Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 239 (2024) 209–216

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2023

Reducing cost overruns through data-driven methods used in uncertainty analyses

Julian Mæhlen[a,*], Jan Petter Bekkevold[a], Morten Welde[b], Nils O. E. Olsson[b]

[a]*Holte Consulting, Drammensveien 145b, Oslo 0277, Norway*
[c]*Department of Civil and Environmental Engineering, Norwegian University of Science and Technology (NTNU), Norway*

Abstract

This article focuses on how data from completed projects can be applied to uncertainty analysis. Cost estimates and risk assessments in public construction projects rely on expert opinions and, by little extent, historical figures. By analysing the relationship between project features and budget deviation through statistical methods, we find that total area and estimated square meter price is significantly negatively correlated to cost overruns. Smaller projects tend to have higher cost overruns than larger ones. We argue that cost risk analyses can be improved by such insight.

Keywords: Uncertainty analysis; cost risk; cost overruns; construction projects; data-driven methods.

1. Introduction

Uncertainty analysis is a systematic method for identifying, describing and calculating cost risk in projects [1]. The results provide a basis for decision-making, management regime and management of the project. Uncertainty analyses are often based on subjective assessments rather than data from completed projects. It takes an inside view, rather than

* Corresponding author.
E-mail address: jum@holteconsulting.com

10.1016/j.procs.2024.06.164

an outside view, where data from completed projects is used as a reference to estimate the uncertainty of future projects. One disadvantage of an inside view is that uncertainty analyses may be vulnerable to human bias relying on pure intuition and individual experience [2]. Previous studies show that projects systematically underestimate costs [3, 4, 5]. Thus, we see a potential for an improved method, relying less on expert opinions and more on data from completed projects.

Various causes of cost overruns have been suggested in the literature, and considerable discussion exists about which factors most drive the overruns. For example, it has been argued whether management-related reasons are prominent [6], the size of the project [7, 8] or whether cost overruns are due to strategic underestimation [9]. Different countries thus use different methodologies to uncover uncertainty and provide unbiased estimates [10].

This article focuses on what data on completed projects can say about cost uncertainty, thus avoiding some human factors. This could provide benchmark values in a reasonableness assessment and contribute to better cost control through realistic budgets. The study is limited to construction projects, as both contract models, calculation structure and investment regime are relatively standardised in the sector. Therefore, it makes it easy to normalise data to a satisfactory level for analysis purposes.

To quantify the effect of uncertainty in a project, we study the relationship between the project's features and relative discrepancies between the estimated and actual construction cost of the project, as has been done in previous studies [5, 11]. The relative deviation is defined as

$$Relative\ deviation = \frac{Actual\ construction\ cost\ after\ completion - Estimated\ construction\ cost}{Estimated\ construction\ cost} \quad (1)$$

This corresponds to the percentage difference between the budget at the time of investment decision and the actual cost after completion. In Section 5, we discuss how relative deviation may correspond to expected additions in projects, thus being applicable as a benchmark value in uncertainty analyses for future projects.

The article proceeds as follows. First, we present a description of the data and methodology used for our empirical findings, followed by the results. Thereafter, we discuss our findings and conclude on its implications. The article ends with our recommendations for how uncertainty analyses can be improved through data-driven methods.

2. Data

We have collected data on construction projects run by public construction project clients responsible for property development and construction. This includes municipalities and state-owned companies with decision-making authority in public construction projects. The investment regime is broadly similar, with a development phase, pre-project phase and implementation phase. A basic calculation is established in the pre-project phase, representing each component's most likely cost. However, the basic calculation does not capture cost increases due to uncertainty. This means that the basic calculation is often lower than the actual cost [5], which the uncertainty analysis attempts to take into account [1].

Data have been collected on 489 projects from three public construction project clients in Norway. Each project is designated as a data point for analysis purposes. Each data point has its data relating to size, contract form, building type and investment type, i.e., whether it is a new building or a renovation. In addition, we have stated the basic calculation and final cost for each project. The basic calculation allows us to calculate an estimated square meter price, which is also an interesting project feature. The dataset consists of projects carried out in the period 2003 to 2020. To sensibly compare cost figures for projects distributed over the period, the price level for estimates and final costs has been adjusted to the turn of 2021/2022.

