to make meaningful comparisons.

We summarized all the benchmarks we know in Table 12 and sorted them based on their median EQF value. The table contains two benchmark values from DeMarco. The first value is the widely known benchmark that he gives in a footnote in his book [14, p. 157]. We contacted DeMarco and requested the data that he used to derive the results. Unfortunately, he was unable to reproduce his data. Fortunately, Little analyzed a graph in DeMarco's book that contained the data and approximated the data points. We gratefully thank Little for providing us with these data, that he also used in his article [48]. The data consist of EQF values of 20 projects. Due to the approximation of the data points, the average value of these data is slightly different from the one DeMarco reported in his book.

Through personal communication, we obtained details on the figures reported by Lister [47], but again there were no data available. The values of our case studies were calculated using Formula (1). Of each benchmark, we give the median and average EQF value. We also state the number of projects that were used to derive each benchmark.

Let us discuss Table 12, now. In his book, DeMarco states that the values he found are not that great. Indeed, when compared with the other benchmarks the EQF values of DeMarco are relatively low. He also describes that an average EQF of 10 should be attainable. However, the benchmark of DeMarco is an average value. Recall that we argued that the median value is better for comparisons as it gives a more accurate description of the quality of the forecasts. Since large values influence the median not as much as an average, the difference between the two can become quite large with skewed distributions. We see this, for instance, with the cost forecasts of organization Y: whereas the median is 8.5, the average amounts to 36.9. Therefore, the average EQF values given by DeMarco are difficult to use for comparisons.

The benchmarks of Lister are found in his article [47]. This article mentions a norm of 4 and the highest sustainable scores of 8–9. However, the article does not state what is meant by the figures reported. We contacted Lister and learned that they represent median values. He indicated that the figures are derived based on his own experience. By helping companies for over twenty years to use the EQF metric, he has seen a lot of EQF values. This allowed him to make an assessment of the overall EQF values. Although there are no data supporting his figures, they are in line with our values found in the case studies. Namely, if we consider Landmark Graphics to be the norm, we also find a median EQF value of 4 and a highest sustained median score of 8.5 by organization Y for costs. In the literature, we even find a highest median score of 9.4.

When we compare the case studies with the benchmarks from the literature, we conclude that organization X has poor forecasting quality, Landmark Graphics seems reasonable, and organization Y has good forecasting quality. However, due to the lack of sufficient benchmarks, we are unable to assess how good or how bad the forecasting qualities really are.

The benchmarks shown in this section are a start to reference the quality of your forecasts with these organizations. Of course, it would be interesting to see more organizations calculate their EQF values and report on these figures. This will allow for better comparisons in the future. Needless to say, it must always be reported how these values are calculated and whether the values represent average or median values. The benchmarks that we have given are calculated using Formula (1) of Section 4.2.1 and represent median values.

7.2. Project success

The second benchmark related to forecasting we discuss, is the successful and challenged project figures published by Standish group [29]. In another article [18], we argue in detail that these figures are meaningless for benchmarking. For

Table 13
Standish project benchmarks over the years.

Year	Success	Challenged	Impaired
1994	16%	53%	31%
1996	27%	33%	40%
1998	26%	46%	28%
2000	28%	49%	23%
2004	29%	53%	18%

Table 14
Comparing Standish success to real estimation accuracy.

Source	Success	Challenged	Median EQF of initial forecasts
Organization X	67%	33%	1.1
Landmark Graphics	5.8%	94.2%	2.3
Organization Y cost	59%	41%	6.4
Organization Y functionality	55%	45%	5.7
Organization Y combined	35%	65%	6.5
1/Landmark Graphics	94.2%	5.8%	2.3

completeness sake, we will give a short overview of the findings of that article. It turns out that the Standish figures are not at all in accordance with reality.

To alleviate the problems with Standish's definitions found in our other article, in this article we propose alternative definitions without these drawbacks. These new definitions use our reference cone to incorporate forecasting quality. We applied our notion of plan accuracy to the case studies, and now the results are well in line with reality. Note that we do not discuss Standish Group's definition of failing projects as this is not related to forecasting.

7.2.1. Chaos definitions

Here, we briefly discuss the findings of another article [18]. In that article, we discuss project success as Standish group defined it in their Chaos reports. For completeness sake, we recall that Standish group classifies projects into three groups [29] as follows.

- Resolution Type 1, or project success: The project is completed on-time and on-budget, with all features and functions as initially specified.
- Resolution Type 2, or project challenged: The project is completed and operational but over-budget, over the time estimate, and offers fewer features and functions than originally specified.
- Resolution Type 3, or project impaired: The project is cancelled at some point during the development cycle.

We also recall the various benchmarks of the above project categories that are published by Standish Group. We enumerate them in Table 13. These figures are published in various articles [27,29–31]. They are rather alarming, and as such widely cited.

In our other article, we argue that the Standish definitions have four major problems. First, the definitions are misleading as they are only about the forecasting accuracy of cost, time and functionality. However, they named the categories successful and challenged projects, implying much more than forecasting accuracy of these dimensions.

Second, the definitions do not account for underruns of cost and time and overruns of functionality. Using the case studies of organization Y, we illustrate that applying the Standish definitions leads to unrealistically low rates.

Third, steering on the definitions perverts forecasting accuracy. In the case study of organization X, we found that they used the Standish definitions to determine the success of projects. This resulted in overstating budgets to increase the margin of success. Yet, it degraded the quality of the forecast as we showed using our tools.

Fourth, applying the definitions leads to meaningless rates that do not reflect the reality. We applied Standish's definitions to our case studies, which we discussed in this article, to compute their success rates. Since we want to compare these rates with those we will derive using our alternative definitions later on, we recall them in Table 14.

The table shows that organization X is considered highly successful by Standish. However, the quality of the initial forecasts in terms of EQF is lowest of all the case studies. Moreover, the table introduces a fictitious organization 1/Landmark Graphics, which represents an organization with the exact forecasting accuracy of Landmark Graphics, but an overrun becomes an underrun and vice versa. The success rate of that organization shows that by inversing the bias of Landmark Graphics, they become highly successful. We found that these success rates of our case studies do not reflect the reality at all.

7.2.2. Proposed definitions

Above, we addressed a number of issues that render the Standish figures meaningless. But, how do we improve the Standish Group definitions then in order to derive meaningful rates? Such definitions must at least take into account

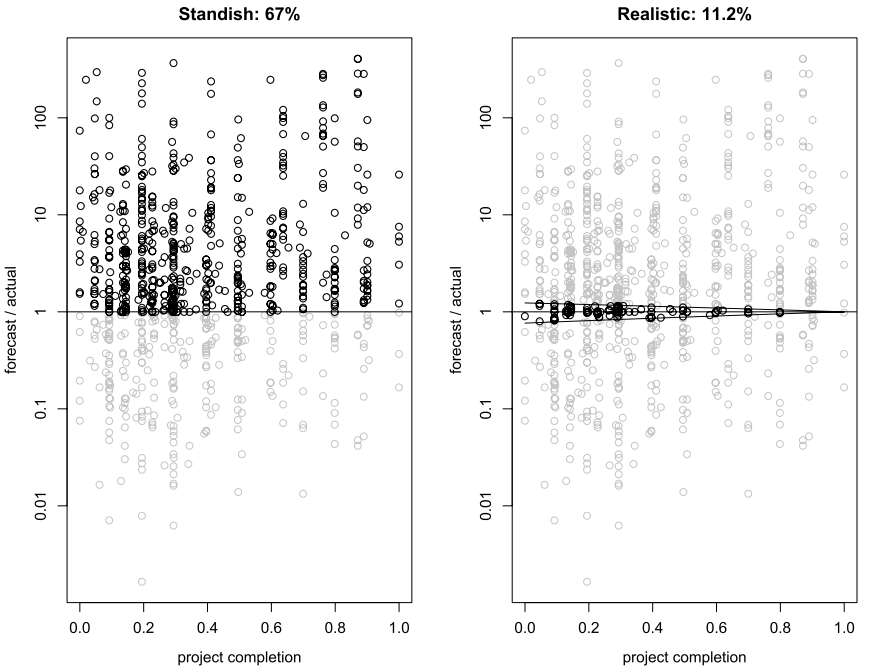

Fig. 17. Real-world data illustrates that Standish' project success definition is not realistic. Example was drawn using data from organization X.

underruns for cost and time and overruns for functionality. Also, they should preferably account for the political nature of the forecasts. Note that we will use the term *plan accuracy* instead of project success to correct the first problem of the Standish definitions.

