Finnish Transport
Infrastructure Agency

Thesis
3/2020

Hanna Sandell

FORECASTING INDIRECT COSTS IN FINNISH PUBLIC TRANSPORT INFRASTRUCTURE PROJECTS

Applications with Machine Learning Models

Hanna Sandell

Forecasting Indirect Costs in Finnish Public Transport Infrastructure Projects

Applications with Machine Learning Models

Thesis 3/2020

Finnish Transport Infrastructure Agency

Helsinki 2020

Cover picture: FTIA's photo archives

Online publication pdf (www.vayla.fi)

ISSN 2490-1202
ISBN 978-952-317-803-8

Finnish Transport Infrastructure Agency
P.O.Box 33
FIN-00521 HELSINKI, Finland
Tel. +358 (0)295 34 3000

Hanna Sandell: Forecasting Indirect Costs in Finnish Public Transport Infrastructure Projects: Applications with Machine Learning Models. Finnish Transport Infrastructure Agency 2020. Thesis 3/2020. 51 pages. ISSN 2490-1202, ISBN 978-952-317-803-8.

Key words: Machine learning, Cost estimation, Transport infrastructure

Abstract

The inaccuracy of transport infrastructure projects' cost estimation has become large issue especially because the amount of large mega projects has been increasing during past few years. The cost estimation inaccuracy is problematic because it biases the results of cost-benefit analysis, which is used to measure the profitability of a project. Subsequently this bias can lead to the misallocation of scarce resources. Besides the construction costs, cost estimation includes the computation of owner's indirect costs. In this thesis, owner's indirect costs cover the construction management costs and design costs of the project. According to the current instructions, indirect costs are calculated using fixed default values. As the currently used calculation method does not take the project's individual properties into account, need for alternative approach has increased.

Thus the objective of this thesis is to forecast owner's indirect costs in the late phases of the infrastructure projects by applying two different machine learning models: linear multiple regression model and artificial neural network model. Additionally, the aim is to study, whether machine learning models provided can outperform the currently used instructions in the prediction of indirect costs. Aim of well-functioning forecast model would be to improve the cost estimation's accuracy level. In this thesis, owner is defined as the government and indirect costs are only forecasted in later phases of the project.

Research question is attempted to solve by applying two commonly used machine learning models: artificial neural network and multiple regression model. Neural network used in this thesis is a feedforward network, which learning mechanism is based on backpropagation algorithm. Multiple regression model utilizes traditional OLS method in the estimation of parameters' values. Models are constructed with data provided by Finnish Transport Infrastructure Agency. Data includes infrastructure projects' initial data and the actual shares of design and construction management costs of each project.

As an outcome, this thesis provides two preliminary forecast models for owner's indirect costs. The results also indicate that the neural network and regression model are able to forecast owner's indirect costs in both categories with higher accuracy compared to the current instructions. Furthermore, study aided to recognize influential variables affecting the indirect costs. During the research process, also few improvements for further development of the forecast models were identified.

From the machine learning models, neural network performs better in forecasting the design costs and regression model is able to forecast the construction management costs with slightly better accuracy. These results support the conclusion that costs with uncertain and missing information can be forecasted more precisely with more complex machine learning models, such as the artificial neural network. On the other hand, costs with comprehensive knowledge can be accurately predicted with simpler models, such as the multiple regression.

Hanna Sandell: Tilaajan hanketehtävien ennustaminen infrahankkeissa: sovellutuksia koneoppimismenetelmillä. Väylävirasto. Helsinki 2020. Opinnäytetyö 3/2020. 51 sivua. ISSN 2490-1202, ISBN 978-952-317-803-8.

Asiasanat: koneoppiminen, kustannusarvio, liikenneinfrastruktuuri

Tiivistelmä

Infrahankkeiden kustannusarvioiden epätarkkuus on merkittävä ongelma etenkin, kun suurten hankkeiden määrä on lisääntynyt voimakkaasti viime vuosina. Epätarkkuus vaikuttaa paisti tuottavuuden arvioimiseen käytettävän kustannus-hyötyanalyysin tarkkuuteen, mutta myös rajallisten resurssien allokoitumiseen. Rakentamisen kustannuksien lisäksi, kustannusarviosta voidaan erotella tilaajalle koituvat epäsuorat kustannukset, tilaajan hanketehtävät. Kustannusarvioiden nykyisen ohjeistuksen mukaan hanketehtävät lasketaan hankkeelle vakioituja prosenttiosuuksia käyttäen. Nykyinen hanketehtävien arvioimiseen käytetty malli ei kuitenkaan kykene ottamaan huomioon hankkeisiin liittyviä yksilöllisiä muuttujia ja tarve vaihtoehtoiselle arviointimenetelmälle on kasvanut.

Tämän tutkielman ensisijaisena tavoitteena on tutkia tilaajalle koituvien kustannuksien ennustamista koneoppimisen menetelmiä hyödyntäen. Lisäksi tavoitteena on selvittää, onko infrahankkeista tilaajalle koituvia kustannuksia mahdollista ennustaa olemassa olevaa ohjeistusta tarkemmin hyödyntäen koneoppimisen menetelmiä. Toimivan ennustemallin tavoitteena on parantaa kustannusarvioiden tarkkuutta nykyisestä tasostaan. Tutkielmassa tilaajaksi määritellään valtio ja tilaajan hanketehtävistä tarkastellaan ainoastaan hankkeen myöhäisten vaiheiden aikaisia suunnittelun ja rakennuttamisen kustannuksia.

Tutkimusongelmaa pyritään ratkaisemaan hyödyntämällä kahta yleisesti käytettyä koneoppimismallia: keinotekoista neuroverkkoa ja usean muuttujan lineaarista regressiomallia. Neuroverkko on rakenteeltaan eteenpäin syöttävä verkko, jonka oppiminen perustuu vastavirta-algoritmiin. Regressiomallissa parametrien estimointi tapahtuu puolestaan perinteisellä OLS-metodilla. Tutkielman mallien rakentamisessa hyödynnetty aineisto on saatu Väylävirastolta. Hyödynnetty aineisto koostuu hankkeiden perustiedoista sekä suunnittelu- ja rakennuttamiskustannuksien toteumatiedoista.

Tutkielman tuloksena saatiin rakennettua kaksi alustavaa ennustemallia tilaajan hanketehtäville. Tulokset osoittavat, että neuroverkko sekä regressiomalli kykenevät ennustamaan suunnittelu- sekä rakennuttamiskustannuksia paremmin kuin nykyinen malli. Lisäksi tutkimus auttoi tunnistamaan tilastollisesti merkittäviä muuttujia, jotka vaikuttavat tilaajan hanketehtävien kustannuksiin. Tutkimusprosessin aikana tunnistettiin myös malleissa olevia kehityskohteita, joihin mahdollisessa jatkotutkimuksessa tulisi keskittyä.

Koneoppimismalleista neuroverkko suoriutuu paremmin suunnittelukustannuksien ennustamisesta ja regressiomalli puolestaan ennustaa rakennuttamisen kustannuksia tarkemmalla tasolla. Lisäksi tutkimuksen tulokset tukevat aikaisempaa johtopäätöstä siitä, että kustannuksia, joihin liittyvä tieto on vähäistä, on kannattavampaa ennustaa monimutkaisemmilla koneoppimismenetelmillä, kuten neuroverkolla. Toisaalta kustannuksia, joihin liittyvä tieto on kattavaa, voidaan ennustaa tarkasti hyödyntämällä yksinkertaisempia koneoppimisen menetelmiä, kuten usean muuttujan regressiomallia.

Hanna Sandell: Prognostisering av indirekta kostnader i infrastrukturprojekt i Finland. Trafikledsverket. Helsingfors 2020. Lärdomsprov 3/2020. 51 sidor. ISSN 2490-1202, ISBN 978-952-317-804-5-803-8.

Sammandrag

Bristen på precision i kostnadsberäkningarna i infrastrukturprojekt är ett betydande problem särskilt då stora projekt blivit avsevärt vanligare under de senaste åren. Inexaktheten påverkar inte bara precisionen av den kostnads-nyttoanalys som används för att uppskatta produktiviteten, men även allokeringen av de begränsade resurserna. I en kostnadsberäkning kan utöver byggkostnaderna även de indirekta kostnaderna för beställaren, dvs. beställarens projektuppgifter specificeras. Enligt de nuvarande anvisningarna om kostnadsberäkning beräknas projektuppgifter för projektet med hjälp av standardiserade procentandelar. Den nuvarande modellen som används för att uppskatta projektuppgifter kan dock inte ta hänsyn till individuella variabler i projekten, vilket lett till ett ökat behov av en alternativ uppskattningsmetod.

Det primära målet med denna avhandling är att studera prognostisering av beställarens kostnader med hjälp av maskininlärningsmetoder. Ett ytterligare syfte är att utreda huruvida det är möjligt att med maskininlärningsmetoder förutsäga kostnader som infrastrukturprojekt medför för beställaren med högre precision än den som de nuvarande anvisningarna medger. Syftet med en välfungerande prognosmodell är att förbättra kostnadsberäkningarnas precision från dagens nivå. I avhandlingen definieras staten som kunden, och av beställarens projektuppgifter studeras endast kostnaderna för planering och byggande i projektets sena skeden.

För att lösa forskningsproblemet används två vanliga modeller för maskininlärning: ett artificiellt neuronnätverk och en multipel linjär regressionsmodell. Ett neuronnätverk är ett transfernätverk vars inlärning bygger på en feedback-algoritm. I regressionsmodellen estimeras parametrarna i sin tur med den traditionella OLS-metoden. Underlaget som användes för att bygga modellerna i avhandlingen har erhållits av Trafikledsverket. Underlaget består av grundläggande uppgifter om projekt samt av uppgifter om planerings- och byggkostnadernas utfall.

Som ett resultat av avhandlingen byggdes två preliminära prognosmodeller för beställarens projektuppgifter. Resultaten visar att neuronnätverket och regressionsmodellen kan prognostisera planerings- och byggkostnader bättre än den nuvarande modellen. Undersökningen hjälpte också med att identifiera statistiskt signifikanta variabler som påverkar kostnaderna för beställarens projektuppgifter. Under undersökningsprocessen identifierades dessutom områden i modellerna som kräver utveckling och på vilka eventuell fortsatt forskning borde inriktas.

Av maskininlärningsmodellerna är neuronnätverket bättre när det gäller att prognostisera planeringskostnader, medan regressionsmodellen i sin tur förutspår byggkostnader med högre precision. Undersökningens resultat stödjer dessutom den tidigare slutsatsen om att det är lönsammare att använda mer komplexa maskininlärningsmetoder, till exempel neuronnätverk, för prognostisera kostnader om vilka kunskaperna är ringa. Å andra sidan kan kostnader om vilka kunskaperna är täckande med hög precision genom att använda enklare metoder för maskininlärning, såsom multipel linjär regression.

Preface

During recent years, there has been major investments on improving the quality of the transport infrastructure projects' cost estimations. In cost estimations, indirect costs of infrastructure projects are estimated on top of the other implementation costs and the current policy advises to estimate the indirect costs by using constant default values. However, the policy does not take into account project's individual properties which affect directly to the proportion of the indirect costs. Thus the objective of this research was to study potential ways to forecast the indirect costs by applying machine learning methods. Study aimed to develop forecast models which could be able to forecast the proportion of indirect costs more reliably compared to the current policy. Forecast models constructed in this thesis will be developed further in the future as the amount of collected data increases.

This master's thesis was written by Hanna Sandell at University of Helsinki, Faculty of Social Sciences, Economics. The thesis was advised and guided by Ari Huomo from Finnish Transport Infrastructure Agency. The thesis was supervised by Mika Meitz from University of Helsinki."

Helsinki September 2020

Finnish Transport Infrastructure Agency

Contents

1 Introduction

Investments on infrastructure are extremely relevant in promoting economic growth. The effect of how public infrastructure investments impact economic well-being is a well researched topic (Munnel, 1992; D´emurger, 2001). In these studies, it was noticed that the government's participation to the funding of infrastructure development is necessary in order to enable effective and wellfunctioning infrastructure system. Furthermore, results of these studies have been clearly showing that the use of public resources on infrastructure development increases the output on private sector. Funding the infrastructure projects is profitable for the government because the infrastructure investments can have a large positive effect on country's gross domestic product (GDP) (Cantarelli, Flyvbjerg & Molin 2010). For instance, Honkatukia and Antikainen (2004) estimated that if investments on infrastructure worth of 3.5 billion euros would be implemented, it could increase Finland's GDP even more than 0.3% in a year.

However, the decision whether a new infrastructure project is implemented or not is based on the cost-benefit analysis (CBA), which measures the project's profitability (Layard & Glaister 1994). Functionality of CBA is based on the assumption that benefits and costs of a project can be measured reliably. Unfortunately, producing reliable cost estimations is extremely difficult. Wrong calculations of costs lead to severe inaccuracies in cost estimations and these falsely estimated costs lead to a large bias in CBA. Thus correct cost estimations could benefit the CBA process substantially. In addition, inaccurate cost estimation does not only bias the CBA but also often leads to the misallocation of scarce resources (Flyvbjerg, Holm & Buhl 2002; Button, 2010). More exact cost analysis could therefore lead to better resource allocation and as Button (2010) mentions, solving this misallocation is an important task in transport economics.