The data have several areas for improvement. The lack of standards for reporting final costs, unique preconditions and different designations weaken the basis for comparison. Important challenges not considered in the study include changes in scope and maturity on the basis of investment licensing and reporting routines. Earlier studies confirm that such factors influence the relative deviation between the estimate and final cost [12]. For example, significant changes may have been made to the project following an investment decision that affects the relative deviation without this being due to cost consequences of uncertainty. Despite weaknesses in the dataset, the volume and sample of projects should be representative to provide credible insight into relationships between project features and variances in estimated and actual project costs.

3. Methodology

The dataset is analysed through descriptive methods and regression analysis to reveal relationships between project features and relative deviations. Of the 489 projects on which data have been collected, 215 are suitable for the study's analyses. The criteria for the sample have been that they have data on

- Estimate at investment decision and actual final cost
- Gross total area
- Type of intervention (new build, rehabilitation or both)
- Building type (school, prison, museum, etc.). These have further been categorised by degree of distinctiveness / peculiarity. A low degree indicates the types of buildings built frequently, and vice versa for a high degree.
- Project execution model (design-build contract, build-only contract or collaborative contract).

The methodology used in the study is shown in Fig. 1.

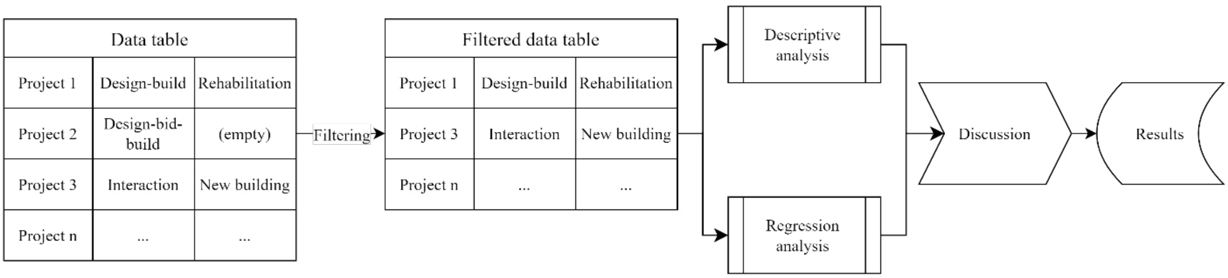

Fig. 1. Process for filtering and analysing data in the study.

The database enables descriptive analyses by filtering the sample data by measure type, project execution model, client, building type, total cost, total area and square meter price. In addition, the 10th, 50th and 90th percentile for relative deviations (shown in Table 1 as P10, P50 and P90) are studied for the sample. Thus, one can see how the discrepancy between estimated and actual project costs is affected by different filtering.

A weakness of the descriptive analysis is that it is difficult to assess whether the studied relationships are due to chance or other features and variables. Covariation between several variables is not captured. Regression analysis assesses the effect of the relative deviation of several variables combined. The relative deviation between estimated and actual project cost is set as the target variable.

To determine the best regression line, the least squares method is used with the formula

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_n X_{ni} + \epsilon_i \quad (2)$$

Where β is the regression coefficient and X variable for each project feature n. The least squares method attempts to minimize the sum of the squares of the residuals ϵ for the list of projects i. The total area is a continuous variable given in the number of thousands of square feet. The square metre price is divided into categorical (dummy) variables, which compare with the interquartile area, i.e., projects with a square meter price between the 25th and 75th percentile of the square meter prices in the data set. Investment type, contract form and client organisation are categorical variables comparing to the most common type (mode).

The degree of explanation R^2 shows how much variation is explained by independent variables, and tests for multicollinearity and heteroscedasticity have been done to validate the models. When interpreting the estimated relationships from the regression models, we focus on the significance level of the estimated coefficients associated with the explanatory variables included in the model. Significance is illustrated by asterisks in the regression tables.

