IT Forecast Quality in Measure and Number

J. Laurenz Eveleens

Nederlandse Organisatie voor Wetenschappelijk Onderzoek

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under the Jacquard project number:

638.004.405 *EQUITY: Exploring Quantifiable Information Technology Yields*

"exploring quantifiable IT yields"

This thesis was sponsored by PwC.

Cover illustrated by Annemieke van Dongen.

ISBN: 978-94-6191-050-9
© Laurenz Eveleens 2011

VRIJE UNIVERSITEIT

IT Forecast Quality in Measure and Number

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op 13 december 2011 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Johan Laurenz Eveleens

geboren te Amsterdam

promotor:	prof.dr. C. Verhoef
copromotor:	dr. R.J. Peters

Dankwoord

Het schrijven van dit proefschrift is een bijzondere en leerzame ervaring geweest. In het bijzonder wil ik twee personen bedanken, zonder wiens bijdrage dit proefschrift niet gemaakt had kunnen worden.

In de eerste plaats is dat Chris Verhoef, die mij de kans heeft gegeven dit promotieonderzoek te doen en die mij met zijn hulp, toewijding en voortdurende inspanning gedurende het traject heeft geholpen. Hij was altijd bereikbaar en maakte tijd om mijn werk door te nemen en het werk te verbeteren. Hij heeft mij geleerd onderzoek te doen, artikelen te schrijven en commentaar te verwerken. Ik ben blij Chris te hebben leren kennen en met hem te hebben kunnen samenwerken.

In de tweede plaats wil ik Rob Peters bedanken. Met zijn ervaring en achtergrond, heeft Rob een zeer belangrijke bijdrage geleverd aan het onderzoek. Hij heeft veelvuldig de tijd genomen mijn werk grondig en aandachtig te lezen. In onze talloze discussies heeft hij het mij mogelijk gemaakt het onderzoek steeds te verbeteren en aan te scherpen. Ook op het persoonlijke vlak heb ik goed contact gehad met Rob.

Naast Chris en Rob, wil ik Peter Kampstra, Erald Kulk, Łukasz Kwiatkowski en de andere collega's van de VU bedanken voor de samenwerking en gezellige tijd. Ook bedank ik de organisaties die meegewerkt hebben aan dit onderzoek door hun vertrouwelijke data met mij te delen. Zonder de bijdrage van deze organisaties, was dit proefschrift niet mogelijk geweest. In het bijzonder, wil ik Mark van der Pas bedanken voor het meedenken en meeschrijven aan een artikel dat onderdeel vormt van dit proefschrift. Daarnaast ben ik de leescommissie, bestaande uit Hakan Erdogmus, Wan Fokkink, Bert Kersten en Philippe Kruchten, dankbaar voor het lezen en beoordelen van dit onderzoek.

Tot slot, wil ik mijn vrouw, familie en vrienden bedanken. Mijn ouders hebben mij in staat gesteld mijn studie te volgen, te kunnen afronden en gestimuleerd de uitdaging van het promotieonderzoek aan te gaan. In de persoonlijk moeilijke omstandigheden tijdens mijn promotie, hebben mijn ouders, vrouw, familie en vrienden mij gesteund en gesterkt. Dit heeft het voor mij mogelijk gemaakt dit proefschrift, ondanks alles, tot een goed einde te brengen. Hiervoor ben ik hen allen enorm veel dank verschuldigd.

Laurenz Eveleens

Contents

CHAPTER 1

Introduction

1.1 Setting expectations

Consider a man who is a big fan of cycling. On his holiday, he decides to cycle a stage, as shown in Figure 1.1, of a famous cycling event. Before starting his trip, he makes a forecast of the time he will require to complete the 200 kilometer long stage. Taking into account the terrain of the course, the weather conditions and potential traffic delays, he predicts the entire stage will take him approximately 10 hours. After a long day of cycling the trip turns out to take him 10.5 hours.

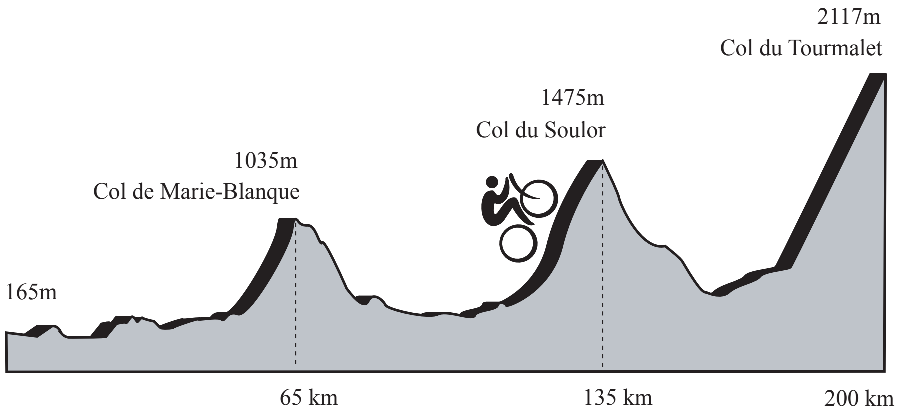


Figure 1.1: Example stage of a famous cycling event.

Invigourated by the journey, the next day he plans to cycle the exact same stage again. However, his partner just arrived and wants to know when he will be back. She intends them to have diner together later that day. The man expects that, although it normally would take about 10 hours, if everything goes well, it should be possible within 8 hours. Therefore, not wanting to disappoint his partner, he says he expects to be back in 8 hours. In the end, the stage still takes him 10 hours cycling.

On the third day, the man wants to cycle the stage one last time. However, the day before, his partner was not pleased with him arriving 2 hours later than he had said. He now knows that it is likely the trip will take him about 10 hours. Yet, he wants to avoid being late again. Therefore, he tells his partner it will take him 12 hours before coming home. This way, if the trip does take longer he will still be home on time. And if he is back earlier, his partner will only be happy about it. In the end, he arrives after 9.5 hours.

In the above example, on the first day the prediction of the required time was based on a computation of different factors that are of influence on the required time. In the example, these factors were the length and terrain of the stage, the weather conditions and traffic. However, the other two days illustrate that expectations not only depend on quantitative factors. Although it is possible to properly assess the required time based on the influencing factors and previous experiences, the man adjusts his prediction to suit his own desire to go cycling and the wishes of his partner, making his forecasts biased. Not only in every day life forecasts can be biased, but also in a business environment forecasts are sometimes biased.

1.2 Expectations in IT

Organizations need to undertake IT projects every year to consolidate and expand their business. These IT projects often play a profound role in modern companies due to their size and impact. 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.

IT executives have the task to identify those projects that will be most valuable to the organization. This decision is based on both qualitative and quantitative information. An example of qualitative information is the strategic alignment of a project. Does the project help in achieving the organizational goals? The quantitative information concerns costs and benefits over time. The questions are: will the project deliver value to the organization? Are the costs acceptable with respect to the benefits the asset will yield over time? Is the expected functionality in line with the costs needed to create the asset? These questions are the basis for any decision making and rely heavily on the forecasts of the underlying Key Performance Indicators. Therefore, their forecast quality is crucial. The forecasts also serve as budget and time constraints in many occasions.

Frequently, IT projects are discussed in the news because they fail to meet the expectations that were initially set. For example, in 2004 the Protestant Church Netherlands or PKN, decided to implement a new web-based member registration system. This system was expected to be operational in 2007 and would require an investment of 1.8 million Euro [38, 46]. After many problems with the implementation, in June 2008, PKN decided to cancel the project after a delay of 1.5 years and having spent 5.5 million Euro [107].

Another example is the implementation of a system to assist in the execution of the "Wet werk en inkomen naar arbeidvermogen"[1] or WIA legislation in the Netherlands. The WIA legislation provides government support for employees who become (partially) incapable to perform work. In 2005, the WIA-project was estimated to cost 76 million Euro and to finish by January 1st, 2008 [120, 129]. In June 2008, a half year after the intended delivery date, the project was cancelled having cost 87 million Euro and with hardly any reusable result [118].

Unfortunately, many more of such examples can be discussed. In these examples, projects either spend more, take longer, deliver less functionality than forecasted or do not deliver anything at all. In the examples, the initial expectations were considered valid and accurate. It is assumed that the project could have finished within the time and budget that was given. However, one can wonder how accurate these forecasts actually were. Namely, as we illustrated in the example of the cyclist, these forecasts may also have been influenced by biases. This begs the question: how reliable are forecasts? How does a manager know whether the quality of the forecasts is not just influenced by the forecast model chosen, but also by biases?

These questions stress the importance of adequate management by the Chief Information Officer (CIO). More and more, voices are raised [36, 92] that the CIO must manage IT projects to maximize their business value and return, instead of controlling their cost. For instance, in 1998 a research project known as Beyond Budgeting was started to change the current management model [44]. Where previously the management was focussed on planning and control of cost and the technical aspect of IT, Beyond Budgeting emphasizes the importance of the business value.

Pisello et al. [98] argue that the CIO needs to become the Chief Financial Officer (CFO) of IT to improve the organization's value. But, what does managing the business value of IT entail? An article by Lorie et al. [85] states that executives face three tasks in achieving good financial management, among which the correct forecasting of the expected cash flows. In this thesis, we focus on that particular task.

In information technology, the META group [36] showed that forecasting, especially of benefits, is far from common practice. Of the organizations that were surveyed, 84% indicated that no business cases were made for their IT projects or only for a select few projects. And if forecasts of the cash flows are made, the question is how to assess their validity.

That proper forecasting is indeed a challenging task, is shown, for instance, by a survey of the Kellogg School of Management [81]. The survey found that 82% of the responding CIO's regarded forecasting of IT benefits as a major challenge. On top of that, the survey showed that 68% do not track project benefits at all and only

[1] Law of work and income according to labour capacity

26% track actual financial metrics after having made an investment decision. So, how do we know that forecasts of the IT business value are accurate, unbiased and reliable enough to support the decision making process? And are organizations able to adequately predict the business value of IT investments?

1.3 Research questions

The above considerations lead to a number of research questions. How is a manager able to assess the quality of the forecasts made for IT projects? What is the accuracy of the forecasts made for cost, time, functionality and benefits in real-world organizations? Is it possible to compare the forecasting accuracy between organizations? Which benchmarks does the literature provide for IT forecast quality? How is it possible to utilize the knowledge of the forecast quality to enhance decision making?

1.4 Thesis outline and contributions

In this thesis we address the above questions. The main body of the thesis is divided into four chapters. Below we give an overview of these chapters and their contributions.

1.4.1 Terminology

In Chapter 2 we introduce terms and notations that we use throughout the thesis. Although most terms are used frequently in common English, the meaning and understanding of them differs among people. Therefore, we clearly state how we comprehend them.

1.4.2 IT forecast quality and biases

In Chapter 3 we address how to quantify forecast quality of positive valued entities such as project cost, time and functionality. First, two statistics previously proposed to analyze IT forecast data are examined: Boehm's cone of uncertainty [8] and DeMarco's Estimating Quality Factor [20]. We generalize the cone as a family of distributions that predict IT forecasts on the basis of expected accuracy and predictive bias. With these distributions, IT executives are able to compare the forecasts made, with a required accuracy. We illustrate that plotting forecast-to-actual ratios against a reference cone, which visualizes the required accuracy by the IT manager, reveals potential biases involved with IT forecasting. The comparison provides critical information on IT forecast quality to IT executive, allowing them to assess biases and the uncertainty of the forecasts made.

This approach is illustrated by applying it to data of four real-world organizations that in total consist of 1824 projects with a total actual cost of 1059+ million Euro and 12287 forecasts. Moreover, we illustrate how to use the information to enrich forecast information for decision making. We show that IT executives are able to correct for possible biases and use the uncertainty to determine the likelihood of specific scenarios. For instance, what is the likelihood, based on historical forecast information, that a project will cost more than a certain amount?

Finally, we survey benchmarks related to forecasting and will find that none of them reckon with biases.

Contributions of this chapter

- We show that there are theoretical problems with the cone of uncertainty. For example, contrary to popular belief, the conical shape of Boehm's cone is not caused by improved estimation, but can also be found when estimation accuracy decreases.

- We generalize the cone as a family of distributions that predict IT forecasts on the basis of expected accuracy and predictive bias. With these distributions, IT executives are able to compare the forecasts made, with a required accuracy.

- We show that the distribution of forecast-to-actual ratios varies 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.

- Compared with the 24 [10] or 25 [9] forecasts used to support Boehm's cone of uncertainty, the 20 data points used for DeMarco's EQF [20] and the 106 projects analyzed by Little [84] for both topics, our case studies form a sizable addition to the research done in the literature.

- We illustrate how IT executives can use the quantified forecast information to enhance decision making. In Section 3.7 we provide for a practitioner's guide on how to obtain the required data, and how to use the data to obtain additional information.

1.4.3 Forecast quality and Chaos project succes

Chapter 4 discusses in more detail the consequences of forecast quality and their potential biases on often-quoted rates of project success. In particular, the figures published by Standish Group in their Chaos reports [47, 48, 49, 50, 51] are discussed. Contrary to popular belief, we show that the Standish figures are in fact about forecast quality.

In 1994, Standish published the Chaos report that showed a shocking 16% project success. This and renewed figures by Standish are often used to indicate that project management of application software development is in a troublesome state.

We demonstrate that Standish's definitions have major problems by applying their definitions to our extensive data consisting of 5457 forecasts of 1211 real-world projects of in total hundreds of millions of Euros. It turns out that the Standish figures do not reflect the reality of the case studies at all.

To alleviate for the problems, we propose alternative definitions based on our extensive research.

Contributions of this chapter

- We show that Standish's definitions have four major problems.
  - First, they are misleading, since they are solely based on estimation accuracy of cost, time and functionality.
  - Second, their estimation accuracy measure is one-sided leading to unrealistic success rates.
  - Third, steering on Standish's definitions perverts good estimation practice.
  - Fourth, the resulting figures are meaningless because they averaged numbers with an unknown bias, which are introduced by different underlying estimation processes.
- We propose alternative benchmarks that take into account the effect of possible biases in IT forecasts.

1.4.4 IT business value and its quality

Chapter 5 discusses how to quantify the forecast quality of IT business value. We address a common economic indicator often used to determine the business value of project proposals, the Net Present Value (NPV).

The method described in Chapter 3 is extended, by developing a generalized method that is able to account for asymptotic cases and negatively valued entities. We assess the generalization with real-world data of four organizations together consisting of 1435 IT assets with a total investment cost of 1232+ million Euro for which 6328 forecasts were made.

Using the generalized method, the forecast quality of the NPV is determined, along with the benefits and costs, using real-world data of another 102 IT assets with a total business value of 1812 million Euro.

Also, a sensitivity analysis is performed to investigate the impact on the quality of an asset's forecasted NPV when the forecast quality of benefits or costs improves. Finally, we show how to use the quantified forecast information to enhance decision information using two simulation examples.

Contributions of this chapter

- We develop a generalized method to quantify forecast quality that is able to account for asymptotic cases and negatively valued entities.

- For the real-world case study in Chapter 5, we will find that the quality of the forecasted NPV's is lower than that of the forecasted benefits, which is again lower than that of the forecast quality of the costs.

- We will find that the forecast quality of Cost Reduction assets are different from the forecast quality of New Product Development assets. In our case study, the forecasts of Cost Reduction assets are of higher quality.

- Counterintuitively, it turns out in this case study that if the quality of cost forecasts would improve, the overall quality of its NPV predictions would degrade. This is caused by the bias of the benefit forecasts that is negated by the bias of the cost forecasts. Therefore, reducing the bias of the cost forecasts results in less accurate NPV forecasts. This effect underlines the importance of both accurate cost and benefit predictions.

- We demonstrate how IT executives can use the quantified forecast information to enhance decision making with the aid of Monte Carlo simulations.

1.4.5 Summary in dutch

Finally, in Chapter 6 we conclude the thesis with a summary in dutch.

1.5 Case studies

We will make use of extensive real-world data throughout this thesis. We analyze the IT forecast 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 1926 projects with total investment cost of 1232+ million Euro for which 12389 forecasts were made.

Of a large financial service provider, Y, we use data consisting of 667 forecasts of 140 project costs and 100 functionality forecasts of 83 assets. A multinational organization, X, provided us with 3767 forecasts for 867 project costs. From Landmark Graphics, LGC, we obtained data containing 6245 forecasts of 121 project durations. Finally, a large telecommunication organization Z provided 1508 forecasts made for 613 project costs.

For the purpose of evaluating the forecast quality of IT business value in Chapter 5, we analyze another data set from the large telecommunication organization, Z. This data consists of 102 NPV forecasts made for 102 IT assets that together represent discounted benefits of 4714 million Euro and an investment value of 173 million Euro.

1.6 Tools of the trade

In this thesis we chose to use known tools in the IT economics literature to quantify IT forecast quality. These tools consist, among others, of Boehm's cone of uncertainty and DeMarco's Estimating Quality Factor or EQF. The cone of uncertainty is a plot that depicts forecasts divided by actuals. The EQF is a measure that quantifies forecast quality.

Using project data on completed projects of several organizations, these models are extended and applied to the data. In this thesis we make these tools accessible and usable for IT executives, allowing them to gain information about IT forecast quality of their organization in measure and number. Finally, we make extensive use of exploratory data analyses and simulations.

1.7 Related work

IT forecasting has been an important topic for numerous years. In statistical mathematics, assessing the quality of estimation methods is a well-discussed topic [29]. There are well-defined criteria that determine the quality of these methods. The generalized method we will develop in this thesis makes use of such criteria and does not provide new statistical ways to determine forecast quality.

However, the statistical methods and metrics are often not accessible to IT executives. Therefore, in this thesis, we discuss how to present, summarize and visualize the IT forecast accuracy in such a way that executives are able to assess their quality and use it to enhance decision information. With the generalized method we aim to make quantifying the forecast quality more readily available to IT executives. Furthermore, the method allows executives to acquire knowledge on the forecast quality of their organization.

Methods for software estimation IT forecasting methods and their accuracy are frequently discussed in the literature to achieve correct forecasting of project proposals. For example, many books [8, 14, 17, 56, 77, 89] have been written describing issues and guidelines to achieve accurate estimates. Moreover, numerous estimation tools exist and are used in practice, among others COCOMO [8], SLIM [102], SEER, SPQR/20 and KnowledgePlan [57]. These tools assist in forecasting relevant project values, such as cost, effort and durations.

Numerous articles [12, 59, 68, 111] compare different estimation methods to determine which of them are most accurate under certain circumstances. Using tools such as the MRE [19, 29, 45] to quantify forecast quality, the purpose of these authors is to compare software cost estimation methods. However, many of these studies use small data sets and result in inconclusive or difficult to generalize findings about the estimation methods [12]. Moreover, the tools used to quantify the quality of IT forecasts are not subject to scrutiny and merely facilitate a way to make the comparison. Finally, these articles do not address business value forecasts, but forecasts such as cost, size, functionality or duration.

In this thesis, the purpose of the quantification is to assist in governing IT. We illustrate what information IT executives are able to extract using our approaches to improve decision making based on the forecasts. Therefore, this thesis describes more extensively how to quantify IT forecast quality and challenges known results in this area.

Moreover, in this thesis we will make use of a large sample of data regarding 1926 IT projects with 12287 IT forecasts. Next to that, we will discuss and assess the tools used to quantify forecast quality.

Finally, we analyze forecasted and re-estimated NPVs of 102 IT assets to assess the accuracy of the initial forecasts of IT business value. We are unaware of articles that quantify the quality of business value forecasts of IT investments. A book by Bower [79] did quantify the quality of NPV predictions for 50 assets in another industry. We will compare our case study to the one described by Bower.

Tools For comparing software cost estimation methods, authors seldom use the EQF [20]. Only a small number of articles [82, 83, 84, 110] refer to the EQF. We will argue that a popular method, the *magnitude of relative error*, or MRE [19], has a severe drawback when compared with the EQF. Therefore, in this thesis, we elaborate on the EQF and its uses.

Moreover, we will generalize the well-known cone of uncertainty [8]. The cone of uncertainty is cited by many [18, 67, 76, 86, 88, 114, 121]. 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 [83]. Little uses 121 projects of Landmark Graphics to draw the cone of uncertainty and compute EQF values based on the time forecasts made for these projects. The data shows underestimations and convergence to the actual. Little argues that the cone of uncertainty should converge to the actual by definition.

In Chapter 3 we assess the cone of uncertainty as well and extend the analyses done by Little [83], by using his data in one of our case studies. Moreover, we will use additional data sets and generalize the cone as a family of distributions. We will illustrate that the cone of uncertainty converges to the actual under certain circumstances, but does not necessarily need to.

Biases In this thesis, we will frequently discuss biases. We want to make potential biases in forecasts transparent by plotting the deviations from the actual against a reference cone. Although various authors [7, 8, 41, 54, 96, 121] state that the politics of forecasting and other biases are highly influential for forecast quality, none have ventured to visualize their effects on forecasts. In this thesis, we will show how this can be done using our reference cone, which is a generalization of the cone of uncertainty [8].

In many occasions in this thesis we will speak of a bias caused by political reasons. However, the methods we will present are not restricted to this type of bias. Moreover, in this thesis we will not address why biases exist or how they can be avoided. The literature provides for extensive research on these subjects. Below, we provide the interested reader with a short overview of this subject.

Several articles [24, 39, 64, 112, 115] describe cognative biases, such as anchoring, framing and confirming-evidence traps that distort judgement and cause one to consistently and predictable err. Also, many articles [20, 27, 42, 60, 69, 80, 127], discuss biases in the context of software engineering and software cost estimation.

In a book by McConnell [89], 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. McConnell [89, pg.47] illustrated this with an experiment in which 100 groups of estimators applied 17 effort multipliers in Cocomo II to the same estimation problem. The results showed wide variations that were only caused by the biases due to the estimators' experiences.

There are many reasons why biases, intentional or unintentional, occur. For example, Lederer et al. [80] 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 [27]. 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 [20, 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 [42, 60].

Weinberg et al. [127] 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.

Apart from acknowledging that biases exist, several articles also address how biases can be avoided [2, 39, 42, 64].

Literature benchmarks Finally, numerous benchmarks are given in the literature [6, 7, 20, 47, 48, 49, 53, 82, 88, 96] related to the quality of forecasts. 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 thesis, we illustrate how to incorporate the political nature of IT forecasting, which makes true comparisons possible.

1.8 Origin of the chapters

This thesis is based on three published articles. In this thesis, the articles have been rearranged to increase readability. The following overview indicates for each chapter of this thesis on which article it is primarily based.

- **Chapter 3**
  J.L. Eveleens and C. Verhoef. Quantifying IT forecast quality.
  Science of Computer Programming, vol. 74: pag. 934 – 988, 2009.
- **Chapter 4**
  J.L. Eveleens and C. Verhoef. Rise and fall of the Chaos report figures.
  IEEE Software, vol. 27 (1): pag. 30–36, 2010.
- **Chapter 5**
  J.L. Eveleens and C. Verhoef. Quantifying forecast quality of IT business value.
  Article accepted for publication on July 8, 2011. To appear in Science of Computer Programming.

1.9 EQUITY Research

In the field of IT economics, Verhoef wrote a number of papers [122, 123, 124, 125] that discuss methods for organizations to utilize available data to obtain useful quantitative insights in managing IT. Based on his initial work, a research program called Exploring QUantifiable Information Technology Yields, or EQUITY, was started. The EQUITY project aims at studying the possible connections between yields and information technology, so that competition with software will be enabled in a calculated manner. The goal is to develop a quantitative approach which is both sufficiently accurate, and can be used in software-intensive organisations to facilitate rational decision making on software investments.

EQUITY's objective is to provide managers with information from the bit to the board level. That is, to obtain information from all data sources available to an organization ranging from the source code, project administrations and the financial situation of the organization. Notable work based on the information contained in source code was performed by Kwiatkowski et al. [74, 75]. Based on project administrations, Kulk et al. analysed IT risks such as requirements creep and reasons behind misestimations [72, 73]. Eveleens et al. and Kampstra et al. researched IT processes on an organizational level, quantitively analysing the impact and benefits of software process improvements [25, 66].

In this thesis we research IT forecast quality in measure and number. We assess how to quantify the quality of IT forecasts and use the information to enhance decision making. The methods provide IT executives with information that enhances rational decision making of software investments. Just like basing the duration of a cycling trip on the length and terrain of the stage, the weather conditions and traffic, rather than biased expectations.

Acknowledgments This research received partial support by the Dutch *Joint Academic and Commercial Quality Research & Development (Jacquard)* program on Software Engineering Research via contract 638.004.405 *EQUITY: Exploring Quantifiable Information Technology Yields* and via contract 638.003.611 *Symbiosis: Synergy of managing business-IT-alignment, IT-sourcing and off shoring success in society*. Furthermore, we would like to thank a number of organizations that will remain anonymous and Todd Little of Landmark Graphics Corporation for kindly sharing their forecasting data with us. Moreover, we thank Brad Appleton, Capers Jones, Magne Jørgensen, Steve McConnell, George Tillmann, Nicholas Zvegintzov and Sandjai Bhulai for their comments. Finally, we thank PwC for sponsoring this thesis.

CHAPTER 2

Terminology

In this chapter, we introduce terms and notations that we use throughout the thesis. In Figure 2.1, some of the terms and their relations are depicted. Although most terms are used frequently in common English, the meaning and understanding of them differs among people.

For example, in 2006 there was a lively discussion in IEEE Software [37] between Little, McConnell, Gryphon and Kruchten about the cone of uncertainty in general and Little's article [84] on this subject in particular. Since in this thesis we will not only find results similar to Little's, but also other results, we analyzed their discussion, which is cited and misunderstood by others [128]. We learned 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 this thesis 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 uniform terminology. We transpose various views to this terminology in order to assess their validity.

Asset An asset is defined by the Oxford dictionary [119] as an item of property owned by a person or company, regarded as having value and available to meet debts, commitments, or legacies. In this thesis, we assume that projects need to be executed to create these assets. Even if a project modifies an existing asset, we will consider the altered asset as being a new asset. The assets can be both tangible and non-tangible.

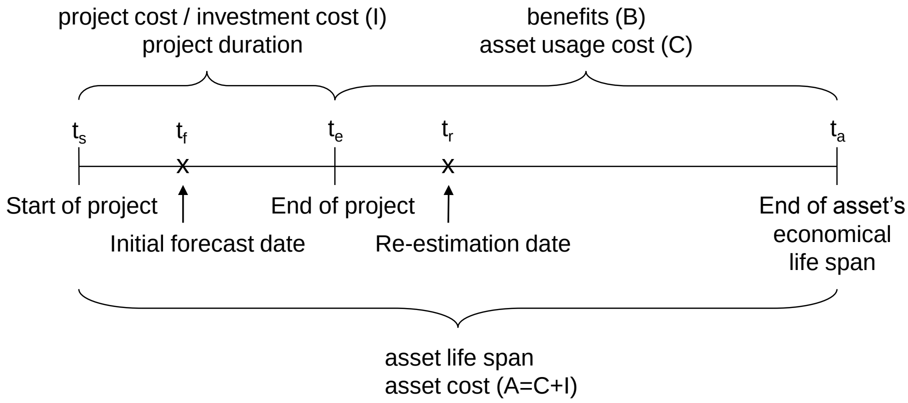


Figure 2.1: Relevant terms in the life span of an IT asset.

Entity In this thesis, we will use the word *entity* to denote any quantifiable aspect of an asset that is of interest. For instance, entities that we will consider are the project cost or benefits. Other examples are economic indicators, such as the Net Present Value or Internal Rate of Return.

Asset usage cost and project cost We will use the term cost in different ways. The total asset costs, denoted by A is the full range of costs that is made for an asset. These costs consist of the project costs, denoted by I for investment costs, and the asset usage costs, denoted by C. We will consider project costs to constitute merely the costs for executing the project. The asset usage costs are all costs excluding the project costs. This entails, for instance, marketing, network usage and maintenance costs.

We note that in this thesis we will only discuss costs that are computed in their present value. When we refer to A, I or C, we refer to the discounted asset costs, discounted project costs or discounted asset usage costs.

We make a distinction in costs as most organizations try to manage, control and contain project costs. This is done mainly due to the relative ease with which the project costs can be measured. In Chapter 5, we investigated to which extent control of this subset of the costs is useful in the context of the entire benefits and costs of an asset.

We note that for this thesis, it is not important how the asset costs are derived. We consider the asset usage costs as given. We will not question how the costs are computed, what precisely should be quantified and how it should be quantified. We assume that the estimators within an organization use equal definitions of how and what to incorporate in determining the costs.

Benefit In the Oxford dictionary [119], a benefit is defined as an advantage or profit gained from something. In the context of this thesis, we consider a benefit, denoted by B, as a *quantified monetary* advantage or profit gained from something. Thus, we are only discussing benefits that have been quantified and represent a monetary gain.

We note that in this thesis we will only discuss benefits that are computed in their present value. When we refer to B, we refer to the discounted benefits.

In this thesis, we will not dive into the question how the benefits have been quantified. Similar to the asset costs, we do not consider what precisely should be quantified and how it should be quantified. We assume that estimators within an organization apply the same definitions to compute the benefits.

Cash flow In this thesis, a cash flow CF of a certain time period p is equal to the benefits B minus the asset costs A in that period. Since the benefits B and asset costs A are assumed to be discounted, summing the cash flows leads to the Net Present Value, which will be discussed in Chapter 5.

Duration and life span As with costs, there are a number of durations we will consider. The different durations are displayed in Figure 2.1. The first is the project duration, given by $t_e - t_s$. The second duration that is of interest is the economical life span of an asset, given by $t_a - t_s$. We define this life span of an asset as the period over which benefits and asset costs are forecasted.

Forecast We define a forecast (or forecasting) of a certain entity as the prediction of the value that entity will have in the future. A forecast consists 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 costs and costs that are still to be made. The burned costs are the ex-post part and the future costs form the ex-ante part; their sum is the forecast of the total IT costs. In other words, a forecast (or forecasting) of a certain entity is a prediction of that entity.

In Chapter 3, the entities that we will analyze and focus on are the cost, 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. In Chapter 5, we extend our approach to allow for indicators that are able to assume positive and negative values, such as the Net Present Value.

Since we will address forecasts of different entities, we introduce a notation for forecasts. When we discuss forecasts, we will denote forecast f of entity e as f_e. If there is no ambiguity about the entity in question in a particular paragraph or it is of no relevance, we will simply use f.

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 [119], 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 [37]. Both ex-post and ex-ante are also used in a book on forecasting [4], although there they have a slightly different interpretation.

Ex-ante In this thesis, 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 executive 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 chapter, 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.

We 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 [84]. 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.

Point forecast The forecasts we discuss in this thesis are point forecasts. A point forecast is a single prediction that is often a summary of a large range of possible outcomes.

When a prediction is made, an estimator considers multiple scenarios that may occur and all relevant risks for the entity in question. For example, consider an estimator that needs to estimate the business value of an IT asset. The estimator should consider the risk that the project required to create the IT asset gets canceled. A study by Capers Jones found that of software applications in the 10.000 function point size range, about 36% are canceled and never completed [57]. The cancelation of a project significantly impacts the project outcome up to a swap to negative business value. This risk should be accounted for in the range of possible outcomes when forecasting the business value early on.

Another example is the possibility of litigation when the asset is developed under a contract. In 2001, Capers Jones and his colleagues at SPR observed that 5% of projects within the United States that were outsourced, were probable to result in litigation or had litigation in progress [55]. Capers Jones states that an average lawsuit in the U.S. costs both the plaintiff and the defendant so much money that all applications ending up in court have negative values. If an IT application is going to be developed under contract, a formal risk assessment is needed plus very strong contracts with penalties for non performance.

If one third of large applications are canceled, and 5% of outsourced projects may result in litigation, the CIO needs more certainty than exists today that applications receiving funds will be developed using best practices. This implies that an early risk analysis should be part of the funding equation.

A risk analysis considers the likelihood the risk will occur and its impact on the entity to be forecasted. Through personal communication from Capers Jones, we received a list containing 200 potential risks that can influence the outcome of software projects [58]. Table 2.1 describes 10 risks from that list.

Table 2.1: Ten potential risks that can influence the outcome of software projects [58].

Potential risk
Risk of dilution of ownership due to multipe funding rounds
Risks of difficult data migration from legacy applications
Risk of significant layoffs of project team
Risk of inadequate warranties for quality and security
Risk of security flaws in application
Risk of late start in deploying risk solutions
Risk of estimates being rejected due to lack of benchmarks
Risk of software raising hardware warranty costs
Risks from disconnected "stove pipe" applications
Risk that requirements are not kept updated after release

These scenarios and their chance of occurrence lead to a range of possible outcomes. This range or interval of possibilities is the prediction of the value of interest and provides information on the risks related to the project. The interval allows the management to set adequate targets and make commitments based on their risk averseness or appetite.

However, in practice this interval is rarely given to the management. In many cases, the interval is summarized to a single point forecast, for instance, the most likely scenario to occur. As the management is confronted with these point forecasts, we assess their quality in this thesis. Next to that, we discuss ways to recreate the interval based on historical point forecasts in Section 5.6.

Actual We define an actual of a certain entity to be the final realization of that entity. That is, the actual is the true value of that which has been forecasted. For instance, if a project costs 1 million Euro in expenses the actual is 1 million Euro.

The notation that we will use for actuals is the same as with the forecasts, a_e with a an actual of entity e. Again, we also use the shorter version a when it is clear which entity is referred to.

In this thesis, 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 3.4, we will 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 executives want estimators to provide a prediction of the final realization. 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.

It is important to realize that, in Chapter 3, we will use the final realization as reference to assess the quality of forecasts made. In Chapter 5 we will use the final realization or the best available approximation of the final realization as reference.

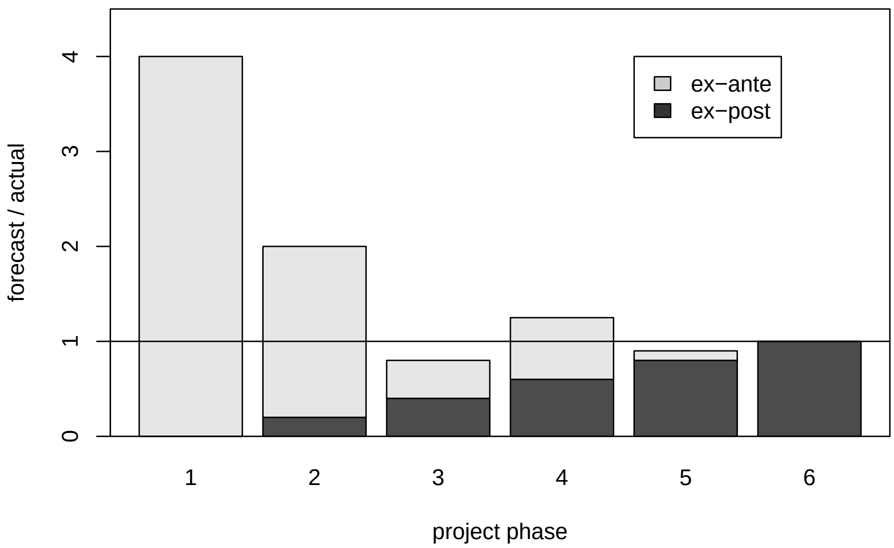


Figure 2.2: Example forecasts.

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 executive wants a conservative forecast to determine the budget, the executive can use the forecast of the final realization and add a percentage to serve as budget. If the IT executive 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 Figure 2.2 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 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.

With these definitions we are able to assess the discussion in IEEE Software [37] between Little, McConnell, Gryphon and Kruchten about the cone of uncertainty in general and Little's article [84] on this subject in particular. In Table 2.2, 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.

Table 2.2: Translation of different views in terms of this thesis.

Author	forecast	ex-post	ex-ante	actual
Boehm	estimate	-	-	actual
McConnell	estimate	-	-	actual
Little	estimate	-	remainder	actual
Gryphon	-	-	estimate	-
Kruchten	absolute overall uncertainty	-	-	-

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 forecast quality quantified by function G_k, to be better forecasted than other collections. Let E be a collection of p projects with their forecast quality quantified by function G_k, and let F be a similar collection of q projects. That is, the forecast quality of the projects in both collections E and F are quantified using the same underlying function g. 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 [20, p. 14] advises to use the median value exactly for this reason. Additionally, in other publications, e.g. [117], 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 forecast quality.

Forecast-to-actual ratio A forecast-to-actual ratio or f/a ratio is a measure to assess the quality of forecasting. In the IT-context, this ratio was introduced by Barry Boehm [8]. This term is also used in other research areas [90]. It measures the forecast quality by dividing the forecast by its actual. We will analyze f/a ratios of several entities. When necessary, we denote f_e/a_e also as $(f/a)_e$ for forecast f and actual a of entity e.

Forecast-to-actual plot To assess the quality of forecasts, we plot the f/a ratios in what is known as the forecast-to-actual plot, or f/a plot. The f/a plot depicts f/a ratios against the relative time at which the forecasts are made. The f/a ratios that are depicted should be a homogeneous set of forecasts.

Reference point We define a reference point as the reference to which we compare the value of forecasts. An example of a reference point is the actual itself. In Chapter 3, we will only use the actual as reference point. In Chapter 5, we also use other reference points. For instance, we will use re-estimations of the benefits and the asset usage costs made a year after project completion as reference point. In Section 5.3, we will discuss the implications of using different reference points.

Bias Another term we will frequently use in this thesis is *bias*. In the Oxford dictionary [119], 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.

We note that the above definition does not exclude any sort of bias. For instance, cognative or other biases as discussed in Chapter 1, can cause one to consistently err and thereby introduce systematic deviations in a forecast. Therefore, all potential biases that occur intentionally or unintentionally and affect the forecasts are considered in this thesis.

Target and commitment An important reason for political bias is the difference between forecast, target and commitment. A number of authors [3, 61, 77, 84, 89, 126] 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 [61] notes that many people mix up these terms, even within the same project.

Armour [3] 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 executive 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 executives to assess the risk they take by setting certain targets or commitments. In Section 3.5, 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. In this thesis, we will make use of the Estimating Quality Factor or EQF to compute forecast quality. The EQF, which we will discuss extensively in Chapter 3, computes deviations between forecasts and actuals and weighs these deviations with time. Therefore, the EQF is a function dependent on time. Still, we will not give definitions of the start or end date of a project, as this is outside the scope of this chapter. However, 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 thesis.

Now that we set the terminology and precisely defined the notions that we will use throughout this thesis, we are able to commence with quantifying IT forecast quality.

CHAPTER 3 Quantifying IT forecast quality

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 [20]

3.1 Introduction

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 can lead 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.

In this chapter, we propose how to quantify IT forecast quality, so that IT executives know their forecast quality and what bias they can expect. We discuss how to visualize and quantify the accuracy of forecasts. There are two well-known tools dealing with this assessment. We discuss the merits and limitations of both these tools in this chapter. One tool, the Estimating Quality Factor (EQF) [20], 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 [8]. This figure depicts deviations between forecasts and actuals by plotting forecast-to-actual ratios on a logarithmic scale for different phases of a project. With the logarithmic axis, the plot shows a symmetric conical shape indicating a decrease in the deviations as a project progresses.

In this chapter, 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, 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 of the costs. This leads to a different shape than forecasts without political bias, as in Boehm's case.

Especially, when one never assesses IT forecast 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 executives up to the highest level assumed the forecasts to be accurate. Needless to say that this is an entirely unwanted situation. In this chapter, we propose a method to reveal the bias by making the deviations of the forecast from the actual transparent.

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 chapter, we evaluate a number of benchmarks found in the literature related to IT forecasting.

Organization of this chapter The remainder of this chapter is organized as follows. In Section 3.2, 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 3.3 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, for instance, 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 3.4. We display several plots of forecast-to-actual ratios against a reference cone with different forecasting patterns, and provide the accompanying EQFs.

In Section 3.5, we describe how to use the analyses proposed in this chapter to enhance forecast information for decision making. We discuss three approaches that provide additional information about the uncertainty of newly made forecasts.

Section 3.6 challenges the credibility of benchmarks related to forecasting in the literature. We review benchmarking information regarding EQFs and provide new benchmarks based on the four presented case studies: best, worst and mid-case benchmarks.

On request of practitioners, in Section 3.7, we provide an overview of the lessons learned in this chapter. 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 3.8, we conclude.

3.2 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 [8, p. 310–313], which describes methods and procedures for software cost estimation. We recall this famous result in Figure 3.1. 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 [10, 88], the interpretation of the cone is that when one knows more about a software project, your forecasts will improve.

