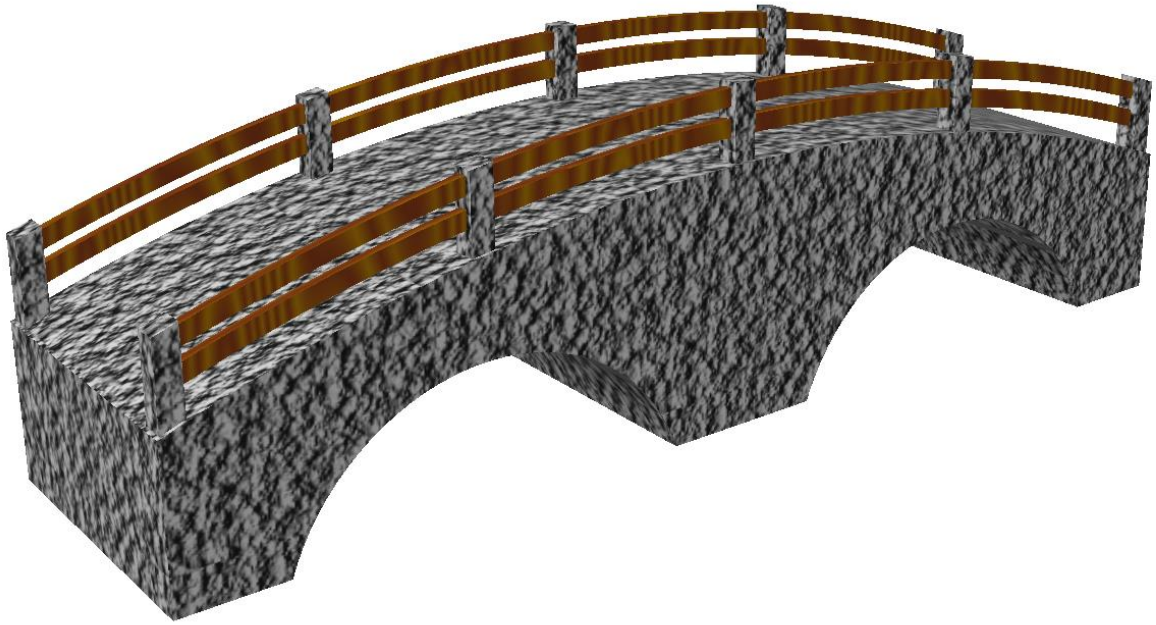
CHALMERS
UNIVERSITY OF TECHNOLOGY



AI-Supported Bridge Tendering

A Process for Transforming Data for Improved Decision Making in Tendering

Master's thesis in the Master's Programme Management and Economics of Innovation

Daniel Boman
Lisa Wallin

DEPARTMENT OF TECHNOLOGY MANAGEMENT AND ECONOMICS
DIVISON OF INNOVATION AND R&D MANAGEMENT

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020
www.chalmers.se
Report No. E2020:004

REPORT NO. E 2020:004

AI-Supported Bridge Tendering

A Process for Transforming Data for Improved Decision Making in Tendering

DANIEL BOMAN
LISA WALLIN

Department of Technology Management and Economics
Division of Innovation and R&D Management
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2020

AI-Supported Bridge Tendering
A Process for Transforming Data for Improved Decision Making in Tendering
DANIEL BOMAN
LISA WALLIN

Report no. E2020:004
Department of Technology Management and Economics
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Cover: Image of bridge made in CAD.

Gothenburg, Sweden 2020

Acknowledgements

This thesis was conducted at the department of Technology Management and Economics at Chalmers University of Technology and at a large construction company in Gothenburg, Sweden. We would like to thank everyone involved in performing this thesis. Especially we would like to thank our tutors Kaj Suneson at Chalmers University of Technology, Christina Claeson-Jonsson and Rasmus Rempling at NCC for all their support, feedback and guidance throughout this process.

Daniel Boman and Lisa Wallin

AI-Supported Bridge Tendering
A Process for Transforming Data for Improved Decision Making in Tendering

DANIEL BOMAN
LISA WALLIN

Department of Technology Management and Economics
Chalmers University of Technology

ABSTRACT
The tender process is an important part of the construction industry where most projects are procured using a competitive tender. A tender bid is costly to calculate which means that for each tendering without receiving a contract, the cost of it needs to be accounted for in following projects, which makes it increasingly difficult to place competitive tenders. Due to the national principle of public access, all tender documents procured from the Swedish transport administration, Trafikverket, are public which underlines the perception that the construction industry deals with large amounts of data. However, the industry is well known for its fragmented data practices. Hence, a large potential lies within utilizing the available data and finding ways to transform data into information. Today's tender process is mainly based on the intuition of experienced practitioners. Given the importance of competitive tenders and the amount of existing data, a construction company would benefit from a process using data to support tendering.

This study aims to create a step-by-step process to transform data to information with the overall purpose to guide tendering. The first step is to investigate Trafikverket as a data source in terms of structure and quality of public data. Then, the data is processed into a neural network model predicting the number of tender bids for projects. By performing these two steps and combining them, a sketch of a process divided into steps for each required action is formed. By decomposing all actions and evaluating options and choices along the transformation, a process is constructed.

Collecting data from Trafikverket is time consuming due to unstructured archiving. Unless Trafikverket standardizes a digital praxis for archiving documents, or if natural language processing techniques significantly increase in their performance, Trafikverket is not advisable as a source for large quantities of data. Moreover, the neural network model does not have enough predictive power given the data size and input variables used in this study. The performance of the neural network model is Root-Mean-Squared Error of 2.45 given only 41 observations. However, performance can likely be increased further by adding more data or by further identifying input variables that affect the number of tender bids. Finally, this study proposes and eight-step process which differs from previous processes since it accounts for the fact that much data in the construction industry are in documents rather than in a proper database.

Keywords: Data Transfer, Data-Information-Knowledge-Wisdom, Knowledge Discovery in Databases, Neural Network Model, Construction Industry, Tendering, Trafikverket

Contents

1 Introduction

The majority of all advanced construction works are procured using a competitive tender (Ng, Cheung, Skitmore, & Wong, 2004). In a competitive tender, contractors compete on who can offer to build a construction for the lowest price. Each tender bid a contractor places is associated with a cost due to labor and resources used for preparing and calculating the tender (Ngai, Drew, Lo, & Skitmore, 2002). The costs associated with each lost tender needs to be covered by the income from tenders that are won (Ngai et al., 2002). Moreover, projects with fewer competing tender bids tend to have higher project bid prices (Carr, 2005), which favors the contractor. Thus, optimizing project selection to bid on projects with less competition, or selecting the projects with the highest likelihood of winning, is of great importance for contractors in order to maintain a healthy margin and placing competitive tenders. However, today the project selection and price levels of tender bids are based mainly on experience and intuition of practitioners (Cheng, Hsiang, Tsai, & Do, 2011), and could, therefore, benefit from an approach based on data from previous tenders.

Ismail, Bandi, and Maaz (2018) state that the construction industry is known to deal with large amounts of data; hence, there is a big potential in utilizing the data for further development within the industry. Additionally, the construction industry is also well known for fragmented data practices such as non-standardized values (Bilal et al., 2016). This implies that the industry does not extract the full value of accessible data and that it could gain from focusing on data quality and processing (Bilal et al., 2016). Abdullah and Ahmad (2013) claim that in order to efficiently manage data, organizations must establish a process for transforming unstructured data to structured data. A possible remedy for the previously mentioned subject is found in the data-information-knowledge-wisdom (DIKW) hierarchy and the Knowledge Discovery in Databases (KDD) process. The widely recognized DIKW hierarchy model offers a methodology to contextualize data, information, knowledge and wisdom to describe the process of transforming entities of a lower level to entities of a higher level in the hierarchy (Rowley, 2007). Similarly, the KDD process proposes a nine step process as a way to obtain knowledge from data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). However, the DIKW hierarchy and the KDD process are general processes, and contractors could benefit from a process tailored towards the current situation of fragmented data in the construction industry.

One of Sweden's fundamental laws, the Freedom of the Press Act, has incorporated the principle of public access to official documents from government authorities (Government Offices of Sweden, 2020). One of these authorities is Trafikverket, the Swedish Transport Administration, which is responsible for building, operating and maintaining public roads and railways (Trafikverket, 2015). Trafikverket uses competitive tenders to procure infrastructure construction projects, and the documents produced for each tender become public documents (Trafikverket, 2020a). Therefore, it could be of interest for construction companies to collect these documents and utilize the data in the documents to guide decisions in the tender process.