One asterisk represents a significant correlation at a 10 per cent level, two asterisks 5 per cent and three asterisks 1 per cent. Our analyses interpret P-values greater than 10 per cent as an indication of a correlation without establishing significance.
The regression models used to achieve an explanatory degree of around 0.2, which means that the included project features do not explain more than 20 per cent of the variation in relative deviations. Future studies can hopefully access more standardised data for analyses and more explanatory variables, thus achieving a higher degree of explanation. In this study, it is important to understand that results must be interpreted given the assumptions and sources of error that apply and that rather than allowing quantitative conclusions, this provides a basis for indications for tendencies and hypotheses for further research.

4. Analysis

In this section, the primary analysis results are presented. For each of the project features, there is a sub-section that communicates the results. Quantitative results from the descriptive analysis can be read in Table 1, while results from the regression analysis can be read in Table 2.

Table 1. Filtered data and corresponding percentiles for relative deviation. *n* is the number of observations.

		P10	P50	P90	n
Gross total area	Q1 (Lower quartile)	-5 %	12 %	42 %	48
	Q2-3	-12 %	4 %	30 %	94
	Q4	-13 %	-1 %	19 %	48
Estimated square meter price	Q1 (Lower quartile)	-4 %	13 %	43 %	48
	Q2-3	-11 %	6 %	27 %	94
	Q4	-16 %	0 %	14 %	48
Distinctiveness	Low	-11 %	3 %	28 %	128
	High	-8 %	10 %	34 %	87
Project execution model	Design-bid-build contract	-13 %	3 %	32 %	51
	Design-build contract	-8 %	7 %	30 %	100
	Collaborative contract	-11 %	9 %	22 %	15
Type of measure	New building	-11 %	4 %	30 %	135
	New building and refurbishment	-15 %	6 %	27 %	26
	Renovation	-6 %	11 %	52 %	54

Table 2. Regression results.

Independent variable	Coefficient	Robust std. Err.
Total area (1000 m2)	-0.005***	(0.002)
Square meter price, Q1	0.068**	(0.034)
Square meter price, Q4	-0.077***	(0.021)
New build and renovation	0.004	(0.033)
Rehabilitation	0.027	(0.039)
Collaborative contract	-0.008	(0.031)
Design-bid-build contract	0.012	(0.028)
High distinctiveness	0.015	(0.038)
Constant	0.111***	(0.040)
R^2	0.194	

4.1. Total area

We have studied whether the gross total area is related to the relative deviation for the projects in the data set. Based on the observations, projects with a larger gross total area may have a lower relative deviation than projects with a smaller area. The regression model shows that the total area coefficient is significantly negative at the strictest level (1 per cent). This reinforces the correlation from the descriptive analysis and indicates a correlation between total area and relative deviation of construction costs. The interpretation of the regression coefficient is that for every 1,000 square meters increase in total area, the relative deviation decreases by 0.5 percentage points.

This means that the cost estimate is underestimated more often for small projects than for large ones.

4.2. Square meter price

We have investigated whether the estimated square meter price for the contract given in NOK/m2 is related to the relative deviation in construction costs. We observe that projects with a low estimated square meter price have a higher relative deviation than projects with a higher square meter price. The standard deviation is also greatest for the projects with the lowest estimated square meter price. Projects with the highest estimated square meter price are better on the estimatefor total costs and have the least standard deviation.

The regression model shows that the included indicator variables (dummy) representing the upper and lower quartiles for square meter price are reflected in relative deviations. At the five per cent level, there is a significant correlation between a low square meter price and a high relative deviation. Conversely, the quarter of the projects with the highest estimated square meter price has a lower relative deviation. This finding is significant at the 1 per cent level, and the model thus indicates a clear negative correlation between the estimated square meter price and the relative deviation of construction costs.

The analysis shows that the estimated square meter price and relative deviation are negatively correlated. A low estimated square meter price correlates significantly with a high relative deviation and vice versa.

