LAPPEENRANTA-LAHTI UNIVERSITY OF TECHNOLOGY LUT

School of Business and Management

Master's Thesis, Business Analytics

Antti Virtanen

FORECASTING CASH FLOW CURVE OF CONSTRUCTION PROJECTS USING SUPPORT VECTOR REGRESSION AND PROJECT COST COMPOSITION

1st Examiner: Professor Mikael Collan

2nd Examiner: Post-doctoral researcher Jyrki Savolainen

ABSTRACT

Lappeenranta-Lahti University of Technology LUT

School of Business and Management

Master's Programme in Business Analytics

Antti Virtanen

Forecasting cash flow curve of construction projects using support vector regression and project cost composition

Master's thesis

2021

80 pages, 12 tables, 10 figures and 3 appendices

Examiners: Professor Mikael Collan and Post-doctoral researcher Jyrki Savolainen

Keywords: Project cash flow forecasting, nomothetical modeling, support vector regression, construction project management

Cash flow management is a crucial factor of construction project profitability and its negligence contributes a significant portion of contractor bankruptcies. This study proposes a novel cash outflow forecasting model. The model applies a machine learning method, support vector regression (SVR), on historical data of similar projects to forecast the current project's cash outflow from the beginning to the end of construction. In the proposed model project key characteristics are identified via project k-means clustering and project cost composition before producing the cash outflow forecast. The model is tested and verified using actual data from 33 projects of a Finnish general contractor. The forecasting model and its intermediate versions are benchmarked against the current state-of-the-art approaches found in the literature.

A systematic literature review of the current cash outflow models in the construction industry is conducted. The review shows that cash outflow is forecasted indirectly by estimating a cost commitment curve with a linear logit model and applying a fixed timelag based on project cost composition. The issues with this approach is that it cannot fit non-linear relationships and assumes that different cost categories are incurring at a uniform rate which results to a systematic error. The proposed model addresses the identified issues by applying non-linear methodology to forecast cash outflow directly and utilizing project cost composition to estimate the cash outflow curve profile which makes it novel from the theoretical perspective.

The results of the proposed model performance are promising. Forecasting cash outflow directly reduced the average error by 5.41% compared to the often used indirect approach. The use of SVR improved the model's ability to fit an individual project and utilization of project cost composition had a similar effect in the pre-construction phase reducing the root mean squared error (RMSE) to 7.75% from 10.25% RMSE observed with the standard approach. Within the construction phase, the average error reduced from -2.33% pre-construction level to an average of -0.67%.

Tiivistelmä

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Master's Programme in Business Analytics

Antti Virtanen

Rakentamisprojektien kassavirtakäyrän ennustaminen tukivektoriregression ja projektin kustannusrakenteen avulla

Kauppatieteiden pro gradu -tutkielma

2021

80 sivua, 12 taulukkoa, 10 kuvaa ja 3 liitettä

Tarkastajat: Professori Mikael Collan ja Tutkijatohtori Jyrki Savolainen

Avainsanat: Projektin kassavirran ennustaminen, tukivektoriregressio, rakentamisprojektien hallinta

Kassavirran hallinta on ratkaiseva tekijä rakentamisprojektien kannattavuudessa ja sen laiminlyönti aiheuttaa merkittävän osan urakoitsijoiden konkursseista. Tutkimus ehdottaa uudenlaista kassavirran ennustemallia, jota voidaan käyttää ennen rakentamisen aloittamista sekä sen aikana. Malli soveltaa koneoppimismenetelmää (tukivektoriregressio) ennustamaan nykyisen projektin kassavirtaa rakentamisen alusta loppuun käyttäen vastaavien projektien historiallisia tietoja. Se tunnistaa projektin ominaisuudet projektien ryhmittelyn (k:n keskiarvon klusterointimenetelmä) ja kustannusrakenteiden avulla. Mallin toimivuus on testattu ja todennettu käyttäen toteumatietoa suomalaisen pääurakoitsijan 33:sta projektista. Ennustemallia ja sen väliversioita verrataan kirjallisuuden johtaviin lähestymistapoihin.

Rakennusteollisuuden nykyisistä kassavirtamalleista tehdään systemaattinen kirjallisuuskatsaus, joka osoittaa, että kassavirta ennustetaan epäsuorasti arvioimalla kustannuskäyrä lineaarisella mallilla (logaritminen lineeariregressio) ja käyttämällä kiinteää aikaviivettä, joka perustuu projektin kustannusrakenteeseen. Lähestymistavan ongelmana on, että se ei sovellu mallintamaan epälineearisia suhteita ja se olettaa kustannuskategorioiden samantahtisen toteutumisen, mikä johtaa systemaattiseen virheeseen. Ehdotettu ennustemalli vastaa tunnistettuihin ongelmiin soveltamalla epälineaarista menetelmää kassavirran suoraan ennustamiseen ja arvioimalla kassavirtakäyrän muotoa projektin kustannusrakenteiden avulla. Tämä tekee mallista uuden teoreettisesta näkökulmasta.

Ehdotetun mallin suorituskyvyn tulokset ovat lupaavia. Kassavirran ennustaminen suoraan pienensi keskimääräistä virhettä 5.41% verrattuna yleisesti käytettyyn epäsuoraan ennustamiseen. Tukivektoriregression käyttö paransi mallin kykyä ennustaa yksittäinen projekti sekä projektin kustannusrakenteen hyödyntämisellä oli samanlainen vaikutus rakentamista edeltävässä vaiheessa, jossa ne paransivat mallin keskineliövirheen neliöjuuren (RMSE) 7.75%:iin tavanomaisen lähestymistavan 10.25%:sta. Rakenamisvaiheessa keskimääräinen virhe pieneni rakentamisvaihetta edeltävästä -2.33%:sta -0.67%:iin.

Table of contents

1 Introduction

Construction projects are identified as unique and they typically last long periods, especially when building new and large structures. Nam & Tatum (1988) list the main five characteristics of construction products that are immobility, complexity, durability, costliness, and a high degree of social responsibility. As a result of these qualities and their implications, construction has been classified as a high-risk industry.

This study attempts to build a mathematical cash flow forecast model that can be used to control the financial risk involved in contracting. This is done by quantifying required financing for ongoing and future known projects. The proposed model concentrates on the cash outflow component as it can be predicted mathematically with a satisfactory error rate, whereas the inflow component is heavily correlated with contractual terms. In addition to the above-mentioned predictive abilities, it also benefits contractors as it demands only general data of the projects. Therefore, it requires a minimal amount of site-level interaction, thus reaching a high level of automation.

1.1 Background

Contractors are constantly bidding on new projects in their tender phase after which they move on to a planning phase that has a varying duration depending on the contract. This follows with the actual construction phase that ends in a project handover and guarantee phase. As construction companies have numerous contracts in various phases simultaneously, they must prepare their cash flow regardless of the project phase. The information that a bidding contractor has in a tender phase or even after winning the contract (planning phase) is very different compared to the construction phase when project plans are available. Therefore, the required forecasting model should be able to generate predictions for both the pre-construction (tender and planning) and construction phases.

One of the risk-increasing factors in the construction industry is that typically contractors are competing for projects with an emphasis on the lowest price which has

resulted in low and unreliable profit margins (Sorrell, 2003; Teerajetgul et al., 2009). This has led to alternative ways of increasing profitability through efficient project cash flow management and cash farming. Increasing the amount of positive cash flow from a project raises profitability in two ways. First, the required amount of capital that a contractor invests into a project is smaller, hence the return on investment percentage is higher. Second, the positive cash flow that is generated at the beginning of the project via unbalancing of the contract is available for reinvestment. However, in the latter case seeing this money as a profit instead of trade credit has led to increased insolvencies in the industry. (Kenley, 1999)

Boussabaine & Kaka (1998) and Hwee & Tiong (2002) state that the construction industry has proportionally higher number of bankruptcies than any other sector. For this reason, bank managers are often reluctant to grant loans to contractors with a liquidity problem, and even if they do, the cost of the loan will most certainly reflect the conceived risk with the loan (Navon, 1996). For the above-stated reasons, adequate financial management and accurate forecasting are essential in the construction industry to make sufficient provisions and guarantee the financing of the contracts that include periods of negative cash flow.

Due to the distinct characteristics of the construction industry, its financial traits are of their own kind. Tserng et al. (2014) list some of these characteristics, such as, a need for large cash supply, short-term financing caused by running simultaneous projects, large inventories that are filled with in-progress construction and materials in addition to high book value inflated with valuable machines and equipment. The fact that the contractor's capital is invested in illiquid assets while its operations require extensive amounts of cash makes the management of working capital and cash flow indispensable in the construction industry. Hwee & Tiong (2002) state that cash flow is the most important factor of profitability for in-progress construction projects. A questionnaire for construction contractors conducted by Shash & Qarra (2018) indicates that 40% of the respondents encounter financial failure in some of their contracts annually due to poor cash flow management. Therefore, contractors cannot manage their financials only in terms of revenue and costs as they also need to

consider actual cash in and cash out which are, for later clarified reasons, two highly dissimilar concepts.

Finance is in fact identified as the most important resource in the construction process (Mawdesley et al. 1997, cited in Odeyinka et al. 2008). Singh & Lokanathan (1992) state that more construction firms fail through lack of liquidity than by inadequately managing other resources, which makes cash the most important one. Similarly, Peer & Rosental (1982, cited in Navon 1996) find that lack of working capital causes more failures in construction companies than does their profitability. Overall, four out of five most common reasons, why construction businesses fail, are budgetary issues (Arditi et al. 2000). However, the provisions that are taken should be adequate to finance projects but not cause a permanent surplus of funds which is itself also an uneconomic state of affairs (Kaka, 1990).

1.2 Motivation

The motivation of this thesis is to offer an efficient cash flow forecasting model for a central organization of a construction company. The need for a mathematical cash flow forecasting model has been also noted in the literature. However, the previous research has its focus on modeling client-side cash flow and tender phase in addition to using conventional methodology that is based on linear relationships. A more sophisticated mathematical model is therefore needed to reduce the systematic error that is caused by the previous models.

An alternative to mathematical forecasting would be compiling the forecast at the site level. Even though site engineers and project managers can compose accurate project cash flow predictions with very detailed site-level information of the project, this is often cumbersome work because of the complex linkage between cost items and project schedule. In addition, these undergo frequent changes during the project and taking the later specified cash flow affecting factors into account increases the complexity of cash flow forecasting. This is true especially for large projects. Altogether, an efficient cash flow forecasting method is not only needed in the tender phase as limited

resources and complex relationships that affect cash flow are still causing inaccuracies in manually derived the construction phase cash flow forecasts.

Mathematical models, on the other hand, can offer close approximates and their errors are consolidated in a company-level cash flow forecast. Navon's (1996) survey discloses that all of the surveyed construction companies prepare their cash flow at a company level. In addition, the majority of the contractors, that do project-level cash flow predictions in parallel, do them centrally (Navon, 1996). Similarly, Kaka (1993) describes that cash flow and working capital forecasts are usually done on an overall basis. This indicates that there is a need for a mathematical model that is efficient and able to provide sufficiently accurate forecasts with general data, instead of site-level information.

The uniqueness of construction projects offers its own kind of difficulties in project forecasting. On top of the above-mentioned financial requirements, construction projects have numerous variables affecting their outcome and their relationships are often unclear. Kenley & Wilson (1986) argue that in addition to direct construction and contract-related factors, others such as economic, political, managerial, union and personality-related variables cause variation in project outcomes. Chan et al. (2009) also state that project duration and cost are reliant on many uncertain factors like productivity, resource availability and weather. When modeling project cash flow, Zayed & Liu (2009) identified 43 factors that affect it. In addition, construction managers have a control over none or just a few of these variables.

The ambiguity related to project cash flow makes forecasting difficult for estimators or project managers. This makes a simple and fast cash flow forecasting technique important especially in the tender phase where detailed schedules are rarely planned because time is lacking and information is limited (Kaka & Price, 1993). There is also some evidence that statistical models with large training data can offer superior forecasts compared to contractor's initial estimates (Mills & Tasaico, 2005). The results of Shash & Qarra (2018) also suggest using quantitative forecasting models in the

tender phase. They find that the vast majority of contractors do cash flow forecasting only before bidding with a focus on surviving throughout the contract instead of getting a measurable financial view on the project cash flow (Shash & Qarra, 2018). Without a quantitative and time-bound financial forecast, it is difficult if not impossible to get an accurate view of the contractor's financial requirements which is why there is a need for a forecasting model in the tender phase of a project.

Project forecasts are typically done by budgeting and deriving income from the project schedule. However, cash flow prediction is just a bit more tedious as different cost categories have distinct time lags concerning their cash disbursement and using the project schedule makes income forecasts particularly exposed to delays. Cui et al. (2010) list several reasons that make revenue and expenses differ significantly from actual cash flows. Some of these reasons are investing (for example in equipment) and depreciations related to it, front-end loading techniques which include unbalanced pricing and overbilling, accrual accounts (for example prepaid expenses, receivables and inventories), payment lags, retainage, deferring payments for subcontractors or using pay-when-paid clause with suppliers (Cui et al., 2010). Park (2004) criticizes the traditional approaches as they often do not consider these factors, especially after the planning stage, but they rather use cost and earned value directly in forecasting cash flow.

Park et al. (2005) find that models, that use monthly cost and earned value forecasts as cash flow prediction basis, entail a possibility of inaccurate predictions if the used forecasts are imprecise. This can often be the case as keeping the monthly financial forecasts up to date is time-consuming and may not be the highest priority in a construction site. Park (2004) finds that during the construction phase the relative portion of different cost categories is fluctuating from the original project budget. However, in practice, this is often not reflected in cash flow forecasts which should be done by adjusting cost categories' relative weights with respect to the actuals (Park, 2004). This causes that the time lag related to the remaining costs is distorted and the cash flow forecast is incorrect. Therefore, it is highly beneficial that the used mathematical cash flow forecasting model can be also used in the construction phase.

1.3 Research objectives

This study aims to contribute to the literature by applying machine learning together with the (moving) cost category weights approach. This is something that has not been suggested previously in the literature to the author's knowledge. This approach benefits the industry as it offers a mathematical model that is better able to capture complex relationships by applying a more sophisticated algorithm compared to the traditional approach. It requires only general data (for example estimated total cost and weights of cost categories in terms of financial data) which are often available in an applicable form as opposed to project schedule, monthly budgets or earned-value planning data. The proposed model can be used to predict project cash outflow from the tender phase to the end of the construction and it is tested with a comprehensive, heterogeneous dataset that is required to study the model's ability to capture individual project's uniqueness.

There has been a slow trend towards artificial intelligence (AI) and machine learning (ML) in the construction management literature, but Hua (2008) points out that in construction economics and project budget and cash flow area, conventional methods are generally applied more often than in other construction management topics.

Figure 1 illustrates, how this study combines three research areas in construction management and economics. It does not only insert a new machine learning method into an old model, but it also complements the traditional approach by exposing it to some previously less researched data and suggests a new forecasting model. In a similar manner, the study does not only remain in the management area which is often focused on analyzing causal relations in construction data by machine learning, but it offers a usable, quantitative model for production use. Last but not least, the proposed model offers a higher level of automation compared to site-level models via machine learning as it does not require site-level interaction apart from categorized project end forecasts which should be accessible also for the central organization.

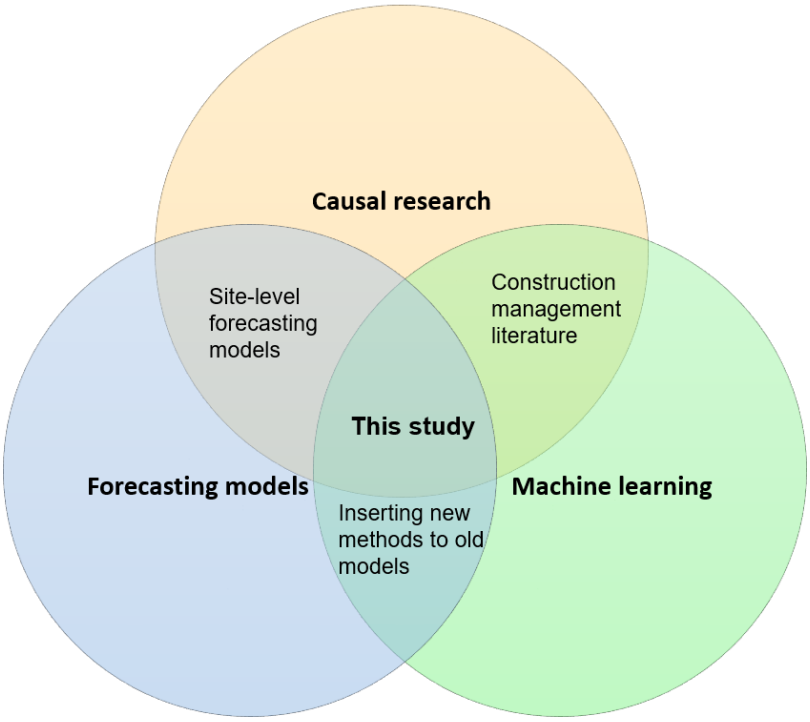


Figure 1. *The combination of research areas in construction management and economics involved in this study*

1.4 Research questions

The study aims to answer two questions:

1) According to the literature, how cash flow forecasting of construction projects is performed and what are the central issues?
    - How will direct forecasting of the cash outflow curve perform compared to forecasting the cost curve and applying a fixed time lag?
    - How will cost categorized project end forecasts affect the accuracy of cash outflow predictions in different phases of a construction project?

2) How to improve the current support vector regression based cash flow models?
    - How is support vector regression currently applied?
    - Can support vector regression be used and how to better capture the relationships between cash outflow and other financial data compared to the standard approach?

To answer the first research question, this study seizes two of its sub-questions that revolve around identified issues in current cash flow forecasting models and proposed solutions to them. The intuition behind these questions is that different cost categories have different time lags regarding their cash disbursement and their relative size is therefore affecting project cash flow profile. In an extreme case at the end of construction, the weight of guarantee provisions might take up most of the remaining cost budget and this category is usually not cashed out in the construction phase, if at all. Combined with an observation made by Park (2004), that different cost categories are not occurring at a uniform rate, the hypothesis for the first sub-question is that applying fixed time lag to the cost curve will introduce systematic error to the cash flow model. Therefore, the hypothesis for the first sub-question is that forecasting the cash outflow curve directly will outperform the traditional approach.

Similarly, for the second sub-question, the hypothesis is that cash outflow predictions should improve when weights of budgeted cost categories are known. To answer this question, tender phase data is enriched with weights of different costs in project end forecast that are later modified with respect to actuals and forecast changes in the construction phase. Therefore, the second sub-question can be divided into two:

a. How will budgeted cost category distribution affect tender phase predictions?
b. What is the effect of adjusting cost category weights in the construction phase predictions?

As the standard approach of generating an S-curve with a logit model by Kaka & Price (1991) is not suitable for multiple variables, this study explores the possibility of using support vector regression to generate it. S-curve is used as a graphical representation that shows the project's cumulative progress against time. Additionally, the numerous variables affect project cash flow with complex relationships. Therefore, linear regression and one independent variable (time) might not be the best basis for mathematical forecasts. Sapankevych & Sankar (2009) observe that support vector regression is not dependent on linear and stationary processes. Therefore, the

hypothesis is that the suggested approach using support vector regression can make better predictions than the logit model, which is based on one independent variable, log transformation and linear regression.

Simultaneously, the research must critically assess the model's ability to capture the uniqueness of individual projects by analyzing the benefits that are gained by clustering the projects. This is because projects are fundamentally unique and average curves will certainly lead to systematic error which can be reduced only by accurate project grouping (Kenley & Wilson, 1986; Kaka & Price, 1993).

1.5 Limitations

In terms of financial data, the study uses only categorized project actuals and project end forecasts which limits some special characteristics in the project schedule outside of the model. Because of this, the results cannot be directly compared with models that use monthly budgets and earned-value forecasts or cost-schedule-integrated models.

Even though the research studies project cash flow, it focuses only on the cash outflow component. This is justified by the findings of Kaka & Price (1993) and Evans & Kaka (1998), who conclude that a standard value curve cannot be fitted even for a specific group of projects because value curves are uniquely distorted by unbalancing and over-measure. Therefore, if the proposed model needs to be expanded into a net cash flow model, cash inflow should be derived from the project schedule because contractual terms are giving too much weight on the profile of the value curve.