The precise effects of inaccurate cost estimation in infrastructure investments has been researched in many previous studies (Flyvbjerg, Holm & Buhl 2002; Lundberg, Jenpanitsub & Pyddok, 2011; Cantarelli, Flyvbjerg, Molin & Wee, 2010). The common conclusion in these studies has been that despite the large-scale effects of cost over- and underruns, the accuracy in cost estimation has not improved during the past few decades. Although it is clear that the accuracy of cost estimation plays a key role in the development of the infrastructure field, accurate cost estimation has turned out to be extremely difficult. Due to the problematic nature of cost estimation, methods that enable accurate measurement of costs in construction projects have become an interesting topic as the technology has developed and the transport infrastructure industry is changing towards more automatized practises. As a result, the use of machine learning models in cost estimation have increased rapidly (Bouabaz & Hamami, 2008; Kim, An & Kang, 2004). Especially, the increased use of artificial neural network (ANN) models has been a successful step in the development of cost estimation. ANN models have enabled the use of corrupted, noisy and incomplete data in cost estimation problems and therefore the new models are capable to address problems with missing information (Kim, An & Kang, 2004).
To meet the increasing expectations set by the developing infrastructure industry Finnish Transport Infrastructure Agency (FTIA) has launched a program where the goal is to design new cost management system. Objective of the cost

management system is to create reliable and open cost data in order to achieve accurate cost estimations. Besides the traditional construction costs, also the estimation of indirect cost of the project has proven to be complicated. Alternative approach is required because the calculation of the indirect costs is not possible with the similar input logic that is applied to the calculation of building costs and as Emsley et al. (2002) note, the construction costs should always be modeled separately from the indirect costs, due to the large variety related to these costs.

Thus the objective of this thesis is to forecast owner's indirect costs in the late phases of the infrastructure projects by applying two different machine learning models: linear multiple regression model (MR) and artificial neural network model (ANN). Additionally, the aim is to study, whether machine learning models provided can outperform the currently used instructions in the prediction of indirect costs. Main reason that the accurate estimation of indirect costs has been difficult so far is the lack of existing analysis and thus the intention is to provide new insight on the topic. To my knowledge, there exists no studies, which examine how the indirect costs in FTIA's projects are formed and what variables effect on these costs. The current instructions estimate the size of indirect costs to be as much as 20-50% of the overall costs depending on the phase of the project (Liikennevirasto, 2013). Thus deeper knowledge and more accurate prediction ability are necessary to ensure more reliable cost estimations. Although, in this thesis I will cover only late phases of the owner's indirect costs and exclude the early phases and contractor's indirect costs from the analysis.

In this thesis, the owner is defined as the government who acts as a financier of the infrastructure project and owner's indirect costs are defined as the project management costs that are not part of the actual building costs. Project management costs include construction management, design and risk reserves of the project. Machine learning models applied in this thesis are based on data collected from various previous infrastructure projects implemented in Finland. Aim of these models is to predict the proportional values of indirect costs and examine whether machine learning models are able to perform better than the fixed default values. Ultimate aim of this thesis is to create a preliminary model for the upcoming cost management system and ensure the production of reliable cost data and accurate cost estimates. In ideal situation the model created in this study could be utilized in the research process of the earlier design phases as well. However, this paper does not take into account the possible effects of policy decisions. As it has been showed in earlier researches (Flyvbjerg et al. 2002) the political situation and the incentives of the policy makers do effect the cost overruns. Unfortunately, these effects are impossible to measure reliably and thus we have to exclude these policy effects from the analysis of this research.

This thesis will first examine on general level how the profitability of infrastructure projects is measured and how the cost estimation inaccuracies bias the profitability estimations. After this I will view the previous studies regarding application of machine learning models in cost estimation and how the use of machine learning methods can improve the cost estimation accuracy from its current state. Third section will discuss more detailed of the current system and instructions, which are written for the calculation of indirect costs in FTIA's projects. After this follows the data description and detailed explanations of how the models used in this study are constructed and how they function. Following the model descriptions, thesis will discuss how the performance of these models is estimated. Results of the study are addressed in two different subsections. First of these subsections represents the results of the construction management model and the second section focuses merely on the design cost model. The next chapter covers discussion of the results with respect to earlier similar studies and short consideration how the preliminary forecast models provided by this study could be improved. Final section summarizes the overall conclusions we can draw from this study and discusses possible topics for further research.

2 Literature review

2.1 Measuring the profitability of infrastructure projects

CBA is one of the most common tools used in a decision-making process (Munger, 2000; G´omez-Lobo, 2012). Fundamental idea of the CBA is to evaluate whether the project should be implemented or not by comparing the benefits and costs of the project. Simply put, the goal of CBA is to achieve efficient allocation of scarce resources (World Bank, 2004). Several institutions and governments have set demands for the use of CBA in the evaluation of infrastructure investment's profitability. For example, if a European Union (EU) member country wishes to receive funding from structural funds or cohesion fund, CBA must be carried out before applying the subsidy (Jones, Moura & Domingos, 2013). CBA is calculated by discounting the costs and benefits of a certain project into a present value. Net present value (NPV) is generally represented as:

$$\mathrm{NPV} = \sum_{t=0}^{\infty} \frac{B_{t-C_t}}{(1+r)^t}, \qquad (1)$$

where *B* represents the benefits of the project, *C* represents the costs of the project and *r* denotes the associated discount rate. Challenge in this otherwise simple equation is defining the exact values for *B* and *C*. Generally, the interest of economic research has been focused on issues related to the measurement of benefits like determination of monetary value for human life. Although the precise evaluation of benefits is crucial, also the accurate evaluation of costs is important in order to receive non-biased results from CBA. (Jones, Moura & Domingos, 2013)

As acknowledged in previous literature, CBA can also be used inappropriately (Flyvbjerg, Skamris Holm & Buhl, 2003; C´omez-Lobo, 2012). For instance, public institutions and politicians may disregard the results provided by the CBA in order to pursue their own political interests (Flyvbjerg, Skamris Holm & Buhl, 2003). Furthermore, there have been a few known cases in history (Persky, 2001) where the estimates of benefits have been manipulated in order to achieve the desirable outcome. The inappropriate use of CBA ultimately leads to the inefficient use of public funds. Former literature in this matter has tackled the question of how the correct use of CBA could be guaranteed (C´omez-Lobo, 2012). For simplicity, this paper assumes that CBA is used appropriately as a decision-making tool.

Besides the possibility of manipulating the results of CBA, there are several other weaknesses related to the analysis. Vickerman (2007) have shown that the estimation of CBA in larger mega projects is harder and inaccuracies tend to happen more often as the size of a project increases. As the amount of mega projects is constantly increasing, the importance of accurate CBA results is growing larger. In order to reply to this change happening in infrastructure industry, the structure of CBA needs to change rapidly. Jones et al. (2013) researched the detected factors effecting the failure of CBA. Weaknesses recognized by Jones et al. (2013) are related to factors such as overlooking the residual value, safety, value of time, environmental impacts and traffic forecast. In addition, one of the major weakness identified is the cost estimation inaccuracy, which Jones et al. (2013) have define as a common problem in the use of CBA. Effects and reasons behind these cost over- and underruns have been examined closely (Skamris & Flyvbjerg, 1997; Flyvbjerg, Skamris Holm & Buhl, 2003; Mayer & McGoey-Smith, 2006).

Nevertheless, CBA is still widely recommended as an evaluation method and it is the generally accepted tool to evaluate the profitability of a single project. Although, researchers are advised to take these limitations mentioned above into account (Wee, 2012; Jones, Moura & Domingos, 2013). Therefore, in order to develop CBA further, aim should be in repairing detected flaws. The next chapter will examine more generally the effects of cost overruns examined in earlier literature and evaluate their significance to the economic development.

2.1.1 The effect of cost estimation inaccuracy

It is clear that the inaccuracy of cost estimation biases CBA and the current cost estimation system is not able to control this bias. This section will overview the consequences of cost estimation inaccuracy and explain how previous studies have studied this problem. As Flyvbjerg et al. (1997) state in their research paper, the most common overruns in cost estimations of infrastructure projects are between 50 to 100% and there are tendency even for overruns above 100%. So far, this trend of large overruns has not taken any turn and in the case of large overrun, the result gained from CBA will be deceptive (Jones, Moura & Domingos, 2013).

Study of Flyvbjerg et al. (1997) revealed that from 41 projects 32% of them fell between 50-100% in overruns. As the implementation decisions are based on the cost estimations, these inaccuracies inevitably lead to the misallocation of government's funds (Flyvbjerg, 1997). Furthermore, Flyvbjerg et al. (2002) note that there are cost estimation inaccuracies in nine projects out of 10. Cost overruns are typically large and Flyvbjerg et al. (2002) estimated that average for all cost escalations is approximately 28%. However, there are large differences on the average cost escalations between project types. For instance, cost escalations for rail projects are on average 45% on average and for road projects, this same figure is only 20%. These are alarming proportions of the infrastructure investments and as Flyvbjerg et al. (2002) indicate, there has been no progress in decreasing the amount of cost overruns. Most worrying part is that the amount of large mega projects constantly increasing. This means that in the long-run the absolute amount of cost overruns will increase substantially if we do not learn how to control them.

Cantarelli et al. (2010) limits the explanations for cost escalations in four categories: psychological, political, economical and technical. Psychological reasons are often related to optimism bias, which creates the illusion that construction project would be cheaper than it actually is. On the contrary, cost inaccuracy caused by political reasons, is often deliberate and it sometimes involves the manipulating of the forecasts. Economical explanations are related to insufficient resources and overall bad financing. The focus of this thesis is in the last category, technical explanations, which is considered as the most influential explanatory factor for cost overruns (Cantarelli, Flyvbjerg, Molin & Wee, 2010).

There are several technical reasons why costs are escalated in infrastructure projects. Delays, uncertainty, increases in market prices and inadequate data are only few examples of these (Cantarelli, Flyvbjerg, Molin & Wee, 2010). It is important to be able to separate, which risks related to cost escalation are the most harmful. Mayer et al. (2006) made a risk-ranking chart for costs. It shows how important certain risks are for overall costs. This chart includes risk events like the delays caused by a bad weather, issues related to hauls and relocation of the borrow pits. However, in the first place of this risk ranking is the uncertainty of soft costs. Soft cost are the G&A costs of the project. In other words, these are the costs, which are not directly related to the project's building. The mitigation of the uncertainty in soft costs is necessary, because it produces the greatest risk for cost escalation. This thesis specifically aims to decrease the uncertainty of these soft costs and that way reduce the cost estimation inaccuracy.

The policy implications done in the previous literature have been straightforward. According to the previous studies, policy decision makers should focus more on the length of implementation phase and cost controlling rather than to the political view. In addition, the developing of less deceptive and more precise cost estimates is considered as a key factor in solving problems related to cost escalations (Flvybjerg, Skamris Holm & Buhl, 2002). In the following chapters, this thesis will familiarize with few possible solutions for cost estimation problems and evaluate how well these models have functioned in the previous studies regarding similar cost estimation problems.

2.2 Cost estimation with machine learning models

Regardless of the significance of accurate cost estimation, performance of the current cost management system is poor. The main reason for inaccurate results is the problematic and complex nature of cost estimation. Central problems with cost estimation are the lack of existing database and absence of adequate estimation methods. Estimation is challenging especially because models forecasting costs need to be able to handle incomplete information (Bouabaz & Hamami, 2008). Research papers focused on improving the cost estimation in construction projects, have applied several different forecasting methods to achieve better cost estimates (Kim, An & Kang, 2004; Bouabaz & Hamami, 2008; Bayram & Al-Jibouri, 2016).

Besides the more traditional cost estimation applications, the use of machine learning models has increased rapidly as a tool for predicting construction costs. For instance, multiple regression model, artificial neural network model and case-based reasoning model have been utilized in previous cost estimation studies (Kim, An & Kang, 2004). Primary reason for applying machine learning models is that they are assumed to increase the accuracy of the cost estimation forecasts substantially. The advantage of machine learning is based on the idea that exact structure of the model is chosen by the data. In other words, the model is directly based on the statistics and information offered by the available sample set (Mullainathan & Spiess, 2017). There are various different machinelearning models with diverse strengths but in this thesis we will compare the suitability of MR and ANN models as the prediction methods for indirect costs.

2.2.1 Multiple regression model

Regression models have been popular in cost estimation especially because of their strong mathematical background and ability to evaluate the suitability of a curve to a specific data set (Kim, An & Kang 2004; Sodikov 2005). Especially MR models have shown adequate accuracy in cost estimation problems and regression analysis is convenient in detecting the dependencies between variables. However, the accuracy level of a regression model is often smaller compared to other machine learning models (Sodikov. 2005).

In their study, Kim et al. (2004) measured the performance of different cost estimation models in forecasting the construction costs. The study shows, that the mean absolute error rate (MAER) for multiple regression model is higher than for the neural network or case based reasoning model. MAER measures model's percentage error between actual and estimated values and higher value of MAER implicates lower accuracy of the model. In Kim et al. (2004) study, MAER for multiple regression model was 6.95, for best neural network model 2.97 and for case based reasoning model 4.81. Accuracy level of MR analysis is clearly the weakest one of these models, but the error is still reasonably small and therefore it is able to produce sufficiently accurate results if the use of other models is not possible. In addition, on contrary to ANN model, MR model is able to explain what the coefficients for independent variables are.

Other studies have also indicated that the multiple regression model performs adequately with cost estimation. Sodikov (2005) examined how the accuracy of cost estimation in the transport infrastructure projects of developing countries could be improved. Study compared the performance of ANN and MR model in cost estimation of Poland and Thailand's road projects. From the Poland data set, 38 projects were included to the final model and from Thailand data set, the number of included road projects was 42. In Sodikov's (2005) study the mean magnitude of relative error for MR model was 36% for Poland and 30% for Thailand. Considering Flyvbjerg's et al. (2002) study, where it is indicated that most of the cost overruns fell between 50-100%, the accuracy appears to be better in Sodikov's (2005) MR model. However, in Flyvbjerg et al. (2002) study it was also acknowledged that the average cost estimation error in all projects is 28%. Compared to this the multiple regression model does not perform well enough. By separating the road and rail projects to two different categories, the multiple regression model used in Sodikov's (2005) study would be suitable for estimating the costs of rail projects, as the average error size for all rail project is 42%.