Let us first consider what it means for a project to be plan accurate. A project is plan accurate when the initial forecast does not deviate too much from the actual, for both underrun and overrun. But, when do we consider an initial forecast to be 'too far' from the actual?

To answer this question, we use two tools discussed in this article: the EQF and the reference cone. Recall that the EQF is a measure of quality that quantifies the deviation of a forecast to the actual. If the deviation becomes greater, the EQF becomes smaller and vice versa. Recall that the reference cone is a description of how the forecasts should behave at given predefined conditions and quality. The quality of a reference cone is expressed in terms of the EQF.

These tools allow us to describe whether an initial forecast is 'too far'. First, we assess under which cone conditions, as we discussed in Section 3.1, the initial forecasts are made. This translates to a family of reference cones. Subsequently, we determine the quality we wish the forecasts to adhere to in terms of an EQF value. This leads to a corresponding reference cone, with the bandwidths of which we use to determine what f/a ratios are plan accurate. We consider an f/a ratio within the bandwidths of the reference cone as plan accurate. Every f/a ratio outside the predefined reference cone is considered plan inaccurate.

We illustrate this idea in Fig. 17. The plot on the left shows Standish' project success definition applied to data from organization X. All f/a ratios larger or equal to 1, the black dots, are considered by Standish to be successful projects. The f/a ratios smaller than 1, the grey dots, are according to Standish unsuccessful. The right-hand side plot illustrates our idea to the same data. The plot contains a reference cone as described by Formulas (4) and (5) with a predefined EQF of 8.5. This means that the reference cone describes how the f/a ratios should behave when there is no political bias, and the knowledge of what has been done, is used in making the forecast. The EQF value is based on a realistically obtainable median quality, which we found for our best case study and is in line with Lister's experience. All f/a ratios within the reference cone, the black dots, are considered plan accurate. The grey dots outside the reference cone are plan inaccurate.

Note that there is no incongruence with what we discussed in Section 4.2.2. There, we argued that the reference cone only gives an *indication* of the quantified quality of the f/a ratios in terms of the EQF. That is, it is possible for a project to have all f/a ratios in the reference cone, but have an EQF value lower than that of the reference cone. Similarly, it is possible for a project to have some f/a ratios outside the reference cone, but still have a higher EQF value.

However, in this case, we are only interested in the initial f/a ratios and not the progression of these ratios during the project. We judge the initial ratio by assessing whether it falls within bandwidths we find acceptable at any given moment. The cone specifies these bandwidth. We assume that the initial f/a ratio has the potential quality to achieve the predetermined quality level if it falls within the bandwidth and is thus considered plan accurate.

The data used in the figure consist of the initial forecasts of organization X. In theory, one would expect the initial forecast at or near the start date of the project. However, the figure shows that in reality this need not be true. In the other case studies,

we also found spread in the time of the initial forecasts, except for Landmark Graphics. There, all initial forecasts were made at the start date of a project.

The figure clearly illustrates the difference between the Standish definitions and our idea. With our methods, we solve the second and third problem of the Standish definitions, which are one-sided leading to unrealistic rates and perverting forecasting accuracy. Our reference cone defines which deviations we consider 'too far', not only in case of overruns, but also for the underruns. The right-hand side plot already shows a more realistic assessment of the plan accuracy of organization X than with Standish's definitions, even though this particular reference cone does not yet take into account the political bias of the organization. We will show how this is done below, but first we translate our idea into proper definitions.

Plan accuracy definitions. Informally stated, our idea to compute the percentage of plan accuracy for a data set is as follows. First, we pick an EQF value, preferably a realistic one. We choose the EQF in such a way that the value represents the quality of the forecasts that we find acceptable. Then, we describe the conditions to which the forecasts should adhere, and draw the corresponding reference lines. For example, in Section 3.1, we assumed that the forecasts had no political bias and the ex-post part was known and used. In Section 4.2.1, we showed how these conditions lead to a mathematical description of the lines, resulting in Formulas (4) and (5). With the reference cone drawn, we plot the f/a ratios of the initial forecasts in the same picture. Every f/a ratio that is within or equal to the reference cone is plan accurate, while every ratio outside is not. The number of ratios inside the cone divided by the total amount of f/a ratios, is the plan accuracy percentage.

The above is more formally stated in the following definitions. Assume that the initial forecasts and their actuals of time, cost and functionality are given. Furthermore, we assume the quality of the forecasts with respect to a preselected EQF value, is given by a reference cone. Then we define:

- A project is plan *accurate* with respect to a preselected EQF if the initial forecasts of time, cost and functionality divided by the actuals are contained in or equal to the reference cone.
- A project is plan *inaccurate* with respect to a preselected EQF if at least one initial forecasts of time, cost or functionality divided by the actual is not contained in the reference cone.

The advantage of the Standish definitions is that they are simple and the drawback is that one cannot use them. The drawback of our definitions is that the reference cone that is chosen must be carefully described, but the advantage is that they are useful.

The proposed definitions allow for two types of comparison. In the first type of comparison, we choose an EQF value and prespecify a reference cone in order to compare each organization with these preselected conditions. In this situation, we assume that there is a desirable way of forecasting and see how well the organizations are able to forecast in such a way. This comparison takes both underruns and overruns into account, however does not account for the politics involved in the organizations.

In the second type of comparison, we choose an EQF value, but compute organization-specific reference cones for every organization. In this case, we assume there are no optimal conditions the forecasts should adhere to. We acknowledge forecasts of different organizations that are made under different conditions, and we want to compare the quality of the forecasts that are made under the conditions present. This comparison takes into account underruns and overruns as well as the politics involved in forecasting. Below, we show the different types of comparisons by applying them to our case studies.

7.2.3. Preselected cone conditions

In the first type of comparison, we apply the same reference cone to each organization. This means that we state how forecasts should be made, and compare the case studies to see how well they are able to forecast according to this standard. For this comparison, we choose as an example an EQF value of 8.5 as the quality the forecasts should adhere to. We use the conditions of Section 3.1 to define how the forecasts should be made. This means, the goal of the forecast should be to quickly and accurately predict the actual value and the ex-post part is known and used. In Section 4.2.1, we showed this results in reference lines as given by Formulas (4) and (5). With the EQF value chosen to 8.5, the reference cone that we apply is thus given by:

$$l(t) = t + \left(1 - \frac{2}{8.5}\right) \cdot (1 - t) = \frac{4}{17}t + \frac{13}{17}$$
$$u(t) = t + \left(\frac{2}{8.5} + 1\right) \cdot (1 - t) = \frac{21}{17} - \frac{4}{17}t.$$

Table 15 shows the results of applying this reference cone to each case study. The table shows results more in accordance with the real situation of our case studies.Landmark Graphics shows limited plan accuracy when compared with organization Y, since its median EQF is lower, and since their forecasting is politically biased. Organization X behaves rather poor when compared with organization Y, since they grossly overestimate with a very poor EQF. Organization Y has a reasonable plan accuracy, since they are able to forecast according to the standard we uphold here.

Note that the rate of cost and functionality combined for organization Y is lower than the individual accuracy rates. Partly, this is caused by the chosen EQF of 8.5. If we would have chosen a smaller EQF value, the difference in rates will diminish.

Table 15
Plan accuracy with an EQF quality of 8.5.