4.3. Building type

This analysis examines the relationship between building type and deviations in construction costs. Since the data sample for each building type is limited, the building types have been sorted by degree of distinctiveness, where the building type has been categorized with either a high or low degree of distinctiveness. We observe that projects with high distinctiveness have higher relative deviations, but the regression model shows no significant correlation. This means we must assume that there is a fair chance that the correlation is due to other project features.

Descriptive analyses suggest a correlation, but this is not unambiguous. The regression analysis does not provide a basis for concluding that the degree of peculiarity is associated with relative deviations.

4.4. Project execution model

This analysis examines whether the project execution model is related to relative deviation. For analysis purposes, we categorize three distinct execution models, but in reality, there may be different hybrids of these. Part of the project may be executed as Design-build and another as Design-bid-build, with a varying degree of collaboration.

The results suggest a correlation between the project execution model and relative deviation, but the regression analysis shows no significant correlation. The average relative deviation is approximately the same for different project execution models. Therefore, the relationship between contract form and relative deviation is not clear-cut.

4.5. Investment type

This analysis examines whether investment type new-build or rehabilitation is associated with relative deviations. We observe that the rehabilitation projects in the dataset have a higher relative deviation than new build. Furthermore, P90 is far higher than for new buildings, which means that the largest overruns are often renovation projects. The regression analysis also tells us that the relationships from the descriptive analysis can, to a small extent, reject the null hypothesis that the correlation is not due to chance.

The descriptive analysis indicates that investment type is correlated with relative deviations in construction costs, where the extent of rehabilitation in a project is related to a higher relative deviation. However, the regression analysis shows no significant correlation and suggests that the correlation is due to other project attributes.

5. Discussion

To use results from the analysis in uncertainty analysis for future projects, the results must be interpreted and transformed into knowledge about uncertainty in projects. Initially, we discuss whether empirical data on relative deviations in construction costs are a good indicator of expected additions in future projects.

The expected addition is the expected relative deviation between the base estimate and the actual final cost for the entire investment. As we have defined relative deviation in Equation 1, we compare the budget to actual construction costs. However, construction costs are only a part of the total investment, as the client will have internal costs as a share of the total investment and, therefore, the uncertainty. Our view is that in most projects, cost elements of construction costs constitute by far the largest share of the costs in a construction project. This means that the uncertainty in the contract dominates the contribution to overall uncertainty. This suggests that, in relative terms, it is a good approach to consider total uncertainty in construction costs corresponding to the total uncertainty in the project.

The data consists mainly of data from two of the largest state-owned client organizations in Norway, and several of the large municipalities and county authorities. With large investment portfolios, market participants have established project models and routines to ensure good project management. This argues that the projects in the database are representative of typical project implementation in the public sector. At the same time, the dataset does not represent all types of public construction projects. For example, 90 per cent of the projects in the database are smaller than 11,000 square meters, and some public works projects can have sizes of several tens of thousands of square meters. This may apply to hospitals, universities or larger public office buildings. For such megaprojects, the data set will not be representative. The same can be said for facilities with a non-ordinary high or low square meter price. For example, a laboratory may have a very high square meter price, and the uncertainty in such a project lacks a good reference basis in the study's data. Therefore, we argue that the data used in the study represent most public construction projects, but specific buildings are not sufficiently represented.

Finally, we discuss the findings and their applicability.

Gross total area: The analyses show that the total area has a significant correlation with relative deviations in the data set, and a negative correlation is demonstrated. This implies that smaller construction projects receive a higher cost increase than larger projects. This may seem surprising, as it could intuitively be expected that the implementation of smaller construction projects has a lower complexity; The project scope is more transparent, implementation takes place in a more limited area, and there are often fewer actors and stakeholders. To try to explain the correlation, we develop hypotheses as possible explanations. The first hypothesis concerns different maturity levels in investment decisions for large and small projects. Large projects often involve the most significant risk exposure in an investment portfolio. The requirements imposed on the project when making an investment decision may therefore be stricter the larger the project becomes. This may entail a more detailed design, more robust corporate governance and professional resources throughout the planning phase. A client may be willing to make an investment decision on a more immature basis for a smaller project because, at the portfolio level, there is a lower risk in absolute costs if the estimate should be underestimated. Thus, it is conceivable that smaller construction projects have a lower level of maturity when making investment decisions than larger projects, contributing to a larger relative deviation.