More recent study of Bayram and Al-Jibouri (2016) compared the performance of traditional cost estimation methods to the performance of nontraditional methods such as neural network model and regression analysis. Aim of this study was to estimate which methods can produce the most realistic predictions of the final costs of building projects. Final costs were represented in the form of three different models: unit area costs, client detailed costs and contract sums. Sample size was larger compared to studies estimating infrastructure projects as 420 finished public building projects were used to establish the models. Bayram and Al-Jibouri (2016) evaluated the performance using four different measures such as root-mean-square error (RMSE) and mean absolute percentage error (MAPE). On contrary to other studies, Bayram et al. (2016) concluded that regression analysis is able to produce very accurate results in more detailed design stages of the project. Although the MAPE values for regression analysis were relatively high, between 20% and 60%, small RMSE values supports the fact that model actually performs well. For the client detailed costs model, the performance level of regression analysis was the highest.

Although, multiple regression model is a popular and simple tool for predicting, it has also received considerable amount of criticism. As former literature has stated, MR model is not able to help choosing the right cost model fitting best to the available data and the model is not appropriate for describing nonlinear relationships (Kim, An & Kang, 2004; Sodikov 2005). Furthermore, Kim et al. (2004) note that there is certain expectations for the data before it can be fitted to the regression model and variables used in the model need to be evaluated before using them in analysis. In many cases, this is not possible and therefore the use of regression model is challenging. In addition, MR model has problems working with high dimensional models where the number of input variables is relatively large (Kim, An & Kang, 2004).

Problem with multiple regression model is often fulfilling all these required assumptions. Luckily, neural network models can fight with these problems more effectively (Smith & Mason, 1996). Due to the problems related to multiple regression and especially to the prediction accuracy of this model, this thesis will also study the artificial neural network model as an alternative for traditional multiple regression models.

2.2.2 Artificial neural network model

As noted in the previous chapter, multiple regression requires a lot of knowledge about the data beforehand the analysis and the estimation accuracy in more complex problems is often poor. Data for cost models is typically incomplete and therefore attempt to build the best possible model is based on this inaccurate information (Smith & Mason, 1996). One solution to overcome this problem has been replacing multiple regression model with artificial neural network models. The results of previous studies indicate that ANN models are superior compared to other cost estimation models in the terms of accuracy (Kim, An & Kang, 2004; Sonmez, 2004; Sodikov, 2005; Sonmez & Ontepeli, 2009). In their study, Kim et al. (2004) compared how well three different models could predict future costs. In this study, it was observed that the performance of ANN model with MAER of 2.97 was superior compared to the accuracy of case based reasoning model and a standard multiple regression model. ANN model aims to learn like a human brain and this makes it useful tool in high dimensional cost estimation problems where the number of input variables is large (Kim, An & Kang, 2004). The suitability of ANN models to construction projects and cost estimation problems have been researched thoroughly during past twenty years (Boussabaine, 1996; Bode, 2000; Bouabaz 2008). Waziri et al. (2017) notes that the use of ANN models in recent studies has increased because its ability to forecast cost estimates with high accuracy.

Bouabaz et al. (2008) studied how well ANN could model the exact construction costs of bridge repairing. Results of the researches clearly implicated the superior accuracy of ANN models in construction cost estimation. Regardless of the sparse data used in the study, model was able to achieve accuracy of 96% for relatively small sample size including 40 bridge prepairing projects.

Mean error of the best ANN model was -0.25% and the maximal errors shifted between -6.63% and 5.52%. One of the main achievements of this study was to show the benefits of ANN model compared to regular cost estimation systems. Main conclusion drawn from the study was that well designed ANN model is able to overcome large number of uncertainties, which are common especially in the early stages of construction projects and produce prediction results of high accuracy. When considering the cost management system used in Finland, ability to forecast cost in early stage of project would be an enormous improvement. In addition, results gained from Bouabaz et al. (2008) study, appear to agree with earlier research papers. Comparison of the Bouabaz et al. (2008) results with the study of Kim et al. (2004), where the mean absolute error rate for the best neural network model was 2.97, shows clearly that the ANN model produces very accurate results even for highly complex problems with small sample size.

Similarly to earlier research, Sodikov's (2005) study also indicates good performance of ANN models. In this study, mean magnitude of relative error for the smaller data set, which includes 38 highway projects, is 24% and for the larger data set including 42 projects, the relative error is 26%. With these results, ANN outperforms the multiple regression model's equal outcomes of 30 and 36%. ANN model is also able to achieve more precise cost estimations compared to the current estimations and therefore it could decrease the amount of cost overruns rapidly. Furthermore, Sodikov (2005) has categorized the mean error rates so that less than 25% is considered as a good score and error rates between 25-50% are considered as fair. Errors over 50% are considered poor and therefore these should not been taken into account. This result furthermore indicates the good performance of ANN model. Furthermore, Sodikov (2005) notes that MR model is a helpful tool in the building of ANN model because it is able to identify the most central variables and therefore able to limit the variables used in ANN.

Polat (2012) studied how the contingency costs of a project can be forecasted with machine learning methods. Study was executed applying 195 building construction projects. This study applied several ANN models with different number of neurons in the hidden layers as the research methodology. The most accurate model was ANN with 3 hidden layers and 4 neurons in the hidden layers. MAPE for this model was 9.13% and RMSE 0.87. As the RMSE measures the standard deviation of the residuals, the error range is 0.87 from the actual value. Models with large amount of neurons inside the hidden layers produced even better results but these are invalid as this leads to the overfitting of the model Main conclusion of Polat (2012) is that ANN is able to predict the amount of contingency costs satisfactorily. However, it is important to note that this research does not compare ANN model with any other existing models and therefore the relative performance ability is unknown.

Furthermore, Bayram and Al-Jibouri (2016) concluded that neural networks works best for the post-design stages. In their study, the neural network models where superior for the contract sum model. MAPE values for the best models were between 10% and 35% and similarly RMSE values were also relatively small meaning that the model did not produce any significant over- or underruns. Benefit of the model is the possibility to change the number of hidden layers, which enables the model to adapt to the existing conditions. Although, neural networks worked better than the regression analysis for most of the models, it should be noted that both of the models were considerably better than the traditional method. However, Bayram and Al-Jibouri (2016) caution that the use of neural networks requires more input parameters than other models and applying the model in practice is more difficult.

ANN model has also been widely criticized for its problem with interpretability (Smith & Mason, 1996; Kim, An & Kang, 2004; Bayram & Al-Jibouri, 2016). Compared to multiple regression model ANN model is harder to interpret because of its black box feature. This means that the exact values of predictors stay unknown and therefore it is extremely hard to identify which predictors are the most important ones. In multiple regression models, this dependency is easy to find and interpret. Due to the parametrized structure of ANN model, it is able to produce results with outstanding accuracy but the interpretability of the results is weak and correlations between independent and dependent variables are harder to observe compared to MR model. Bayram et al. (2016) also noted that one problem with ANN models is their need for several input parameters compared. High dimensional models are in danger of overfitting the model and it leads to the model's poor generalization ability. In other words, the model would be incapable to forecast costs outside the sample set.

Although applications of ANN bare some faults, the use of these models will most likely only increase in the future because of its superior level of accuracy (Smith & Mason, 1996; Kim, An & Kang, 2004). Furthermore, the ability to handle incomplete data makes the ANN model suitable especially for cost estimation problems as Kim et al. (2004) and Zwainy et al. (2017) highlight.

3 Background of the current system

Currently used cost management instructions in FTIA's projects are described in the Liikennevirasto (2013) guide. This chapter will only examine the main features of the whole cost estimation calculation process and merely the calculation process of indirect costs is explained in more detail as these costs are the main interest in this thesis. Another objective of this chapter is to highlight the obvious problem related to the current calculation method for indirect costs.

The cost estimations in FTIA's infrastructure projects are calculated by using Fore –cost management system. Fore consist out of four different applications: project part calculation (Hola), construction element calculation (Rola), project program (Scope) and project management –data analyzing (Arena). All costs of the project are calculated using Hola and Rola applications. Hola is an application that contains all methods and price lists, which are used in the determination of a single objective's price. With Rola, the planned costs can be measured up against project's objectives and it is used to perform cost assessment of plans. With the help of these applications, Fore –system calculates the preliminary cost estimation. After calculating building costs, the indirect costs are added on top of the building costs using default values specified in advance. For instance, in FTIA's projects owner's and contractor's indirect costs are estimated to be 20-50% from the building costs depending on the planning phase. (Liikennevirasto, 2013)

Calculation of the owner's indirect costs in Fore is merely directional guideline, which is based on the proposition suggested in Liikennevirasto (2013) instructions. The instructions give default values for each category of indirect costs depending on which design phase of the project is currently running. Indirect costs are thus instructed to calculate with the same percentage in every infrastructure project and most of the project properties are disregarded completely in the estimation process. However, the user of Fore is able to modify the percentages suggested by the instructions. Problem is that before this study there existed no measured values for the actual indirect costs and therefore the instructions do not include any suggestions whether specific qualities of the project require increasing or decreasing of the proportion of indirect costs.

Table 1: Owner's indirect costs in the construction phase

	Owner's tasks	**Construction plan**
5650	Design during implementation	1
5710	Construction management	4
5730	Other owner's tasks	1
5761	Reserves	5
	Total	**11**

Existing research indicates the importance of considering the properties of the project in order to achieve cost estimations that are more accurate. In the study of Flyvbjerg et al. (2003), it was discovered that there is a significant difference in costs between projects with different characteristics. Table 1 is adapted and translated from Liikennevirasto (2013). It describes the proportions for owner's indirect costs in the construction phase of the project. Column on the left represents denominations which are consistent with the official infrastructure nomenclature. Middle column includes the names of the denominations. Contents of design tasks and construction management will be explained in detail in the following chapter.

Forecasting the proportion of other owner's tasks and reserves are excluded from this study. Especially reserves are complicated to measure and thus the accuracy of forecast would be hard to estimate. Reserves are commonly defined as costs estimated in advance for surprising expenses that might occur during the project implementation. Reserves include changes in price level, risk reserves, such as business or insurance risk, and the additional work caused by unexpected circumstances. Although the risk reserves are important part of the owner's indirect costs, the data related to these costs is impossible to gather from the existing system and therefore this study will primarily focus on design and construction management costs. (Rakennustieto, 2015)

Other owner's tasks include some simple project related tasks, which are included in the owner's duties as the financier of the project. The right side column describes the percentage proportions of indirect costs out of the overall costs and the primary aim of this study is to produce more reliable methods to calculate indirect costs than these currently used default values.

Examining the table 1 shows that the proportion of indirect costs during implementation phase are 1% for the design costs and 4% for the construction management costs. These default values for indirect cost are same for each projects in this particular phase. However, collected data includes some projects where also construction planning is included to the design costs. In these projects, the instructions estimate the design proportion to be 5% of the overall costs. Thus in these particular projects, we compare the actual values of design proportions to this percentage in order to obtain more reliable comparison.

Collected data already indicates that the proportions of actual indirect costs vary a lot in the construction phase depending on the project qualities and therefore the use of default values ultimately biases the final cost estimation. The level of accuracy of the current cost estimation method for the indirect costs is poor and the system is not able to produce enough exact results for the cost-benefit analysis. Goal of the new cost management system is to fix these estimation errors and produce more reliable cost data. Contribution of this thesis is to estimate the owner's indirect costs with more reliable statistical methods. New cost management system could increase the accuracy of the cost estimations, improve the reliability of CBA and eventually decrease the misallocation of resources.

4 Data

Data used in this thesis is received from FTIA. The data is divided into separate categories based on the official infrastructure nomenclature. Categories for owner's indirect costs examined in this thesis are design costs and construction management costs. The data covers merely the costs happened in the late phases of the project and thus costs related to previous design phases of the project are not taken into account in this study. As Skitmore (1987) showed in his study, the amount of information available related to each project is positively correlated with the accuracy of the construction project's cost estimation. Therefore, in this study it is essential to focus also to the information availability.

The category of the design costs include all the expenses, which are related to the designing of the project in any period during the project's life cycle. Additionally, design costs include following sub-categories: initial data for the plans, general design, design required by authority, construction design and designing done during the construction phase. All measurements and additional research done during project's life cycle are also part of the design costs. For instance, soil examinations are often required during the project to ensure the quality of the ground. The data used in this thesis covers merely the design done during the construction phase and in some projects also the construction design. In addition, the data includes costs for required initial data. Large part of the design costs are related to the research required to do the actual designing. (Rakennustieto, 2015; Liikennevirasto, 2013)

Second category is the construction management costs. These costs include the administration tasks of the project. Administration tasks include the surveillance of the construction, preparations related to the bidding and other administrative tasks of the project. Following the requirements set by the law such as the evaluation of the environmental effects is as well included to the construction management costs. Often these tasks are managed by the consultant hired for the project or in some cases the owner itself. In addition, all expenses related to the general project management are part of this category.
These expenses include the maintenance of project portals, traveling expenses required by the project and the use of required information services. In this study, also the communication costs are considered as part of the construction management costs. (Rakennustieto, 2015)

4.1 Description of the data

Data is constructed from information related to different types of infrastructure projects, which include road, waterway and railway projects. All projects included in the sample set are implemented in Finland and these projects are finished between 2015-2018. Older projects than this are not included as the data related to these projects is not comprehensive enough. Total amount of FTIA's projects finished during this time period is 283 but as this thesis is focused on the construction phase of the project, majority of these projects need to be excluded as these are not implementation projects. In addition, projects where it is impossible to separate the indirect costs from the building costs are not included to the sample. Thus, after validation of suitable projects the final sample size ended up to be 60 projects.