Few studies have been performed in optimizing project selection and tendering processes within the construction industry by systematically analyzing tender data. The ones performed are limited to isolated geographical areas with homogeneous projects. These stud-

ies use different statistical and predictive methods such as regression analysis and correlation between factors to reduce costs and increase efficiency in tendering (Ballesteros-Pérez, González-Cruz, Fernández-Diego, & Pellicer, 2014; Drew & Skitmore, 1997; Flanagan & Norman, 1982; Minli & Shanshan, 2012; Ngai et al., 2002; Skitmore, 2002). Furthermore, a few studies in adjacent research fields have used a neural network model as the prediction tool. The neural network model is inspired by the human brain and it can learn to make predictions or solve classification tasks by using training examples. Additionally, the neural network model is becoming increasingly popular across industries as the performance improves and the fields of application grows (Lee, 2019).

Duff, Emsley, Gregory, Lowe, and Masterman (1986) and Kim, Yoon, An, Cho, and Kang (2004) argue that a neural network is better suited while modeling with incomplete data sets compared to other common methods such as regression analysis. Due to the fragmented nature of data practices in the construction industry, complete data sets are not guaranteed, which emphasize the value of this ability for the model in this study, making the neural network model particularly interesting. Additionally, a neural network model is applicable to large data sets (Duff et al., 1986) which makes the method germane to model tender data from Trafikverket. Further, Kim et al. (2004) claim that a neural network can model interdependencies and non-linear relationships which are likely to appear in cost parameters in construction projects. However, training a neural network model requires data processing and choices of variables in the setting of the model which the current literature lack consensus about. No study, to our knowledge, has combined these steps and explored the overall process from identifying data sources to transform it into information supporting tendering.

1.1 Purpose

The overall purpose of this study is to explore how tender data can be used by contractors to guide decisions in the tender process. More specifically, we aim to increase the understanding of three important issues for contractors to start using data to guide decisions. The first issue is obtaining tender data. We investigated Trafikveket as a source for historical tender data by assessing the available data and its structure and quality for bridge projects in Sweden. The purpose of investigating Trafikverket as a data source is to determine if they have relevant data with sufficient structure and quality to be used as input data. The second issue is the choice of model. We investigate how suitable a neural network model is as a quantitative method to predict the number of tender bids of bridge projects. The third issue concerns the process, specifically the lack of knowledge about the Trafikverket public databank and the transformation of this data into useful information. Based on the course of action used in the first two issues, we aim to propose a process that is more tailored towards tender decisions than the DIKW-hierarchy and the KDD process. The purpose of the process is to create a better understanding of how to include tender data in the decision making process.

1.2 Delimitations

Due to time constraints and the time-consuming process of acquiring data from Trafikverket, the data is limited to publicly available tender data of bridges in Sweden for the limited time period 2011-2015.

1.3 Research Questions

Based on Section 1.1, several research questions are to be explored in the thesis:

1. What characteristics do tender documents from Trafikverket contain in terms of data available, quality, and structure?

2. How suitable is a neural network model for predicting the amount of competing tender bids on projects procured by Trafikverket?

3. What are the steps involved in a process to transform tender data to information capable to support decisions in the tender process?

1.4 Organization of the Study

First in this study, a background of the topic could be found, explaining the possibility for development in the field. Thereafter followed the purpose of the study, the delimitations of the scope and the chosen research questions for the report. In Section 2, a theoretical foundation is presented where the authors elaborate on technical concepts and previous studies regarding data transformation and the use of tender data. Section 3 is the methods chapter where the methodology of the performed study is presented. In Section 4, the data that has been used for the study and its quality is presented, thereby answering research question one. Thereafter, in Section 5, the results are presented to answer research question two and three and is presented as the process suggested for transforming data to information, and as performance from the model with the neural network. Finalizing the study are conclusions, which can be found in Section 7, including further research suggestions to build upon this thesis in the future.

2 Theoretical Background

In this Section, the theoretical framework for the thesis is presented. First, two perspectives of the process of transforming data into higher level information is described. Thereafter, literature describing how data can be profiled, cleaned and structured is presented. Finally, the methods and goals of several previous studies using tender data is summarized.

2.1 Transforming Data Into Information

The process of transforming data into actionable information has been studied by several authors (Fayyad et al., 1996; Rowley, 2007). In this Section, two proposed models for the process of turning data into information are presented.

2.1.1 The Data-Information-Knowledge-Wisdom Hierarchy

The Data-Information-Knowledge-Wisdom (DIKW) hierarchy is a model describing how data, information, knowledge and wisdom relate to each other (Rowley, 2007). The underlying sentiment is that wisdom is created from knowledge, knowledge is created from information, and information is created from data (Rowley, 2007). The DIKW hierarchy is often quoted, either directly or implicitly (Baskarada & Koronios, 2013), but the definitions of its contents varies. Data is often described as records, facts or observations that have no meaning or structure (Rowley, 2007). One definition of data provided by Rowley (2007, p. 170) is "discrete, objective facts or observations, which are unorganized and unprocessed, and do not convey any specific meaning". Information is often defined as data that has been made meaningful, processed for a purpose, or can be understood by the recipient (Rowley, 2007). Rowley (2007, p. 170) defines information as "data that have been organized so that they have meaning and value to the recipient". The definitions of knowledge and wisdom are more complex and diverse than those of data and information (Rowley, 2007). Some authors argue that knowledge is simply information that has been processed to convey understanding while others argue that the term is ambiguous and that there is no consensus on the definition of the concept of knowledge (Rowley, 2007). The concept of wisdom can be defined as "accumulated knowledge, which allows you to understand how to apply concepts from one domain to new situations or problems" (Rowley, 2007, p. 174). However, wisdom is rarely defined in DIKW literature and has been argued to be an elusive concept (Rowley, 2007).

While the definitions of the levels in the DIKW hierarchy varies, common to most conceptualizations is that the challenge is considered to be understanding how data can be transformed into information, and consequently into the following steps (Rowley, 2007). There is no consensus regarding the processes that transforms data into information (Rowley, 2007). However, information is seen by some to be structured data, indicating that the transformation process between data and information consists of structuring data into a schema (Rowley, 2007). In accordance with this, Jin, Anderson, Cafarella, and Jagadish (2017) propose that raw data is usually not in a format that allows it to be entered into a database directly. Therefore, a common first step in data analysis tasks is to process the raw data into a format suitable for downstream applications (Jin et al., 2017). Curtis and Cobham (2008) propose five processes that data undergoes before being used as information: classification, sorting, aggregating, calculations, and selection. Bocij, Greasley, and

Table 1: Description of five data transformation processes (Bocij et al., 2008).

Data Transformation Process	Description
Classification	Dividing data points into categories.
Sorting	Grouping data or putting data into a particular order.
Aggregating	Summarising data in terms of totals, averages or other methods.
Calculations	Performing calculations on data in order to retrieve information implicitly included in the data.
Selection	Including or discarding data based on a set of criteria.

Hickie (2008) agree with the five processes but elaborate that any action taken to make data meaningful is a data transformation process. See Table 1 for a description of each of the five data transformation processes, based on Bocij et al. (2008).

2.1.2 Knowledge Discovery in Databases

A similar process for transforming data is the Knowledge Discovery in Databases (KDD) process proposed by Fayyad et al. (1996). According to the authors, the objective of the KDD process is to obtain knowledge from data. However, it is important to note that the definition of knowledge used by Fayyad et al. (1996) in the KDD process differs from the general definition of knowledge. Fayyad et al. (1996) define knowledge in a completely user oriented way as: any pattern that exceeds some threshold of interestingness. In this case, interestingness is an overall measure that combines validity, novelty, usefulness, and simplicity (Fayyad et al., 1996). KDD includes the entire process of turning data into knowledge such as data storage, choice of algorithms, result visualization and the interaction between man and machine during the process (Fayyad et al., 1996). Two important concepts in the KDD process are data cleaning and data access (Fayyad et al., 1996). Data cleaning includes tasks such as creating a uniform naming and representation system for data, assessing the amount of missing data and possible aids, and addressing other errors in the data (Fayyad et al., 1996). Data access refers to the creation of access paths to the data (Fayyad et al., 1996).

Fayyad et al. (1996) propose a nine step process for transforming data into knowledge which is illustrated in Figure 1. The first step is to acquire the necessary domain expertise, and establishing the goals of the data transformation process. The second step is to select the target data i.e. the desired data that will be collected, processed and on which analysis eventually will be done. The third step is to clean and process the data in order to address missing data, and ensure that data is uniformly represented. The fourth step consists of identifying features to represent the data and reducing the dimension of the feature vector. The fifth step is to identify data-mining methods that can be used on the data to achieve the goals that was specified in step 1. The sixth step is a screening and selection of models and associated parameters to be applied to the data. The seventh step is the actual execution of the data-mining methods such as clustering and regression. The eighth step is interpreting,

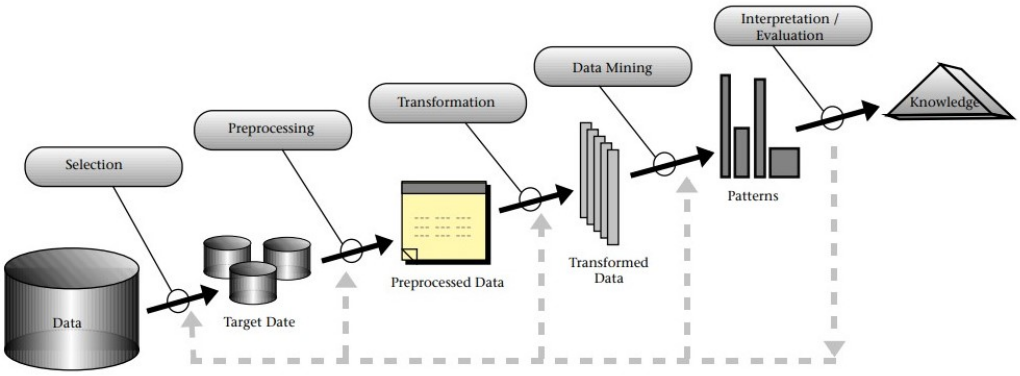


Figure 1: The Knowledge Discovery in Databases process (Fayyad et al., 1996)

and if desired visualizing, the results for interpretation. The ninth and final step is acting on the resulting discoveries from the process, or documenting the results for the appropriate stakeholders.