Organization	Plan accurate	Plan inaccurate
Landmark Graphics	19.0%	81.0%
Organization X	11.2%	88.8%
Organization Y cost	56.4%	43.6%
Organization Y functionality	51.8%	48.2%
Organization Y combined	25.5%	74.5%
1/Landmark Graphics	15.7%	84.3%

But, partly it is also inherent to the definition of the plan accuracy rate. The cost and functionality forecast for a single project must both be accurate enough to consider the project accurate. If either one does not comply, the project already becomes inaccurate. Therefore, the combined accuracy rate will always be lower than the individual accuracy rates.

In the table, we again added 1/Landmark Graphics. Recall that this fictitious organization resembles the exact opposite of Landmark Graphics. That is, the deviations to the actual are the same, but an underrun becomes an overrun and vice versa. We showed that with the definitions of Standish, 1/Landmark Graphics is highly successful even though the deviations to the actual on a logarithmic scale are the same as Landmark Graphics. Table 15 shows that with our definition of plan accuracy, 1/Landmark Graphics is about as successful as Landmark Graphics. In fact, the fictitious organization is slightly less plan accurate, but this is due to the asymmetric nature of the f/a plot. We allow more leniency for understating forecasts as a result, which is unfavorable for the fictitious organization. Finally, the values of Table 15 seem to be low, but this is because we set the forecasting quality quite high with an EQF value of 8.5.

This comparison is similar to Standish's benchmark, with the exception that we account for underruns as well. However, we argued in our other article [18] that the Standish figures are unreliable, since they combined the results of different organization that have different biases. Similarly, with this comparison we are not able to make generalized statements about the plan accuracy of these organizations. We are only able to state how well organizations are able to adhere to the standard we set. In order to make generalized statements about plan accuracy, we need to account for the fourth problem of the Standish definitions, namely, the political bias. This is done in the second type of comparison that we describe below.

7.2.4. Organization-specific cone conditions

In the second type of comparison, we acknowledge different biases exist in forecasting and we do not have an opinion on which one is better. Instead of comparing with a chosen standard, we want to know what the quality of the forecasts is given their bias. To obtain such a comparison, we need to debias the f/a ratios. We do this by drawing reference lines similar to the comparison above, however, each reference cone is different for each organization. Thus, we need to construct reference cones for the political bias found in the case studies. In Section 4.2.1, we showed how to perform such calculations. Given a number of cone conditions, we are able to compute corresponding reference lines. Here, we will apply the same methodology, however, we will use varying cone conditions.

One of the cone conditions used to create the reference lines in Section 4.2.1, is the goal condition. We assumed the goal of the forecast is to predict without bias as quickly and accurately as possible the actual value of interest for the project. In the case studies, we found different goals. For instance, in Landmark Graphics the goal was to forecast the minimal value and with organization X it was to predict the actual value plus a safety margin. In these cases, the initial forecasts will not center around $f/a = 1$, but some other value, for instance, $f/a = 0.5$ for Landmark Graphics or $f/a = 3$ for organization Y. Therefore, in the calculations of the reference lines, the actual is no longer 1, but rather 0.5 or 3 at the start of the project.

Another cone condition we used to compute the reference cone of Section 4.2.1, is the ex-post inclusion. We assume that each consecutive forecast incorporates the ex-post part and we assume this part is known with certainty causing the convergence of the reference lines. However, in the case study of organization X we did not find convergence to the actual. Instead, the forecast accuracy remained constant as the project progressed. In such circumstances, the ex-post part is not used and, for the sake of our computations, it is zero.

Deriving the reference lines based on these varying cone conditions results in reference cones that are in line with the forecasting practice of the organizations. Then, these lines also account for, for instance, the political bias. Below, we will show how to make the calculations given the cone conditions of our case studies.

7.2.4.1. Landmark Graphics. In the Landmark Graphics case study, the ex-post part was known and used, but the goal was to predict the earliest possible date the project could finish. However, it is not possible to objectively determine this goal. Therefore, we need to approximate the earliest possible date pragmatically. The median f/a ratio of the initial forecasts of Landmark Graphics is 0.56. As the project progresses, the earliest possible date will gradually increase to 1 since the ex-post part is known and used. Thus the new reference point, formerly the actual, at time t becomes $0.56 + 0.44t$. We assume the ex-post part growth is constant, leading to an ex-post part of t. This makes the ex-ante part $0.56 + 0.44t - t = 0.56 - 0.56t$. We are able to predict this within $1/c_1$ to c_2 times leading to, for the lower limit line, $1/c_1 \cdot (0.56 - 0.56t)$ and, for the upper limit line, $(0.56 - 0.56t) \cdot c_2$. We reiterate the procedure with the EQF as we did in Section 4.2.1, but now the surface below the reference point is not 1, but $\int_{t=0}^{1} 0.56 + 0.44t \, \mathrm{d}t = 0.78$. Then the reference lines that account for the biases in

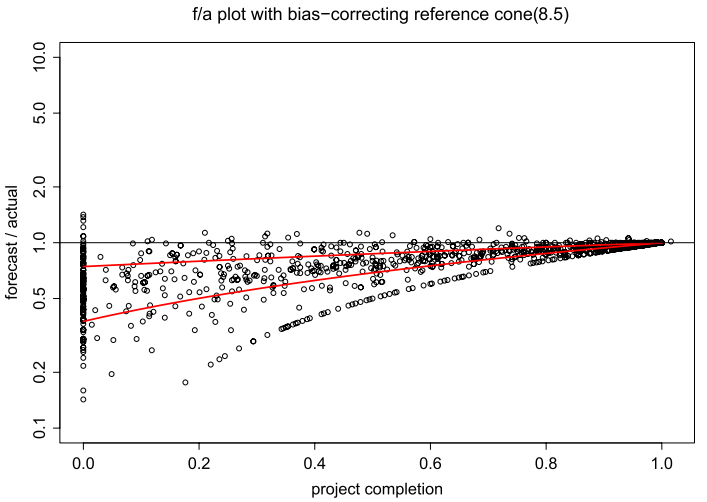

Fig. 18. f/a plot of Landmark Graphics of all data with a bias-correcting reference cone.

Landmark Graphics are given by:

$$l(t) = t + \left(1 - \frac{0.78}{0.28 \cdot \mathrm{EQF}_l}\right) \cdot (0.56 - 0.56t) \approx 0.376 + 0.624t$$
$$u(t) = t + \left(1 + \frac{0.78}{0.28 \cdot \mathrm{EQF}_u}\right) \cdot (0.56 - 0.56t) \approx 0.744 + 0.256t.$$

To illustrate that the reference cone now accounts for the bias of Landmark Graphics, we depict in Fig. 18 the same f/a plot as used in the case studies. However, in this case the above reference cone that accounts for the political bias is drawn. The figure shows that the reference lines are well in line with the bias we find in the data. Similar to the calculations of Landmark Graphics, we repeat the method for the other organizations to find their biased reference cones.

Organization X. For organization X, different cone conditions hold. In this organization, we found the ex-post part was not used and the goal is not to forecast the actual, but rather the actual plus a large safety margin. However, it is not possible to objectively determine the safety margin used. Yet, it is possible with pragmatic assumptions to approximate the goal. The median f/a ratio of the initial forecasts of organization X is 1.625. Since the f/a ratios do not converge to the actual, the forecasts at time t are aimed at predicting the actual value times 1.625. With these assumptions, we repeat the calculations. Since the ex-post part is not used, the ex-post part is 0. The ex-ante part is the reference point minus the ex-post part, in this case 1.625–0. We reiterate the procedure and find that the reference lines that account for the biases in organization X are given by:

$$l(t) = 1.625 - \frac{1.625}{\mathrm{EQF}_l} \approx 1.434$$
$$u(t) = 1.625 + \frac{1.625}{\mathrm{EQF}_u} \approx 1.816.$$

Organization Y cost. For organization Y, we found the cost and functionality forecasts to be in line with the cone conditions as we described in Section 3.1. That is, the goal of the forecasts is to predict as quickly and accurately as possible the actual without bias and the ex-post is known and used. Therefore, we expect the reference lines for this organization to be similar to Formulas (4) and (5) we found in Section 4.2.1. However, to make the comparison fair, we will apply the same procedure used for the other case studies to this organization.