The sample sizes in earlier similar transport infrastructure studies have been relatively small and models constructed in these papers have included less than 50 projects and therefore we can assume that the use of similar models with small sample size should be appropriate in this research as well (Sodikov,2005; Bouabaz & Hamami, 2008). Moreover, the goal of this study is to suggest possible improvements for the current instructions and experiment the suitability of machine learning methods for the calculation of indirect costs rather than create perfectly functioning model. However, the limitations of small sample size should always be recognized.

Table 2: Description of the dependent variables

Dependent variable	Min	Max	Mean
Construction management	0.49	8.92	3.22
Design	0.04	10.04	3.13

In each project included in the data set, costs were sorted by the instalments. Project managers, engineers or consultants have divided these instalments into two categories mentioned above: construction management and design costs. Table 2 represents the minimum, maximum and mean values for design and construction management costs. Proportional shares of construction management costs and design costs from the overall costs are the dependent variables used in the models. To clearly distinct models for construction management and design costs, the models are addressed in this study in separate sections. In the case of design costs values vary between 0.04% and 10.04% and for construction management costs these changes are between 0.49% and 8.92% depending on the project. Mean value for design costs was 3.13% and for construction management costs the similar count was 3.22%. The proportions of construction management and design costs from the overall expenses are defined similarly to current instructions (Liikennevirasto, 2013). This enables the comparison of the current system and machine learning methods tested in this research. Proportion of design costs is calculated by subtracting the design and construction management costs from overall costs of the project and dividing the design costs with this result. Proportion of the construction management costs is calculated similarly but in this case the design costs are included as part of the overall costs (Liikennevirasto, 2013).

The values of the variables were received from FTIA's systems, report cards and with question forms sent to the project managers. Some of these variables such as contract type, project costs and project type have been used in similar cost estimation studies before (R´onai, 2001; Irfan et al., 2011). Other variables utilized in this study are the quality of land, number of contracts and the scope of the land. Large amount of independent variables is not uncommon as Emsley et al. (2002) noted that the number of variables related to cost estimation in construction projects is typically very large.

Table 3: Description of the independent variables

Description	Input variable	Value range
Overall costs	OC	296 355 to 208 235 216
Number of contracts	NC	1 to 110
Overall area	AREA	0.1 to 633
Quality of the land	LAND	Class I, Class II, Class III
Size of the project	SIZE	Small, Large
Area of the project	SA	Small, Large
Project type	PT	Road, Railway
Design & implementation	ST	St, Mixed, Implementation
Pure implementation	IMPLEM	St, Mixed, Implementation

Table 3 represents a summary of the independent variables. Number of contracts is the amount of contracts included in the projects implementation. Overall costs include building costs and indirect costs during the implementation phase. Quality of the land describes the environmental circumstances of the project. Grading from classes one to three is based on the evaluation given by project manager, engineer or consultant. Class one indicates that uncertainties related to the quality of the land are estimated to have small cost effects, class two indicates middle-sized effects and class three large effects.

Size of the project and project type are dummy variables which describe main characteristics of the projects. Label of the size is based on the overall costs. With construction management costs, the project is determined as small if the costs are less than two millions. With design cost model the project is small if the overall costs are less than five millions and large otherwise. Overall area is measured and denoted as kilometers. The area of the project is binary variable and project is determined small if the overall area is less than two kilometer and large otherwise. Contract types are divided into implementation contracts, design and implementation contracts and mixed contracts. Thus there is two different dummy variables to separate the different contract groups from each other.

Especially the MR model is sensitive for too many variables when the sample size is small and therefore the number of variables should be kept relatively small in these models and thus only the most significant variables will be chosen. ANN model does not have similar restrictions and several input variables only benefit the model. However, ANN model requires that all information is known and therefore input variables can not include any missing values. Only variables which increase the accuracy level of the model are considered to the model.

5 Methodology

As previous studies have shown, MR and ANN models work typically well with cost estimation problems. Thus, this thesis will apply these two models to forecast the indirect costs. The first part of this chapter will explain in detail how the models are formulated. The second part will focus on how the performance of the models is evaluated and how it is compared to the performance of the existing system. All model applications used in this thesis are formulated using R software environment.ö

5.1 Models

5.1.1 Multiple regression

Multiple regression model is often preferred as cost estimation model especially because of its ability to work with different problem types (Kim, An & Kang, 2004). Furthermore, multiple regression analysis is able to explain the significance of each variable and the dependencies between independent variables and dependent variable. This feature is valuable especially because ANN model does not enable it. Equation for multiple regression is commonly represented as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + u, \qquad (2)$$

where Y is the dependent variable, $x_1, x_2, \ldots, x_n$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, .., \beta_n$ are the coefficients estimated with the regression analysis and u represents the residual terms of the model. In this thesis we will apply multiple regression analysis because compared to simple regression with only one independent variable it gives higher accuracy for the cost estimates as multiple independent variables are able to explain bigger proportion of the variation in dependent variable (Sayadi, Lashgari & Paraszczak, 2011).

To determine the estimates of β_0 and $\beta_1, \ldots, \beta_n$ we will apply the ordinary least squares (OLS) method. To obtain the values for these parameters, we need to determine the estimation process. Obtaining the OLS estimates for multiple regression differs from the typical OLS estimation for single regression as the number of independent variables and subsequently the number of required estimates changes.

Now the objective is to choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1 \ldots \hat{\beta}_n$ such that the sum of squared residuals is minimized. Sum of squared residuals is represented as:

$$\sum_{i=1}^{n} \hat{u}^2{}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2, \qquad (3)$$

where the first subscript in independent variable describes the number of observations $i = 1,2,..,n$. The second subscript is merely a distinction measure to separate the independent variables from each other. The sum of squared residuals can be minimized by calculating the first order conditions for OLS estimates. In other words, we need to find the minimization problem for equation 3. To obtain the first order conditions for OLS we assume that:

$$E(u) = 0, \tag{4}$$

and

$$Cov(x_j, u) = E(x_{j}\,, u) = 0, \tag{5}$$

where *j* = 1,2,..,*k*. This means that the expected value of the residuals is zero and that the covariance between independent variable x_j and residual *u* is also zero. Now the minimization problem related to the sum of squared residuals can be written as (Woolridge, 2013):

$$\begin{aligned}
\textstyle\sum_{i=1}^{n}(y_i - \beta\hat{}_0 - \beta\hat{}_1 x_{i1} - \cdots - \beta\hat{}_k x_{ik}) &= 0 \\
\textstyle\sum_{i=1}^{n} x_{i1}(y_i - \beta\hat{}_0 - \beta\hat{}_1 x_{i1} - \cdots - \beta\hat{}_k x_{ik}) &= 0 \\
\textstyle\sum_{i=1}^{n} x_{i2}(y_i - \beta\hat{}_0 - \beta\hat{}_1 x_{i1} - \cdots - \beta\hat{}_k x_{ik}) &= 0 \\
\vdots \\
\textstyle\sum_{i=1}^{n} x_{ik}(y_i - \beta\hat{}_0 - \beta\hat{}_1 x_{i1} - \cdots - \beta\hat{}_k x_{ik}) &= 0.
\end{aligned} \tag{6}$$

Equation 6 formulates the first order conditions for OLS and these enable us to find the intercept estimate $\hat{\beta}_0$ and estimates $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$. Furthermore, we need to assume that these linear equations can be solve uniquely to $\hat{\beta}_j$ (Woolridge, 2013). With first order conditions we can minimize the sum of squared residuals and formulate the following OLS regression equation:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k, \tag{7}$$

Retaining the values for OLS estimates is straightforward but there is few disadvantages in the simplicity of MR model. In order to multiple regression function properly we need to make various assumptions on its properties. Firstly, the regression model assumes homogeneity of the residuals' variance, which means that the variance of the residuals is assumed to be constant. Homoskedasticity can be tested in various different ways. However, in this study we will apply heteroskedasticity robust standard errors, which allows us to leave out any assumptions on the structure of the heteroskedasticity. White (1980) proved how standard errors can be calculated while there is presence of heteroskedasticity. In practice, this means that instead of using the usual ordinary least squares standard errors, the model will calculate the heteroskedasticity consistent standard errors. Heteroskedasticity robust standard errors for multiple regression are given by (Woolridge, 2013):

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \mathrm{r}\hat{}^2_{ij}\hat{\mathrm{u}}^2_i}{SSR^2_j}, \tag{8}$$

where r_i is the residual received from regressing variable x_j on all other independent variables, $\hat{}u_i$ represents the residuals from the original regression of dependent variable on independent variable and *SSR* is the sum of squared residuals. Equation 3 describes how *SSR* is formulated. (Woolridge, 2013)

Secondly, multiple regression assumes normally distributed residuals. However, although MR model assumes the normality of residuals there can exists some deviation from normal distribution and thus the assumption of normality is not strict. We also need to assume that the data selected is representative of the whole population, which indicates random sampling. In addition we need to assume that the expected value of error term is zero with all independent variables $x_1,..,x_n$. This assumption can be satisfied by assuming that on average the independent variables are not related to the other unobserved variables. (Woolridge, 2013)

Next assumption which need to be examined in the case of multiple regression model is that there exists no multicollinearity. In other words, there should not exist any linear relationship between independent variables and furthermore there should be some variation in the sample of each of these variables. Finally, the regression model assumes linearity. Although the regression model itself should be linear, multiple regression does not assume that the relationships between underlying variables are linear. The linearity of the MR model depends actually on the linearity of the parameters b_n (Woolridge, 2013). This assumption means that the model representing the population can be written in the form of equation 2. All required assumptions are verified during model construction using appropriate diagnostic checks to ensure the validity of the results.

5.1.2 Artificial neural network

Compared to the simple multiple regression, ANN is more complex model and as earlier research has shown (Bayram & Al-Jibouri 2016; Kim, An & Kang 2004; Polat, 2012; Sodikov 2005) it usually produces more precise cost estimates. Furthermore, it does not have any restrictions concerning linearity of the variables. ANN models are used in variety of different tasks because of the model’s ability to adjust to different problem types. For instance, neural networks are utilized in forecasting oil prices or weather phenomenons and these are also used for complex pattern recognition tasks such as facial recognition. In order to approach the research problem in question successfully, the structure of the network need to be specified carefully. The following chapter will describe the detailed structure and functioning of the ANN model applied in this study.

In general, neural network models support several learning types such as supervised, unsupervised and reinforcement learning. However, supervised learning has been discovered to be most suitable with regression problems and thus this study will only focus on this particular learning type. Supervised learning means that the model’s learning task is to produce outputs from the chosen inputs. Thus input variable *x* forms non-linear descriptions of output $y = f(x;w)$. In this equation, *w* represents the parameters given to the associated neural network. (Babinec, 1994)

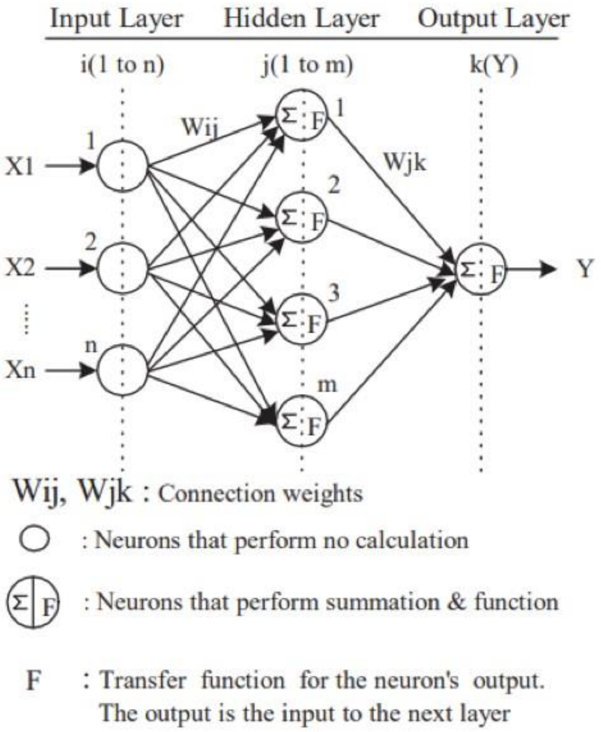
Input Layer
Hidden Layer
Output Layer
i(1 to n)
j(1 to m)
k(Y)
X1
X2
Xn
1
2
n
Wij
Wjk
1
2
3
m
Σ F
Y
Wij, Wjk : Connection weights
: Neurons that perform no calculation
: Neurons that perform summation & function
F : Transfer function for the neuron's output. The output is the input to the next layer

Figure 1: Source: Kim, An & Kang (2004)

Simply put, the intelligence of ANN model is based on its ability to recognize patterns in data and subsequently creating noiseless output (Bouabaz & Hamami, 2008). Fig 1. reflects the basic formulation of neural network model with only one hidden layer. The network is formulated with different levels of layers and neurons, which work inside these layers. The first level of ANN shown in the Fig 1. is the input layer, where the independent variables X_1,X_2....X_n are located. One of the many strengths of this model is that there are no restrictions for amount of inputs used in the model. In addition, input layer itself does not perform any calculation and its task is therefore merely to transfer the given inputs to the next layer.