Another limitation considers the source of data. The study uses heterogeneous data in terms of project classifications as it contains infrastructure and building projects in multiple segments. The data is retrieved from a general contractor with a long history and well-defined processes which makes different projects' data comparable. However, as the data is collected from only one contractor, it cannot assess whether the model can find similarities and differences between distinct contractors' projects.

1.6 Structure of the study

The second chapter goes through the methodologies used in this study. The third chapter reviews the relevant literature after which the data collection process and the proposed model are described in the fourth chapter. In the fifth chapter, empirical results of the model are presented and followed by results analysis. Finally, conclusions are represented in the sixth chapter.

2 Methodology

This chapter goes through the key methodologies that are used in this study which are K-means clustering, support vector regression, kernel functions and hyperparameter optimization.

2.1 K-means clustering

To capture the uniqueness of projects while maintaining predictive abilities, projects need to be clustered based on their attributes. Cheng et al. (2009) use k-means clustering to identify similar projects. K-means clustering separates dataset $\{X_1, \dots, X_N\}$ with *N* observations of random *D*-dimensional Euclidean variable x into *K* number of clusters. The goal of the algorithm is to set $\{\mu_1, \dots, \mu_K\}$ *D*-dimensional vectors as cluster centers and assign data points to the nearest cluster center in a way that the sum of squares of the distances between each datapoint and its respective cluster is minimum. The assignment of each datapoint can be indicated with a binary variable $r_{nk} \in \{0,1\}$, where $\mathrm{k} = 1, \dots, \mathrm{K}$ represents cluster k which datapoint x_n is assigned to, so that $r_{nk} = 1$ and $r_{nj} = 0$ for $j \neq k$. The objective function is defined by:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2 \quad (1)$$

which represents the sum of squares of the distance between each datapoint and its assigned cluster center μ_k. The objective is to minimize *J* by finding optimal values for $\{r_{nk}\}$ and μ_k. This can be achieved iteratively by first assigning initial values for μ_k and minimizing *J* with respect to r_{nk} while keeping μ_k fixed. Second, *J* is minimized with respect to μ_k while keeping r_{nk} fixed. This process is looped until convergence. The first step and second step are described by Equations 2 and 3, respectively:

$$r_{nk} = \begin{cases} 1 & if\ k = \arg min_j \left\| x_n - \mu_j \right\|^2 \\ 0 & otherwise \end{cases} \quad (2)$$

which illustrates that each datapoint can be optimized separately by choosing *k* which gives the minimum of $\|x_n - \mu_k\|^2$.

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \quad (3)$$

which describes that the vector μ_k can be assigned to be the mean of all points assigned to the cluster. The process is stopped after there are no changes in assignments. The risk of the k-means clustering is that the solution may converge to a local instead of global minimum, *J*. (Bishop, 2006)

The k-means clustering algorithm is not evaluating the number of appropriate clusters. Therefore in order to perform k-means clustering, the required number of clusters (*K*) is needed. This can be determined using the Elbow Method. Liu & Deng (2020) describe the core idea of Elbow Methods as computing the objective function (Equation 1) for an increasing number of clusters until the benefit of an additional cluster is sharply reduced. If *K* is smaller than the number of required clusters, an additional cluster will significantly decrease *J*. After reaching the true number of clusters, increasing *K* will reduce *J* just slightly. Therefore, plotting *J* and *K* will form a shape of an elbow where the *K* value of the elbow will be the required number of clusters.

2.2 Support vector regression

Support vector regression (SVR) is an application of support vector machines that focuses on regression analysis applications. Some of the advantages of SVR are that it is guaranteed to converge to the optimal solution as opposed to artificial neural networks and it is not dependent on linear and stationary processes. Because optimization is often needed to enhance the performance of the model, it is also beneficial that it has a small number of free parameters left to optimize. (Sapankevych & Sankar, 2009)

When using regression analysis for non-linear regression applications, a function $f(x)$ can be formed so that its outputs are equal to the predicted value:

$$f(x) = \big(w \cdot \phi(x)\big) + b \quad (4)$$

where time-series data $x(t)$ is mapped to higher dimensional feature space via kernel function $\phi(x)$ after which linear regression can be performed with weights w and threshold b. Performing linear regression in high dimensional feature space corresponds to non-linear regression in low dimensional input space. (Müller et al. 1997)

The objective is finding optimal weights for w and threshold b in addition to defining criteria for finding an optimal set of weights. Those can be found by, first, minimizing the flatness of weights that can be ensured by minimizing the Euclidean norm (i.e. $\|w\|^2$). Second, the empirical risk, that is the error generated by the estimation, must be minimized. (Sapankevych & Sankar, 2009) Empirical risk is defined as:

$$R_{emp}(f) = \frac{1}{N}\sum_{i=0}^{N-1} L(x(i), y(i), f(x(i), w) \quad (5)$$

where i is an index of discrete time-series $t = \{0,1, \dots, N-1\}$ and y(i) refers to training data of the predicted value. L(.) is the loss function. (Sapankevych & Sankar, 2009)

However, minimizing empirical risk with no control will lead to overfitting and bad generalization performance. Therefore, a capacity control term $\|w\|^2$ should be introduced. This will lead to regularized risk functional:

$$R_{reg}(f) = R_{emp}(f) + \frac{\lambda}{2}\|w\|^2 \quad (6)$$

where term λ is called regularization constant. (Smola & Schölkopf, 2004)

To find optimal weights for w and minimize regularized risk, a quadratic programming problem can be formed using Vapnik's ε – intensive loss function :

minimize $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} L(y(i), f(x(i), w))$

where

$L\big(y(i), f(x(i), w)\big)$

$$= \begin{cases} |y(i) - f(x(i), w)| - \varepsilon & if \ \ |y(i) - f(x(i), w)| \ \geq \varepsilon \\ 0 & otherwise \end{cases} \quad (7)$$

The constant C in the objective function includes a summation normalization factor $1/N$ and the term ε refers to how accurately the function will be approximated. Equation 7 assumes that function $f(x)$ exists and is feasible. However, in some cases to make the function feasible some errors may need to be accepted which is why some slack variables are introduced. Determining optimal weights and bias values is a problem to be solved with convex optimization which can be done using Lagrange multipliers and dual optimization problem:

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)\langle x(i), x(j)\rangle \\ & -\varepsilon\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{N} y(i)(\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \sum_{i-1}^{N}(\alpha_i - \alpha_i^*) = 0 \text{: } \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \quad (8)$$

Karush-Kuhn-Tucker conditions state that at the point of the optimal solution, the product between variables and constraints equals zero. The solution of the weights can be based on that. Therefore, function $f(x)$ can be approximated as the sum of optimal weights times the dot product between datapoints:

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\langle x, x(i)\rangle + b \quad (9)$$

The datapoints that lie on the limit or outside the ε range with non-zero Lagrange multipliers α are defined as Support Vectors. The optimal weights that are associated with non-zero Lagrange multipliers are usually not the entire dataset. Therefore, this provides sparseness as one does not need the entire dataset to define $f(x)$ which is a significant advantage of support vector regression. (Sapankevych & Sankar, 2009)

As defined in Equation 4, in order to perform non-linear regression, the input space needs to be mapped to higher dimensional feature space using kernel function $\phi(x)$. Any symmetric kernel function *k* that satisfies Mercer's condition corresponds to a dot product in some feature space (Müller et al. 1997). As the algorithm depends on dot products between patterns $x(i)$, it does not need to know ϕ explicitly as knowing that $k(x, x') = \langle \phi(x), \phi(x') \rangle$ is sufficient (Smola & Schölkopf, 2004). This can be

substituted back into Equations 8 and 9 which result to Equations 10 and 11, respectively:

$$\text{maximize} \quad -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right)k\left(x_i, x_j\right) - \varepsilon\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{N} y(i)(\alpha_i - \alpha_i^*)$$

$$\text{subject to} \quad \sum_{i-1}^{N}(\alpha_i - \alpha_i^*) = 0 \text{: } \alpha_i, \alpha_i^* \in [0, C] \tag{10}$$

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{11}$$

2.3 Kernel selection and hyperparameter optimization

As described in the earlier section, the key to non-linear regression is the kernel function. There are multiple kernel functions that satisfy Mercer's condition, and the choice of appropriate kernel function is typically done by empirical analysis (Sapankevych & Sankar, 2009). Waters & Vanhoucke (2014) list some of the most commonly used non-linear kernel functions:

$$k(x, x') = (\gamma\langle x, x'\rangle + r)^d \tag{12}$$

$$k(x, x') = e^{-\gamma\|x - x'\|^2} \tag{13}$$

$$k(x, x') = tanh(\gamma\langle x, x'\rangle + r) \tag{14}$$

Equations 12, 13 and 14 represent polynomial, radial basis and sigmoidal functions, respectively. Depending on the function number of parameters needs to be optimized. All of the functions need to tune gamma γ, sigmoidal and polynomial functions need to optimize the coefficient *r*. Additionally, when using a polynomial kernel, its degree *d* needs to be determined.

This study has chosen to use the radial basis function (RBF) kernel for multiple reasons. First, Lin & Lin (2003) show that the sigmoid kernel resembles the RBF kernel

with certain parameters. Based on this and other unfavorable properties of the sigmoid kernel, they suggest not to use it and to use RBF as the first choice instead (Lin & Lin, 2003). Second, the RBF kernel has only one hyperparameter whereas the polynomial kernel has three. Therefore, using RBF will significantly decrease model complexity. Third, the RBF kernel has fewer numerical difficulties (Bao et al. 2005; Cheng & Wu, 2009; Wauters & Vanhoucke, 2014).

There is little guidance on how to determine parameter values for the chosen kernel (Wauters & Vanhoucke, 2014). Similarly, Sapankevych & Sankar (2009) state based on their literature review that there is no optimal method for choosing free parameters of support vector regression. Hsu et al. (2003) suggest using grid-search and cross-validation. This has also been the commonly applied approach in the literature as it does not make the algorithm overfit to training data, see for example, Espinoza et al. (2005), Bao et al. (2005), Sousa et al. (2014) and Wauters & Vanhoucke (2014). When using grid-search, Hsu et al. (2003), Bao et al. (2005) and Wauters & Vanhoucke (2014) suggest using exponentially growing sequences of C and γ to determine optimal parameters, for example, $C = 2^{-5}, \dots, 2^{15}$, $\gamma = 2^{-15}, \dots, 2^{3}$.

Cross-validation is implemented in a way where the k-fold cross-validation algorithm partitions the training dataset into k folds. After this, the model is trained using $k - 1$ folds as the training data and the resulting model is validated with the remaining fold. The same procedure is applied for each of the folds. Finally, the performance of k-fold cross-validation is measured by the average error in validation sets in the above-described loop. This way the actual test set does not “leak” to the model, and it uses training data efficiently while maintaining generalization performance. The grid-search applies the k-fold cross-validation algorithm for all possible combinations of C and γ after which their optimal values can be determined based on the average error of k-fold cross-validation.

3 Literature review

The literature search process is illustrated in Figure 2. Webster & Watson (2002) suggest a systematic search for literature review in order to get a complete view on the subject especially because the field of information systems is quite an interdisciplinary field.

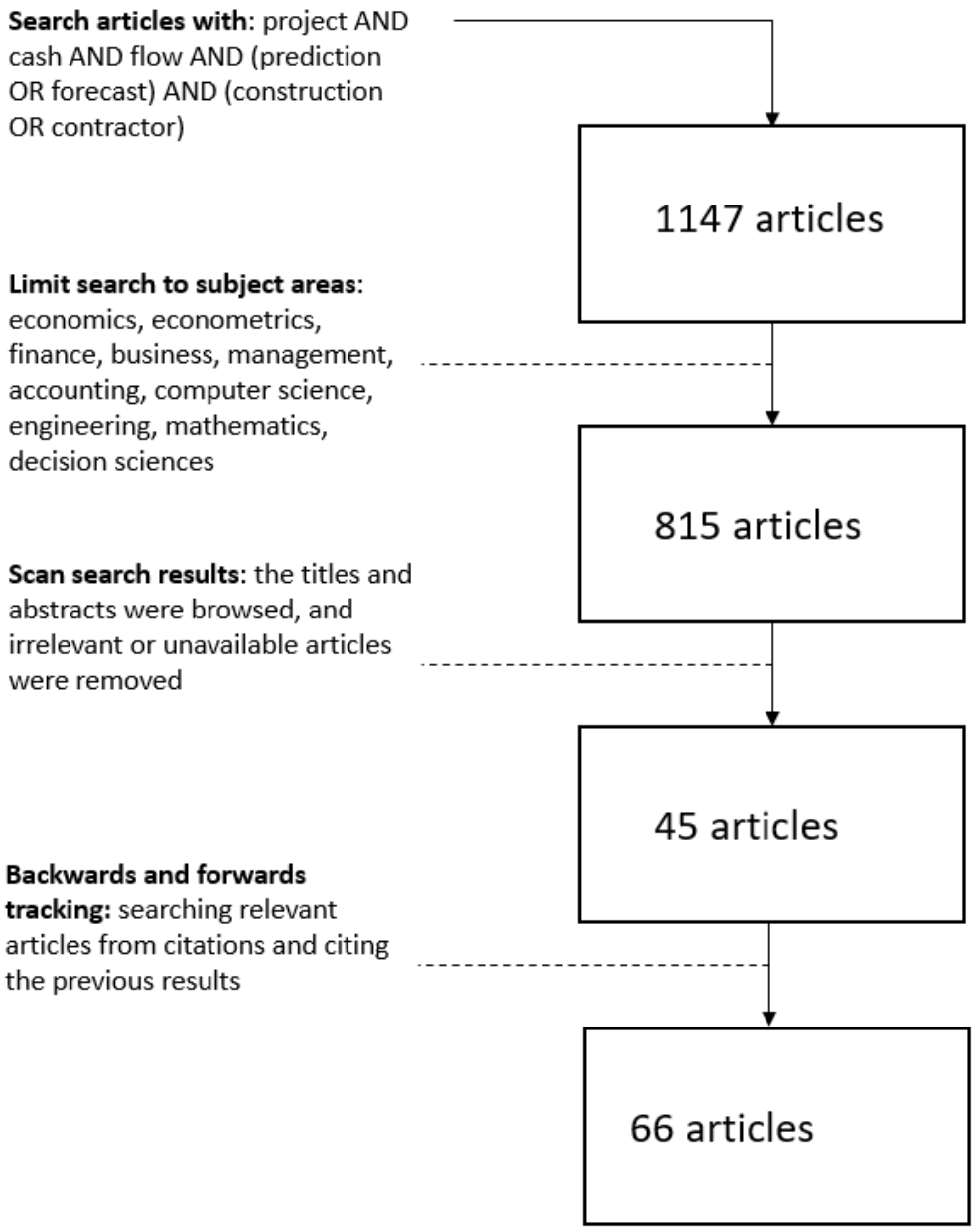


Figure 2. *The literature search process*

The first step of the search was defining a keyword that would describe the construction project cash flow forecasting field thoroughly but restrict irrelevant articles from the search. The used keyword was *"project AND cash AND flow AND (prediction OR forecast) AND (construction OR contractor)".* Cash and flow were used separately as some articles might focus on cash but use terms like cost or expenditure flow instead of cash flow. A database that was used for the search was SCOPUS that compiles literature from multiple fields.

As the search results from the first step contained some irrelevant subject areas, such as medicine and chemistry, the second step of the process was to limit the search to relevant subject areas. However, the selected subject areas were still quite broad in order to get an interdisciplinary perspective. For example, engineering literature might help get a better understanding from site-level planning models whereas finance, accounting, business, management and economics areas could focus more on company-level forecasting. Finally, subject areas such as econometrics, computer science and mathematics were included in the search as they might contain some progressive models that are not applied in a broader scope in the industry.

In the third step of the process, the remaining articles' titles and abstracts were scanned. Only the articles, that were found to be useful in the study or as a connecting reference for other relevant literature, were saved. Some topics from the construction management area were excluded, such as management decision-making, optimization and risk management. However, if the articles on these subjects were also exploring causal relationships related to cash flow, they were included. Additionally, unavailable articles were removed.

In the last step of the process, articles' introduction, literature review or similar sections were skimmed through and relevant citations were added to the literature review material. Additionally, all the articles that were citing the previous step's results were scanned and applicable ones were collected. Similar criteria as in the third step were used.

3.1 S-curve method

S-curve is used as a graphical representation that shows the project's cumulative progress against time. The cumulative progress can be measured, for example, in project value (value curve) or project cost (cost curve). Boussabaine et al. (1999) generalize the cost accrual of a construction project to three phases that form the S-curve:

1) In the first third of project duration, one-quarter of forecasted total costs incur in a parabolic pattern.
2) In the second third of project duration, costs incur in a linear fashion so that three-quarters of forecasted total costs have accumulated.
3) In the last third of the project duration, costs incur as a mirror image of the first third, so that all of the forecasted costs have accumulated.

The initial nomothetical net cash flow model is proposed by Nazem (1968 cited in Kenley & Wilson, 1986) who uses historical project financial data to deduce a standard S-curve that is used to obtain predictions for all of the future projects. He argues that contractors have multiple projects going on simultaneously, and therefore their standard curve would yield capital requirements for the given company. Figure 3 illustrates that the general idea of using a standard curve makes sense in project portfolio forecasting. Aggregating projects' A, B, C and D cash flow together would produce the same result as multiplying the standard curve by four. It would be tempting to use the standard curve as a forecasting basis for future projects as it is easily accessible whereas forecasting the projects in a periodic manner would require a considerable amount of effort.

This approach seems very ideal, especially for large contractors and clients (in terms of clients' cash outflow). This is because small errors in individual projects would not cause significant variation in the total forecast. The intuition behind this idea is that the errors between individual project S-curves and a standard S-curve are random. The randomness of the errors would mean that they are eliminated in an aggregate

forecast. Therefore, the ultimate goal of mathematical models is to get rid of systematic error so that the remaining error is random.

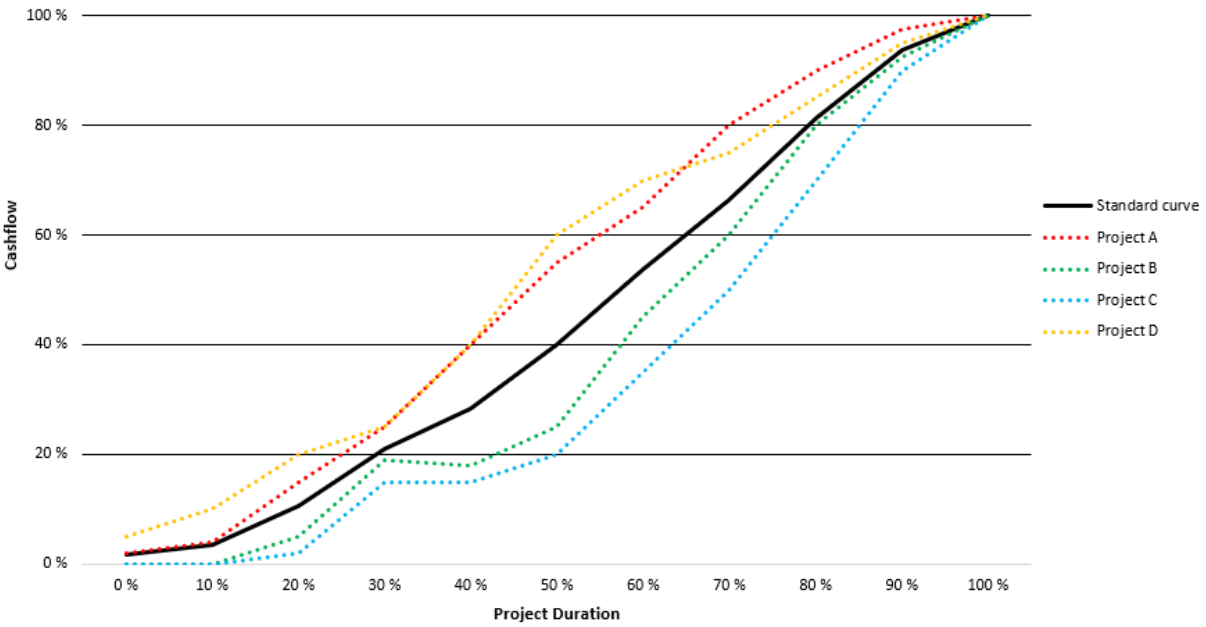


Figure 3*. Standard curve of projects A, B, C and D (nomothetical approach)*

3.2 Uniqueness of construction projects

Kenley & Wilson (1986) argue that construction projects are unique, and the nomothetical approach has only been a temporary solution as the research has been trending towards an idiographic cash flow model the whole time. For example, Hudson's and Maunick's 1974 study tries to search patterns within groups and categories of projects and Berny's and Howe's 1982 model reflects a specific form of an individual project (cited in Kenley & Wilson, 1986). Similarly, Peterman's 1970 and Allsop's 1980 papers take an idiographic approach by pioneering planning data models by basing their value curves on bar charts of bill items and contract schedules, respectively (see, Kaka & Price 1991).