Estimated square meter price: Like total area, square meter price is negatively correlated with relative deviation in construction cost. This implies that projects with a high estimated square meter price have a lower expected premium than projects with a low estimated square meter price.

A low estimated square meter price implies either that the project is a low-cost building or that the project is underestimated. An underestimated project means that elements have been omitted, documentation is immature, or the basis for the estimate is otherwise incomplete. A low estimated square meter price may thus explain a larger relative deviation. On the other hand, for projects where the estimated square meter price is high, it can be assumed there is greater room for manoeuvrability to reduce in qualities. This means that in projects where the estimate appears to be tight, larger additions must be expected than in projects where the estimate seems to be spacious.

Building type: The descriptive analysis of building types indicates a positive correlation between the peculiarity of the building type and relative deviation, but we find no significant correlation in the regression model. This may indicate that it is not the distinctive character of the building type that correlates with relative deviations but other project attributes. Projects with a building type of low peculiarity must, in many cases, satisfy defined requirements in accordance with the building regulations (i.e., a school building) and thus have little flexibility in managing scope in the event of cost overruns. To a lesser extent, such requirements are defined for building types of higher features. Although projects with high distinctiveness have uncertainty associated with a lack of experience, flexibility in scope can thus contribute to lower costs, which means there is no indicative correlation between peculiarity and relative deviation. However, the results mean we cannot conclude on a correlation, so it should be studied more closely why the descriptive analysis shows a correlation.

Investment type: From the analysis, we see a positive correlation between the extent of rehabilitation and relative deviation, but the regression model finds no significant correlation. A common perception from project implementation is that there is greater uncertainty associated with rehabilitation projects than new buildings. This is because it is more difficult to anticipate all necessary measures in a rehabilitation project and because the condition of existing buildings is not always well-mapped. The study does not provide a definitive answer as to whether this is the case. Still, the deviation in the 90th percentile for the category «Rehabilitation» of as much as 52 per cent shows that rehabilitation projects may, in the worst case, be far more expensive than estimated. This does not necessarily have much impact on expected cost increases, but it should be taken into account in the uncertainty provision.

Project execution model: The analysis shows no significant correlation between the project execution model and relative deviation. This may be because the project's distinctive nature and uncertainty profile affects the choice of project execution model and not the other way around. The choice of project execution model is based on the features of the project. Design-build contracts are suitable, for example, in many projects where it is easy to define precise requirement specifications and where few changes are expected. Design-bid-build contracts are suitable where the client has great expertise about the project to be implemented and will have control over all changes. A collaborative contract is often used in complex projects where the scope could be clearer to the developer. The fact that the analysis shows little correlation between the project execution model and relative deviation is logical.

6. Conclusion

Filtering the data set provides historical figures on relative deviations in projects with similar features. The median deviation from such a sample is the historical benchmark value for expected additions in a project with the same features. For example, let's say a new school is to be built, and it is based on an expected addition similar to the median deviation that can be extracted from the data set for schools. This will not serve as a blueprint for predicting the relative deviation between the estimate and final cost when the project is completed. On the other hand, it is a good starting point for the uncertainty assessment, where subject matter experts can argue why this specific project differs from the historical median.

From the analysis and discussion results, we believe that the project execution model, building type and investment type are not a basis for determining a benchmark for expected additions. Still, the correlation we see in the descriptive analysis should be used to qualitatively assess the project's uncertainty profile, both in terms of expected cost and size of uncertainty provision. On the other hand, area and estimated square meter price are good indicators for determining whether the basic calculation is likely over- or underestimated and to what extent.

7. Recommendations

We believe that uncertainty analyses can be improved through data-driven methods where data from completed projects are used as input in the estimation of new projects.