The horizontal axis in Figure 3.1 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 Figure 3.1 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 Figure 3.1, is known as a confidence interval. McConnell [88, p. 169] popularized the use of this interval for IT forecasts to give an indication of their uncertainty. Tockey [117, p.351–355] further explored the method by showing how to compute the uncertainty based on historical data. Others, for instance [128], 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

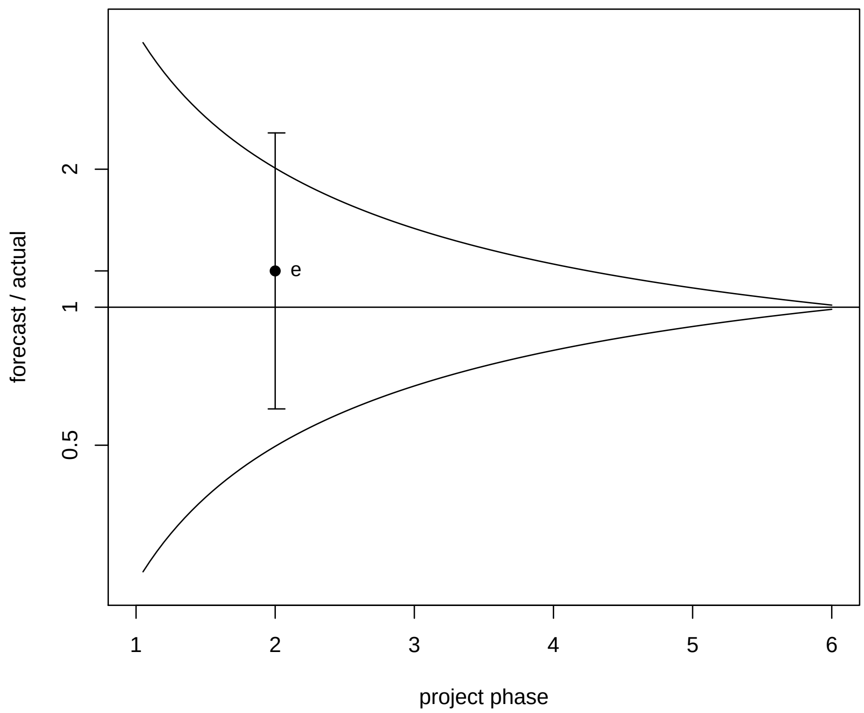


Figure 3.1: Boehm's cone of uncertainty.

of the confidence interval.

Different views In a 2006 IEEE Software article [84], Little questions the implication of Boehm's cone of uncertainty [10] 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 [37] 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 [8]. 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 [15] and Laird [76].

In later work [10], 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 forecast quality improves, this is only true when the ex-post part is known and used in making the forecast. 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 and use all information available, which we will show in this section. Later on in Section 3.4, we will show that if the ex-post part is not used, forecasts do not necessarily improve. Recall that with forecasts we mean the total of the ex-post and ex-ante part.

- McConnell [88] endorses the viewpoint of Boehm that one can forecast a project more accurately as more knowledge becomes available. McConnell states in an article [37] and in his book [89, 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 [3]. According to McConnell, Boehm's cone therefore does not promise an improvement of the forecasts. In Section 3.4 and Section 3.5.2, 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 uncertainty of the ex-ante part does not need to decrease 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 [130], 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.

3.2.1 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 [8, 99].
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 on a logarithmic scale 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 costs of a project, we also take into account all activities regarding the costs for creating the requirements and performing the business study.

One author [37] 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 [20] stated, the estimator should construct the forecast in such a way that the development method is reflected in it. However, in this chapter, 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 costs 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 costs 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 [25], 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 half way 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 chapter we consider only duration.

In the other article [25], 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 follows 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 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 [20], 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.

3.2.2 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 [103]. 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
factor.of.deviation = 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) *
           (factor.of.deviation - 1/factor.of.deviation) +
           1/factor.of.deviation)
   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 `factor.of.deviation` 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 factor of deviation. 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 Figure 3.2. The asymmetry around the actual value of the cone of uncertainty is also observed by Laranjeira [78] and Cohn [18]. 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 Figure 3.2, 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 `factor.of.deviation` $= 4-2i/100$, 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 Figure 3.2, the ex-ante accuracy *decreases* as time progresses. We used as theoretical model a linear decrease of the accuracy using the formula `factor.of.deviation` $= 4 + 2i/100$, 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 Figure 3.2, 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 does not need 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

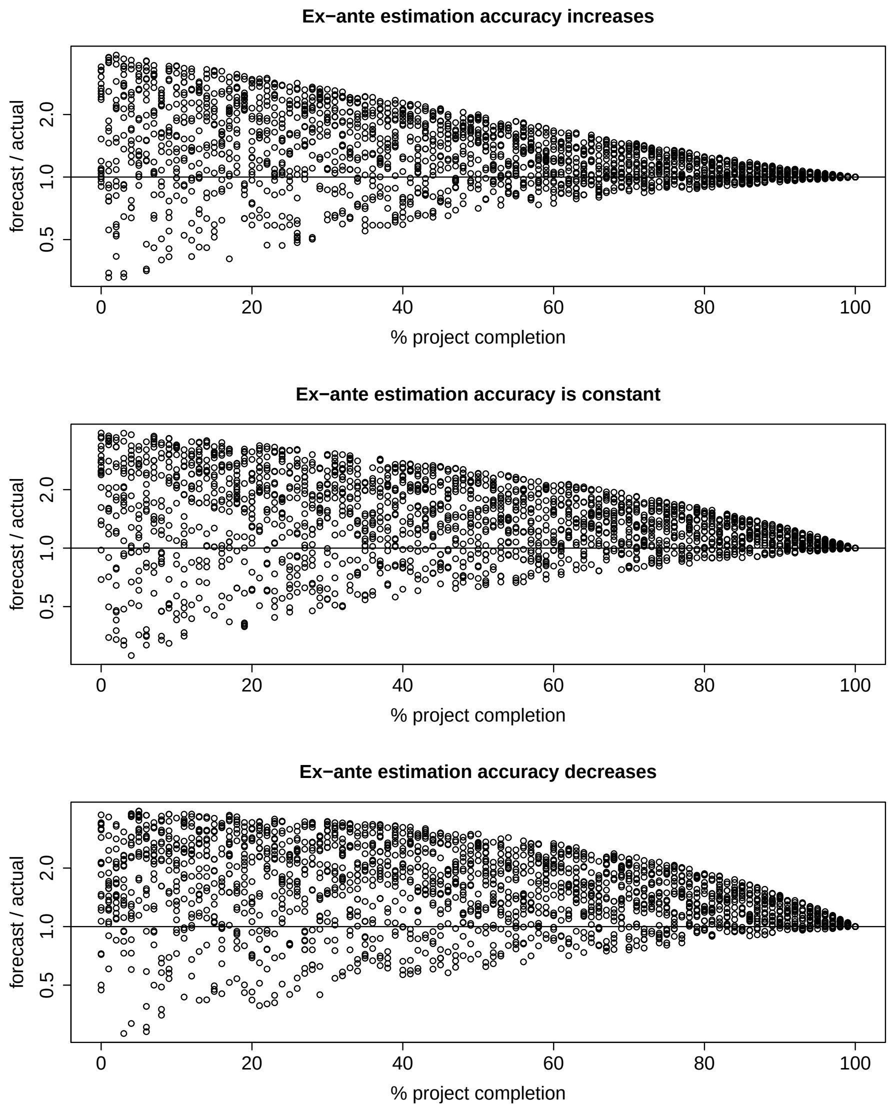


Figure 3.2: Simulation of the cone of uncertainty with increasing, constant and decreasing ex-ante estimation accuracy.

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.

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 a factor of deviation of 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. 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 3.3.2.1, we will explain this in more detail. Other possible symmetry is discussed in Section 3.5.3, where we will discuss the work done by Little [83]. 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 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

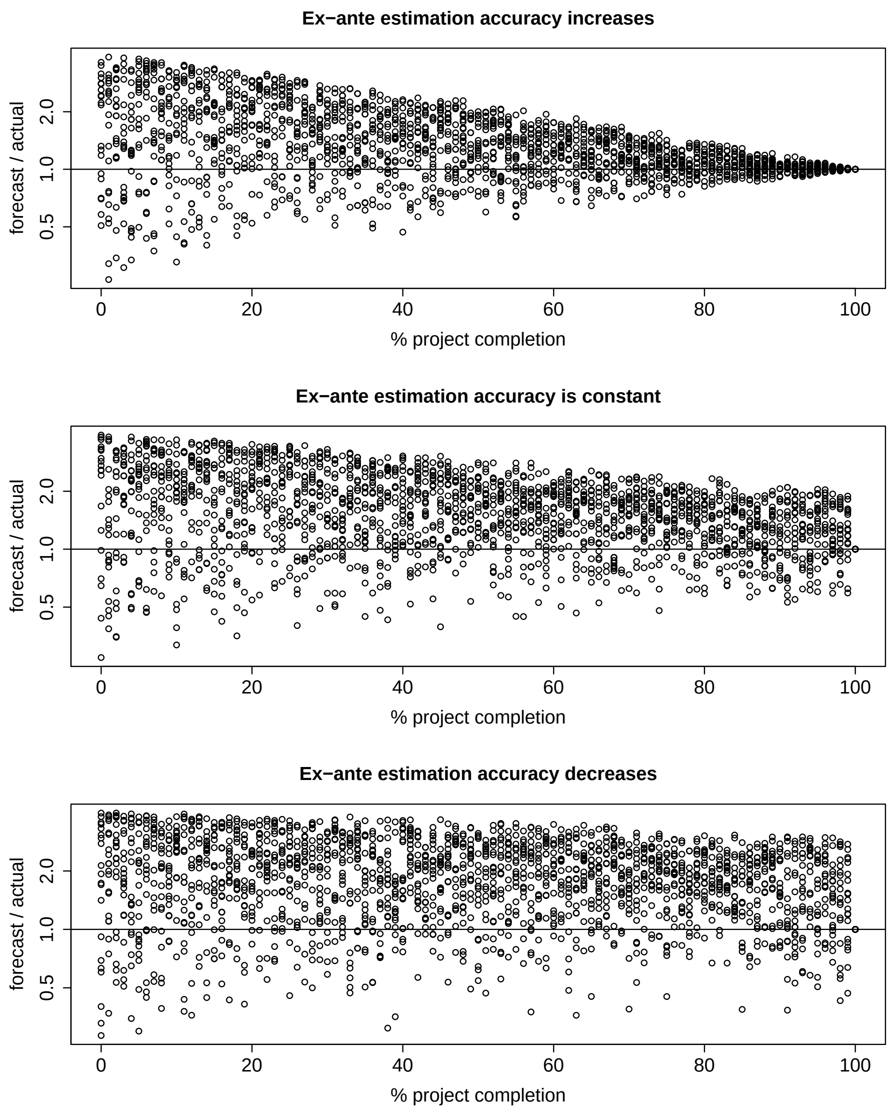


Figure 3.3: Simulation of cones of uncertainty with increasing, constant and decreasing estimation accuracy of the ex-ante part, assuming a uniform function for the ex-post part growth and *not* knowing the ex-post part exactly thus far.

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 [99, 101]. We changed the simulation to use the Rayleigh function with a peak of the effort at 60% of the project's progress [8, 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 Figure 3.2. 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 3.4, 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 Figure 3.3. 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 [84]. 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.

3.3 Quality of forecasts

In the introduction, we argued that the quality of forecasts 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 forecast 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 [20]. We will show that the EQF is a forecast 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 forecast 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 unbiased 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 forecast 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.

3.3.1 Estimating Quality Factor

The EQF was defined by DeMarco in his book, *Controlling software projects* [20]. 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 [20, 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 [125], 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_s the start date of the asset, t_a the end date of the asset and $e(t)$ the value of the forecast at time t ($t_s \leq t \leq t_a$). Then, the EQF is represented mathematically by:

$$\begin{aligned}
\text{EQF} &= \frac{\text{Area under actual value}}{\text{Area between forecast and actual value}} \\
&= \frac{\int_{t_s}^{t_a} a\; dt}{\int_{t_s}^{t_a} |a - e(t)|\; dt} \\
&= \frac{\int_{t_s}^{t_a} 1\; dt}{\int_{t_s}^{t_a} |1 - e(t)/a|\; dt}. \qquad (3.1)
\end{aligned}$$

In Section 5.3, we will generalize the EQF for other reference points than the actual and discuss the impact of this generalization.

The assumption is that $e(t)$ is known for the range $[t_s, t_a]$. That is, we know at all times 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 $t_s < x < t_a$. 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 $[t_s, x)$ equals $e(x)$.

Figure 3.4 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 the area between the upper line and the actual value (horizontal line at 1), and the area between the lower line and the actual value both compute to an EQF value of 3.5, a little less than our just found 3.6.

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. In case of positive valued entities, 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 area 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

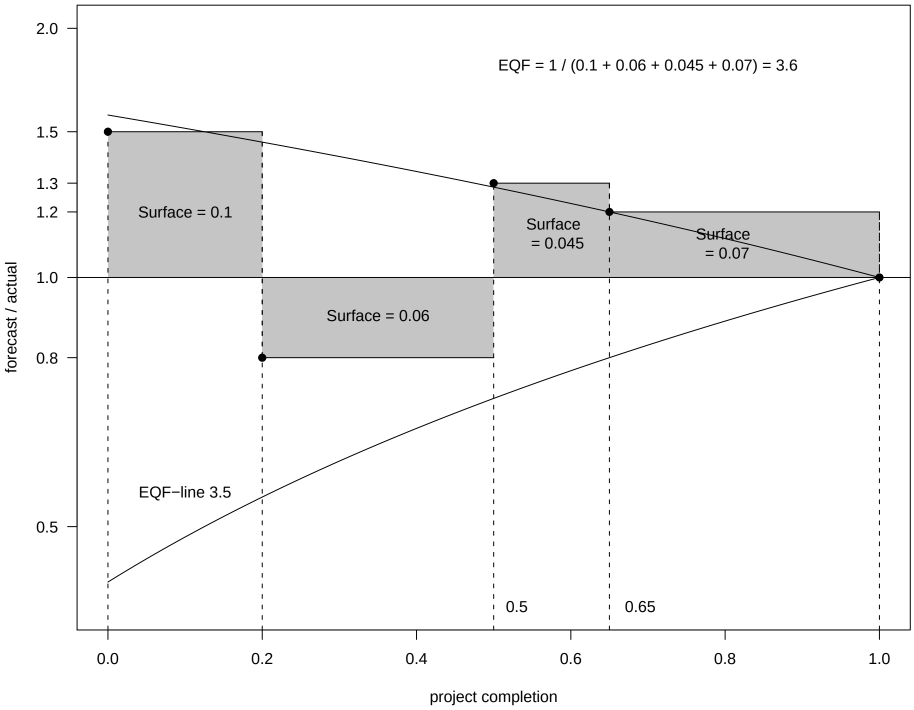


Figure 3.4: An example EQF calculation of a single project with lines of a reference cone.

bound of one for solely underestimation holds for every situation for positive valued entities. 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 Chapter 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 In statistics, other measures are known that quantify the quality of forecasts, such as the MSE, MAPE and MRE or Magnitude of Relative Error [19, 29, 45]. We found that the MRE is used more often than the EQF in the literature [19, 59, 68, 71, 89, 91, 93, 116]. The MRE was introduced in 1986 in a textbook by Conte et al. [19]. 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

$$\begin{aligned} \text{MRE} &= \frac{|a - f|}{|a|} \\ &= |1 - f/a|. \end{aligned}$$

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/MRE. 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 forecast quality of a project with the MRE, it is possible, for instance, to take the average MRE of all forecasts, or mean MRE (MMRE). This means the MMRE weighs each forecast the same, even though the forecasts are possible not equally spaced in time. The drawback of this approach 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 halfway during the project. This allows to boost forecast 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 time-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 forecast quality than the MRE.

It is possible to use the other measures such as the MSE, MAPE or MRE instead of the EQF. However, the benefit of the EQF is that it is defined as a time-weighted average deviation to the actual. In our analyses, we assess the forecast quality of assets that can have multiple forecasts. For decision making, it is important for these forecasts to be as quickly as accurately as possible. Therefore, it is important to account for the timing of subsequent forecasts. The EQF is defined to incorporate this effect.

We 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 Chapter 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 3.4, 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 [20, 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 calculates 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 this chapter, we will use the EQF as defined in Formula 3.1, 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 Figure 3.4, we depicted an example calculation of an EQF value for a single project. This figure uses the same axes as in Section 3.2 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 Chapter 2, $E = \{G_i : i = 1, \ldots, p\}$ to be a collection of p projects with their forecast 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 forecast 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 forecast 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 forecast 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 forecast quality can be rewarded.

- 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 [19, 59, 71, 91].

- We can compare between organizations. Different organizations can benchmark their own IT forecast 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 forecast quality.

EQF variance In a number of publications [20, 82, 83, 84, 125], we found that some of the comparisons were done solely based on single values without considering their spread. Articles by Lister [82] and Verhoef [125] state that projects with EQF values in the order of 10 can be considered good. DeMarco [20, p. 157] reports an *average* EQF value for a group of projects. Articles by Little [83, 84] use the *median* value of a group of projects to compare the forecast quality with forecast quality of other industries. In this chapter, 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.

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 Figure 3.5. 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. [71] 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 Figure 3.5, 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 is to illustrate that using box plots is more insightful than using just a single aggregated EQF.

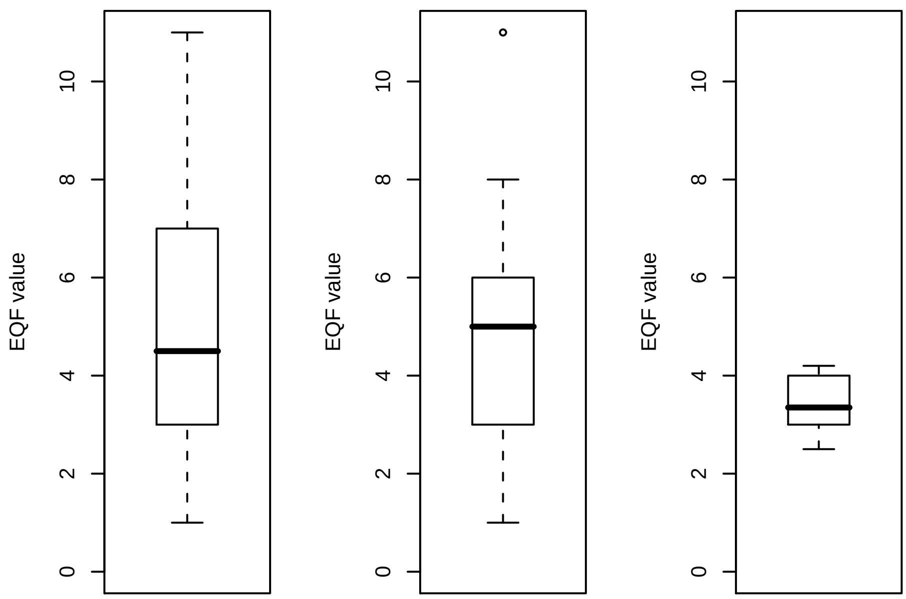


Figure 3.5: Three box plots of constructed example portfolios with ten projects each. The median EQF values from left to right are 4.5, 5 and 3.35.

In summary, by quantifying the forecasts through a box plot of the EQF forecast metric, management has a tool to assess IT forecast 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 [20] 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, by assessing the estimators solely on the EQF value, they are inclined to make 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 of the discrepancy, the estimator benefits from taking it into account thus leading to an increasing number of forecasts per project.

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 using additional data from the ex-post portion is an improvement. Also, if the forecast takes into account velocity [18], 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 executives 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.

We 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.

3.3.2 f/a plots

In this section, we discuss f/a plots. In Section 3.2, 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 3.2.1 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 3.2.2, 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 3.2.2, 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 [83]. 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, we 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). We 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).

3.3.2.1 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. Alternatively, we could also use the first approach to find out which assumptions are the cause for the difference in shape.

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 [87] to compute the results.

In our example case, we want the forecasts to abide by the conditions as formulated in Section 3.2.1. 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 Chapter 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. Let t be any time during a project with $t_s < t < t_a$. Let a be the actual value at time t_a and let x be the project's progression relative to its project duration with $x = (t - t_s)/(t_a - t_s) \in [0, 1]$. Now, we need to specify the forecast at relative time x, $e(x)$.

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 $\int_0^1 g(x)dx = 1$ which results in $y = 1$. We find the relative size of the ex-post part by solving the integral $p(x) = \int_0^x g(x)dx = x$. Thus, at $x = 20\%$ project completion, $p(20\%) = 20\%$ of the work has been done. Therefore, the size of the ex-post part is $a \cdot p(x)$.

The estimation accuracy of the ex-ante part remains the same as the project progresses. The ex-ante part at any time x of the project is of size actual – ex-post, with ex-post being $a \cdot p(x) = ax$. Therefore, we find the size of the ex-ante part to be $a - a \cdot p(x)$.

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 $(a - a \cdot p(x))/c$ and the upper bound is $c(a - a \cdot p(x))$.

With the ex-post and ex-ante part defined, we are able to specify $e(x)$. In case of solely underestimations, $e_l(x)$ is given by $a \cdot p(x) + (a - a \cdot p(x))/c$. For solely overestimations, we find $e_u(t) = a \cdot p(x) + c(a - a \cdot p(x))$.

However, the f/a plot does not depict $e(x)$, but it illustrates f/a ratios. Therefore, we need to formulate the reference lines in terms of f/a ratios. Thus, the lower bound l and upper bound u of a forecast made at time x are mathematically defined by:

$$\begin{aligned} l(x) &= \tfrac{e_l(x)}{a} = p(x) + \frac{1 - p(x)}{c} \\ u(x) &= \tfrac{e_u(x)}{a} = p(x) + c \cdot (1 - p(x)). \end{aligned}$$

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 3.4. 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 3.2, 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:

$$l(x) = p(x) + \frac{1 - p(x)}{c_1} \tag{3.2}$$

$$u(x) = p(x) + c_2 \cdot (1 - p(x)). \tag{3.3}$$

We note that $p(x) = x$ 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 Figure 3.4. 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:

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

Solving c_2 from this equation yields:

$$
\begin{aligned}
c_2 - 1 &= \frac{2}{\mathrm{EQF}_u} \\
c_2 &= 1 + \frac{2}{\mathrm{EQF}_u}.
\end{aligned}
$$

Via 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:

$$
l(x) = x + \left(1 - \frac{2}{\mathrm{EQF}_l}\right) \cdot (1-x) \tag{3.4}
$$

$$
u(x) = x + \left(1 + \frac{2}{\mathrm{EQF}_u}\right) \cdot (1-x). \tag{3.5}
$$

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 3.4 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 at a constant rate.

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 3.5 holds for each EQF value. We note that Formulas 3.4 and 3.5 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 this chapter, 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 3.4 and $\mathrm{EQF}_l = 4.5$ and an upper bound formulated with Formula 3.5 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 chapter, it is equal to using Formula 3.2 and 3.3 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 methods 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 $p(x) = x$. In Section 3.2, 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 [8, p. 93], the growth function $g(x)$ becomes more complex:

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

with b 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 $\int_0^1 g(x)dx = 1$ we conclude that

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

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

$$p(x) = \int_0^x g(x)dx = -be^{\frac{-x^2}{2p^2}} + b.$$

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

$$\begin{aligned} l(x) &= 1 + \frac{1}{\mathrm{EQF}} \cdot (0.58 - 2.32e^{-1.39x^2}) \\ u(x) &= 1 + \frac{1}{\mathrm{EQF}} \cdot (-0.58 + 2.32e^{-1.39x^2}). \end{aligned}$$

We 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 [87] 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 value $p(x)$. That means the ex-post part becomes $a \cdot p(x)/m$ for the lower bound and $n \cdot a \cdot p(x)$ 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 Equations 3.2 and 3.3. For a given c_2, we need c_1 to be as follows

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

  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:

$$\begin{aligned} l(x) &= \frac{1}{p(x) + c_2 \cdot (1 - p(x))} \\ u(x) &= p(x) + c_2 \cdot (1 - p(x)). \end{aligned}$$

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 $1/x$. 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 3.2.2 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 half way 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 to be 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

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

  This leads to

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

  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 $x = 0.5$. 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.

3.3.2.2 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 Figure 3.4 in Section 3.3.1, 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 Formula 3.4 and 3.5. 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 than 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 area 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 forecasts 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 forecast quality and their bias of four large organizations.

3.4 Case studies

In this section, we apply the methods we proposed in this chapter 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 [83]. 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 Chapter 4, 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, 667 forecasts were made of the total costs 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 [20]). 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 costs 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 costs 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 cost forecasts either.

3.4.1 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. Figure 3.6 illustrates these typical patterns. We 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.

Moreover, 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 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.

Finally, we note that the naming conventions of the different patterns given below are ambiguous. For instance, for entities used in this chapter as project cost or project durations, a forecast larger than the actual is a pessimistic forecast. That is, the forecast is a pessimistic projection of what really happened. However, in case of entities such as benefits used in Chapter 5, a forecast larger than the actual is an optimistic forecast. Although the labels are ambiguous, the patterns remain the same.

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 Figure 3.6. Then, we will give examples of situations in which such a pattern can occur. The typical patterns shown in the figure are summarized below.

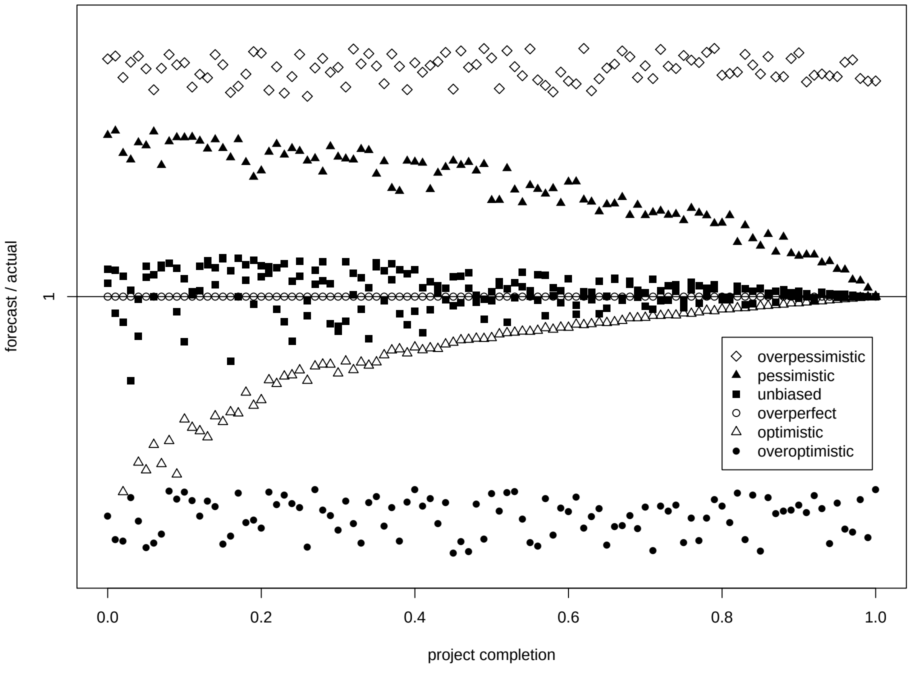


Figure 3.6: Typical patterns in an f/a plot.

- 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 [20]. With an independent metrics group judged on forecast 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.

3.4.2 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 calculate the deviation between the forecast and the actual. We will also derive the EQF values for each project.

Little [84] 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 3.2, we illustrated with simulations that this effect is caused by the use of the ex-post part in making the forecast.

We want to note that the f/a plot by Little [84] is slightly different from the one we present in this chapter. 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 3.3.1, we discussed 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.

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 [20]. 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 3.6.

We extend Little's analyses by applying the methods described in this chapter to his data. We depict the result in Figure 3.7, 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 Figure 3.7, we note that the reference cone exactly follows the conditions that we used to derive Formulas 3.4 and 3.5 of the reference cone in Section 3.3.2.1. 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, for visual comparison purposes, we want an EQF value that most projects have attained. Boehm [8] 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. We note that, although we use this EQF value in this situation, it is always possible to choose other EQF values that one wants to compare with.

The box plot of the EQFs in Figure 3.7 has a median value of 4.7. The solid gray line in the box plot corresponds to the 20% quantile with an EQF value of

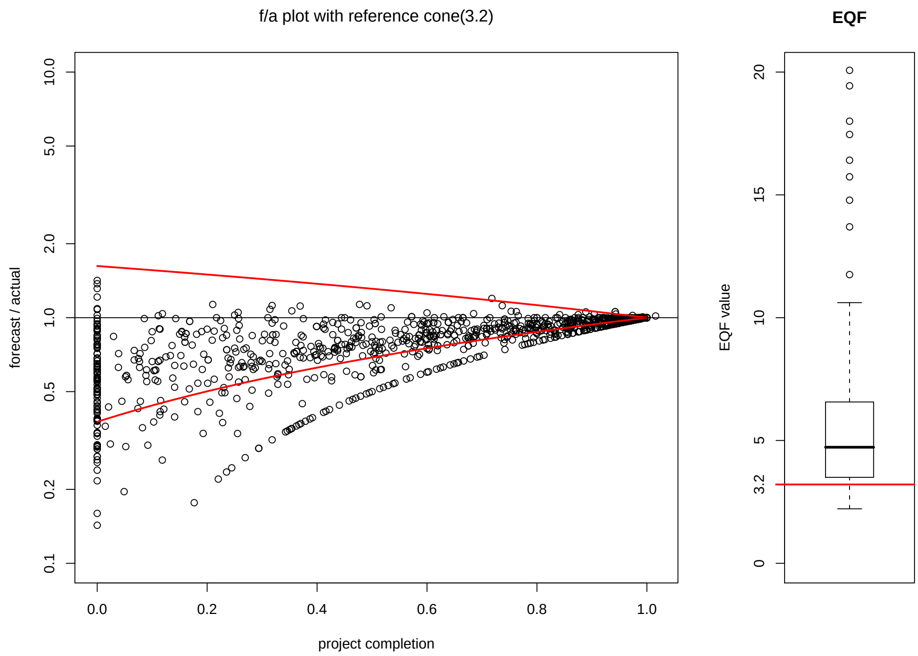


Figure 3.7: f/a plot with reference cone(3.2) and EQF box plot of 121 projects.

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.

Indeed, Little [84] 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 [21] 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 forecasts 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 Chapter 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 3.5, we show how to do this.

3.4.3 Large multinational company X

In the second case study, we assess 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 investigate 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 Figure 3.8. 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.

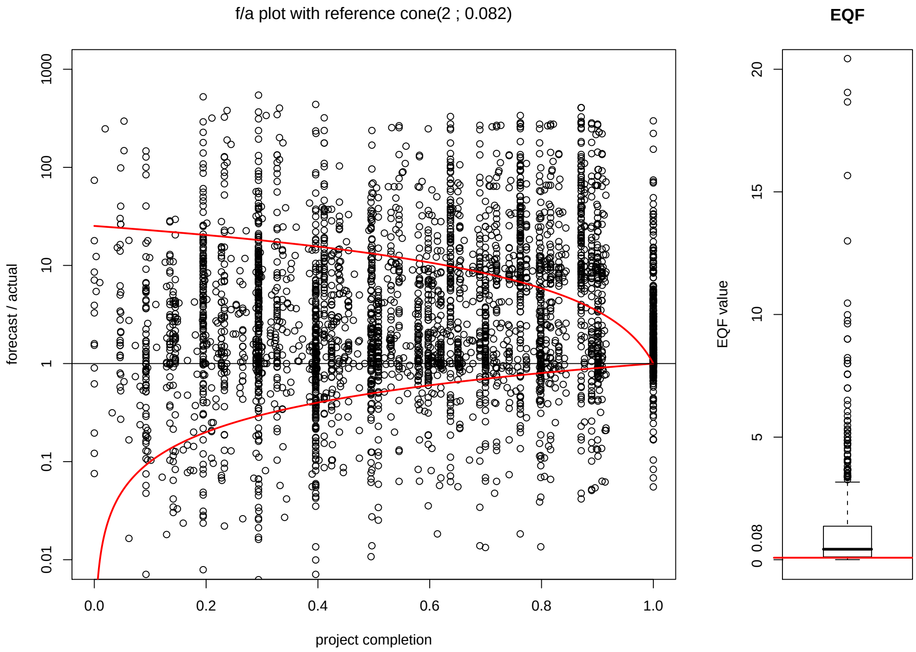


Figure 3.8: f/a plot with reference cone$(2, 0.08)$ and EQF box plot of 867 projects.

The plots in Figure 3.8 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 3.2, we have seen a similar shape in Figure 3.3, 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 costs 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. This leads to overstating budgets, thus increased the safety margin of success. Therefore, forecasters in this organization predicted much higher costs instead of the most probable actual costs.

Also, since reapplying for the budget was difficult, project managers refrained from lowering the forecasts using the ex-post part. If they would update 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 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 forecast 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.

3.4.4 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 Figure 3.9. 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.

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

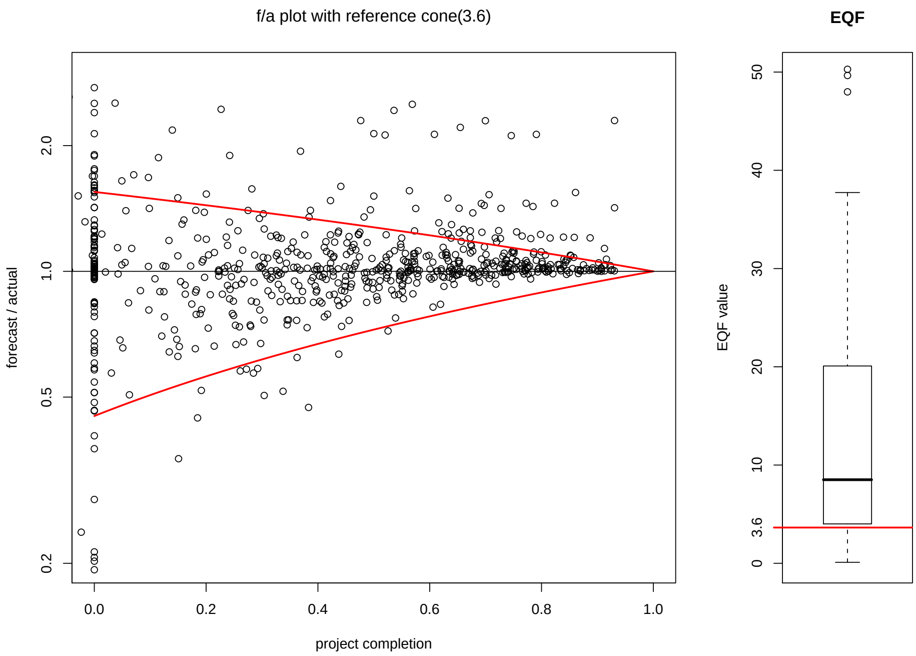


Figure 3.9: f/a plot with reference cone(3.6) and EQF box plot of 140 projects.

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 [23, 32], 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 Figure 3.9.

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 forecasts, which were made for the functionality of the projects measured in function points [23, 32]. The same data set is also used in an article by Kulk et al. [72], 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 chapter, we analyze the deviations themselves to quantify the quality of the forecasts made.

In Figure 3.10, 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.

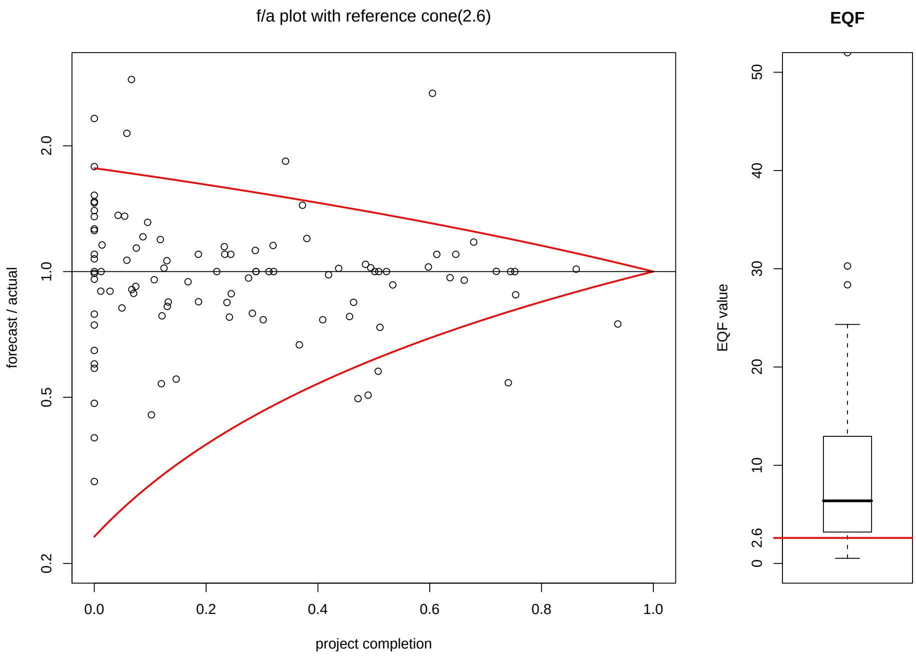


Figure 3.10: f/a plot with reference cone(2.6) and EQF box plot of 83 projects.

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.

Case summary The f/a ratios in this organization compare well with our reference cone as defined in Section 3.3.2.1 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.

3.4.5 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 costs 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.

We note that we only obtained the approved costs and not the actual costs. In this organization, the actual costs of projects were aggregated to higher levels without storing information on the actual costs 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 costs 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 Figure 3.11, 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

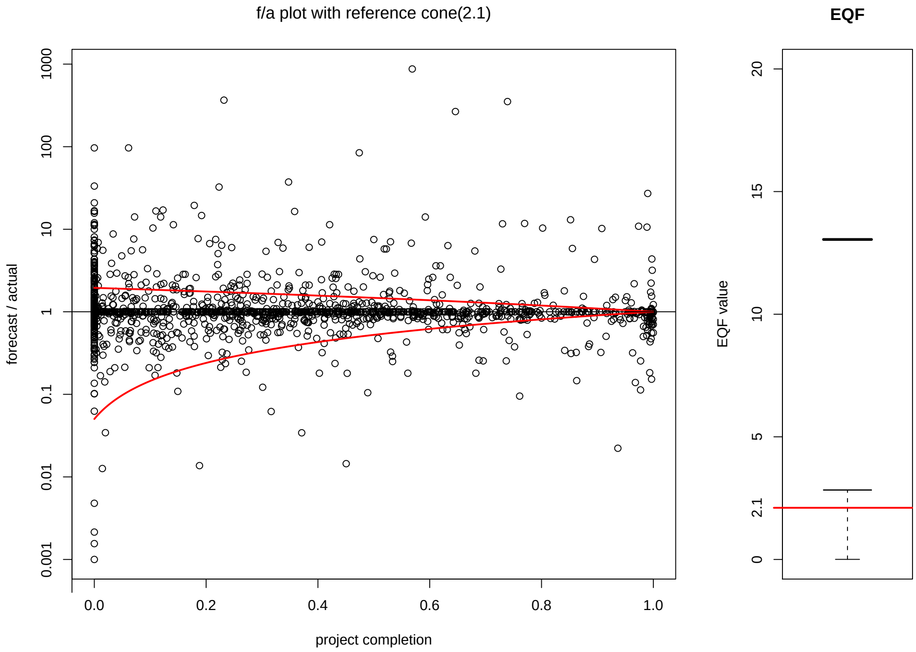


Figure 3.11: f/a plot with reference cone(2.1) and EQF box plot of 613 projects.

than 25% of all the projects are able to perfectly predict the approved costs of the project perfectly. 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 costs 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.
- 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.

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 Figure 3.12. 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

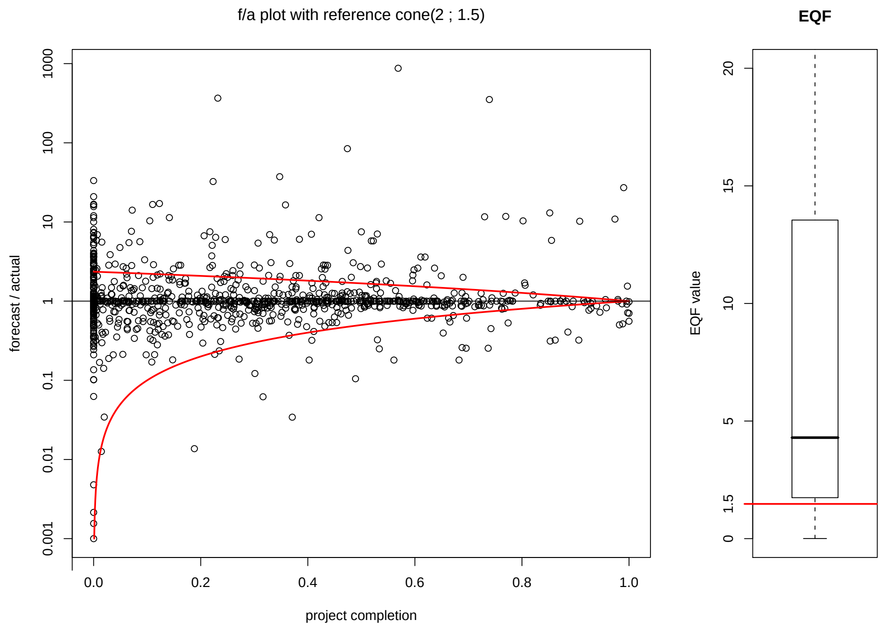


Figure 3.12: f/a plot with reference cone$(2, 1.5)$ and EQF box plot of projects started after 2004 and without projects with an infinite EQF value.

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.

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 costs 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 costs, we note that it is very relevant. The forecasts of costs 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 forecast data. We were able to use our approach to identify the differences in the quality of forecasts 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 costs, less attention was given to the quality of the forecasts of the costs. We agree that the value of a project is important, but we do note that the value is influenced by the costs. Therefore, it is equally important to check and improve the quality of the forecasts of costs as well.