Kenley & Wilson (1986) suggest that the variation in S-curves is caused by a multiplicity of factors in addition to direct construction and contract-related ones, such as economic and political climate, managerial structure and actions, union relations and

personality conflicts. After deriving a threshold value for acceptable standard deviation in estimates, Kenley & Wilson (1986) illustrate that a standard S-curve seems to be fitting a group of projects quite well in terms of how random the errors seem when looking at the whole portfolio of projects. However, only 20% of the projects fit the average model in terms of the determined threshold value. Therefore, this would suggest that using the nomothetical approach would leave significant systematic error in the model which can be removed only with the idiographic methodology that considers the uniqueness of each project. Kaka (1994) highlights that the accuracy of cash flow and working capital forecasts are usually dependent on how sustained the segments in the contractor's project portfolio are compared to last year. Therefore, this is an important observation also for large contractors as their relative distribution of different kinds of projects is most likely not constant throughout the time which implies a risk of systematic errors when using the nomothetical approach.

As the uniqueness of construction projects has solid evidence behind it, the problem at hand is, how to account for the individuality in project forecasts. The idiographic model suggested by Kenley & Wilson (1989) is only suitable for post-hoc analysis, as it fits a single S-curve for each project after its completion. In order to be able to forecast, the model applied should utilize historical data that is collected from past projects which leaves systematic error in the model, and at the same time, recognize uniqueness. As an alternative, the individuality of a project can be captured with detailed planning data, and therefore manual labor that is required to obtain it. This has caused a trade-off situation between the amount of manual work that is put into the forecasting and accepting systematic error that is caused by averaging projects.

The models presented in Figure 4 have settled the trade-off in different points. The highest amount of work is required in cost-schedule integrated models, where each cost item in a bill of quantities is associated with a respective activity in the project schedule (Navon, 1995). Many authors suggest that this is an ideal approach, but at the same time acknowledge that it is practically very hard to maintain (Hwee & Tiong, 2002; Banki & Esmaeili, 2009). The cost-schedule integrated approach also requires increased technical complexity in terms of systems integration, in addition to manual

work. This approach utilizes highly detailed information on each project but requires constant labor as project schedules tend to fluctuate and the bill of quantities is often not compatible with scheduled activities (Navon 1995). These obstacles have caused a significant gap between academic research and practice, as cost-schedule integrated models are rarely applied in the industry (Cho et al. 2020).

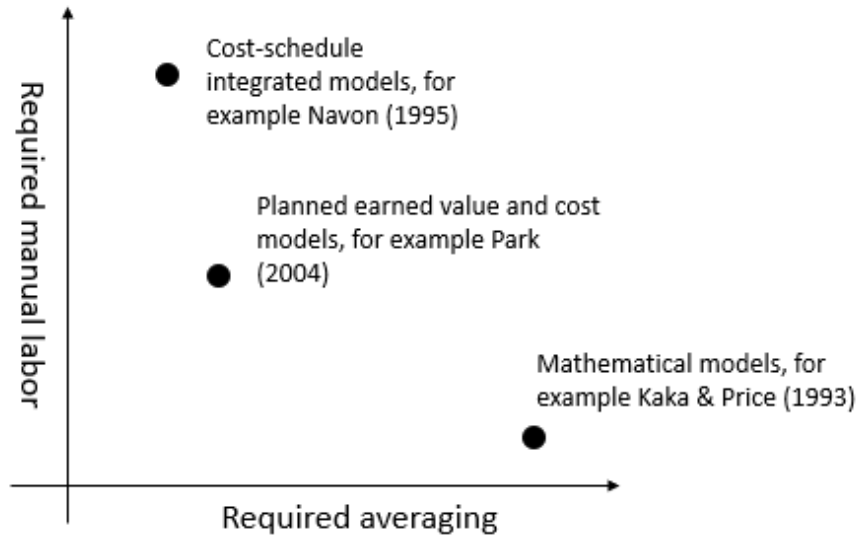


Figure 4. *The trade-off between required manual labor and averaging in different cash flow models.*

Navon (1996) categorizes used cash flow prediction models only into mathematical and cost-schedule integrated models, but later on, a third category has risen which is planned earned value and cost models. Compared to cost-schedule integrated models, Park's (2004) model uses slightly less detailed data as he applies monthly earned value and cost forecasts separately and the costs are represented on category level instead of individual cost items. As the model still follows an individual project's monthly forecast it is able to reflect a specific form of a project and it must average only in terms of cost categories.

Park et al. (2005) recognize an issue in the planned earned value and cost models as they are dependent on the accuracy of monthly planning values which might result in inaccurate cash flow forecasts if the planned values are not accurate. In terms of

required manual labor, monthly forecasts are still very detailed level information and require constant maintenance because of changes in schedules and costs. Additionally, according to Kaka & Price (1993) and Shash's & Qarra's 2018 questionnaire these models are unlikely to be used in the tender phase as detailed planning is rarely done prior to bidding.

Mathematical models distinguish from cost-schedule integrated and planned earned value and cost models because they estimate the shape of the S-curve whereas the last two follow a proposed project plan. This is also what characterizes the research on idiographic and nomothetic approaches. The developments made on cost-schedule integrated and earned value and cost models are focusing on laws that relate to a specific project. For example, Chen et al. (2011) enhance the cost-schedule integrated model by developing a coordination mechanism that accounts for different payment conditions and payment irregularity. Mathematical modeling, on the other hand, has its focus on better estimation techniques and attempts to explore general laws that apply to construction projects.

As a result of their focus, mathematical models require substantially less manual work as they only need general data. For example Kaka's & Price's (1993) cost commitment model needs only the type of the project, size of contract, company (if there are more than one), type of contract and project duration as an input. For the same reason, mathematical models can be applied in practice with fewer difficulties. On the other hand, the estimated shape of the S-curve is solely dependent on past projects. The model uses only project characteristics and project end forecasts and therefore does not reflect unique details in the project schedule. This weakness of mathematical models is substantially less prominent when forecasting cash flow for a project portfolio. For example, Kaka & Price (1993) suggest their model for evaluating company-level cash flow as individual project errors are then consolidated.

Kaka & Price (1993) argue that poor project groupings are one of the key reasons why earlier research into mathematical models has failed to predict accurate S-curves.

Similarly, Boussabaine et al. (1999) claim that the accuracy of mathematical forecasts is dependent on whether the standard curve's conditions represent a forecasted project. As a solution to this problem, Skitmore (1998) suggests utilizing an increased number of variables that represent the characteristics of a project.

The findings of Kaka et al. (2003) highlight the importance of accurate project clustering, as they find that differences caused by averaging a group of projects is causing higher errors than differences arising from actual project planning. They claim that that cost profiles of construction projects are different because of project characteristics instead of people undertaking them. Therefore, in order to reduce the systematic error that is caused by averaging projects by using historical data, projects must be grouped together accurately based on project characteristics. Even if done so, the shapes of S-curves might still differ substantially because construction projects are fundamentally unique.

A common approach has been to classify projects based on their attributes. For example, Kaka & Price (1993) and Evans & Kaka (1998) base their groups on project duration and type, and Chao & Chien (2009) use location and type of work. The findings of Banki & Esmaeeli (2008) indicate that using a homogenous project portfolio results in lower errors compared to earlier research. This supports the common understanding that accurate grouping of projects based on their characteristics is an important contributor in improving mathematical forecasts. This is also supported by findings of Kaka & Price (1993) who report that the difference in average curves of grouped projects is higher than the variability between individual projects within groups.

There are some variables that are known to affect the shape of the S-curve, for example, Ross et al. (2013) find that type of construction, procurement route and type of work will affect the cash flow forecast directly. Skitmore (1992) suggests that a fitting model should use different parameter values for different types of construction and find that the most notable predictors for accurate groupings are contract value, project type

and duration. Similarly, Kaka & Price (1993) find that project duration and type are affecting S-curve shape on a statistically significant level. They provide a further explanation that in short contracts, the costs are piling in the beginning because the work is often started so that resources are already on the site. Similarly for a type of contract, in design-and-build projects, the costs are naturally higher at the start, and on the contrary, management contracts have slow starts since subcontractors are chosen only at the beginning of the project (Kaka & Price, 1993).

3.3 Utilizing cost curve

The S-curve approach has been widely adopted in the literature. Most of the research from the 1970s to the early 1990s has utilized it in a way where the cash in and outflow curves were composed separately from a value curve. The net cash flow would then be the difference between these curves. The use of the value curve was originating from investments in early research by construction clients who wanted to forecast their expenditure flow, and later this approach was applied in contractors' net cash flow forecasting. (Kaka & Price, 1993) However, Kaka & Price (1991) find that value curve models are not sensitive to the choice of value curve but the variability of net cash flow curves are a result of variability in systematic delays of cash-out and cash-in. Therefore, they suggest a model where cash-in and cash-out are separately deduced from value and cost curves, respectively.

As opposed to the earlier approach that has used the value curve as the initial basis for the cash-out and cash-in curves, Kaka & Price (1991) use cost commitment data to obtain value and cost curves. They argue that the cost commitment curve could be estimated more accurately because contractors do different kinds of loading to unbalance the contract. These actions are taken in order to improve the contractor's cash inflow and they include, for example, loading scheduled items that might have large variation and front-end loading the schedule. Similar measures are not taken at the same rate for the cost of items. (Kaka & Price, 1991)

The overall distortion of a value curve can be evidenced by comparing bills of quantities of several contractors for the same contract (Kaka & Price, 1993). In their research on the effects of project planning variability on cost commitment curves, Kaka et al. (2003) conclude that even though different planners' construction programs vary significantly, it does not impact the profile of the cost curve considerably whereas their value curves are most likely different. The hypothesis, that the cost commitment model is more suitable for cash flow prediction than value curve models, is tested by Kaka & Price (1993) and Evans & Kaka (1998) who conclude that the value curve is causing higher errors in estimates and it cannot be fitted for even a specific group of projects, respectively. Altogether, the evidence gives a strong reason to exclude estimation of value curve outside of the mathematical models and focus on the cost curve, or alternatively, some income planning or contract data would be needed to obtain value curve.

In an idiographic approach, in order to obtain cash-out from the cost curve, a respective time lag needs to be introduced as proposed by Kaka & Price (1991). Park (2004) notes that different cost categories generally have different time lags associated with their payment. He suggests that a common budget ratio for general contractors is 50-70% of subcontract costs, 25-35% of material costs, 5-15% of labor costs, 10-25% of equipment costs and 5-15% of indirect costs. Additionally, the budgeted total cost might also include provisions. As the total budget is distributed in various cost categories with highly different time lags the used model should utilize cost categories separately instead of the total cost. For example, equipment costs might only include depreciations with no actual cash flow outflow, subcontractors may have pay-when-paid clauses and employee salaries are booked at the same moment that they are paid.

In the past research, only a few mathematical project cash flow forecasting models have utilized different cost categories (Kaka & Price, 1991; Kaka, 1996). Additionally, these models are distinct from traditional mathematical models as they use an overwhelming number of parameters, for example, Kaka (1996) uses over 50 variables. Kaka & Price (1991) estimate the cost and value curve first and apply

systematic delays to them to obtain cash flow. However, this method has an assumption that project cost composition would be stable throughout the project which is not true as observed by Park (2004). Therefore, they introduce a systematic error into their model by assuming that different costs incur at a uniform rate. Dozens of suggested new mathematical models use only total cost, value or cash flow when it comes to financial data that is used for predictions or S-curve generation (Kaka & Price, 1993; Boussabaine & Kaka, 1998; Boussabaine & Elhag, 1999; Chao & Chien, 2009; Chao & Chien, 2010; Cheng et al., 2011; Cheng & Roy, 2011; Cheng et al. 2012; Chiao et al., 2012; Cristóbal et al. 2015; Cheng et al. 2015; Cheng et al. 2020). Reflecting on his earlier model, Chao (2013) hypothesizes that the S-curve model could be improved if additional input variables would be introduced and they would reflect project conditions.

On the contrary, the research on idiographic models (namely cost-schedule integrated and planning data models) has given significant attention to time lags related to different cost categories and payment conditions. For example, Park (2004), Chen et al. (2005) and Tabyang & Benjaoran (2013) conclude that payment lags are needed in order to predict cash flow accurately. Meanwhile, mathematical models have concentrated on finding more accurate ways to predict project cash flow, the used data has been quite consistent for the last 30 years in the literature, although advanced information technology has enabled recording and using more precise data. Even though nomothetical models are not suited for such sophisticated cost-payment coordination methods as idiographic ones, the research on adjacent subjects suggests that utilizing project cost composition may increase mathematical models' ability to capture project cash outflow profiles more accurately. Therefore, this study looks into the possibility of forecasting the project cash flow curve directly.

3.4 Mathematical methods

All the way to the late 1980s, most of the papers use polynomial regression in estimating the S-curve (Kaka & Price, 1993). Kenley & Wilson (1986) criticize this approach for violating regression model's assumptions and using a large number of constants. As an alternative, they suggest a logit model that utilizes log transformation

and linear regression. Kaka & Price (1991) utilize the proposed logit model and cost commitment data in their cash flow forecasting model that is designed for tender phase predictions. They find a linear equation by logit transformation for dependent and independent variables:

$$Logit = Ln\ \frac{Z}{1-Z} \tag{15}$$

where Logit is the transformation and Z is the variable to be transformed. They express logistic equation for cost commitment flows as:

$$Ln\ \frac{c}{1-c} = \alpha + \beta \times Ln\ [t/(1-t)] \tag{16}$$

where cost (c) is the dependent variable and time(t) is the independent variable. α and β are constants. Cost can be derived (expressed in percentages):

$$c = \frac{100 \times F}{1+F} \quad \text{where} \quad F = e^{\alpha} \times (\frac{t}{100-t})^{\beta} \tag{17}$$

The model suggests that cost values can be approximated using Equation 17. After all the values of t and c are transformed (to X and Y, respectively), the data should approximate a line described by Equation 18, from which parameters α and β which can be derived with linear regression:

$$Y = \alpha + \beta X \tag{18}$$

where $Y = Ln\ \frac{c}{1-c}$ and $X = Ln\ \frac{t}{1-t}$

The logit approach has been one of the most used and accurate models when comparing conventional methods (Skitmore, 1992; Navon, 1996). However, it cannot estimate progress from start to the end, and the common approach has been to exclude 10% from both ends. As it is designed for tender phase predictions, it is not meant to be updated during construction to reflect the actual progress.

When forecasting with standard S-curves that are based on historical data, the results are dependent on how accurately the chosen curve represents an individual project. The problem is difficult especially because it is not clear which variable affects it and

to which degree. Chao & Chien (2009) illustrate this problem by demonstrating that linear correlations between quantitative input and optimized parameters of their model are very weak and therefore these are most likely nonlinear. Similarly, Odeyinka et al. (2012) find indication that the relationships of risk factors affecting cost flow forecasts might be strictly non-linear. Boussabaine & Kaka (1998) conclude that because the relationships are complex and often nonlinear, a regression model might not be the best solution. Hua (2008) suggests exploring AI approaches in quantitative analysis of construction economics as they offer a possibility to take the complexity into account.

A few artificial neural network (ANN) models have been developed to solve this problem. Boussabaine & Kaka (1998) use ANN to predict cumulative costs. Their inputs begin from 10% of project completion until 50% and giving output for the remaining tenths of the project until 90% of completion. The model can be used for tender phase predictions only if the cumulative cost at 10% of completion is an estimated input. Boussabaine et al. (1999) reduce the output of their ANN model to only give three outputs for total cost at 70, 80 and 90% of project completion. Chao's & Chien's (2009) model uses a polynomial function in addition to neural networks to forecast project progress and is the only ANN-based model that is suitable for tender phase predictions.

A similar short-term prediction trend has been consistent also for other methods that are used in mathematical models. All of the support vector machine (SVM) based models (Cheng et al. 2009; Cheng & Roy, 2011; Cheng et al. 2012; Cheng et al. 2013; Cheng et al. 2015), Grey prediction models (Cheng et al. 2011; Cristóbal et al. 2015) and deep learning models (Cheng et al. 2020) are focusing only on short-term predictions even though some of the models could be modified for slightly extended forecasting intervals.

As these models have different prediction intervals it is difficult to compare their accuracy. Hongjiu et al. (2012) compare the performance of artificial intelligence based cash flow prediction methods and found that SVM's performed the best and have the

strongest robustness with small samples. This is especially important in project cash flow prediction because project samples are relatively small.

In their overview of the literature on project control and earned value management, Willems & Vanhoucke (2015) recognize only four papers on SVM. One was applied in forecasting project time and cost based on periodic earned value management data (Wauters & Vanhoucke, 2014) and two of them (Cheng et al. 2010; Cheng & Roy, 2011) focus on short-interval cash outflow forecasting. Wauters & Vanhoucke (2014) conclude that the proposed support vector machine (SVM) method is superior to compared forecasting methods if the used training set resembles the test set. Similarly, Peško et al. (2017) compare artificial neural networks and SVM in estimating project cost and duration. They report that SVM models are more able to generalize input data and thus, providing more accurate estimations for both cost and duration.

A comparison between different times series prediction methods by Sapankevych & Sankar (2009) suggests using ANN's or SVM's in project cash flow prediction as they are not dependent on linear, stationary processes. However, SVM's differ from ANN's as they are guaranteed to converge to the optimal solution (Sapankevych & Sankar, 2009). Overall, SVM's are less researched in construction management and economics area compared to ANN's and none of the existing models utilize SVM for cash flow predictions for the whole project duration. Additionally, there is some evidence that SVM's perform better in the project forecasting area compared to ANN's. Therefore, the existing literature encourages to development of an SVM-based cash flow forecasting model for the whole project duration.

3.5 Summary

Figure 5 summarizes different kinds of cash flow models that are present in the literature. Project level cash flow forecasting divides into two sorts of models: mathematical and planning-based ones. Mathematical models only require general data of the projects and then estimate periodic forecasts with an S-curve based on historical data. Planning-based models on the other hand are dependent on project

schedules and financial forecasts that are generated by project personnel. Therefore, they require detailed level information of the project. This study takes advantage of mathematical models' ability to generate predictions with only a little effort when it comes to the required manual work of the user.

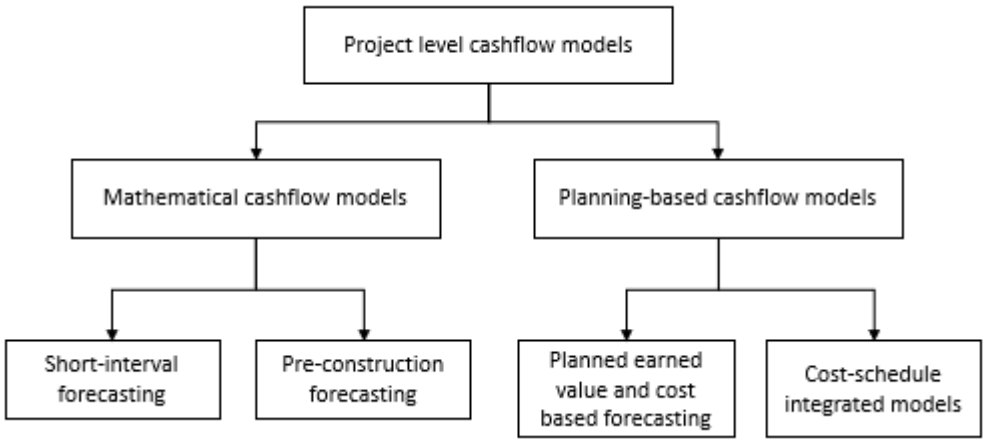


Figure 5. *Project-level cash flow models used in the literature.*

Mathematical models can be roughly split into two groups: short-interval and pre-construction forecasting models. Short-interval forecasting models have made great progress in terms of capability to map complex, non-linear relationships. However, they offer only a little help to contractors as they cannot be used to predict the whole project period. On the contrary, pre-construction models are highly important for construction organizations, but they mainly use logit transformation and linear regression. Chao' & Chien's (2009) progress prediction model makes an exception, but it is not yet applied to cash flow forecasting. Therefore, there is a significant gap in the literature as none of the existing research explores the possibility of forecasting the whole project period using methods that are studied in a short-interval forecasting context.