2.2 Data Profiling

Assessing the quality of data is called data profiling (Ganti & Sarma, 2013). The process of data profiling usually includes calculating various data quality metrics, which provide information about the quality of the data (Ganti & Sarma, 2013). Moreover, a data profile is often created both before and after data cleaning takes place, in order to evaluate both the need for, and the results of the data cleaning activities (Ganti & Sarma, 2013). Data quality can be defined as "data that is fit for use by data consumers" (Strong, Lee, & Wang, 1997, p. 104). Due to the increased amount of digital data available, data quality is becoming important for the private and public sector alike (Batini, Cappiello, Francalanci, & Maurino, 2009). Data quality dimensions have been proposed by several authors (Batini et al., 2009; Pipino, Lee, & Wang, 2002; Scannapieco & Catarci, 2002; Strong et al., 1997; Wang & Strong, 1996). While the proposed dimensions to measure data quality varies, most of them include the dimensions of accuracy, completeness, consistency, timeliness, interpretability, and accessibility (Scannapieco & Catarci, 2002). Moreover, most additional dimensions that have been proposed can be considered redundant due to being very context specific or only capturing secondary features (Scannapieco & Catarci, 2002). In Table 2, the six quality dimensions are described in more detail.

When assessing data quality Pipino et al. (2002) propose that the set of metrics to use should be specific to the required situational needs. Batini et al. (2009) propose several metrics for how to measure each dimension of data quality, including the metrics listed in Table 3.

2.3 Data Cleaning

Data cleaning refers to the identification and resolving of data quality issues (Rahm & Do, 2000). Tools for cleaning data are called Extraction, Transformation, and Loading (ETL)

Table 2: Six quality dimensions.

Dimension	Description
Accuracy	The degree to which data is error-free, accurate and correct (Wang & Strong, 1996) in comparison to its real value (Batini et al., 2009).
Completeness	Describes the scope of the data in terms of breadth, depth (Wang & Strong, 1996), and missing values (Batini et al., 2009).
Consistency	Degree to which data points of the same type comes in the same format (Pipino et al., 2002).
Timeliness	Measures if the data is sufficiently up-to-date in order to be used for a given task (Pipino et al., 2002).
Interpretability	Degree to which the data is sufficiently well-defined and in appropriate languages for it to be interpreted (Pipino et al., 2002).
Accessibility	How available the data is in terms of ease and quickness of retrieval (Pipino et al., 2002).

Table 3: Metrics for measuring data quality dimensions according to (Batini et al., 2009)

Dimension	Metric
Accuracy	Amount of correct data points/Total number of data points
Completeness	Amount of non-missing data points/Total number of data points
Consistency	Amount of consistent data points/Total number of data points
Timeliness	$(\max(0;1\text{-Currency/Volatility}))^s$
Interpretability	Amount of data points that can be interpreted
Accessibility	max (0;1-(Delivery time-Request time)/(Deadline time-Request time))

tools (Galhardas, Florescu, Shasha, Simon, & Saita, 2001). Creating a data cleaning process consists of two steps: designing the process at the logical level and at the physical level (Galhardas et al., 2001). Designing the process at the logical level refers to a conceptualization of which data cleaning algorithms will be used and in what order (Galhardas et al., 2001). The goal of the logical design is to define an algorithm that will transform the raw data into clean data with as high accuracy as possible (Galhardas et al., 2001). The design on the physical level refers to implementing the algorithms to perform each transformation in the logical design (Galhardas et al., 2001). The objective of the physical design is to achieve as good processing speed as possible without affecting the overall accuracy (Galhardas et al., 2001).

Pigott (2001) present approaches when working with data of inadequate completeness: complete case analysis and available case analysis. Complete case analysis entails omitting all observations that are missing any data point that is included in the model (Pigott, 2001). Available case analysis refers to using all available data where it is possible (Pigott, 2001). However, for calculations including multiple variables, only the observations including all variables required for the calculation are used (Pigott, 2001). If the consistency of data is not sufficient for use, standardization might be necessary. Standardization is the activity of transforming the values of a dataset to all conform to the same conventions (Ganti & Sarma, 2013). An example of data standardization is transforming the dimensions of a product into the same measurement unit (Ganti & Sarma, 2013).

2.4 Data Structure

There exist unstructured, semi-structured and structured data (Li, Ooi, Feng, Wang, & Zhou, 2008). Structured data is organized, the elements are addressable and it is the most processed to manage information and effective analysis (Rusu et al., 2013). Semi-structured data exists to ease space and has some organizational properties and unstructured data is data without organizational properties (Rusu et al., 2013). Unstructured data can be found scattered through organizations and in various documents and is estimated to contain 85% of all business information as part of documents text bodies and can consists of e.g. e-mail messages, memos, reports, contracts and spreadsheets (Abdullah & Ahmad, 2013).

McCallum (2005) state that the main part of the information in the world is in an unstructured form, in a so called natural language text and must be transformed into a structured and normalized database form. Abdullah and Ahmad (2013) claim that in order to efficiently manage data, organizations must establish a process of mapping unstructured data to structured data. Abdullah and Ahmad (2013) propose a method consisting of four processes:

1. Extraction
2. Classification
3. Repositories Development
4. Data Mapping

Where extraction regards identifying different sources and formats of unstructured data for further extraction of entities and facts. The second process, classification is the process of categorizing the unstructured data depending on the nature and the format of the data. The main data classes are Text, Image, Audio and Video and the identification of main data classes of unstructured data within an organization is determined by observing the collected data available. The third process, Repositories Development regards the way the unstructured data should be stored and managed to support organizational functions through access, usage and innovation of it. The last process in the method suggested by Abdullah and Ahmad (2013) is Data Mapping which content preparing unstructured data subjects and identification of thematic topics to which the subjects belong. Obtaining and accessing information from unstructured data is important and in order for the unstructured data to grow and be more meaningful, it needs to be associated with a thematic topic (Abdullah & Ahmad, 2013).

One proposition on how to structure a dataset is presented by Wickham et al. (2014) called Tidy data which is a standardized way of mapping the meaning of a dataset into a structure. The dataset is a group of values whom each belong to a variable and an observation, where the variable includes all values describing the same attribute over all units and an observation contains all values describing the same unit over all attributes (Wickham et al., 2014). The authors define Tidy data as a dataset that fulfills the following three conditions:

1. Every variable is a column.
2. Every observation is a row.

3. Every observational unit has its own table.

A Tidy dataset enables easy extraction of variables due to the standard way of structuring the data (Wickham et al., 2014). They claim that it suits programming languages such as R especially well, since the layout of the data pairs all variables from the same observation.

2.5 Tender Data Analysis

In this Section, information from previous studies using tender data will be presented.

2.5.1 Factors Determining Competition

In a study by Drew and Skitmore (1997) the effect of contract type and size is used to model contractors competitiveness in bidding and their competitive behaviour. There is a significant correlation between competitiveness and variation in bidding and one suggested reason for this is that contractors have preferred contract types and sizes on which they bid more competitively (Drew & Skitmore, 1997). Drew and Skitmore (1997) claim the construction market to consist of a set of defining factors, namely contract size and complexity, the type of contract, the geographical location of the project and additionally stated by Flanagan and Norman (1982) the current workload for the contractor. Drew and Skitmore (1997) have used similar historical tender data to analyze the relative competitiveness for construction companies, in order to better understand their strengths and weaknesses to further optimize their tender bid processes and behaviour.

Drew and Skitmore (1997) derive the diversity of bids between contractors to different levels of efficiency. Efficient bidders tend to put lower bids and will reach a higher level of competitiveness over time (Drew & Skitmore, 1997). Additionally, different contractors have different preferences in terms of a contract's size, complexity and location, which will determine the level of competitiveness in their bid (Drew & Skitmore, 1997).