3.4.6 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 3.1. We 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.

Table 3.1: Case comparison based on EQF values and f/a patterns.

Organization	20% quantile	median EQF	mean 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

To properly compare the organizations, we need a criteria to determine what is good and what not. In Chapter 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 3.3.1, we showed that the EQF is suitable as function G_i. Thus, the forecasts of one organization are better than those of others, if 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 3.6, 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 executives 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.

3.5 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 executives 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 approaches that provide such enhanced forecast information. The first approach 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 approach 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 [88] 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 approach 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 approach 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 approaches use historical data to make statements about future forecast uncertainty. The underlying assumption of the three approaches is that the historical 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 approaches must be regarded with extra care.

3.5.1 Basic calculations

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

Consider an IT executive 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 executive 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 3.3.2.1 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.

$$
\begin{aligned}
l(x) &= x + \left(1 - \frac{2}{\mathrm{EQF}_l}\right) \cdot (1 - x) \\
u(x) &= x + \left(1 + \frac{2}{\mathrm{EQF}_u}\right) \cdot (1 - x).
\end{aligned}
$$

In this case we are interested in the deviation at $x = 0$. 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 $l(0)$. We 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 Chapter 4 we show how to correct for biases if more information is available.

Given our median EQF value of 5, we find that $l(0) = 0 + \left(1 - \frac{2}{5}\right) \cdot (1 - 0) = 0.6$. Thus, assuming the median EQF value and optimistic forecasts, 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.

We note that this heuristic 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 $[0.7, 1.7]$ too wide a range of values.

Still, this type of information will enhance the quality of the forecast information provided to the IT executive. 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 approaches that further enhance the forecast information. We note that these approaches 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 3.3, to be cautious in using these approaches. If one has not performed these analyses or does 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 3.4 and other EQF benchmark values, which we will describe in the next section.

3.5.2 The confidence interval

The second approach we discuss to enhance the forecast information is well-known in statistics and is known as a confidence interval. Boehm [8] argues that each forecast should include an indication of its degree of uncertainty. McConnell [88] proposes a method that gives a prediction of this uncertainty by creating an interval around a forecast using the cone of uncertainty. Tockey [117] explored this method further, while others [16] 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. We note that both Kitchenham et al. [71] and Tockey [117, 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 3.4. 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 Figure 3.1, in Section 3.2. 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 [8, p. 310], we know that these intend to represent 80% confidence limits. This means that 80% of the time, the actual 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 Figure 3.1, 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 Euro. 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 Euro. 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 [113], 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 heuristic 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 Euro. 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 Euro. 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 [83]. If two confidence intervals have the same confidence level, we prefer the interval with the lowest U/L ratio.

3.5.2.1 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 s_i to be the f/a ratio of forecast i contained within that group and $s = \{s_i\}$. Let l be the desired confidence level for the confidence interval with $0\% \leq l \leq 100\%$. Let v be the $(100\% - l)/2$ quantile of s and w the $100\% - (100\% - l)/2$ quantile of s. For instance, if we take a confidence interval with confidence level of $l = 80\%$, v is the 10% quantile and w the 90% quantile.

We calculate the confidence limits by $U = 1/v$ and $L = 1/w$. We note that the upper confidence limit is determined by the lower quantile and the lower confidence limit by the upper quantile. The confidence limits are limits for the true value of the actual, or θ. Given the ratios s and the lower quantile v of these ratios, we find θ by $v = f/\theta \rightarrow \theta = f/v = f \cdot (1/v) = fU$. Therefore, the upper confidence limit is determined by the lower quantile and vice versa.

Table 3.2: 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

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. Using the historical f/a ratios, we computed the reciprocals of the 10% and 90% quantiles (v and w). These values form the lower and upper confidence limits and are summarized in Table 3.2. 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. We will also explain in detail how we derived this specific grouping.

The values given in Table 3.2 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 Graphics, it is likely the project will cost between 1.11 and 3.3 million Euro. In case of organization X, it can cost between 0.06 to 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. We note that it is also possible to apply this heuristic 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.

3.5.2.2 Benchmark values

In his book [88], McConnell gives benchmark values for the confidence limits with a confidence level of 80%. McConnell uses the same values as given by Boehm [8]. The values are summarized in Table 3.3, which also shows the U/L ratio. We 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.

Table 3.3: Confidence intervals with confidence level 80% derived from Boehm as used by McConnell [88, 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

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 3.3 with those of our case studies in Table 3.2, 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.

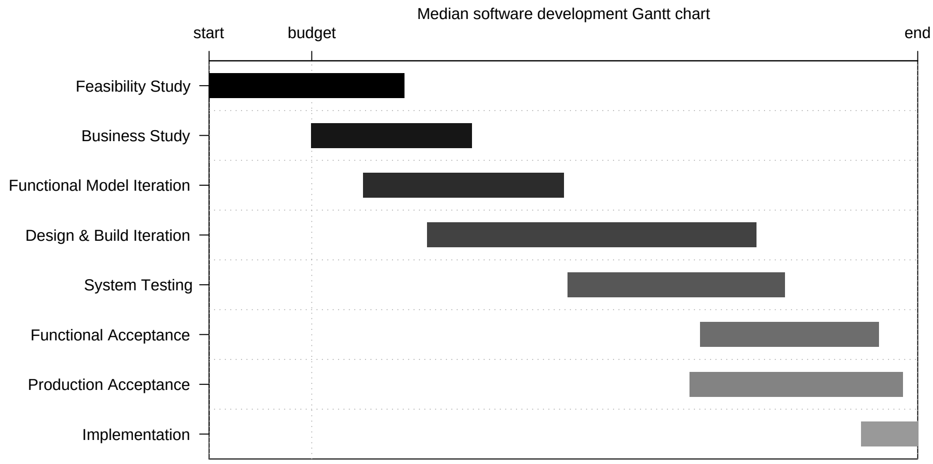


Figure 3.13: An aggregated Gantt chart of median start and end times of different phases for 318 projects normalized for their duration (taken from a article by Eveleens et al. [25]).

Fortunately, in an article by Eveleens et al. [25], 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

Table 3.4: 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%

completeness, we recall Figure 3.13 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 3.4. We note that the translation presented in this table corresponds to the grouping of the forecasts we chose in Table 3.2.

In the translation to 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 the 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 3.3. 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 3.5, we enumerate these confidence levels for each case study. We note that according to Boehm and McConnell this should be around 80% in all cases, and not a lot more or less.

Table 3.5: The confidence levels for the case studies with the confidence intervals proposed by McConnell applied to them.

% of completion	LGC	X	Y cost	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%

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 low or high confidence level is not helpful at all. Therefore, in these cases, the intervals do not provide much information. Only of 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 3.6. 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 3.7. 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.

Table 3.6: 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 3.7: 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%

Although Table 3.7 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.

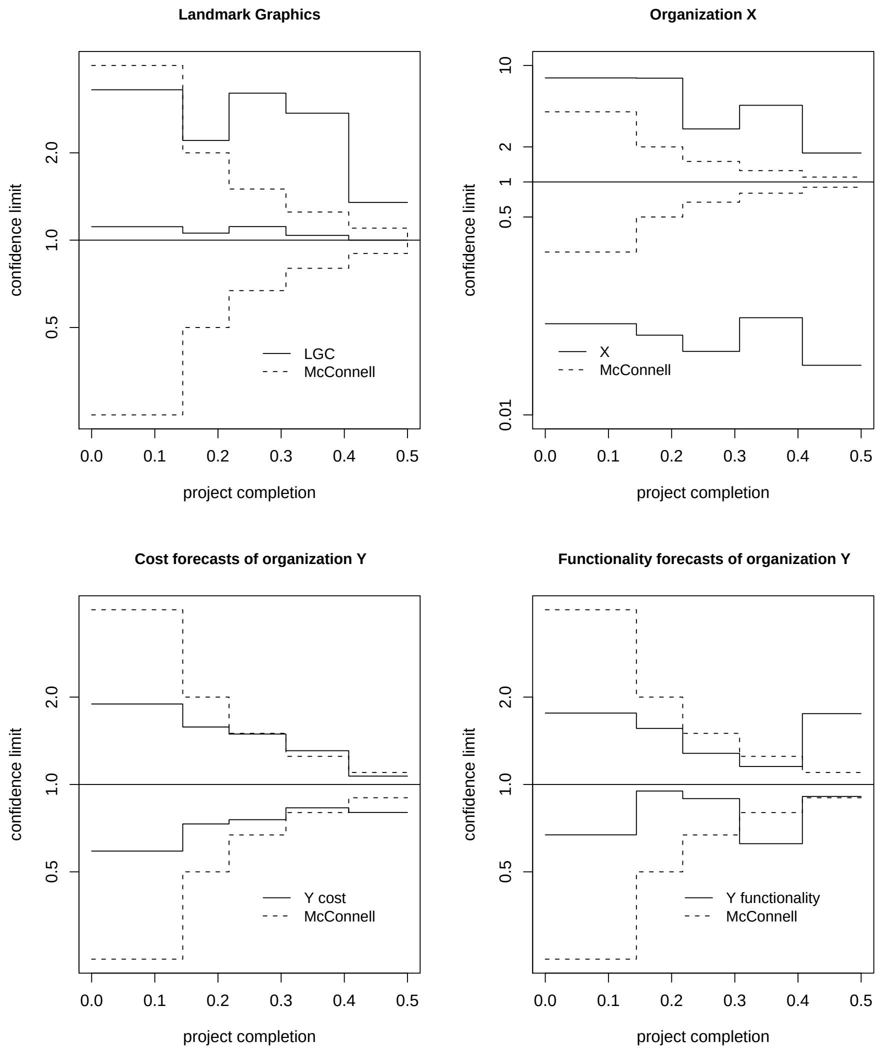
The results are explained by the conditions under which the forecasts are made in each case study. The conditions that apply to organization Y are in alignment with the assumptions underlying Boehm's cone of uncertainty. And indeed, applying the confidence intervals gives reasonable results. In the other 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 Figure 3.14. 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 3.2.

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 forecast 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

Figure 3.14: Visualizing the differences between McConnell/Tockey confidence limits and organization-specific confidence limits.



use. They are only useful in organizations that make forecasts under similar circumstances as McConnell assumed. This implies that there is no bias, forecast 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.

3.5.3 Distribution of ratios

The third approach that enriches forecast information is the distribution of the f/a ratios. This distribution is a generalization of the confidence interval. In that approach, 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 historical 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 historical information and does not require additional assumptions about the data. In an article by Feller [28] and in a book by Fisz [29], 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

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

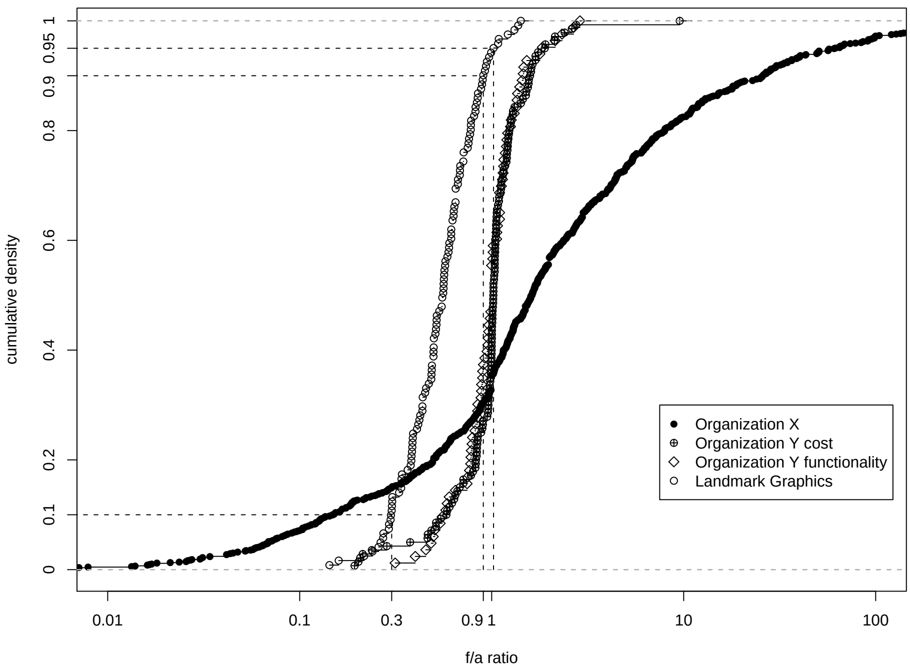
This definition enables deriving the cumulative empirical distribution function of the f/a ratios. In Figure 3.15, we show these functions of the initial forecasts of Landmark Graphics, organizations X and Y. The horizontal axis shows the value of the f/a ratios. The vertical axis depicts the cumulative 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 [103], these figures are easily created using the command `plot(ecdf(dataset))`.

We 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 executive 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 Chapter 2. For instance, a commitment could be made that the project should cost no more than 1 million Euro.

Figure 3.15: 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.



However, the odds of achieving this commitment are slim. The IT executive 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.

3.5.3.1 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 an empirical distribution always fits the actual data set without making additional assumptions about the data. In the literature, a number of alternative distributions are proposed. In an article by Pescio [93], distributions such as the beta, triangular and log-logistic distribution are suggested. Both Putnam [100] and Laranjeira [78] 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 [84]. 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 [77] 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 chapter, 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.

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. Statistically speaking, for low p-values, we reject the null hypothesis. A 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 3.8. 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. We note that removing a couple of projects does not significantly improve the p-values of these tests.

Table 3.8: Test results whether forecast/actual ratios of Landmark Graphics are lognormally distributed.

	Initial		Planning	
test name	p-value	result	p-value	result
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	result	p-value	result
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

Moreover, we also tested the initial f/a ratios of the other organizations as depicted in Figure 3.15 for lognormality. The p-values in Table 3.9 indicate that none of these data sets are statistically lognormal.

Table 3.9: 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	result	p-value	result	p-value	result
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

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 3.10. 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.

Table 3.10: Test results whether ex-ante ratios are lognormally distributed.

	Initial		Planning	
test name	p-value	result	p-value	result
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	result	p-value	result
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

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 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 Figure 3.15, 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 an empirical distribution derived from that data to provide management information.

Summary We illustrated that the distribution of the f/a ratios can provide valuable information to IT executives. 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.

3.6 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 [20]. We also discuss the median benchmark values given by Lister [82]. Although there is no data supporting the benchmark values of Lister, they appear to be an adequate indication of forecast quality at the time of writing this chapter. The values given by Lister compare well with the values we find in our case studies.

Second, we 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.

3.6.1 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. [73]. 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 3.11 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 [20, 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 [83]. 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 [82], but again there were no data available. The values of our case studies were calculated using Formula 3.1. 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.

Table 3.11: 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 [83]	7.0	-	-
Landmark Graphics 2002 [83]	7.6	-	-
Landmark Graphics 2003 [83]	8.4	-	-
Organization Y cost	8.5	36.9	140
Kulk et al. [73]	9.4	-	221

Let us discuss Table 3.11. 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 [82]. That 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 forecast quality, Landmark Graphics seems reasonable, and organization Y has good forecast quality. However, due to the lack of sufficient benchmarks, we are unable to assess how good or how bad the forecast 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 3.1 of Section 3.3.2.1 and represent median values.

3.6.2 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 costs were lower than the actual costs, 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 3.12, 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. We 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 possible bias toward overruns. Overruns take on values anywhere from 0% to infinity. However, underruns only take values between −100% to 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 3.12 are not only found in IT. Similar overruns are reported in other industries, for instance transportation [30]. However, the table does not contain figures reported on by Standish [47, 41, 48, 49, 54]. Namely, these figures were not calculated in the same way as done by the other authors. From the first Chaos report [47], we learned that Standish analyzed overruns of combined challenged and impaired projects. We found that the definitions only account for overruns as will be discussed more extensively in Chapter 4. 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 the 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 [62], to compare these figures of Standish with those of other authors.

Table 3.12: 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
[6] Augustine	1979	-	33%	100 projects	time
[53] Jenkins	1984	33.5%	66.7%	72 projects	cost
[7] Topping	1985	26%	40%	22 projects	effort
[96] Phan	1988	-	33%	191 respondents	cost
[7] Bergeron	1992	-	33%	89 projects	effort
Landmark	2002	79%	105%	121 projects	time
[108] Sauer	2002	-	13%	412 respondents	cost
[108] Sauer	2002	-	20%	412 respondents	time
[108] Sauer	2002	-	−7%	412 respondents	functionality
Org. Y funct.	2005	0%	10%	140 projects	functionality
Org. X	2006	−38%	286%	867 projects	cost
Org. Y cost	2006	−2%	16%	140 projects	cost

In subsequent Standish publications [48, 49, 54], 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 3.13, 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 [34, 35] and Jørgensen [62] state that the figures reported by Standish are inconsistent with other reported figures in the literature. Zvegintzov [131] 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

Table 3.13: 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
[47] Standish	1994	-	189%	365 respondents	cost
[47] Standish	1994	-	222%	365 respondents	time
[54] Johnson/Standish	1996	-	142%	-	cost
[54] Johnson/Standish	1996	-	131%	-	time
[49] Standish	1998	-	69%	-	cost
[54] Johnson/Standish	1998	-	79%	-	time
[49] Standish	2000	-	45%	-	cost
[49] Standish	2000	-	63%	-	time
[54] Johnson/Standish	2002	-	43%	-	cost
[54] Johnson/Standish	2002	-	82%	-	time
Landmark Graphics	2002	85%	113%	114 projects	time
[54] Johnson/Standish	2004	-	56%	-	cost
[54] Johnson/Standish	2004	-	84%	-	time
Org. Y functionality	2005	−17%	−22%	37 projects	funct.
Org. X	2006	173%	1001%	284 projects	cost
Org. Y cost	2006	20%	67%	57 projects	cost

are unreliable and must not be used. In Chapter 4, we will show 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 forecast 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 [8] noted on Augustine's results that high forecasts lead to confronting situations which people want to avoid by giving lower forecasts. Bergeron [7] stated that estimators are aware that lower forecasts have no immediate consequence and that additional budget will often be made available later. Phan [96] found that overly optimistic forecasts are one of the main causes for the project delays and cost overruns. Johnson [41] 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 3.12 and Table 3.13 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 Chapter 4, we will show how to quantify biases 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 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.

3.7 Practitioners' guide

In this chapter, 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 [8] and DeMarco [20]. Over the years, many authors have referred to and used these works as a small subpart of their research. In this chapter, 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 chapter. 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 executive is able to obtain by following the approach proposed in this chapter.

3.7.1 Lessons learned

The main findings of this chapter 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 this chapter allow for comparisons to be made between different 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 executives 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 executives 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 forecasts when compared with the other case studies. This supports DeMarco's statement, that such a department will improve forecast 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 biases.

- The distribution of f/a ratios can provide valuable insight to IT executives. 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 overrun benchmarks reported by others are meaningless, as none account for the effect of biases.

3.7.2 How to quantify forecast quality

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 executives 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 chapter, the following data must be recorded for each newly made forecast: the start date of the project (t_s), the end date of the project (t_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 $(t - t_s)/(t_e - t_s)$. 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 3.1 described in Section 3.3.1.

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 3.4 and 3.5 described in Section 3.3.2.1. 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 3.2 and 3.3 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 3.11 in Section 3.6.

With these steps, an organization is able to assess the quality of IT forecasts. 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 forecast 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 3.5. 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 median quality of the forecasts in terms of EQF is 5. Using formula 3.4 of Section 3.3.2.1, we find that the initial f/a ratio corresponding to that EQF value is $l(0) = 0.6$. This indicates that there is roughly 50% chance 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. It depends on when decisions need to be made or forecasts are made. If you do not have a clear idea on how to form the groups, we suggest to use those we created for organization Y in Table 3.4 in Section 3.5.2.

   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 3.5.2. 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 executive 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 Figure 3.15 in Section 3.5.3. 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 3.2. 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 3.6 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 benchmarks 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 approaches, the first priority is to collect one's own data.

3.8 Conclusions

The quality of IT forecasts 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 chapter, we showed how to quantify the quality and potential bias of IT forecasts to improve decision making using these 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 executive 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

12287 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 forecasts. 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 forecast quality 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 forecast quality 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 approaches 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 forecasts 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. Therefore, these figures are meaningless without further information on the bias of the forecasts.

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

CHAPTER 4

The rise and fall of the Chaos report figures

4.1 Introduction

In the previous chapter we showed how to quantify forecast quality. In this chapter we discuss the consequences of forecast quality and their potential biases on often-quoted rates of project success.

For many years researchers and practitioners are analyzing how to successfully manage IT projects. Among them is the Standish Group, which regularly publishes its findings in their so-called Chaos reports. In 1994, Standish reported a shocking 16% project success rate, another 53% were challenged and 31% of the projects failed outright. In subsequent reports Standish updated their findings, yet the figures remained troublesome. These reports, derived from their longitudinal data, suggest that many efforts and best practices to improve project management hardly help to increase project success. Over the years their figures attracted tremendous attention.

However, there are four major problems with the Standish definitions of a "successful" and a "challenged" project. First, their definitions are misleading, since they are solely based on estimation accuracy of cost, time and functionality. Second, the definitions are one-sided, causing highly underrated success rates. Third, steering on the definitions causes the inverse effect: they pervert estimation accuracy. Fourth, Standish's definitions lead to meaningless figures because they averaged numbers with an unknown bias, which are introduced by different underlying estimation processes.

Other authors, e.g. Glass [34, 35] and Jørgensen et al. [62] indicated that the only way to assess the credibility of the Chaos results is to use Standish's data and reiterate their analyses. But there is another method. Obtain your own data and reproduce their research to assess its validity. We will demonstrate this by applying *their* definitions to *our* extensive data consisting of 5457 forecasts of 1211 real-world projects of in total hundreds of millions of Euros. Moreover, we will propose new definitions that do account for potential biases.

4.2 Misleading definitions

We recall Standish's definitions in order to show they are misleading. In 1994 the Standish Group published the first in a series of reports, called the Chaos report [47]. This report is a summary of their research findings, which aimed to investigate causes of software project failure and to find the key ingredients to reduce such failures. Moreover, they intended to identify the scope of software project failures. To achieve this latter goal, Standish defined three project categories. We recall them verbatim from their report [47].

- 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.

To find answers to their research questions Standish sent out questionnaires. Their total sample size was 365 respondents representing 8380 applications. Based on the responses Standish published overall percentages for each of the above categories. Standish updated their figures in subsequent years as shown in Table 4.1. These figures were published in various white articles [41, 47, 48, 49, 50] and most recently on their website [51].

Table 4.1: Standish project benchmarks over the years.

year	success	challenged	failed
1994	16%	53%	31%
1996	27%	33%	40%
1998	26%	46%	28%
2000	28%	49%	23%
2004	29%	53%	18%
2006	35%	46%	19%
2009	32%	44%	24%

The figures indicate large problems with software engineering projects, and as such have had an enormous impact. They suggest that the many efforts and best-practices to improve the way we develop software seem hardly successful. The numbers have been widely cited in scientific articles and the media. Many authors use the figures to point out that project management of software development is in a crisis. The numbers even found their way to a report for the President of the United States [63], to substantiate the claim that both software products and processes in the US are inadequate.

The impact of the figures and their widespread use, indicate that thousands of authors have accepted the Standish findings. They are perceived to be impeccable and unquestionable. However, the Standish definitions of "successful" and "challenged" projects are problematic. If you read the definitions twice, you notice that Standish defines a successful project solely by adherence to an initial forecast of cost, time and functionality. If you read the definitions thrice, you will realize that the latter is only defined by the *amount* of features and functions, not functionality itself. Indeed, Standish discussed this in their report [47]: "For challenged projects, more than a quarter were completed with only 25% to 49% of originally-specified features and functions." So, Standish defines a project to be a success based on how well it did with respect to its original estimates of the amount of cost, time and functionality. Therefore, the Standish successful and challenged definitions are equivalent to the following.

- Resolution Type 1, or project success: The project is completed and the forecast-to-actual ratios of cost and time are ≥ 1 and the f/a ratio of the amount of functionality is ≤ 1.

- Resolution Type 2, or project challenged: The project is completed and operational but for cost and time both $f/a < 1$ and for the amount of functionality $f/a > 1$.

The reformulated definitions illustrate that they are only about estimation deviation.

We note that Jørgensen et al. [62] show that the definitions do not cover all possibilities. For instance, a project which is within budget and time, but has less functionality does not fit any category. In this chapter we assume a project that does not comply to one or more of the criteria for success belongs to the challenged project category.

The Standish success measure is calculated by counting the number of projects that have an *initial* forecast larger than the actual for cost and time, and smaller for functionality. This is divided by the total number of projects to arrive at their success rates. In fact, Standish Group equals their success measure as a measure of estimation accuracy of cost, time and functionality.

In reality, the part of a project's success that is related to estimation deviation is highly context dependent. In some contexts, for instance, 25% estimation error does no harm and does not impact what we normally would consider project success. While in other contexts, for instance, only 5% overrun would cause much harm and make the project "challenged". In that sense, there is no way around inclusion of more context (or totally different definitions) when assessing successful and challenged projects. However, the Standish definitions do not consider context, such as usefulness, profit and user satisfaction of a software development project.

Therefore, this illustrates the first problem with the definitions. They are misleading since they are solely based on estimation accuracy of cost, time and functionality. But they call them successful and challenged projects, suggesting much more than these estimation deviations.

4.3 Unrealistic rates

Now we address the second problem of the Standish definitions. If we take for granted that their definitions are misnomers, the next issue we address is whether their estimation accuracy definitions are sound. This is not the case. Their measures are one-sided since they neglect underruns for cost and time and overruns for the amount of functionality.

In Chapter 3 we assessed estimation accuracy with two tools. The first is derived from Barry Boehm's now famous *cone of uncertainty* [8] and the second is Tom DeMarco's Estimating Quality Factor (EQF) [20]. The cone of uncertainty is a plot which depicts forecast-to-actual ratios against project progression. This plot gives information on how the forecasts are made, what deviations they contain and whether there are biases.

With the EQF we are able to quantify the quality of the forecasts. The EQF is a time-weighted estimation accuracy measure. The higher the EQF value of a forecast, the higher its quality. An EQF value of 5 means the time-weighted forecasts of a single project deviate on average 1/5 = 20% from the actual.

We applied Boehm's and DeMarco's work to our own extensive data. We detected large biases, which the organizations were not aware of. Below we recall two data sets from Chapter 3 to prove our claim that the one-sided definitions lead to unrealistic rates.

Cost The first case study we recall is organization Y, from which we obtained data on 140 software development projects. In total 667 forecasts were made for the total costs of these projects. We analyzed the data in Section 3.4.4. To ease readability, we summarize the findings below and recall the f/a plot as shown in Figure 4.1. The horizontal axis represents project progression. The start of a project is depicted at zero and project completion is represented by one. The vertical axis shows the value of the f/a ratio. For instance, a data point at project completion 0.2 and f/a ratio 2 indicates a forecast was made when the project was completed for one fifth. This forecast was two times the actual, meaning the project turned out to be 50% of the estimated cost.

The f/a ratios in the figure resemble Boehm's conical shape with the forecasts centered around the actual value. This is supported by a median f/a ratio of 1.0. The quality of the forecasts is relatively high with a median EQF value of 8.5. This indicates that half of all projects have a time-weighted average deviation of 12% or less from the actual. It turned out that in this organization the forecasts were assessed by an independent metrics group. This group made its own cost calculations next to those of project managers. If large discrepancies arose these

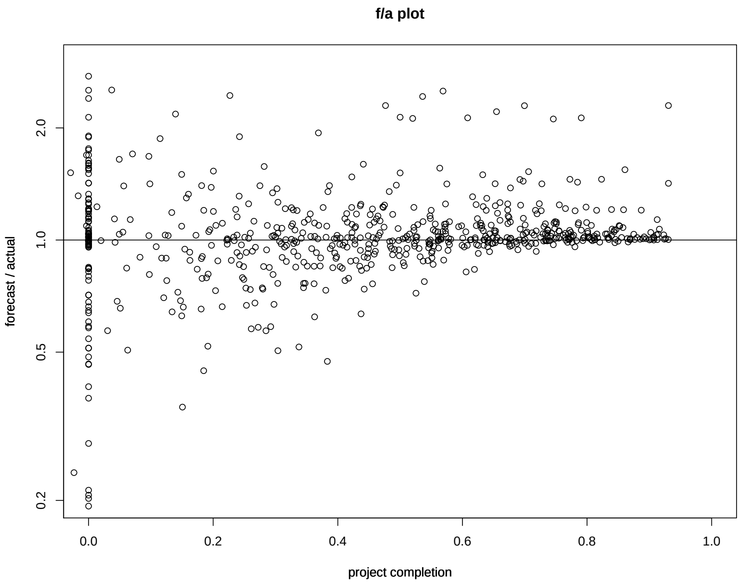


Figure 4.1: 667 f/a ratios for 140 project costs of organization Y.

needed to be resolved before any budget was approved. This caused forecasts to aim at predicting the actual value. Yet, even though this organization is accurate in its cost forecasts, when we apply the Standish definitions to the initial forecasts we find only a 59% success rate.

Functionality From the same organization Y we also obtained data of 83 software development projects in the period 2003 to 2005. In total 100 forecasts were made for the functionality of the projects counted in function points [32].

The functionality f/a plot in Figure 4.2 shows a situation similar to the f/a ratios for the costs. Also here the bias is negligible based on the figure and a median f/a ratio of 1.0. Except for some outliers, the f/a ratios converge to the actual value. The functionality forecasts have a median EQF value of 6.4. This means that the function point forecasts of half of all projects have a time-weighted average deviation of 16% or less from the actual amount.

The functionality of the projects was counted by multiple experienced function point counters. As they were not involved with the execution of the projects, their only incentive was to predict the actual value. However, despite the accuracy of the forecasts, when we apply the Standish definitions to the initial forecasts we find only a 55% success rate.

Combined 55 software development projects contained forecasts and actuals of both cost and functionality. There were in total 231 cost forecasts and 69 function-

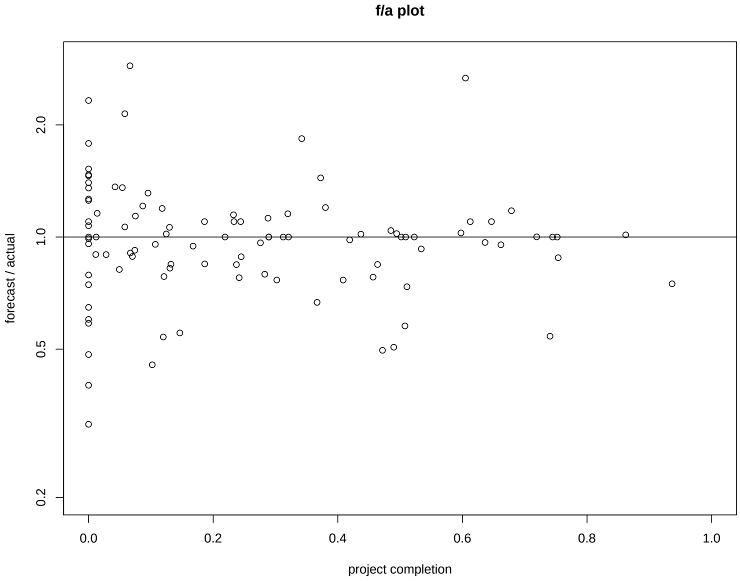


Figure 4.2: 100 f/a ratios for 83 project function points of organization Y.

ality forecasts. Both cost and functionality forecasts were unbiased and converged to the actual value. The cost forecasts have a median EQF of 9.0 and the functionality forecasts 5.0. Thus, half of the projects have a time-weighted average deviation of 11% for the costs and 20% deviation for functionality.

We applied the reformulated Standish definitions to the initial forecasts of the combined data. Even without taking failed projects and the time dimension into account, the best-in-class organization Y obtains a success rate of 35%. Yet, the median EQF of both initial forecasts of costs and functionality is 6.5, showing that half of their projects have an average time-weighted deviation of only 15% to the actuals. If this organization in two dimensions is already so unsuccessful according to Standish, it is hardly surprising they found only a 16% success rate in their first report [47].

These case studies show that the best-in-class organization Y obtains unrealistically low success rates for the individual cost and functionality forecasts due to the one-sidedness of the definitions. Combining these already low rates further degrades the success rate. Clearly, the Standish success rates do not give an accurate indication of true estimation accuracy of cost and functionality in the case of an unbiased best-in-class organization, proving our second claim.

4.4 Perverting accuracy

In this section we discuss the third problem with the Standish definitions. Namely, steering on them causes large cost and time overestimations (and large functionality underestimations) perverting estimation accuracy, rather than improving it. We will illustrate this by recalling another case study from Chapter 3 to prove that the one-sided definitions reach the opposite of estimation accuracy.

Cost In Section 3.4.3 we analyzed data from a large multinational organization X comprising of 867 IT-intensive projects. In total 3767 forecasts of the costs of these projects were made. To ease readability, we summarize the findings below and recall the f/a plot as shown in Figure 4.3.

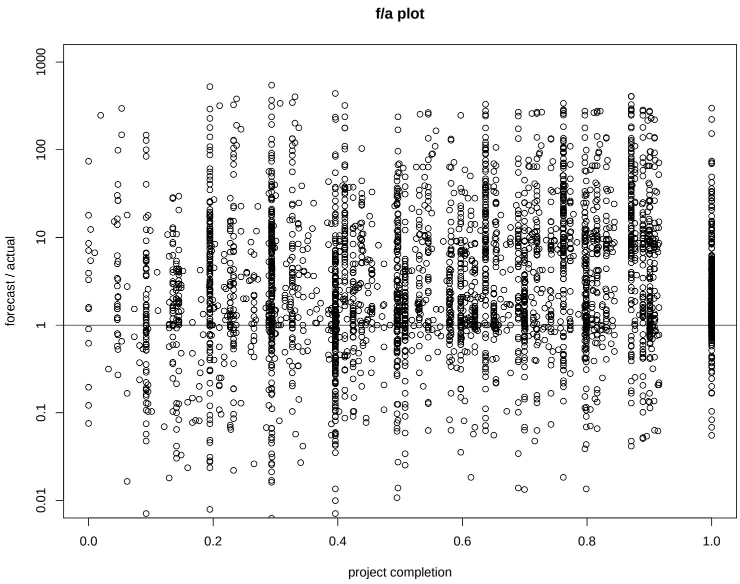


Figure 4.3: 3767f/a ratios for 867 project costs of organization X.

The f/a ratios in Figure 4.3 show that the forecasts made in this organization are in general higher than the actual. Also, the data does not show a conical shape as we would expect from Boehm's *cone of uncertainty*. Projects even had surplus budget *after* completion. After discussion with the organization, we found that this organization steered on Standish project success indicators. They adopted the Standish definitions to establish when projects were successful. This caused project managers to overstate budget requests to increase their safety margin for success. However, this practice perverted the quality of the forecasts. The quality is low with a median EQF value of 0.43. So, 50% of all projects have a time-weighted average deviation of 233% or more from the actual.

This case study proves the third major problem of the Standish definitions. It shows that steering on the Standish definitions causes the inverse effect. The one-sided definitions lead to large overestimates to increase the margin of success. Yet, this causes the estimation accuracy to deteriorate.

4.5 Meaningless rates

In this section, we will address the fourth major problem of the Standish figures, namely that they are meaningless. The second organization showed that large biases occur in practice. Even if you do not steer on Standish KPIs there are biases. We will show this by recalling another case study from Chapter 3. With all case studies together we will prove that the Standish figures are highly problematic. Namely, without taking forecast biases into account, it is almost impossible to make any general statement about estimation accuracy across institutional boundaries.

Time In Section 3.4.2 we analyzed data from Landmark Graphics that we obtained from Todd Little. The data consists of 121 software development projects with 923 distinct forecasts that predict the duration of these projects. Again, to ease readability, we summarize the findings below and recall the f/a plot as shown in Figure 4.4.

Most forecasts made in this organization are lower than the actual. So, projects take longer than initially anticipated. The median EQF value is 4.7. This means that half of all projects have a time-weighted average deviation of their forecasts of 21% or less from the actual. The institutional bias of Landmark Graphics was to forecast the minimum value instead of the actual value. This caused most forecasts to be lower than the actuals.

4.5.1 Applying Standish's definitions

In two of the three organizations the forecasts were significantly biased. In the first organization we determined the institutional bias was negligible. It turned out that the cost and functionality forecasts were assessed by an independent metrics group. In the second organization the forecasts were much higher than the actual values as large safety margins were taken into account. In the last organization most forecasts were lower than the actual values since they predicted the minimal time required to finish the project.

To illustrate how the forecast biases, introduced by the different underlying estimation processes, affect the Chaos report figures, we applied Standish's definitions to the cases described in this chapter. Since Standish deals with initial forecasts, we also used the initial forecast of each project. We note that this is a subset of all data points shown in the f/a plots above.

Also, our resulting figures are an upper bound for the Chaos successful project figures. We explain why. First, our figures do not incorporate failed projects. If

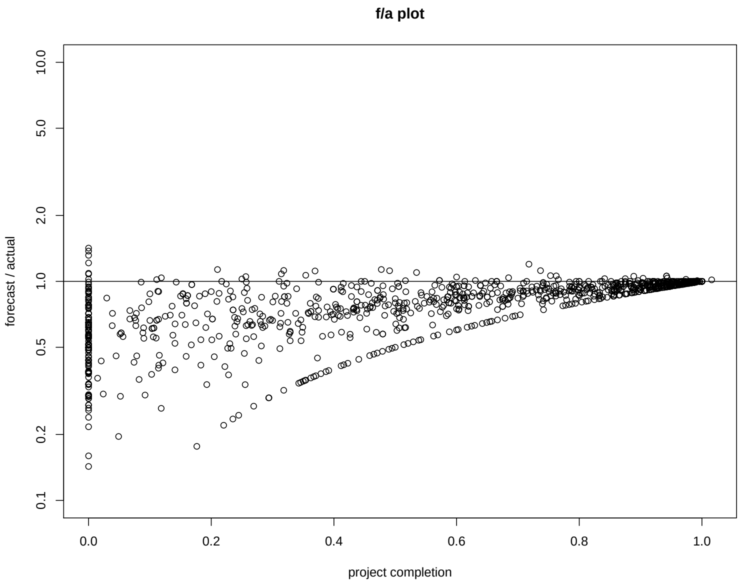


Figure 4.4: 923 f/a ratios for 121 project durations of Landmark Graphics.

failed projects would be taken into account, the success rates of our case studies will always be equal or lower than the current percentages.

Second, in each of the case studies we present data of only cost or time or functionality and once both cost and functionality. In our analysis we assume the remaining dimensions to be 100% successful, meaning our percentages are only influenced by one or two dimensions. If data for all three dimensions (cost, time and functionality) of these organizations is available and taken into account, again the success rates will always be equal or lower than the successful percentages calculated for only one or two dimensions. Still, these rates suffice to prove that Standish's success and challenge rates do not at all reflect the reality, since it is bad practice to process erroneous numbers of different dimensions to reach a proper result.

Table 4.2 shows the numbers calculated according to Standish's definitions for our case studies plus a fictitious organization having the opposite bias of Landmark Graphics. The table provides an interesting insight in the Standish figures. It shows that organization X is very successful compared to the other case studies. Nearly 70% of the projects is successful according to the definitions of Standish. On the other end Landmark Graphics has only a 6% success rate. Organization Y is in between with a success rate of 59% for costs, 55% for functionality, and 35% for the combination.

However, the f/a plots and their median EQFs clearly show that this is far from reality. Landmark Graphics's and organization Y's initial forecasts deviate

Table 4.2: Comparing Standish success to real estimation accuracy.

source	Standish' success	Standish' challenged	estimation accuracy measured in 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

much less from their actuals than in the case of organization X. For, that organization overestimates tenfold times up to hundredfold times as shown in Figure 4.3. Also, the estimation quality of the other organizations outperform that of organization X, which is illustrated by the median EQF of their initial forecasts: 2.3 for Landmark Graphics, 6.4 for costs and 5.7 for functionality of organization Y versus 1.1 for organization X. So, half of Landmark Graphics's initial forecasts only deviate 43% from the actual value, 16% for costs and 18% for functionality of organization Y versus 91% for organization X. Still organization X is considered highly successful compared the other organizations according to Standish.

To further illustrate how easy it is to become highly successful in Standish's terms, we also presented 1/Landmark Graphics. This fictitious organization resembles the exact opposite of Landmark Graphics. That is, the deviations to the actuals remain the same, but an overrun becomes an underrun and vice versa. Suddenly, 1/Landmark Graphics becomes highly successful with a success rate of 94%. Thus, with the opposite institutional bias, Landmark Graphics would improve their Standish success rate from 6% to 94%.

These case studies show that the Standish figures for individual organizations are not reflecting the reality, and are highly influenced by forecasting biases. Since the underlying data has an unknown bias, any aggregation of that data is unreliable and meaningless. So, without accounting for biases, Standish cannot average the numbers to reach a proper result.

The influence of biased forecasts on the Standish figures is not only evident from our figures. Standish, by means of their chairman Jim Johnson, clearly indicates it is easy to manipulate their figures [54]:

> In 1998, they [the respondents] had changed their [estimating] process so that they were then taking their best estimate, and then doubling it and adding half again.