This study aims to contribute to the literature by introducing a cash flow model that uses support vector regression to predict the whole project period. It can be also updated during construction as opposed to current pre-construction forecasting models. This is not only a nice-to-have feature as one construction project might take

years and contribute a significant proportion of the contractor's revenue. Therefore, reflecting the changes of actuals and project end forecast to the cash flow predictions are also required.

Whereas the research on mathematical forecasting has focused on advances in methodology and few concepts (mainly project grouping and differences in cost and value curves), planning-based models have focused on causal relationships. This has been possible because the used models are more data-intensive compared to mathematical cash flow models that have been stuck on one independent variable (time) in S-curve generation in the pre-construction forecasting area. Short-interval forecasting models have also used cumulative cost or value in addition to time. However, the findings of research on planning-based models suggest that the relative share of different cost categories is greatly affecting the S-curve profile. Additionally, the cash disbursements of different cost categories are not occurring at a uniform rate. Therefore, current state-of-the-art mathematical models that apply fixed time lag to a predicted total cost curve are introducing a systematic error in their predictions. This study aims to investigate the possibility of directly forecasting the cash outflow curve utilizing project cost composition for the above-stated reasons.

4 Data and proposed model

The data is collected from a Finnish general contractor's enterprise resource planning system. Overall, the obtained dataset consists of 33 projects of which 28 and 5 are building and infrastructure construction, respectively. The requirements for selected contracts were that they are valued over one million euros, have cost categorized project end forecasts and at least 80% completed when measuring with an estimated completion date. The earliest contract is started 4.1.2019 and the latest estimated completion date is 30.9.2021.

Table 1. *General statistics of collected projects*

	Minimum	Mean	Maximum
Contract sum (€)	1 111 000	8 062 399	23 848 000
Duration (days)	207	518	905
Gross area (m^2)*	1 927	7016	40 170

* *Reported only for building projects*

General quantitative variables of collected projects are presented in Table 1. The smallest contract value is 1.1M€ and the largest is 23.8M€ with the mean being 8.1M€. The shortest and longest contracts are 207 and 905 days long, respectively. The average duration of a contract is 518 days. The gross area is the sum of floor areas in a building, and it is reported only for building projects. The mean gross area is 518 square meters while the minimum and maximum floor areas are 1 927 and 40 170, respectively. As it can be seen from Table 1, there are large variations in all of the variables which make project clustering necessary in order to get groups of similar projects.

The financials of the projects are collected in monthly periods. Obtained data includes total cash outflow and categorized actual costs and current forecast. Examples of collected actuals and forecasts are presented in Tables 2 and 3, respectively. The first

collected period is the month when construction is estimated to begin. The initial period contains all of the actualized costs before the construction. The last collected month is the month of the estimated project handover date.

Table 2. *Example of collected projects actuals data*

Period	3-2019	4-2019	…	2-2020	3-2020	…	10-2020	11-2020
Cash outflow	**5 886,22 €**	**41 606,30 €**		**3 787 252,41 €**	**4 214 491,48 €**		**7 375 904,51 €**	**7 555 425,67 €**
Labor	- €	3 997,08 €	…	466 252,75 €	515 340,37 €	…	749 890,33 €	750 728,97 €
Materials	- €	1 154,07 €	…	1 385 275,83 €	1 512 177,80 €	…	2 033 757,41 €	2 035 339,62 €
Subcontract	- €	38 820,00 €	…	1 246 309,89 €	1 473 518,32 €	…	3 457 322,30 €	3 665 838,12 €
Others	12 824,22 €	41 865,15 €	…	722 002,94 €	811 177,99 €	…	1 194 677,17 €	1 195 191,66 €
Social cost	- €	2 388,26 €	…	277 274,69 €	305 907,53 €	…	450 891,63 €	451 380,81 €
Add. work	- €	- €	…	9 962,00 €	9 962,00 €	…	12 804,30 €	12 804,30 €
Risk prov.	- €	- €	…	278 000,00 €	447 000,00 €	…	- €	- €
Guarantee prov.	- €	- €	…	- €	- €	…	314 225,89 €	174 191,07 €

The actualized costs consist of eight cost categories. These are labor, materials, subcontracting, other non-calculatory costs, additional works related costs, social costs, deferrals and provisions. The last three are strictly calculatory costs. The project end cost forecasts are reported in eleven different cost categories. Six of them are non-calculatory: labor, materials, subcontracting, other non-calculatory costs, additional works related costs and financing. Five of them are calculatory: social costs, risk provision for a specified item, non-allocated risk provision, guarantee period provision and construction period provision.

Table 3. *Example of project end forecasts*

Period	3-2019	4-2019	…	2-2020	3-2020	…	10-2020	11-2020
Labor	451 720,00 €	451 720,00 €	…	621 913,00 €	643 817,00 €	…	763 680,00 €	763 680,00 €
Materials	2 221 205,00 €	2 221 205,00 €	…	2 519 541,00 €	2 497 885,00 €	…	2 555 777,00 €	2 555 777,00 €
Subcontract	2 959 444,00 €	2 959 444,00 €	…	3 330 389,00 €	3 369 812,00 €	…	3 734 470,00 €	3 734 470,00 €
Others	301 349,00 €	301 349,00 €	…	494 654,00 €	501 754,00 €	…	636 013,00 €	636 013,00 €
Social cost	270 029,50 €	270 029,50 €	…	368 502,49 €	380 551,05 €	…	450 724,82 €	450 724,82 €
Add. work	- €	- €	…	10 670,00 €	13 773,00 €	…	33 292,00 €	33 292,00 €
Spec. risk	18 877,00 €	18 877,00 €	…	2 000,00 €	2 000,00 €	…	2 000,00 €	2 000,00 €
Risk	- €	- €	…	30 580,00 €	33 580,00 €	…	104 634,00 €	104 634,00 €
Finance	31 970,00 €	31 970,00 €	…	32 000,00 €	32 000,00 €	…	- €	- €
Guarantee prov.	- €	- €	…	13 740,00 €	13 740,00 €	…	59 740,00 €	59 740,00 €
Construction prov.	71 740,00 €	71 740,00 €	…	92 000,00 €	92 000,00 €	…	- €	- €

In addition to the data that the proposed model needs, time lags for each non-calculatory cost category are collected in order to test the logit model. As the earliest project starts at the beginning of 2019, payment times for cost categories are collected from years 2017 and 2018 to prevent information leak of time lags to the logit model.

Payment times are represented in a similar monthly fashion as in Kaka & Price (1991) in Table 4. This is because project actuals are measured only at the end of each period. Now that both project cost composition and time lags are collected, an average time lag can be computed. As there are also calculatory costs in the forecasts, the total time lag percent is less than 100 as some costs are not cashed out. For the initial cost composition forecast in Table 3, the average time lag would be 7%, 80% and 7% for months 0,1 and 2, respectively. This is later used to obtain cash outflow from the forecasted cost curve.

Table 4. *Time lags for each non-calculatory cost category*

	No. of months delay		
Cost category	**0**	**1**	**2**
Labor	100 %	0 %	0 %
Materials	0 %	80 %	20 %
Subcontract	0 %	100 %	0 %
Others	7 %	93 %	0 %
Add. Work	13 %	87 %	0 %
Finance	0 %	60 %	40 %

4.1 Cash outflow model

The fact, that numerous variables are affecting a project's financial outcome with non-linear and unknown relationships, would suggest that AI and ML would be more suitable to solve these kinds of problems. There are some successful models that outperform traditional approaches, for example, Cheng & Roy (2011). However, they only predict short intervals and are designed to be used only in the construction phase. In addition, they have only been tested with small and homogenous datasets. This problem is also observed by Kenley & Wilson (1986). They suggest an idiographic approach for project cash flow modeling and point out, that nomothetical models which have been providing good results usually use a narrow set of projects.

Even though it has been noted in the research as early as in the 1970s that different cost categories have different time lags in respect to their cash disbursement (for example Ashley & Teicholz 1977, cited in Park 2004), a standard approach in the mathematical models has been to utilize total costs without categorizing. Chen et al. (2005) find out from their analysis of cost-schedule integrated cash flow models that using payment categories and time lags is necessary in order to obtain accurate cash outflow predictions. Park (2004) applies this idea successfully by introducing an idiographic cash flow model using moving weights of cost categories. However, his model requires monthly cost budget and earned value planning data which are often not available.

Mathematical models can still use these observations to improve their accuracy. The most recognized cash flow model, developed by Kaka & Price (1991), makes an improvement to earlier models by utilizing cost categories and their time lags. However, the cash outflow is derived from a cost curve with fixed weights of cost categories and thus fixed time lag. The observations of Park (2004) and Park et al. (2005) clearly indicate that different cost categories are not incurring at a uniform rate. Therefore, the fixed cost category distribution is not justifiable as it applies the same time lag to incurred costs for the whole project period. This study suggests forecasting the cash outflow curve directly as a solution. This way project cost distribution's

relationship with the rate of cash disbursements can be derived in a more accurate manner.

The suggested cash outflow prediction model is presented in Figure 6. It can be used to obtain cash outflow predictions prior and during construction. It requires three or four descriptive variables for project clustering depending on the type of work. These are the type of work, contract value, contract duration and gross area. In terms of financial data, the model requires cost categorized project end forecast and project time interval. Project actuals that include initial-to-date cash outflow are needed for training the model and for the construction phase predictions.

Project grouping is first performed in two phases. As K-means clustering is applicable only for interval or ratio scale variables, the first phase is to group projects based on different kinds of work. This is introduced as an addition to K-means clustering because multiple studies have found type of work to be a significant predictor of S-curve shape (Skitmore, 1992; Kaka & Price, 1993; Ross et al. 2013). After the initial grouping, K-means clustering is performed for each group based on normalized contract value, contract duration and gross area. However, if the gross area is not reported for a given group, K-means clustering is performed only based on contract value and duration. Normalization is done in order to get all the variables to a common scale (0 to 1).

Contract value and duration are also found to be accurate predictors of project grouping and are commonly used in the literature (Skitmore, 1992; Kaka & Price, 1993; Evans & Kaka, 1998). The gross area of a building project is added to describe the project and control for price differentiation of building in different parts of Finland. It gives further detail on project complexity that contract value and duration might not be able to describe. For example, two similar residential buildings in Southern and Eastern Finland may have highly different contract values because of price differentiation but their similarity can be captured by the gross area of the building.

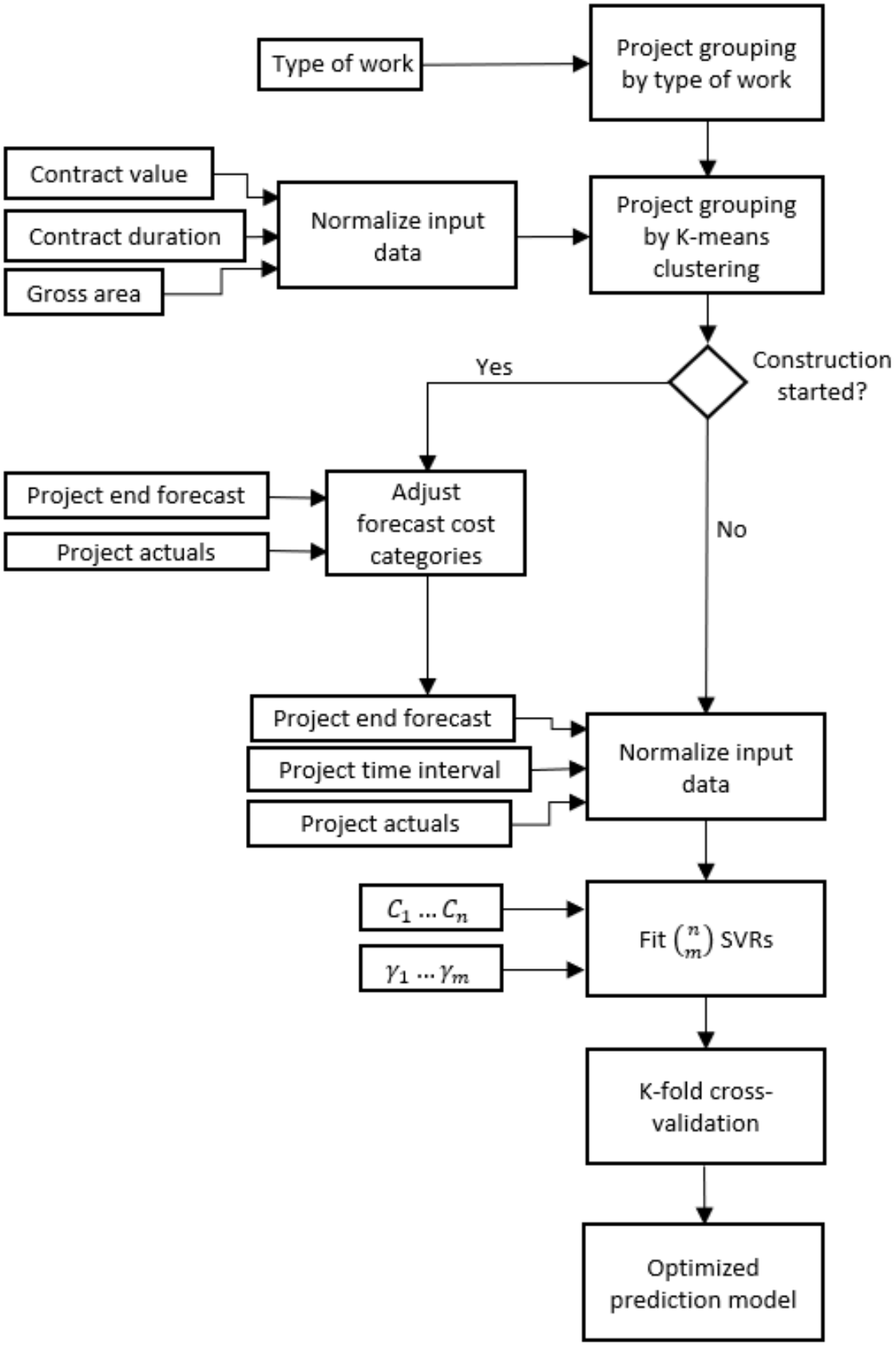


Figure 6. *Proposed cash outflow model using SVR and cost composition*

After project grouping has been performed, input data is processed to be ready for SVR training. If construction has started, project actuals are deducted from project end forecasts so that only non-actualized costs are left of the forecast. Also, the cumulative cost and cash flow that has already been incurred are also added to the input data. The normalization of input data includes transforming cost categories of project end forecast and actuals to the relative share of total forecasted cost and modifying dates of project duration to a percentage of completion in time.

Next, SVR is fitted $\binom{n}{m}$ times as optimal parameters of *C* and γ are explored via grid-search where the model is inputted two lists of different hyperparameters ($C_1 \dots C_n$ and $\gamma_1 \dots \gamma_m$). Then k-fold cross-validation algorithm is performed for each fitted SVR with $k = 5$. After getting average errors of each k-fold cross-validation, the set of hyperparameters with the lowest error is chosen and the model is optimized to perform predictions.

5 Empirical results

Applying the model begins by clustering the projects. First, the dataset is divided into building and infrastructure projects. Next, the input data for clustering is normalized to range from 0 to 1. As there are only four infrastructure projects, K-means clustering is not performed for their group. For building projects, the appropriate number of clusters is first estimated. A visualization of the Elbow Method can be found in Appendix 1. Based on the results from Elbow Method, the projects are clustered into three different groups based on contract value, duration and gross area. Table 5 represent cluster centers and the number of projects for each cluster.

Table 5. *Cluster centers for building and infrastructure projects.*

Cluster	No. projects	Contract value	Duration	Gross area
Building 1	18	5 880 020 €	453	4226
Building 2	10	15 335 954 €	728	8020
Building 3	1	5 604 564 €	323	40170
Infrastructure	5	1 863 424 €	453	-

Two of the three clusters form sensible groups. The ten projects that are assigned to cluster 2 can be categorized as large based on all of their attributes, whereas the projects of cluster 1 are relatively small. However, one project is forming its own cluster. Its contract value and duration are close to cluster 2 but its gross area is tenfold. This indicates its unique project type and therefore it is left out of forecasting. Infrastructure projects form their own cluster. Its mean contract value is distinctively smaller compared to building projects and its average duration is the same with small building projects.

In summary, after the two-phased project grouping, there are three separate groups to analyze: infrastructure projects, large building projects and small building projects that consist of 4, 10 and 18 projects, respectively. As the projects are now grouped, next follows the training of the prediction model. From each group, 80% of the projects are used for training and 20% for testing the optimized model. The projects are split into

training and test sets randomly. This results in 3, 8 and 14 training projects and 1, 2 and 4 test projects for infrastructure, large building and small building projects, respectively. In total, there are 443 and 115 observations in the training and testing datasets. Next, the empirical section divides into two parts as does the model. First, pre-construction predictions are represented which follows with the construction phase predictions.

5.1 Pre-construction forecasting

As multiple changes are proposed compared to the logit cost commitment curve model, the transition to the introduced model is done in multiple phases. The suggested phases are:

1) Forecasting standard cash outflow curve directly instead of obtaining it from cost commitment curve with time lags.
2) Using support vector regression as opposed to logit transformation and linear regression.
3) Free hyperparameters of support vector regression are optimized with grid search and k-fold cross-validation.
4) As an alternative to obtaining a fixed time lag, project cost composition is used directly in S-curve prediction.

First, the logit model is used to get baseline forecasts. This is done by forecasting standard cost commitment curves and then applying fixed time lag that is determined by the project's cost composition and time lags presented in Table 4 (Logit_CCC). Second, to answer the first sub-question of research question one, logit transformation and linear regression are applied to forecast the cash outflow curve directly (Logit_COC). Data exclusion range is 10% from both ends for both of the logit models. This results that for logit models there are only 359 and 88 training and test observations.

The cash outflow curve is then estimated with support vector regression with default and optimized hyperparameters, SVR and SVR_OPT, respectively. Used default parameters for SVR are $C = 1$ and for gamma as in Equation 19.

$$\gamma = \frac{1}{n_f \, \sigma^2(X)} \tag{19}$$

where n_f is the number of features and $\sigma^2(X)$ is the variance of predictive variables. Grid-search ranges for C and γ are 2^{-5} ... 2^{15} and 2^{-15} ... 2^{5}, respectively. After this, the input data is enriched with project cost composition and the cash outflow curve is predicted using SVR with default (SVR_CC) and optimized hyperparameters (SVR_CC_OPT). The error of different methodologies is measured with root mean squared error (RMSE) that is presented in Equation 20.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{20}$$

where y_i is the actual cash outflow, $\hat{y}_i$ is predicted value and n is the number of observations.

The training results for different models are presented in Table 6 which contains RMSE for all the training projects, all data points in the given cluster and all data points in the entire training set. The standard deviation of errors (STD) is also presented for the last two. Unsurprisingly, the cost commitment curve logit model has the highest training RMSE of 10.6% for all predictions as it is not directly fitted to actual cash outflow. The cash outflow curve logit model proves this point as it lowers the training RMSE of all observations to 9.83%. However, SVR with default and optimized hyperparameters are better able to fit the training data with respective errors of 9.25% and 8.85%.

After the training data has been enriched with project cost composition, SVR has a lot more variables to base its predictions on. This can be seen with significantly lower error rates with both default and optimized SVR. The most notable drop can be observed in infrastructure project predictions and in projects that could be considered as outliers, namely, projects 3, 7, 14 and 21. As the RMSE drops from 7.7% (SVR_CC) to 6.27% (SVR_CC_OPT) by optimizing, its effect is huge compared to SVR with the time as the only predictive variable (SVR and SVR_OPT). These observations can also indicate from overfitting.