The following layer, where input layer transfers the inputs is called the hidden layer. In Fig 1. W_{ij},W_{jk} represent the connection weights related to each neuron between the layers. There is typically one or more hidden layers and these layers process inside the model receiving and distributing inputs. Inside hidden layers the actual calculation takes place. Equation 9 represents the neuron's calculation which happens between the weights, independent variables and threshold. Task of the neurons is to calculate the sum of weighted inputs, subtract the threshold and transfer this into an output (Bouabaz & Hamami, 2008). Neurons calculate the output inside the layers using function:

$$y_i = f_i(\sum_{j=1}^{n} w_{ij}x_j - \theta_i), \tag{9}$$

where y_i is the output produced by neuron, w_{ij} is the weight, which connects to the input j, f_i is the hidden layer's transform function, x_i is the value of input variable and θ_i is the value of the threshold. The last layer in Fig 1. is the output layer where neurons calculate the actual output for the model.

Above description explains how the simple forward mechanism in the artificial neural network functions. The second step is to define the architecture of the model and how the model learns during the training process. Chosen architecture defines the training method on which the learning process of the network is based on. ANN model used in this thesis will be based on backpropagation architecture. Backpropagation is one of the most commonly used algorithm in neural network models and especially studies regarding cost estimation have applied this algorithm diligently. Backpropagation includes two steps. The first step is the forward propagation, which is the earlier described iteration through network that produces the output of the model. The next step is the backward propagation, which is the actual learning step of the process. Task of this phase is to reduce the error of the output by teaching the model to learn from the first iteration. (Wythoff, 1992)

In order to improve the model's accuracy from the forward propagation, backpropagation algorithm utilizes the loss function. Loss function simply denotes the difference between the outcome of the model and the targeted outcome. The task of backpropagation algorithm is to calculate the loss function and process the model backwards while calculating new values for each weight. Objective is to improve the model by training it to learn from its own mistakes and enhance the estimated output towards the actual output. (Kim, An & Kang, 2004; Wythoff, 1992)

After defining the error of the model, the values for the connection weights need to be redefined. Redefining of the new adjusted values for weights is task left for the activation functions. In other words, activation functions modify the output produced by the neurons and enhance the model output closer towards the actual value. There exists several different activation functions, which can be used in the task of adjusting weights and selection of the specific activation function determines how the backpropagation architecture will function. Backpropagation algorithm in this study uses Sigmoid transfer functions as activation functions, which are also utilized in former similar studies (Kim, An & Kang, 2004). Sigmoid transfer function is a logistic function which produces only positive outputs between 0 and 1. Sigmoid function is typically used for training data which similarly receives values only inside this range (Sibi, Allwyn Jones & Siddarth, 2013). In this paper, we will normalize the values of training data to fit this range. Normalization is done utilizing function:

$$j_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{10}$$

where x_i denotes the actual value of the variable for observation *i*, *max*(*x*) denotes the maximum value from all observations of *x* and similarly *min*(*x*) is the minimum value for the same group. Thus j_i represents the normalized value of the variable for observation *i*. After normalization, all our variables receive the value inside the range 0 to 1. With normalized data we are able to apply the Sigmoid transfer function. In this study Sigmoid transfer function handles the calculation for new values for weights. Sigmoid transfer function can be represented with two equations:

$$f(x_j) = 1/(1 + \exp(-(\sum_{i=1}^{n} w_{ij} \cdot x_i - \theta_i j))), \tag{11}$$

$$f(x_k) = 1/(1 + \exp(-(\sum_{k=1}^{n} w_{kj} \cdot x_j - \theta_k j))), \tag{12}$$

where Eq. (11) represents the output produced by hidden neurons and subsequently Eq. (12) represents the output produced by output neurons. Here w_{ij} is the weight connected to inputs *i* and *j*, x_i is the input variable and θ_i denotes the value of threshold. Use of backpropagation algorithm often increases the learning of ANN model and simultaneously decreases the error of the model by providing more accurate values for the weights. (Kim, An & Kang, 2004; Sodikov, 2005)

All different neural network architectures have certain learning rule on which the model's learning is based on. Learning rule associated with back-propagation algorithm is the generalized delta rule. Objective of generalized delta rule is to minimize the error of the model outcome. This is done by applying the loss function and activation function. Kasabov (1996) noted that the generalized delta rule defines regulation how the change of connection weight between neurons *i* and *j* is represented. Generalized delta rule can be written as (Kasabov, 1996):

$$\Delta w_{ij} = \eta[a_i - y_i]f'(x_n), \tag{13}$$

where η is the learning rate set for the model, a_i is the output produced by the network, y_i is the desired output and thus the term $a_i - y_i$ describes the loss function. Finally $f'(x_n)$, where $n = j,k$ represents the derivative of the associated activation function. If we adjust each weight with this rule, the value of each connection weight moves towards its own minimum (McClelland & Rumelhart, 1988). The system achieves its own equilibrium when all weights reach their minimum. Thus if the system is built properly, the error between output produced by the network and actual output is minimized. The learning process of backpropagation is done until it reaches the convergence. It is important to notice that backpropagation process can not be performed if the convergence is not reached and thus the model is unable to function (Ciaburro & Venkateswaran, 2017). This time consuming and complex training process of the ANN model often results as accurate predictions and thus the time consumed in the development of the model is exploited well.

After defining the model's architecture we need to set the values for each of the network's parameters. ANN model allows to experiment with different numbers of hidden layers, neurons and learning rate in order to find the most suitable model. In this study, the objective is to experiment with multiple different parameter combinations to find the most suitable model. However, experimenting with different combinations of neurons, hidden layers and learning rates takes considerable amount of time, which is according to Kim et al. (2004) one of the main issues related to ANN models as there exists no specific rules on how to determine the exact number of these parameters. However, some conditions for the selection process of the different parameters can be set.

Number of hidden layers effects directly on the complexity of the model. For instance, very high number of hidden layers produces accurate results for the training set but with the test set the results are poor (Kasabov, 1996). This indicates that the model is overfitting and thus the number of hidden layers need to be reduced. On the other hand, only one hidden layer indicates very simple neural network and it usually works best for simpler prediction problems. When building ANN model with small sample set, the number of hidden layers should be kept minimum in order to avoid overfitting as the model becomes more complex as more hidden layers are added into it. The exact number of hidden layer is nearly impossible to determine beforehand but the number of layers can be easily experimented with and experimenting enables us to find the correct number of hidden layers.

The optimal number of hidden neurons is similarly hard to determine without experimenting different combinations. However, some studies have suggested methods to help limit the range of the neurons. Kasabov (1996) suggested one way for calculating the minimum number of neurons used in the neural networks:

$$h \geq \frac{p-1}{n+2}, \tag{14}$$

where h denotes the number of hidden neurons, p denotes the number of training sample and n denotes the number of input nodes. The selected number of hidden neurons affects directly to the model's accuracy. Larger the number of hidden neurons, the more features network is able to model from the training data. However, if the number of hidden neurons is set too high the model takes substantial amount of time to go through the whole learning process (Kasabov, 1996). This suggested solution can be seen as a good limitation but the final number for hidden neurons needs often be determined by experimenting the model with different combinations.

Learning rate determines how much the weights are enhanced during the training process. Kasabov (1996) proposed that there is inversely proportional connection between the hidden neurons and learning rate. However, there is no commonly acknowledged customs on how to determine the optimal learning rate. Especially in regression studies where there is only one output neuron, the selection of the learning rate is ultimately often based on experimenting. Nevertheless, there is a clear trade-off, which need to be considered when choosing between small and large learning rate. Small learning rate converges the model toward more optimal weights but smaller learning rates take longer to train. If the learning rate is too small, the model does not converge and therefore it is not able to function. If the learning rate is large, it takes short time to train the model but the weights are not as optimal. Thus, the key is to find optimal learning rate where the training does not take too much time but the weights are as optimal as possible. (Kasabov, 1996)

The process of building the ANN model described above takes considerable amount of time but unlike the MR model, neural networks do not require verifying of various assumptions. Nevertheless, there are some concerns which we need to take into consideration. Fitness of the both MR and ANN models is important for the reliability of the models. Overfitting is the primary problem, which typically appears with machine learning models. This means that the model fits too well to the training set and therefore the generalization of the data is poor. Thus, the model forecasts training values accurately but the predictions for values outside the training set are poor. In other words, models overfitting have typically a large variance and it is mainly caused by the excessive complexity of the model. With neural networks this problem usually incurs when minimized error is the local instead of the global minimum (Kasabov, 1996). Another problem related to the generalization of the data is underfitting. Models, which underfit are too simple and thus not flexible enough to model the reality. This leads to the large bias of the model and therefore to bad prediction accuracy. (Babinec, 1994)

Luckily, poor fitness of the model can be tested and often removed if observed. In this study the goodness of fit will be tested with the root mean squared error measurement (RMSE) comparison between the training and test sets. If the model is observed to underfit the data, it is simple to fix by increasing the complexity of the model. In the case of clear overfitting, we have few methods to reduce it. For MR model easiest trick is to decrease the amount of independent variables which directly decreases the amount of overfitting. On the other hand, for ANN model there is several options to help to reduce the overfitting. Firstly, one can build a smaller network. This means that the number of hidden layers is reduced and thus there is fewer degrees of freedom in the model which ultimately leads to less overfitting. Second option is to utilize regularization, which can be done for instance with early stopping. Early stopping means taking part of the training set as test set and while running the model and calculating the errors for test set, we stop the iterations immediately when the error starts increasing. Third method to avoid overfitting is creating penalty function, which alarms when the sum of squared errors grows too large. (Babinec, 1994)

5.2 Evaluation of the model performance

Evaluation of the model performance is important because it gives us reliable knowledge how well the models are able to forecast indirect costs and whether the model's generalization ability is sufficient. There are several ways to measure the accuracy of the forecast models and in this thesis, the performance of the models is measured by using two standard error measures. First standard error measure is root mean squared error, which is given by:

$$\mathrm{RMSE} = \frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2, \qquad (15)$$

where A_t is the actual value, F_t is the predicted value and n is the number of observations. RMSE measures the standard deviation of the residuals and it is commonly used measurement method for model accuracy (Bayram & AlJibouri, 2016). Furthermore, use of RMSE is beneficial as it is valid indicator of the model performance. In this thesis the RMSE values between training and test set are also compared to each other in order to be able to detect critical overfitting or underfitting of the models.

Second measurement tool is the mean absolute error (MAE). It measures the average amplitude of the prediction error and it is useful tool in interpreting the results of the model as it gives good perspective how accurately the model is able to predict the actual costs. MAE is calculated in the following way:

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|, \qquad (16)$$

Calculation of these error values is simple but they give reliable and generally approved results on estimating the model accuracy (Bayram & Al-Jibouri, 2016). Results provided by regression model and ANN will be compared to current system's MAE and RMSE values. MAE and RMSE values for the current system will be calculated by assuming that the user utilizes the default values defined by the system. This comparison between models formulated in this thesis and the current system allows us to argue, which method should be applied in the cost estimation process. Furthermore, it allows us to examine how well the current cost management system is able to forecast the costs.

However, the estimation of the model performance can not be left solely to the simple error measure tools. It is also important to be able to evaluate the interpretability of the models. As these results should be part of the new cost estimation system, the knowledge of how the results for indirect costs are gained is vital. With multiple regression model, interpreting of the results is effortless. The functioning of the model and values for the parameters b_n are known, which increases the credibility and reliability of the model. On contrary, ANN model is much more complicated to interpret as the neural networks are so called black box models. In other words, the individual effect of each independent variable is unknown, which makes it difficult to explain how each input variable impacts to the output. Fortunately, this tension between the model accuracy and interpretability has been studied throughly during past few years.

For instance, Lundberg et al. (2017) studied the use of shapley additive explanations (SHAP) as the interpreting method for more complex machine learning models. Idea of SHAP is that it determines values of importance for each variable. These values are able to explain a relative importance of each feature and therefore eases the interpreting of the model. The use of SHAP and other tools is, however, complicated and the benefit from using these models in this study would probably stay relatively small. Thus this study will utilize the results from multiple regression to interpret the functioning of ANN.

6 Results

This chapter represents the results of the empirical study and reviews the findings from model comparisons. The results are viewed in two sections in order to separate the models for construction management and design costs.

6.1 Construction management costs

The final model for construction management costs included 48 projects from the 60 project sample. Data set was divided into training set which includes 80% of the projects and into test set which included the rest 20% of the projects. The data was divided by random number, which is determined with random selection and this is made to assure that the results are as credible and reliable as possible. Test set was left aside from the training phase of the model to enable us to measure the ability of the model to generalize outside the training sample. Model's ability to generalize was measured by examining the accuracy of prediction results and by comparing RMSE values between the training and test sets. The performance ability of MR and ANN models were tested 10 times by dividing the sample set with random number into training and test data. Objective of repeating the test was to examine how different training and test samples affect the prediction accuracy and whether the deviations between different tests stayed small.

MR model was first experimented using different variable combinations to ensure that the final model includes only statistically significant variables. In the final MR model, six statistically significant variables were identified. The OLS regression equation related to the MR model can be represented as:

$$\widehat{CM} = \widehat{\beta_0} + \widehat{\beta_1}PT + \widehat{\beta_2}SIZE + \widehat{\beta_3}ST + \widehat{\beta_4}IMPLEM + \widehat{\beta_5}NC + \widehat{\beta_6}log(AREA), \quad (17)$$

where $\hat{\beta}_0$ represents the OLS intercept estimate and $\hat{\beta}_1,\ldots,\hat{\beta}_6$ represent the computed OLS estimates. From these variables project type, size of the project and contract type are binary variables. Contract type is divided into design and implementation contracts, pure implementation contracts and to mixed contracts, which include only projects that have applied both contract types. Project type is divided into road and railway projects. Variable gets value 1 when it is road project and 0 otherwise. Size of the project is divided into small projects with overall costs less than two million euros and into larger projects with overall costs over or equal to two million euros. Small projects are notated with 1 and large projects with 0. Area of the project and number of contracts are modeled as numerical variables. To fit the area variable better to the model, the variable is altered with logarithmic transformation.