Similarly, Flanagan and Norman (1982) emphasize on the need for optimizing tendering processes due to the cost, effort and time of unsuccessful tendering. Flanagan and Norman (1982) examine relative competitiveness of contractors in the same geographical area with contracts similar in complexity and type. The different contractors are of different size and would mostly be tendering in different project value ranges (Flanagan & Norman, 1982). By mapping the tenders for each contractors bidding performance, Flanagan and Norman (1982) present how competitive the contractors are and conclude that one company considered contract size and type; another's competitiveness was unrelated by contract size and type; and one company was more successful on contracts in a higher value range. Further Flanagan and Norman (1982) conclude that the patterns shown supports the idea that contractors tendering strategy is affected by the type of project and by the contracts value range.

In another study, estimating future bidding performance, Ballesteros-Pérez et al. (2014) state that many bidding strategy models already have been developed to predict the probability of successful tendering. However, Ballesteros-Pérez et al. (2014) argue that the complexity of public tendering requires additional tools to support decision making and the selection process for bidders. Bid Tender Forecasting Model (BTFM), a model presented in 2012, uses a statistical procedure based on relevant historical bidding performance to

improve bidding strategies and further support the success rate in tendering (Ballesteros-Pérez et al., 2014). Since studying a company's rivals behaviour and how one positions in comparison is valuable, Ballesteros-Pérez et al. (2014) show how the BTFM is used to map competitors and subgroups in their patterns for bidding.

The method presented by Ballesteros-Pérez et al. (2014) describe how to measure the performance of a bidder in terms of their score out of the maximum score of the projects tenders and the position achieved by the bidder out of number of other bids. In order to assess the bidder, the method requires a dataset containing previous bids, past scores and positions out of the maximum score and number of bidders (Ballesteros-Pérez et al., 2014). Further, Ballesteros-Pérez et al. (2014) state that the model requires the dataset to consist of tenders homogeneous to the forecasted tender in terms of e.g. scope of works and geographical area as well as being fairly recent in execution. In case of older datasets, Ballesteros-Pérez et al. (2014) mean that factors such as change in regulations and economic situations can affect the level of the tender price.

Given that many companies know their closest competitors, it is valuable to be able to assess the scoring and positioning within a group of bidders (Ballesteros-Pérez et al., 2014). Ballesteros-Pérez et al. (2014) further develop the the model to plot the performance for different actors in a group of bidders for companies to assess in what range of scores the probability of being surpassed lies. With this information, Ballesteros-Pérez et al. (2014) claim, a single bidder can utilize their economic and technical capability to increase competitiveness.

2.5.2 Number of Tenders In Competition

A study by Ngai et al. (2002) create an idea for a framework to determine the minimal number of bidders for a construction contract in order to get competitive bids using a cost effective approach. For every tendering there is a cost thus for each unsuccessful tender there are costs that must be accounted for by the contractor in the next project (Ngai et al., 2002). This increases the price for projects in the long run, resulting in an unnecessary waste of limited resources (Ngai et al., 2002).

According to Ngai et al. (2002), the determination of price of a tender is cyclical with the economy. In booming economic years, the number of tenders will be lower per project since more projects will be performed and companies will tend to bid with higher profit margins (Ngai et al., 2002). Conversely, the competition will be higher in slump periods when the utilization of capacity goes down due to decreasing amounts of projects and companies will bid with lower profit margins (Ngai et al., 2002). Hence, Ngai et al. (2002) claim that the market condition affects the price level of the bid and the number of competitors. Based on this, Ngai et al. (2002) perform an empirical analysis, predicting the number of competitors for projects by using the rate of change for the tendering price index (TPI) and the average number of competitors in a regression analysis creating a model, a time series, to predict the number of average tenders for a project (Ngai et al., 2002).

2.5.3 Predicting Tender Price

In a study performed by Minli and Shanshan (2012), the application of an artificial neural network model is examined as a decision-making model for tender offers by predicting the

tender price. Minli and Shanshan (2012) state that many previous studies have established models for estimation of tender offers, however, applying the artificial neural network model makes for a more precise and effective predictions. The intensity in competition in construction markets makes it more difficult to successfully bid on projects hence an increased ability to determine the relationship between affecting factors and final offer for a project increases the ability for an enterprise to win the bidding (Minli & Shanshan, 2012). The factors affecting the tender price is according to Minli and Shanshan (2012) often uncertain, ambiguous and affected by psychological factors of the decision-maker of the tender. Further, Minli and Shanshan (2012) claim that an experienced bidder roughly can estimate the tender price based on experience without having to perform complex and extensive calculations. In this study, Minli and Shanshan (2012) present 14 factors affecting the characteristics of bidding for a project. In Table 4, these factors are presented. The amount of observations used by previous studies for training neural network models to predict tender prices has varied. Elhag and Boussabaine (1999) trained a neural network model to predict tender prices using 36 observations but concluded that more observations were needed. However, Emsley, Lowe, Duff, Harding, and Hickson (2002) used 288 observations to train a neural network model to predict construction costs with an average error of 16.6%.

2.6 Neural Network Models

Initially neural network models were built in the 1950's to represent the brain's organization represented as a computer system (Lee, 2019). In 2012, the neural network models significantly improved, and over the recent years the error rate has decreased further due to a combination of larger data sets, larger models and better computational abilities (Lee, 2019; Sun, Shrivastava, Singh, & Gupta, 2017). At the speed of decreasing error rate, the interest in neural network has increased across industries and the fields of application are currently growing (Lee, 2019).

A neural network model consists of layers of connected neurons with weighted connections between them (Kriesel, 2007). Each neuron is a mathematical function that calculates the weighted sum of its inputs (Lee, 2019). As illustrated in Figure 2, in a feedforward neural network model, the neurons are divided into an input layer, hidden layers and an output layer, and the neurons are directly connected to all neurons in the following layer (Kriesel, 2007).

Moreover, a feedforward neural network model with n input neurons and m output neurons transforms an input vector of dimension n into an output vector of dimension m (Kriesel, 2007). The power of a neural network model is that it learns from examples and is trained by adjusting neuron input weights based on how it performs. If the network predicts well, that input neuron will be weighted more, if it predicts poor, it will be reduced (Lee, 2019).

The first layer in a feedforward neural network model takes the entries from the input vector and forwards it to all neurons in the first hidden layer (Kriesel, 2007). Before being used as input in the first hidden layer, each entry from the input vector is multiplied by a weight corresponding to each connection between the input layer and the first hidden layer (Kriesel, 2007).

Table 4: Factors determining the bid of a project (Minli & Shanshan, 2012).

Location	Presence of the company geographically.
Construction Condition	The construction condition of the engineering such as technical equipment, human resources etc.
Market Condition	Current market situation in the industry.
The Level of Competition	Number of competitive bidders on the project.
The Following Projects	Planned future projects.
Profit Status	Previous projects profit.
Current Task	Number of current projects in progress.
Return Rate	Predicted rate of return based on performed constructions.
Type	The operating domain for the enterprise.
Scale	Price value range of the project.
Owner	Relationship between company and the contractor.
Risks	Potential risks and severity of consequences.
Lasting Time	Time frame of the project.
Complexity	Complexity of the project.

In the hidden layers of a feedforward neural network model, each neuron takes the weighted sum of all outputs from the previous layer as input (Kriesel, 2007). The input is then transformed using an activation function and forwarded to the next layer, multiplied by weights corresponding to each connection to the neurons of the next layer (Kriesel, 2007). A common form of activation function is an S-curve such as the logistic function (Kriesel, 2007). The procedure is then repeated for each layer until the last hidden layer is reached (Kriesel, 2007).

The output layer is calculated in a similar manner as the hidden layers, but the activation function of the output layer can differ depending on the type of problem (Kriesel, 2007). The output vector serves as a prediction of the desired output given an input vector, and when compared to the desired output it can be used to train the network (Kriesel, 2007).

One of the uses of neural network models is supervised learning in which a set of observed input and output vectors is used to train the network (Kriesel, 2007). In supervised learning, the objective is to adjust the weights of the network until it can predict the desired output vector given only the input vector (Kriesel, 2007). Moreover, it is desired that the weights resulting from training the network can be used to predict output vectors based on input vectors not included in the training set i.e. generalizing the predictions beyond the data provided to the network during the training (Kriesel, 2007).

The most common approach used to train neural network models is the backpropagation algorithm which uses gradient decent to minimize the error as a function of the weights in the neural network model (Kriesel, 2007). The error measures the discrepancy between the predicted outputs produced by the neural network model and the desired outputs (Kriesel, 2007). One common error function is the sum of squared differences between predicted and desired outputs (Kriesel, 2007).

2.6.1 K-fold Cross Validation

Working with small data sets, to avoid to over-fit the model yet using enough data points, the k-fold cross validation is the ideal choice (Yadav & Shukla, 2016). K-fold cross validation parts the data set in k equally sized folds and iterates k times with shuffled data to include heterogeneous data points in each fold (Yadav & Shukla, 2016). For each iteration k is used for testing and $k - 1$ is used to train the model until all folds have been used respectively as test data (Yadav & Shukla, 2016). Further is the average of the result used to evaluate the performance of the model (Yadav & Shukla, 2016).