Johnson made this statement with respect to the drop in their reported average cost overruns between 1996 (142%) and 1998 (69%). In that article Johnson continues that he does not believe this change of process is the cause of the drop. However, our case studies show that forecast biases have a giant influence on such figures. Therefore, we believe the drop in the reported cost overruns is most likely caused by this change in the estimating process.

We evidently showed that by ignoring the potential bias and forecast quality, the figures of Standish Group do not give an adequate indication what, according to their definitions, constitutes a successful or challenged project. Since some organizations have the tendency to overestimate and others to underestimate, their success and challenge rates are in fact meaningless as they do not account for these clearly present biases. With that we proved our fourth claim.

4.6 Discussion

This chapter is not the first to challenge the credibility of the Chaos report figures. A number of authors also “questioned the unquestionable” [34] and raised concerns about the validity of the figures [35, 62, 131].

For instance, Zvegintzov [131] placed low reliability on information where the actual data and data sources are kept hidden. He argued since Standish has not explained, for instance, how the organizations surveyed were chosen, what survey questions were asked or how many good responses were received, there is little to believe.

Also, Glass [34, 35] felt the figures do not represent reality. Without plentiful successful projects, he asserts, it is not possible to live in what some describe as the current computer era.

Moreover, Jørgensen et al. [62] expressed their doubt about the numbers. They unveiled a number of issues with Standish’s definitions and argue the resulting figures are therefore unusable. For instance, they argue that the definitions of successful and challenged projects focus on overruns and discard underruns. We additionally showed that one-sidedness leads to highly unrealistic rates, and steering on them even perverts estimation accuracy.

Despite the valid questions raised by our predecessors, to date no-one was able to definitely refute the credibility of the Standish figures. We now clearly showed Standish’s definitions suffer from four major problems which undermine the validity of their figures.

We communicated our findings [26] to Standish Group, who responded by means of their chairman Jim Johnson. He wrote: “All data and information in the CHAOS reports and all Standish reports should be considered Standish opinion and the reader bears all risk in the use of this opinion.” We fully support this disclaimer, which to our knowledge was never stated in the Chaos reports.

4.7 Proposed definitions

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 underruns for cost and time and overruns for functionality. Also, they should preferably account for the biases of the forecasts. We 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 Chapter 3: 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 3.2.1, 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. We use the bandwidths of this reference cone 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 Figure 4.5. 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 3.4 and 3.5 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 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 as described in Section 3.6.1. 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.

We note that there is no incongruence with what we discussed in Section 3.3.2.2. 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.

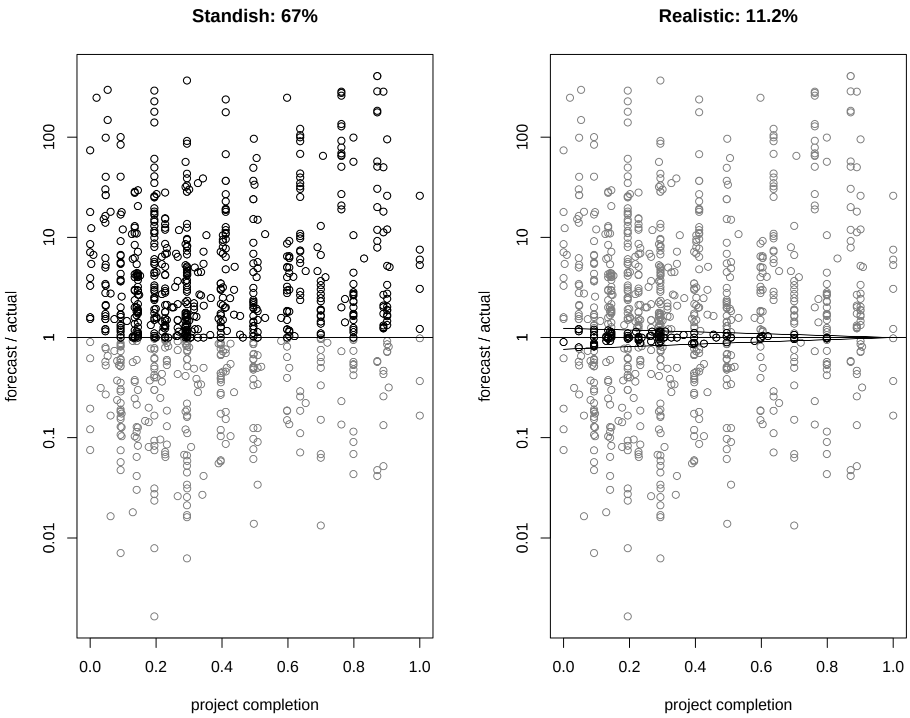


Figure 4.5: Real-world data illustrates that Standish' project success definition is not realistic. Example was drawn using data from organization X.

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 3.2.1, we assumed that the forecasts had no bias and the ex-post part was known and used. In Section 3.3.2.1, we showed how these conditions lead to a mathematical description of the lines, resulting in Formulas 3.4 and 3.5. 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 biases 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 biases involved in forecasting. Below, we show the different types of comparisons by applying them to our case studies.

4.7.0.1 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 3.2.1 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 3.3.2.1, we showed this results in reference lines as given by Formulas 3.4 and 3.5. With the EQF value chosen to 8.5, the reference cone that we apply is thus given by:

$$
\begin{aligned}
l(x) &= x + (1 - \tfrac{2}{8.5}) \cdot (1 - x) &&= \frac{4}{17}x + \frac{13}{17} \\
u(x) &= x + (\tfrac{2}{8.5} + 1) \cdot (1 - x) &&= \frac{21}{17} - \frac{4}{17}x.
\end{aligned}
$$

Table 4.3: 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%

Table 4.3 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 forecasts are 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.

We 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. 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 4.3 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 4.3 seem to be low, but this is because we set the forecast 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 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 biases. This is done in the second type of comparison that we describe below.

4.7.0.2 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 biases found in the case studies. In Section 3.3.2.1, 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 3.3.2.1, 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 X.

Another cone condition we used to compute the reference cone of Section 3.3.2.1, 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 bias. Below, we will show how to make the calculations given the cone conditions of our case studies. In the calculations we again choose an EQF value of 8.5 as the quality the forecasts should adhere to.

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 a since the ex-post part is known and used. Thus the new reference point, formerly the actual, at time x becomes $0.56 \cdot a + 0.44a \cdot x$. We assume the ex-post part growth is constant, leading to an ex-post part of ax. This makes the ex-ante part $0.56a + 0.44ax - ax = 0.56a - 0.56ax$. We are able to predict this within $1/c_1$ to c_2 times leading to, for the lower limit line, $1/c_1 \cdot (0.56 - 0.56x)$ and, for the upper limit line, $(0.56 - 0.56x) \cdot c_2$. We reiterate the procedure with the EQF as we did in Section 3.3.2.1, but now the area below the reference point is not a, but $\int_{x=0}^{1} 0.56a + 0.44axdt = 0.78a$. Then the reference lines that account for the biases in Landmark Graphics are given by:

$$\begin{aligned} l(x) &= x + (1 - \tfrac{0.78}{0.28 \cdot \mathrm{EQF}_l}) \cdot (0.56 - 0.56x) &\approx 0.376 + 0.624x \\ u(x) &= x + (1 + \tfrac{0.78}{0.28 \cdot \mathrm{EQF}_u}) \cdot (0.56 - 0.56x) &\approx 0.744 + 0.256x. \end{aligned}$$

To illustrate that the reference cone now accounts for the bias of Landmark Graphics, we depict in Figure 4.6 the same f/a plot as used in the case studies. However, in this case the above reference cone that accounts for the 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 x 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,

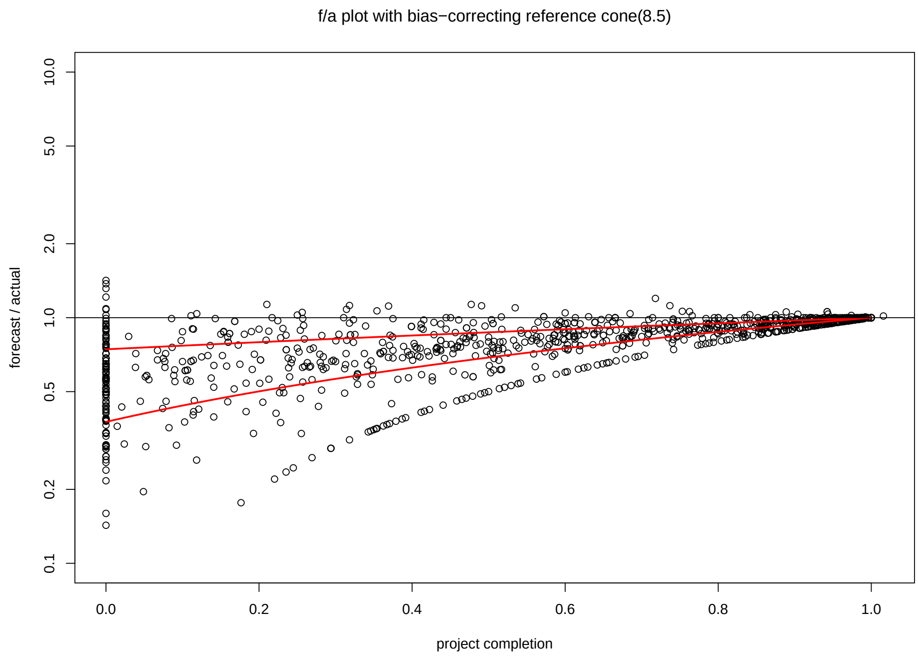


Figure 4.6: f/a plot of Landmark Graphics of all data with a bias-correcting reference cone.

in this case $1.625a - 0$. We reiterate the procedure and find that the reference lines that account for the biases in organization X are given by:

$$\begin{aligned} l(x) &= 1.625 - \tfrac{1.625}{\mathrm{EQF}_l} &\approx 1.434 \\ u(x) &= 1.625 + \tfrac{1.625}{\mathrm{EQF}_u} &\approx 1.816. \end{aligned}$$

Organization Y cost For organization Y, we found the cost and functionality forecasts to be in line with the cone conditions we described in Section 3.2.1. 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 3.4 and 3.5 we found in Section 3.3.2.1. 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 reference point will gradually decrease to a, thus the new reference point at time x becomes $1.02a - 0.02ax$. We assume the ex-post part growth to be constant, thus the ex-post part is equal to ax. This makes the ex-ante part $1.02a - 0.02ax - ax = 1.02a - 1.02ax$. We are able to predict this within $1/c_1$ to c_2 times leading to bounds of the ex-ante part of $1/c_1 \cdot (1.02a - 1.02ax)$ and $(1.02a - 1.02ax) \cdot c_2$. 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:

$$\begin{aligned} l(x) &= x + (1 - \tfrac{1.01}{0.51 \cdot EQF_l}) \cdot (1.02 - 1.02x) &&\approx 0.782 + 0.218x \\ u(x) &= x + (1 + \tfrac{1.01}{0.51 \cdot EQF_u}) \cdot (1.02 - 1.02x) &&\approx 1.258 - 0.258x. \end{aligned}$$

We note that these reference lines only differ slightly from the theoretical lines given by Formula 3.4 and 3.5. 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 3.4 and 3.5 we found in Section 3.3.2.1. Therefore, the reference lines accounting for the biases of the functionality forecasts are given by:

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

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:

$$\begin{aligned} l(x) &= x + (1 - \tfrac{1.02}{0.52 \cdot EQF_l}) \cdot (1.04 - 1.04x) &&\approx 0.800 + 0.200x \\ u(x) &= x + (1 + \tfrac{1.02}{0.52 \cdot EQF_u}) \cdot (1.04 - 1.04x) &&\approx 1.280 - 0.280x. \end{aligned}$$

The reference lines of the functionality forecasts is given by:

$$\begin{aligned} l(x) &= x + (1 - \tfrac{1.005}{0.505 \cdot EQF_l}) \cdot (1.01 - 1.01x) &&\approx 0.774 + 0.226x \\ u(x) &= x + (1 + \tfrac{1.005}{0.505 \cdot EQF_u}) \cdot (1.01 - 1.01x) &&\approx 1.246 - 0.246x. \end{aligned}$$

Applying organization-specific cones With the different 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. Figure 4.7 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 4.4, we summarize

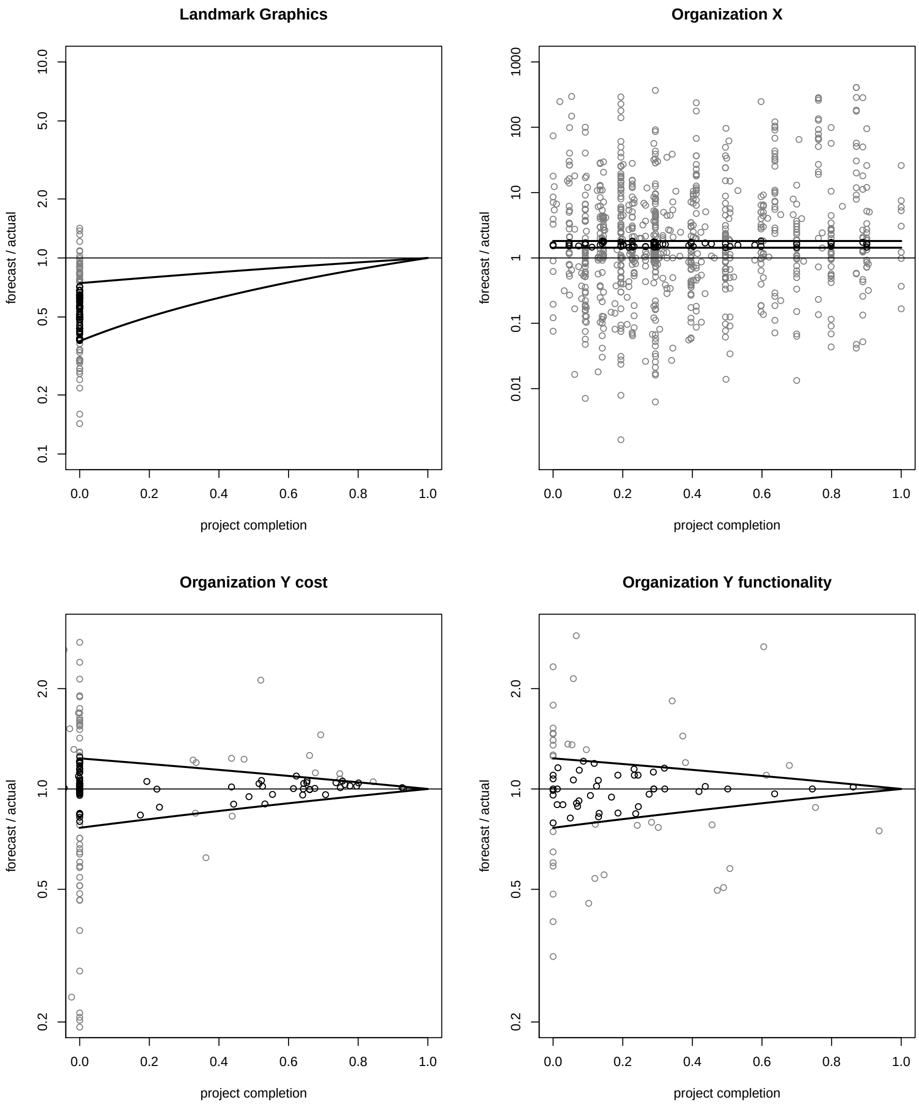


Figure 4.7: True plan accuracy comparison of the three organizations while accounting for the biases.

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

Table 4.4: 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%

rate of 58%. Although they predict the earliest possible date instead of the actual, they are capable of making very reasonable forecasts of 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 3.5, 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 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 bias of the organization. Similar to the other comparisons, we find this organization to have a reasonable forecast quality for both cost and functionality when compared with the other organizations.

Since the figures of Table 4.4 do account for the 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.

Conclusions 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 forecast quality between organizations.

CHAPTER 5

Quantifying forecast quality of IT business value

5.1 Introduction

In this chapter we will quantify the forecast quality of IT business value. We discuss common economic indicators often used to determine the business value of proposals. We elaborate on one of these indicators, the Net Present Value (NPV). For this indicator, we will assess the forecast quality. Since the overall forecast quality of the NPV is determined by the forecast quality of its components, such as benefits and costs, their forecast quality is assessed as well.

We make use of the method developed in Chapter 3 to quantify and visualize the quality of IT forecasts. That method assumes positive values, like budgets, durations, and size in function points or lines of code. However, common economic indicators such as the NPV can take any value. By applying the method of Chapter 3 to real-world data of NPV calculations, it turned out the model was inadequate to deal with zero's and negative values. Indeed, it is natural for an NPV to be zero valued or negative.

Next to that, the actuals of all entities are only known with certainty after some time. For instance, the actual project costs are only known after the project is completed. Or, the actual NPV of an asset is known after the economic life span of an investment, for instance, 5 years after the initial forecast. However, to support decision making, we prefer recent information about the current forecasting practice; something that the quality of 5-year old forecasts will not provide for. Moreover, in case of some entities, for instance, the NPV, it is possible the actual will never be objectively measured. Therefore, we cannot always compare forecasts with actuals, making other reference points necessary.

For these reasons, we need adaptations of the forecast assessment method developed in Chapter 3. We note that the existing method is perfectly fit for positively valued indicators, so that method is not at all outdated by this chapter.

But for business value a more sophisticated method is necessary.

Therefore, to assess the quality of NPV forecasts, in this chapter we extend and generalize the existing method that is intuitive for positively valued forecasts and actuals. We will develop a method that incorporates other reference points than the actual and allows for negative valued reference points or forecasts. Moreover, we resolve visual limitations in asymptotic cases.

At first sight, these issues appear trivial, but they are not. Adjusting the model to accommodate for these issues, entails an extensive and detailed discussion. For readers interested in the use of the model rather than the elaborate discussion of the adjustments, it is possible to skim the generalization of the method.

In this chapter, we assess the quality of NPV forecasts, benefits and costs using real-world data of 102 IT assets. We obtained this data from a telecommunication organization, Z, that structurally makes business cases for its IT investments. Moreover, this organization uses procedures to evaluate completed investments. The data represents an NPV value of 1812 million Euro, with discounted benefits of 4714 million Euro and an investment value of 173 million Euro. Combined with data from other organizations, this chapter discusses 1620 IT assets with an investment cost of 1232+ million Euro for which in total 6513 forecasts are made.

For the data of organization Z, we determine the accuracy of the forecasts and check for potential biases. We found that the quality of the forecasted NPVs is lower than the forecasted benefits, which is again lower than the forecast quality of the costs. Also, there turned out to be a significant difference in forecast quality between assets classified as Cost Reduction or New Product Development. The NPV, benefit and cost forecasts of Cost Reduction assets were more accurately predicted than that of the New Product Development assets. Moreover, whereas the forecasts of Cost Reduction assets showed no biases, the New Product Development forecasts of benefits and costs were generally overestimated in the real-world case study.

We also performed a sensitivity analysis to investigate the impact on the quality of an asset's forecasted NPV when the forecast quality of benefits or costs improves. Counterintuitively, it turned out that if the quality of the cost forecasts would improve, the overall quality of the NPV predictions would degrade. This is caused by the bias toward overestimation of both benefit and cost forecasts. The overestimation of the benefits is compensated for by an overestimation of the costs. This illustrates that increasing control over the costs without measures to ensure the quality of the benefits, may yield a lower forecast quality of the overall NPV.

Finally, we will illustrate how it is possible to use the quantified forecast quality further to enhance decision information. This is done by describing two basic simulation examples. The first example shows how to acquire additional information for rationing the capital budget over various project proposals. The second example provides insight in the forecasted business value that will be generated from project proposals when accounting for the forecast quality and bias.

Organization of this chapter In Section 5.2 we discuss economic indicators that are used to quantify the business value of project proposals. More particularly, we cover in detail the well-known Net Present Value.

For the interested reader, in Section 5.3 we extend the method developed in Chapter 3 to assess IT forecast quality and make it fit to assess economic indicators. We generalize the method to allow for different reference points and negatively valued entities. We assess the impact of the generalization using data of four organizations.

Section 5.4 explores the real-world NPV data we obtained from the telecommunication organization, Z. We compare this data to benchmark data from the literature. We also investigate the data for possible heterogeneity.

After the exploration, we commence with the assessment of the forecast quality of the NPV, benefits and costs in Section 5.5. Moreover, we perform a sensitivity analysis to investigate the impact on the quality of NPV forecasts when the forecast quality of benefits or costs improve.

Section 5.6 illustrates how the quantified forecast information is used to enhance decision information. This is illustrated by using two simulation examples.

In Section 5.7 we discuss limitations of our research. Finally, Section 5.8 concludes this chapter.

5.2 Asset business value

Pisello et al. [98] state that the Chief Information Officer (CIO) must become the Chief Financial Officer (CFO) of IT. To do so, the CIO must manage IT projects in such a way that their business value is maximized. But how to determine the business value of IT assets?

To specify the business value of IT assets, in their decision making executives are provided with information about the advantages, disadvantages and risks of each project proposal. Project proposals contain two kinds of information: qualitative and quantitative. Qualitative arguments are, for instance, corporate social responsibility or strategic alignment. According to multiple surveys [5, 31, 97], many organizations consider these arguments as important criteria for project selection.

The contribution of this chapter is to the quantitative kind of information. This part of the available information consists, among others, of forecasts of quantified benefits and costs. Often, these predictions are summarized using economic indicators. Well-known examples of such indicators are the Net Present Value, the Internal Rate of Return, the Return on Investment, and the Payback Period. We describe these indicators briefly below.

- Net Present Value (NPV). The Net Present Value is a summation of the predicted monetary benefits and costs of a project discounted to current value. If the NPV is positive this indicates the project is estimated to provide for monetary gain given the discount rate. If the NPV is zero, the project is a neutral investment: it generates enough discounted benefits to cover the discounted costs. If the NPV is negative the project proposal is expected to result in a monetary loss given the rate used. In the next subsection we will discuss the NPV in more detail.

- Internal Rate of Return (IRR). The Internal Rate of Return is the discount rate for which the NPV is equal to 0. The IRR is compared to the rate we would receive for similar investments, which is also known as the opportunity cost of holding capital. If the IRR is lower than the opportunity cost of capital, the investment should not be funded from a quantitative point of view. If the IRR is high compared to other rates, the project will generate more yield than other similar projects in the market.

- Return on Investment (ROI). In a book by Bierman [40], the Return on Investment is defined as an average income after depreciation divided by its investment. The ratio is also known as rate of return and there are many different ways of computing it. In all cases the higher the ratio the more profitable the project.

- Payback Period (PBP). The Payback Period is the amount of time needed after project completion to generate an equal amount of cumulated cash flows to cover the initial investment. Often these calculations are not discounted to today's currency. The shorter the Payback Period, the sooner the initial investment is paid back.

These economic indicators give IT executives an indication of the business value of the project proposals. It is possible to use them to make predictions and to evaluate the final realization. Therefore, we are able to use these indicators to answer our question: how do we know that the quantitative business value forecasts are accurate, unbiased and reliable enough to support the decision making process?

We will answer this question by assessing the forecast quality of one of the indicators, the NPV. We will only use the NPV, simply because our case study of organization Z uses this indicator to support its decision making. However, the generalized method we will propose later on, is applicable to the other indicators as well.

But, before answering the question, we need to better understand the NPV. First, we discuss the limitations of the different indicators with respect to each other. Each indicator has its theoretical and/or practical disadvantages. Then, we will identify the different components of the NPV and discuss each of them.

5.2.1 Indicator limitations

We briefly discussed the NPV, IRR, ROI and PBP. Brealey et al. [11] state that theoretically the use of the NPV leads to better investment decisions than these other well-known indicators. They argue that in some situations the IRR leads to different results than the NPV. We note that although both the IRR and NPV are derived from the same formula, the way they are derived can cause different outcomes. One situation in which the IRR is ineffective, is when there is a mixture of multiple positive and negative cash flows. For instance, consider an asset that, besides the initial investment, has some final fee to be paid at the end of its life span. In that case it may occur that there are two realistic IRR ratios for which the NPV is 0. This is also observed by, among others, Lorie et al. [85]. Moreover, Brealey et al. discuss that the IRR does not discern between borrowing or lending money, and has problems when the opportunity cost of capital is different over several years.

Brealey et al. also discuss that the Return on Investment leads to worse decisions, since it does not account for the timing of the cash flows. Since the cash flows are generally averaged, the indicator places no importance on whether the cash flows are earned in the first year or the last year. Yet, in reality this is a crucial aspect of the investment decision.

Finally, Brealey et al. argue that the Payback Period ignores all cash flows generated after the initial investment is paid back. However, these cash flows can make an investment highly lucrative or not. Therefore, Brealey et al. suggest to use the NPV to justify investment decisions.

However, the NPV is also not without limitations. A practical disadvantage of the NPV is that it does not consider the scarceness of the available resources. For instance, an asset with a predicted NPV value of 100 Euro is considered a better investment than one with a predicted 80 Euro, even though the former asset may involve investment costs of 10 million Euro and the latter 0.1 million Euro. We note that this aspect is accounted for by the IRR, as well as in the ROI indicator by dividing by the investment costs.

Another disadvantage of the NPV is that the determination of the discount rate, which is needed in the calculations, is difficult. Later on in this section, we discuss this discount rate in more detail. It is possible to derive organization-specific discount rates. However, determining the required project-specific discount rate is not trivial.

Moreover, one should consider that a discount rate that is applicable now, may not be applicable next month as is pointed out by Ross [105]. A project proposal with a negative NPV given the discount rate this month, can be a very interesting opportunity some time later. Ingersoll et al. [52] developed a method to account for the value of optionality with respect to the uncertainty of interest rates. An investment should only be made when the NPV is sufficiently positive to forego the option to delay the investment.

Indicators in practice A few decades ago, numerous surveys [31, 33, 95, 97, 109] found that the NPV was not the method of choice of most Chief Financial Officers (CFO). In 2002, an article by Ryan et al. [106] shows that till 1996 studies generally indicated the method most used by organizations was the IRR. Both Ryan et al. [106] and Arnold et al. [5] find that only just after the year 2000, organizations have adopted the NPV as preferable indicator. The survey of Ryan et al. [106] found that 85% of the respondents of the survey indicated to use the NPV often.

Still, all these surveys indicate that the organizations frequently use multiple indicators to support decision making. Although theory suggests the NPV should suffice to make investment decisions, in practice a combination of the NPV and other common indicators are used. Ryan et al. [106] also found that there is a correlation between the capital budget and the use of NPV and IRR. The larger the budget the more likely the use of either one of these methods. The Payback Period is found to be more frequently used by organizations with smaller capital budgets. These surveys show that many economic indicators are used for decision making.

5.2.2 Net Present Value

In this section we elaborate on the Net Present Value. First, we show how to compute the NPV. Then, we discuss its components, assumptions and interpretation.

Informally stated, the NPV determines the monetary value an asset adds to an organization. The cornerstones of this economic indicator are the predicted benefits, costs and economic life span. Simply put, the NPV determines whether the benefits outweigh the costs, both of which are computed in todays worth. If the NPV is positive, it means the asset will generate value for the organization. If it is negative, creating the asset will result in an overall loss.

Formally, the NPV is described by Brealey et al. [11] as follows. Denote CF_p as the cash flow predicted for time period p and r_p the discount rate of time period p. Let N be the total amount of time units that are used. Then, the NPV is calculated in the following way:

$$\mathrm{NPV} = \sum_{p=1}^{N} \frac{CF_p}{(1+r_p)^{p-1}}.$$

Below we discuss the elements of the formula in more detail.

5.2.2.1 Discount rate

The discount rate r_p is also known as the opportunity cost of capital. Often, the discount rate is chosen identical for each time unit. That is, $r = r_p, \forall p$.

The purpose of the discount rate is two-fold. First, it accounts for the time value of money. It is better to have 100 Euro today than it is to have 100 Euro tomorrow, since the former can be invested immediately to generate additional income. By discounting the future cash flows, we acknowledge this time value of money.

Second, any investment must be funded with capital. The providers of this capital require compensation for making their capital available. This is the cost of the capital that the organization intends to use. Any investment should aim to have a higher return than the cost of capital. If not, the organization would waste the capital of the investors. They would have done better to return the money to the investors and let them invest it otherwise.

But how to determine this cost of capital or discount rate? A number of books and articles [1, 11, 40, 104, 109] explain methods, among others the weighted average cost of capital or WACC, to find the discount rate. A survey [13] found that the WACC is used most often in practice. The WACC is organization-specific.

However, the discount rate used in the calculations of the NPV is investment-specific. Not all assets of an organization will have the same risk as the entire organization. Some assets will have higher risk and other assets will have lower risk. The organization-specific WACC should be changed to account for the particular asset risk. For instance, Dewan et al. [22] and Verhoef [124] suggest to increase the WACC in case of IT assets.

Determining this correction to the WACC is not a trivial task. Moreover, it is often difficult to establish which other assets are equivalently risky. It is therefore not surprising that the survey by Bruner et al. [13] found that many organizations do not adjust the WACC for individual investments. Petty et al. [95] contended that the use of sophisticated risk-adjustment techniques would be limited until risk can be measured more precisely and one can show the impact of additional risk upon the firm's cost of capital.

Discount rate forecast The discount rate that is used to compute a particular NPV, is based on an assessment of the risk of that investment. In most cases the risk involved is not objectively measurable and is thus only a prediction of the actual risk.

The accuracy of this forecasted discount rate directly impacts the forecasting accuracy of the NPV. However, in this chapter, we will not assess the forecast quality of the discount rate. We consider the cost of capital as given and will not question its derivation. In our case study of organization Z, all calculated NPVs of a particular asset are based on the same discount rate.

5.2.2.2 Cash flow

Another crucial element of the NPV calculations are the forecasts of the cash flows. The cash flows CF_p in the equation are the expected cash flows for each time period p. These predictions should account for the likelihood and impact of the risk of different scenarios on the benefits and costs, such as cost overrun, project failure and/or late delivery. These scenarios lead to a probability distribution of possible cash flows.

Surveys of Gitman et al. [33] and Fremgen [31] found that estimating the cash flows is considered the most critical stage of the capital budgeting process. In 1978, Schall et al. [109] surveyed that individual project risk is assessed by means of a probability distribution of cash flows by 25% of the respondents and another 10% using sensitivity analysis. In most cases, the risks were assessed implicitly. That is, the distribution is not made explicit, but is implicitly incorporated in the predictions of the cash flows by the estimator. In 2000, Arnold et al. [5] found that 94% of the organizations required a formal risk evaluation. This was done in

85% of the cases using sensitivity analysis, often in conjunction with a subjective assessment. A probability analysis was performed by 31% of the organizations.

In stark contrast are the findings in the information technology sector. In 2002, the Meta Group [36] surveyed that 84% of the organizations do not use business cases at all for their IT-projects or only for selected projects. The Kellogg School of Management [81] found that 68% do not track benefits. The numbers show that organizations have difficulty determining the benefits of IT assets, let alone to formally account for the risks in the forecasts.

Unbiased A critical assumption of the cash flow predictions is that they are unbiased. However, forecasts of, for instance, duration and cost can be biased [7, 8, 96]. It is conceivable that the same applies to forecasts of cash flows. The decision to invest in a project is highly dependent on these forecasts. Those making the proposal and those with interests in executing it, may be inclined to overestimate the cash flows to make the proposal more appealing. It is therefore crucial to check whether the assumption of unbiased cash flow forecasts holds.

To investigate this assumption, we have to consider what the cash flow is composed of. A cash flow is the resultant of the projected benefits minus the forecasted costs. If we find the cash flows to be unbiased, this may imply both benefits and costs are unbiased. But it could also mean that they are both highly biased, but counter each other's effect. Although the effect is overall the same, the latter situation is unwanted. In that case, it is mere luck and not good forecasting practice. Luck may change any instant, a good forecasting practice not. Therefore, to assess whether cash flows are unbiased, we should not only analyze the cash flows, but consider the components that it consists of.

To better illustrate the different components of the cash flow, we reformulate the above equation of the NPV. We define b_p to be the predicted benefits of period p, c_p the predicted asset usage costs and i_p the project costs, with $CF_p = b_p - c_p - i_p$. Let N be the total number of periods in which the economical asset life span is divided and r_p the discount rate of time period p. Then, we write the NPV using the following equation:

$$\begin{aligned}
\mathrm{NPV} &= \sum_{p=1}^{N} \frac{b_p - c_p - i_p}{(1+r_p)^{p-1}} \\
&= \sum_{p=1}^{N} \frac{b_p}{(1+r_p)^{p-1}} - \sum_{p=1}^{N} \frac{c_p}{(1+r_p)^{p-1}} - \sum_{p=1}^{N} \frac{i_p}{(1+r_p)^{p-1}} \\
&= B - C - I. \qquad (5.1)
\end{aligned}$$

In Formula 5.1, B amounts to the cumulated discounted benefits, C the cumulated discounted asset usage costs and I the cumulated discounted project costs. Recall that in Figure 2.1 in Chapter 2, we illustrated that the benefits and asset usage costs are not present during project execution. That is, the summation of the benefits and asset usage costs usually only have values over the interval

$p \in y, y+1, \ldots, N$, where y is the period in which the end of the project, t_e, falls. The summation of project costs usually has values over the interval $p \in 1, 2, \ldots, y$.

To assess whether the cash flows and NPV are unbiased, we have to investigate each of these components. We make a distinction between project costs and asset usage costs, since we wish to know whether the quality of their forecasts are different. Many organizations record and have insight in the project costs. However, these costs often amount to only a relative small portion of the entire asset costs. We want to see whether accurate forecasts of the project costs pays off, or that it is wise to put more effort in the correct prediction of the asset usage costs.

We note that it remains helpful to analyze the NPV directly. Assessing the quality of the forecasted NPVs shows the impact of the interactions of the individual elements. If we find both the benefits and the costs to be overestimated, we do not know their combined effect on the accuracy of the NPV. Therefore, a combination of the individual analyses and the overall quality of the NPV is most insightful. In this situation, the interdependences are contained in the NPV analysis and the individual analyses provide answers as to where the variance comes from.

In the case study of organization Z, we will assess the forecast quality of the NPV. Above, we discussed that it is useful to increase the depth of the analysis by also investigating the components of the NPV, that is, B, C and I. It is possible to further increase the depth of the analysis by also considering the components of B, C and I. For instance, B is the sum of b_p over all time periods p. Therefore, it is possible to investigate the forecast quality of each b_p separately.

The level of detail that is required depends on the goal of the analysis. In this chapter, the primary goal is the forecast quality of the NPV. Moreover, we wish to determine whether control of the project costs is useful without sufficient control over the other elements. For these purposes, we will only analyze the NPV, B, C, and I.

5.2.2.3 Time

Besides the discount rate and the predicted cash flows, time is another variable in the equation. Time is captured in the predicted total number of time periods denoted by N. This total amount is a forecast of the asset's economical life span. The time periods are commonly expressed in years, but can also be different, for instance, in months.

An accurate prediction of this variable is important for the resulting NPV. If the life span is too long, we may unjustly attribute additional benefits and costs to the asset. On the other hand, if the predicted life span is too short, we will ignore future cash flows of the asset in our calculations. For instance, suppose the initial forecast predicted the economic life span to be 5 years. It may occur that when the NPV is re-estimated, it is estimated the economic life span is 5.5 years. When in this half year a positive cash flow is generated, the initial NPV forecast will be underestimated. These cash flows can make the difference between a positive or negative NPV.

However, in this chapter, we will not assess the forecast quality of the asset's economic life span. We consider the life span, N, as given and will not question its derivation. In our case study of organization Z, the two estimated NPVs of a particular asset used the same economic life span. Moreover, all estimated NPVs of a particular asset are discounted to the same present moment.

5.2.2.4 Indirect influences

Besides the described elements in the NPV equation, there are other factors that influence the outcome of the NPV indirectly.

Indirect time effect Apart from the direct impact, time also influences the NPV calculation in indirect ways.

For example, consider an asset that is delivered three months later than expected. In the first three months no benefits or asset usage costs occur, making their forecasts overestimations. Moreover, all predicted cash flows will become different. Namely, because the cash flows occur later, they will be discounted differently. Also, the timing of the forecasts can be crucial, for example, due to growing competition on the market. Therefore, delays can severely influence the NPV, benefits and costs, making their accuracy forecasting difficult. These inaccuracies in the forecasts are also discussed in an article by Peters et al. [94].

Moreover, in an article by Putnam [102] and a book by Boehm [8, p. 472] it was found that shortening the project duration beyond its optimum can significantly increase the project costs. Boehm observed a similar effect when the project duration was stretched beyond its optimum.

Therefore, the accuracy of the prediction of the project duration is reflected in the accuracy of the forecasts of benefits, asset usage costs and project costs. Any inaccuracy we find in the analysis of the other variables may partly be caused by the inaccurate forecast of the project duration.

Functionality Although functionality is not mentioned in the formula of the NPV, it has an indirect impact on the variables in the equation. For instance, if a project is executed and delivers less functionality than anticipated, this may prevent certain forecasted benefits to materialize (solution underdelivery). Similarly, if the resulting IT program has more functionality, it is possible additional benefits are generated as a result. An increase in requirements is also known as requirements creep [14, 56]. On the other hand, an increase in functionality may also result in higher project and asset usage costs.

These effects should be weighted in the overall forecasts of the benefits, asset usage costs and project costs. This was done, for instance, in an article by Peters et al. [94] by considering requirement creep scenarios. Due to these effects, it is possible the forecasts of functionality are correlated with the forecasts of the other components. For instance, if the functionality is underestimated, the benefits and costs may be underestimated as well.

To investigate a possible correlation between the accuracy of the project cost forecasts and the functionality forecasts, we analyzed the data from organization Y. In Figure 5.1, we depicted the initial functionality, denoted by F, $(f/a)_{\mathrm{F}}$ ratios against the initial project cost $(f/a)_{\mathrm{I}}$ ratios of 55 projects of which we had all the relevant data. The functionality forecasts were measured using function point countings [23, 32].

The correlation coefficient of the two data sets is -0.05. Since values of -1 or 1 represent perfect correlation and 0 depicts no correlation, the coefficient shows that there is no correlation. This result reveals that the relation between functionality and project costs may only be marginal with respect to their forecast quality. That is, underestimating the functionality does not directly cause an underestimation of the project costs.

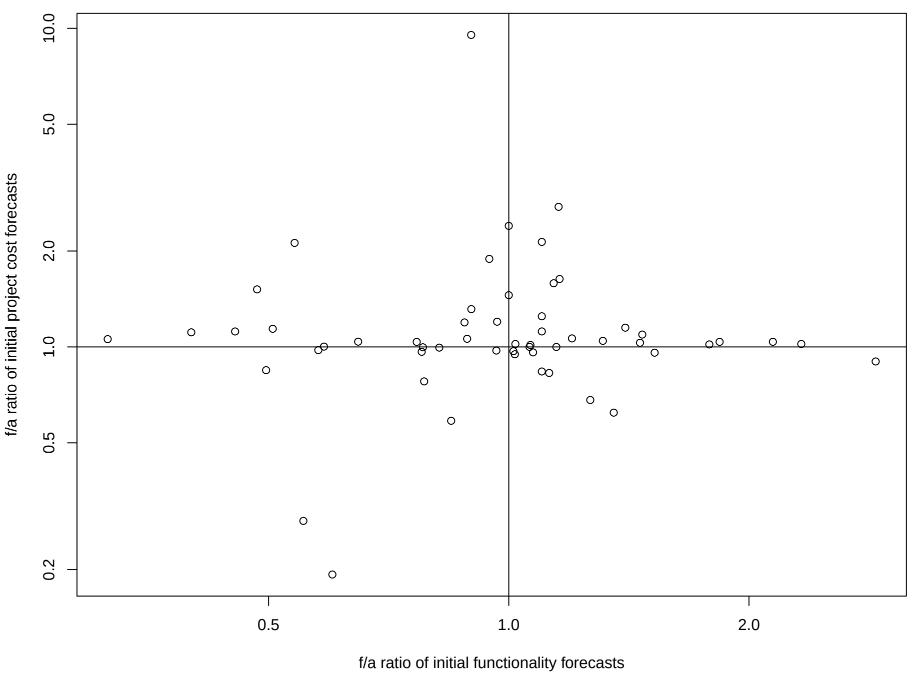


Figure 5.1: No evidence of correlation between forecasting accuracy of initial project cost and functionality forecasts.

We note that Boehm's cone of uncertainty [8] showed that the time at which the forecasts are made is relevant for their quality. Although the forecasts in our real-world data set of project costs and functionality are not made at the same time, the differences between the moments they are made are small. The median of the difference between the moment the project cost forecast was made and the functionality forecast was made, divided by their project duration is 0 and the average difference is 0.08. Therefore, there is no indication that the results of this analysis are influenced by the different timing of the forecasts.

Software quality Software quality is highly relevant to value prediction of both the benefits and the costs. The benefits derived from an asset are based on the perceived value of the customer. If the software contains many defects, the customer may value the asset less, thereby generating less benefits.