The summary of the regression model in table 4 shows us that all independent variables denote significance at 5% level. The standard errors represented in parenthesis are calculated as heteroskedasticity robust standard errors. The R-squared value for the model was 0.74, which means that the independent variables are able to explain 74% out of the variation that happens in construction management costs. Correspondingly adjusted R-squared in our model equals 0.69. Following merely the R-squared in MR analysis can *lead* to false conclusions as the R-squared value never diminishes when additional variables are added. Furthermore, the R-squared is not able to define whether predictions or estimates of the model are biased. On the contrary, adjusted R-squared only increases as significant variables are added to the model and thus it is better indicator of single variable's significance in MR model than the R-squared as adding insignificant variables to the model leads to a decrease in adjusted R-squared.

The intercept of the model was at 5.38%. If the overall costs of the project were less than two millions, the project was determined as small and thus construction management costs increased by 2.8 percentage points. Table 4 also shows that road projects decrease the construction management costs by 1.16 percentage points compared to railway projects. In the case of contract types, the design and implementation contract decrease the costs compared to the mixed contract type by 2.3 percentage points. If the contract type is implementation contract, the proportion of construction management costs shrink by 1.8 percentage points compared to mixed contract type. Thus we are able to deduce that the construction management costs are higher for pure implementation projects than for projects that are executed as design and implementation projects.

Table 4: Regression model summary

	CM model
(Intercept)	0.0538^{***} (0.007)
PT	-0.0116^{**} (0.006)
SIZE	0.0279^{***} (0.007)
ST	-0.0257^{**} (0.011)
IMPLEM	-0.0173^{**} (0.007)
NC	0.0009^{***} (0.000)
log(AREA)	-0.0036^{**} (0.002)
R_2	0.74
Adj. R^2	0.69
RMSE	0.01
P-value	1.16e-07

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 4 also indicates that number of contracts is statistically highly significant variable as it denotes significance at 1% level. Number of contracts is positively correlated with construction management costs. If the number of contracts increases by 1, the share of construction management costs increases by 0.09 percentage points. Conversely, the correlation between construction management costs and area of the project is negative. Large area of the project thus implies smaller construction management costs. Because of the logarithmic transformation done for the area variable, the increase of area by 10% decreases the construction management costs on average by 0.0036 percentage points.

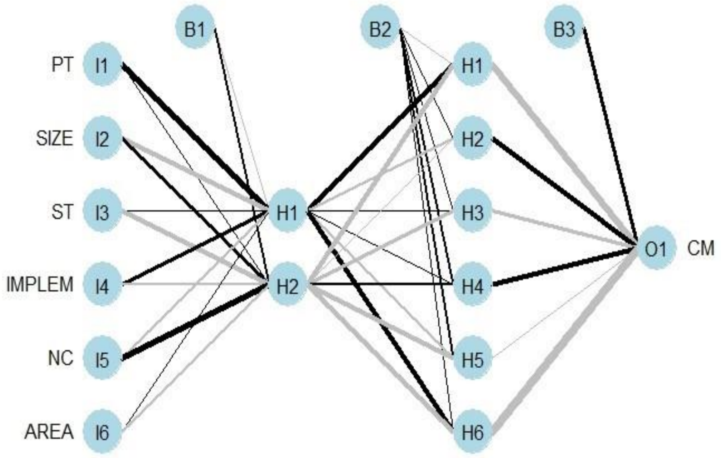

Figure 2: Neural network structure for construction management costs

The results provided by MR and ANN models were very similar although the structure of the models differs notably from each others. Figure 2 represents the structure of the artificial neural network used to forecast the construction management costs. I1-I6 represent the input variables, H1-H2 represent the hidden neurons, B1-B3 represent the intercept weights attached to each hidden neuron and output neuron and finally O1 represents the output produced by the model. Lines between the neurons in fig 2. describe properties of the connection weights. Thickness of the line represents the magnitude of weight and the color of the line reveals whether the weight sign is positive or negative. In fig 2. black lines represent positive weights and grey lines represent the negative weights. Neural network was experimented with 50 different combinations changing the number of layers, neurons and learning rate. The final model was chosen based on it's prediction accuracy. The figure shows that in the model we have 6 input variables, two hidden layers with 2 hidden neurons inside the first hidden layer and 6 neurons inside the second layer. In addition, the learning rate of the model was set to 0.1 and threshold for the error functions' partial derivatives was set at 0.01, which operates as stopping criteria for the model.

The model can be described by its number of neurons inside each layer. In our case the best model 6-2-6-1, outperformed all other models experimented in this study. Other models produced higher MAE and RMSE values and thus the generalization ability outside the training data was poorer. Variables used in ANN model are same as the ones used in MR model. MR model gives good insight on which independent variables affect the dependent variable and therefore it helps in the construction process of ANN. The model was tested with other variable combinations as well but the accuracy of the model only decreased. Model also utilized backpropagation algorithm to adjust the weights and to achieve better prediction accuracy.

Table 5 summarizes the connection weights between each neurons in different layers. As the first hidden layer includes only two neurons, the original input variables have weight attached to these two neurons. H1-H6 denote the weight connections between neurons. For instance, project type has two weights attached to neurons in the first hidden layer with values of 2.17 and 2.50. Thus the neurons in first hidden layer, H1 and H2, have respectively weight connections to each six neurons located in the following hidden layer. Finally, the neurons in the second layer are all attached to the output layer with different weights.

Table 5: Summary of network's connection weights

Parameter		Weight					
		H1	H2	H3	H4	H5	H6
Input layer	PT	2.17	2.50	-	-	-	-
	SIZE	-1.57	1.47	-	-	-	-
	ST	0.31	-1.98	-	-	-	-
	IMPLEM	1.07	-0.70	-	-	-	-
	NC	-0.86	2.22	-	-	-	-
	AREA	0.23	-0.91	-	-	-	-
Hidden layer	H1	-0.29	-0.91	0.20	0.06	-0.69	2.01
	H2	-1.56	-0.08	-1.17	0.83	-1.81	-1.98
Output layer	CM	-2.67	1.95	-1.23	1.78	-0.28	-3.04

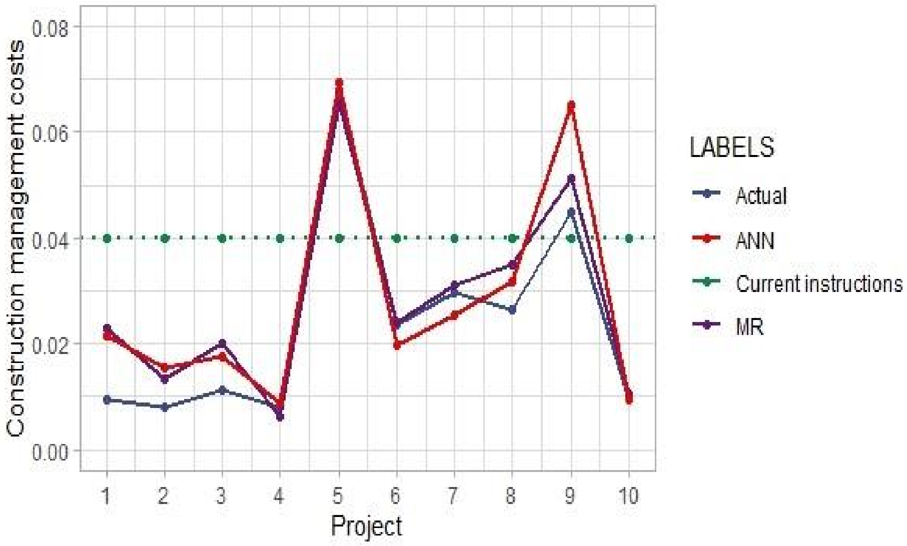

Figure 3: Prediction results for construction management costs

After building the models it was necessary to test how well the models are able to forecast indirect costs with data excluded from the training. To test the generalization ability of the model, I utilized the test sample, which was intentionally left aside from the training phase of the model. By doing this, it is possible to evaluate how well the model could predict construction management costs in reality. Actual values of the test set were compared to the predicted values produced by ANN model, MR model and the current system. Fig 3. represents prediction results for the test set with the most accurate ANN and MR models. Each point in the graph describes one executed project and its construction management costs as a proportion of the overall costs. The picture indicates that MR and ANN models are able to predict the results with notably higher accuracy compared to the currently used default values. Furthermore, figure indicates that the actual values for construction management vary considerably between different projects and thus default values determined in current instructions are not able to reflect the reality well. Figure shows that in all the test cases at least one of the forecast models is able to predict the construction cost better than the default values. Comparing the results of ANN and MR curves, we detect that in most of the projects the predicted values are very close to the actual values producing relatively high accuracy level.

Table 6: Error measurements for construction management costs

Model	MAE	RMSE	Min MAE	Max MAE
ANN	1.07	1.52	0.61	1.30
MR	0.89	1.16	0.47	1.15
Current instructions	2.21	2.40	1.72	2.53

Table 6 represents MAE and RMSE values for test samples of ANN and MR model. Mean values are calculated for the ten randomized tests performed. Table includes as well the corresponding values for default values imposed by the current instructions. MAE and RMSE values in table 6 are presented as percentage points to ease the interpretability of the results. Smallest MAE for MR model was 0.47 percentage points, highest MAE for tests was 1.15 and the mean MAE for all test was 0.89. For currently used instructions the measured error were notably higher as the mean MAE was 2.21 percentage points, smallest MAE value was 1.72 and highest 2.53. In addition, table 6 reveals that ANN model did not perform in the end as well as the MR model. The mean MAE for all tests was 1.07 and values of MAE varied between 0.61 and 1.30. The mean RMSE for MR model was 1.16, for ANN model 1.52 and for the current instructions it was 2.43. Differences between MAE and RMSE values stayed relatively small and it is important to notice that the RMSE is always higher compared to corresponding test set's MAE. Although the difference between MR and ANN model's prediction accuracy was small, both models outperformed the current instructions distinctly.

Table 7: RMSE values for training and test sets

Model	Training RMSE	Test RMSE
ANN	1.28	1.52
MR	1.06	1.16

Table 7 represents the mean RMSE for training and test sets for both models. RMSE values were measured for the training and test set of the model separately to help recognize whether there exists any notable and critical overfitting or underfitting of the model. If RMSE value is low for trainining set and high for test set, it usually implicates that the model is overfitting. On the contrary, if RMSE value is high for both training and test sets it usually implicates that the model is underfitting. Low RMSE values for both categories indicate that the model is not overfitting or underfitting and thus it's ability to generalization is good. Values in the table are again represented as percentage points in order to ease the interpretability of the table. There is no absolute limit how small RMSE should be so we need to settle for comparative view. RMSE for training set of ANN model was 1.28 when the RMSE for test set was 1.52. For MR model training RMSE was 1.06 and test 1.16. These are all relatively small values and especially the difference between the test and training RMSE is small, which indicates that the machine learning models are not notably underfitting or overfitting. However, it is important to notice that both RMSE values are higher for ANN model than for MR model. This would indicate slightly better generalization ability of MR model. RMSE comparison does not include the current instructions as the fitness of this model is not the main interest in this study.

6.2 Design costs

The final model for design costs included 47 projects from the 60 project sample set. Similarly to construction management model, data set was divided into training set, which included 80% of the projects and into test set, which included the rest 20% of the projects. Test set was left aside from the training phase of the model similarly as we did in the setup of construction management model. The models were tested 10 times by dividing the sample set with random number into training and test data. The process of constructing the model is same for construction management costs and design costs in order that the processes would obey identical line.

MR model for design costs was built similarly to MR model for construction management costs. In the case of design costs, four statistically significant variables were found during the model construction: contract type, project type, size of the project and area of the project. The associated OLS regression equation can be represented as:

$$\widehat{DESIGN} = \widehat{\beta_0} + \widehat{\beta_1} PT + \widehat{\beta_2} SIZE + \widehat{\beta_3} ST + \widehat{\beta_4} SA, \qquad (18)$$

where $\hat{\beta}_0$ denotes the OLS intercept estimate and $\hat{\beta}_1,\ldots,\hat{\beta}_4$ denotes the computed OLS estimates. Table 8 represents the structure of the final regression model for forecasting the design costs. As table 8 reveals, contract type and area of the project were not significant at 5% level. However, p-value for both of these variables denote significance at 10% level. Other model's variables, size and project type, are statistically highly significant as these denote significance even at 1% level.

The summary of the regression model in Table 8 shows us that all the heteroskedasticity robust standard errors denoted in parenthesis are relatively small and similar to the errors in the construction management model. R-squared in our model equals 0.62, which means that the independent variables are able to explain 62% out of the variance that happens in dependent variable. Adjusted R-squared was correspondingly 0.57 and thus close to the R-squared. However, the explanatory power in this model is lower than in the construction management model, which indicates that the design cost model is not able to explain as large proportion of the variation of the dependent variable.