2.6.2 Learning Curve

A learning curve shows the training, and generalization error of a model as a function of the number of training examples used (Sammut & Webb, 2011). Most empirical studies use a data set of fixed size and, therefore, does not say anything about the performance of the model with more or less data (Sammut & Webb, 2011). Using a learning curve can provide an overview of how the performance of a model is affected by adding more data, and is therefore a useful tool in finding out if additional data is needed to improve a model (Sammut & Webb, 2011).

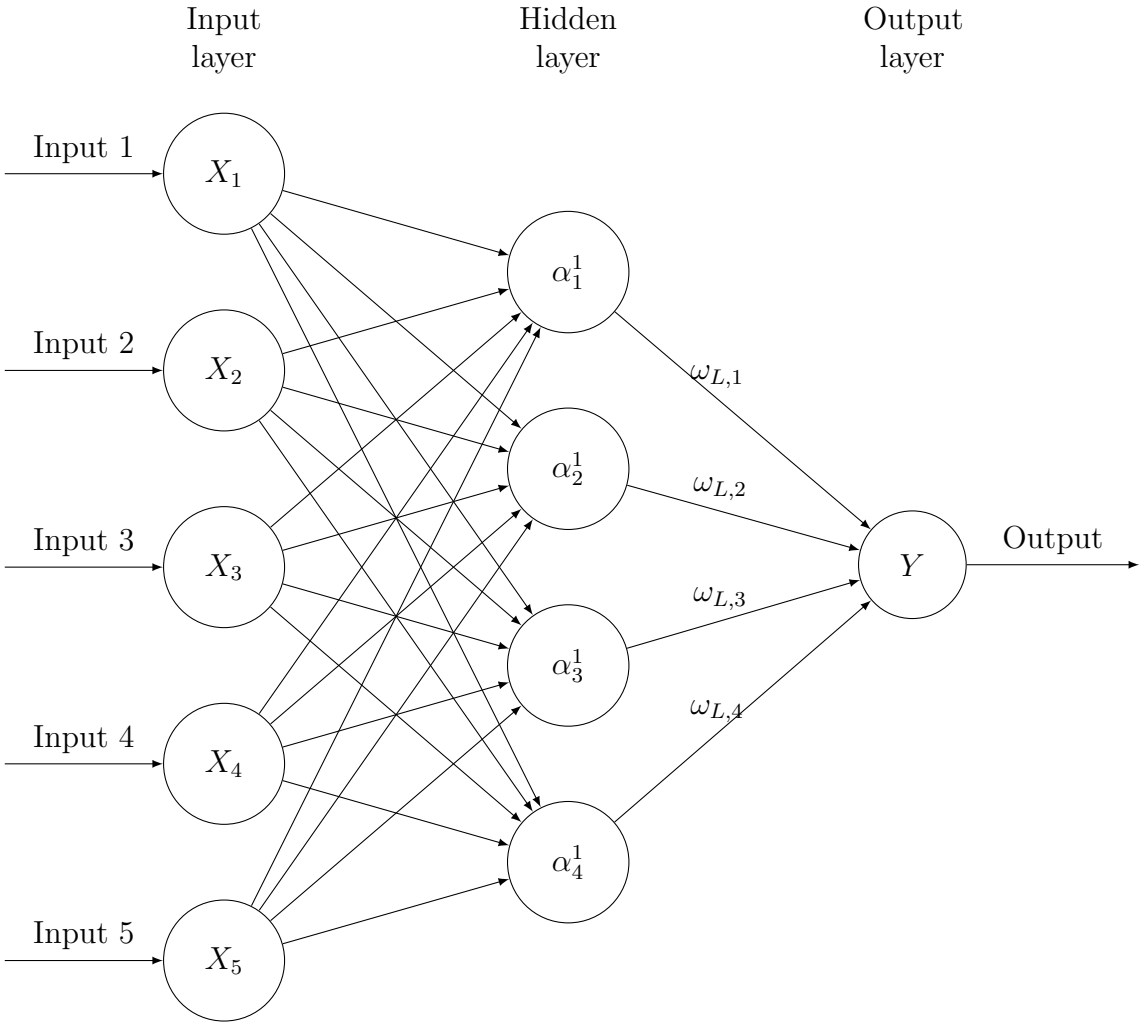


Figure 2: A neural network model with an input layer of dimension 5, one hidden layer of dimension 4 and a one dimensional output layer. Note that, in the Figure, only the connections between the hidden layer and the output layers have weights, but in reality each connection carries a weight.

3 Method

In this Section, the method of the study is explained. First, the scientific approach is described, followed by a detailed description of each component of the study.

3.1 Scientific Approach

This study employs an exploratory research approach, with a mix of qualitative and quantitative methods. The components of answering each research question of the study is illustrated in Figure 3. The first research question was answered using tender documents from Trafikverket belonging to 41 construction projects procured by Trafikverket, and an interview with a professional working with tender offers at a construction company. The purpose of the interview was for the authors to familiarize themselves with the decisions involved in the tender process, and to get first hand information about what factors affect the number of tender bids on construction projects. The findings from the interview, in combination with suggestions from literature, were used to identify data available in the tender documents that could potentially be useful when applying a neural network model to predict the number of tender bids for a project. Following the identification of data in the documents, the data was assessed in terms of structure and the quality of the data in terms of completeness and accessibility.

The second research question was answered by applying a neural network model in R, to data from the 41 projects collected in this study. R is a software environment and coding language for statistical computing (The R Foundation, 2020). First, the data identified while answering the first research question, was manually transferred to an Excel sheet by the authors. The data was then imported into R from the Excel sheet, and the package 'neuralnet' (Günther & Fritsch, 2010) was used to train the model. Due to the limited amount of data in this study, k-fold cross validation was used to evaluate the accuracy of different setups of the neural network model in terms of number of hidden layers, number of neurons, different input data, and calibration settings of the 'neuralnet' package.

The third research question was answered by proposing a process based on the course of action used for research question 1 and 2. The findings from the first research question was used to propose a process for collecting, and turning tender data from its original PDF-format to a format of sufficient quality, and proper structure to be used as input in a neural network model. The findings from the second research question was used to propose a process for processing and evaluation measures which can be used when applying a neural network model on tender data.

3.2 Data Collection

For this study, data was collected from Trafikverket and by conducting one interview, which is described in this Section.

3.2.1 Interview

A semi-structured interview was conducted for the study. The decision to use a semi-structured interview was due to the exploratory nature of the research questions, requiring

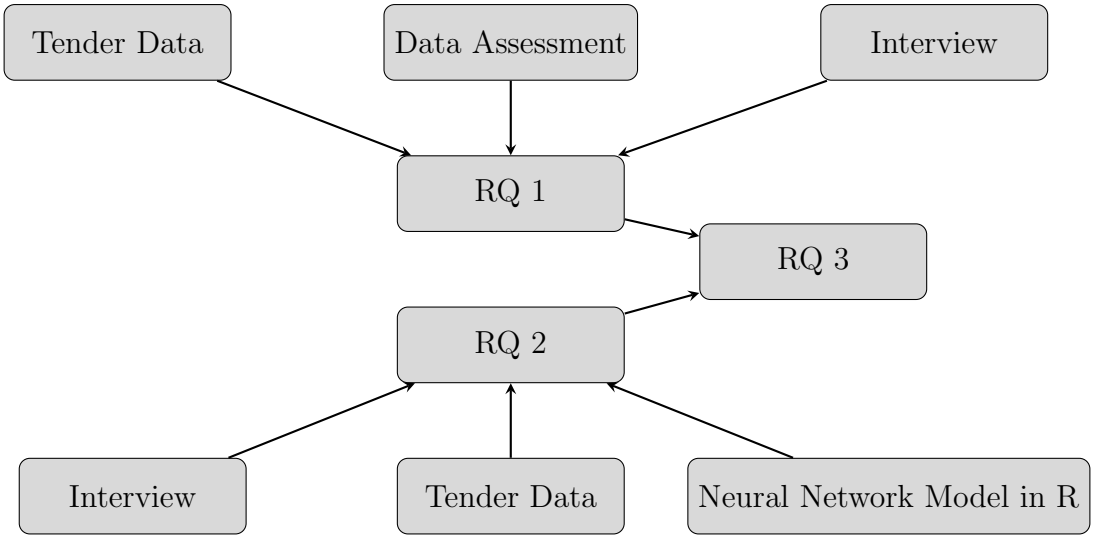


Figure 3: Logic for answering the research questions of the study.

a more open ended interview approach. The interview was held with a professional involved with both selecting projects and placing tender bids. The interview lasted for 30 minutes and was recorded and transcribed to allow for analysis.