- **Further develop a database of public works projects and revise analysis results.** The study results are based on only a sample of data on public construction projects that have been available for this study. In addition, many of the data points have not been used for analysis purposes as there is a lack of information that makes them inapplicable to the study's analyses. Therefore, we assume there is a potential to establish an even more complete database with higher accuracy.

- **Standardized documentation of completed public construction projects.** We see the potential for data analysis of completed projects to provide great value in uncertainty analysis and cost management in public construction projects using good, high-resolution data. Robust findings require high data quality. To improve data quality, it will be of great value to ensure consistent and standard documentation of project data. This is possible by establishing a standard for documenting public construction projects that is applicable for analysis purposes. Today, there is a big difference in how public actors retain data about their completed projects. A standard could facilitate the exchange of experience across the board. The preparation and introduction of such a standard may, at best, be of great economic value.

- **Standardized requirement for documentation of costs incurred by the contractor.**
  It can be virtually impossible for the client to document the final costs for different building components in a Design-build contract. To have better predictability and traceability of costs, it could therefore be advantageous to require the contractor to submit actual final costs for different building components. This will give public builders a better understanding of the costs incurred.

References

[1] Austeng, K., Torp, O., Midtbø, J.T., Helland, V. and Jordanger, I., 2005. Usikkerhetsanalyse - Metoder. Concept-rapport no. 12. Trondheim: Norwegian University of Science and Technology.

[2] Chadee, A., Hernandez, S.R. and Martin, H., 2021. The Influence of Optimism Bias on Time and Cost on Construction Projects. Emerging Science Journal, 5(4), pp. 429-442.

[3] Odeck, J., 2017. Variation in cost overruns of transportation projects: an econometric meta-regression analysis of studies reported in the literature. Transportation, 46, pp. 1345–1368

[4] Sarmento, J.M. and Renneboog, L., 2017. Cost Overruns in Public Sector Investment Projects. Public Works Management & Policy, 22(2), pp. 140-164.

[5] Welde, M., Jørgensen, M., Larssen, P. F. and Halkjelsvik, T., 2019. Estimering av kostnader i store statlige prosjekter: Hvor gode er estimatene og usikkerhetsanalysene i KS2-rapportene? Concept-rapport no. 59. Trondheim: Norwegian University of Science and Technology.

[6] Adam, A., Josephson, P.E.B. and Lindahl, G., 2017. Aggregation of factors causing cost overruns and time delays in large public construction projects: Trends and implications. Engineering, Construction and Architectural Management, 24(3), pp. 393-406.

[7] Cantarelli, C.C., van Wee, B., Molin, E.J.E. and Flyvbjerg, B., 2012. Different cost performance: different determinants?: The case of cost overruns in Dutch transport infrastructure projects. Transport Policy, 22, pp. 88-95.

[8] Lundberg, M., Jenpanitsub, A. and Pyddoke, R., 2011. Cost overruns in Swedish transport projects. CTS Working Paper 2011:11.

[9] Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Skamris Holm, M., Lovallo, D., Lunn, D., Molin, E., Rønnest, A., Steward, A. and van Wee, B., 2018. Five things you should know about cost overrun, Transportation Research Part A: Policy and Practice, 118, pp. 174-190.

[10] Welde, M. and Klakegg, O.J., 2022. Avoiding Cost Overrun Through Stochastic Cost Estimation and External Quality Assurance. IEEE Transactions on Engineering Management, doi: 10.1109/TEM.2022.3173175

[11] Berg, H., Bukkestein, I. and Nyhus, O. H., 2022. Kostnadskontroll i statlige prosjekter med og uten ekstern kvalitetssikring. Concept arbeidsrapport 2022-1. Trondheim: Norwegian University of Science and Technology.

[12] Love, P.E.D., Smith, J., Simpson, I., Regan, M. and Olatunji, O., 2015. Understanding the landscape of overruns in transportation infrastructure projects. Environment and Planning B: Planning and Design, 42 (3), pp. 490–509.