Table 8: Regression model summary

	Design model
(Intercept)	0.0377^{***}
	(0.006)
SIZE	0.0206^{***}
	(0.008)
PT	-0.0247^{***}
	(0.007)
ST	-0.0214^{*}
	(0.011)
SA	0.0103^{*}
	(0.008)
R_2	0.62
Adj. R^2	0.57
P-value	0.00
RMSE	0.02

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

From examining the results showed in table 8, we are able to conclude that larger projects tend to have smaller proportions of indirect costs. If the overall costs of the project were less than 5 million and project was determined small, the design costs increase by 2.06 percentage points compared to larger project. In the case of project types, road projects decrease the design costs compared to railway projects. With road projects, costs diminish by 2.47 percentage points compared to railway projects. This indicates that in railway projects the proportional share of the design costs from overall costs is notably larger compared to the same share in road projects. In the design model the contract type was divided only into two categories, design and implementation contracts and other contract types. Design and implementation contracts decrease the value compared to other contract types by 2.14 percentage points. The geographical size of the project was this time denoted as binary variable. Variable obtains value 1 when the combined area of the project is smaller than two kilometers and value 0 otherwise. Model summary shows us that when the area of the project is small, the design costs are increased by 1.03 percentage points.

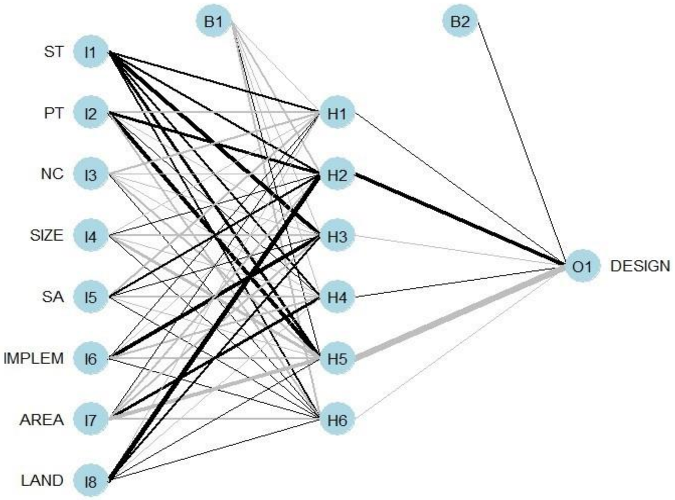

Figure 4: Structure of the neural network model for design costs

Figure 4 represents the structure of the artificial neural network related to the design model. I1-I6 represent the input variables, H1-H2 represent the hidden neurons, B1-B3 represent the intercept weights attached to each hidden neuron and output neuron and finally O1 represents the output produced by the model. Lines between the neurons in figure 4 describe properties of the connection weights. Thickness of the line represents the magnitude of weight and the color of the line reveals whether the weight sign is positive or negative. In figure 4 black lines represent positive weights and grey lines represent the negative weights. The figure shows that we have eight input variables and only one hidden layer with six neurons. Number of input variable in ANN model is larger than in MR model because the model's performance improved by adding the variables to the model. Compared to the ANN model built for construction management costs, this model is less complex as it only includes one hidden layer. In the final ANN model learning rate was set to 0.1 and threshold was set to 0.01. Similarly to the construction management model, artificial neural network was tested with 50 different combinations changing the number of layers, neurons and learning rate and the final model was chosen based on it's accuracy level.

Table 9: Summary of network's connection weights

Parameter		Weight					
		H1	H2	H3	H4	H5	H6
Input layer	ST	-0.17	1.31	0.37	-1.37	0.33	-1.23
	PT	-1.18	1.81	-0.53	1.05	-0.48	-0.94
	NC	0.65	-0.57	2.18	1.41	-0.76	1.48
	SIZE	-0.41	-2.58	-0.18	-1.50	-1.51	0.14
	SA	0.13	0.79	-1.61	-0.85	0.64	1.09
	IMPLEM	-0.47	-2.51	0.33	-0.98	-0.91	1.09
	AREA	-0.70	1.03	-0.82	0.85	0.34	0.68
	LAND	0.92	0.81	-0.24	-0.47	-0.25	1.51
Output layer	DESIGN	0.60	-3.58	-1.40	0.35	-0.92	2.39

Experiments with different values of the parameters exposed the differences between models. The best model forecasting design costs was, 8-6-1, which outperformed all other models with different parameter values. Other models produced slightly higher MAE and RMSE values, which indicates that the generalization ability in these models were poorer. Table 9 summarizes the connection weights between each neurons in different layers. In the ANN model for design costs there is only one hidden layer and thus all the input variables have weights attached to each of the six neurons inside the hidden layer. For instance, number of contracts have six weights attached to neurons H1-H6 in the hidden layer and correspondingly neurons H1-H6 all have weights connected to the output layer.

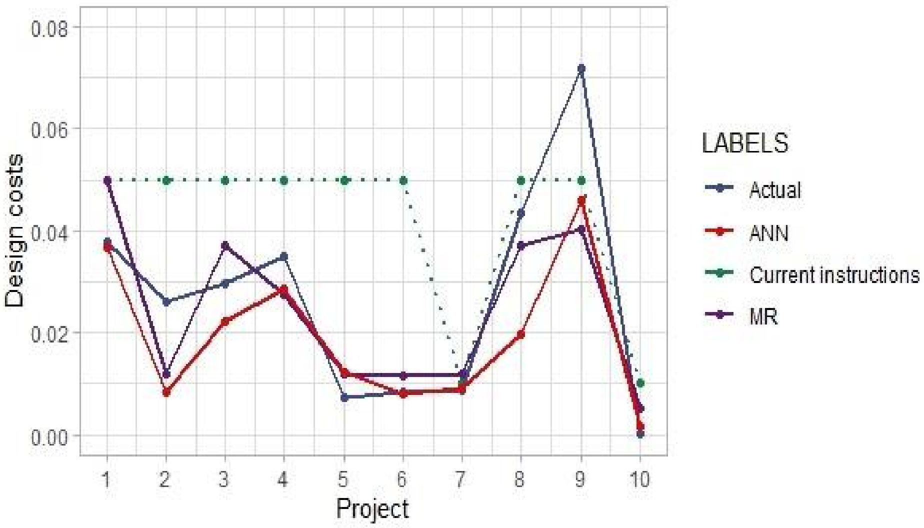

Figure 5: Prediction results for design costs

Similarly to the construction management model, the generalization ability of the model was tested by utilizing the test sample left aside from the training phase. Actual design costs of the test set were compared to the predicted values produced by ANN model, MR model and the current system. Figure 5 represents prediction results for the best performed MR and ANN models. Similarly to figure 3, each point in the graph describes one executed project and its design costs as a proportion of the overall costs. As the figure 5 shows, the design costs vary intensely. The prediction accuracy for current instructions is higher compared to construction management costs and this is mostly caused by the feature of current instructions receiving two different values depending whether the data included construction planning. Figure indicates that the ANN model and MR model forecast the design costs quite equally. However, in most of the test cases the ANN model's forecasts are more accurate compared to MR model, which indicates better accuracy of the ANN model. Comparing figure 5 to the figure 3 in previous chapter, it seems clear that the predictions for design costs is not as accurate as for construction management costs. Figure 5 also indicates that ANN and MR model are both able to outperform the current instructions although not as distinctly as with the construction management costs.

Table 10: Error measurements for design costs

Model	MAE	RMSE	Min MAE	Max MAE
ANN	1.11	1.57	0.78	1.41
MR	1.19	1.59	0.81	1.57
Current instructions	1.81	2.29	1.12	2.46

Table 10 represents MAE and RMSE values for test samples of the models and for the current instructions. Mean values are calculated as an average from the ten randomized tests performed and all values in table 10 are denoted as percentage points to ease the interpretability. Mean MAE for ANN model was 1.11 and for the same model minimum MAE reached was 0.78 and maximum 1.41.

For MR model mean MAE was correspondingly 1.19 reaching 0.81 minimum and 1.57 maximum values. These results comply with the previous studies as the ANN model is able to outperform the accuracy level of the MR model. In addition, similarly to construction management model MAE for the current instructions was slightly higher than either of the models'. Mean MAE for current instructions was 1.81 and it shifted between 1.12 and 2.46.

Table 11: RMSE values for training and test sets

Model	Training RMSE	Test RMSE
ANN	1.34	1.57
MR	1.66	1.59

Table 11 shows the measured RMSE values for the separated training and test sets of the ANN and MR model. As mentioned in the earlier chapter, comparison of RMSE between the test and training set is good way to indicate that there is no larger issues with the fitness of the model. Similarly to construction management model, the differences stayed relatively small and the RMSE values were not particularly high. For ANN model the difference between training set RMSE 1.34 and test set RMSE 1.57 is remotely larger than the one for MR model. With the MR model test results, difference between training RMSE 1.66 and test RMSE 1.59 is very small and can be kept as an indicator of good fit. It is hard to set any strict limits for the values that would indicate overfitting or underfitting. However, the differences between RMSE values and the actual values in both models are small enough to conclude that there is not at least any clearly observable overfitting or underfitting to be identified in these two models.

7 Discussion

The results of this study indicate that the machine learning methods are able to forecast owner's indirect cost more accurately compared to the currently applied instructions. However, the results of this research did not fully comply with the findings from earlier similar studies where same machine learning methods have been utilized to forecast construction costs. In this chapter, I will discuss the results of this study compared to earlier research and examine the possible reasons and explanations behind the contradictory results. In addition, I will discuss the possibility of developing further the forecast models represented in the previous chapter.

One clear conflict with earlier studies was that artificial neural network was not able to predict the construction management costs in implementation phase as well as multiple regression model could. Although this seems to be on clear contrary compared to earlier researches (Kim, An & Kang, 2004; Sodikov, 2005; Bayram & Al-Jibouri, 2016), the same literature has also given few possible explanations why the neural network does not necessarily always reach the same accuracy level as MR model. Firstly, as Sodikov (2005) concluded in his research, ANN model is proved to be superior compared to MR model merely in the earlier phases of the project where the information is still insufficient and the number of uncertainties is larger. In addition, he implied that MR model could be sufficient alternative for ANN model in the later phases of the project. This claim complies with the results of this study, because the construction management costs can be measured with vast knowledge as we are focused on the implementation phase of the project. Furthermore, Bayram and Al-Jibouri (2016) support this claim as well by noting that the simple multiple regression model is often the best choice in situations where the knowledge is comprehensive. This observation could explain at least some of the incompatibility between the results of this research and earlier studies.

Secondly, Kim et al. (2004) noted that the construction of ANN model is problematic as it takes considerably more time compared to other machine learning models. In this study, we tested 50 different ANN models with different combinations of hidden layers and neurons and varying the learning rate. To achieve even better accuracy level, the model would require more time and experiments with different values of parameters (Sodikov, 2005). Therefore, by consuming even more time to the model construction, ANN model might outperform MR model as the difference in accuracy between the two models is already very small. However, trying different network models is very arduous and possibly unnecessary as the MR model was able to produce fairly accurate results.

These explanations seem credible especially if we take a look at the results of the design cost model. Model forecasting design costs is also focused on later design phases of the project but in the case of design, the earlier design phases affect directly on the design costs of the subsequent phases and thus the knowledge related to design costs is less comprehensive. For instance, if the general design is done thoroughly enough the proportion of design costs in the later phases are inevitably smaller than if the general design has been more careless. The obvious problem is that the goodness of the plan is hard to estimate beforehand and thus the effect caused by the quality of previous design phases is challenging to measure reliably. This problem could be eased by collecting data on the actual design costs in the earlier phases and comparing their effect on the design costs in later phases. However, we can draw conclusion that the data applied in this thesis contains more missing information related to the design costs compared to the construction management costs. Thus the results of this study also favour the hypothesis that more complex machine learning models are suitable in the cases where the quality of the data is poorer. This strongly supports Sodikov's (2005) claim that data with more uncertainties and incomplete information is more suitable for ANN model and more comprehensive data is more suitable for simpler models. Although ANN model outperforms MR model in the prediction of design costs, the updating of neural network is extremely tedious. Model updating is necessary while receiving additional data from finished projects. Kim et al. (2004) made a conclusion that the model update would be considerably more complex for ANN model than for MR model or other similar simpler models. Updating of ANN model is arduous because as the amount of data increases, values for each parameter needs to be adjusted to maintain the best network structure. As described earlier, this is time-consuming process and thus the benefits from the use of neural network should be significantly larger compared to other available models. Benefits of applying MR model in forecasting the indirect costs during implementation phase seem apparent. Especially this is the case with construction management costs, where the multiple regression model outperforms the ANN model. With design costs there is a trade-off between the model accuracy and interpretability of the model. As the difference between error measurements of these two models is small and the interpretability of MR model is superior we can conclude that the benefits of using MR model would exceed the ones of ANN model. Also the easy implementation of MR model in practice is beneficial.

The use of neural network should be, however, considered as an option when modeling the indirect costs in earlier design phases of the project. As Bayram and Al-Jibouri (2016) noted, earlier design phases do not include as much knowledge compared to the implementation phase and thus the benefits of applying MR model would not be as distinct. On the contrary, ability to function with flawed and missing information is one of ANN models' many benefits and therefore it could provide more accurate predictions for indirect costs in the earlier phases. Interpretability problem with neural network could be eased with the use of sensitivity analysis similarly as Sodikov (2005) did. In this study the sensitivity analysis was not necessary as the MR analysis gave the necessary information regarding the importance of the variables but if the study would be replicated without MR model, alternative solution similar to sensitivity analysis would be essential.

Sodikov (2005) noted that rough approximate estimations in construction projects typically lead to large inaccuracies and that with the help of machine learning methods this difference between actual costs and estimations can be decreased notably. This conclusion is supported by the results provided by this study. As the results clearly indicate, both forecast models applied in this study are able to outperform the current instructions. Results showed that instructions for indirect costs have caused notable estimation errors and that the both machine learning methods applied are able to decrease the average error between actual and predicted costs. Thus applying machine learning models more in the cost estimation process of transport infrastructure projects could have significant positive effect on the accuracy of the cost estimation.