Table 6. *Pre-construction training RMSE and STD of different models.*

	TRAINING RMSE & STD					
	Logit_CCC	Logit_COC	SVR	SVR_OPT	SVR_CC	SVR_CC_OPT
Small building projects						
Project 1	3,47 %	7,51 %	9,84 %	9,40 %	10,44 %	6,92 %
Project 2	15,37 %	10,93 %	6,24 %	6,98 %	5,54 %	6,02 %
Project 3	11,35 %	22,11 %	23,17 %	22,66 %	16,92 %	7,64 %
Project 4	4,70 %	7,04 %	6,91 %	6,37 %	5,18 %	5,64 %
Project 5	15,00 %	13,08 %	8,89 %	9,13 %	7,92 %	7,73 %
Project 6	12,37 %	8,95 %	6,19 %	6,35 %	5,62 %	6,25 %
Project 7	13,91 %	11,51 %	13,87 %	13,73 %	15,12 %	8,34 %
Project 8	3,03 %	7,55 %	9,57 %	9,01 %	8,53 %	7,19 %
Project 9	9,76 %	6,83 %	3,69 %	4,17 %	5,13 %	6,47 %
Project 10	2,65 %	8,60 %	10,82 %	10,43 %	7,48 %	6,36 %
Project 11	9,98 %	7,20 %	6,80 %	6,35 %	6,81 %	6,45 %
Project 12	10,76 %	7,13 %	3,87 %	4,20 %	4,11 %	6,08 %
Project 13	8,32 %	6,23 %	4,62 %	4,09 %	5,28 %	3,70 %
Project 14	18,89 %	15,04 %	11,09 %	11,54 %	9,30 %	6,66 %
RMSE	**11,50 %**	**10,67 %**	**9,82 %**	**9,71 %**	**8,90 %**	**6,68 %**
STD	**8,87 %**	**10,49 %**	**9,80 %**	**9,74 %**	**8,88 %**	**6,62 %**
Large building projects						
Project 15	7,56 %	3,91 %	5,72 %	3,85 %	6,40 %	5,78 %
Project 16	13,88 %	8,65 %	7,26 %	6,28 %	5,63 %	5,59 %
Project 17	8,06 %	3,77 %	7,80 %	7,22 %	7,52 %	7,38 %
Project 18	4,83 %	8,02 %	7,34 %	6,95 %	6,89 %	6,41 %
Project 19	8,35 %	4,45 %	3,93 %	1,33 %	3,16 %	2,13 %
Project 20	3,17 %	5,91 %	6,25 %	5,20 %	6,22 %	5,08 %
Project 21	16,50 %	13,13 %	10,21 %	9,56 %	7,58 %	7,11 %
Project 22	3,29 %	6,50 %	8,71 %	7,57 %	7,56 %	5,14 %
RMSE	**9,43 %**	**7,62 %**	**7,48 %**	**6,55 %**	**6,58 %**	**5,84 %**
STD	**7,69 %**	**7,55 %**	**7,22 %**	**6,56 %**	**6,06 %**	**5,85 %**
Infrastructure projects						
Project 23	18,02 %	5,51 %	8,53 %	7,95 %	6,02 %	6,64 %
Project 24	4,86 %	8,37 %	9,44 %	8,41 %	3,74 %	4,37 %
Project 25	20,14 %	19,36 %	14,34 %	14,58 %	7,97 %	6,04 %
Project 26	5,57 %	12,85 %	14,70 %	14,04 %	7,97 %	7,73 %
RMSE	**14,41 %**	**13,55 %**	**12,19 %**	**11,87 %**	**6,58 %**	**5,99 %**
STD	**12,58 %**	**13,55 %**	**12,19 %**	**11,99 %**	**6,58 %**	**5,99 %**
RMSE (all)	**10,60 %**	**9,83 %**	**9,25 %**	**8,85 %**	**7,77 %**	**6,27 %**
STD (all)	**8,56 %**	**9,70 %**	**9,14 %**	**8,86 %**	**7,66 %**	**6,27 %**

The standard deviation of errors is tightly following the RMSE value with all the other methodologies except for Logit_CCC as the standard deviations of Logit_COC, SVR, SVR_OPT, SVR_CC and SVR_CC_OPT are 9.7%, 9.14%, 8.86%, 7.66% and 6.27%,

respectively. For Logit_CCC standard deviation is significantly lower than RMSE in all the project clusters with a standard deviation of all errors being 8.56%.

Table 7. *Pre-construction test RMSE and STD of different models.*

	TEST RMSE & STD					
	Logit_CCC	**Logit_COC**	**SVR**	**SVR_OPT**	**SVR_CC**	**SVR_CC_OPT**
Small building projects						
Project 1	6,10 %	9,61 %	11,31 %	10,75 %	11,27 %	9,64 %
Project 2	8,20 %	6,20 %	4,49 %	4,02 %	3,78 %	4,70 %
Project 3	2,87 %	7,26 %	8,04 %	7,80 %	9,03 %	7,98 %
Project 4	5,01 %	4,00 %	8,33 %	6,86 %	4,99 %	11,29 %
RMSE	**5,78 %**	**7,59 %**	**8,78 %**	**8,19 %**	**8,53 %**	**8,70 %**
STD	**5,85 %**	**5,65 %**	**5,70 %**	**5,60 %**	**6,44 %**	**7,90 %**
Large building projects						
Project 5	10,15 %	8,53 %	7,66 %	7,43 %	8,53 %	9,64 %
Project 6	7,61 %	5,56 %	3,94 %	2,39 %	3,24 %	2,17 %
RMSE	**8,78 %**	**6,98 %**	**5,74 %**	**5,07 %**	**5,98 %**	**6,36 %**
STD	**4,16 %**	**6,97 %**	**5,26 %**	**5,11 %**	**4,92 %**	**6,03 %**
Infrastructure projects						
Project 7	20,94 %	18,92 %	15,43 %	16,18 %	17,91 %	11,23 %
STD	**11,23 %**	**11,65 %**	**11,84 %**	**11,12 %**	**10,58 %**	**8,93 %**
RMSE (all)	**10,25 %**	**9,76 %**	**9,05 %**	**8,80 %**	**9,57 %**	**8,33 %**
STD (all)	**8,73 %**	**9,81 %**	**8,60 %**	**8,68 %**	**9,35 %**	**8,17 %**

The test results of different models are presented in Table 7 which contains RMSE for all the test projects, all data points in the given project cluster and all data points in the whole test set. The standard deviation of errors (STD) is also presented for the former two. Overall error rates of Logit_CCC, Logit_COC, SVR, SVR_OPT, SVR_CC and SVR_CC_OPT are 10.25%, 9.76%, 9.05%, 8.8%, 9.57% and 8.33%, respectively. Measuring with RMSE, the rank order of the models is the same as in the training set with the exception of SVR_CC that is defeated by SVR and SVR_OPT. Therefore, SVR_CC_OPT and SVR_OPT models have the highest performance in the respective order.

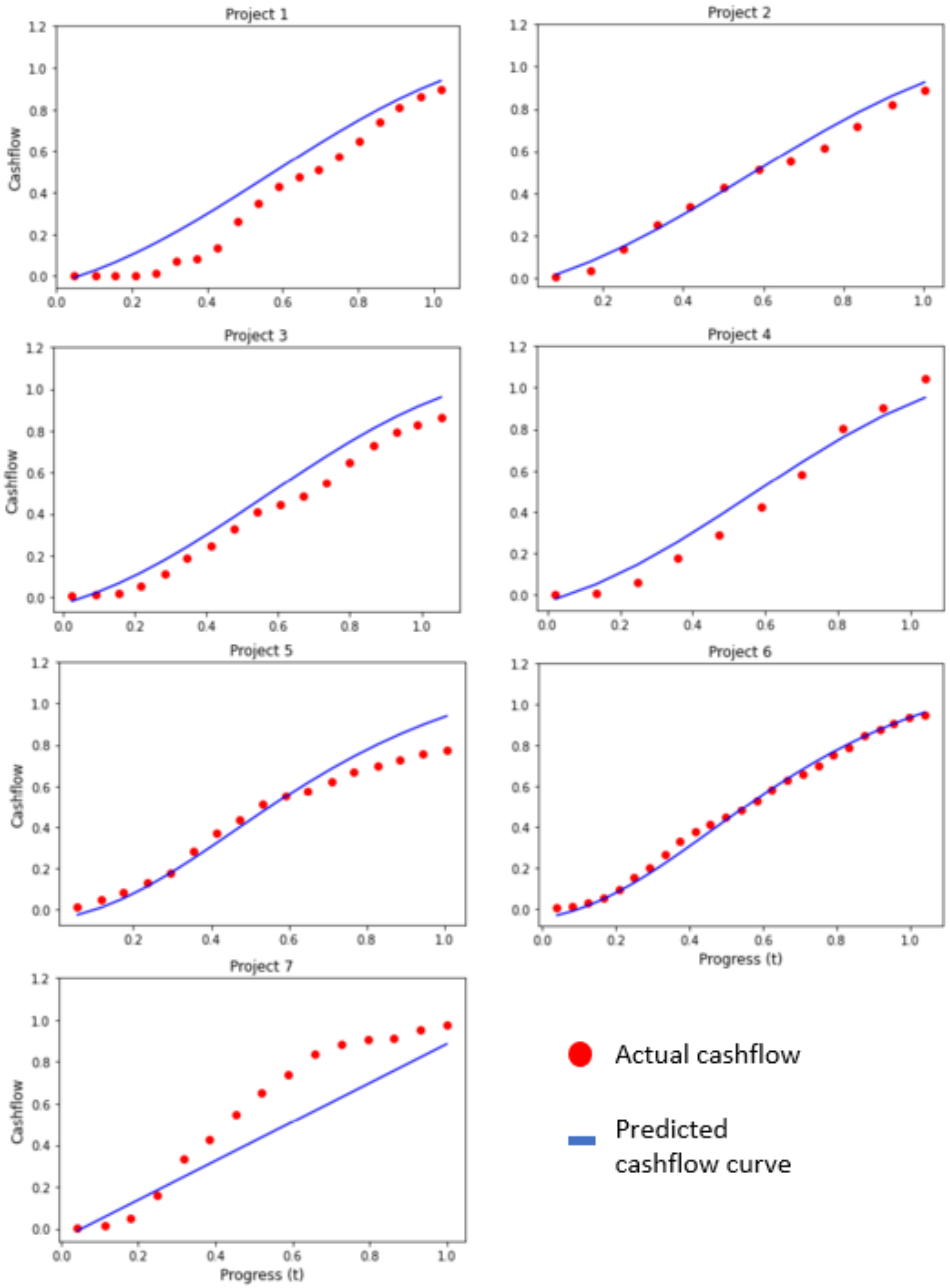


Figure 7. *Predictions of optimized support vector regression (SVR_OPT) and the actual cash flow for test projects in the pre-construction phase.*

The predicted cash outflow curves of SVR_OPT can be found in Figure 7. It is noteworthy that SVR_OPT is performing better with projects whose cash outflow curve is following a predictable pattern and therefore have low error rates across the different models. This can be seen in small and large building project groups which have lower RMSE values in both training and test sets compared to infrastructure projects. Even though SVR_OPT predictions have the lowest RMSE only for one of seven projects

(project 5), it stays consistently in low- and mid-level error rates. As an example of opposite behavior, Logit_CCC generates overwhelmingly best predictions for small building projects (RMSE 5.78%) but performs the worst in large building and infrastructure project categories (RMSEs 8.78% and 20.94%). This indicates that the model is able to forecast only the common and gentle S-curve pattern that resembles a linear line (projects 1,2,3, 4 and 6) but fails terribly with abrupt S-curves (projects 5 and 7). Therefore, there is not enough steepness in the S-curves of Logit_CCC which is most likely caused by linear methodology.

However, it can be seen from both the RMSE values and Figure 7 that a standard curve with one predictive variable does not conform very well to abnormal cash outflow curves of projects 1 and 7 even though the methodology is non-linear. As a matter of fact, SVR_OPT cannot find any predictable pattern in infrastructure projects and settles for predicting a linear line for project 7. The results imply that hyperparameter-optimized SVR is able to predict a well-generalized cash flow curve. Interestingly, it generates nearly identical curves for both small and large building projects even though the optimized hyperparameters are different.

The predicted cash outflow curves of SVR_CC_OPT are presented in Figure 8. As opposed to SVR_OPT, predicted curves are noticeably different between and within project groups. The benefit of using SVR_CC_OPT can be seen from the predictions of projects 1 and 7 where the other models have produced significantly less accurate predictions except logit models for project 1. However, the predictions generated by SVR_CC_OPT can be classified even as bad for projects 4 and 5 as their RMSEs are, respectively, 4.43% and 2.21% higher compared to SVR_OPT and generally higher than in other models. The results imply that SVR_CC_OPT is able to conform to different kinds of cash flow profiles but with a cost of lower performance with the common patterns.

It can be seen from Figures 7 and 8 that some of the predictions are negative at the beginning of the project as optimized SVR has not been able to fit to the training data

without crossing the x-axis. This issue sticks out especially in SVR_CC_OPT's prediction at the beginning of project 4, where the prediction is almost -24% of the forecasted total cost. As there is no reason to assume that there would be any credit memos or cost compensations cashed in without corresponding cost right at the beginning of the construction, it would be reasonable to limit the predictions to zero.

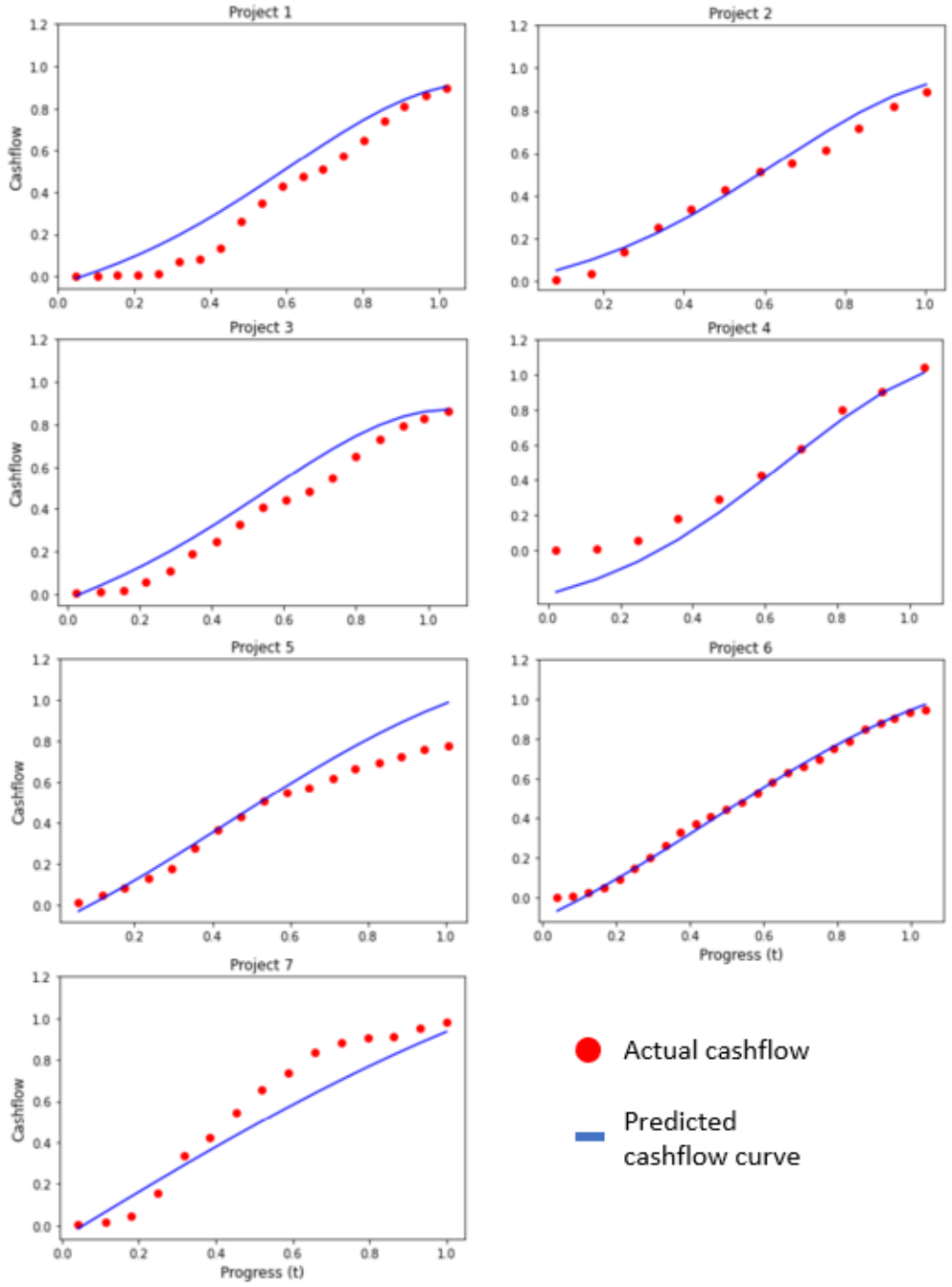


Figure 8. *Predictions of optimized support vector regression using project cost composition (SVR_CC_OPT) and the actual cash flow for test projects in the pre-construction phase.*

With this adjustment, the overall RMSE of SVR_OPT and SVR_CC_OPT drop to 8.78% and 7.75%, respectively. Therefore, there is only a minuscule improvement in SVR_OPT but the RMSE of SVR_CC_OPT improves by 0.58%, making the gap of over 1% to SVR_OPT. The predictions generated by SVR_CC_OPT with the abovementioned limitation are presented in Appendix 2 and RMSE values for small building, large building and infrastructure projects are 7.65%, 6.22% and 11.22%, respectively.

Errors with the abovementioned limitation are presented in Figure 9 for all of the models. The mean errors for Logit_CCC. Logit_COC, SVR, SVR_OPT, SVR_CC and SVR_CC_OPT are 5.46%, 0.05%, -2.92%, -1.75%, -2.23% and -2.33%. The peaks of error distributions are around zero only for Logit_CCC, SVR_OPT and SVR_CC_OPT. However, the rest of the Logit_CCC distribution is located mostly at the positive side of the x-axis which explains the high RMSE and relatively low STD. This also results in an exceptionally high average error which implies a high systematic error. When the logit model is shifted to forecast cash outflow curve, the errors are quite balanced with an extremely low average error of Logit_COC. On the contrary to Logit_CCC, the error distributions of SVR, SVR_OPT, SVR_CC are SVR_CC_OPT positively skewed. When comparing the profile of the distributions, SVR_CC_OPT is the closest to the normal distribution and has a shorter right tail compared to other models.

The abovementioned observations of error distributions imply that a lower standard deviation of errors is a positive trait at least for SVR_OPT and SVR_CC_OPT as their errors are distributed around zero. As presented in Table 7, the standard deviation of errors for SVR_CC_OPT is the lowest (8.17%). After restricting the predictions to be zero or above, the standard deviation of errors drops to 7.42% for SVR_CC_OPT and 8.64% for SVR_OPT. For other models, there is no effect. These observations would suggest that SVR_CC_OPT can be considered the best performing model also in terms of error distribution as it has only a small disadvantage in mean error to SVR_OPT but is better in terms of other traits. This especially is important in the context of mathematical forecasting.

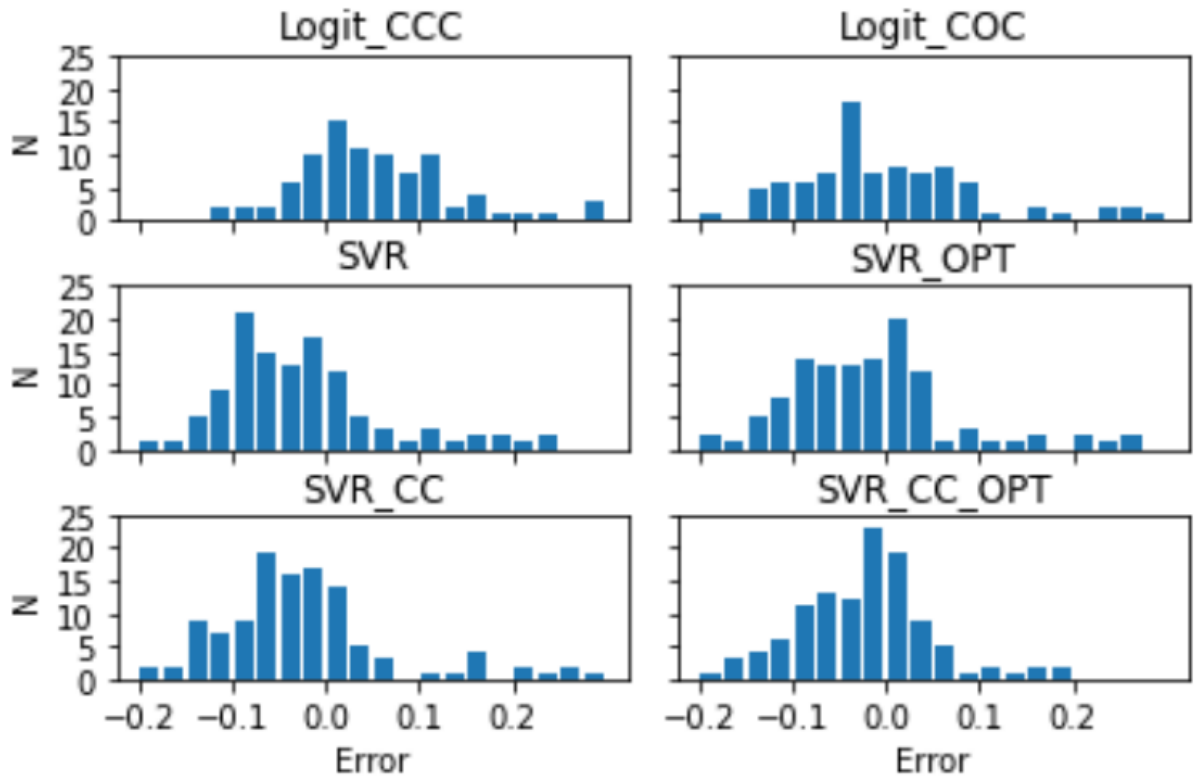


Figure 9. *Test error distributions of different models in the pre-construction phase.*

Table 8 represents the actual payment times for non-calculatory cost categories from the observation period. When comparing these with the ones from Table 4, it can be noted that all except labor and finance have increased. For these two, labor has stayed the same and finance costs have been paid in the same month instead of 60% and 40% in the second and third month, respectively. However, the financing costs are very marginal in the overall budget. On the other hand, the largest cost categories, materials and subcontracting, have had longer payment times compared to the used ones. In addition, other miscellaneous costs had longer payment lags.