Capers Jones [57] found that the U.S. average for IT software quality is about 5.0 defects per function point combined with 85% defect removal efficiency. This results in delivery of about 0.75 bugs or defects per function point. Best in class IT software quality produces less than 3.0 defects per function point combined with more than 95% defect removal efficiency. This results in delivery of about 0.15 defects per function point.

Jones argues that increasing the defect removal efficiency from 85% to 95% saves money and shortens development schedules. This is supported by others [14]. Applications using state of the art software quality methods have development costs about 15% lower than average projects, and maintenance costs about 55% lower than average projects. Total cost of ownership is about 40% lower than average projects [58]. Therefore, it is beneficial to produce assets with a relative high quality.

This illustrates that the forecasts of benefits and costs are influenced by the delivered software quality of the asset. However, in this chapter, we will not consider the software quality of the assets. In this chapter, we focus on assessing the resulting forecast quality of the NPV, benefits and costs.

5.2.3 Summary

To support decision making, IT executives have access to numerous economic indicators that summarize the expected business value of project proposals. Despite its limitations, theory suggests that the NPV method is superior compared to other methods. However, our discussion of the indicator shows that the NPV is far from certain. Each of its components, needs to be predicted. Therefore, it is not surprising that in practice many organizations use multiple economic indicators to gain insight in the value of an asset proposal.

Like the NPV, any economic indicator is highly dependent on the accurate forecasting of its elements. There are numerous components that influence the final NPV either directly, such as the benefits or costs, or indirectly, such as project duration, functionality or software quality. Therefore, in this chapter, we assess the forecast quality of the NPV, and its components: the benefits, asset usage costs and project costs.

In the next section, we discuss a method with which it is possible to assess the forecast quality of the NPV and its components. Those not interested in the mathematical elaborations, may skip the next section and continue with Section 5.4.

5.3 Generalized method

In this chapter, we wish to determine whether organizations are able to adequately assess the business value of project proposals. In the previous section, we discussed the NPV and other economic indicators, which represent the business value of assets. To assess whether organizations are capable to accurately forecast the business value of IT proposals, we need to investigate the forecast quality of the NPV and its components. Recall that the Net Present Value is a summation of the predicted monetary benefits and costs of a project discounted to current value.

A method to assess forecast quality is described in Chapter 3. As discussed that method uses the f/a plot, the reference cone and the EQF. Recall that the EQF is a measure of the deviation between forecast and actual. The method is applicable for entities that are positively valued, that is, $f, a > 0$. We wish to evaluate the quality of the forecasts of the NPV, benefits and costs in the same manner. However, if we want to do so, a number of problems arise.

Asymptotic behavior The first problem is a visual complication in case of asymptotic behavior. With asymptotic behavior we refer to situations in which forecast and/or actual are zero. For project costs, usually the forecasts and actuals are greater than zero. In practice, project proposals that cost nothing or assets that cost nothing hardly ever occur. However, in the context of economic indicators, a forecast of zero is important. For instance, a NPV of zero is the turning point between an asset being yield or loss generating. Also, it is not unlikely to find assets that have no benefits. For example, consider an asset that was developed, yet on completion it turned out there was no longer a market for it.

Zero actual benefits cause visual problems for the f/a plot. Let us explain. The f/a plot visualizes potential biases by plotting f/a ratios on a logarithmic scale. If the actual is zero, the ratio becomes infinite, making the logarithm also infinite. Therefore, in the f/a plot this point cannot be visualized in a normal way.

This problem also arises if the forecast is zero. In this case, the f/a ratio is zero. However, the logarithm of zero is not defined. Thus, in that case the ratio can also not be depicted in the plot.

Normally, not many of such zero forecasts will be made. In most cases, a forecast of zero indicates no forecast is made at all. If this is the case, it is best to remove these forecasts from the analysis all together, as they will not reveal information on the quality of the forecasting process. However, sometimes an entity is truly forecasted to be zero, or remains interesting for analysis in conjunction with other forecasts. For instance, consider the case of a forecast of benefits and no forecast made for the costs of the project. In that case, it may be interesting to incorporate the data point in the analysis, to assess the quality of the resulting NPV forecast.

Why is it a problem that the f/a plot does not visualize ratios with a forecast or actual of zero? If there are many such ratios in a data set, the f/a plot may point to a potential bias in the data that does not exist. For example, consider an extreme situation in which 51% of the data in a data set consists of zero actuals and the remainder of the forecasts are underestimations, that is $f/a < 1$. The f/a plot would only depict the latter half of the data, which will show a bias toward underestimation. In reality, there is no particular bias given the data, since 51% of the data, the zero actuals, are overestimations. Clearly, the ability of the f/a plot to detect biases is hampered by zero forecasts and/or actuals.

Reference point A second technical problem is caused by the reference point with which we compare the quality of the forecasts. With the f/a ratio this reference point is the actual. However, it is questionable whether the actual is useful to support decision making. For instance, suppose one finds that 2-year old project cost forecasts or 5-year-old forecasted benefits were generally overestimated. Although this sheds light on the forecast quality of 2 or 5 years ago, in most cases this is hardly information that is useful to apply to today's forecasts. By the time the actuals are known, the forecasting practice may already have been changed. To support decision making, we prefer more recent information about the current forecasting practice. Moreover, in case of, for instance, the NPV, the actual may never be objectively measured.

A solution is to re-estimate, for instance, the benefits before the end of the economic life span. For instance, it is possible to re-estimate the benefits a year after project completion. At that time, it is more clear to the estimator which of the many possible scenarios is unfolding. The ex-post part is considerably larger than in the previous forecasts and the ex-ante part becomes smaller and smaller. This way we are able to approximate the actual, which allows us to derive more recent forecast information that we are able to use for today's project proposals.

However, in this case the value of the re-estimation is no longer objectively measurable. In fact, it is a forecast in itself. If we compare the forecast quality of earlier forecasts with this approximation of the actual, we use a different reference point than the actual. We will investigate how the assessment of forecast quality is affected by such alternative reference points.

Negative values Finally, the NPV can be both positive and negative. In practice, most negative NPVs will occur in re-estimations. For example, consider an asset with a positive forecasted NPV that grossly overestimated the benefits. Afterward, the asset was recomputed and the costs turned out to be greater than its benefits, resulting in a negative NPV.

It is also possible to have negative forecasted NPVs. For example, mandatory assets or assets with a negative NPV that are interesting for their qualitative features. We note that mandatory assets can have a positive value as well. If the asset would not be performed, the organization risks a fine or other sanctions, which potentially make such assets beneficial to undertake. Even in case of negatively forecasted NPVs, their forecast quality remains interesting to investigate to contain the predicted losses.

The negative values cause two problems for the f/a plot. The first problem is that the f/a plot uses a logarithmic axis to depict f/a ratios. However, the logarithm is not defined for negative values. Therefore, we are unable to depict them. We could abandon the logarithmic scale and use a linear axis, allowing us to depict negative ratios as well. However, a linear axis does not allow for easy distinction of biases.

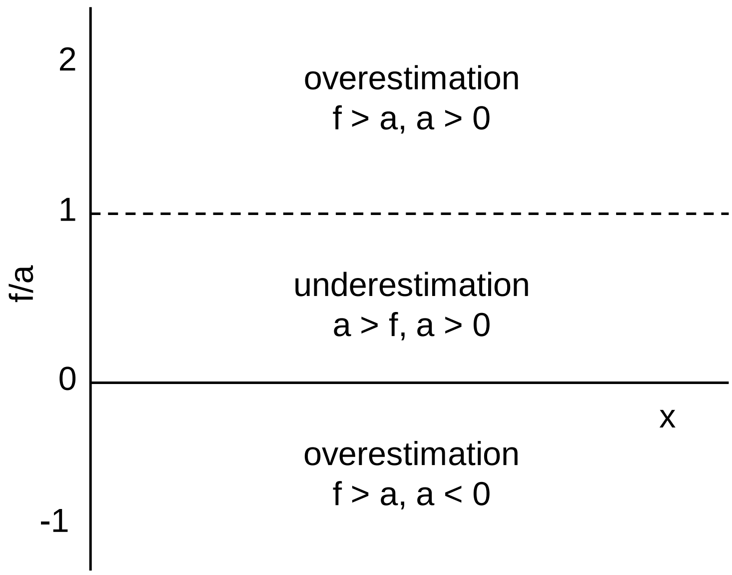


Figure 5.2: Illustration of an f/a plot with a linear vertical axis, positive forecasts and allowing for positive and negative actuals.

Let us explain. Assume we have positive forecasts and both positive and negative actuals, that is $f > 0$ and $a \in \mathbb{R}$. Suppose we would use a linear axis for the f/a plot and allow for negative actuals. In this case, all the negative ratios would be depicted below the line $f/a = 0$. This situation is illustrated by Figure 5.2. These negative ratios indicate that, since forecast f is positive, the forecast is larger than the actual and is thus an overestimation. For all f/a ratios in the interval $[0, 1]$, the f/a ratio indicates that the forecasts are smaller than their actuals, or equivalently, they are underestimations. Finally, the f/a ratios above the line $f/a = 1$ depict forecasts that are larger than the actuals, which are again overestimations. Thus, when we use a linear axis and assume non-negative forecasts, and both positive and negative actuals, the f/a plot consists of three sections of which the middle one shows underestimations and the remainder overestimations.

A similar explanation applies when we assume non-negative actuals, and both positive and negative forecasts. Thus, such figures make it difficult to adequately distinguish biases, not to mention obtaining a visual impression of the quality of the forecasts.

The second problem of the negative values is that the f/a ratio becomes ambiguous. For instance, consider an f/a ratio of -1. Such a ratio is possible when $f < 0 < a$, but also when $a < 0 < f$. Thus, a ratio of -1 can both mean the forecast is larger than the actual as well as the actual is larger than the forecast.

A similar problem arises with positive f/a ratios. Consider an f/a ratio of 2. This occurs when $f > a$ with $f, a > 0$ and when $f < a$ with $f, a < 0$. Thus, it is no longer possible to determine whether the forecast is smaller or larger than the actual value. To cope with these problems, we will propose other ratios than Boehm's f/a ratio that deal satisfactorily with negatively valued entities.

In the remainder of this section, we address each of these issues separately. This results in a generally applicable method to assess forecast quality of entities that range over the entire real numbers.

5.3.1 Asymptotic behavior

Above we stated that in case of asymptotic behavior, that is, forecast or actual is zero, a problem of visualization arises. Kitchenham et al. [70] also discuss visualizations problems that can arise when using ratio measures. For the f/a ratio, there are three possible scenarios. First, suppose we have a positive valued forecast and an actual of zero. In this case the f/a ratio is infinite. In an f/a plot such points cannot be depicted in a normal way.

Second, consider a positive valued actual and a forecast of zero. Now, the f/a ratio is zero. Since the f/a plot uses a logarithmic axis and the logarithm of zero is undefined, this ratio cannot be drawn.

Finally, it is possible to have a forecast and actual of zero. In this situation, the f/a ratio of $0/0$ itself is undefined. Again, we are unable to visualize this ratio in the f/a plot.

If many zero forecasts and/or actuals are present in a data set, this hampers the ability of the f/a plot to visualize biases. Therefore, we define the desired behavior of an f/a ratio in the described asymptotic cases.

Zero actual If the actual is zero, we wish to depict a ratio in the f/a plot that is infinite. A possible solution is to draw these points an order greater than the maximum value of the other ratios in the data set that are smaller than infinity. For instance, suppose the maximum f/a ratio smaller than infinity is 20. Then, we plot the infinite f/a ratios as f/a ratios one order larger, that is, 200. To allow for a better distinction of the infinite f/a ratios, we visualize them differently from the other ratios. Suppose the normal ratios are shown as dots, then we plot the infinite ratios, for instance, as squares. In this way, we are able to visualize these ratios, resulting in an f/a plot to adequately depict potential biases.

Zero forecast Suppose we have a forecast of zero and positive actual. In this situation, the f/a ratio is zero. The f/a plot uses a logarithmic scale for the f/a ratios. However, the logarithm of zero is not defined. Therefore, the logarithm prevents us from drawing f/a ratios of zero.

We could draw these points without taking the logarithm. However, this is unwanted as then the points are no longer scaled, which is very useful in recognizing potential biases and evaluating the accuracy of the forecasts. Namely, this scaling creates equal spacing for under- and overestimation. That is, in the figure the distance between 1/2 and 1 is equal to that of 2 and 1. This property is useful to assess the accuracy of both under- and overestimation.

Therefore, we propose to visualize these points similarly to f/a ratios with a zero actual. When the forecast is zero and the actual is positive, this indicates that the forecast was smaller than the actual. Thus, we draw these points one order smaller than the minimum value of the f/a ratios in the data set that are larger than zero. For instance, suppose the minimum value of f/a ratios larger than zero is 1/20. Then, we plot all f/a ratios that are zero at an f/a ratio of 1/200. Again, to allow for better distinction, we depict these points differently from the other f/a ratios.

Zero forecast and actual Finally, the situation could occur that both forecast and actual are zero. The f/a ratio of 0/0 is undefined. In this scenario, the forecast that was made was extremely accurate. It precisely predicted the final outcome. Therefore, a solution is to define such an f/a ratio as 1, that is, $0/0 \equiv 1$. For better distinction, these points should be visualized differently from the other ratios.

Besides these solutions, we point out that there is a statistic that, in conjunction with the f/a plot, is useful to assess biases. This is the median f/a ratio of the data set. The median f/a ratio is not influenced if either a forecast or an actual is zero. Only if they are both zero, there is a potential influence, since the f/a ratio of 0/0 is not defined and is therefore not used to compute the median.

Nonetheless, it is useful, besides the f/a plot, to consider the median f/a ratio. We will visualize this median in the f/a plot by adding a horizontal dashed line at the value of the median.

5.3.2 Reference point

All f/a ratios use the actual to compare the forecast quality with. However, when we want to use the forecast quality to enhance decision information, the actual is not necessarily the most useful reference point. For instance, in our case with the NPVs, we know their value with certainty as soon as the economic life span is over, and then it is too late to enhance decision information.

Then, we need an alternative: a reference point that is not the actual, but most probably a better prediction of the actual than the initial forecast. For instance, it is possible to approximate the actual two or three years after the exploitation of the asset.

In our case study, the telecommunication organization, Z, is accustomed to approximate actuals by re-estimating them after, say, a year upon project completion. This is in itself a new forecast, usually of better quality than of other previous forecasts. Namely, at the time of the re-estimation, there is more information available. For the estimator it becomes more clear which of the many possible scenarios is unfolding. The project is finished and operational, initial benefits are received and initial asset usage costs are made. This should give a more reliable indication of the final realization than at the start of the project. Of course, in theory this may not be true, but in most cases this approach may be the most relevant prediction of the actual we are able to obtain.

If we wish to use such re-estimates, an important consideration is how to determine a reasonable period after which to perform the re-estimation. For instance, one may wonder whether it is beneficial to re-estimate the NPV as soon as possible. If the re-estimation is made earlier, it will potentially deviate less from other forecasts as things may not have changed as much. Of course, the closer the reference point r is to the actual a, the better the results of the analysis using r should compare to the analysis using the actual.

To minimize such influences, preferably the ex-post part of the re-estimation must be as large as possible. Unfortunately, there is no easy guideline for what a reasonable period is.

Later on, we investigate the influence on the analysis of the forecast quality when we use other reference points than the actual. But first, we discuss the technical consequences for the f/a plot, EQF and reference cone in case of alternative reference points.

5.3.2.1 Generalization

To account for reference points instead of the actual, we need to generalize the f/a ratio. So far, we have used the term *f/a ratio* to refer to a forecast divided by its actual. We now replace the actual by a more general reference point r. This leads to the notation f/r ratio to denote a forecast divided by a reference point r. When this point is the actual a, this is equal to the f/a ratio as proposed by Boehm [8].

f/r plot We use the notation f/r plot, instead of f/a plot, to denote a forecast to reference point plot. The replacement of the actual by a reference point affects both the horizontal and the vertical axis.

In the current f/a plot, the horizontal axis was scaled using t_a, which indicates the date of the actual. For instance, if the actual occurred after two months and a forecast was made after one month, the f/a ratio would be depicted at the horizontal axis at $(t_f - t_s)/(t_a - t_s) = 1/2$. In the generalized method, we have obtained the reference point at t_r, some time before the date of the actual t_a. Since we refer to reference point r in the f/r plot, we will also use t_r as scaling factor instead of t_a. That is, the horizontal axis of the f/r plot depicts the project progression based on the reference point, defined by $(t_f - t_s)/(t_r - t_s)$.

For the vertical axis, the f/r plot depicts f/r ratios instead of f/a ratios.

EQF_r The change of reference point not only affects the f/r plot, but also the way we compute the EQF. Let us explain. The EQF is defined by the following formula:

$$EQF = \frac{\int_{t_s}^{t_a} 1\, dt}{\int_{t_s}^{t_a} |1 - e(t)/a|\, dt}.$$

However, in this formula we do not have the values for the actual a or the value of t_a. We *do* have the values of the reference point r and the time at which this re-estimation was made t_r. Since we use this re-estimation as reference, we substitute a with our more general reference point r and t_a with t_r. This leads to the following notation of the EQF:

$$EQF_r = \frac{\int_{t_s}^{t_r} 1\, dt}{\int_{t_s}^{t_r} |1 - e(t)/r|\, dt}. \tag{5.2}$$

We denote this EQF with EQF_r to indicate the difference between the formulas. Note that for EQF_r, the deviations to the reference point are not time-weighted over the entire economic life span of the asset, yet, only over the time span that we have seen thus far.

Reference cone The reference cone is not affected by the change of reference point. The only necessary change is that the original EQF should be replaced by EQF_r. This leads to the following reference lines.

$$l(x) = x + \left(1 - \frac{2}{\mathrm{EQF}_{r,l}}\right) \cdot (1 - x) \tag{5.3}$$

$$u(x) = x + \left(1 + \frac{2}{\mathrm{EQF}_{r,u}}\right) \cdot (1 - x). \tag{5.4}$$

We note that x, given by $x = t_f - t_s/t_r - t_s$, is relative in these formulas and lies in the interval $[0, 1]$. These lines use t_r instead of t_a. Still, the slopes of the lines remain the same even though we changed the reference point.

5.3.2.2 Interpretation

Above we described the technical changes of the generalization, yet not their interpretation. With the f/a plot, EQF and reference cone, we are able to quantify forecast quality by assessing the data for potential biases and its accuracy. These tools use an objectively measurable actual as reference point. Since it is objectively measurable, the actual is therefore unbiased.

With the f/r plot, EQF_r and reference cone, we use an alternative reference point. This reference point is only partially objectively measurable. That is, it is only possible to objectively measure the ex-post part. The remainder, the ex-ante part, needs to be predicted. Therefore, the reference point is also a forecast. As we know, the result of a forecast is uncertain. Moreover, forecasts are potentially biased. Therefore, it is important to determine the influence of the change of reference point on the analysis of the forecast quality.

For instance, is the ability of detecting potential biases affected by the change of reference point? If the reference point itself is biased, the overall bias we find in the analyses may be lessened or even nullified. And, how does using another reference point affect the outcome of the EQF_r? For instance, it is possible the change of reference point creates a positive as well as a negative effect on the EQF_r compared to the EQF. Namely, when we use t_r instead of t_a, we give more weight to the earlier forecasts. These earlier forecasts may appear more accurate if the reference point is closer to the forecast than the actual. On the other hand, it is also possible they appear less accurate when the reference point is further away.

Beforehand, in an unbiased situation there is no way to determine whether the reference point will be below or above the actual. Namely, any forecast and therefore the reference point should be centered around the actual. Therefore, it is equally likely to be above or below, or oscillate around the actual. Due to this oscillating behavior, the EQF_r will not necessarily improve with respect to the EQF. It may just as well be lower than the EQF.

To investigate the effect of the generalization, we analyze the data of Chapter 3, of which we have multiple forecasts and of which we also know the actual. This allows us to compute the forecast quality using the actual as reference point. Subsequently, we use forecasts made before the final realization as alternative reference points. Using these reference points, we are again able to assess the forecast quality. The differences between the outcomes of both analyses, provides insight in the effect of replacing the actual by a reference point in real-world cases.

As said, the data analyzed in Chapter 3 was obtained from four organizations. The first data set from a large financial organization, Y, consist of 667 forecasts of in total 140 project costs. The second organization is a multinational organization, X, for which we have 3767 forecasts for in total 867 project costs. The third data set is of a large telecommunications provider, Z, consisting of 971 forecasts of in total 307 project costs. The final data set is of Landmark Graphics containing 923 forecasts of in total 121 project durations.

In the data sets, we aim to use forecasts as reference points that are in the range of 20%–60% project completion. We chose this range to mimic the other data set of organization Z containing NPV calculations, which we will analyze in the next sections. There, we find that many re-estimations of benefits and asset usage costs are made between 0.5 and 1.5 years after project completion. With an average project duration of 1 year and an average life span of 5 years, this computes to roughly the same 20%–60%. The chosen range is thus a realistic reflection of the reference points that we will use in our case study to assess the forecast quality of NPVs.

Some projects may not have a forecast available in the range 20%–60% to be used as reference point. In most cases, the actual is then replaced by a forecast in the range 60%–100%.

Below, we show for each data set the analysis using the actual and alternative reference points, and discuss the differences that occur in these analyses.

Organization Y Of the total 140 projects, 125 projects had enough forecasts, to allow us to use one of their forecasts as reference point. The range in which 80% of these reference points were made, is 36%–74% project completion. Of the 125 projects, we have a total of 352 forecasts for the f/r plot. Since we wish to compare the quality of the assessments, we use the same forecasts for the f/a plot. We note that in the f/a plot, we compare these forecasts with the actual and in the f/r plot with an alternative reference point.

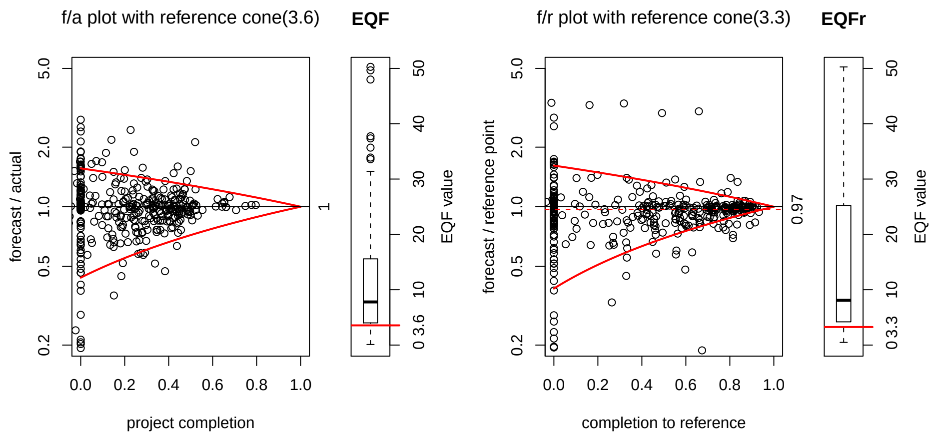


Figure 5.3: The f/a plot and f/r plot of 352 forecasts of in total 125 project costs. Both are unbiased, show convergence to the reference point and have similar median forecast quality.

In Figure 5.3 we depict both the f/a plot and the f/r plot of the forecasts. The f/a plot shows no bias, whereas the f/r plot reveals a potential optimistic bias. However, the f/a ratios with a median of 1.00 and the f/r ratios with a median of 0.97 indicate there is no particular bias in either case.

The forecast quality of the f/r plot is slightly higher than that of the f/a plot. Although the 20% quantile of 3.3 versus 3.6, is slightly lower, the median EQF_r of 7.8 versus 8.1 for the EQF is higher. Interestingly, the differences between EQF and EQF_r are a median 20% of the original EQF, indicating the EQF is generally higher. In fact, only 38% of the projects have a higher EQF_r than the EQF value. Most of these projects have relatively small EQF values. Not surprisingly, in these cases, the EQF_r almost always improves, thereby increasing the overall EQF_r quality. For individual projects, the EQF and EQF_r differ significantly. It is not unlikely to find the EQF_r being half to two times the original EQF. Compared to benchmarks from the literature [82, 20] and Chapter 3, the EQF values in both analyses are relatively high.

In this organization with unbiased and relatively accurate forecasts, the f/a plot and f/r plot lead to the same conclusion. Both the bias and the forecast accuracy show similar results. Therefore, the analyses do not appear to be hampered by changing from reference point in this case. However, due to the large differences between the EQF and EQF_r for individual projects, it is more difficult to determine which projects have inaccurate forecasts. Projects with low EQF values are likely to have a better EQF_r value when a different reference point is used.

Organization X From the 867 projects, we selected 713 projects that had sufficient forecasts to allow for a change of reference point. Of these alternative reference points, 80% are in the range 40%–87%. For the f/r plot we have 1373 forecasts available. We use the same forecasts in the analysis of the f/a ratios. Figure 5.4 depicts the results of both analyses.

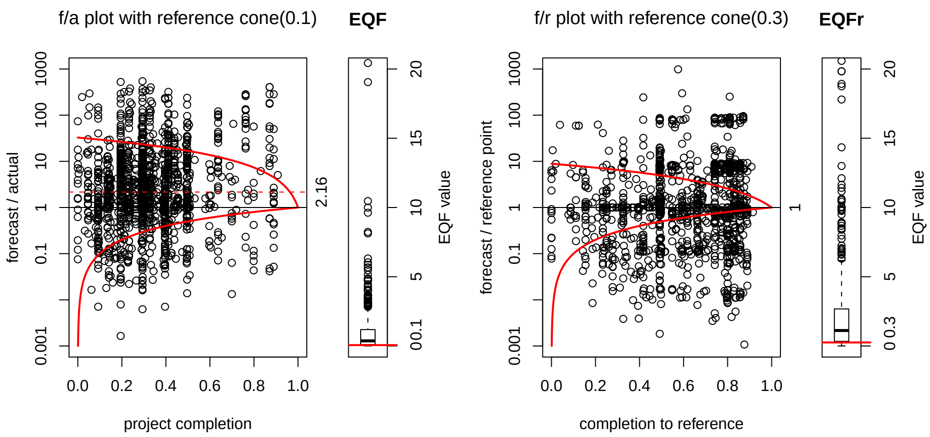


Figure 5.4: The f/a plot and f/r plot of 1373 forecasts of together 713 project costs. The former shows an overpessimistic bias, while the latter reveals no bias. Both plots show no convergence to the reference point and their median forecasting accuracy is similar.

The plots show a different bias. The f/a plot reveals an overpessimistic pattern supported by a median f/a ratio of 2.16. However, the f/r plot shows no particular bias with a median f/r ratio of 1.00. The explanation is that the reference point itself is equally biased as the other forecasts. By comparing the forecasts to their biased reference point, the bias disappears.

We note that, in this case, the bias is equal for both forecasts and reference points, since there is no convergence to the reference point. This result is evident from both plots. If there is convergence, the bias of the reference point should be less than that of the forecasts. In that case we would still see the bias, as we will see in the next case studies. However, if there is no convergence to the reference point, the forecasts will have the same bias as the reference point, thereby nullifying the bias in the f/r plot.

The assessment of the forecasting accuracy shows considerable differences. The EQF_r is higher than the EQF, in this case for both the 20% quantile, with 0.3 and 0.1 respectively, and the median value, with 1.11 versus 0.37. For 67% of all projects, the EQF_r value is higher than their EQF value. For individual projects, the EQF_r can be up to 0.7 to 10 times the original EQF.

This example shows an organization with low forecast quality and an overpessimistic bias. By using an alternative reference point, the overall forecast quality improves, yet remains relatively low. Generally, the EQF_r values are higher than their corresponding EQF values. The f/r plot no longer reveals any bias. The reason is that the alternative reference point itself is equally biased as the forecasts. When the f/r plot shows no convergence to the actual, the plot no longer reveals the true bias of the forecasts. Of course, in this case, the most important problem is the absence of convergence, which the f/r plot still shows.

Organization Z In this data set, 253 projects had more than one forecast. We selected 253 forecasts as alternative reference points, of which 80% are in the range of 11%–60%. In total, 554 forecasts remain for the f/r plot and the f/a plot. Figure 5.5 depicts the results.

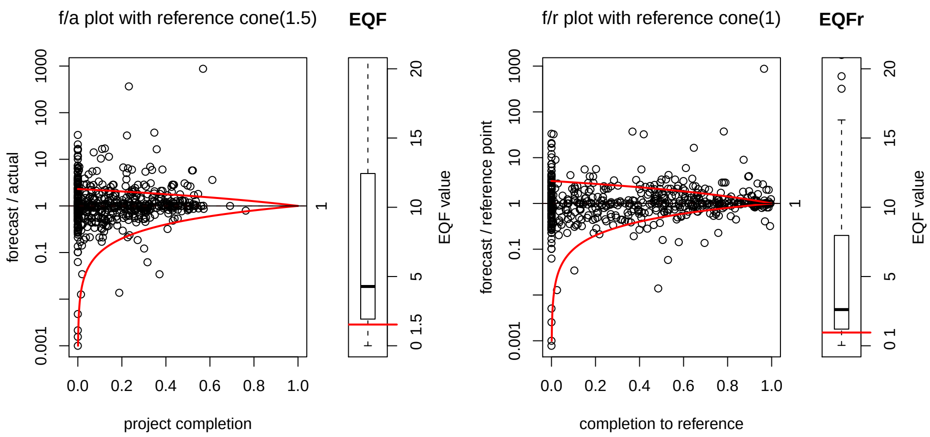


Figure 5.5: The f/a plot and f/r plot of 554 forecasts of in total 253 project costs. Both plots indicate no particular bias is present and the forecasts converge to their reference point. The median forecast quality is about the same in both analyses.

Both plots show no sign of a particular bias. This is supported by the median f/r ratio of 1 and the median f/a ratio of 1. Since the forecasts have no particular bias, the chosen reference points do not have a bias either.

The forecasting accuracy in terms of EQF$_r$ is lower than that in terms of EQF. Both the 20% quantile, with 0.96 versus 1.5, and the median of 2.6 versus 4.3, illustrate that for this organization, the accuracy using an alternative reference point is lower than the accuracy compared to the final realization.

For this organization, the effect of changing the reference point is noticeable in the forecast quality. The EQF$_r$ values are considerable lower than the EQF values. However, both analyses show that the forecasts are unbiased and converge to the reference value.

Landmark Graphics In this data set, all the 140 projects had sufficient forecasts that could be used as alternative reference points. The alternative reference points are for 80% contained in the 31%–71% interval. In total, 352 forecasts remained that we used in both the f/r plot and f/a plot.

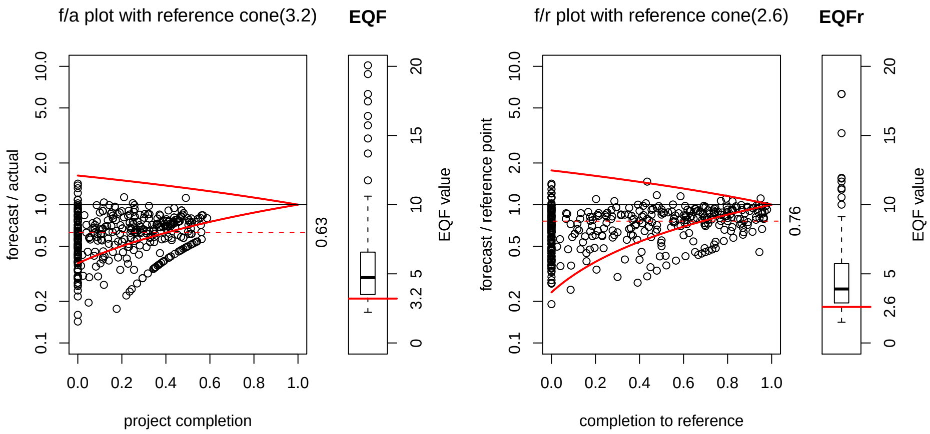


Figure 5.6: The f/a plot and f/r plot of 381 forecasts of 121 project durations. Both plots illustrate an optimistic pattern and convergence to the reference point. The EQF$_r$ values are slightly lower than the EQF values.

Figure 5.6 depicts the results of the analyses. The f/r plot and the f/a plot indicate the forecasts have an optimistic bias. This result is confirmed by the median f/r ratio of 0.76 and the median f/a ratio of 0.63. Although both show the bias, the bias of the f/a plot is larger than in case of the f/r plot. This is not surprising as the forecasts that we use as reference points in the f/r plot are themselves biased. Since the forecasts converge to the reference points, visual in both f/a plot and f/r plot, the bias of the reference point is smaller than for earlier forecasts. Thus, the bias remains visible, but is less pronounced than in the f/a plot.

For this organization, the EQF$_r$ values are generally lower than the EQF values.

The f/r ratios show a 20% quantile of 2.6 and a median of 3.9, whereas the f/a ratios have a 20% quantile of 3.2 and a median of 4.7. The impact for individual projects is relatively small in this organization. The difference between EQF and EQF_r is in 80% of the cases between 0.3 and 1.5 times the EQF.

The analyses using the actual and alternative reference points, lead to similar conclusions. The forecasts have an optimistic bias and converge to the reference value. The bias is less conspicuous in the f/r plot than in the f/a plot. The EQF_r values are generally lower than the EQF values. Of all projects combined, the forecasting accuracy in both analyses is relatively similar.

Summary The replacement of the actual by a reference point affects the analysis of forecast quality. In both cases it is possible to distinguish whether the forecasts converge to the reference point or not. If the forecasts do not converge, it is no longer possible to distinguish biases in the f/r plot. If the forecasts do converge, the f/r plot will show a bias similar to the f/a plot. However, the bias derived from the f/r plot will be less than that of the f/a plot. Therefore, the true bias will generally be worse than the one we find in the f/r plot.

With respect to the effect on the forecasting accuracy measured in EQF and EQF_r, there is no way to determine the impact based on the f/r plot. For some of our case studies, the EQF_r values were generally higher than the EQF values, while for others they were lower. Beforehand, in an unbiased situation there is no way to predict this behavior, as the reference point may be both higher and lower than the actual value. Therefore, for individual projects it is no longer possible to identify outliers with respect to forecasting accuracy. Still, when we analyze the forecasting accuracy of all projects combined, the overall conclusions remain similar.

The results show that care should be taken when using the EQF_r values for benchmarking purposes. The EQF_r values do not lend themselves particularly well for benchmarking between different organizations. Such comparisons are hampered by the moment at which re-estimations are made and the biases of the organizations. For benchmarking between organizations, the EQF based on the actual as reference point should be used. The EQF_r values, however, do have their value for comparisons within an organization.

5.3.3 Negative values

Finally, a problem arises for the f/a ratio and f/r ratio in case of negative values of f, a, and/or r. If we allow for negative reference points, the f/r plot becomes dissected into three parts of which one indicates underestimation and two overestimation. If negative forecasts f are possible as well, the f/r ratio itself becomes ambiguous, as it is not clear whether the forecast or reference point is larger than the other. To this end, we need to find a different way of dealing with negative values.

This different way should preferably allow for an identical assessment of the forecast quality as the f/r ratio. Therefore, before considering alternatives, we first enumerate the characteristics the f/r ratio has.

- Divisible: The ratio shows whether the forecast is larger than its reference point or vice versa. The ratios are strictly separated by some value that divides the ratios into two groups of under- and overestimation. For the f/r ratio this means that all values greater than 1 indicate the forecast is larger than the reference point and all values smaller than 1 indicate the forecast is smaller. This is useful as it allows to detect potential biases in the forecasts made.

- Difference: The f/r ratio shows the difference between the forecast and its reference point. It allows us to distinguish between accurate forecasts, in this case ratios close to 1, and inaccurate forecasts, which are very large ratios or ratios close to zero. Thus, the ratio displays the accuracy of the forecast with respect to its reference point.

- Scaling: Not only does the f/r ratio show the difference, but it also shows how accurate or inaccurate the forecasts are. Using the logarithmic scale, the f/r plot depicts equal inaccuracies for under- and overestimation at an equal distance. That is, both a ratio of 1/2 and 2, which represent an equal under- and overestimation for positive entities, have an equal distance to the reference point in the plot. This equal distance is relevant as it allows for a visual assessment of the quality of forecasts and the severity of potential biases.

- Relative: By dividing the forecast with its reference point, the ratio becomes independent of the size of the asset. Namely, in case of positive valued entities, the reference point is often a good indication of the size. For instance, the higher the actual costs, the larger the project. Similar arguments can be made, for instance, for the project duration and effort of a project. The advantage of accounting for the size is that both small and large projects are assessed equally by allowing similar leniency. Thus, a ratio of 2 indicates the same for different sized projects.

  We note that we assume r to be a good approximation of size. If in certain cases r is not, the f/r ratio may not have the relative property.

This enumeration gives us a guideline to develop alternatives to deal with negative values. We wish to find an alternative that retains the listed properties and resolves the problems caused by negative values. Below, we address a number of ways, among others, alternative ratios, to cope with negative forecasts and reference points.

5.3.3.1 Alternative approaches

Below, we discuss a number of alternative approaches. We note that this enumeration is not exhaustive.

Mathematical translation A possible solution to deal with negative values is to perform a mathematical translation. A mathematical translation is known as the movement of a point, line, or shape a constant distance in a specified direction.

By adding a large number M to each forecast and reference point, it is possible to remove the negative values. For instance, let M be defined as the absolute of the minimum value of forecast and reference point in the data set plus 1. That is, $M = |\min(f, r)|+1$. Then, we add M to both forecast and reference point, effectively making the ratio $(f + M)/(r + M)$.

This translation solves the problems that arise from negative values by translating the ratios to the solution space of only positive valued entities. The translation has the properties difference and divisible, as it enables distinction between forecast and reference point.

However, due to adding a large number to both numerator and denominator, the ratio does not have the scaling or relative property. Although the ratio is translated to the solution space of positive entities, using the logarithmic scale does not yield similar results. To again allow for equal spacing, we would need to somehow translate the ratios back, so that it retains the same interpretation as the f/r ratio. However, how to do this is not clear.

Taking the difference An alternative approach is to take the difference between the forecast and the reference point. This alternative does not have a problem with negative values like the f/r ratio. If $f > r$ then $(f - r) > 0$ regardless of the signs, resulting in a positive valued ratio. If $f < r$ then $(f - r) < 0$ resulting in a negative valued ratio. Thus, the difference is either positive or negative depending on whether the forecast is larger or smaller than the reference point.

Besides the divisible property, the value also complies with the difference and scaling property. For a number of economic indicators, such as the ROI and IRR, it is sufficient to use this value to assess the forecast quality. Namely, the relative property for the $(f - r)$ value is needless as the underlying indicators are already relative themselves. For instance, the ROI relates the net benefits of the asset to its investment costs. Therefore, an ROI value of 10% means the same for an asset of 100 Euro and of 1 million Euro. As a result the $(f - r)_{\mathrm{ROI}}$ value is automatically relative. It is thus possible to analyze the forecast quality of these entities using the value $f - r$.

However, other economic indicators, such as the NPV, are not relative themselves. Namely, if the difference between forecasted NPV and reference NPV is 2 Euro, the interpretation is completely different when the reference point NPV is 100 Euro or 1 million Euro. Therefore, we need another alternative ratio that is relative for all entities.

Taking the relative difference An alternative to the $(f - r)$ value, is to divide the difference between forecast and reference point by an approximation of size, denoted by s. By dividing by an approximation of size, we make the differences between forecast and reference point relative. This means that the $(f - r)/s$ ratio has all the properties we listed for the f/r ratio and is thus suitable as an alternative. Furthermore, the $(f - r)/|s|$ ratio has the same interpretation as the f/r ratio. Namely, overestimations are represented above the horizontal line, while underestimations are depicted below this line.

There are many candidates for s. One example is the absolute size of the reference point, $|r|$. However, in that case the ratios do not necessarily indicate the same for a small and a large asset. For instance, the re-estimated NPV value does not adequately describe the size of an asset. Namely, suppose the re-estimation of the NPV results in an NPV of 1. It is possible for a large multi million Euro asset to result in an NPV of 1, but this can also apply to an asset of 100 Euro. Still, the ratio would yield the same result if the re-estimated NPV would turn out to be 2, even though the difference in the forecast is relatively small for the multi million Euro asset and relatively large for the 100 Euro asset. Therefore, in case of the NPV, this is not the best approximation.

For the $((f - r)/|s|)_{\mathrm{NPV}}$, other alternatives are, for example, the forecasts or re-estimations of the benefits, project costs, asset usage costs or asset costs. In case of entities that are already relative themselves, such as the IRR or ROI, the approximation of size s can be chosen 1, or $s = 1$.

Notation Above we introduced a family of alternative ratios that is usable for positive and negative valued NPVs. Since the notation is cumbersome, we introduce a new notation here. We will use the term $(f - r)_e/s$ ratio, when we analyze the $(f - r)_e/|s_d|$ ratio for entity e. In this case, s_d is the size of entity d and should always be further specified. We will use $(f - r)/s$ ratio, when it is clear which entity is referred to.

Approximating size We stated that candidates for approximations of size for the NPV are, for example, the forecasts or re-estimations of the benefits, project costs, asset usage costs or asset costs. The choice between these discounted benefits, project costs, asset usage costs and asset costs is arbitrary as long as they are adequate approximations of the size of an asset. In this chapter, we will choose the asset costs, as it incorporates all the costs that are made for the asset over its entire life span.