Although this study gave promising results, it is necessary to evaluate the results of this study critically and consider ways to improve the forecast models in the future. For instance, there still exists several projects of which indirect costs the model is unable to predict accurately. Thus, it is important to understand limitations in the accuracy of these models and to what extent these forecast models are able to improve the cost estimations. Firstly, as the previous studies have noted, the machine learning models are not able to reduce the inaccuracies of cost estimations infinitely and thus there will exist cost estimation bias reflected to CBA. However, this bias can be decreased from its current state and as Sodikov (2005) showed, with the use of machine learning methods cost estimations are even able to achieve Flyvbjerg's et al. (2003) definition of good cost estimation accuracy. Secondly, the limitations of relatively small sample size should not be disregarded and therefore enhancing the models in the future with larger sample would be essential.

Provided that this study would be reproduced, especially the design cost model could be improved even further. The model for design costs does not take into account that some projects include also construction design costs instead of solely design costs during implementation. Including this factor into the model the accuracy of the model could be improved. In this thesis, the separation between these projects was not done and therefore we were not able to address this issue. However, I have few suggestion how this issue could be examined more closely while reproducing this study. One option is creating a control variable representing these two separate design categories. For instance, variable could get value 1 when the project includes only design done during implementation and 0 otherwise. Control variable could help to separate the effect of construction design. However, concern with exploiting control variable is that it takes merely the average construction design into account and thus the variation in construction design costs between projects should be relatively small for this method to be robust. Without separation of the costs, we cannot say this with certainty. Another alternative choice could be reproducing two completely different models after separating the design costs. In other words, we could create one model forecasting the design done during implementation and other model to forecast the construction design. This solution also requires precise separation of design costs and thus it is difficult to execute in practice. By excluding the effect of design phase on the model we could possibly obtain more accurate models with smaller residuals.

There is also other machine learning methods, which could be utilized in order to solve the research question. For instance, Kim et al. (2004) applied the case-based reasoning model. As case-based reasoning has different strengths and weaknesses compared to the ANN and MR models, replication of this study with case-based reasoning model could provide interesting results. However, considering the sample size and complexity of the problem, ANN and MR model produced by themselves reasonably accurate results. The overall success of machine learning models in cost estimation has been remarkable but the use of these methods in the current cost estimation process of transport infrastructure projects still appears to be rare. The results of this study, especially the ability to improve the prediction accuracy, are yet another evidence that the application of machine learning methods more widely in the cost estimation process could lead to major improvements in the field of cost management.

8 Conclusions

Aim of this thesis was to propose new forecasting methods for the estimation of owner's indirect costs in transport infrastructure projects. In addition, the objective was to examine whether there exists machine learning models that could outperform currently applied instructions, which suggests the use of constant default values in the prediction of indirect costs. Study applied two commonly used machine learning models: artificial neural network and multiple regression analysis. Results produced by these two models clearly implicate that forecasting of owner's indirect costs can be notably improved by applying machine learning models. Ultimately this study aided to build two preliminary forecast models for owner's indirect costs. In addition, the study helped us to recognize the variables affecting most to the indirect costs and revealed that the prediction of design costs is actually more challenging compared to the prediction of construction management costs.

The most essential results of this study were the two forecast models proposed. Models' superiority compared to the performance of current instructions was clear especially in the case of construction management costs. The accuracy level, which forecast model for construction management costs achieved was notable improvement to the current state. However, prediction of design costs turned out to be rather complex and the model turned out to be less accurate compared to the construction management model. Depending on the project's properties, the cost estimation's shift towards actual costs of the project can be quite notable when using the forecast models developed in this thesis instead of the current instructions. It is important to notice that the better accuracy level of cost estimation does not mean direct savings for the financier. Rather it leads to more realistic results received from CBA and ultimately it could lead to improvements in the allocation of scarce resources. However, in reality also policy effects do have an impact on the CBA and although the cost estimation would be accurate, the results of CBA could still be biased due to other reasons.

Regardless of this, the benefits of accurate cost estimations are still significant. Future studies should focus on developing the forecast models suggested in this study. For instance, separation of design costs during implementation and construction design costs would enable to examine the individual effect of these two separate phases. This additional research could subsequently increase the accuracy of the design costs' forecast model. Replication of this study with larger sample and examining the effect of additional variables, which were excluded from this study could provide even more accurate results than the models proposed here. Although, the both forecast models built in this study could be improved in the future with more comprehensive and larger sample, the gains of this study are still considerable. Besides applying the proposed forecast models in practice, the research can be utilized even further. For instance, this study covered merely infrastructure projects financed by the government but also municipalities finance large amounts of transport infrastructure projects, which face similar owner's indirect costs. Applying machine learning models in these projects is one example how the results of this study could be used. In addition, this study could be utilized directly to the prediction of indirect costs in earlier design phases.

Modelling the indirect costs in earlier construction phases of the project is also another interesting further research topic. Reproducing this study with earlier design phases could enable accurate prediction of the indirect costs throughout the project's life cycle. These earlier phases include more missing information and the data quality is often poorer and the prediction accuracy smaller. Thus, applying ANN model in the forecasts of earlier design phases should be taken into consideration. Deeper knowledge of the earlier design phases could help also forecasting the design costs in later phases with better accuracy because designing done during these phases affects the amount of designing required later on.

This study has only scratched the surface of indirect costs and there remains a lot of room for further research in achieving more accurate and reliable cost management. However, as the earlier studies have concluded, the use of machine learning methods should be increased in the field of cost estimation in order to increase the accuracy of the estimations. This study supports this claim and additionally suggests that the selection of the particular machine learning method should be strongly related to the phase of the project. Earlier phases require more complex models, such as the artificial neural network but the subsequent phases include enough information to work with less complex machine learning models, such as simple multiple regression model.

References

Al-Zwainy, F. & Aidan, I. (2017). Forecasting the cost of structure of infrastructure Projects Utilizing Artificial Neural Network Model. Indian Journal of Science and Technology. Vol 10, pp. 1-12.

Babinec, T. (1997). Neural networks and statistical models. Proceedings of the Sawtooth Software Conference. pp. 333-341.

Bayram, S. & Al-Jibouri, S. (2016).Efficacy of Estimation Methods in Forecasting Building Projects' Costs. Journal of Construction Engineering and Management. Vol. 142, Issue 11.

Bode, J. (2000). Neural networks for cost estimation: simulations and pilot application. International Journal of Production Research. Vol. 38, Issue 6, pp. 1231–54.

Bouabaz, M. & Hamami, M. (2008). A cost estimation model for repair bridges based on artificial neural networks. American Journal of Applied Sciences. Vol. 5, Issue 4, pp. 334–339.

Boussabaine, A. H. (1996) The use of artificial neural networks in construction management: a review. Construction Management and Economics. Vol. 14, Issue 5, pp. 427–36.

Button, K. (2010). Transport economics. Third edition. Edward Elgar Publishing Limited. pp. 116-205

Cantarelli, C. C., Flyvbjerg, B., Molin, E.J.E. & van Wee, B. (2010). Cost Overruns in Large-Scale Transportation Infrastructure Projects: Explanations and Their Theoretical Embeddedness. European Journal of Transport and Infrastructure Research. Vol. 10, Issue 1, pp. 5-18.

Ciaburro, G. & Venkateswaran, B. (2017). Neural Networks with R. Packt Publishing Ltd. Chapter 2, 5.

Emsley, M., Lowe, D., Duff, R., Harding, A. & Hickson, A. (2002). Data modeling and the application of a neural network approach to the prediction of total construction costs. Construction Management and Economics. Vol 20, pp 465-472

Flyvbjerg, B. & Skamris, M. (1997). Inaccuracy of traffic forecasts and cost estimates on large transport projects. Transport Policy. Vol. 4, Issue 3, pp. 141-146.

Flyvbjerg B., Skamris Holm, M., & Buhl S. (2003). How common and how large are cost overruns in transport infrastructure projects? Transport Reviews, Vol. 23, no. 1, pp. 71-88

Flyvbjerg, B., Skamris Holm, M., & Buhl, S. (2002). Underestimating Costs in Public Works Projects: Error or Lie? Journal of the American Planning Association, Vol. 68, no. 3, pp. 279-295.

G´omez-Lobo, A. (2012). Institutional Safeguards for Cost Benefit Analysis: Lessons from the Chilean National System. Journal of Benefit-Cost Analysis. Vol. 3, Issue 1, Article 1

Gwilliam, K.M. & Mackie, P.J. (2017). Economics and transport policy. Taylor & Francis Group. Vol. 7, pp. 53-113.

Honkatukia, J. & Antikainen, R. (2004). Väylähankkeiden kansantaloudellinen merkitys, VATT-keskustelualoitteita 341. Valtion taloudellinen tutkimuskeskus.

Irfan, M., Khurshid, M., Anastasopoulos, P., Labi S. & Moavenzadeh F. (2011). Planning stage estimation of highway project duration on the basis of anticipated project, project type and contract type. International Journal of Project Management. Vol. 29, Issue 1, pp.78-92

Jones, H., Moura, F. & Domingos, T. (2013). Transport infrastructure project evaluation using cost-benefit analysis. Procedia - Social and Behavioral Sciences. Vol. 111, pp. 400 – 409.

Kasabov, N. (1996) Foundation of neural Networks, Fussy Systems, and Knowledge Engineering. The MIT Press Cambridge, Massachusetts. pp. 251328.

Kim, G-H., An, S-H. & Kang, K-I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Building and Environment. Vol. 39, pp. 1235-1242.

Lasshof, B. & Stoy, C. (2016). Estimation models for heating energy and electricity costs. Construction Management and Economics. Vol. 34, Issue 9, pp. 622-640.

Layard, R. & Glaister, S. (1994). Cost-benefit analysis. Cambridge, UK: Cambridge University Press. Second edition, pp. 1-56.

Liikennevirasto. (2013). Väylähankkeiden kustannushallinta. Liikenneviraston ohjeita 46/2013.

Lundberg M., Jenpanitsub, A. & Pyddoke, R. (2011).Cost overruns in Swedish transport projects. Working papers in Transport Economics. 2011:11, CTS - Centre for Transport Studies Stockholm.

McClelland, J.L. & Rumelhart, D.E. (1988). Training hidden units: the Generalized Delta Rule. Explorations in Parallel Distributed Processing. Chapter 5, pp. 121-159

Mullainathan, S. & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives. Vol. 31, Nro 2 – Spring 2017 pp. 87-106.

Munnel, A.H. (1992). Policy Watch: Infrastructure Investment and Economic Growth. Journal of Economic Perspectives. Vol. 6, Number 4.

Nijkamp, P. & Ubbels B. (1998). How reliable are estimates of infrastructure costs? A comparative analysis. Research Memorandum 1998-29.

Odeck, J. (2004). Cost overruns in road construction — what are their sizes and determinants? Transport Policy, Vol. 11, Issue 1, pp. 43-53.

Polat, G. (2012). ANN approach to determine cost contingency in international construction project. Journal of Applied Management and Investments. Vol. 1, pp. 195–201.

Rakennustieto Oy. (2015). Infra 2015 Rakennusosa- ja hankenimikkeistö Määrämittausohje. Rakennustietosäätiö RTS. pp. 5-20, 157-165.

R´onai, P. (2001). Cost estimation method of transport infrastructure projects. Periodica Polytechnica Ser. Transp. Eng. Vol 29. No. 1-2, pp. 107-116.

Sibi, B., Allwyn Jones S. & Siddarth P. (2013). Analysis of different activation functions using backpropagation neural networks. Journal of Theoretical and Applied Information Technology. Vol 47, No.3, pp- 1264-1268

Skitmore, M. (1987) The effect of project information on the accuracy of building price forecasts. In: Brandon PS, editor. Building cost modelling and computers. London: E& FN Spon. pp. 327–336.

Smith, A.E. & Mason A.K. (1996). Cost estimation Predictive Modeling: regression versus Neural Network. The Engineering Economist. Vol 42, No. 2, pp. 137-161.
Sodikov, J. (2005). Cost Estimation of Highway Projects in Developing Countries: Artificial Neural Network Approach. Journal of the Eastern Asia Society for Transportation Studies. Vol 6, pp. 1036-1047

Van Wee, B. (2012). How suitable is CBA for the ex-ante evaluation of transport projects and policies? A discussion from the perspective of ethics. Transport Policy, Vol 19, Issue 1, pp. 1-7.

Vickerman, R. (2007). Cost — Benefit Analysis and Large-Scale Infrastructure Projects: State of the Art and Challenges Environment and Planning B: Planning and Design, Vol.34 (4), pp. 598-610

Waziri, B.S., Bala, K. & Bustani, S.A. (2017). Artificial Neural Networks in Construction Engineering and Management. International Journal of Architecture, Engineering and Construction, Vol 6, No. 1, March 2017, pp. 50-60

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 48(4), pp. 817–838.
Woolridge, J. (2013). Introductory Econometrics: A Modern Approach. Fifth Edition. pp. 268-294.

World Bank (2004). Monitoring and evaluation: Some tools, methods and approaches. Evaluation Capacity Development working paper series; ECD.

Wythoff, B.J. (1993). Backpropagation neural networks. A tutorial. Chemometrics and Intelligent Laboratory Systems. Vol. 18, pp. 115-155.

VÄYLÄ

Finnish Transport
Infrastructure Agency

ISSN 2490-1202
ISBN 978-952-317-803-8
www.vayla.fi