3.2.2 Tender Data Obtained From Trafikverket

Trafikverket is the government administration responsible for long-term planning of the Swedish transport system (Trafikverket, 2020b). Documents belonging to 41 construction projects procured by Trafikverket during the period 2011-2015 were collected under the principle of public access to official documents (Government Offices of Sweden, 2020). The common denominator among the projects are that they all include at least one bridge. The process of requesting the documents from Trafikverket was divided into two steps. First, a list of all projects procured by Trafikverket during the time period 2011-2015 was requested. In order to narrow down the search, a request for all projects with the keyword "bridge" was requested.

The second step was to use the list from step one to identify the tender ID of the projects of interest. The tender ID is the identification number of Trafikverket's tenders and can be used to further request documents belonging to that tender. When the tender ID of the projects had been identified, a request was sent to Trafikverket. The request included the tender ID of the projects of interest and a declaration of what documents belonging to the project were desired. For this thesis, the following documents were requested for each tender: tender enquiry documents, tender bids, and post-tender documents. In Section 4, the documents are described in detail. For this study, documents belonging to the 41 projects were obtained after a duration of 18 working days from the request using tender ID. However, based on the correspondence with Trafikverket, the tender bid documents were the main contributor to the long lead time.

3.2.3 Business Cycle Data

To incorporate the effect of the current business cycle in Sweden at the time of a project, an indicator representing economic activity was extracted from the statistic database at the Swedish business cycle institution (Konjunkturinstitutet, 2020).

3.3 Data Assessment

The data received from Trafikverket was originally in PDF-format and sorted into folders based on the tender ID of the project that the document belonged to. The first step of the data assessment was mapping out the completeness in terms of percentage of requested documents included for each project. This allowed the authors to get an initial understanding of which documents had a tendency to be missing and therefore might not be appropriate to include in the analysis without further data gathering. The results of the mapping are displayed in Tables 7, 8, 9, and 10. The next step was to map out each type of document in terms of what the document describes and the data that potentially could be extracted from the document, which is illustrated in Table 5 and 6. The data assessment was done in order for the authors to understand whether further data processing was needed before application of the neural network model, and in order to understand the required steps in a more generalized process of using tender data to support decision in the tender process.

3.4 Data Structuring

The data collected from Trafikverket has some organizational properties and can be argued to be information for the purpose of the tendering process. However, for extracting data for the purpose of guiding tender decisions it is mainly unstructured. The data contain mostly documents with natural language text and needs to be transformed into a structured database form. This was performed by combining elements from the proposed process by Abdullah and Ahmad (2013) and the Tidy data approach presented by Wickham et al. (2014). The structure of the data set was made up by assigning each variable a column and every observation a row as the Tidy data structure suggest. Extracting the data included identifying different variables from the data and the Repositories Development guided the storage of the data to support organizational functions. In this case, storing it in an Excel spreadsheet was satisfactory for the amount of data in this study. From that format it was deemed easily accessible to transform the data into other more flexible formats for further applications.

3.5 Model Construction

The model used in this study uses the number of tender bids as the output variable. The input variables used to predict the number of tender bids are project size, location, contract type, if there is a max penalty, and the level of economic activity at the time of the tendering. Project size is the value of the winning tender offer measured in the Swedish currency SEK. The location of the project indicates what county the project was built in. Contract type refers to if the contract is a design-and-build contract or a traditional contract. The existence of a max penalty indicates whether there is an upper-limit to penalties for delays and other incidents. The level of economic activity is measured using the business cycle index from Konjunkturinstitutet. The variables location and contract type were represented

using one binary variable for each county and contract type, where 1 indicated true and 0 indicated false.

Previous studies using tender data have predicted different variables in order to strive to support tendering. Our choice of using the number of tender bids as the output variable came mainly from findings in the interview, in which this variable was stated to be useful. The input variables for the model was chosen both from findings in the literature and by the data retrieved from the interview. Furthermore, the variables were limited and chosen due to the availability to retrieve them from online sources or the documents from Trafikverket. The literature identified the variables: location, type of contract, project size and the level of economic activity. The interview supported the three first mentioned variables and added the max penalty variable.

3.6 Training of Neural Network Model

In order to evaluate the suitability of using a neural network model as prediction method of the amount of tender bids on a construction project, a neural network model was applied to the data using the software R. The findings from the interview and suggestions from literature were used in order to determine what input parameters to use. The data was imported from Excel and then normalized before being used as input in a neural network model using the existing library 'neuralnet'. As previously mentioned, k-fold cross validation was used due to the low amount of data, and the accuracy was measured using the root-mean-squared errors. In order to find the optimal number of hidden layers and neurons, a trial-and-error approach was used and the best performing combination was chosen. In order to visualise the performance, a graph with observed values and predicted values was used. Finally, a learning curve of the training and test error scores was plotted to be used for recommending possible improvements to the model.

3.7 Process for Transforming Data to Information

To answer the third research question the methodology from research question one and two are combined and thoroughly evaluated. Firstly, a decomposition of the actions from identifying projects, requiring and retrieving documents, assessing the content and structuring the data is performed in a methodological design dividing it into necessary steps. Between the assessment and the representation of the data the parameters are chosen, influenced by the choice of model, collected data and existing literature. Thereafter are the steps from making the structured data compatible to be used as input in the model extracted. Further in the process, after running the model, the steps to validate the settings and evaluate the results are decomposed and extracted as the last steps finalizing the suggested process.

4 Data

In this Section, the data collected from the interview and the documents from Trafikverket are presented.

4.1 Interview

For a construction company, the tender process consists of three steps: first a meeting where all parties gather to discuss potential projects, then the actual tender bids are calculated for the decided projects, and finally the tender offers are sent to management for approval. During the initial meeting, a bid/no-bid decision is made concerning the projects currently accepting tender bids. The interviewee reported that the decisions are mainly based on the intuition of the participants, about factors such as expected number of tender offers and complexity of the project. The interviewee also mentioned that projects are often selected to maximize utilization of the workforce and that projects similar to the ones that the firm has previously performed well on are often selected. When the bid/no-bid decision has been made, the second step is to calculate the amount of the tender bid.

Futhermore, the interviewee mentioned that the calculation procedure varies depending on what type of contract is being worked on. A design-and-build contract requires the participation of a structural engineer which, according to the interviewee, increases the effort required for projects with design-and-build contracts. On the other hand, a traditional contract requires more man hours than a design-and-build contract, but is still deemed to require less effort. The final step of calculating the tender bid value is to add a margin to the calculated costs. The interviewee mentioned that depending on the current utilization, the size of the margin may be adjusted, and that all participants in the calculations also tend to make small adjustments themselves, making the adjustment of the margin bigger than intended. Once the tender bid value has been calculated, the last step is to get approval from management where the size of the project determines how high in the organization the approval needs to come from.

Moreover, the interviewee mentioned three types of information that is deemed useful for supporting decisions in the tender process: expected amount of tender bids on a project, spread in tender bid values for different types of projects, and company performance on different types of projects. The interviewee elaborated that given this information, the participants in the tender process can identify which projects are likely to have many competing offers, or which projects they have low performance on, and thus spend less resources on placing tender bids on projects where the likelihood of winning is low. When asked about what distinguishes construction projects, the interviewee mentioned complexity, location and the size of the project. In this case, size refers to the price level of the winning tender bid. Moreover, it was mentioned that one way to measure complexity is depending on if the projects include water or just roads. The interviewee also concluded that while reviewing the inquiry documents, one thing in particular that is reviewed, is the maximum penalty for the project since projects without a limit on penalties for incidents such as delays might not be worth tendering on.

According to the interviewee, sometimes Trafikverket adds bonuses depending on parame-

ters such as societal factors or sustainability. To illustrate they mention a "green project" initiative that Trafikverket has used. However, the interviewee continued by stating that these types of extra bonuses for reaching certain criteria are only viewed as bonuses and not something they calculate into the tender bid to lower it. Further, the interviewee argued that due to the frequent occurrence of appeals against procurements of Trafikverket by the trade union, bids with more sustainable choices rarely got chosen over the lowest price.

4.2 Documents from Trafikverket

The data collected from Trafikverket consists of tender documents belonging to the 42 construction projects used in this study. In this Section, the properties and completeness of the tender documents are presented, and then the the data, extracted from the documents and used in the neural network model, is described.

4.2.1 Tender Documents

The tender documents received from Trafikverket can be placed into three categories: tender enquiry documents, tender bid documents and post-tender documents. Tender enquiry documents are the documents sent out by Trafikverket when starting a request for tenders. The tender enquiry documents have six main categories which together describe the project in question, and the requirements on companies making an offer. The tender bid documents are the actual tender offers sent in by each company, consisting of a signed tender form confirming their tender, and a priced content specification which lists the offered price for each item in the project. The post-tender documents are made when the request for tender is completed and consists of a tender protocol listing all tender offers received, and a signed contract between Trafikverket and the company who provided the winning offer. In Table 5 and 6, the documents within each category is described along with the identified data included in each document. The documents were originally in PDF format, sorted into folders based on the project that they belonged to. Within each project folder, the archiving of the documents varied between being sorted into folders based on what category they belonged to and not being sorted at all. Moreover, in several cases, all documents belonging to a project was not included, meaning that the concerned documents were not sent to us from Trafikverket, who claimed not to have them. Tables 7, 8, 9, and 10 describes the extent to which the requested documents were included or missing.