Then, the question remains whether we use the forecasted or re-estimated asset costs. That is, do we use f_{A} or r_{A}? This choice mainly depends on the purpose for which we wish to use the ratio.

The use of f_{A} is best if we wish to enhance decision information. Let us explain. With the $(f-r)/s$ ratio and using $s = f_{\mathrm{A}}$, it is possible to apply historical information to new project proposals, for instance, by means of a confidence interval. Suppose we have an 80% confidence interval of $[-0.25, 0.75]$ of historical $(f - r)_{\mathrm{NPV}}/s$ ratios for the benefits with $s = f_{\mathrm{A}}$. Consider a new project proposal that forecasted the NPV to be 5 million Euro and the total asset costs to be 2 million Euro. This would allow us to enrich our decision information in the following manner. We want to know r. We know that $(f - r)_{\mathrm{NPV}}/s = -0.25$ leading to $(5 - r)/2 = -0.25 \rightarrow r = 5.5$. For the upper bound of the range, we find $(5 - r)/2 = 0.75 \rightarrow r = 3.5$. Thus, based on historical data we would expect this new project proposal in 80% of the cases to yield an NPV between 3.5 and 5.5 million Euro.

This enhancement of decision information is not possible if we would use $s = r_{\mathrm{A}}$, instead of $s = f_{\mathrm{A}}$. For example, suppose we have a 80% confidence interval of the $(f - r)_{\mathrm{NPV}}/s$ ratio with $s = r_{\mathrm{A}}$ of $[-0.25, 0.75]$ for the NPV. Suppose we have a new project proposal that forecasts the NPV to be 5 million Euro and the asset costs to be 2 million Euro. We want to assess r, thus we would try $(2 - r)_{\mathrm{NPV}}/r_{\mathrm{A}} = -0.25$. Unfortunately, we cannot find r_{NPV}, since we do not know r_{A} at that time. Therefore, it is best to use f_{A} when we want to enhance decision information.

In Section 5.4 we will use $s = r_{\mathrm{A}}$ and in Section 5.5 we use $s = f_{\mathrm{A}}$.

Economic interpretation We propose to use the $(f - r)/s$ ratio with either $s = f_{\mathrm{A}}$ or $s = r_{\mathrm{A}}$ to analyze the forecast quality of the NPV. This ratio is not entirely arbitrary in the economic context of the NPV.

Suppose we would choose as approximation of size s the forecasted investment costs I of the asset. In this case, the $(f - r)_{\mathrm{NPV}}/s$ ratio is closely related to the ROI. We note that the ROI can be calculated using the formula $(f_{\mathrm{B}} - f_{\mathrm{C}})/f_{\mathrm{I}}$.

Let us explain. When we dissect the $(f - r)_{\mathrm{NPV}}/s$ ratio, we find that it consists of the difference between $f/|s|$ and $r/|s|$. Rewriting the ratio gives $f_{\mathrm{NPV}}/|s| = \frac{(f_{\mathrm{B}} - f_{\mathrm{C}} - f_{\mathrm{I}})}{f_{\mathrm{I}}} = f_{\mathrm{ROI}} - 1$. Similarly, $r_{\mathrm{NPV}}/|s| = \frac{r_{\mathrm{B}} - r_{\mathrm{C}} - r_{\mathrm{I}}}{f_{\mathrm{I}}}$. This fraction has the interpretation $r_{\mathrm{ROI}} - 1$. This means that using the $(f - r)_{\mathrm{NPV}}/s$ ratio with this approximation of size, we basically analyze $(f - r)_{\mathrm{ROI}}$.

Therefore, the proposed $(f - r)/s$ ratio for the NPV makes sense given the economic context. In this chapter, we will choose either the forecasted or re-estimated asset cost for s.

5.3.3.2 Extension

The proposed $(f - r)/s$ ratio makes it necessary to evaluate the way we construct the tools that we use to analyze forecast quality in case of negative values. In the previous section, we showed how to use the f/r plot, EQF_r and reference cone when assessing forecasts and reference points with non-negative values. There we altered the f/a ratio to account for other reference points than the actual, leading to the f/r ratio. However, the tools discussed there are not applicable in case of negative values. Therefore, we discuss how the f/r plot, EQF_r and reference cone should be altered to account for the proposed $(f - r)/s$ ratio in situations that negative values occur.

$(f - r)/s$ **plot** In the $(f - r)/s$ plot we depict the $(f - r)/s$ ratios against their progression relative to the reference point. Therefore, the $(f - r)/s$ plot has the same horizontal axis as the f/r plot. We use the date of the reference point t_r to scale the dates of the other forecasts.

The vertical axis of the $(f-r)/s$ plot depicts the $(f-r)/s$ ratios. We note that for these ratios the line $(f-r)/s = 0$ is the line between underestimation and overestimation. With the f/r ratio this is the line $f/r = 1$.

Furthermore, since negative values are possible for this ratio, the logarithmic transformation that we used for the f/r plot no longer applies. However, there is no problem abandoning the transformation in this case and use a linear axis, since the $(f-r)/s$ ratio allows for scaling. That is, for this ratio the distance between -2 and 2 on the linear axis indicates an equal accuracy of under- or overestimation.

$(f-r)/s$ ratio versus f/r ratio We described that we will use the $(f-r)/s$ ratio for entities with both positive and negative valued forecasts or reference points, and the f/r ratio for non-negative entities. One may wonder, why we do not used the $(f-r)/s$ ratio for the non-negative entities as well. For instance, instead of f/r we could use the $(f-r)/s$ ratio.

However, this is not always a good alternative for non-negative valued entities. Let us explain. Suppose we use the f/r ratio to assess the forecast quality of project costs. With the f/r ratio, all the underestimates, that is a forecast is smaller than its actual, will be in the range $[0,1]$, and all overestimations are in the range $[1,\infty]$. Using the logarithm, these ranges are made equal. Namely, the logarithm of the underestimations transforms the range to $[-\infty,0]$, and for the overestimates to $[0,\infty]$.

Now suppose we would use the $(f-r)/s$ ratio to assess the forecast quality of project cost. As the reference point of the project cost is a good approximation of size, the most likely choice for s is the reference point itself, thus $s = r$. With the $(f-r)/r$ ratio all the underestimates, that is forecast smaller than the actual, will be in the range $[-1,0]$, and all overestimation are in the range $[0,\infty]$. This range has no equal scaling. Since there are now negative values, it is not possible to use the logarithm to make the ranges comparable. Therefore, in such cases, the f/r ratio is more appropriate for non-negative valued entities.

We note that this argument holds when r is a good approximation of the size of the entity.

EQF_s The original EQF divided the size of the actual by the difference between the forecast and its actual. In this case, we use an alternative size measure to compare the forecast quality with. Therefore, we need to adapt the EQF for the new ratio. The EQF of the $(f-r)/s$ ratio, which we denote with EQF_s, is given by the following formula.

$$\mathrm{EQF}_s = \frac{\text{Area under } s}{\text{Area between forecast and reference value}}$$

$$\mathrm{EQF}_s = \frac{\int_{t_s}^{t_r} s\,dt}{\int_{t_s}^{t_r} |e(t) - r|\,dt} \tag{5.5}$$

$$\mathrm{EQF}_s = \frac{\int_{t_s}^{t_r} 1\,dt}{\int_{t_s}^{t_r} |(e(t) - r)/s|\,dt}. \tag{5.6}$$

This formula calculates 1 divided by the distance between the $(f-r)/s$ ratios and 0. This is also illustrated in Figure 5.7 that shows an example EQF_s calculation. The computation is similar to the EQF_r, which computes 1 divided by the distance between the f/r ratios and 1.

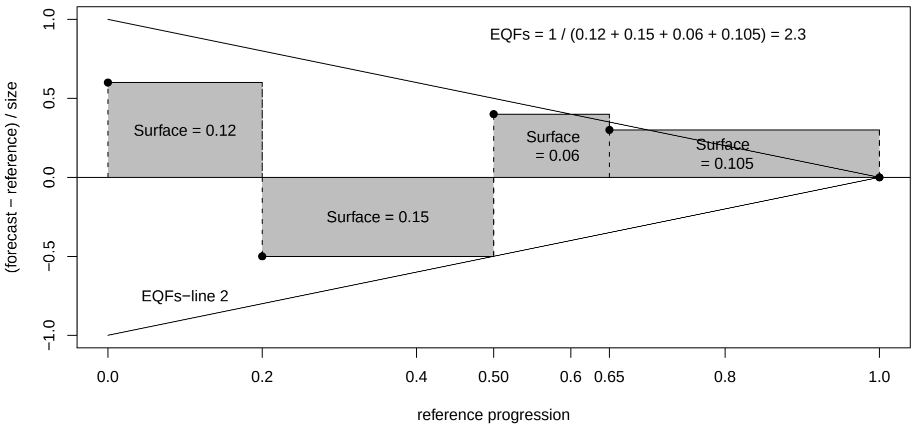


Figure 5.7: An example of an EQF_s calculation for four forecasts.

Reference cone Finally, we need to re-evaluate the reference cone as discussed in Chapter 3. The original reference cone no longer holds in this situation, since its reference lines are centered around an f/r ratio of 1 and are positive for both under- and overestimation. However, the $(f-r)/s$ plot is centered around 0 and depicts underestimations as negative ratios. Therefore, we need to recompute the reference lines.

Most assumptions discussed in Chapter 3 that were used to draw the reference lines, are still applicable. That is: we assume that each consecutive forecast incorporates the ex-post part and we assume this part is known with certainty; the accuracy of the ex-ante part is assumed to remain constant as the project progresses; and the goal of the forecast is to predict without bias and as quickly and as accurately as possible the actual value of interest for the project.

However, the ex-post growth assumption needs further consideration. We assumed the growth of the ex-post part to be described by a constant function. In many cases, the NPV will not comply to this assumption. Namely, during project execution investment costs are made until the time the project is completed. From that point on, benefits are potentially generated with the asset that together are hopefully higher than the asset usage costs. This causes the NPV to grow, which continues over the economic life span of the asset. This common growth of the NPV is not a constant growing function.

However, we will make the crude assumption that the NPV does grow following a constant function. We assume the initial project costs are made right at the start of the project and the benefits are generated immediately. That is, the NPV will grow constant from the moment the project starts. This means the ex-post follows the constant growing line from $-I$ to r. Even though this is a crude assumption, we use it since, due to its simplicity, it is useful to illustrate the calculations required to compute the reference lines. If other growth functions are used the calculations become more involved, but the method not.

We are able to derive alternative reference lines as follows. First, we need to specify the forecast at relative time x, $e(x)$. We note that $x = (t - t_s)/(t_r - t_s) \in [0, 1]$. Each forecast consists of an ex-post and an ex-ante part. For the ex-post part, we incorporate as much information as possible and we know this information exactly. We describe the ex-post part with a linear function that is given by $p(x) = (r + I)x - I$. Moreover, we define $p(0) = 0$.

The ex-ante part is not known with certainty. We assume we are able to predict the ex-ante part with a forecasting accuracy of c_1 for underestimation and c_2 ($c_2 \geq c_1$) for overestimation. This indicates that we are able to predict the ex-ante part within c_1 and c_2 times its value. The size of the ex-ante part is defined by the reference value minus the ex-post part. Mathematically, the ex-ante part is given by $r - p(x)$.

With the ex-post and ex-ante part defined, we are able to specify $e(x)$. In case of solely underestimations, $e_l(x)$ is given by $p(x) + c_1 \cdot (r - p(x))$. For solely overestimations, we find $e_u(t) = p(x) + c_2 \cdot (r - p(x))$.

However, the $(f - r)/s$ plot does not depict $e(x)$, but it illustrates $(f - r)/s$ ratios. Therefore, we need to formulate the reference lines in terms of $(f - r)/s$ ratios. This leads to the following formulas.

$$
\begin{aligned}
l(x) &= \tfrac{e_l(x)-r}{|s|} &= \frac{(r + I)x - I + c_1 \cdot (r - (r + I)x + I) - r}{|s|} \\
u(x) &= \tfrac{e_u(x)-r}{|s|} &= \frac{(r + I)x - I + c_2 \cdot (r - (r + I)x + I) - r}{|s|}.
\end{aligned}
$$

Preferably, similar to the current reference lines, we wish to express these lines in terms of EQF_s. We find the relation between EQF_s, c_1 and c_2 by computing the EQF_s. Below, we illustrate the calculations in case of solely overestimations.

$$
\begin{aligned}
EQF_{s,l} &= \frac{\int_0^1 1\,dx}{\int_0^1 |(r - e_l(x))/s|}\,dx \\
&= \frac{\int_0^1 1\,dx}{\int_0^1 \frac{r-(rx+Ix-I+c_1\cdot(r-rx-Ix+I))}{|s|}}\,dx \\
&= \frac{\int_0^1 1\,dx}{\int_0^1 \frac{r-rx-Ix+I-c_1r+c_1rx+c_1Ix-c_1I}{|s|}}\,dx \\
&= \frac{1}{[rx - \frac{1}{2}rx^2 - \frac{1}{2}Ix^2 + Ix - c_1rx + \frac{1}{2}c_1rx^2 + \frac{1}{2}c_1Ix^2 + c_1Ix/|s|x]_0^1} \\
&= \frac{|s|}{r - \frac{1}{2}r - \frac{1}{2}I + I - c_1r + \frac{1}{2}c_1r + \frac{1}{2}c_1I + c_1I} \\
&= \frac{2|s|}{(r+I)(1-c_1)}.
\end{aligned}
$$

From this formula, we find the following.

$$
\begin{aligned}
\frac{2|s|}{EQF_{s,l}} &= (r+I)(1-c_1) \\
\frac{2|s|}{(r+I)EQF_{s,l}} &= 1 - c_1 \\
\frac{2|s|}{(r+I)EQF_{s,l}} - 1 &= -c_1. \\
1 - \frac{2|s|}{(r+I)EQF_{s,l}} &= c_1.
\end{aligned}
$$

Via similar calculations, we find that $c_2 = 1 + \frac{2|s|}{(r+I)\mathrm{EQF}_{s,u}}$. If we substitute these values in the formulas of our reference lines, we find the following expressions.

$$
\begin{aligned}
l(x) &= \frac{(r+I)x - I + \left(1 - \frac{2|s|}{(r+I)\mathrm{EQF}_{s,l}}\right)\cdot(r-(r+I)x+I) - r}{|s|} \\
&= \frac{(r+I)x - I + r - (r+I)x + I - r}{|s|} \\
&\quad - \frac{\frac{2|s|}{(r+I)\mathrm{EQF}_{s,l}}(r-(r+I)x+I)}{|s|} \\
&= \frac{-2(r+I)(1-x)}{(r+I)\mathrm{EQF}_{s,l}} \\
l(x) &= \frac{-2}{\mathrm{EQF}_{s,l}}(1-x) && (5.7) \\
u(x) &= \frac{2}{\mathrm{EQF}_{s,u}}(1-x). && (5.8)
\end{aligned}
$$

These reference lines describe the reference cone that is usable for the $(f - r)/s$ plot. The calculations lead to relatively simple reference lines as also illustrated in Figure 5.7. This is caused by the modest assumptions that we used for the growth of the ex-post part. Using the above steps, it is possible to derive other reference lines given different assumptions for the growth of the ex-post part. As calculations can become involved with more complex assumptions, we recommend using computer algebra packages like Maple [87] to compute the results.

With this generalized method, we derived at an $(f - r)/s$ plot, EQF_s and alternative reference cone that allow for an assessment of forecast quality for negative valued entities.

5.3.4 Summary

In this section, we discussed that three problems may arise in applying the f/a ratio when we use it to quantify forecast quality. The f/a ratio runs into visual complications in asymptotic behavior; fails if the actual is not objectively measurable; and does not cope with the case of non-negative valued entities. To overcome these problems, we extended and generalized the method.

This generalized method is applicable for any entity. The tools to be used depend on the entity in question. In Table 5.1, we give an overview of the tools that are applicable in each situation. Below, we summarize these different situations.

Non-negative valued actual The method for non-negative valued entities with the actuals as reference point consists of an f/a plot, *EQF* and reference cone. The f/a plot depicts f/a ratios against the moment the forecast was made relative to the actual a. The f/a ratios on the vertical axis are drawn on a logarithmic scale.

Table 5.1: Overview of the tools that are applicable in different situations.

entity value	reference point	type of plot	EQF Formula	reference cone Formulas
non-negative	actual	f/a	3.1	3.4 and 3.5
non-negative	reference point	f/r	5.2	5.3 and 5.4
positive or negative	reference point	$(f-r)/s$	5.6	5.7 and 5.8

In case the forecast f is zero, the resulting f/a ratio is visualized by plotting the ratio one order below the minimum f/a ratio. That is, if the minimum f/a ratio is 1/2, the ratio is drawn at 1/20. For better distinction, this point should be depicted differently from the other ratios. When the actual a is zero, the f/a ratio is drawn a clear distance above the maximum f/a ratio. Finally, when both forecast and reference point are zero, the point is visualized by $f/a = 0/0 = 1$.

Non-negative valued reference point The method for non-negative valued entities consists of an f/r plot, EQF_r and reference cone. The f/r plot depicts f/r ratios against the moment the forecast was made, relative to the reference point r. The f/r ratios on the vertical axis are drawn on a logarithmic scale.

Comparing the forecast quality against a reference point r affects the analysis with respect to the f/a plot. In both the f/r plot and the f/a plot, it is possible to distinguish whether the forecasts converge to the reference point or not. If the forecasts do not converge, it is no longer possible to distinguish biases in the f/r plot. In case of convergence, the f/r plot shows a similar bias, which is generally less clearly pronounced than in the f/a plot.

Using data of four organizations, we found that the EQF_r is sometimes higher than the EQF and sometimes lower. Still, when we analyze the forecasting accuracy of all projects combined, the conclusions remain similar.

However, for individual projects it is no longer possible to identify outliers with respect to the forecasting accuracy. For this reason, the EQF values computed using reference points other than the actual are not suitable for benchmarking purposes between organizations. For assessing forecast quality, preferably the actual is used for the analyses instead of a more general reference point.

Positive and negative valued When entities can take both positive and negative values for both their forecasts and reference points, the method consists of the $(f-r)/s$ plot, EQF_s and an alternative reference cone given by Formula 5.7 for the lower reference line and Formula 5.8 for the upper reference line. The $(f-r)/s$ plot depicts $(f-r)/s$ ratios against the relative progression with respect to the reference point r. The approximation of size s of an asset needs to be specified for each figure. For the NPV, candidates are the forecasts or re-estimations of the benefits, asset usage costs, project costs or asset costs. In the analyses of the NPV, we will in this chapter use both f_{A} and a_{A}.

5.4 Prerequisite data analysis

In the previous section, we discussed the necessary tools to assess the forecast quality of the NPV, benefits, asset usage costs and project costs. In this section, we describe the data that we obtained from the large telecommunication organization Z. An investigation of this data is prerequisite for analyzing the forecast quality of the organization in the next section.

First, we give an overview of the data we obtained. Since this data involves reference points and not actuals of the NPV, we discuss at which point in time the re-estimations are made. Next, we assess the data of organization Z by comparing it to benchmark data from the literature. Finally, since our tools assume homogeneous data, we check for heterogeneity in the data set provided.

In this section we will make use of statistical tests to statistically verify possible heterogeneity. For all the tests we will perform, we use a threshold of $\alpha = 0.05$ for accepting or rejecting the hypothesis that is being tested.

5.4.1 Data overview

From a large telecommunication organization Z, we received forecasts and re-estimations of NPVs of 102 projects. These projects were executed in the period of 2001 till 2009. In total, the overall benefits amount to 4714 million Euro and asset costs of 2908 million Euro, which consist for 173 million Euro of investment costs. The median project duration is 327 days and the assets have a median expected economic life span of 5.5 years.

More precisely, the data consists of the following components for each project.

- the actual startdate (t_s) of the project.
- the actual enddate (t_e) of the project.
- the forecasted enddate (t'_a) of the asset.
- the initial forecast date (t_f) of the NPV.
- the re-estimation date (t_r) of the NPV.
- the forecasted asset life span (N).
- the forecasted and re-estimated NPV.
- the forecasted and re-estimated benefits (B).
- the forecasted and re-estimated asset usage cost (C).
- the forecasted and actual project cost (I).
- An asset classification: New Product Development, Cost Reduction or Sales Expansion.

In the above enumeration, the benefits, asset usage costs and project costs refer to the cumulated discounted totals. In our case study, these totals are all discounted using the same discount rate. The forecasts and re-estimations of the totals of an asset are discounted relative to the same point in time in that asset's life span. In Formula 5.1 in Section 5.2.2, we denoted them with B, C and I.

The data contain a large number of assets, that in their re-estimation have zero benefits. A number of IT projects was canceled or was completed successfully, yet the assets they created no longer had any business value. For instance, it is possible projects were stopped during implementation due to changes in the market. These assets are simple to recalculate, namely the benefits are zero in the post calculation. They comprise a relative large portion of the re-estimated NPVs. This confirms that forecasts and actuals with a value of zero are more common than you would expect.

We do note that if a project was stopped, for instance, due to changes in the market place, in our analyses the forecasts of its entities may be considered inaccurate. Since the estimators provide single point forecasts, these predictions cannot account for the possibility of stopping the project. These forecasts will be regarded as inaccurate, even though the cancelation of the project may be beneficial to the organization. Those responsible may be disinclined to cancel a project to avoid inaccurate forecasts. Organizations should make sure that the accuracy of forecasts does not discourage the cancelation of projects when necessary.

We emphasize that, for the NPV, benefits and asset usage costs, the re-estimations are available and not the actuals. Although the data spans a period of 2001 till 2009, it does not contain actual NPVs as the organization only recomputes them before the economic life span has ended. This emphasizes once more that collecting data with truly finalized NPVs is rather scarce.

5.4.2 Re-estimation date

The data we obtained consists of re-estimated NPVs, benefits and asset usage costs. To assess the forecast quality, it is preferable that the ex-post part of the re-estimations is as large as possible. That is, the re-estimation should be made shortly before the final realization. However, from the perspective of gaining relevant decision information, we would prefer the reference point to be made more closely after project completion. Therefore, it is interesting to investigate at which moment the re-estimations are made in the data set.

Organization Z re-estimated the benefits and asset usage costs some time after their initial forecast. The re-estimations were made a median 1.4 years after the initial forecast and a median 0.71 years after project completion. The date was also a median 3.5 years before the asset's predicted economic life span.

To further investigate the re-estimation date, we also assess the relative moment of re-estimation. We define the relative moment of re-estimation as the relative time of the re-estimation over the entire life span of the asset. Mathematically, this moment is given by $(t_r - t_s)/(t_a - t_s)$ with t_s the start date of the asset, t_r

the date of re-estimation and t_a the end date of the assets life span. We note that in this chapter t_a is given by the number of years over which benefits and costs are predicted.

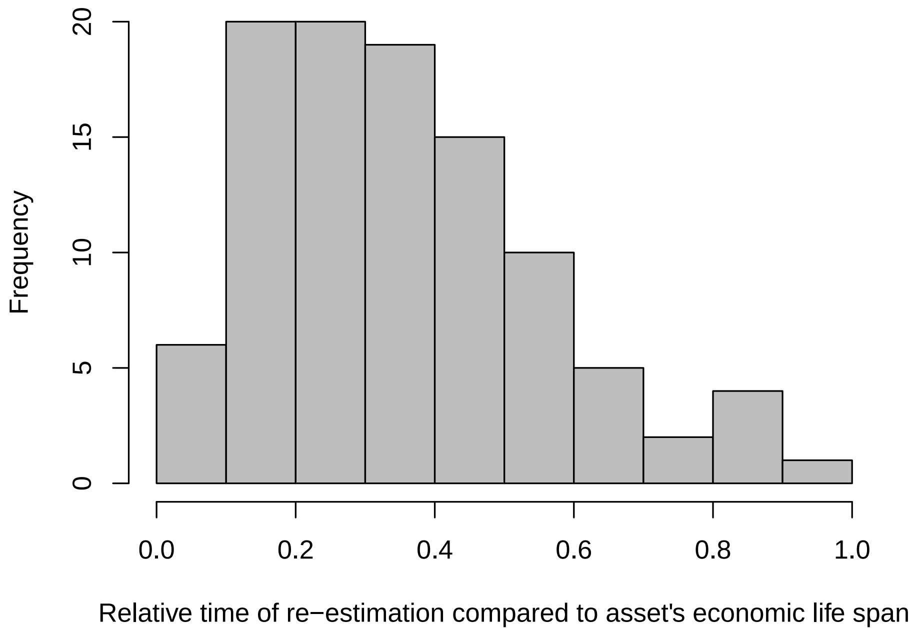


Figure 5.8: Distribution of the relative time of the re-estimations with respect to the predicted asset's economic life span.

In Figure 5.8, we show a histogram of the relative time of the re-estimations compared to the predicted asset's economic life span. For instance, a value of 0.4 means the re-estimation of the NPV was made at 40% of the predicted economic life span of the asset. The figure shows a large spread in the moment of re-estimation. Many re-estimations are made in the range 20%–60%.

5.4.3 Literature benchmark

To assess the quality of the data of our real-world case study, we benchmark it against a case study found in the literature. The only benchmark related to the forecast quality of NPVs we are aware of, is given in a book by Bower [79]. In this book, results of the difference between forecasted and actual NPVs are described for a case study consisting of 50 assets.

Table 5.2: Summary of a/f ratios of NPVs for in total 50 Cost Reduction, Sales Expansion and New Product Development assets as described in a book by Bower [79].

Type of asset	Mean a/f ratio	standard deviation
Cost reduction	1.1	narrow
Sales expansion	0.6	reasonable ≈ 0.2
New products	0.1	very large

Table 5.3: Summary of the r/f ratios of in total 100 assets for our telecommunications organization. For sake of comparison two assets are left out, since their NPV forecast is zero. These results should not be used for other comparisons.

Type of asset	mean r/f ratio	standard deviation	median r/f ratio	number of assets
Cost reduction	2.32	5.00	1.03	35
Sales expansion	1.07	0.66	0.81	4
New products	0.42	4.95	0.11	61

For the sake of completeness, we show Table 5.2 taken from the book. The table contains the mean a/f ratio of the NPV for three asset classifications. Unfortunately, the book reveals little detail about the analysis. Through personal communication, we understood that the underlying data is no longer available. Still, the author was able to give an indication of the variation of the figures, allowing us to add the column *standard deviation*.

However, we must be cautious about this benchmark for three reasons. First, since the book was published in 1970, the assets are not IT assets. Therefore, these findings may have limited application in our IT-context.

Second, the case study examines the a/f ratio of the NPV. In Section 5.3, we discussed that this ratio is ambiguous when used for entities with both positive and negative outcomes, such as the NPV. Bower communicated to us, that the case study contained negative actuals, yet, not how many or their impact on the results. Therefore, the results depicted in the table do not reveal anything about potential biases or the quality of the forecasts.

Third, the case study of Bower uses the actual as reference point. We assume that these are objectively measured final realizations. In our case study, we use re-estimations of the NPV as our reference points.

However, we are still able to use the benchmark to assess whether our data behaves similarly for this ratio. The data we obtained contains twice the amount of assets as Bower and uses the same asset classification. In Table 5.3, we show summaries of the r/f ratios of our case study. We left two assets out of this overview, since there was a Sales Expansion and a New Product Development asset with a forecasted NPV of zero. We note that we use the r/f ratio and not the $(f - r)/s$ ratio we proposed as alternative ratio for the NPVs for the sake of comparison.

Both the mean and the standard deviations of the r/f ratios of our case study do not resemble the a/f ratio benchmarks of Bower. For all classifications, the mean of the ratios is higher than those observed by Bower. Moreover, in our case study, the standard deviation of the Cost Reduction assets and the New Product Development assets are comparable. The literature benchmark on the other hand, found this standard deviation to be quite different for these classifications. This indicates that our case study involves more variation than found in the literature benchmark. Because the means in our case study are significantly influenced by the variation in the data, we also incorporated the median value of the r/f ratios.

Based on the findings, the data behaves differently to the benchmark of the literature. Both the mean and the variation of the distributions in our data set are considerable larger.

Still, similar to Bower, we find that the categories are different. In both studies, the mean r/f ratio or a/f ratio of the Cost Reduction assets is larger than that of the Sales Expansion and New Product Development assets. This indicates that our data set is possible heterogeneous.

5.4.4 Heterogeneity

Finally, we analyze the data for heterogeneity. The proposed tools to analyze forecast quality, assume that the underlying data is homogeneous. Therefore, before we are able to apply the tools, we verify if the data complies to this assumption.

Namely, forecasts may have different uncertainties for different types of assets. For instance, a forecast for the costs of a standard application may have less uncertainties than a forecast for a newly developed product. Therefore, a deviation of 2 times the actual may be a high quality forecast for one asset, whereas it is a poor forecast for another given its range of possible outcomes. In case of a significant difference between certain assets, the forecast quality should be assessed separately. Of course, a discussion is in order as to whether the found difference is justified or is simply caused by making less accurate predictions.

The data we obtained, contains a number of variables that are interesting to investigate in this respect. In the comparison with a data set from the literature, we found that there is a potential significant difference between the asset classifications. Therefore, first, we determine whether the different asset classifications cause heterogeneity in the data set. Moreover, we investigate whether the forecast quality is correlated with the moment of re-estimation or contains a yearly trend.

5.4.4.1 Asset classification

Organization Z assigned each asset to one of three categories, being: *Cost Reduction*, *Sales Expansion* or *New Product Development*. The categories consist of 35 Cost Reduction assets, 5 Sales Expansion assets and 62 New Product Development assets. Unfortunately, the amount of assets available in the Sales Expansion category is insufficient to investigate for significant differences with the other categories. Therefore, we leave this category out of the analysis.

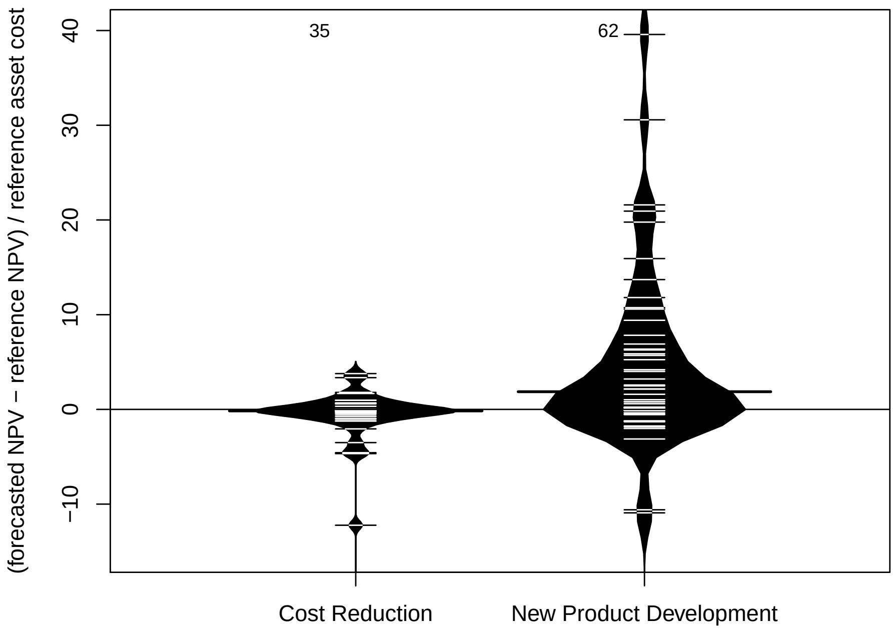


Figure 5.9: Potential heterogeneity between Cost Reduction assets and New Product Development assets with respect to the forecast quality of the NPV. Each line in the beans [65] represents a single forecast. The bean shape is an approximation of the underlying probability distribution of the data. The large horizontal line in each of the bean plots illustrate the median ratio of the data.

To detect potential heterogeneity, we plot the $(f-r)/s$ ratios of the NPV forecasts for both Cost Reduction and New Product Development assets in Figure 5.9. In this case, we use $s = r_{\mathrm{A}}$, since we only want to assess whether differences exist. At this time, we are not interested in enhancing decision information. Therefore, we do not need to use the initial forecast of the asset costs, but are able to use the best approximation of the size that is available.

The figure depicts two so-called bean plots [65], one for each category. The vertical axis shows the $(f-r)/s$ ratios of the NPVs. Each line in the beans represents a single forecast. The bean shape is an approximation of the underlying probability distribution of the data. The large horizontal line in each of the bean plots illustrate the median ratio of the data. A couple of extreme values are not depicted in the range we chose for this figure.

The bean plots illustrate that the location of the distributions is different. The Cost Reduction assets appear unbiased, whereas the New Product Development assets show a slight bias toward overestimation. We used the Kruskal-Wallis test and the Wilcoxon test to verify whether the location of the distributions is significantly different. As shown in Table 5.4, with a p-value of 0.00018 and a threshold of $\alpha = 0.05$ for the Kruskal-Wallis test, we reject the hypothesis that the location of the distributions is equal. A one-sided Wilcoxon test results in a p-value of 0.000009, which rejects the hypothesis that the location of the Cost Reduction assets is greater than or equal to that of the New Product Development assets. Thus, we find there is a significant difference of location for the $(f - r)/s$ ratios of the asset classifications.

Table 5.4: An overview of the outcomes of the statistical tests for the locations of the distributions for Cost Reduction and New Product Development assets.

same locations	Kruskal-Wallis test	reject / accept	Wilcoxon test	reject / accept
NPV	0.00018	reject	0.000009	reject
benefits	0.0004	reject	0.0002	reject
asset usage cost	0.003	reject	0.002	reject
project cost	0.51	accept	0.26	accept

Also, the bean plots show the variance of the $(f - r)/s$ ratios for the Cost Reduction category is smaller than that of New Product Development. We tested whether the variance is significantly different using the Brown-Forsythe test. As shown in Table 5.5, with a p-value of 0.20 for the Brown-Forsythe test, we cannot reject the null hypothesis. Therefore, the variance of the asset categories is not statistically different.

Table 5.5: An overview of the outcomes of the statistical tests for the variation of the distributions for Cost Reduction and New Product Development assets.

same variance	Brown-Forsythe test	reject / accept
NPV	0.20	accept
benefits	0.03	reject
asset usage cost	0.42	accept
project cost	0.08	accept

We also analyzed the f/r ratios of benefits and asset usage costs and the f/a ratios of project costs for heterogeneity with respect to the asset classifications. The analyses of the f/r ratios of benefits show similar results as the $(f-r)/s$ ratio of the NPVs as illustrated in Tables 5.4 and 5.5. The tests indicate that the location of the benefit f/r ratios shows significant differences between the categories. Also, the variance of the asset categories for the benefits turns out to be significantly different. This indicates that not only the location, but also the accuracy of the benefit forecasts differs between the asset categories.

The tests for the asset usage cost f/r ratios show heterogeneity in location, but not for the variation. The f/a ratios of project costs show no heterogeneity for the asset classifications. Moreover, we do not find a statistical difference of the variance of the f/a ratios of project costs. In this case study, the forecast quality of the project execution is not dependent on the type of assets that is being made.

Based on these findings, we conclude that the data of the forecasted NPVs, benefits and asset usage costs are heterogeneous for the asset classifications Cost Reduction and New Product Development in this case study. Therefore, we have to analyze the forecast quality of the NPV, benefits and asset usage costs for each of these classifications separately. When analyzing the forecast quality of project costs in our case study, there is no necessity to assess the asset classifications separately, since we found no proof of heterogeneity. However, we will still do so, to make the different assessments analogue.

Also, due to the low number of Sales Expansion assets, in the remainder of this chapter, we focus on the forecast quality of Cost Reduction and New Product Development assets.

5.4.4.2 Re-estimation date

In Section 5.4.2, we discussed the re-estimation dates in the data set and found that the data has a large spread in the relative time of the re-estimations. Therefore, here we assess whether this spread results in heterogeneity of the data. For instance, it is possible that earlier re-estimations result in more accurate ratios as things may not have changed as much. We investigate whether the $(f-r)/s$ ratio of the NPV and the f/r ratios of benefits and asset usage costs are influenced by the timing of the re-estimations. We note that we do not assess the f/a ratios of project costs, since these are actuals and not re-estimations.

To investigate for correlations, we plotted the ratios against the relative time and statistically confirmed the findings using Kendall's correlation coefficient test. We note that we performed the analyses for the combined asset classifications and the classifications separately. Most of these analyses show no correlations for the NPV, benefits and asset usage costs and the relative time of the re-estimations.

In two cases, Kendall's correlation coefficient indicated significant correlations. However, these correlations are influenced by the asymptotic behavior of numerous f/r ratios in the data set. Let us explain. We illustrate the two analyses in Figure 5.10. The left-hand plot depicts on the vertical axis the f/r ratios of the benefit forecasts for New Product Development assets. The vertical axis of the right-hand plot shows the f/r ratios for the asset usage costs for New Product Development assets. The horizontal axes represent the relative time of re-estimation. The dashed diagonal line is the potential correlation of the data. This line is computed without taking the asymptotic cases of zero forecasts or zero reference points into account. Namely, the correlation test is unable to incorporate data points with value infinite.

Kendall's correlation coefficient test gives a p-value of 0.033 for the benefits

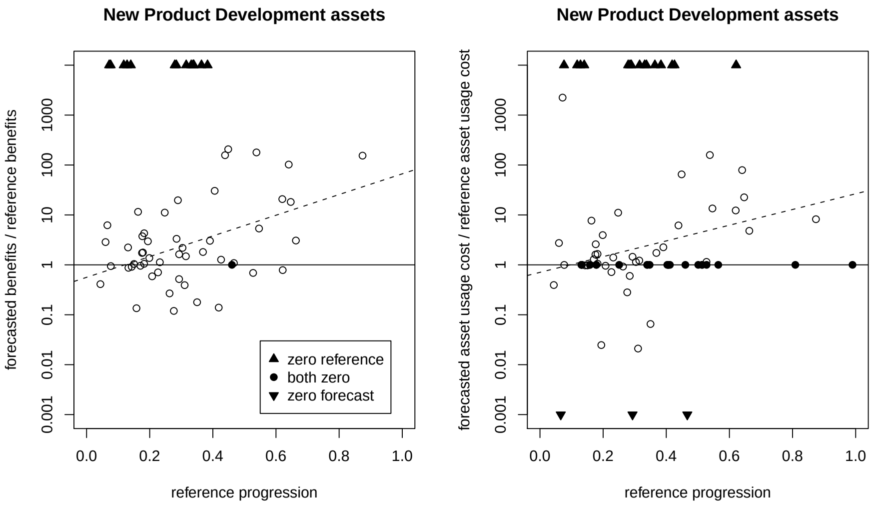


Figure 5.10: No correlation between f/r ratios of benefits and asset usage costs and the relative time of re-estimation.

and 0.015 for the asset usage costs. This indicates that the correlation is statistically significant given the threshold $\alpha = 0.05$.

However, the plots also illustrate numerous asymptotic f/r ratios that are not taken into account in the test. For instance, in the left-hand plot we find many assets that have no benefits in their re-estimations. In the right-hand plot, we also find numerous assets that both have a zero forecast and reference point. For instance, a number of IT projects were completed successfully, yet the assets they created no longer had any business value. These asymptotic f/r ratios highly influence our test for correlation.

Based on the analyses and accounting for the asymptotic cases, we concluded that there is no conclusive evidence for significant correlation between the forecast quality of the NPV, benefits and asset usage costs, and the relative time at which the re-estimation is made. These analyses once more underscore the relevance of depicting ratios with asymptotic behavior when assessing forecast quality.

5.4.4.3 Yearly trend

Besides the asset classification and re-estimation date, we also analyzed the data for a potential trend of the forecast quality. Since we obtained data over 9 years, we analyzed whether forecasts of older assets had a significantly different forecast quality than forecasts of newer projects. We do note that the available data for each year is relatively small. The figures and tests give no indication there exists a significant time-dependence in the forecast quality for either the Cost Reduction projects or the New Product Development projects.

5.4.5 Summary

We found the data to be heterogeneous with respect to the asset classification. For Cost Reduction and New Product Development projects, the forecast quality of the NPVs and its components, the benefits and asset usage costs are significantly different. Although the project costs showed no significant difference, to remain consistent, we will assess their forecast quality separately for each asset classifications.

There were too few Sales Expansion assets to find differences with this category. Therefore, in the remainder of our analyses, we do not consider Sales Expansion assets.

The investigation of other variables did not result in further heterogeneity. With this knowledge we are able to commence with the analysis of the forecast quality of the NPV, benefits, asset usage costs and project costs in the next section.

5.5 Case study

In the previous sections, we developed a generalized method and performed a prerequisite data analysis on the data that we obtained from organization Z. In this section, we will use the generalized method to assess the forecast quality of the NPV, benefits, asset usage costs and project costs. Due to the heterogeneity caused by the asset classification, we consider Cost Reduction and New Product Development assets separately. Moreover, we leave the five Sales Expansion assets out of the analyses.

First, we investigate the forecast quality of the NPV. Recall that the Net Present Value is a summation of the predicted monetary benefits and costs of a project discounted to current value. For this purpose, we use the $(f - r)/s$ ratio, the reference cone and the EQF_s. Recall that the EQF_s is a measure of the deviation between forecast and the reference value. As approximation of the asset size s, we use the first made forecast of the asset cost, or $s = f_{\mathrm{A}}$.