Again, for the cost forecasts we pragmatically approximate the goal of the organization by taking the median f/a ratio of the initial forecast. We find the median to be 1.02. Since the ex-post part is known and used as the project progresses, the ratio will gradually decrease to 1, thus the new reference point at time t becomes 1.02–0.02t. We assume the ex-post part growth to be constant, thus the ex-post part is equal to t. This makes the ex-ante part $1.02 - 0.02t - t = 1.02 - 1.02t$. We are able to predict this within $1/c_1$ to c_2 times leading to bounds of the ex-ante part of $1/c_1 \cdot (1.02 - 1.02t)$ and $(1.02 - 1.02t) \cdot c_2$. We reiterate the procedure and find that the reference lines that account for the biases of the cost forecasts in organization Y are given by:

$$l(t) = t + \left(1 - \frac{1.01}{0.51 \cdot EQF_l}\right) \cdot (1.02 - 1.02t) \approx 0.782 + 0.218t$$

Table 16
Plan accuracy accounted for biases with an EQF quality of 8.5.

Organization	Plan accurate	Plan inaccurate
Landmark Graphics	57.9%	42.1%
Organization X	7.7%	92.3%
Organization Y cost	56.4%	43.6%
Organization Y functionality	51.8%	48.2%
Organization Y combined	25.5%	74.5%

$$u(t) = t + \left(1 + \frac{1.01}{0.51 \cdot EQF_u}\right) \cdot (1.02 - 1.02t) \approx 1.258 - 0.258t.$$

Note that these reference lines only differ slightly from the theoretical lines given by Formula (4) and (5). This again indicates the high quality of the forecasting process at this organization.

Organization Y functionality. For the functionality forecasts, we find the median f/a ratio of the initial forecasts to be 1.00. In this case, the reference lines are thus equal to Formulas (4) and (5) we found in Section 4.2.1. Therefore, the reference lines accounting for the biases of the functionality forecasts are given by:

$$l(t) = t + \left(1 - \frac{2}{\mathrm{EQF}_l}\right) \cdot (1 - t) \approx 0.765 + 0.235t$$
$$u(t) = t + \left(1 + \frac{2}{\mathrm{EQF}_u}\right) \cdot (1 - t) \approx 1.235 - 0.235t.$$

7.2.4.2. Organization Y combined. In case of the combined cost and functionality forecasts, we have a median f/a ratio of 1.04 for the costs and 1.01 for functionality. Following the same procedure as above, we find the reference line for the costs to be:

$$l(t) = t + \left(1 - \frac{1.02}{0.52 \cdot EQF_l}\right) \cdot (1.04 - 1.04t) \approx 0.800 + 0.200t$$
$$u(t) = t + \left(1 + \frac{1.02}{0.52 \cdot EQF_u}\right) \cdot (1.04 - 1.04t) \approx 1.280 - 0.280t.$$

The reference lines of the functionality forecasts is given by:

$$l(t) = t + \left(1 - \frac{1.005}{0.505 \cdot EQF_l}\right) \cdot (1.01 - 1.01t) \approx 0.774 + 0.226t$$
$$u(t) = t + \left(1 + \frac{1.005}{0.505 \cdot EQF_u}\right) \cdot (1.01 - 1.01t) \approx 1.246 - 0.246t.$$

Applying organization-specific cones. With the different political biases accounted for in each reference cone, we are able to compare the quality of the forecasts made with respect to the conditions present in each organization. We apply our proposed definition to the case studies with each using their own reference cone. Fig. 19 shows the results of applying the definitions to Landmark Graphics, organization X and organization Y. In the figure, the black dots represent the plan accurate projects and the grey dots the plan inaccurate projects. In Table 16, we summarize the percentage of f/a ratios of each organization that are contained within their own reference cone.

The percentages in the table indicate the plan accuracy in each organization with respect to their forecasting process. Landmark Graphics, in contrast to our previous comparison, turns out to have a very reasonable plan accuracy with a rate of 58%. Although they predict the earliest possible date instead of the actual, they are capable of making very reasonable forecasts for this goal. As we noted before, executives can decide not to change forecasting policies to remove the bias found in their organization. Since Landmark Graphics is capable of making reasonably accurate forecasts, the management could account for the bias in their own calculations and leave the forecasting process as it is. In Section 6, we showed how executives are able to account for the bias in their calculations.

Organization X remains to have poor plan accuracy. Even though we account for the political bias found in this organization, we use a rather high EQF value of 8.5. The spread of the f/a ratios of organization X is much larger than the quality we compare with. Therefore, the plan accuracy of this organization remains poor.

Of organization Y, the percentages did not change with respect to our previous comparison. The reason is that the reference cones we used in that example, are nearly identical to the reference cone that accounts for the political bias of the organization. Similar to the other comparisons, we find this organization to have a reasonable forecasting quality for both cost and functionality when compared with the other organizations.

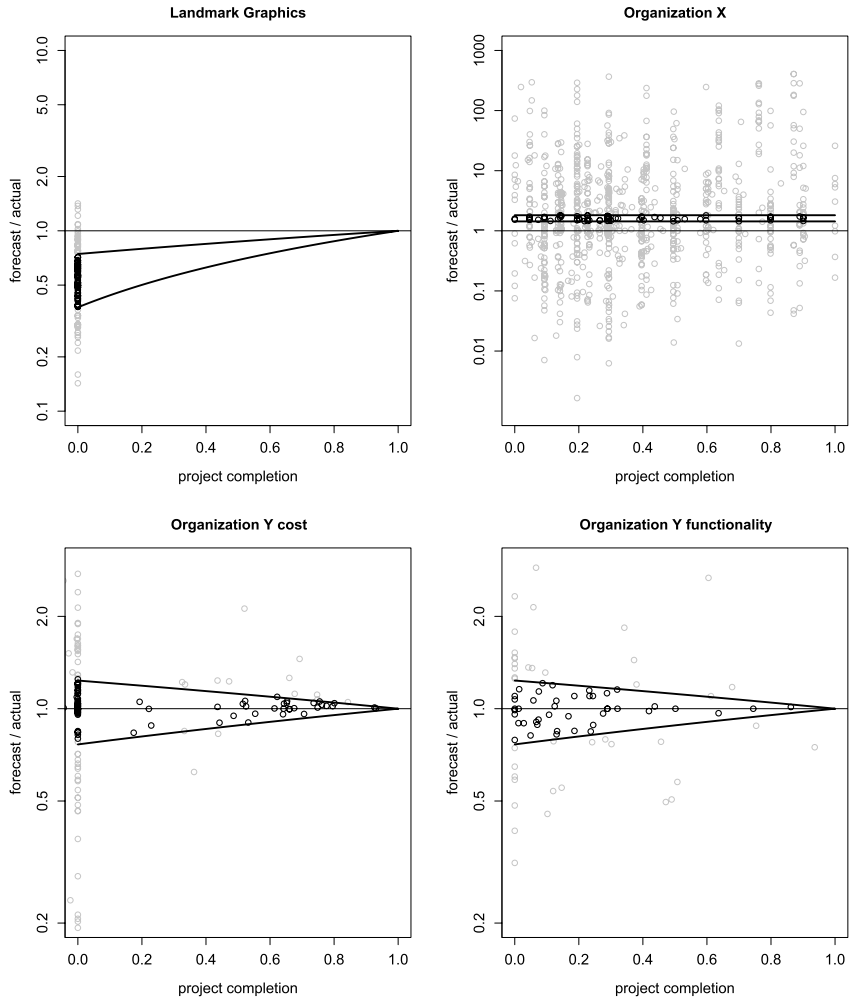

Fig. 19. True plan accuracy comparison of the three organizations while accounting for the biases.