Reflecting the comparison between Tables 4 and 8 to the error distribution of Logit_CCC in Figure 9, the performance of Logit_CCC would have been worse if more accurate payment lags would have been used. This is because a majority of its errors are positive, meaning that the predictions have been smaller than the actual cash outflow. Simultaneously the actual payment lags have been longer when the differences of Tables 4 and 8 are weighted with the relative sizes of the cost categories.

Therefore, the applied delay to the cost commitment curve would have been even larger.

Table 8. *Actual payment times of non-calculatory cost categories in the observation period.*

	No. of months delay		
Cost category	**0**	**1**	**2**
Labor	100 %	0 %	0 %
Materials	0 %	70 %	30 %
Subcontract	0 %	97 %	3 %
Others	3 %	97 %	0 %
Add. Work	7 %	93 %	0 %
Finance	100 %	0 %	0 %

5.2 The construction phase forecasting

As the logit model is designed to be used with only one variable and not to be updated, this section observes only SVR, SVR_OPT, SVR_CC and SVR_CC_OPT models. All of them are extended with two variables: progress at the time of prediction and normalized total cash outflow that is actualized in the time of prediction. Also, project end forecast changes are updated to the construction phase data which will affect all of the models through the normalization denominator. The weights of different cost categories are also adjusted with respect to actualized costs for SVR_CC and SVR_CC_OPT models.

As there may be plenty of periods before the period to be predicted, multiple instances of the same period are generated. These all have different values for progress and actualized cash flow at the time of prediction in addition to the remaining weights of cost categories for SVR_CC and SVR_CC_OPT. In the training data, there are as many instances of the same period as there are periods that have progress smaller or equal to 50% before it. This results that in total, there are 1154, 1608 and 244 training observations for small building, large building and infrastructure projects, respectively.

This is implemented in order to avoid the models from overfitting and give them better generalization abilities.

Training and test results are presented with 10% increments in respect of progress in time from 10% to 50%. As the financials are collected monthly, the data is not available for these exact points of time. Because of this, the closest month before the given decimal is chosen. Forecasts are then generated for the rest of the project duration with this input.

Results for training projects are presented in Table 9. At the end of the table, RMSE and standard deviation of errors are also presented for all the of the datapoints as the model is trained with all of them. When measuring with all the observations and RMSE, the order from the fittest to the worst fitting is SVR_CC_OPT, SVR_CC, SVR_OPT and SVR with respective RMSEs of 5,68%, 5,94%, 6,98, 7,15%. In relative means, the results are quite the same as in pre-construction training but the difference between SVR_CC and SVR_CC_OPT is not as large. Also, the overall difference in training RMSE between models is not as large which can be explained by increasing the predictive variables of SVR and SVR_OPT from one to three.

The training RMSE for all the data points in the construction phase is smaller than in the pre-construction phase for all of the models. This can be explained by enriching the predictive variables with data on project progress. The difference of overall training RMSE between pre-construction and the construction phase is 2.1%, 1.87%, 1.83% and 0.59% for SVR, SVR_OPT, SVR_CC and SVR_CC_OPT, respectively. Therefore, there is a large difference between the latter three and SVR_CC_OPT in respect of improving the training RMSE.

As mentioned above, the decrease in RMSE for SVR and SVR_OPT could be explained by increasing the predictive variables which will give them the ability to adjust for different kinds of cash flow profiles. However, the drop in RMSE for SVR_CC is almost at the same level even though it had this ability also in the pre-construction

phase. If the decrease would be caused by adjusted weights of cost categories, it would also drop the RMSE of SVR_CC_OPT with the same magnitude which is not the case. There are also large differences between the hyperparameters of SVR_CC and SVR_CC_OPT as their respective C's and γ's for different project groups are 1 and 1.633, 1 and 1.546, 1 and 1.23 for SVR_CC and 2.177 and 1.043, 27.779 and 0.486, 11.888 and 0.176 for SVR_CC_OPT. This could indicate that a good training fit can be achieved with various hyperparameters as more detailed data from project progress is available but the k-fold cross-validation algorithm is toning down the overfitting of SVR_CC_OPT which keeps the training RMSE relatively high. This would mean that generalization of the construction phase cost category data is a more difficult task compared to the pre-construction phase.

When looking at the RMSE of the models in different points of progress, there is a clear distinction between the models with and without project cost composition. SVR_CC and SVR_CC_OPT begin with 1.89-2.34% lower RMSE and they continue to improve till the progress is over 30%. However, after this both of their RMSE start rising rapidly so that when project progress is less than 50% their RMSE is higher than in the beginning. This could be caused by project cost composition not having such a large effect on the cash outflow profile after the construction is in full motion which is around the third of project duration. Training these models with all of the training data still appears to give them better generalization ability, because when experimenting with training data that has only the datapoints with prediction time right before 50% progress, test RMSE increases by 1.4%.

On the contrary, SVR and SVR_OPT models begin with quite high training RMSE, but which is still lower than the pre-construction phase training error and they continue to improve training RMSE for the whole observed interval. This could indicate that SVR and SVR_OPT are utilizing their input more efficiently but they have less information to begin with it as they have a lower number of variables. However, as they have slightly higher training errors compared to SVR_CC and SVR_CC_OPT even when progress is under 40%, their overall RMSE is significantly higher.

Table 9. *The construction phase training RMSE and STD of different models.*

TRAINING RMSE & STD				
	SVR	**SVR_OPT**	**SVR_CC**	**SVR_CC_OPT**
Progress < 10%				
Small building projects	8,33 %	8,53 %	6,93 %	6,90 %
Large building projects	7,24 %	7,23 %	6,19 %	5,57 %
Infrastructure projects	12,20 %	12,29 %	6,14 %	5,20 %
RMSE	**8,44 %**	**8,54 %**	**6,55 %**	**6,20 %**
STD	**8,42 %**	**8,53 %**	**6,54 %**	**6,21 %**
Progress < 20%				
Small building projects	7,86 %	7,87 %	6,35 %	6,32 %
Large building projects	7,56 %	7,34 %	5,61 %	5,36 %
Infrastructure projects	7,15 %	6,52 %	6,76 %	5,96 %
RMSE	**7,66 %**	**7,51 %**	**6,11 %**	**5,90 %**
STD	**7,63 %**	**7,48 %**	**6,09 %**	**5,86 %**
Progress < 30%				
Small building projects	6,20 %	6,13 %	4,89 %	4,93 %
Large building projects	7,46 %	7,16 %	5,46 %	5,30 %
Infrastructure projects	6,28 %	6,09 %	7,38 %	6,05 %
RMSE	**6,75 %**	**6,56 %**	**5,47 %**	**5,23 %**
STD	**6,73 %**	**6,52 %**	**5,46 %**	**5,12 %**
Progress < 40%				
Small building projects	5,99 %	5,88 %	6,15 %	6,04 %
Large building projects	6,83 %	6,40 %	5,69 %	5,50 %
Infrastructure projects	6,07 %	6,20 %	7,99 %	6,25 %
RMSE	**6,35 %**	**6,13 %**	**6,20 %**	**5,85 %**
STD	**6,36 %**	**6,08 %**	**6,19 %**	**5,86 %**
Progress < 50%				
Small building projects	4,93 %	4,23 %	8,01 %	7,73 %
Large building projects	6,12 %	5,64 %	6,35 %	5,20 %
Infrastructure projects	6,59 %	5,79 %	8,38 %	6,82 %
RMSE	**5,63 %**	**5,02 %**	**7,46 %**	**6,74 %**
STD	**5,61 %**	**5,01 %**	**6,70 %**	**6,45 %**
All				
RMSE	**7,15 %**	**6,98 %**	**5,94 %**	**5,68 %**
STD	**7,14 %**	**6,98 %**	**5,93 %**	**5,67 %**

Test results of different models are presented in 10% progress increments in Tables 10 and 11. When comparing the overall RMSE in different levels of progress the best performing models are SVR, SVR_CC_OPT, SVR_OPT at 10%, 20-30% and 40-50% of progress, respectively. The best RMSE for each point of progress is 7.82%, 8.19%,

7.84%, 6.79% and 6.08%. SVR_CC sticks out from the other models as it has the worst performance in each point of progress.

Combining predictions of all five periods, the RMSEs for SVR, SVR_OPT, SVR_CC and SVR_CC_OPT are 7.78%, 7.80%, 8.61% and 7.81%, respectively. Therefore, although SVR, SVR_OPT and SVR_CC_OPT are within a 0.03% range, SVR has overall the lowest RMSE. When comparing SVR to SVR_OPT, it is better when progress is in the 10-20% range after which SVR_OPT performs better. As there are more periods to predict at the beginning of a project, there are also more observations then, which weighs just enough that SVR has the lowest total RMSE.

Distinctively, the pre-construction phase RMSE of 7.75% (when there is a lower bound of zero for predictions) of SVR_CC_OPT is outperformed only until the progress is at 40%. When looking at the overall RMSE of all five prediction periods, none of the models beat the RMSE of SVR_CC_OPT in the pre-construction phase. However, after the threshold is exceeded at 40% all of the models surpass it by far as their total RMSE then ranges from 6.79% - 7.15%. When looking at the development of RMSE by the model, SVR and SVR_OPT both are decreasing their RMSE from pre-construction to 50% of the project duration with the exception of higher errors when progress is 20%. Compared to the pre-construction phase, SVR_CC and SVR_CC_OPT, on the other hand, both start with higher RMSE’s at 10% of progress and they gradually lower their error until surpassing the pre-construction level at 30% and 40% of progress, respectively.

Table 10. *RMSE and STD values for test projects with progress from <10% to <30%.*

	TEST RMSE & STD			
	SVR	**SVR_OPT**	**SVR_CC**	**SVR_CC_OPT**
Progress < 10%				
Small building projects				
Project 1	10,69 %	11,10 %	9,99 %	10,22 %
Project 2	3,67 %	3,60 %	6,41 %	6,29 %
Project 3	6,39 %	6,36 %	8,51 %	8,56 %
Project 4	9,35 %	9,74 %	8,11 %	7,73 %
RMSE	**8,25 %**	**8,49 %**	**8,61 %**	**8,63 %**
STD	**7,19 %**	**7,40 %**	**8,03 %**	**8,02 %**
Large building projects				
Project 5	3,32 %	4,53 %	5,62 %	8,15 %
Project 6	3,31 %	3,65 %	3,50 %	3,61 %
RMSE	**3,31 %**	**4,04 %**	**4,49 %**	**5,91 %**
STD	**3,25 %**	**3,79 %**	**3,09 %**	**5,24 %**
Large building projects				
Project 7	13,22 %	13,20 %	18,30 %	12,87 %
STD	**10,48 %**	**10,50 %**	**11,53 %**	**10,28 %**
All	**7,82 %**	**8,06 %**	**9,42 %**	**8,48 %**
STD	**7,75 %**	**7,94 %**	**9,41 %**	**8,36 %**
Progress < 20%				
Small building projects				
Project 1	8,56 %	8,55 %	7,89 %	8,07 %
Project 2	3,70 %	3,65 %	7,64 %	7,68 %
Project 3	4,48 %	4,40 %	6,61 %	6,57 %
Project 4	9,45 %	9,53 %	9,76 %	10,01 %
RMSE	**6,93 %**	**6,92 %**	**7,85 %**	**7,96 %**
STD	**6,86 %**	**6,88 %**	**7,89 %**	**8,00 %**
Large building projects				
Project 5	3,60 %	5,09 %	6,32 %	8,93 %
Project 6	3,14 %	3,98 %	2,81 %	3,49 %
RMSE	**3,33 %**	**4,45 %**	**4,55 %**	**6,26 %**
STD	**3,14 %**	**4,21 %**	**3,50 %**	**5,24 %**
Large building projects				
Project 7	19,23 %	19,74 %	19,31 %	12,80 %
STD	**8,01 %**	**7,53 %**	**8,00 %**	**7,93 %**
All	**8,67 %**	**8,98 %**	**9,26 %**	**8,19 %**
STD	**8,64 %**	**8,97 %**	**9,28 %**	**8,23 %**
Progress < 30%				
Small building projects				
Project 1	7,79 %	7,68 %	6,22 %	6,40 %
Project 2	4,34 %	4,16 %	3,71 %	3,75 %
Project 3	4,11 %	4,04 %	8,10 %	8,40 %
Project 4	10,14 %	9,90 %	10,47 %	10,70 %
RMSE	**6,80 %**	**6,66 %**	**7,27 %**	**7,48 %**
STD	**6,80 %**	**6,72 %**	**7,34 %**	**7,56 %**
Large building projects				
Project 5	3,11 %	5,26 %	6,12 %	9,42 %
Project 6	4,12 %	4,23 %	1,82 %	2,91 %
RMSE	**3,75 %**	**4,67 %**	**4,12 %**	**6,37 %**
STD	**3,21 %**	**3,83 %**	**3,77 %**	**5,04 %**
Large building projects				
Project 7	16,89 %	15,75 %	18,35 %	11,80 %
STD	**6,37 %**	**5,98 %**	**5,59 %**	**5,98 %**
All	**8,14 %**	**7,95 %**	**8,80 %**	**7,84 %**
STD	**7,97 %**	**7,90 %**	**8,73 %**	**7,88 %**

Table 11. *RMSE and STD values for test projects with progress from <40% to <50%.*

	SVR	SVR_OPT	SVR_CC	SVR_CC_OPT
Progress < 40%				
Small building projects				
Project 1	8,53 %	8,68 %	6,07 %	6,04 %
Project 2	5,18 %	5,47 %	3,52 %	3,11 %
Project 3	4,86 %	4,78 %	9,12 %	9,80 %
Project 4	9,85 %	9,24 %	11,42 %	11,36 %
RMSE	**7,22 %**	**7,18 %**	**7,78 %**	**7,96 %**
STD	**7,21 %**	**7,24 %**	**7,79 %**	**7,94 %**
Large building projects				
Project 5	2,59 %	4,67 %	4,89 %	8,62 %
Project 6	4,53 %	4,95 %	2,38 %	1,40 %
RMSE	**3,86 %**	**4,84 %**	**3,62 %**	**5,61 %**
STD	**3,13 %**	**2,91 %**	**3,68 %**	**5,36 %**
Large building projects				
Project 7	10,83 %	9,60 %	11,23 %	7,10 %
STD	**4,48 %**	**4,77 %**	**3,62 %**	**4,71 %**
All	**6,81 %**	**6,79 %**	**7,15 %**	**7,07 %**
STD	**6,78 %**	**6,84 %**	**7,17 %**	**7,08 %**
Progress < 50%				
Small building projects				
Project 1	5,79 %	5,33 %	6,38 %	5,97 %
Project 2	4,68 %	4,08 %	4,54 %	3,89 %
Project 3	5,59 %	3,85 %	6,05 %	6,43 %
Project 4	11,14 %	10,42 %	16,08 %	16,07 %
RMSE	**6,70 %**	**5,89 %**	**8,37 %**	**8,28 %**
STD	**6,81 %**	**5,94 %**	**8,51 %**	**8,40 %**
Large building projects				
Project 5	4,03 %	4,55 %	1,89 %	5,38 %
Project 6	4,96 %	3,43 %	4,60 %	2,77 %
RMSE	**4,60 %**	**3,93 %**	**3,73 %**	**4,05 %**
STD	**3,26 %**	**1,97 %**	**2,92 %**	**4,13 %**
Large building projects				
Project 7	10,81 %	10,23 %	10,12 %	5,30 %
STD	**3,36 %**	**4,46 %**	**2,63 %**	**4,44 %**
All	**6,77 %**	**6,08 %**	**7,36 %**	**6,67 %**
STD	**6,82 %**	**6,11 %**	**7,06 %**	**6,72 %**

Similar features of the models as in the pre-construction phase can be observed also in the construction phase as SVR_CC_OPT can generate better predictions for uncommon cash flow profiles. This can be seen from project 7 throughout the observed periods although SVR's and SVR_OPT's predictions are almost as accurate when progress is less than 10%. Also, SVR_CC_OPT and SVR_CC are able to reflect the delay in the beginning of project 1 in their forecasts when progress is less than 30%

whereas SVR and SVR_OPT catch up only when progress is smaller than 50%. However, SVR_CC_OPT is performing significantly worse in the common cash flow profiles of small and large building projects thorough the observed time frame.

Comparing the test results to the ones from training, there are both differences and similarities. The most notable distinction is that the performance of SVR_CC and SVR_CC_OPT starts to deteriorate after 30% of progress in the training phase, but when testing the model, both of their RMSE is over 1% better at 50% compared to 30% of progress. Another noteworthy difference is that the generalization ability of SVR_CC_OPT is notably better than SVR_CC as the gap between their performance increase from 0.26% to 0.80% when moving from the training phase to the test phase. The similarity of the phases, on the other hand, can be seen from the incremental improvement of SVR and SVR_OPT.

Standard deviations of the errors from all prediction periods are 7.77%, 7.80%, 8.60% and 7.79% for SVR, SVR_OPT, SVR_CC and SVR_CC_OPT, respectively. SVR, SVR_OPT and SVR_CC_OPT are again within a 0.03% range and there is only a 0.01% difference between each of the four model's RMSE and STD. Looking at the total STDs of each prediction period and model, there are not notable differences compared to the overall results.

The distributions of errors are presented in Figure 10. The mean errors for SVR, SVR_OPT, SVR_CC and SVR_CC_OPT are 0.51%, 0.59%, 0.65% and -0.67%. Notably, the mean errors have decreased significantly from the pre-construction phase forecasting. Similar to distributions in the pre-construction phase, SVR_CC_OPT's distribution has its peak near zero but for SVR_OPT it has moved more to the left side. Even though SVR, SVR_OPT and SVR_CC have longer right tails compared to SVR_CC_OPT, in the construction phase the difference is minuscule. Also, the positive skewness of SVR and SVR_OPT compliments their negative peaks of errors. On the other hand, the distribution of SVR_CC_OPT appears platykurtic and has the majority of its highest bars on the negative side. In addition, SVR_CC_OPT has a

higher density in its left tail compared to other models. This has also decreased the performance of SVR_CC_OPT the worst in terms of average error.