4.2.2 Tender Data

As described in Table 11 and 12, five types of variables were extracted from the documents of each project to be inserted into the neural network model. The number of tender offers on each project, which was chosen to be output variable in the neural network model, was extracted. Based on the results of the interview, project size, location, contract type and max penalty was also extracted from the documents in order to be used as input variables in the model. For variables describing location or contract type, one binary input variable for each possible location and contract type was used. A "1" was used when the location or contract type was true and a "0" was used for false. For project size, the amount of the winning tender offer, expressed in SEK, was used. In addition, a business cycle index from Konjunkturinstitutet, indicating the economic activity on the date the project was contracted was used as input parameter into the neural network model.

Table 5: Description of documents and data included in tender enquiry documents.

Document	**Description**	**Data included**
Tender Enquiry Documents		
Administrative Requirements (AR)	Specification of the conditions under which the project is to be conducted, including a loose description of the project, specifications of the tender process, and other demands such as work-times and salary requirements.	• Type of project • Project location • Type of contract
Tender Form (TF)	The tender form is the central document of each tender bid where bidders fill in information about their firm, bid amount, and confirm that they have the economic and technical capacity required for the project.	• Required credit score and/or risk class • Required technical capacity
Contract	The contract to be signed by Trafikverket and the winning tender once the procurement is finished. The contract specifies the project loosely along with payment conditions, penalties, and the identity of the winning tender.	• Type of contract • Project location
Invitation Letter (IL)	The original call for tenders sent out by Trafikverket, specifying the name and location of the project along with information about how to leave a tender bid.	• Tender ID • Project location
Content Specification (CS)	A detailed breakdown of the project in the form of a table of items that are being requested and in what amount.	• Items included in the project • Amount of each item in the project
Building Plans (BP)	Detailed descriptions of each part of the project in the form of blueprints.	• Visualization of objects in the project

Table 6: Documents and data included in tender bids and post-tender documents.

Document	Description	Data included
Tender bids		
Signed Tender Form	The tender form provided by Trafikverket filled out by each bidder with identify of bidder, tender amount, attestation of economic and technical capability.	• Type of contract • Project location • Identity of bidder • Tender value • Required credit score and/or risk class • Required technical capability • Confidentiality request (Y/N)
Priced Content Specification	The content specification provided from Trafikverket, filled out by each bidder with the price for each item listed in the form	• Items included in the project • Amount of each item in the project • Unit price for each item
Post-tender documents		
Tender Protocol	A summarizing protocol that briefly specifies the project in question, the identity of all participating bidders and the value of their bids.	• Tender ID • Project location • Identity of bidders • Tender value • Winning tender • Date
Signed Contract	The contract provided by Trafikverket, signed by the winning bidder.	• Type of contract • Project location

Table 7: Overview of missing documents from tender data of Trafikverket (I=Included, M=Missing).

Tender Enquiry Documents	Tender Bids	Post-Tender Documents	Cases	Share [%]
I	I	I	31	74
I	I	M	6	14
I	M	I	4	10
M	M	M	1	2
Share included [%]	Share included [%]	Share included [%]		
98	88	84		

Table 8: Overview of missing tender enquiry documents from Trafikverket (I=Included, M=Missing).

AR	TF	Contract	IL	CS	BP	Cases	Share [%]
I	I	I	I	I	I	16	38
I	M	I	I	I	I	12	29
I	M	I	I	M	I	5	12
I	I	I	M	I	I	3	7
I	M	I	M	I	I	2	5
I	I	I	I	I	M	2	5
I	I	I	I	M	I	1	2
M	M	M	M	M	M	1	2
Included [%]	Included [%]	Included [%]	Included [%]	Included [%]	Included [%]		
98	52	98	86	84	93		

Table 9: Overview of missing post-tender documents from Trafikverket (I=Included, M=Missing).

Tender Protocol	Signed Contract	Cases	Share [%]
I	I	39	93
M	I	2	5
M	M	1	2
Share included [%]	Share included [%]		
93	98		

Table 10: Overview of missing tender bid documents from Trafikverket (I=Included, M=Missing).

Signed Tender Form	Priced Content Specification	Cases	Share [%]
I	I	50	32
M	I	47	30
I	M	22	14
M	M	37	24
Share included [%]	Share included [%]		
46	62		

Table 11: The variables used to train the neural network model. For location and contract type, one binary variable was used for each location and contract type.

Variable	Description	Type
Project Size	The amount of the winning tender offer	Continuous
Business Cycle Index	Measure of level of economic activity	Continuous
Location	County of the project	Binary
Contract Type	Design-and-build contract or traditional contract	Binary
Max Penalty	Is there a maximum penalty or not	Binary
Tender Offers	Amount of tender offers	Integer

Table 12: Properties of the data used as input for the neural network model.

Data	**Mean**	**Standard Deviation**
Project Size	48 067 529,7	104 570 353,5
Tender Offers	3,8	1,5
Business Cycle Index	98,2	4,8
Data	**Amount 1's**	**Amount 0's**
Build-and-design contract	11	30
Traditional contract	30	11
Max Penalty	33	8
Uppsala county	1	40
Jönköping county	2	39
Blekinge county	1	40
Halland county	4	37
Västra götaland county	11	30
Värmlands county	1	40
Örebro county	3	38
Västmanland county	2	39
Dalarna county	2	39
Västerbotten county	9	32
Norrbotten county	5	36

5 Results

In this Section, the process identified in this study is presented, using Trafikverket's tender documents and a neural network model for predicting number of tender bids as an illustration. Thereafter, the results of using a neural network model to predict number of tender bids is presented.

5.1 Process for Transforming Tender Data

In Figure 4, the process for gathering and transforming tender data into information supporting tender decisions is illustrated. The first step of the process is collection of documents belonging to the projects of interest. In order to collect documents from Trafikverket, a list of projects for the desired time period needs to be retrieved. The list is then used to decide what projects, and documents are desired for the rest of the process. Then an inquiry needs to be sent to Trafikverket, listing the projects and documents that are decided upon. For this study, the process of collecting documents from Trafikverket's archives took 18 working days for documents belonging to 41 projects. An important issue in the document collection step is the choice of what projects to collect documents for. Before having the possibility to review the documents, there is relatively little information available on each project. Therefore, the initial choice of what projects to collect documents for can easily become subjective.

The next step is to examine the available data in the documents, and assess the structure and quality of that data. Examining the data available is done in order for the practitioner to know what potential data can be extracted from the documents and how it can be used. Table 5 and 6 shows the available data identified in Trafikverket's documents. Assessing the structure and quality of the data is done in order to determine the feasibility of different potential applications of the data. For example, an application requiring data that is too unstructured or missing many observations could be excluded based on the structure and quality assessment. As illustrated in Table 7, 8, 9, and 10, the completeness varied between the different document types obtained from Trafikverket. While most documents had an inclusion rate of over 80%, tender bid documents and tender form had an inclusion rate of less than 65%. Thus, for applications requiring tender bid documents or the tender form, a significantly larger set of data would have to be collected than for applications using the other documents.

Once documents have been collected and assessed, the next step is to decide what model and parameters to use. This includes the decision of which input and output variables to use in the model as well as what type of model to use. The choice of input variables refers to the data which the model will be calibrated to, and the output variables refers to the information that the model will provide using the input variables. Decision of model type means deciding on what method to apply the data to, in this study a neural network model was used. It is necessary to choose a set of variables and a model which are compatible with each other and with the overall goals of the process. For example, in this study the goal was to support the decisions made in the tender process and therefore a prediction model was chosen and a set of variables that were compatible with that.

Having decided on a model and variables, the next step is to transfer the variables from

the documents to a format that can be imported into the used computing software. The transfer can be done manually or using automated software depending on the amount of projects and variables used, and what structure the data is within the documents. For unstructured data, automating the data transfer is more difficult, and for large data sets or many variables, a manual transfer is time consuming. Initially we aimed to automatically add data and use software to read text from PDF which would be the only efficient way to approach this step with large data sets. As the study unraveled it became clear that the most time efficient way was to manually transfer data from the unstructured format from Trafikverket into a structured way, using the tidy data framework as reference for building the structure. This became the best solution for two reasons: the small amount of data in our data set and the fact that creating an algorithm to read the PDFs and structure them automatically was outside of the scope of this study. Though it is crucial for the scaling of this process to efficiently collect and structure data from unstructured format with an algorithm or developed software for the purpose.

The next step is to process the data into a format that can be used as input to the chosen model. Depending on the choice of model and variables, the processing required may differ. Some examples from the present study is transforming the variables location, contract type, and max penalty into binary variables where 1 indicates true and 0 indicates false. Moreover, since location and contract type consists of several locations and contract types, one variable for each possible location and contract type had to be made. Moreover, since the different input variables had different scales, all input variables were normalized to values between -1 and 1 before being used. When the processing is complete, the resulting variables are put together in a matrix for use in the model in the next step.