Since the figures of Table 16 do account for the political biases present, we are able to make more general statements about the plan accuracy of the organizations. However, in our case studies we only had either time or cost or functionality and in one case cost and functionality at our disposal. Therefore, summarizing these figures will not result in meaningful benchmarks.

Concluding, our analyses based on the historical data clearly illustrate that our proposed definitions create meaningful results that are usable for true comparisons between organizations. We encourage Standish to choose an EQF value and define reference cones for the different organizations, so they are able to reiterate their analyses using our definitions.

7.3. Overruns

Another series of public benchmarks related to forecasting deal with cost or time overruns. A cost overrun, or equivalently cost underestimation, means the forecasted cost were lower than the actual cost, or $f/a < 1$. In this section, we discuss the numerous benchmarks reported on in the literature. We argue that the large variation in these figures could very well be caused by the politics of forecasting. Although most authors underscore the importance of politics as the main problem with their data, none of them quantify its influence on the quality of the forecasts underlying their benchmarks. We suggest using a desired median EQF and reference cone as benchmarks instead of the figures reported on overruns in the literature.

As said, many researchers reported benchmarks on cost and time overruns. In Table 17, we give an overview of the articles we are aware of on this subject. The figures reported in the table are calculated by taking the difference between the actual value and the initial forecast. This difference is divided by the initial forecast and multiplied by 100% to obtain overrun percentages. Note that for the calculation of the figures in the table all projects were used. That is, projects with both overrun and underrun were considered. A value of 100% means the actual is twice the forecast. A value of −50% means the actual is half the forecast. We also note that using this definition, the average value is not very insightful as there is a bias toward overruns. Overruns take on values anywhere from 0% to infinity. However, underruns only take values between

Table 17
Overrun figures reported in the literature and our case studies taking into account all projects. Organization Z was excluded as in this case study we analyzed approvals instead of actuals. These figures should not be used for benchmarking, since the biases in the data sets are unknown.

Source	Year	Median	Average	Amount	Unit
[4] Augustine	1979	–	33%	100 projects	Time
[32] Jenkins	1984	33.5%	66.7%	72 projects	Cost
[5] Topping	1985	26%	40%	22 projects	Effort
[57] Phan	1988	–	33%	191 respondents	Cost
[5] Bergeron	1992	–	33%	89 projects	Effort
Landmark Graphics	2002	79%	105%	121 projects	Time
[63] Sauer	2002	–	13%	412 respondents	Cost
[63] Sauer	2002	–	20%	412 respondents	Time
[63] Sauer	2002	–	−7%	412 respondents	Functionality
Organization Y functionality	2005	0%	10%	140 projects	Functionality
Organization X	2006	−38%	286%	867 projects	Cost
Organization Y cost	2006	−2%	16%	140 projects	Cost

Table 18
Overrun figures reported in the literature and our case studies taking only projects into account with overrun. These figures should not be used for benchmarking, since the biases in the data sets are unknown.

Source	Year	Median	Average	Number of data	Unit
[29] Standish	1994	–	189%	365 respondents	Cost
[29] Standish	1994	–	222%	365 respondents	Time
[33] Johnson/Standish	1996	–	142%	–	Cost
[33] Johnson/Standish	1996	–	131%	–	Time
[31] Standish	1998	–	69%	–	Cost
[33] Johnson/Standish	1998	–	79%	–	Time
[31] Standish	2000	–	45%	–	Cost
[31] Standish	2000	–	63%	–	Time
[33] Johnson/Standish	2002	–	43%	–	Cost
[33] Johnson/Standish	2002	–	82%	–	Time
Landmark Graphics	2002	85%	113%	114 projects	Time
[33] Johnson/Standish	2004	–	56%	–	Cost
[33] Johnson/Standish	2004	–	84%	–	Time
Organization Y functionality	2005	−17%	−22%	37 projects	Functionality
Organization X	2006	173%	1001%	284 projects	Cost
Organization Y cost	2006	20%	67%	57 projects	Cost

−100% and 0%. Therefore, averaging these values is meaningless. The median value is less sensitive to this phenomenon and must be used if no information on the underlying data is available.

The figures in Table 17 are not only found in IT. Similar overruns are reported in other industries, for instance transportation [22]. However, the table does not contain figures reported on by Standish [29,27,30,31,33]. Namely, these figures were not calculated in the same way as done by the other authors. From the first Chaos report [29], we learned that Standish analyzed overruns of combined challenged and impaired projects. Recall our discussion on these definitions from the previous section. We found that the definitions only account for overruns. Following the definitions given by Standish, all challenged and impaired projects have overruns for time and cost and have less functionality. Thus, projects with underruns are not taken into account, as is done in other articles. Basically, Standish reports the extent of the overrun when an overrun occurs. Therefore, it is incorrect, as is done in an article by Jørgensen [37], to compare these figures of Standish with those of other authors.

In subsequent Standish publications [30,31,33], it is not defined for which projects the overruns are calculated. However, assuming Standish computed the figures consistently, we assume that these only take into account projects with overrun as well. In Table 18, we give an overview of these Standish figures together with figures from our case studies derived for only projects with overrun.

However, the validity of some of the figures is disputed. For instance, the figures reported by Standish have been under debate lately. Glass [24,25] and Jørgensen [37] state that the figures reported by Standish are inconsistent with other reported figures in the literature. Zvegintzov [77] places low reliability on information, where the actual data and data sources are kept hidden. They argue that Standish does not clearly describe which projects were investigated and how they calculated their results. Therefore, they feel the figures of Standish are unreliable and must not be used. In our other article [18], we showed that the Standish figures are indeed meaningless.

The reason that the various benchmark figures display large variations is that none of these figures account for the biases present. More precisely, it is not quantified in any of the cases we encountered in the literature what the political undercurrent or quality of the forecasting is. However, many of the authors felt that there was something troublesome with their findings, since many instigations to such doubts were reported.

For instance, Boehm [6] noted on Augustine's results that high forecasts lead to confronting situations which people want to avoid by giving lower forecasts. Bergeron [5] stated that estimators are aware that lower forecasts have no immediate consequence and that additional budget will often be made available later. Phan [57] found that overly optimistic forecasts are one of the main causes for the project delays and cost overruns. Johnson [27] described that Standish takes into account what he calls sand bagging: overstating project budget to avoid failure. But he does not explain how.

In this respect, what do the overrun figures in Table 17 and Table 18 represent? Clearly, they are not an indication of how good or bad the projects are managed, but an indication of the politics involved in forecasting. For instance, of organization X, the figures represent the result of steering on Standish indicators.

In the previous subsection, we have shown how to quantify the politics and in combination with an EQF, create more insightful benchmarks than the current average overrun figures. We suggest to use our benchmarking methodology for comparing organizations.

Summary. We agree with Glass, Jørgenson and Zvegintzov that the overrun figures of Standish are meaningless, which we elaborately argued elsewhere [18]. But for that matter, so are the other benchmarks reported in the literature. Of all the published benchmarks that we are aware of, it is unclear what the political nature of the forecasts comprises. Therefore, their overrun figures give an unclear picture of the true situation of IT projects. The low overrun figures reported by Augustine, Phan, Bergeron and others may well be caused by sand bagging. And the Standish figures could also be explained by overstating budgets or deadlines (as we actually observed). Without information on the politics of forecasting, it is hard to draw any conclusions on the published benchmarks.

8. Practitioners guide

In this article, we have extensively discussed the use and limitations of the tools necessary to quantify IT forecast quality. These tools, the f/a plot and the EQF, are based on the methods developed 25 years ago by Boehm [6] and DeMarco [14]. Over the years, many authors have referred to and used these works as a small subpart of their research. In this article, we have seen this has not always been done correctly. Therefore, educators and practitioners requested a short summary of our results so that they are able to use this in their textbooks and practice. In this section, we provide them with an overview of the main contributions of this article. We will also provide guidelines that will allow organizations to collect the necessary data and develop the tools to adequately quantify IT forecast quality. We show what kind of information an IT governor is able to obtain by following the approach proposed in this article.