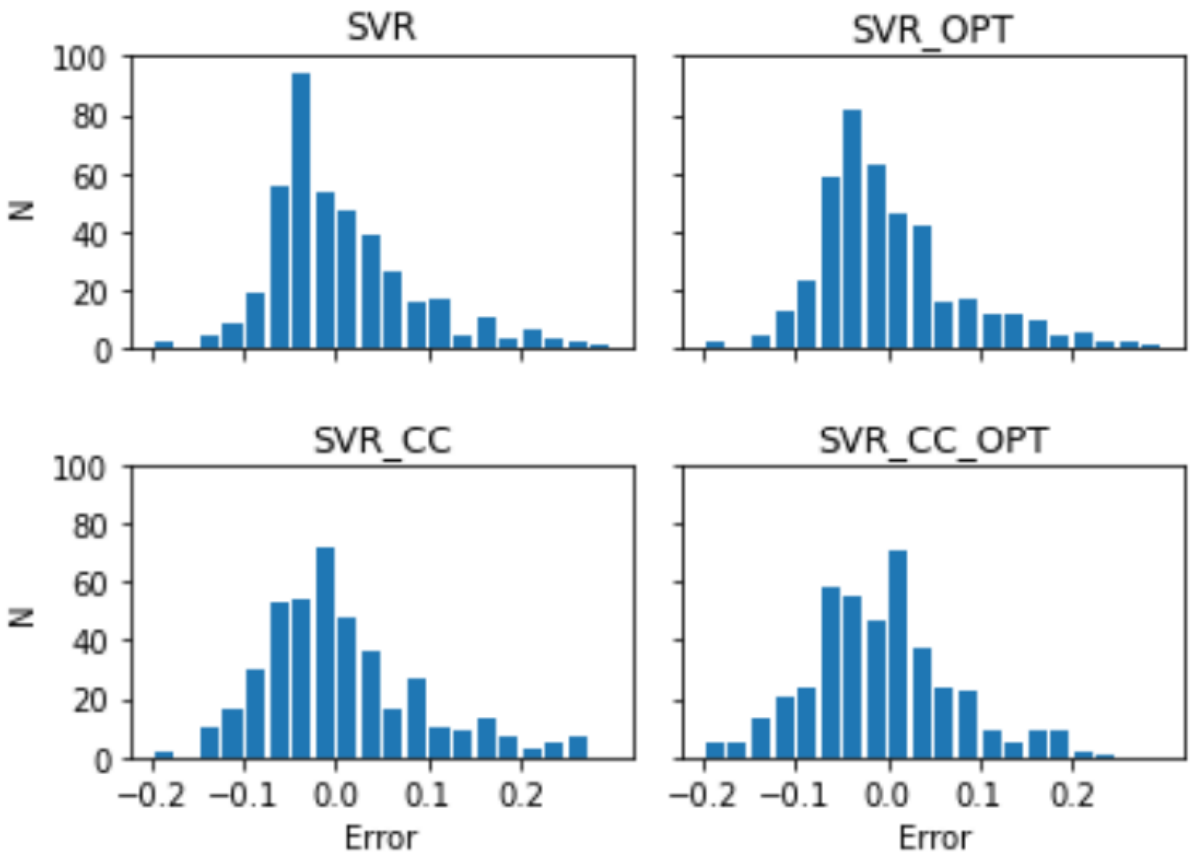


Figure 10. *Test error distributions of different models in the construction phase.*

5.3 Results analysis

As the literature review suggested that there are multiple unresolved issues in the mathematical forecasting models and that they are still in demand, the study presented a cash outflow model in chapter 4.1 that addresses all of the identified flaws in pre-construction models. As multiple enhancements were proposed, the study examined various intermediate versions of the model to estimate the impact of each suggested change. The benchmark model (Logit_CCC) along with four intermediate models (Logit_COC, SVR, SVR_OPT and SVR_CC) and the final model (SVR_CC_OPT) are presented in Table 12 in ascending order with respect to their overall performance.

Table 12. *Summary of the models and their performance in ascending order.*

Model	Description	Advantages	Disadvantages
Logit_CCC	Estimates a time-dependent cost commitment curve using linear regression and logit transformation. Derives cash outflow curve with a timelag based on project cost composition.	- Costflow forecasting is a well-researched topic - The estimates can be compared to periodic cost budgets (if available).	- Linear methodology - Exposes cash outflow forecasts to systematic error by assuming that distinct cost categories incur at a uniform rate - Cannot be updated during construction
Logit_COC	Estimates a time-dependent cash outflow curve using linear regression and logit transformation.	- Estimates cash outflow curve direcly which results in low systematic error	- Linear methodology - Does not account differences in project cost composition - Cannot be updated during construction
SVR_CC	Estimates cash outflow curve depending on time and project cost compositon using support vector regression.	- Estimates cash outflow curve direcly which results in low systematic error - Can be updated during construction - Non-linear methodology - Accounts differences in project cost composition	- Overfits the model and is therefore unable to generalize project cost composition data
SVR	Estimates a time-dependent cash outflow curve using support vector regression.	- Estimates cash outflow curve direcly which results in low systematic error - Can be updated during construction - Non-linear methodology - Good generalization ability	- Does not account differences in project cost composition
SVR_OPT	Estimates a time-dependent cash outflow curve using support vector regression. Optimizes hyperparameters with grid-search and k-fold cross-validation	- Estimates cash outflow curve direcly which results in low systematic error - Can be updated during construction - Non-linear methodology - Excellent generalization ability	- Does not account differences in project cost composition
SVR_CC_OPT	Estimates cash outflow curve depending on time and project cost compositon using support vector regression. Optimizes hyperparameters with grid-search and k-fold cross-validation	- Estimates cash outflow curve direcly which results in low systematic error - Can be updated during construction - Non-linear methodology - Accounts differences in project cost composition	- Occasional difficulties to generalize project cost composition data

The most significant modification made in this study is forecasting the cash outflow curve directly instead of using the traditional approach of forecasting the cost curve and deriving the cash outflow by applying a fixed time lag. This is a major change because after the publication of the net cash flow model proposed by Kaka & Price (1991) all of the research on mathematical pre-construction forecasting has focused solely on forecasting the cost commitment curve. However, this model imposes a systematic error to the cash outflow forecast by assuming that different costs are incurring at a uniform rate. As there is no research available on forecasting the cash outflow for the whole project duration, the results of this study are novel.

The impact of the proposed modification can be best evaluated by comparing the results of the logit model with cost commitment curve and fixed time lag (Logit_CCC) and logit model with cash outflow curve (Logit_COC) in the pre-construction phase forecasting. It must be noted that Logit_CCC suffers from more averaging compared to the approach of Kaka and Price (1991) as the used time lags are general instead of project-specific. However, it is difficult to assess the impact of this because some of the project-specific time lags may still be uncertain in the pre-construction phase which will cause errors in the project-specific approach. For example, subcontractor payment terms may be unknown because they are not hired yet. On the other hand, some of the time lags are fixed regardless of the project because the procurement is done at a company level. A comparison between Tables 4 and 8 actually suggests that the chosen methodology has supported the performance of Logit_CCC because the actual payment times have been longer than the used ones and the model has already underestimated the cash outflow.

The respective overall RMSEs of Logit_CCC and Logit_COC for test projects are 10.25% and 9.76%. Therefore, the difference is 0.49% to the advantage of Logit_COC. By project, Logit_COC has more accurate results for five of the seven projects. It can be seen from Figure 9 that most of the errors of Logit_CCC are on the right side of the y-axis which also explains the low standard deviation of the model. On the contrary, the errors of Logit_COC lay on both sides. This is an important quality as the main justification for using mathematical models is that the errors are canceled out in the

consolidated forecast of a project portfolio. The systematic error that Logit_CCC imposes on the cash outflow curve can be observed quantitatively by comparing its average error of 5.46% to the corresponding 0.05% of Logit_COC. The same effect can be also observed for all of the other models which is why it is listed as their advantage in Table 12. Therefore, it can be concluded that direct forecasting of the cash outflow curve will most likely yield better results especially in terms of lower systematic error.

Another modification of the proposed model explores the possibility of using project cost composition in defining the cash outflow profile in different phases of a project and its impact on forecast accuracy. In order to be able to use project cost composition in the proposed manner, the logit model must be abandoned, as it is designed to be used with only one predictive variable. Choosing SVR also tackles the problem of using linear models in pre-construction forecasting. The proposed enhancement also addresses the issue of using only the total cash outflow in current short-interval SVR models.

Comparing overall the pre-construction phase RMSE of, SVR, SVR_OPT and SVR_CC_OPT to the baseline of Logit_CCC and Logit_COC, there is no question of the superiority of SVR based model. Their differences in RMSEs compared to Logit_COC are 0.71%, 0.98% and 2.01%, respectively (when predictions are limited to zero or above). It also must be noted that while the logit model is able to forecast only 80% of the project duration, SVR is applied for the whole project duration. As Logit_COC still has the lowest mean error in the pre-construction phase, the advantage of using non-linear methodology is stemming from a better fitting ability instead of reducing systematic error before construction. However, there is a significant drop in the RMSE of SVR-based models in the construction phase. This indicates that the ability to be updated during construction is reducing systematic error. These two features are marked as an advantage to all of the SVR-based models in Table 12.

Applying project composition together with SVR in the pre-construction phase resulted in the best overall RMSE. However, some issues emerged as SVR_CC_OPT performed relatively worse with projects that have predictable cash outflow curves and generated negative predictions at the beginning of some projects. This effect is even worse for SVR_CC which has caused it to be performing worse than SVR and SVR_OPT even though it has more information available on the predicted projects. These overreactive traits of SVR_CC and SVR_CC_OPT indicate overfitting and giving too much weight to project cost composition.

The results from the construction phase are more ambiguous as the overall RMSE of all prediction periods for SVR, SVR_OPT and SVR_CC_OPT are within a 0.03% range. It seems like SVR_CC_OPT's reactivity to cost composition is still present and it proves to be useful in multiple instances during project 7 progress. This can be seen especially in the shift from 10% to 20% of progress where the predictions of SVR and SVR_OPT deteriorate to the all-time worst for this project but SVR_CC_OPT is still able to improve its accuracy. On the other hand, apart from this shift, SVR and SVR_OPT are able to improve their accuracy in each increment. In addition, they overwhelmingly beat SVR_CC_OPT in small and large building projects.

The test results also show that SVR_CC_OPT is not only unable to predict the surge in cash outflow in the last third of project 4 but is also defeated by SVR and SVR_OPT that do not have the information of adjusted weights of cost categories. This is an important observation because project 4 has 91.5% of its remaining cost budget allocated to materials and subcontracting prior to 50% of progress. In comparison, the percentage is 75.2%, 78.4% and 78.5% for projects 1, 2 and 3, respectively. Despite this information, SVR_CC_OPT is predicting that the accumulation of cash disbursements is decelerating drastically faster for project 4 compared to projects 1, 2 and 3. This, of course, is against common sense because over 90% of the remaining costs are allocated to the two largest categories of payables for project 4. The predictions of SVR_CC_OPT for projects 1-4 prior to 50% of progress are presented in Appendix 3. Overall, the construction phase results indicate that the utilization of

project cost composition does not seem to produce any improvements with the suggested methodology in the given phase.

Multiple observations indicate that this is caused by increasing importance in the level of project progress which dilutes the influence of project cost composition. Firstly, the training results of SVR_CC and SVR_CC_OPT actually get worse in the construction phase after 30% of progress. Even though the effect is smaller in the test phase, the relative performance of SVR_CC_OPT gets weaker in 40-50% of progress compared to 20-30% progress. Secondly, the performance of SVR_CC_OPT gets worse when exposing it to progress data from 10-30% of project duration compared to the pre-construction phase.

This follows with a question, that shouldn't SVR_CC_OPT still have the best performance even though the importance of project cost composition is lower, as long as it provides meaningful information that is not exposed to SVR and SVR_OPT. The results indicate that learning from cost composition data is a much harder task compared to information on progress in time and cash outflow. This can be seen from the large benefit that is gained by optimization of SVR_CC and then observing the results of SVR and SVR_OPT that are around the same level. It may be that the model is not sophisticated enough to fit the data correctly. On the other hand, this may also be caused by too little training data.

Along with the above observations, the decrease in the relative performance of SVR_CC_OPT when moving from pre-construction to the construction phase would suggest that this is an issue with the fitting of the model. This is because in the pre-construction phase all of the projects are in a similar stage and the data is as comparable as it can be. However, in the construction phase, the stages of different projects may be highly dissimilar which causes more fluctuation in actualized cash outflow and weights of the remaining cost categories. Consequently, the model needs to be able to fit a higher number of scenarios that result from different situations in various parts of the project.

6 Conclusions

The study has proposed a new cash outflow forecasting model to adjust the current state-of-the-art pre-construction models' weaknesses. The suggested improvements are supported by the literature on adjacent subjects of planning-based cash flow modeling and short-term interval forecasting. The key features of the model are mapping nonlinear and multivariate relationships, recognizing project uniqueness and minimizing systematic error. This is achieved by using support vector regression, clustering projects, utilizing project cost composition, updating the model during construction and forecasting cash outflow curve directly from historical data. The proposed model can be used to forecast a diverse project portfolio of ongoing and future known projects for the whole project duration.

In order to answer research questions one and two, the study has identified the central issues in different construction project cash flow forecasting models and the usage of support vector regression in this context. In mathematical pre-construction forecasting models, the issues are using linear models, deriving cash flow from a cost curve with a fixed time lag and using models that cannot be updated during construction nor be used to forecast the whole project duration. Another sub-category of mathematical models is short-interval forecast modeling which does not entail the latter two issues of pre-construction models, but on the other hand, is not very beneficial to the industry as financial forecasting is usually done on a more extended time horizon.

Short-interval forecasting is the only cash flow forecasting category where SVR is currently applied. It is also applied in other construction project control areas such as cost and time forecasting with earned value planning data in addition to cost and duration estimation. The results obtained from these studies have strongly favored SVR. On a comparison made outside of the construction industry, SVR outperformed other AI methods in forecasting financial time series and also proved to be the most robust with small samples. This also suggests that SVR is applicable to the construction industry as the project samples are often small.

On top of examining mathematical forecasting models, the study has also shortly explored planning-based cash flow models, namely planned earned value and cost-based forecasting and cost-schedule integrated models. The common benefit of using these idiographic models is being able to forecast with much greater detail on each project which will naturally lead to more accurate predictions. However, this benefit comes with a cost of manual work that is required from the project personnel. This is a major issue as it has been addressed that cost-schedule models are rarely applied in the industry because of this and the compatibility issues of the approach. In addition, these models are prone to human errors as they are entirely dependent on the accuracy of project schedules and financial planning.

The empirical results from the pre-construction phase suggest that the answer to the first sub-question of research question one is that direct forecasting of cash outflow curve reduces the systematic error that is caused by the standard practice of estimating cost commitment curve and applying a time lag that is based on project cost composition. There is also a slight improvement in individual project fit. The second sub-question of research question two can be answered positively as there are multiple improvements that are gained by applying support vector regression. The empirical results suggest that the largest advancement in improving individual project fit can be attributed to using support vector regression. It also allows the model to be expanded into a multivariate one. First, this way the model can be updated during construction which reduces the systematic error. Second, it allows project cost composition to be used directly in estimating cash outflow curve profile.

This also allows answering the second sub-question of research question one. Cost category weights improve individual project fit in the pre-construction phase but there are no signs of reduced systematic error. In the construction phase, their importance decreases along with project progress and therefore does not improve forecasting results significantly. However, the results indicate that cost category weights can be used to predict uncommon cash flow profiles throughout the project duration and the major issue is with overfitting to cost composition. Therefore this may also be caused

by too small training sample and some improvements may be gained with a larger dataset.

As one of the limitations of this study is using project data from only one contractor, further research with more construction companies could assess the above hypothesis along with testing whether the results can be generalized for more than one contractor. Another significant limitation of the study is focusing only on the cash outflow profile. Therefore, it cannot be used to forecast net cash flow. This could be achieved by conducting research on cash inflow curve estimation in a similar manner utilizing support vector regression that enables multivariate and non-linear modeling which has not been suggested before in the literature. The final limitation of the study considers the level of detail in the used data. As the model uses only general data and total budgets, construction companies should assess whether they have or should acquire more detailed data of the projects for practical applications. Comparative study between planning-based and mathematical models would make this assessment easier.

REFERENCES

Arditi, D., Koksal, A. and Kale, S. (2000) 'Business failures in the construction industry', *Engineering, Construction and Architectural Management*, 7(2), pp. 120–132. doi: 10.1108/eb021137.

Banki, M. T. and Esmaeeli, B. (2008) 'Using historical data for forecasting s-curves at construction industry', *2008 IEEE International Conference on Industrial Engineering and Engineering Management, IEEM 2008*. IEEE, pp. 282–286. doi: 10.1109/IEEM.2008.4737875.

Banki, M. T. and Esmaeili, B. (2009) 'The effects of variability of the mathematical equations and project categorizations on forecasting S-curves at construction industry', *International Journal of Civil Engineering*, 7(4), pp. 258–270.

Bao, Y. K. *et al.* (2005) 'Forecasting stock composite index by fuzzy support vector machines regression', *2005 International Conference on Machine Learning and Cybernetics, ICMLC 2005*, (August), pp. 3535–3540. doi: 10.1109/icmlc.2005.1527554.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*, *EAI/Springer Innovations in Communication and Computing*. New York: Springer ScienceBusiness Media. doi: 10.1007/978-3-030-57077-4_11.

Boussabaine, A. H. and Elhag, T. (1999) 'Applying fuzzy techniques to cash flow analysis', *Construction Management and Economics*, 17, pp. 745–755.

Boussabaine, A. H. and Kaka, A. P. (1998) 'A neural networks approach for cost flow forecasting', *Construction Management and Economics*, 16, pp. 471–479.

Boussabaine, A. H., Thomas, R. and Elhag, T. M. S. (1999) 'Modelling cost-flow forecasting for water pipeline projects using neural networks', *Engineering, Construction and Architectural Management*, 6(3), pp. 213–224. doi: 10.1108/eb021113.

Chao, L.-C. and Chien, C.-F. (2009) 'Estimating Project S-Curves Using Polynomial Function and Neural Networks', *Journal of Construction Engineering and Management*, 135(3), pp. 169–177. doi: 10.1061/(asce)0733-9364(2009)135:3(169).

Chao, L. C. (2013) ‘Estimating project S-curve based on project attributes and conditions’, *Proceedings of the 13th East Asia-Pacific Conference on Structural Engineering and Construction, EASEC 2013*.

Chao, L. C. and Chien, C. F. (2010) ‘A Model for Updating Project S-curve by Using Neural Networks and Matching Progress’, *Automation in Construction*. Elsevier B.V., 19(1), pp. 84–91. doi: 10.1016/j.autcon.2009.09.006.

Chen, G. W. *et al.* (2010) ‘Application of project cash management and control for infrastructure’, *Journal of Marine Science and Technology*, 18(5), pp. 644–651.

Chen, H. L., Chen, W. T. and Wei, N. C. (2011) ‘Developing a cost-payment coordination model for project cost flow forecasting’, *Journal of Civil Engineering and Management*, 17(4), pp. 494–509. doi: 10.3846/13923730.2011.604540.

Chen, H. L., O’Brien, W. J. and Herbsman, Z. J. (2005) ‘Assessing the accuracy of cash flow models: The significance of payment conditions’, *Journal of Construction Engineering and Management*, 131(6), pp. 669–676. doi: 10.1061/(ASCE)0733-9364(2005)131:6(669).

Cheng, M. Y. *et al.* (2013) ‘Enhanced time-dependent evolutionary fuzzy support vector machines inference model for cash flow prediction and estimate at completion’, *International Journal of Information Technology and Decision Making*, 12(4), pp. 679–710. doi: 10.1142/S0219622013500259.

Cheng, M. Y., Cao, M. T. and Herianto, J. G. (2020) ‘Symbiotic organisms search-optimized deep learning technique for mapping construction cash flow considering complexity of project’, *Chaos, Solitons and Fractals*. Elsevier Ltd, 138. doi: 10.1016/j.chaos.2020.109869.

Cheng, M. Y., Hoang, N. D. and Wu, Y. W. (2012) ‘Prediction of project cash flow using time-depended evolutionary LS-SVM inference model’, *2012 Proceedings of the 29th International Symposium of Automation and Robotics in Construction, ISARC 2012*, (August), pp. 0–8. doi: 10.4017/gt.2012.11.02.223.00.

Cheng, M. Y., Hoang, N. D. and Wu, Y. W. (2015) ‘Cash flow prediction for construction project using a novel adaptive time-dependent least squares support vector machine inference model’, *Journal of Civil Engineering and Management*, 21(6), pp. 679–688. doi: 10.3846/13923730.2014.893906.

Cheng, M. Y. and Roy, A. F. V. (2011) 'Evolutionary fuzzy decision model for cash flow prediction using time-dependent support vector machines', *International Journal of Project Management*. Elsevier Ltd and IPMA, 29(1), pp. 56–65. doi: 10.1016/j.ijproman.2010.01.004.