Next to that, we assess the forecast quality of the components of the NPV. We will analyze the f/r ratios for the benefits and asset usage costs, and the f/a ratios for the project costs.

Finally, we perform a sensitivity analysis to investigate the influence of the components on the forecast quality of the NPV. We assess what the impact is on the forecast accuracy of the NPV forecasts, when the forecast quality of each of the components changes.

5.5.1 NPV

First, we investigate the forecast quality of the NPV. In Table 5.6, we provide an overview of the forecasted and re-estimated NPVs of the 97 IT assets.

The table shows that the total of the initial forecasted NPVs in the Cost Reduction category is $194/179 \approx 1.1$ times higher than the total of its re-estimated NPVs. For the New Product Development assets, the total of the forecasts is

Table 5.6: An overview of forecasted and re-estimated NPVs in millions of Euro.

	Cost Reduction	New Product Development
Number of projects	35	62
Total of initial forecasted NPVs	194	2426
Total of re-estimated NPVs	179	1606
Median initial forecasted NPVs	2.0	11
Median re-estimated NPVs	2.3	1.5

$2426/1606 \approx 1.5$ times higher. The medians of the initial forecasts and the re-estimated NPVs show a larger difference. The median forecasted NPV of the Cost Reduction assets is slightly lower, namely $2.0/2.3 \approx 0.87$ times, than its median reference point. Yet, for the New Product Development assets, it is $11/1.5 \approx 7.3$ times higher. At first sight, these statistics indicate that the forecasted NPVs of the New Product Development assets follow an optimistic pattern. The Cost Reduction assets appear unbiased.

To further investigate the initial forecasted NPVs, we have drawn the $(f - r)/s$ plot with $s = f_A$ in Figure 5.11. The reference cone that is drawn, is given by Formulas 5.7 and 5.8 and the EQF_s is as defined by Formula 5.6.

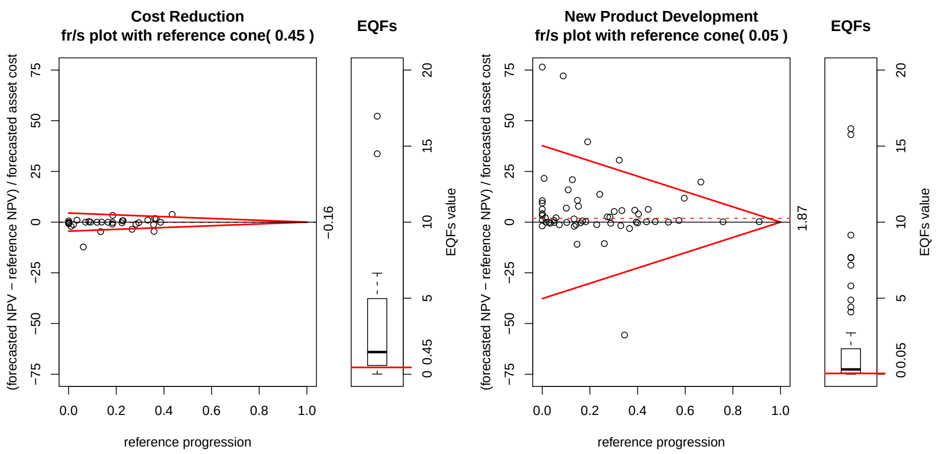


Figure 5.11: $(f - r)/s$ plot, with $s = f_A$, reference cone and EQF_s box plot of the forecasted NPVs for Cost Reduction and New Product Development assets.

Let us explain the figure. The left-hand plot shows the forecast quality of the Cost Reduction assets and the right-hand plot that of the New Product Development assets. The horizontal axis of each figure depicts the reference progression. For instance, suppose a re-estimation was made a year after the start of the asset. And suppose the initial forecast was made four months after the start. This forecast would be depicted at $4/12 = 0.3$ or 30% reference progression. The vertical axis displays the $(f - r)/s$ ratios.

The lines in the figures are the reference lines drawn using Formulas 5.7 and 5.8. The EQF_s value that we chose is the 20% quantile of all EQF_s values. This value indicates the quality that 80% of the assets is able to obtain. The EQF_s box plot gives the EQF_s quality of the forecasts made.

Finally, the dashed horizontal lines in the figures represent the median $(f - r)/s$ ratios of the data. This is also indicated by the value to the right of each plot. For instance, in the left-hand plot, the median $(f - r)/s$ ratio is -0.16.

Most notable, the figure illustrates the difference in accuracy of the forecasted NPVs of the asset categories. The left-hand plot of the Cost Reduction assets show less spread in their $(f-r)/s$ ratios than the New Product Development assets. This is confirmed by the median EQF_s of 1.46 for the former category and 0.32 for the latter. The data shows that the organization is able to more accurately predict Cost Reduction assets than New Product Development assets.

For the Cost Reduction assets, we find a small pessimistic bias indicated by the figure and the median $(f - r)/s$ ratio of -0.16. On the other hand, the New Product Development assets show an optimistic bias in the figure combined with a median $(f-r)/s$ ratio of 1.87. We note, although it does not appear so in the figure, this is a relevant bias. For instance, suppose a new project proposal predicts an NPV of 3.5 million Euro with asset cost of 2 million Euro. The bias indicates that this forecast is possibly overestimated by $1.87 \cdot 2 = 3.74$ million Euro. A relevant deviation indeed, as this indicates the project proposal is likely to result in an NPV of $3.5 - 3.74 = -0.24$ million Euro in this example.

The above analyses show that the organization is able to predict NPVs of Cost Reduction assets almost without bias. Moreover, the accuracy of the Cost Reduction forecasts is better than the forecasts of New Product Development assets. The latter forecasts show an optimistic bias.

5.5.2 Benefits

One of the components of the NPV is the cumulated discounted benefits denoted by B in Formula 5.1. We analyze the forecast quality of these benefits using the f/r plot, reference cone and EQF_r depicted in Figure 5.12. These plots are similar to those we showed for the NPV. However, the reference lines are drawn using Formulas 5.3 and 5.4. We note that for this lower reference line, a minimum EQF_r value of 2 is required. Therefore, in the plots we use $\mathrm{EQF}_r = 2$ for the lower reference lines.

Moreover, the plots also visualize f/r ratios in asymptotic cases, which were not present for the NPV. The f/r ratios described by reference benefits of zero are shown as solid upper triangles. They represent ratios that are infinite, since the forecast is positive and the reference value is zero. If both forecast and reference are zero, the f/r ratio is defined by $0/0 = 1$ and is depicted with a solid dot. Finally, in case the forecast is zero, the f/r ratio is zero, indicating the forecast is underestimated, and is visualized by a downward solid triangle.

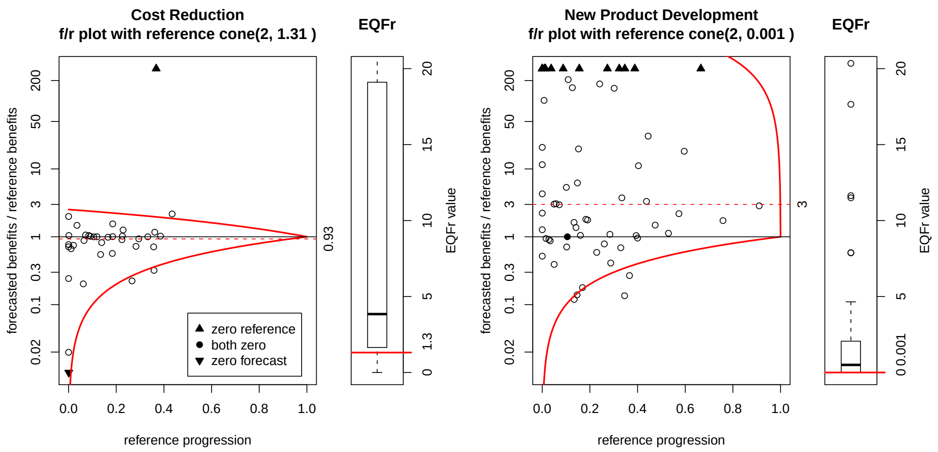


Figure 5.12: f/r plot with reference cone and EQF_r box plot of the forecasted benefits for Cost Reduction and New Product Development assets.

The left-hand f/r plot of the Cost Reduction assets shows no clear bias, similar as for the NPV. The median f/a ratio of the data reveals only a slight bias toward underestimation with a value of 0.93. The forecasts have a median EQF_r of 3.85 indicating that they have an average time-weighted deviation of $1/3.85 \approx 26\%$ to their reference point. We note that in this case, we only have the initial forecast, thus the median deviation to the reference point is also 26%. It is not possible to compare this forecasting accuracy with that of the NPV, since the EQF_s and EQF_r are different. Moreover, we are unaware of other benchmarks to compare these EQF_r values with.

The right-hand plot of the New Product Development shows a larger spread of the f/r ratios than that of the Cost Reduction assets. This is reflected in the forecasting accuracy in terms of EQF_r. With a median EQF_r value of 0.50, the forecasts have a median time-weighted average deviation of $1/0.5 = 200\%$ from their reference point.

The forecasts have an optimistic bias, which is illustrated by the figure and a median f/r ratio of 3. That is, the median of the initial forecasted benefits of New Product Development assets is three times as high as the median of their reference point.

Partly, this is caused by the fact that for a large number of assets the re-estimated benefits are zero. A number of IT projects were canceled or completed successfully, yet the assets they created no longer had any business value. For instance, it is possible projects were stopped during implementation due to changes in the market. The benefits of these assets are simple to recalculate, namely the benefits are zero in the post calculation. They comprise a relative large portion of the re-estimated assets. This confirms that forecasts and actuals with value zero are more common than you would expect.

We do note that if a project was stopped, for instance, due to changes in the market place, in our analyses the corresponding forecasts may be considered inaccurate, even though the accuracy of the forecasts is not the reason for stopping the project. It can be considered unfair to hold the estimators accountable for an inaccurate forecast in this case. Therefore, an organization should carefully consider in these cases whether to judge the forecasts differently or not.

The results of this analysis compare well with the findings for the NPV. The Cost Reduction assets show no particular bias and are more accurately predicted than the New Product Development assets. The quality of the New Product Development forecasts is relatively low and shows a bias toward overestimation.

We note that poor benefit forecasts do not by definition imply forecasting errors of the business domain estimators or the IT domain estimators. The method indicates which forecasts were less accurate than others, yet do not answer the question what the reason is for these inaccuracies. For instance, an inaccurate forecast of the benefits may be caused by an underestimation or overestimation of the sales generated using the asset. Or, no benefits are generated because the asset could not be developed by the IT department. Therefore, inaccurate forecasts have different causes. The method only indicates which forecasts were inaccurate. To determine the cause of the inaccuracies, additional analyses are required.

5.5.3 Asset usage cost

Another component of the NPV is the cumulated discounted asset usage costs. In Formula 5.1 we denoted these usage cost by C. Figure 5.13 shows the f/r plot, reference cone and EQF$_r$ box plot of their forecasts for both the categories Cost Reduction and New Product Development.

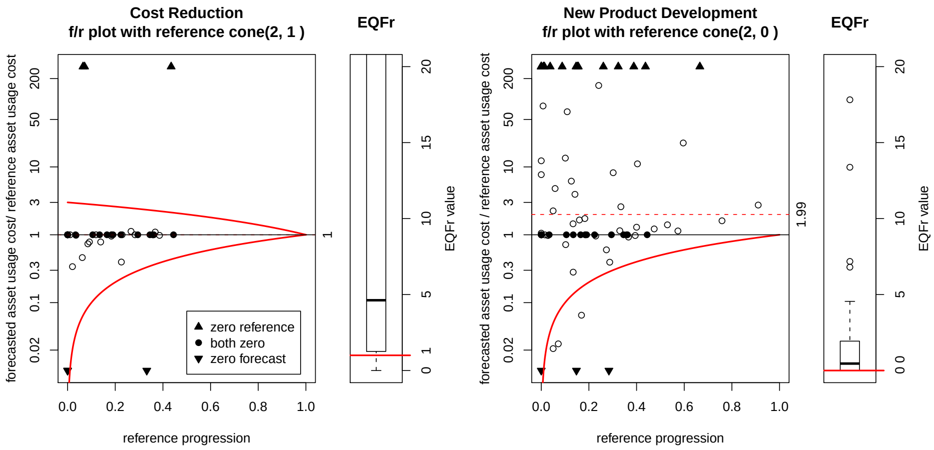


Figure 5.13: f/r plot with reference cone and EQF$_r$ box plot of the forecasted asset usage costs for Cost Reduction and New Product Development assets.

The plots show a large number of f/r ratios for both categories, of which the forecast and reference points are zero. Again, we also find several assets with re-estimated zero reference points. For instance, assets that were stopped during implementation. In these cases, no asset usage costs needed to be made, making the re-estimations zero.

The plots show that the forecasts of the Cost Reduction asset have no bias which is supported by a median f/r ratio of 1. The forecast quality with a median EQF_r value of 4.6 is higher than that of the benefits forecasts.

Similar to the benefits, the asset usage cost forecasts of the New Product Development assets show a bias toward overestimation in the figure and a median f/r ratio of 1.99. The similarity of the bias is partly caused by the relation between the asset usage costs and the benefits. The asset usage costs comprise of direct costs. These direct costs are costs that are incurred in proportion to the activity of the business and can be associated to a particular cost object. For example, if a telephone contract is sold, benefits are generated, yet also direct costs are incurred for, for instance, handset subsidies and interconnect costs. Due to these direct costs, a relation exists between the asset usage costs and the benefits. Therefore, it makes sense to have similar bias and forecast quality for these components.

The quality of the New Product Development forecasts remains low with a median EQF_r value of 0.46. For this category, there is hardly any difference in accuracy between the benefits and asset usage costs.

Summarizing, the analyses of the asset usage costs show similar results as found for the benefit forecasts. The assets of the Cost Reduction category have unbiased forecasts that are of reasonable quality. The New Product Development assets on the other hand have an optimistic bias and have a larger deviation of the forecasts to the reference points.

5.5.4 Project cost

Finally, we investigate the forecast quality of the cumulated discounted project costs denoted by I in Formula 5.1. Figure 5.14 depicts the f/a plot, reference cones and EQF box plots for both asset classifications. We note that for the project costs, we are able to use the actuals as reference point. Therefore, the horizontal axes of the f/a plots depict the project progression instead of reference progression.

The f/a plot of the Cost Reduction assets has a median f/a ratio of 1.05 indicating no significant bias in the forecasts. The forecasting accuracy shows a median EQF value of 10.3. This means the average time-weighted deviation is only 9.7% from the actual.

The f/a plot of the New Product Development assets shows a wider scatter of the f/a ratios than the f/a plot of the Cost Reduction assets. With a median EQF value of 2.0 the quality of these forecasts is significantly less than that of the Cost Reduction assets. The median deviation of the forecast to the actual is 50%. Still, the forecasts are unbiased as shown in the figure and the median f/a ratio of 1.

The forecasts of the project costs have a higher accuracy than those of the benefits and asset usage costs for both categories. However, while the forecasts

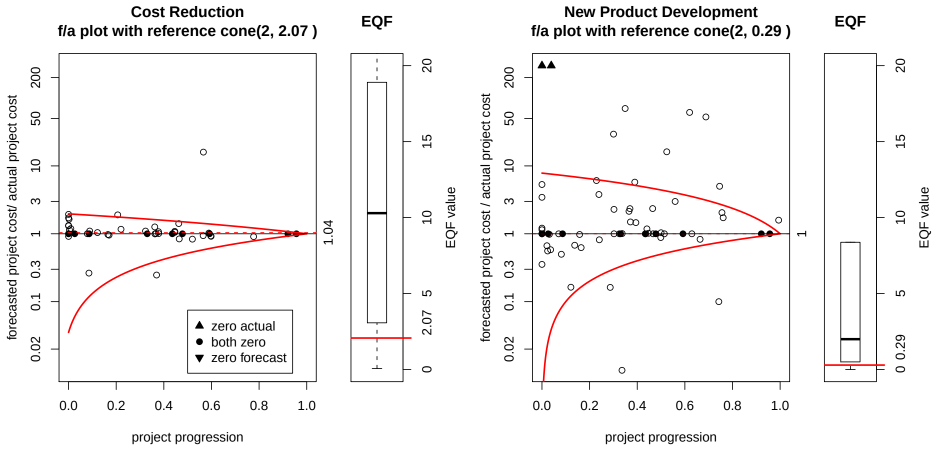


Figure 5.14: f/a plot with reference cone and EQF box plot of the forecasted project costs for Cost Reduction and New Product Development assets.

of the benefits and asset usage costs for New Product Development showed an optimistic bias, we find no bias for the project cost forecasts. Apparently, the organization has found ways to adequately predict the project costs, yet has more difficulties assessing the benefits and asset usage costs.

Benchmark Since we analyzed f/a ratios for the project costs, we are able to compare the quality of these forecasts against the forecast quality of other organizations described in the literature. We will compare our case study with different benchmarks in the literature.

The first comparison is to a literature overview of forecast quality in terms of EQF given in Chapter 3. To ease readability, in Table 5.7 we reiterate that overview.

The benchmark figures in the overview concern only completed projects. However, in the case study of organization Z, there are also canceled projects. To make a fair comparison, we therefore removed the assets that were stopped.

We note that in case of canceled projects, it may be beneficial to have very inaccurate forecasts. Namely, the sooner the project is stopped, the less money may have be spent and the more the forecast will deviate from the actual. Although stopping a project sooner rather than later results in less accurate forecasts, the organization will have lost less money.

We added the results of organization Z to the table. The label *Organization Z - NPD* shows the project cost forecast quality of the New Product Developments assets and the label *Organization Z - CR* for the Cost Reduction assets. However, the case studies of the literature make no distinction in the asset classification. Therefore, we also added the combined quality of both classifications. This is labeled *Organization Z - combined* for the quality of the combined categories. The table is sorted by the median EQF value. It is not possible to use this table for comparisons with the EQF_r or EQF_s.

In our case study, we analyze cumulated discounted project costs. The data in the literature does not contain discounted figures. However, discounting does not play a significant role for the results of organization Z, as the project duration is often smaller than a year. Therefore, it is possible to compare them to the literature data.

We also note that we are unable to give the mean EQF value for the different categories. This is caused by projects with both a forecast and an actual of zero. We defined the f/a ratio to be 1 in this case. Therefore, the forecast has an infinite EQF value. Moreover, in our case study of organization Z, only the initial forecasts were available. Therefore, the EQF values of organization Z are most likely lower than they could have been if multiple forecasts were made.

Table 5.7: Summary of benchmark EQF values found in the literature and our case study. It is not possible to use this table for comparisons with the EQF_r or EQF_s.

source	median EQF	average EQF	number of projects
Organization Z - CR	10.3	-	34
Kulk et al. [73]	9.4	-	221
Organization Y cost	8.5	36.9	140
Landmark Graphics 2003 [83]	8.4	-	-
Landmark Graphics 2002 [83]	7.6	-	-
Landmark Graphics 2001 [83]	7.0	-	-
Organization Y functionality	6.4	9.9	83
Organization Z - combined	6.0	-	83
Lister [82]	4–9	-	-
Organization Z - NPD	4.7	-	49
Landmark Graphics 1999 [84]	4.7	6.3	121
DeMarco - Little [84]	1.9	4.2	20
DeMarco [20]	-	3.8	-
Organization X	0.43	1.6	867

The table shows that the forecasts of project costs for the New Product Development category are of reasonable quality compared to other benchmarks. The forecasts of the Cost Reduction assets are of higher quality than the forecasts of other organizations in the overview. Especially considering the fact that we only account for initial forecasts, the quality of these forecasts are relatively high. We find the quality of the forecasts of the combined classifications comparable to the quality found by Lister and organization Y functionality. This indicates that the telecommunication organization Z has reasonable forecast quality of the project costs compared to the literature.

Another comparison that we make is with data from the ISBSG, or International Software Benchmarking Standards Group. The ISBSG collects benchmarking data from organizations in many countries. As of 2011, the data contains among others sizing information, cost and duration information for more than 5500 projects [43]. For 117 projects the data repository also contains both forecasted and actual costs.

In a book [43] an overview is given of the accuracy of these forecasts. In Table 5.8 these benchmark figures are compared with the forecasting accuracy of our case study. For the comparison, we only use the initial cost forecasts made of our case study. Since the ISBSG data contains different project types, we will compare the results with the combined projects in our case study. Moreover, the ISBSG data concerns completed projects, therefore, we again removed the canceled projects in the case study.

Table 5.8: Summary of benchmark values of the ISBSG repository [43] and our case study.

	ISBSG	Organization Z
number of projects	117	87
forecasts with EQF > 10	44%	46%
forecasts with EQF > 5	64%	57%

The table shows that our case study resembles the project cost forecasting accuracies reported by the ISBSG. The book [43] notes that the ISBSG repository is believed to reflect the best 25% of the industry. This implies that the project cost forecasts made in our case study are of high quality.

However, the possible comparisons between the figures reported by the ISBSG and our case study are limited. As we argue in Chapter 3, biases can significantly influence the accuracy of forecasts and can differ from organization to organization. The database of the ISBSG contains projects from many organizations, of which it is unknown what their organizational bias is. Without accounting for these biases, summarizing forecasting accuracies of different organizations does not result in meaningful benchmarks. Therefore, we restricted comparisons to the one we discussed above.

5.5.5 Sensitivity analysis

In the above analyses, we investigated the forecast quality of the NPV, benefits, asset usage costs and project costs for organization Z. We found that the forecast quality of the project costs for Cost Reduction assets is relatively high. However, that quality is not reflected in the equivalent forecast quality of the NPV.

The reason for this is, that assets do not consist of an equal amount of benefits, asset usage costs and project costs. This is illustrated in Table 5.9, which depicts the total and median values of these components of the NPV. The table shows that the benefits for both asset classifications are more than twice as large as the asset usage costs and project costs.

Table 5.9: Overview of the total and median reference values in millions of Euro for the components of the NPV.

reference	benefits	asset usage costs	project costs
total CR	380	129	72
median CR	3.06	0.05	0.52
total NPD	4104	2314	169
median NPD	6.84	2.04	0.58

Due to the differences in size, the impact on the quality of the NPV forecasts is not the same for each of the components. For instance, making highly accurate project cost forecasts hardly impacts the quality of the NPV, as it only forms a small part of the entire value. In fact, it is even possible that the overall quality of the NPV degrades when the accuracy of the project costs is increased.

To illustrate this, we perform a sensitivity analysis. We change the accuracy of the forecasts of one component and compute the effect on the forecast quality of the NPV in terms of EQF_s. More precisely, we increase the accuracy of the benefit, asset usage cost and project cost forecasts in such a way that their EQF_r or EQF quality is increased by 10%. For instance, if the forecasted benefits of an asset have an EQF_r quality of 4, we adjust the forecast in such a way that its EQF_r quality becomes 4.4.

Let us explain precisely how to derive at these adjusted forecasts. Recall, that the EQF_r is defined in Formula 5.2 in Section 5.3.2 as the area under the reference value divided by area difference between forecast and reference point. Or mathematically, $\mathrm{EQF}_r = \frac{\int_{t_s}^{t_r} 1\ dt}{\int_{t_s}^{t_r} |1-e(t)/r|\ dt}$. To increase the readability of the following calculations, we first rewrite this formula. Using $x = \frac{t-t_s}{t_r-t_s}$ and $f = e(t)$, we find $\mathrm{EQF}_r = \frac{\int_0^1 r\ dx}{\int_0^1 |r-f|\ dx} = \frac{|r|}{|r-f|}$.

We wish to increase the EQF_r by 10%, or similarly the original EQF_r is multiplied by 1.1. Since the reference point does not change, this means we need to decrease the area between the forecast and the reference point. We are able to decrease the area by adjusting the forecast f. We denote the adjusted forecast by f'. The EQF_r based on the adjusted forecast should be equal to 1.1 times the original EQF_r value. To derive at the adjusted forecast f', we solve the following equation.

$$
\begin{aligned}
1.1 \cdot \frac{|r|}{|r-f|} &= \frac{|r|}{|r-f'|} \\
1.1 \cdot |r| \cdot |r-f'| &= |r-f| \cdot |r| \\
|1.1 \cdot r - 1.1 \cdot f'| &= |r-f| \\
|-1.1 \cdot f'| &= |-1.1 \cdot r + r - f| \\
|f'| &= |r - \frac{1}{1.1} r + \frac{1}{1.1} f| \\
f' &= \frac{0.1}{1.1} \cdot r + \frac{1}{1.1} \cdot f \\
f' &= f + \frac{0.1}{1.1}(r-f)
\end{aligned}
$$

In this way, we are able to compute the adjusted forecasts for benefits, asset usage costs and project costs. We note that these computations are the same in case the reference point is the actual.

With these adjusted forecasts, we subsequently recompute the forecasted NPV. For instance, suppose the original NPV was computed using $NPV = f_{\mathrm{B}} - f_{\mathrm{C}} - f_{\mathrm{I}}$ and we adjusted the forecasts of the benefits f'_{B}. Then, the adjusted NPV becomes $NPV' = f'_{\mathrm{B}} - f_{\mathrm{C}} - f_{\mathrm{I}}$. By replacing the original forecast of either the benefits, asset usage costs or project costs by their adjusted forecast, we find the recomputed NPV′ forecast. For the NPV′, we again compute the corresponding EQF'_s value.

For the adjusted NPV′ forecasts, we observe the percentage difference in EQF_s between the original and adjusted situation. We do this by computing $100\% \cdot (\mathrm{EQF}'_s - \mathrm{EQF}_s)/\mathrm{EQF}_s$ with EQF_s the original value of the NPV forecasts and EQF'_s the adjusted value.

In Figure 5.15, we depict the percentage increase of the EQF_s of the NPV forecasts when we improve the forecasts of benefits, asset usage costs and project costs. In Table 5.10, we summarize these results.

Table 5.10: Summary of the improvements of the forecast quality of the NPV when each of its components increases their forecast quality by 10%.

component	% of assets with improved EQF_s quality	median increase EQF_s
benefits CR	77%	9.8%
asset usage costs CR	26%	0.0%
project costs CR	63%	0.19%
benefits NPD	87%	15%
asset usage costs NPD	19%	-4.5%
project costs NPD	10%	-0.38%

For the Cost Reduction assets, we find that improving the benefits, results in a median improvement of 9.8% of the forecast quality of the NPV. Indeed, in 77%

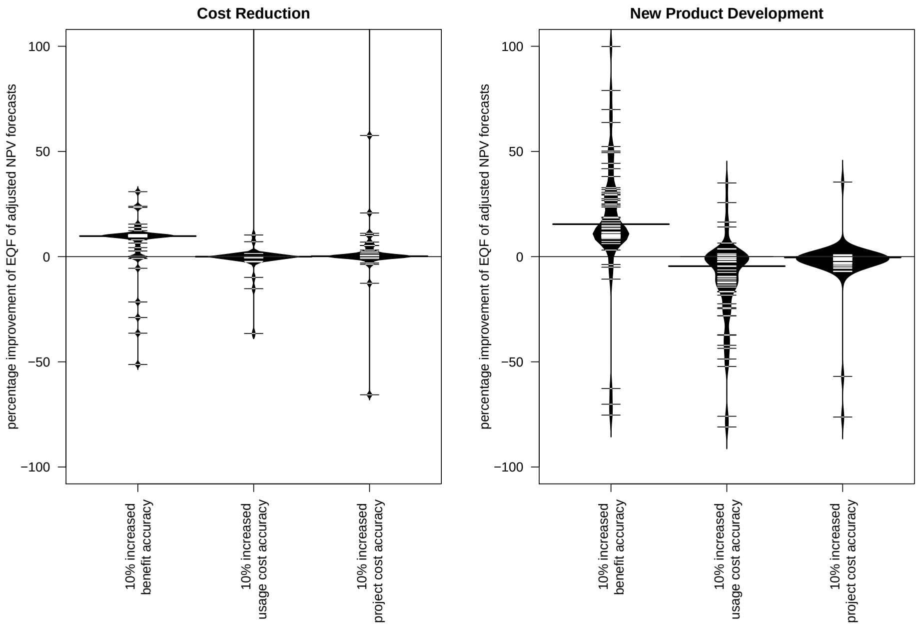


Figure 5.15: The percentage increase of the EQF of the NPV forecasts when we improve either the benefit, usage cost or project cost forecast quality by 10%.

of the Cost Reduction assets, an improvement of the benefit forecasts results in an improved NPV forecast quality.

We note that it is possible that the forecast quality of the NPV decreases while improving other components. For example, consider an asset with forecasted benefits of 100 Euro and a forecasted asset costs of 95 Euro, resulting in an NPV of 5 Euro. Suppose the actual benefits turned out to be 25 Euro and the actual asset costs 19 Euro leading to an NPV of 6 Euro. That is, both the benefits and asset costs are overestimated. If we improve the benefits by 10% this means that the initial forecasted benefits of 100 Euro becomes $100 - (1 - 1/1.1) * (100 - 25) \approx 93$ Euro. Thus, the adjusted NPV forecast is $93 - 95 = -2$ Euro instead of the previous 5 Euro. This adjusted NPV forecast is less accurate than the original NPV forecast, even though the forecast quality of the benefits increased.

Interestingly, when we improve the forecast quality of asset usage costs or project costs for Cost Reduction assets, the NPV quality remains the same with 0.0% and 0.19%. The reason that the improved accuracy of the asset usage costs and project costs is hardly noticeable for the NPV forecasts is the relative size of these elements. The median of the reference benefits of the Cost Reduction assets are five times larger than the median of the combined asset costs. Therefore, a 10% improvement of only a small part does not significantly affect the overall quality. Moreover, there are numerous assets with zero forecast and actual. In these cases, the forecasts were perfect and cannot be further improved.

The plots of the New Product Development assets show that the forecast quality of the NPV improves most by an increase of the accuracy of the benefit forecasts. For 87% of the assets, improved benefits results in an increased accuracy of the NPV. Moreover, improving the forecasting accuracy of the benefits by 10%, yields a median increase of 15% of the EQF_s for the NPV.

On the other hand, an increased accuracy of the asset usage cost and project cost forecasts leads to a *decrease* in the forecasting accuracy of the NPV. For 81% of the assets the NPV forecast accuracy decreases when the assets usage cost forecasts are improved, and 90% for the project costs. For the project costs, this decrease is hardly noticeable. However, for the asset usage costs, the forecast quality of the NPV decreases a median 4.5%.

The reason for this decrease in forecast quality of the NPV, are the biases we found for the benefits and asset usage costs of the New Product Development assets. Both these forecasts have an optimistic bias. Therefore, the overestimation of the benefits is balanced by the overestimation of the asset usage costs. By increasing the accuracy of the latter, we diminish the reduction of the overestimation of the former. Since the benefits are in general larger than the asset usage costs, it is more important for the NPV forecast quality to remain overestimating the asset usage costs to compensate for the overestimated benefits.

These results support the idea [36, 44, 92, 98] that CIO's must manage IT assets to maximize business value, instead of controlling their cost. The analysis shows that focus on control of the project costs or asset usage costs may not be beneficial to the forecast quality of the NPV. In this case study, stressing the forecasting accuracy of the costs could even decrease the forecast quality of the NPV. As the benefits often form the largest part of project proposals, IT executives should focus primarily on the accurate forecasting of the benefits and ultimately the NPV.

5.5.6 Case summary

In this section, we assessed the forecast quality for the telecommunication organization Z. We found that Cost Reduction assets have no biases for the benefit, asset usage cost and project cost forecasts. However, the New Product Development forecasts showed optimistic biases for the benefit and asset usage cost forecasts. Moreover, the New Product Development forecasts were of lower quality in terms of EQF than the Cost Reduction forecasts. Compared to benchmarks from the literature, the forecast quality of the project costs is reasonable.

Finally, we performed a sensitivity analysis to determine the impact on the forecast quality of the NPV of the different components. The analyses showed that an improvement of the benefit forecasts results in the highest improvement of the NPV forecast quality. Interestingly, if the quality of the asset usage cost forecasts improves, it results in a decreased forecasting accuracy of the NPV. These results support the idea that CIO's must manage IT assets to maximize business value, instead of controlling their cost.

5.6 Enhancing decision information

In the previous section, we illustrated how the management is able to investigate the forecast quality of the IT business value. With the analyses, executives are able to find biases, and possible change the estimation process to increase the forecasting accuracy. However, the information of the analyses is also valuable to assist the management in their decision making.

In Chapter 3 approaches were discussed to use the information of the forecast quality and their bias to enhance decision making. With these approaches, IT executives gain information on the accuracy of the forecasts and are able to make risk assessments on project proposals.

This is achieved, for instance, by using a confidence interval or an empirical distribution of f/a ratios. For instance, suppose a project proposal predicts the project costs to be 2 million Euro. Based on the forecast quality of historically forecasted project costs, an approximate interval is determined that shows that the f/a ratios in 80% of the cases have a range of $[0.5, 1.25]$.

The IT executive is able to use this information to derive a likely range of possible project costs for new project proposal. This is done in the following way. The likely range is found by dividing the initial forecast by the bounds of the interval. This leads to $2/1.25 = 1.6$ and $2/0.5 = 4$. Thus, the likely range of possible project costs of the new project proposal is $[1.6, 4]$ million Euro. This additional information allows the executive to determine whether the range of possible outcomes is acceptable to fund the project. We note that this method will also work for the ratios that we described in this chapter.

The approaches to approximate forecast intervals are only necessary when the executives are given point forecasts. As discussed in Chapter 2, it is more interesting for IT executives to immediately receive an interval forecast. Estimators already consider multiple scenarios that may occur, and are therefore able to give a more accurate project-dependent interval. This should allow for better information than these approximated intervals.

In this section, we illustrate how to use forecasts to further enhance decision making. We will do so by describing two Monte Carlo simulation examples. In these examples, we will use the data of our real-world case study of organization Z. The examples provide information for questions, such as: if a certain number of project proposals would be executed, what is the likelihood that the available budget is sufficient to finance them? Or, given a set of proposals, what is a likely range of business value that will be generated?

An important assumption of any simulation, and of any method that uses historical data to make predictions of the future, is that the future will behave similarly as in the past. In the examples we will describe, we will assume that the projects and assets are executed under similar circumstances as before. If this assumption does not hold in a certain situation, the outcomes of a simulation should be judged carefully.

We note that even though we assume the projects will behave similar to the past, it is always possible a result occurs that was not simulated. Namely, it is always possible that the assumption does not hold as initially anticipated. Still, a simulation will provide for useful information.

Moreover, we note that the examples are not meant to discuss in detail how to construct such simulation models. They are only meant to illustrate how it is possible to further utilize the information of the forecast quality assessments to enhance decision information.

5.6.1 Rationing capital budget

In the first Monte Carlo simulation example, we discuss how to utilize the quantified forecast information of forecasted project costs to assist with rationing the capital budget. Suppose that in organization Z, 30 of the 35 Cost Reduction projects and 58 of the 62 New Product Development projects have been completed and their actual project costs are known. For the remaining 5 newest Cost Reduction projects and 7 newest New Product Development projects, the actual project costs are unknown and their project proposals have been submitted. These project proposals have a predicted project cost of 22.0 million Euro. However, the organization has determined that the maximum available budget for new project proposals is less, for instance, 20 million Euro.

The required predicted investment is larger than the budget the organization is willing to spend. Therefore, the IT executives are concerned with the question: how many and which project proposals can be executed, in such a way that the available budget is sufficient to finance them?

One simple solution to the question is that the IT executive selects the most promising project proposals that more or less sum up to the available budget. Yet, with the available information of the completed projects, the executive is able to do more, as we will demonstrate using a simulation model.

We note that the complete answer to the question is not easily given. Namely, it depends on many factors, for instance, the strategic alignment of the proposals to the organization or the business values the proposals predict to yield.

In this example, we simply wish to illustrate the usefulness of information on the forecasting accuracy. Therefore, we simplify the problem by only considering the project costs and their chance of under- and overruns based on historical data. We will not consider other relevant factors necessary to fully answer the question.

Below, we give a basic description of the simulation model. Then, we discuss the results of the simulation and the information the management can derive from such models.

Creating the simulation Using the historical information of completed projects, we are able to construct a Monte Carlo simulation model. A Monte Carlo simulation model virtually executes the 12 new project proposals under similar circumstances as the completed projects.

The procedure boils down to the following. We already have thrown with a dice 30 times for the Cost Reduction category and 58 times for the New Product Development projects. These are the projects that have already been completed.

The outcomes of these throws determine the probability distribution of the dice. With the probability distribution of the dice, we are able to fictively throw with the same dice another 5 and 7 times, once for each new project proposal.

By repeatedly throwing the dice, each time we will have a different outcome of the throws. All these different outcomes will provide information about the possible variation and the likelihood of particular outcomes. We note that since we throw with two different dice with different distributions, the combined distribution of the outcomes is not trivial to compute.

The main assumption of simulation models is that the new projects are performed similarly to those that are already executed. Without this assumption, it is impossible to make any extrapolations or predictions of the future. In this particular case, the assumption is reasonable, since no major changes in the estimating process took place.

Moreover, we assume that the f/a ratios are not correlated with the projects. That is, there is no relation between the f/a ratios and any aspect of the projects. This assumption allows us to create a simple simulation by assuming the f/a ratios are randomly distributed for the projects.

To create the simulation, we compute for each completed project its f/a ratio. For the Cost Reduction proposals, we randomly draw 5 f/a ratios out of the completed Cost Reduction projects. This is done similarly for the New Product Developments. Using these f/a ratios, we are able to simulate the actual project costs for project proposal i using a randomly chosen completed project j, by calculating $u_i = f_i/(f_j/a_j)$. In this formula, f_i is the original forecasted project costs, f_j/a_j the f/a ratio of a completed project and u_i the simulated actual project costs. Summing these simulated outcomes computes to an aggregated simulated project costs for the combined new project proposals. This procedure is iterated 1000 times, resulting in a distribution of possible outcomes.

Simulation results Figure 5.16 illustrates the result of the above described simulation. The figure shows a cumulative distribution function of the simulated outcomes for the new project proposals. The horizontal axis depicts the total simulated project costs for the new project proposals. For example, $x = 30$ indicates that the simulation predicts 30 million Euro is required to execute the project proposals.

The vertical axis depicts the cumulative chance a particular outcome occurred in the simulations. For instance, $x = 20$ has a value of $y = 0.67$. This means that in the simulations the project costs in 67% of the cases were lower or equal to 20 million Euro.

The vertical solid line represents the sum of the forecasted project costs of the new project proposals. Recall that the forecasted project costs for the combined project proposals was 22.0 million Euro. The vertical dashed line is the actual required investment that was used when the projects were actually executed. This turned out to be 30.4 million Euro.

Using this figure, it becomes possible for an IT executive to make a risk assessment for the decision of which projects to invest in. The simulation shows that

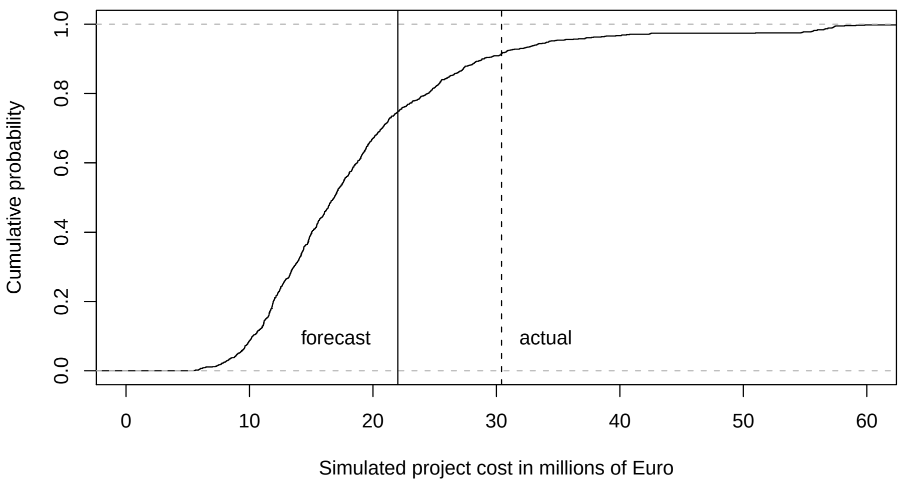


Figure 5.16: Simulation results in the form of a cumulative distribution function of simulated project costs of 5 Cost Reduction project proposals and 7 New Product Development project proposals.

if all project proposals would be executed, there is a 67% chance the projects will cost less than the available budget of 20 million Euro. However, there is also a 33% chance more money is required.

IT executives are able to use the quantified information of the figure to enhance their decision making in many ways. For instance, depending on the risk averseness or appetite of the organization, the IT executive can decide to take the risk and commence with all projects. Or, for example, the executive can determine that a selection needs to be made among the proposals. For each selection, it is possible to create a new simulation to reassess the risk involved when executed that selection.

5.6.2 Asset business values

In the second Monte Carlo simulation example, we illustrate how to enhance decision information when considering project proposals for their asset business value. Suppose, we have the same project proposals as in the previous example. That is, there are 5 Cost Reduction proposals and 7 New Product Development proposals. The predicted asset business value in terms of NPV for the project proposals is 201 million Euro. An IT executive may wonder what the chance is that the predicted value will actually be generated. It is possible to answer this question using another simulation.