8.1. Lessons learned

The main findings of this article are characterized by the following list.

- When the ex-post part is used and no pathological estimation method is used, forecast accuracy improves over time. Therefore, it is useful to monitor the progress of projects by periodically making new forecasts. However, it does not imply that the ex-ante accuracy improves over time as well. This can be achieved by using different estimation methods at different times during the project. The tools described in the article allow for comparisons to be made for such estimation methods.
- It is possible to forecast consistently better than the initial figures reported by Boehm.
- The EQF quantifies IT forecast quality. This allows for comparisons of estimation methods, IT portfolios, IT projects and benchmarking with other organizations. IT governors are able to use this quantified information to support decision making.
- Biases, political and others, become transparent when analyzing an f/a plot. This allows IT governors to either take actions to remove the biases or adjust the forecasts based on the biases found. In either case, with the f/a plot, executives are able to account for the deviations in the decision-making process.
- In our case studies, we found one organization with an independent forecasting department, resulting in a good quality of forecasting when compared with the other case studies. This supports DeMarco's statement, that such a department will improve forecasting quality.
- The confidence limits proposed by McConnell for the confidence interval are a good idea, although their benchmark figures are highly situational due to their dependence on politics.
- The distribution of f/a ratios provides valuable insight to IT governors. Although we found the distribution not to be lognormal or some other theoretical distribution, the cumulative empirical distribution is a valuable alternative when enough data are available. With it, executives are able to make risk/return analyses when deciding on targets and commitments.
- The project success benchmarks of Standish Group and the overrun benchmarks reported by others are meaningless, as none account for the effect of politics.
- It is possible to account for politics in forecasting benchmarks. We gave a new definition for plan accuracy that assumes a given political situation as a yardstick. This allows for true comparisons to be made between organizations with different political biases.

8.2. How to

In this subsection, we want to provide practitioners with guidelines how to quantify IT forecast quality. We will explain what data to be collected and how to analyze them. We also explain how the information from the analyses assists IT governors in their decision making.

Forecast quality check. The first steps to monitor and check the IT forecast quality are as follows.

1. Start collecting data. In order to perform the analyses described in this article, the following data must be recorded for each newly made forecast:the start date of the project (s), the end date of the project (e), the date the forecast is made (t), the value of the forecast (f) and the actual (a). Both a and e are only known when the project is finished and must be linked to the forecast, when available.
2. Compute the f/a ratios and check them for heterogeneity. The ratios are computed by dividing f by a. It is possible that these ratios consist of varying subgroups that consist of significantly different f/a ratios. This can, for instance, be caused by different estimation methods, different project portfolios or different types of projects. If such subgroups exist, the following steps should be performed for each subgroup seperately.
3. Plot the f/a ratios. With the data collected in the previous step, it is possible to compute the f/a ratios and plot them in an f/a plot. The horizontal axis of the f/a plot is the percentage of completion of the project and has a range of $[0, 1]$. The vertical axis of the f/a plot is the value of the f/a ratio depicted on a logarithmic scale. The data points to be plotted are computed as follows: The percentage of completion is found by computing for each forecast $(t - s)/(e - s)$. The corresponding f/a ratio of the project is found by dividing f by a. With this information, each forecast can be plotted in the f/a plot.
4. Compute the EQF values of the projects. For each project, collect all forecasts made for that single project and use them to compute the EQF value using Formula (1) described in Section 4.1.
5. Draw a box plot of the EQF values. When the EQF values for all projects have been computed, it is possible to draw a box plot of these values. This can be done with most statistical packages.
6. Draw a reference cone. Consider to which cone conditions the forecasts ideally adhere to and determine the quality that they should have. We suggest using Formulas (4) and (5) described in Section 4.2.1. These formulas assume that no bias is present and the estimation accuracy of the ex-ante part remains constant. An EQF value must be chosen in order to draw the reference lines in the f/a plot. Also, Formulas (2) and (3) are usable if the quality is easier to express in values c_1 and c_2.

   The reference lines can be used for two possible purposes. First, the lines can be used to assess the shape of the f/a ratios. In this case, we advise to set the EQF value to be the 20% quantile of all EQF values in order to recognize potential biases in the figure. Second, the lines can be used to compare with the quality of the f/a ratios. In this case, we advise to set the EQF value to an adequate quality that is acceptable. To get an idea on what values are attainable in real-world cases, we refer to Table 12 in Section 7.

With these steps, an organization is able to assess the quality of IT forecasting. With the tools, biases are easily detected, allowing for decisions to be made to either remove or work with the biases present. The quality of the forecasts is also quantified, making it possible to compare the quality with other organizations, portfolios and projects. Also, it enables assessing whether improvements made in the forecasting process were successful. Finally, the above steps allow for auditing the forecasting quality of organizations.

Enhance forecast information. The following steps provide for more advanced quantified information to assist in the decision-making process of new projects. To perform these steps, we assume that the previous steps have already been performed. In addition, we advise the following steps.

7. Perform basic calculations as described in Section 6. The analyses described above, enable enhancing forecast information using these basic calculations. Suppose that a decision needs to be made whether to perform a project with a forecasted cost of 2 million Euro. Assume the analyses showed the forecasts resemble an optimistic pattern, that is the forecasts are in general smaller than the actual. Moreover, the quality of the forecasts in terms of EQF is 5. Using formula (4) of Section 4.2.1, we find that the initial f/a ratio corresponding to that EQF value is $l(0) = 0.6$. This indicates that there is roughly a 50% chance that the project will cost no more than $2 \cdot 1/0.6 = 3.3$ million Euro.
8. Derive the confidence interval ranges based on collected data. To do this, first combine the f/a ratios in groups. How to form these groups differs per organization and depends on when decisions need to be made or forecasts are made. If one does not have a clear idea on how to form the groups, we suggest to use those we created for organization Y in Table 5 in Section 6.2.

   In each of the groups, determine the lower bound L by dividing 1 by the 90% quantile of the f/a ratios, as also explained in Section 6.2. Compute the upper bound U by dividing 1 by the 10% quantile. These bounds provide for a confidence interval with a confidence level of 80% and give executives additional information. For instance, a forecast of 2 million Euro is made for the cost of a project. Suppose the confidence interval $[L, U]$ for this forecast is $[0.5, 1.75]$. This interval provides the IT governor with the information that the project will, with high probability, cost between 1 million and 3.5 million Euro.

9. Derive the empirical distribution of the forecasts. We propose to depict the empirical distribution by plotting the cumulative density function as was done for our case studies in Fig. 16 in Section 6.3. This figure provides more extensive information than the confidence intervals discussed above. For example, take the cumulative density function of Landmark Graphics. In the cumulative density function of the organization, we find the confidence levels by looking at the 10% and 90% quantiles. Indeed, for Landmark Graphics the lower bound is $1/0.9 = 1.11$ and the upper bound is $1/0.3 = 3.33$, just as we computed in Table 3. But we are also able to consider other quantiles.

   With the empirical distribution, we are able to make interesting considerations. Consider again the cumulative density function of Landmark Graphics. Suppose an initial forecast is made for the cost of a project to be 2 million Euro. From the cumulative density function we find that with a 95% chance the f/a ratio is smaller or equal to 1. That is, in 95% of the cases the actual value will be higher than the forecast of 2 million Euro. Also, the function indicates that with a 15% chance the f/a ratio will be smaller or equal to 0.3. This means that with a 15% chance the actual will be higher than or equal to $2/0.3 = 6.7$ million Euro. Or equivalently, we are able to say that with a 85% chance the actual will be lower than 6.7 million Euro. The empirical distribution thus allows for risk/return considerations.

The above steps require (minimal) historical data to be available of the f/a ratios. If no such data are available, we advise to gauge the bias and the quality of one's forecasts in terms of EQF. To make an adequate assumption of the quality, consider the EQF values described in Section 7 to obtain an idea. With this information, one is able to perform the basic calculations suggested in the enumeration above.