Cheng, M. Y., Tsai, H. C. and Liu, C. L. (2009) 'Artificial intelligence approaches to achieve strategic control over project cash flows', *Automation in Construction*, 18(4), pp. 386–393. doi: 10.1016/j.autcon.2008.10.005.

Cheng, M. Y. and Wu, Y. W. (2009) 'Evolutionary support vector machine inference system for construction management', *Automation in Construction*. Elsevier B.V., 18(5), pp. 597–604. doi: 10.1016/j.autcon.2008.12.002.

Cheng, Y.-M., Yu, C.-H. and Wang, H.-T. (2011) ' Short-Interval Dynamic Forecasting for Actual S -Curve in the construction phase ', *Journal of Construction Engineering and Management*, 137(11), pp. 933–941. doi: 10.1061/(asce)co.1943-7862.0000358.

Chiao Lin, M. *et al.* (2012) 'A novel dynamic progress forecasting approach for construction projects', *Expert Systems with Applications*. Elsevier Ltd, 39(3), pp. 2247–2255. doi: 10.1016/j.eswa.2011.07.093.

Cho, D., Lee, M. and Shin, J. (2020) 'Development of cost and schedule data integration algorithm based on big data technology', *Applied Sciences (Switzerland)*, 10(24), pp. 1–17. doi: 10.3390/app10248917.

Cristóbal, J. R. S. *et al.* (2015) 'A Residual Grey Prediction Model for Predicting S-curves in Projects', *Procedia Computer Science*. Elsevier Masson SAS, 64(December), pp. 586–593. doi: 10.1016/j.procs.2015.08.570.

Cui, Q., Hastak, M. and Halpin, D. (2010) 'Systems analysis of project cash flow management strategies', *Construction Management and Economics*, 28(4), pp. 361–376. doi: 10.1080/01446191003702484.

Espinoza, M., Suykens, J. A. K. and De Moor, B. (2005) 'Load forecasting using fixed-size least squares support vector machines', *Lecture Notes in Computer Science*, 3512, pp. 1018–1026. doi: 10.1007/11494669_125.

Evans, R. C. and Kaka, A. P. (1998) 'Analysis of the accuracy of standard/average value curves using food retail building projects as case studies', *Engineering,*

Construction and Architectural Management, 5(1), pp. 58–67. doi: 10.1108/eb021061.

Hongjiu, L., Rieg, R. and Yanrong, H. (2012) 'Performance comparison of artificial intelligence methods for predicting cash flow', *Neural Network World*, 22(6), pp. 549–564. doi: 10.14311/NNW.2012.22.034.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. (2003) *A Practical Guide to Support Vector Classification*. doi: 10.1177/02632760022050997.

Hua, G. B. (2008) 'The state of applications of quantitative analysis techniques to construction economics and management (1983 to 2006)', *Construction Management and Economics*, 26(5), pp. 485–497. doi: 10.1080/01446190801998716.

Hwee, N. G. and Tiong, R. L. K. (2002) 'Model on cash flow forecasting and risk analysis for contracting firms', *International Journal of Project Management*, 20(5), pp. 351–363. doi: 10.1016/S0263-7863(01)00037-0.

Kaka, A. (1996) 'Towards more flexible and accurate cash flow forecasting', *Construction Management and Economics*, 14, pp. 35–44.

Kaka, A. P., Lewis, J. and Petros, H. (2003) 'The effects of the variability of project planning on cost commitment curves: A case study', *Engineering, Construction and Architectural Management*, 10(1), pp. 15–26. doi: 10.1108/09699980310466514.

Kaka, A. and Price, A. D. F. (1991) 'Net cash flowmodels: Are they reliable?', *Construction Management and Economics*, 9, pp. 291–308.

Kaka, A. and Price, A. D. F. (1993) 'Modelling standard cost commitment curves for contractors' cash flow forecasting', *Construction Management and Economics*, 11, pp. 271–283.

Kenley, R. (1999) 'Cash farming in building and construction: a stochastic analysis', *Construction Management and Economics1*, 17, pp. 393–401.

Kenley, R. and Wilson, O. D. (1986) 'A construction project cash flow model - an idiographic approach', *Construction Management and Economics*, 4, pp. 213–232.

Kenley, R. and Wilson, O. D. (1989) 'A construction project net cash flow model',

Construction Management and Economics, 7, pp. 3–18.

Lin, H. and Lin, C. (2003) ‘A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods’, *Neural Computation*, (2), pp. 1–32. Available at: http://home.caltech.edu/~htlin/publication/doc/tanh.pdf.

Liu, F. and Deng, Y. (2020) ‘Determine the number of unknown targets in Open World based on Elbow method’, *IEEE Transactions on Fuzzy Systems*. IEEE, 29(5), pp. 1–1. doi: 10.1109/tfuzz.2020.2966182.

Mills, P. and Tasaico, H. A. (2005) ‘Forecasting payments made under construction contracts: Payout curves and cash management in the North Carolina Department of Transportation’, *Transportation Research Record*, (1907), pp. 25–33. doi: 10.3141/1907-04.

Müller, K. R. *et al.* (1997) ‘Predicting time series with support vector machines’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1327(1), pp. 999–1004. doi: 10.1007/bfb0020283.

Nam, C. H. and Tatum, C. B. (1988) ‘Major characteristics of constructed products and resulting limitations of construction technology’, *Construction Management and Economics*, pp. 133–147. doi: 10.1080/01446198800000012.

Navon, R. (1995) ‘Resource-based model for automatic cash-flow forecasting’, *Construction Management and Economics*, 13, pp. 501–510.

Navon, R. (1996) ‘Company-level cash-flow management’, *Journal of Construction Engineering and Management*, 122, pp. 22–35.

Odeyinka, H. A., Lowe, J. and Kaka, A. (2008) ‘An evaluation of risk factors impacting construction cash flow forecast’, *Journal of Financial Management of Property and Construction*, 13(1), pp. 5–17. doi: 10.1108/13664380810882048.

Odeyinka, H., Lowe, J. and Kaka, A. (2012) ‘Regression modelling of risk impacts on construction cost flow forecast’, *Journal of Financial Management of Property and Construction*, 17(3), pp. 203–221. doi: 10.1108/13664381211274335.

Park, H.-K. (2004) ‘Cash flow forecasting in construction project’, *KSCE Journal of Civil Engineering*, 8(3), pp. 265–271. doi: 10.1007/bf02836008.

Park, H. K., Han, S. H. and Russell, J. S. (2005) ‘Cash flow forecasting model for general contractors using moving weights of cost categories’, *Journal of Management in Engineering*, 21(4), pp. 164–172. doi: 10.1061/(ASCE)0742-597X(2005)21:4(164).

Ross, A., Dalton, K. and Sertyesilisik, B. (2013) ‘An investigation on the improvement of construction expenditure forecasting’, *Journal of Civil Engineering and Management*. Taylor & Francis, 19(5), pp. 759–771. doi: 10.3846/13923730.2013.793607.

Sapankevych, N. I. and Sankar, R. (2009) ‘Time series prediction using support vector machines: a survey’, *IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE*, 4, pp. 24–38.

Shash, A. A. and Qarra, A. Al (2018) ‘Cash Flow Management of Construction Projects in Saudi Arabia’, *Project Management Journal*, 49(5), pp. 48–63. doi: 10.1177/8756972818787976.

Singh, S. and Lokanathan, G. (1992) ‘Computer-based Cash Flow Model’, *American Association of Cost Engineers*, p. R.5.1-R.5.14. Available at: http://dx.doi.org/10.1016/j.jaci.2012.05.050.

Skitmore, M. (1992) ‘Parameter prediction for cash flow forecasting models’, *Construction Management and Economics*, 10, pp. 397–413.

Skitmore, M. (1998) ‘A method for forecasting owner monthly construction project expenditure flow’, *International Journal of Forecasting*, 14(1), pp. 17–34. doi: 10.1016/S0169-2070(97)00042-3.

Smola, A. J. and Schölkopf, B. (2004) ‘A tutorial on support vector regression’, *Statistics and Computing*, 14, pp. 199–222. Available at: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1CAD92EF8CCE726A305D8A41F873EEFC?doi=10.1.1.114.4288&rep=rep1&type=pdf%0Ahttp://download.springer.com/static/pdf/493/art%3A10.1023%2FB%3ASTCO.0000035301.49549.88.pdf?auth66=1408162706_8a28764ed0fae9.

Sorrell, S. (2003) ‘Making the link: Climate policy and the reform of the UK construction industry’, *Energy Policy*, 31(9), pp. 865–878. doi: 10.1016/S0301-4215(02)00130-1.

Sousa, J. C., Jorge, H. M. and Neves, L. P. (2014) ‘Short-term load forecasting based on support vector regression and load profiling’, *International Journal of Energy Research*, 38(3), pp. 350–362. doi: 10.1002/er.3048.

Tabyang, W. and Benjaoran, V. (2013) ‘The impact of consideration of payment conditions in cash flow forecasting on financing costs in construction’, *Proceedings of the 13th East Asia-Pacific Conference on Structural Engineering and Construction, EASEC 2013*.

Teerajetgul, W., Chareonngam, C. and Wethyavivorn, P. (2009) ‘Key knowledge factors in Thai construction practice’, *International Journal of Project Management*. Elsevier Ltd and IPMA, 27(8), pp. 833–839. doi: 10.1016/j.ijproman.2009.02.008.

Tserng, H. P. *et al.* (2014) ‘Prediction of default probability for construction firms using the logit model’, *Journal of Civil Engineering and Management*, 20(2), pp. 247–255. doi: 10.3846/13923730.2013.801886.

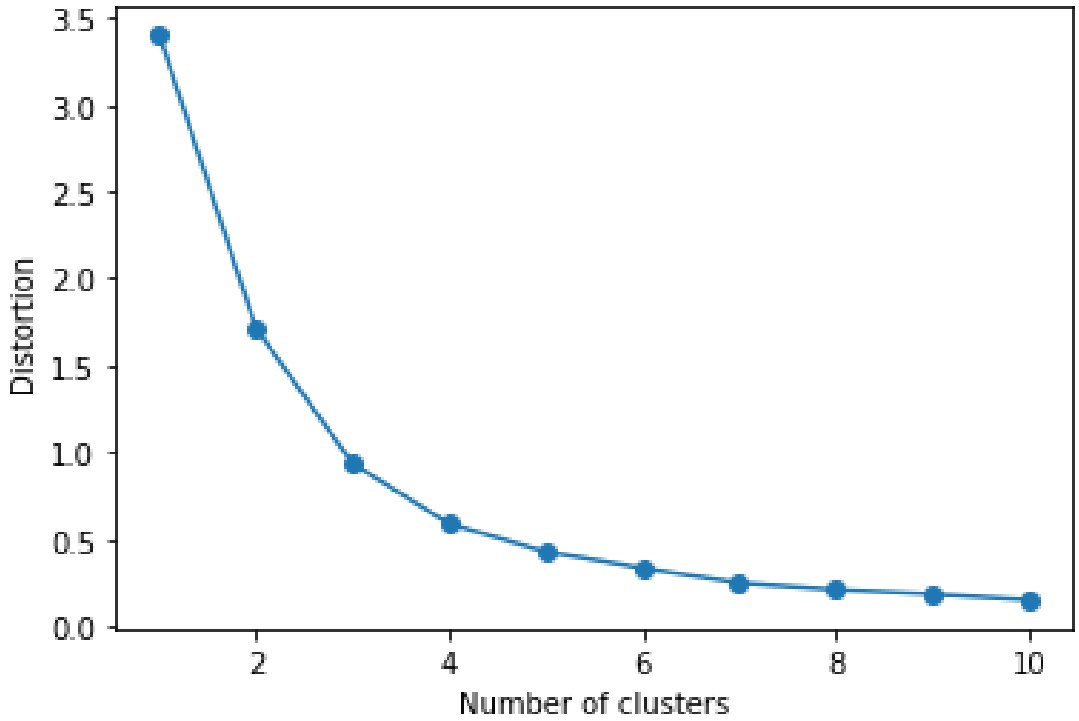
Wauters, M. and Vanhoucke, M. (2014) ‘Support Vector Machine Regression for project control forecasting’, *Automation in Construction*. Elsevier B.V., 47, pp. 92–106. doi: 10.1016/j.autcon.2014.07.014.

Webster, J. and Watson, R. T. (2002) ‘Analyzing Past To Prepare For Future: Writing Literature Review’, *MIS Quarterly*, 26(2), pp. 13–23.

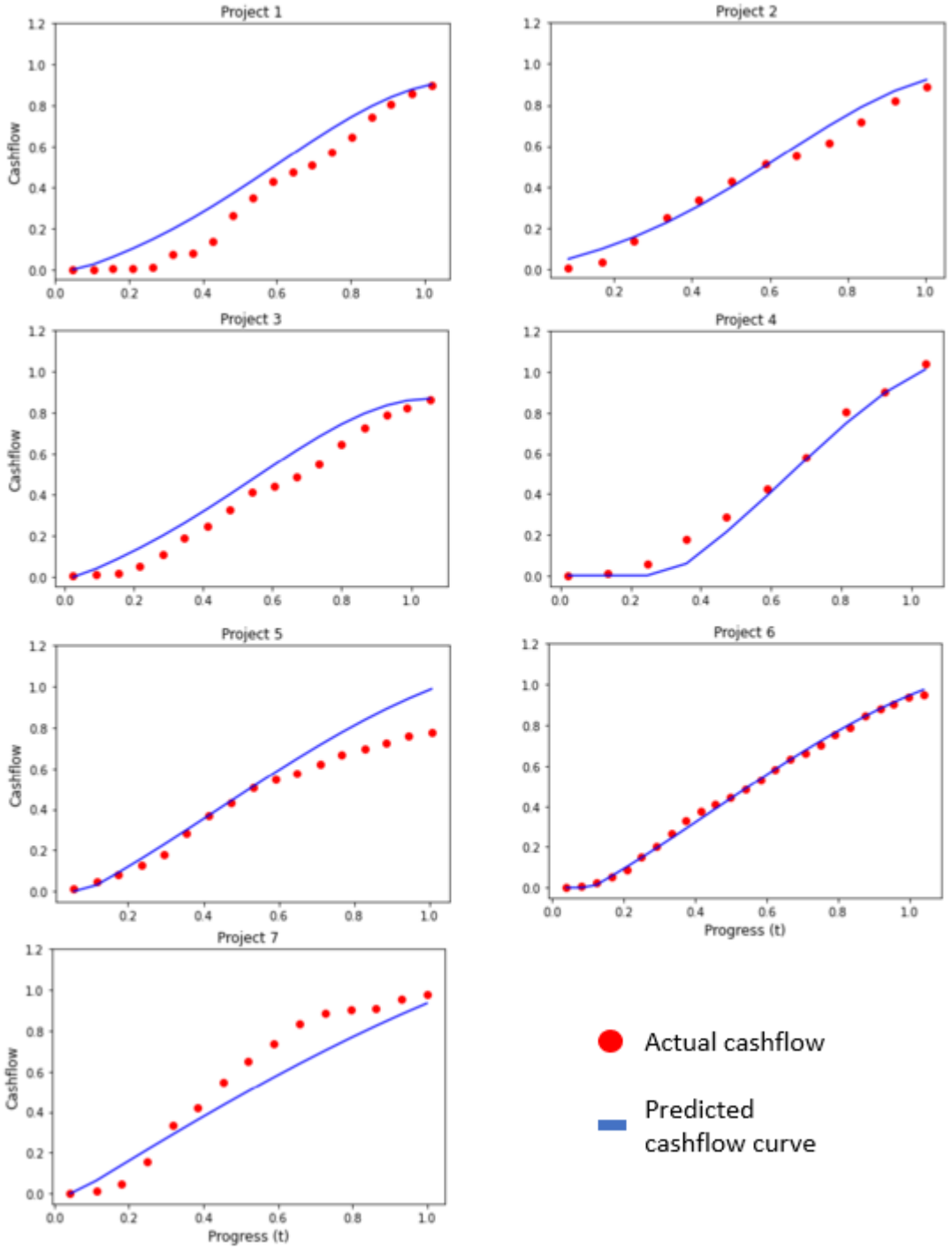
Willems, L. L. and Vanhoucke, M. (2015) ‘Classification of articles and journals on project control and earned value management’, *International Journal of Project Management*. Elsevier Ltd. APM and IPMA., 33(7), pp. 1610–1634. doi: 10.1016/j.ijproman.2015.06.003.

Zayed, T. and Liu, Y. (2014) ‘Cash flow modeling for Construction projects’, *Engineering, Construction and Architectural Management*, 21(2), pp. 170–189. doi: 10.1108/ECAM-08-2012-0082.

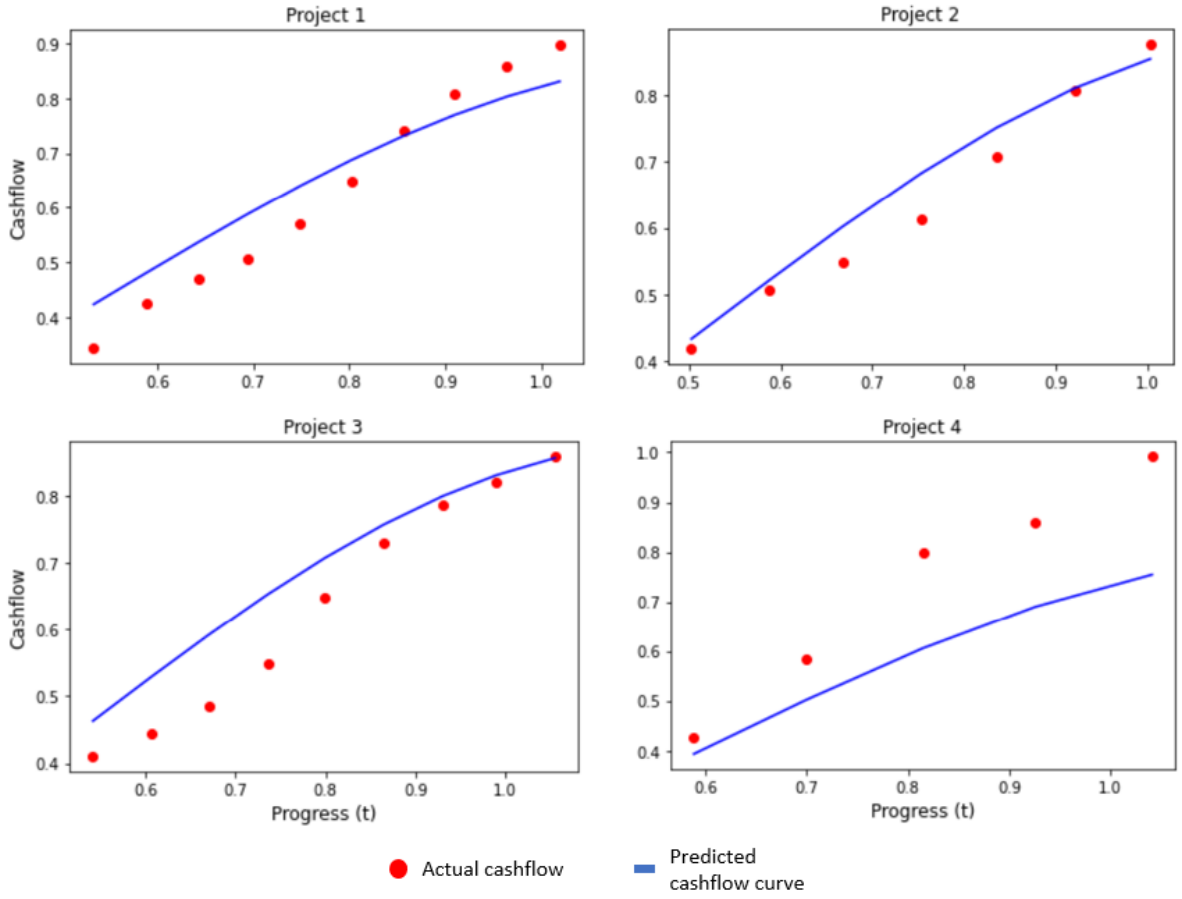
APPENDICES



Appendix 1. *Elbow Method using 1 to 10 clusters*



Appendix 2. *Predictions of optimized support vector regression using project cost composition (SVR_CC_OPT) and the actual cash flow for test projects in the pre-construction phase (predictions limited to zero).*



Appendix 3. *Predictions of optimized support vector regression using project cost composition (SVR_CC_OPT) and the actual cash flow for test projects 1-4 at <50% progress.*