Creating the simulation As in the previous example, we use the historical forecast quality of completed projects to make a prediction of the asset business value of the new project proposals. For the NPV, we need to use the $(f - r)/s$ ratio instead of the f/a ratio. In this case, we use the $(f - r)/s$ ratio with $s = f_{\mathrm{A}}$, the forecasted asset cost.

For each Cost Reduction proposal, we randomly draw a $(f - r)/s$ ratio from the completed Cost Reduction projects. However, the $(f - r)/s$ ratios show a large variation and there are only a limited number of data points. This may cause the extremes of the ratios to have a higher probability of occurring than they have in reality. If we would randomly draw ratios including these extreme ratios, we will find our simulation to produce many unlikely outcomes. Therefore, to avoid this, we remove the upper and lower ten-percent quantiles from the completed projects. This leaves 80% of the completed projects.

With the randomly drawn ratios, we simulate the value generated by each project proposal. This is done for project proposal i using a randomly chosen completed project j, by calculating $u_i = f_i - s_i \cdot (f_j - r_j)/s_j$. In this formula, f_i is the originally forecasted NPV value, s_i the original forecasted asset costs, $(f_j - r_j)/s_j$ the $(f - r)/s$ ratio of a completed project and u_i the simulated actual NPV. Summing these simulated outcomes computes to an aggregated simulated NPV for the combined new project proposals. This procedure is iterated 1000 times.

Simulation of organization Z The results of the simulations are shown in Figure 5.17. The figure shows a cumulative distribution function of the outcomes of the simulation. The horizontal axis depicts the simulated total asset business value of the project proposals. The vertical axis shows the cumulative chance of the outcomes of the simulation. For instance, $x = 0$ corresponds with the value $y = 0.18$, which means that in 18% of the cases the simulation had an adjusted business value smaller or equal to 0.

The figure shows that even with an originally forecasted NPV of 201 million Euro, there is a 18% chance that this selection of proposals will be loss generating. Of course, it is unlikely to actually occur, as the management will undertake actions to prevent the projects to become loss-making. For instance, losses can be reduced by cancelling a project, when it becomes clear the project will not yield the desired results.

The figure also shows that there is a 67% chance that the proposals will yield less than the predicted 201 million Euro. In this case, the actual business value generated by the proposals turned out to be 83 million Euro.

5.6.3 Summary

The two simulation examples illustrate that it is possible to enhance decision information using the quantified forecast information. In the basic examples, we showed how, for instance, the information of the project costs and asset business value can be used. It is also possible to incorporate asset usage cost and benefit

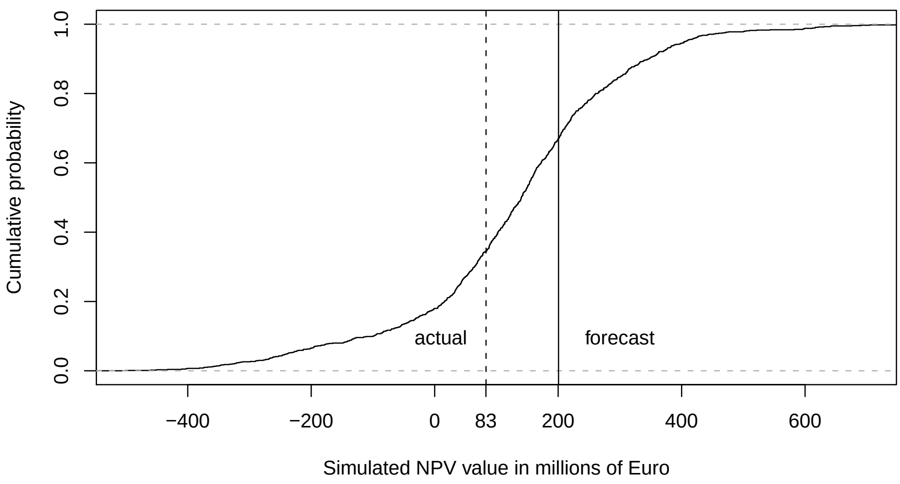


Figure 5.17: Simulation results in the form of a cumulative distribution function of simulated NPVs of 5 Cost Reduction project proposals and 7 New Product Development project proposals.

information in the simulations. These examples show that it is possible to assist IT executives in their decision making.

5.7 Discussion

In the previous sections, we discussed how to quantify the forecast quality of IT business value and use it to enhance decision information. We covered many aspects in detail. The generalized method is not a new statistical way to assess forecast quality. However, it summarizes, visualizes and quantifies the forecast quality in such a way that it is accessible to executives. In this section, we finalize with a number of general remarks about the subject.

In this chapter we addressed quantifying the forecast quality from the perspective of accuracy. For instance, we argued that increasing the accuracy of the forecasts is beneficial to the organization. Namely, more accurate forecast information will allow for better decision making. Therefore, removing a bias to increase the forecasting accuracy is beneficial.

However, biases can have their purpose. For instance, suppose we would ask participants of a tennis tournament up front to raise their hand if they think they are going to win. It is good for the play in the tournament if more than 1 participant will raise their hand. Of course, we know that these forecasts are inaccurate and biased; only one player can actually win. The forecasts show the ambition of the players and their motivation.

Business cases can also be used to communicate ambition. These ambitions should not be killed at the expense of increased forecasting accuracy. IT executives should avoid that, by increasing the forecasting accuracy, the organization becomes risk-averse. However, it remains important for executives to quantify the forecasting accuracy and potential bias to account for the quality in their decision making.

Also, in the chapter we discussed the quality of *forecasts*, not of the estimators per se. Forecasts can be made by several estimators. For instance, one estimator predicts the business domain and another forecasts the IT domain. If a forecast turns out to be less accurate than desired, this may, for instance, be caused by differences in the business domain. In this case, the IT estimator may have made accurate forecasts, yet the overall forecast is still inaccurate. In these situations, caution is needed when assessing the quality of the forecasts. The forecasting accuracy does not necessarily reflect on a particular estimator.

Finally, in this chapter, we discussed it is important to search for heterogeneity in the data. The number of factors at our disposal in the data sets were limited. In the case study of organization Z, we found the data to be heterogeneous with respect to the asset classification. Apart from that, we found no conclusive evidence for factors that influence the accuracy of the forecasts.

However, this does not mean other factors do not exist. Other research [73] is performed to qualitatively or quantitatively determine what influences the accuracy of the forecasts. Although these factors may differ from organization to organization, we stress the importance of investigating them. These factors will assist the estimators in determining how they are able to further improve their forecasting.

5.8 Conclusions

This chapter discussed how to quantify the forecast quality of IT business value. To support decision making, IT business value is captured in numerous economic indicators that summarize the expected business value of project proposals. Although not without its limitations, theory advocates the use of the Net Present Value (NPV) method. Recall that the Net Present Value is a summation of the predicted monetary benefits and costs of a project discounted to current value.

We showed that the NPV is composed of components, such as the benefits, asset usage costs and project costs. Each of the components, needs to be predicted. Therefore, in practice many organizations use multiple economic indicators to gain insight in the value of an asset proposal. Like the NPV, any economic indicator is highly dependent on the accurate forecasting of its elements. In this chapter, we assessed the forecast quality of the NPV, and its components: the benefits, asset usage costs and project costs.

To quantify the forecast quality of the NPV, we generalized a method from

Chapter 3. This method consists of the f/a ratio, the Estimating Quality Factor (EQF) and the reference cone. Recall that the EQF is a measure of the deviation between forecast and actual. For that method, three problems arise when we use it to quantify the forecast quality of IT business value.

First, the f/a ratio has visual complications in asymptotic behavior. For instance, if the actual a is zero, the ratio becomes infinite and cannot be drawn in a normal way.

Second, the f/a ratio compares the forecast with the actual that is objectively measurable. However, when we want to use the forecast quality to enhance decision information, the actual is not necessarily the most useful reference point. For instance, the actual may become known after many years when the original forecasting method may have already changed. Or, in case of the NPV, the actual may not be computed at all. Then, we need an alternative: a reference point that is not the actual, but still a better approximation than when the forecasting started. For instance, it is possible to re-estimate the actual two or three years after the start of the asset.

Third, the f/a ratio is only applicable for non-negative valued entities. However, the NPV can take negative values as well.

To remove these problems, we extended and generalized the method. This generalized method is applicable for any entity. The tools to be used depend on the entity in question.

We assessed the differences between the existing method for positively valued indicators with our generalized method. It turns out that overall the same conclusions are drawn.

Before using the method to quantify the forecast quality of the IT business value for a real-world case study, we performed a prerequisite data analysis. We found the real-world data to be heterogeneous with respect to the asset classification. For Cost Reduction and New Product Development projects, the forecast quality of the NPVs and its components, the benefits and asset usage costs are significantly different. The investigation of other variables did not result in further heterogeneity.

With this knowledge, we performed the analysis of the forecast quality of the NPV, benefits, asset usage costs and project costs for 102 IT assets of a telecommunication organization. The real-world case study represents an NPV value of 1812 million Euro, with discounted benefits of 4714 million Euro and an investment value of 173 million Euro.

We found that Cost Reduction assets have no biases for the benefit, asset usage cost and project cost forecasts. However, the New Product Development forecasts showed optimistic biases for the benefit and asset usage cost forecasts. Moreover, the New Product Development forecasts were of lower quality in terms of EQF than the Cost Reduction forecasts. Compared to benchmarks from the literature, the forecast quality of the project costs is reasonable.

Also, we performed a sensitivity analysis to determine the impact of the different components on the forecast quality of the NPV. The analyses showed that an improvement of the benefit forecasts results in the highest improvement of the

NPV forecast quality. Interestingly, in our real-world case study, if the quality of the asset usage cost forecasts improves, it results in a decreased forecasting accuracy of the NPV. These results support the idea that CIO's must manage IT assets to maximize business value, instead of controlling their cost.

Finally, we discussed how to enhance decision information using the quantified forecast information. We described two simulation examples to show how to use the quantified information of the project costs and asset business value. These examples assists with questions, such as: if a certain number of project proposals would be executed, what is the likelihood that the available budget is sufficient to finance them? Or, given a set of proposals, what is a likely range of business values that will be generated? In these cases, the simulations will provide for additional information to assist the IT executives in their decision making.

This chapter showed how to assess the forecast quality of IT business value. Forecasts are a crucial aspect for IT executives as they determine the potential value and risk related to IT assets. Therefore, knowing the forecast quality and potential bias is essential in the decision making process.

CHAPTER 6

Samenvatting (summary in dutch)

Dit proefschrift gaat over het kwantificeren van de schattingskwaliteit van IT-projecten. Goede schattingen zijn cruciaal voor het welslagen van projecten. Elk project begint met een idee van wat uiteindelijk gemaakt moet worden. Op basis van dat idee worden schattingen gemaakt van de hoeveelheid tijd en geld die nodig is om de gevraagde functionaliteit te realiseren. Daarnaast wordt bepaald wat het project zal opleveren.

Deze schattingen bepalen mede of een project wordt uitgevoerd of niet. Indien de schattingen een positief beeld geven en het project past binnen het beleid van de organisatie, maakt het project goede kans om gefinancierd te worden. De kwaliteit van de schattingen is daarbij cruciaal. Als de kosten, doorlooptijden en opbrengsten verkeerd worden geschat, kunnen onnodige risico's worden gelopen zonder dat de bestuurder zich hiervan bewust is. Onnauwkeurige schattingen kunnen leiden tot hoog gespannen verwachtingen, die later niet haalbaar blijken.

Vaak wordt aangenomen dat schattingen op een objectieve manier tot stand komen. Echter, bij het maken van schattingen spelen bijvoorbeeld politieke motieven een belangrijke rol. Bijvoorbeeld, om een project überhaupt te mogen uitvoeren is het soms noodzakelijk een positieve schatting af te geven van de duur en de kosten van het project, alhoewel bekend is dat het project mogelijk langer zal duren en meer zal kosten. Projectvoorstellen moeten als het ware aan bestuurders worden verkocht. Dergelijke politieke motieven, maar ook andere vertekeningen kunnen zorgen voor afwijkingen in de schattingen.

6.1 Schattingskwaliteit

In dit proefschrift is onderzocht op welke wijze IT managers de kwaliteit van de schattingen kwantitatief kunnen vaststellen. Voor dit doel is gebruik gemaakt van Boehm's cone of uncertainty en DeMarco's Estimating Quality Factor (EQF).

De schattingen worden gevisualiseerd in een forecast-to-actual plot, waarin de ratio van schattingen en bijbehorende uiteindelijke waarde, worden getoond. Een dergelijke plot toont de kwaliteit van de schattingen, die vergeleken kunnen worden met een veralgemeniseerd model gebaseerd op de cone of uncertainty. Hierdoor kunnen vertekeningen inzichtelijk worden gemaakt. Met behulp van de EQF wordt de kwaliteit van de schattingen gekwantificeerd.

De methode wordt geïllustreerd aan de hand van data van vier grote organisaties. In totaal betreft de data 1824 projecten met een totale waarde van meer dan 1059 miljoen Euro, waarvoor in totaal 12287 schattingen zijn gemaakt. Deze studies laten zien dat vertekeningen de kwaliteit van schattingen significant kunnen beïnvloeden.

Tot slot worden drie methoden besproken die, op basis van gemeten schattingskwaliteit, aanvullende informatie verschaffen voor het maken van beslissingen. De methoden stellen het management in staat rekening te houden met mogelijke vertekeningen en de onzekerheid van schattingen, om de risico's van specifieke scenarios in te schatten.

6.2 Benchmarks uit de literatuur

Uit de uitkomsten van onze studies blijkt dat men zich terdege moet realiseren dat gemaakte schattingen van de kosten en duur van projecten vertekend kunnen zijn. Dat realiseert men zich echter lang niet altijd, zoals blijkt uit ons onderzoek naar benchmarks uit de literatuur inzake projectkosten en doorlooptijden van projecten. Bij de berekening van deze benchmarks wordt geen rekening gehouden met de mogelijke vertekeningen van de schattingen van de organisaties die de data hebben aangeleverd. Omdat de mogelijke vertekening per organisatie onbekend is en per organisatie verschilt, zijn de cijfers niet bruikbaar als vergelijkingsmateriaal.

In het bijzonder is gekeken naar veel geciteerde cijfers van Standish Group over project succes. Daaruit bleek dat de definities zoals gehanteerd door Standish vier problemen hebben. Ten eerste zijn de definities misleidend, omdat de definities van project succes alleen zijn gebaseerd op de schattingskwaliteit van kosten, doorlooptijd en functionaliteit. Ten tweede beschouwen de definities alleen onderschattingen en geen overschattingen, waardoor onrealistische succes cijfers ontstaan. Ten derde worden schatters door de definities gemotiveerd om overschattingen te maken. Tot slot worden cijfers van organisaties bij elkaar genomen, waarvan onduidelijk is of de schattingen van de organisaties vertekend zijn. Zonder te corrigeren voor vertekeningen is het onmogelijk vast te stellen wat het resulterende betekent.

De tekortkomingen van de definities worden geïllustreerd door de definities toe te passen op de data van de vier organisaties uit de praktijk. Daaruit blijkt dat de definities geen realistisch beeld geven van de werkelijke schattingskwaliteit.

6.3 Schattingskwaliteit van opbrengsten

Tot slot, gaat dit proefschrift in de op de schattingskwaliteit van de toegevoegde waarde van IT projecten, bijvoorbeeld gemeten door middel van de Netto Contante Waarde (NCW). De eerdere methode voor het kwantificeren van de schattingskwaliteit, gebaseerd op Boehm's cone of uncertainty en DeMarco's EQF, wordt veralgemeniseerd. Hierdoor is de methode in staat om te gaan met asymptotische waarden, negatieve waarden en andere referentiepunten dan de uiteindelijke waarde waaraan de schattingskwaliteit wordt afgemeten. De veralgemeniseerde methode maakt het mogelijk inzicht te verkrijgen in zowel het geheel als de schattingskwaliteit van de afzonderlijke componenten van een schatting. De veralgemeniseerde methode is getest met behulp van de data van de eerdere studies.

Met de veralgemeniseerde methode is de kwaliteit van de schatting van de NCW, opbrengsten en kosten van 102 assets met een totale waarde van 1812 miljoen Euro van een internationaal opererende organisatie geanalyseerd. De studie laat zien dat inzicht in zowel de schattingskwaliteit van de afzonderlijke componenten als het geheel nodig is om tot een goed oordeel te komen van de schattingskwaliteit. Voor deze organisatie blijkt te gelden dat de kwaliteit van de schatting van de NCW lager is dan de kwaliteit van de schatting van de opbrengsten, die op zijn beurt weer lager is dan de kwaliteit van de schatting van de kosten.

Daarnaast is een gevoeligheidsanalyse uitgevoerd, waarbij onderzocht is wat de invloed is van verhoging van de kwaliteit van de schatting van de kosten en de opbrengsten op de schattingskwaliteit van de NCW. Opvallend is dat voor deze studie geldt dat de schattingskwaliteit van de uiteindelijke NCW verslechtert als de schattingskwaliteit van alleen de kosten wordt verbeterd. Dit laat zich als volgt verklaren. Doordat zowel opbrengsten als kosten systematisch te hoog worden ingeschat, neemt het verschil tussen opbrengsten en kosten (de NCW) toe indien de overschatting van de kosten wel wordt aangepakt, maar de overschatting van de opbrengsten niet. In plaats van een verbetering van de kwaliteit van schatting van de NCW, wat men intuïtief zou verwachten, verslechtert ze in dit geval juist. De misschatting van de NCW wordt groter. Dit onderzoeksresultaat onderstreept het belang van goede en politiekvrije schattingen van zowel kosten als opbrengsten.

Tot slot demonstreren we door middel van twee voorbeeldsimulaties hoe de gekwantificeerde schattingskwaliteit kan worden gebruikt voor het verbeteren van management informatie.

Bibliography

[1] Fred Arditti. The weighted average cost of capital: some questions on its definition, interpretation, and use. *The journal of finance*, 28(4):1001–1007, 1973. Cited on page 129.

[2] Hal R. Arkes. Overconfidence in judgmental forecasting. In J.S. Armstrong, editor, *Principles of forecasting: A handbook for researchers and practitioners*, pages 495–515. Springer, 2001. Cited on page 10.

[3] Phillip G. Armour. The inaccurate conception. *Communications of the ACM*, 51(3):13–16, 2008. Cited on pages 21 and 27.

[4] J.S. Armstrong. The forecasting dictionary. In J.S. Armstrong, editor, *Principles of forecasting: A handbook for researchers and practitioners*, pages 761–819. Springer, 2001. Cited on page 16.

[5] Glen Arnold and Panos Hatzopoulos. The theory-practice gap in capital budgeting: evidence from the United Kingdom. *Journal of business finance & Accounting*, 27(5):603–626, 2000. Cited on pages 125, 127, and 129.

[6] N.R. Augustine. Augustine's laws and major system development programs. *Defense systems management review*, 2:50–76, 1979. Cited on pages 11 and 93.

[7] François Bergeron and Jean-Yves St-Arnaud. Estimation of information systems development efforts. *Information & Management*, 22:239–254, 1992. Cited on pages 9, 11, 93, 94, and 130.

[8] Barry Boehm. *Software engineering economics*. Prentice hall PTR, 1981. Cited on pages 4, 8, 9, 20, 24, 25, 26, 28, 36, 51, 60, 76, 79, 94, 95, 104, 130, 132, 133, and 140.

[9] Barry Boehm. Making a difference in the software century. *Computer*, 41(3):32–38, March 2008. Cited on page 5.

[10] Barry Boehm, Bradford Clark, Ellis Horowitz, and Chris Westland. Cost models for future software life cycle processes: COCOMO 2.0. *Annals of software engineering*, 1, 1995. Cited on pages 5, 25, 26, and 27.

[11] R.A. Brealey and S.C. Myers. *Principles of corporate finance*. McGraw-Hill, Inc., 1988. Cited on pages 126, 128, and 129.

[12] Lionel Briand, Khaled El Emam, Dagmar Surmann, Isabella Wieczorek, and Katrina D. Maxwell. An assessment and comparison of common software cost estimation modeling techniques. *Proceedings of the 21st international conference on Software engineering*, pages 313 – 322, 1999. Cited on page 8.

[13] Robert Bruner, Kenneth Eades, Robert Harris, and Robert Higgins. Best practices in estimating the cost of capital: survey and synthesis. *Financial practice and education*, pages 13–28, 1998. Cited on page 129.

[14] Manfred Bundschuh and Carol Dekkers. *The IT Measurement compendium*. Springer, 2010. Cited on pages 8, 132, and 134.

[15] Murray Cantor. Estimation variance and governance, 2006. `www-128.ibm.com/developerworks/rational/library/mar06/cantor/`. Cited on page 27.

[16] Chris Chatfield. Prediction intervals for time-series forecasting. In J.S. Armstrong, editor, *Principles of forecasting: A handbook for researchers and practitioners*, pages 475–494. Springer, 2001. Cited on page 76.

[17] Murali Chemuturi. *Software estimation best practices, tools & techniques*. J. Ross Publishing, 2009. Cited on page 8.

[18] Mike Cohn. *Agile estimating and planning*. Prentice Hall, 2005. Cited on pages 9, 32, and 45.

[19] S.D. Conte, H.E. Dunsmore, and V.Y. Shen. *Software engineering metrics and models*. The Benjamin/Cummings Publishing Company, 1986. Cited on pages 8, 9, 40, and 42.

[20] Tom DeMarco. *Controlling software projects*. Prentice Hall PTR, 1982. Cited on pages 4, 5, 9, 10, 11, 20, 23, 29, 31, 37, 41, 43, 44, 56, 58, 60, 90, 91, 95, 104, 144, and 174.

[21] Tom DeMarco and Timothy Lister. *Waltzing with bears*. Dorset House Publishing, 2003. Cited on page 62.

[22] Sanjeev Dewan, Charles Shi, and Vijay Gurbaxani. Investigating the risk-return relationship of information technology investment: Firm-level empirical analysis. *Management science*, 53(12):1829–1842, 2007. Cited on page 129.

[23] J.B. Dreger. *Function Point Analysis*. Prentice Hall, 1989. Cited on pages 66, 67, and 133.

[24] Nicholas Epley and Thomas Gilovich. The anchoring-and-adjustment heuristic: why the adjustments are insufficient. *Information and Software Technology*, 17(4):311–318, 2006. Cited on page 10.

[25] J.L. Eveleens, P. Kampstra, and C. Verhoef. Quantifying changes in IT metrics after outsourcing transitions. Available via `www.cs.vu.nl/~x/sps/sps.pdf`, 2009. Cited on pages 12, 29, 30, and 80.

[26] J.L. Eveleens and C. Verhoef. Quantifying IT forecast quality. *Science of computer programming*, 74:934 – 988, 2009. Cited on page 111.

[27] Dick Fairley. Making accurate estimates. *IEEE Software*, 19(6):61–63, 2002. Cited on page 10.

[28] W. Feller. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The annals of mathematical statistics*, 19(2):177–189, 1948. Cited on page 85.

[29] Marek Fisz. *Probability Theory and Mathematical Statistics*. John Wiley & Sons, 1963. Cited on pages 8, 40, and 85.

[30] Bent Flyvbjerg, Nils Bruzelius, and Werner Rothengatter. *Megaprojects and risk: an anatomy of ambition*. Cambridge University Press, 2003. Cited on page 93.

[31] James M. Fremgen. Capital budgeting practices: A survey. *Management Accounting*, 54(11):19–25, 1973. Cited on pages 125, 127, and 129.

[32] D. Garmus and D. Herron. *Function Point Analysis – Measurement practices for successful software projects*. Addison-Wesley, 2001. Cited on pages 66, 67, 105, and 133.

[33] Lawrence J. Gitman and John R. Forrester. A survey of capital budgeting techniques used by major U.S. firms. *Financial management*, 6(3):66–71, 1977. Cited on pages 127 and 129.

[34] Robert Glass. IT failure rates—70% or 10–15%. *IEEE Software*, May:110–112, 2005. Cited on pages 93, 101, and 111.

[35] Robert Glass. The Standish report: does it really describe a software crisis. *Communications of the ACM*, 49(8), 2006. Cited on pages 93, 101, and 111.

[36] META Group. The Business of IT Portfolio Management: Balancing Risk, Innovation, and ROI. Technical report, META Group, Stamford, CT, USA, January 2002. Cited on pages 3, 130, and 179.

[37] Stephan Gryphon, Philippe Kruchten, Steve McConnell, and Todd Little. The Cone of Uncertainty. *IEEE Software*, 23(5):8–10, 2006. Cited on pages 13, 16, 19, 26, 27, and 29.

[38] Emiel Hakkenes. 'Ict-plan van PKN deugt echt niet'. News paper. Trouw 27-06-2008. Cited on page 3.

[39] John S. Hammond, Ralph L. Keeney, and Howard Raiffa. The hidden traps in decision making. *Harvard Business review*, 76(5):47–58, 1998. Cited on page 10.

[40] Harold Bierman Harold and Seymour Smidt. *The capital budgeting decision: Economic analysis of investment projects*. Macmillan Publishing Company, 8th edition, 1993. Cited on pages 126 and 129.

[41] Deborah Hartmann. Interview: Jim Johnson of the Standish Group, 2006. www.infoq.com/articles/Interview-Johnson-Standish-CHAOS. Cited on pages 9, 93, 94, and 102.

[42] Nigel Harvey. Improving judgment in forecasting. In J.S. Armstrong, editor, *Principles of forecasting: A handbook for researchers and practitioners*, pages 59–80. Springer, 2001. Cited on page 10.

[43] Peter R. Hill. *Practical software project estimation*. McGraw-Hill, 2011. International software benchmarking standards group. Cited on page 175.

[44] Jeremy Hope and Robin Fraser. Beyond budgeting: building a new management model for the information age. *Management accounting*, 77(1):16–21, 1999. Cited on pages 3 and 179.

[45] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22:679–688, 2007. Cited on pages 8 and 40.

[46] IKON. PKN ontwikkelt Numeri niet verder. online article, 2008. Available via url http://ikonrtv.nl/kerknieuws/nieuws.asp?oId=13174. Cited on page 3.

[47] The Standish Group International Inc. Chaos. Technical report, The Standish Group International Inc., 1994. Cited on pages 5, 11, 93, 94, 102, 103, and 106.

[48] The Standish Group International Inc. Chaos: A recipe for success. Technical report, The Standish Group International Inc., 1999. Cited on pages 5, 11, 93, and 102.

[49] The Standish Group International Inc. Extreme chaos. Technical report, The Standish Group International Inc., 2001. Cited on pages 5, 11, 93, 94, and 102.

[50] The Standish Group International Inc. Ibex healthcheck assessment. Technical report, The Standish Group International Inc., 2006. Cited on pages 5 and 102.

[51] The Standish Group International Inc. Chaos 2009, 2009. `http://www.standishgroup.com/newsroom/chaos_2009.php`. Cited on pages 5 and 102.

[52] Jonathan E. Ingersoll and Stephen A. Ross. Waiting to invest: investment and uncertainty. *The journal of business*, 65(1):1–29, 1992. Cited on page 127.

[53] Milton Jenkins, Justus Naumann, and James Wetherbe. Empirical investigation of systems development practices and results. *Information & Management*, 7:73–82, 1984. Cited on pages 11 and 93.

[54] Jim Johnson. Standish: Why were project failures up and cost overruns down in 1998, 2006. `www.infoq.com/articles/chaos-1998-failure-stats`. Cited on pages 9, 93, 94, and 110.

[55] Capers Jones. Conflict and litigation between software clients and developers. Technical note, April 2001. Version 10. Cited on page 16.

[56] Capers Jones. *Estimating software costs: Bringing realism to estimating*. McGraw-Hill, Inc., second edition, 2007. Cited on pages 8 and 132.

[57] Capers Jones. *Applied Software measurement: Global analysis of productivity and quality*. McGraw-Hill, Inc., third edition, 2008. Cited on pages 8, 16, and 134.

[58] Capers Jones. Personal communication, March 2011. Cited on pages 17 and 134.

[59] Magne Jørgensen. Experience with the accuracy of software maintenance task effort prediction models. *IEEE Transactions on software engineering*, 21:674–681, 1995. Cited on pages 8, 40, and 42.

[60] Magne Jørgensen. A review of studies on expert estimation of software development effort. *Journal of Systems and Software*, 70:37–60, 2004. Cited on page 10.

[61] Magne Jørgensen. Practical guidelines for expert-judgment-based software effort estimation. *IEEE Software*, 22(3):57–63, 2005. Cited on page 21.

[62] Magne Jørgensen and Kjetil Moløkken. How large are software cost overruns? A Review of the 1994 Chaos report. *Information and Software Technology*, 48:297–301, 2006. Cited on pages 93, 101, 103, and 111.

[63] Bill Joy and Ken Kennedy. Information technology research: investing in our future. Technical report, President's Information Technology Advisory Committee, February 1999. Cited on page 103.

[64] Daniel Kahneman, Dan Lovallo, and Olivier Sibony. Before you make that big decision... *Harvard Business review*, 89(6):50–60, 2011. Cited on page 10.

[65] P. Kampstra. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of statistical software*, 28, 2008. Cited on page 163.

[66] P. Kampstra and C. Verhoef. Benchmarking the expected loss of a federal IT portfolio. Available via `www.cs.vu.nl/~x/fed/fed.pdf`. Cited on page 12.

[67] K. Kavoussanakis and Terry Sloan. UKHEC Report on software estimation. Technical report, University of Edinburgh, 2001. Cited on page 9.

[68] Chris Kemerer. An empirical validation of software cost estimation models. *Communications of the ACM*, 30(5):416 – 429, 1987. Cited on pages 8 and 40.

[69] Peeter J. Kirs, Kurt Pflughoeft, and Galen Kroeck. A process model cognitive biasing effects in information systems development and usage. *Information & Management*, 38:153–165, 2001. Cited on page 10.

[70] Barbara Kitchenham, David Ross Jeffery, and Colin Connaughton. Misleading metrics and unsound analyses. *IEEE Software*, 24(2):73–78, 2007. Cited on page 138.

[71] Barbara Kitchenham, Stephan MacDonell, Lesley Pickard, and Martin Shepperd. What accuracy statistics really measure. *IEE Proceedings - Software*, 148:81–85, 2001. Cited on pages 40, 42, 43, and 76.

[72] G.P. Kulk and C. Verhoef. Quantifying requirements volatility effects. *Science of Computer Programming*, 72:136–175, 2008. Cited on pages 12 and 67.

[73] G.P. Kulk and C. Verhoef. Quantifying IT Estimation Risks. *Science of Computer Programming*, 74:900 – 933, 2009. Cited on pages 12, 90, 91, 174, and 186.

[74] L.M. Kwiatkowski and C. Verhoef. Recovering management information from source code. Available via `www.cs.vu.nl/~x/rmsc/rmsc.pdf`. Cited on page 11.

[75] L.M. Kwiatkowski and C. Verhoef. Reducing operational costs through MIPS management. Available via `www.cs.vu.nl/~x/mips/mips.pdf`. Cited on page 11.

[76] Linda M. Laird. The limitations of estimation. *IT professional*, 8(6):40–45, 2006. Cited on pages 9 and 27.

[77] Linda M. Laird and M. Carol Brennan. *Software measurement and estimation: A practical approach*. John Wiley & Sons, 2006. Cited on pages 8, 21, and 87.

[78] Luiz A. Laranjeira. Software size estimation of object-oriented systems. *IEEE Transactions on software engineering*, 16(5):510–522, 1990. Cited on pages 32 and 86.

[79] Joseph L.Bower. *Managing the resource allocation process*. Harvard University, 1970. Cited on pages 9, 160, and 161.

[80] Albert L. Lederer, Rajesh Mirani, Boon Siong Neo, Carol Pollard, Jayesh Prasad, and K. Ramamurthy. Information system cost estimating: A management perspective. *MIS Quarterly*, 14(2):159–176, 1990. Cited on page 10.

[81] I. Leliveld and M. Jeffery. IT portfolio management — Challenges and best practices. Technical report, Kellogg School of Management and Diamond Cluster International Inc., 2003. Cited on pages 3 and 130.

[82] Tim Lister. Becoming a better estimator - An Introduction to using the EQF Metric, 2002. Available via `www.stickyminds.com`. Cited on pages 9, 11, 43, 90, 91, 144, and 174.

[83] Todd Little. Context-Adaptive Agility: Managing complexity and uncertainty. *IEEE Software*, 22(3):28–35, 2005. Cited on pages 9, 34, 43, 46, 55, 78, 91, and 174.

[84] Todd Little. Schedule estimation and uncertainty surrounding the cone of uncertainty. *IEEE Software*, 23(3):48–54, 2006. Cited on pages 5, 9, 13, 16, 19, 21, 26, 36, 43, 59, 62, 87, and 174.

[85] James H. Lorie and Leonard J. Savage. Three problems in rationing capital. *The journal of business*, 28(4):229–239, 1955. Cited on pages 3 and 127.

[86] Karen Lum, Michael Bramble, Jairus Hihn, John Hackney, Mori Khorrami, and Erik Monson. Handbook for software cost estimation. Technical report, Jet Propulsion Laboratory, 2003. Cited on page 9.

[87] Maplesoft. `www.maplesoft.com/Products/Maple/`. Cited on pages 48, 52, and 156.

[88] Steve McConnell. *Rapid development: Taming wild software schedules*. Microsoft Press, 1996. Cited on pages 9, 11, 25, 27, 74, 76, and 79.

[89] Steve McConnell. *Software estimation: Demystifying the black art*. Microsoft Press, 2006. Cited on pages 8, 10, 21, 27, and 40.

[90] M. Shahid Mujtaba and Robert Ritter. Enterprise modeling system: Inventory exposure and delivery performance. Technical report, Hewlett-Packard company, 1994. Cited on page 20.

[91] Ingunn Myrtveit and Erik Stensrud. A controlled experiment to assess the benefits of estimating with analogy and regression models. *IEEE Transactions on software engineering*, 25:510–525, 1999. Cited on pages 40 and 42.

[92] Andy Neely, Mike Bourne, and Chris Adams. Better budgeting or beyond budgeting? *Measuring Business Excellence*, 7(3):22–28, 2003. Cited on pages 3 and 179.

[93] Carlo Pescio. Realistic and useful: toward better estimates, 2007. Available via www.eptacom.net/betterestimate/RealisticAndUseful.pdf. Cited on pages 40 and 86.

[94] R.J. Peters and C. Verhoef. Quantifying the yield of risk-bearing IT-portfolios. *Science of computer programming*, 71:17–56, 2008. Cited on page 132.

[95] J. Petty, D. Scott, and M. Bird. The capital expenditure decision-making process of large corporations. *The engineering economist*, 20(3):159–172, 1975. Cited on pages 127 and 129.

[96] Dien Phan, Douglas Vogel, and Jay Nunamaker. The search for perfect project management. *Computerworld*, 26:95–100, 1988. Cited on pages 9, 11, 93, 94, and 130.

[97] Richard H. Pike. *Capital budgeting in the 1980s: A major survey of the investment practices in large companies*. The institute of cost and management accountants, 1982. Cited on pages 125 and 127.

[98] T. Pisello and P.A. Strassmann. *IT Value Chain Management – Maximizing the ROI from IT Investments*, volume D10 of *Digital Publications from The Information Economics Press*. The Information Economics Press, New Canaan, Connecticut, USA, 2003. Cited on pages 3, 125, and 179.

[99] L.H. Putman and W. Myers. *Measures for Excellence – Reliable Software on Time, Within Budget*. Yourdon Press Computing Series, 1992. Cited on pages 28 and 36.

[100] Lawrence H. Putnam and Ann Fitzsimmons. Estimating software costs. *Datamation*, september, october and november, 1979. Cited on page 86.

[101] L.H. Putnam. A macro-estimation methodology for software development. In *Proceedings IEEE COMPCON 76 Fall*, pages 138–143. IEEE Computer Society Press, 1976. Cited on page 36.

[102] L.H. Putnam. A general empirical solution to the macro software sizing and estimating problem. *IEEE Transactions of software engineering*, 4(4):345 – 381, 1978. Cited on pages 8 and 132.

[103] R-project. www.r-project.org/. Cited on pages 31 and 85.

[104] Stephen Ross, Randolph Westerfield, and Bradford Jordan. *Fundamentals of corporate finance*. McGraw-Hill, Inc., 2000. Cited on page 129.

[105] Stephen A. Ross. Uses, abuses, and alternatives to the net-present-value rule. *Financial Management*, 24(3):96–102, 1995. Cited on page 127.

[106] Patricia A. Ryan and Glenn P. Ryan. Capital budgeting practices of the fortune 1000: How have things changed? *Journal of business and management*, 8(4):355–364, 2002. Cited on pages 127 and 128.

[107] Rik Sanders. PKN verrekent zich fors in ledensysteem. online article, 2008. Available via `http://www.computable.nl/artikel/ict_topics/beheer/2286404/1277800/pkn-verrekent-zich-fors-in-ledensysteem.html`. Cited on page 3.

[108] Chris Sauer, Andrew Gemino, and Blaize Horner Reich. The impact of size and volatility on IT project performance. *Communications of the ACM*, 50(11):79–84, 2007. Cited on page 93.

[109] Lawrence D. Schall, Gary L. Sundem, and Jr. William R. Geijsbeek. Survey and analysis of capital budgeting methods. *The journal of finance*, 33(1):281–287, 1978. Cited on pages 127 and 129.

[110] C.C. Shelley. Practical experience of implementing software measurement programmes in industry. *Transactions on Information and Communications Technologies*, 4, 1993. Cited on page 9.

[111] Martin Shepperd and Chris Schofield. Estimating software project effort using analogies. *IEEE Transactions of software engineering*, 23(11):736–743, 1997. Cited on page 8.

[112] Keng Siau, Yair Wand, and Izak Benbasat. When parents need not have children — cognitive biases in information modeling. *Advanced information systems engineering*, 1080:402–420, 1996. Cited on page 10.

[113] Michael Smithson. *Confidence intervals*, volume 07-140 of *Sage University papers series on quantitative applications in social sciences*. Thousand Oaks, 2003. Cited on page 77.

[114] Ian Sommerville. *Software engineering*. Pearson educcation limited, 7 edition, 2004. Cited on page 9.

[115] Webb Stacy and Jean MacMillian. Cognitive bias in software engineering. *Communications of the ACM*, 38(6):57–63, 1995. Cited on page 10.

[116] Erik Stensrud, Tron Foss, Barbara Kitchenham, and Ingunn Myrtveit. A further empirical investigation of the relationship between MRE and project size. *Empirical software engineering*, 8:139–161, 2003. Cited on page 40.

[117] Steve Tockey. *Return on software*. Addison-Wesley, 2005. Cited on pages 20, 25, and 76.

[118] Diederik Toet. UWV kan 4 procent van WIA-project hergebruiken. online article, 2008. Available via `http://www.computable.nl/artikel/ict_topics/overheid/2728190/1277202/uwv-kan-4-procent-van-wiaproject-hergebruiken.html`. Cited on page 3.

[119] Oxford university press. *Concise oxford english dictionary*. Oxford university press, 10th edition, 1999. Cited on pages 13, 15, 16, and 21.

[120] Sytse van der Schaaf. WIA is schoolvoorbeeld mislukking. *IT executive*, (20):72–75, 2008. Available via http://www.cs.vu.nl/~x/knipselkrant/it-executive-2.pdf. Cited on page 3.

[121] Hans van Vliet. *Software engineering: principles and practice*. John Wiley & Sons, third edition, 2008. Cited on page 9.

[122] C. Verhoef. Quantitative IT portfolio management. *Science of computer programming*, 45:1–96, 2002. Cited on page 11.

[123] C. Verhoef. Quantitative aspects of outsourcing deals. *Science of computer programming*, 56:275–313, 2004. Cited on page 11.

[124] C. Verhoef. Quantifying the value of IT-investments. *Science of computer programming*, 56:281–287, 2005. Cited on pages 11 and 129.

[125] Chris Verhoef. Quantifying the Effects of IT-Governance Rules. *Science of Computer Programming*, 67(2–3):247–277, 2007. Cited on pages 11, 37, and 43.

[126] Frank Vogelezang. Scope management—how uncertainty is your certainty. *International workshop on software measurement*, 2007. Cited on page 21.

[127] Gerald M. Weinberg and Edward L. Schulman. Goals and performance in computer programming. *Human factors*, 16(1):70–77, 1974. Cited on page 10.

[128] Da Yang, Barry Boehm, Ye Yang, Qing Wang, and Mingshu Li. *Coping with the cone of uncertainty: An Empirical study of the SAIV process model*, volume 4470/2007 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2007. Cited on pages 13 and 25.

[129] Ernst & Young. Programma WIA: 'een dure les'. Technical report, Ernst & Young, 2008. Available via http://docs.minszw.nl/pdf/35/2008/35_2008_3_12274.pdf. Cited on page 3.

[130] Yourdon. The ten most important ideas in software engineering, 2006. www.yourdonreport.com/index.php/2006/10/17/the-ten-most-important-ideas-in-software-engineering/. Cited on page 27.

[131] Nicholas Zvegintzov. Frequently begged questions and how to answer them. *IEEE Software*, 20(2):93–96, 1998. Cited on pages 93 and 111.