Also, we do not advise applying confidence intervals or distributions of other organizations to one's own organization. Using these methods of other organizations does provide additional information, however the applicability to one's own organization is unknown. It can give a false sense of accuracy, allowing for advanced but senseless calculations. Before considering either of these methods, the first priority is to collect one's own data.

9. Conclusions

The quality of IT forecasting is crucial for decision making up to the executive level. The forecasts support go/kill decisions for projects and are used to monitor progress. In this article, we showed how to quantify the quality and potential bias of IT forecasting to improve decision making using the forecasts. We elaborately discussed our approach for this purpose: the EQF and an f/a plot with a reference cone of a certain quality derived from the desired EQF.

The well-known and well-established cone of uncertainty of Barry Boehm is viewed differently by many people. In order to confirm or refute different views of various authors on this subject, we made a distinction between the components of a forecast. We argued that the forecasts consist of two components, which we named the ex-post and the ex-ante part. The ex-post part is the part of the total that has been done already. The ex-ante part is the remainder of the work that still needs to be done. With this distinction, we were able to construct simulations that reproduced conical shapes providing insight in the cone of uncertainty. For instance, the simulations illustrated that the conical shape is not derived by improved estimation methods, but is also found when the estimation accuracy of the ex-ante part decreases.

We also illustrated that f/a ratios plotted against a reference cone visualizes bias, for instance political, involved in IT forecasting. Therefore, our pictures provide crucial information for IT governors about the political undercurrent of the forecasts. The EQF quantifies the deviation of forecasts to an actual. This metric allows adequate comparisons to be made between the quality of forecast between different organizations. Our approach provides necessary information to analyze, quantify and monitor the state of IT forecasting in an organization and between organizations.

We illustrated this by analyzing four case studies that in total consisted of 1824 projects with an investment value of 1059+ million Euro that contained 12 287 forecasts. We applied our approach to each case study to assess their IT forecasting practice in a quantitative manner. In the case studies, we found one organization that had good quality and no political bias in their forecasting. Another organization had reasonable quality, but they forecast the minimum value instead of the actual value. Therefore, the forecasts were almost always lower than the actual. The third organization had low quality of forecasting and a large political undercurrent. The forecasts in this organization hardly resembled any relation to the actual value. The last organization had no political bias and low quality of forecasting as that was not given particular attention.

With the information of the analyses, we showed it is possible to enrich forecast information for decision making. We discussed three methods that provide for additional information to assess the uncertainty of newly made forecasts. If sufficient data are available, the methods will allow for risk/return analyses of new project proposals accounting for the uncertainty of forecasts.

Lastly, we discussed a number of benchmarks related to forecasting that we found in the literature. We surveyed the EQF benchmarks found in the literature. We argued that these values are not always useful and added our own values derived from our case studies as new benchmarks. We also showed that the political bias in forecasting has a large influence on some of the overrun benchmarks. For instance, the famous project success figures reported by Standish Group are highly susceptible to the politics involved with the organizations they analyzed. Therefore, these figures are meaningless without further information on the bias of the forecasts. The same applies to the figures reported by others.

In order to obtain more useful figures, we proposed alternative definitions. These definitions take into account the effect of possible biases in IT forecasts. We hope Standish will adopt our alternative definitions, to reassess the earlier reported

figures and publish new findings correcting for the potential bias in their data. We have applied the approach to our own case studies. Our definitions are suitable for true comparisons on IT forecasting quality between organizations.

Finally, we believe that this article is a much needed addition to assess and benchmark IT forecasting, since proper IT forecast quality is indispensable for proper IT governance.

Acknowledgments

This research received partial support by the Dutch *Joint Academic and Commercial Quality Research & Development (Jacquard)* program on Software Engineering Research under contract 638.004.405 *EQUITY: Exploring Quantifiable Information Technology Yields* and under contract 638.003.611 *Symbiosis: Synergy of managing business-IT-alignment, IT-sourcing and off shoring success in society*. Furthermore, we like to thank a number of organizations that will remain anonymous for kindly sharing their forecasting data with us. Also, we are grateful to our colleague Rob Peters for meticulously reviewing and commenting this article numerous times. Moreover, we thank Steve McConnell, George Tillmann, Magne Jørgensen, Brad Appleton, Nicholas Zvegintzov, our colleague Sandjai Bhulai and the anonymous reviewers for commenting this article. Finally, we are grateful to Todd Little of Landmark Graphics Corporation for his comments and providing one of the data sets that is used throughout this article.

References

[1] Hal R. Arkes, Overconfidence in judgmental forecasting, in: J.S. Armstrong (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners, Springer, 2001, pp. 495–515.
[2] Phillip G. Armour, The inaccurate conception, Communications of the ACM 51 (3) (2008) 13–16.
[3] J.S. Armstrong, The forecasting dictionary, in: J.S. Armstrong (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners, Springer, 2001, pp. 761–819.
[4] N.R. Augustine, Augustine's laws and major system development programs, Defense Systems Management Review 2 (1979) 50–76.
[5] François Bergeron, Jean-Yves St-Arnaud, Estimation of information systems development efforts, Information & Management 22 (1992) 239–254.
[6] Barry Boehm, Software Engineering Economics, Prentice Hall PTR, 1981.
[7] Barry Boehm, Making a difference in the software century, Computer 41 (3) (2008) 32–38.
[8] Barry Boehm, Bradford Clark, Ellis Horowitz, Chris Westland, Cost models for future software life cycle processes: COCOMO 2.0, Annals of Software Engineering 1 (1995).
[9] Lionel Briand, Khaled El Emam, Dagmar Surmann, Isabella Wieczorek, Katrina D. Maxwell, An assessment and comparison of common software cost estimation modeling techniques, in: Proceedings of the 21st International Conference on Software Engineering, 1999, pp. 313–322.
[10] Murray Cantor, Estimation variance and governance, 2006. www-128.ibm.com/developerworks/rational/library/mar06/cantor/.
[11] Chris Chatfield, Prediction intervals for time-series forecasting, in: J.S. Armstrong (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners, Springer, 2001, pp. 475–494.
[12] Mike Cohn, Agile Estimating and Planning, Prentice Hall, 2005.
[13] S.D. Conte, H.E. Dunsmore, V.Y. Shen, Software Engineering Metrics and Models, The Benjamin/Cummings Publishing Company, 1986.
[14] Tom DeMarco, Controlling Software Projects, Prentice Hall PTR, 1982.
[15] Tom DeMarco, Timothy Lister, Waltzing with Bears, Dorset House Publishing, 2003.
[16] J.B. Dreger, Function Point Analysis, Prentice Hall, 1989.
[17] J.L. Eveleens, P. Kampstra, C. Verhoef, Quantifying changes in IT metrics after outsourcing transitions. Under review. Available via www.cs.vu.nl/~x/sps/sps.pdf.
[18] J.L. Eveleens, C. Verhoef, The rise and fall of the Chaos report figures, IEEE Software, in press (doi:10.1109/MS.2009.154). Available via www.cs.vu.nl/~x/chaos/chaos.pdf.
[19] Dick Fairley, Making accurate estimates, IEEE Software 19 (6) (2002) 61–63.
[20] W. Feller, On the Kolmogorov-Smirnov limit theorems for empirical distributions, The Annals of Mathematical Statistics 19 (2) (1948) 177–189.
[21] Marek Fisz, Probability Theory and Mathematical Statistics, John Wiley & Sons, 1963.
[22] Bent Flyvbjerg, Nils Bruzelius, Werner Rothengatter, Megaprojects and Risk: An Anatomy of Ambition, Cambridge University Press, 2003.
[23] D. Garmus, D. Herron, Function Point Analysis – Measurement Practices for Successful Software Projects, Addison-Wesley, 2001.
[24] Robert Glass, IT failure rates—70% or 10–15%, IEEE Software (May) (2005) 110–112.
[25] Robert Glass, The Standish report: Does it really describe a software crisis, Communications of the ACM 49 (8) (2006).
[26] Stephan Gryphon, Philippe Kruchten, Steve McConnell, Todd Little, The Cone of Uncertainty, IEEE Software 23 (5) (2006) 8–10.
[27] Deborah Hartmann, Interview: Jim Johnson of the Standish Group, 2006. http://www.infoq.com/articles/Interview-Johnson-Standish-CHAOS.
[28] Nigel Harvey, Improving judgment in forecasting, in: J.S. Armstrong (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners, Springer, 2001, pp. 59–80.
[29] The Standish Group International Inc. Chaos, Technical Report, The Standish Group International Inc., 1994.
[30] The Standish Group International Inc. Chaos: A recipe for success, Technical Report, The Standish Group International Inc., 1999.
[31] The Standish Group International Inc. Extreme chaos, Technical Report, The Standish Group International Inc., 2001.
[32] Milton Jenkins, Justus Naumann, James Wetherbe, Empirical investigation of systems development practices and results, Information & Management 7 (1984) 73–82.
[33] Jim Johnson, Standish: Why were project failures up and cost overruns down in 1998, 2006. http://www.infoq.com/articles/chaos-1998-failure-stats.
[34] Magne Jørgensen, Experience with the accuracy of software maintenance task effort prediction models, IEEE Transactions on Software Engineering 21 (1995) 674–681.
[35] Magne Jørgensen, A review of studies on expert estimation of software development effort, Journal of Systems and Software 70 (2004) 37–60.
[36] Magne Jørgensen, Practical guidelines for expert-judgment-based software effort estimation, IEEE Software 22 (3) (2005) 57–63.
[37] Magne Jørgensen, Kjetil Moløkken, How large are software cost overruns? A Review of the 1994 Chaos report, Information and Software Technology 48 (2006) 297–301.
[38] K. Kavoussanakis, Terry Sloan, UKHEC Report on software estimation, Technical Report, University of Edinburgh, 2001.
[39] Chris Kemerer, An empirical validation of software cost estimation models, Communications of the ACM 30 (5) (1987) 416–429.
[40] Barbara Kitchenham, Stephan MacDonell, Lesley Pickard, Martin Shepperd, What accuracy statistics really measure, IEE Proceedings — Software 148 (2001) 81–85.
[41] G.P. Kulk, C. Verhoef, Quantifying requirements volatility effects, Science of Computer Programming 72 (2008) 136–175.
[42] G.P. Kulk, C. Verhoef, Quantifying IT estimation risks (in press) Available via http://www.cs.vu.nl/~x/qier/qier.pdf, 2009.
[43] Linda M. Laird, The limitations of estimation, IT Professional 8 (6) (2006) 40–45.

[44] Linda M. Laird, M. Carol Brennan, Software Measurement and Estimation: A Practical Approach, John Wiley & Sons, 2006.
[45] Luiz A. Laranjeira, Software size estimation of object-oriented systems, IEEE Transactions on Software Engineering 16 (5) (1990) 510–522.
[46] Albert L. Lederer, Rajesh Mirani, Boon Siong Neo, Carol Pollard, Jayesh Prasad, K. Ramamurthy, Information system cost estimating: A management perspective, MIS Quarterly 14 (2) (1990) 159–176.
[47] Tim Lister, Becoming a better estimator — An Introduction to using the EQF Metric, 2002. Available via http://www.stickyminds.com.
[48] Todd Little, Context-adaptive agility: Managing complexity and uncertainty, IEEE Software 22 (3) (2005) 28–35.
[49] Todd Little, Schedule estimation and uncertainty surrounding the cone of uncertainty, IEEE Software 23 (3) (2006) 48–54.
[50] Karen Lum, Michael Bramble, Jairus Hihn, John Hackney, Mori Khorrami, Erik Monson, Handbook for software cost estimation, Technical Report, Jet Propulsion Laboratory, 2003.
[51] Maplesoft. http://www.maplesoft.com/Products/Maple/.
[52] Steve McConnell, Rapid Development: Taming wild Software Schedules, Microsoft Press, 1996.
[53] Steve McConnell, Software Estimation: Demystifying the black art, Microsoft Press, 2006.
[54] M. Shahid Mujtaba, Robert Ritter, Enterprise modeling system: Inventory exposure and delivery performance, Technical Report, Hewlett-Packard company, 1994.
[55] Ingunn Myrtveit, Erik Stensrud, A controlled experiment to assess the benefits of estimating with analogy and regression models, IEEE Transactions on Software Engineering 25 (1999) 510–525.
[56] Carlo Pescio, Realistic and useful: Toward better estimates, 2007. Available via http://www.eptacom.net/betterestimate/RealisticAndUseful.pdf.
[57] Dien Phan, Douglas Vogel, Jay Nunamaker, The search for perfect project management, Computerworld 26 (1988) 95–100.
[58] L.H. Putman, W. Myers, Measures for Excellence – Reliable Software on Time, Within Budget, Yourdon Press Computing Series, 1992.
[59] Lawrence H. Putnam, Ann Fitzsimmons, Estimating software costs, Datamation (September–October–November) (1979).
[60] L.H. Putnam, A macro-estimation methodology for software development, in: Proceedings IEEE COMPCON 76 Fall, IEEE Computer Society Press, 1976, pp. 138–143.
[61] L.H. Putnam, A general empirical solution to the macro software sizing and estimating problem, IEEE Transactions of Software Engineering 4 (4) (1978) 345–381.
[62] R-project. http://www.r-project.org/.
[63] Chris Sauer, Andrew Gemino, Blaize Horner Reich, The impact of size and volatility on IT project performance, Communications of the ACM 50 (11) (2007) 79–84.
[64] C.C. Shelley, Practical experience of implementing software measurement programmes in industry, Transactions on Information and Communications Technologies 4 (1993).
[65] Martin Shepperd, Chris Schofield, Estimating software project effort using analogies, IEEE Transactions of Software Engineering 23 (11) (1997) 736–743.
[66] Michael Smithson, Confidence intervals, Sage University Papers Series on Quantitative Applications in Social Sciences, vol. 07-140, Thousand Oaks, 2003.
[67] Ian Sommerville, Software Engineering, 7th edition, Pearson Educcation Limited, 2004.
[68] Erik Stensrud, Tron Foss, Barbara Kitchenham, Ingunn Myrtveit, A further empirical investigation of the relationship between MRE and project size, Empirical Software Engineering 8 (2003) 139–161.
[69] Steve Tockey, Return on Software, Addison-Wesley, 2005.
[70] Oxford University Press. Concise oxford english dictionary Oxford University Press, 1999.
[71] Hans van Vliet, Software Engineering: Principles and Practice, 3rd edition, John Wiley & Sons, 2008.
[72] Chris Verhoef, Quantifying the effects of IT-governance rules, Science of Computer Programming 67 (2–3) (2007) 247–277. Available via http://www.cs.vu.nl/~x/gov/gov.pdf.
[73] Frank Vogelezang, Scope management—how uncertainty is your certainty, in: International Workshop on Software Measurement, 2007.
[74] Gerald M. Weinberg, Edward L. Schulman, Goals and performance in computer programming, Human Factors 16 (1) (1974) 70–77.
[75] Da Yang, Barry Boehm, Ye Yang, Qing Wang, Mingshu Li, Coping with the Cone of Uncertainty: An Empirical Study of the SAIV Process Model, in: Lecture Notes in Computer Science, vol. 4470/2007, Springer, Berlin, Heidelberg, 2007.
[76] Yourdon. The ten most important ideas in software engineering, 2006. http://www.yourdonreport.com/index.php/2006/10/17/the-ten-most-important-ideas-in-software-engineering/.
[77] Nicholas Zvegintzov, Frequently begged questions and how to answer them, IEEE Software 20 (2) (1998) 93–96.