When the data has been processed, the next step is the calibration of the model. The calibration step consists of configuring the calibration and executing it. Configuring the calibration refers to deciding what settings to use when calibrating the model to the data. Possible settings are different depending on what model is used, some examples from this study are hidden layers, neurons, threshold, maximum steps, and repetitions. Hidden layers and neurons refers to the decision of how many hidden layers and neurons to use in the neural network model. Threshold is the value of the partial derivatives of the error function used as a stopping criteria for the calibration. Maximum steps is the maximum amount of steps that the gradient decent algorithm will do before stopping the calibration. Repetitions are the number of times that the calibration is performed in order to find the optimal solution. Varying the configuration of the calibration affects both the results of the calibration and the time consumed for calibration.

Having calibrated the model, the next step is to validate the performance of the model. This step includes deciding on one or several validation methods and applying them to the model to examine the performance of the model. The choice of validation method depends on the goals of the study and the model used. In this study, the RMSE of predictions were used as a validation metric. When the validation metrics has been computed, the next step is evaluation. During the evaluation step, the performance of the model is evaluated using the validation metrics calculated previously. The goal is to determine whether the performance is good enough or if more iterations are required. If possible, the performance should also be diagnosed, in order to understand the causes of good or bad performance.

Once performance has been evaluated and diagnosed, a decision needs to be made whether the performance is good enough to end the process or if it is necessary to iterate back to previous steps to improve the model.

At the end of the process, one can iterate back to document collection, document assessment, choice of model & parameters, data processing, or model calibration. Data transfer and validation are not included in possible steps to iterate to since they mainly support the previous steps and therefore iterating to them is not likely to add value. If the size of the data set is deemed to small, or more parameters are needed, iteration to one of the first three steps is appropriate. By iterating to step one, more documents can be collected to increase the size of the data set or collect new types of documents that include new parameters. By iterating to step two, new variables can be identified from the collected documents and eventually added to the model. By iterating to step three, variables that were not included previously can be added to the model in order to improve performance. If performance of the model is diagnosed to be too low due to poor model choice, data processing methods or calibration configurations, one can iterate to step three, five, or six. By iterating to step three, the model type can be changed. By iterating to step five, the methods used for data preprocessing can be adjusted, and by iterating to step six the calibration settings can be adjusted to improve performance.

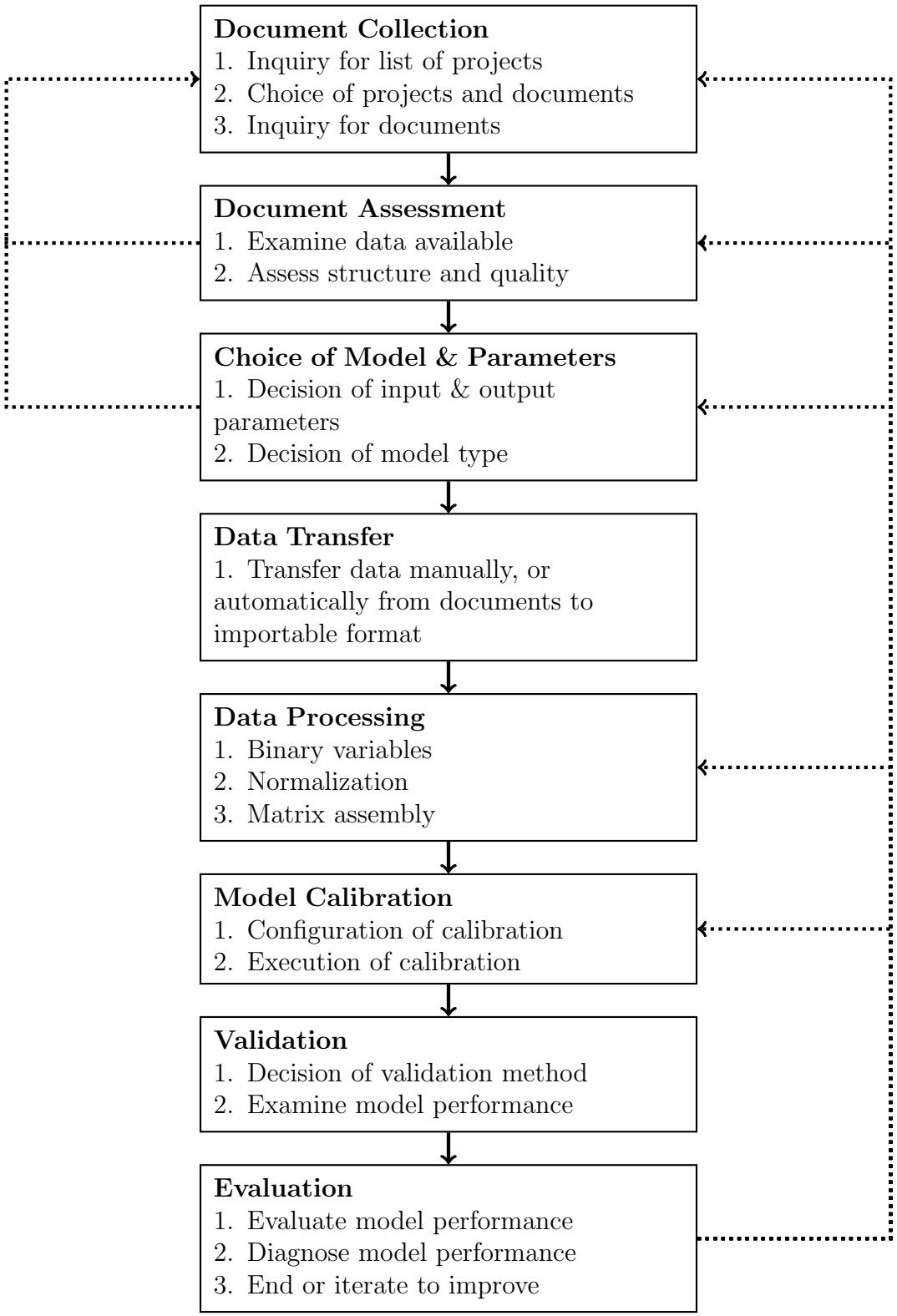


Figure 4: Process for transforming tender documents from Trafikverket into usable information for the tender process.

5.2 Performance of the Neural Network Model

To evaluate the model with the input data presented in Table 11 and 12 it must be calibrated. The step of calibrating the model includes the decision of settings for the model such as number of hidden layers, number of neurons, threshold, maximum steps and repetition. These can be set based on using existing studies with similar prerequisites or as presented in this Section, by performing a number of iterations to optimize the number of hidden layers and number of neurons in each layer. One crucial aspect here is to find the solution giving the best results while not using too much computational power to do so. It is preferable to have as low number of hidden layers, for example, as possible to make the calculations use as little energy as possible. From the perspective of sustainability, this is important if a model like this was to be scaled further.

First in the calibration, the number of hidden layers was varied with an arbitrary fixed number of neurons in each layer. In this case 10 neurons in each layer. In order to validate each iteration, k-fold cross validation with k=4 was used. The root mean squared error (RMSE) between predictions and observations is the quantitative measure that has been used as validation metric for the performance of the predictions. The higher the RMSE, the further away from the observations are the predictions. As can be seen in Table 13, the number of hidden layers resulting in the lowest RMSE is three. Following, the most effective number of neurons in these three layers must be evaluated. To approach this, multiple tests with few to many neurons in each layer where tested. In Table 14 the RMSE for each number of neuron is presented and it is concluded that the initial number of 10 neurons in each of the three layers predicts most accurate at this point. Moreover, when using two and three layers, the algorithm was not able to converge to a solution.

Table 13: RMSE of varying number of hidden layers with 10 neurons in each layer

Variations	**RMSE**
1 hidden layer	6,49
2 hidden layers	2,49
3 hidden layers	2,45
4 hidden layers	2,53
5 hidden layers	2,58

Table 14: RMSE for 3 hidden layers with differing number of neurons in the layers

Variations	**RMSE**
2 neurons in each layer	no result
3 neurons in each layer	no result
5 neurons in each layer	3,03
6 neurons in each layer	2,54
10 neurons in each layer	2,45
15 neurons in each layer	3,02

The predictions with the model of three hidden layers with ten neurons in each layer is presented in Figure 5. In which the observations are plotted as a line and the predictions of the neural network model are plotted as dots. The results in Table 13 indicates that a neural network model with 3 hidden layers of 10 neurons each produces the lowest RMSE of 2,45. Compared to the average number of bidders being 3,8 the RMSE can be considered

significant and the predictive power of the model in its current state is therefore low. Figure 5 shows that a majority of predictions are within the same range as the real values, but several outliers on the lower end are likely to be contributing to the significant RMSE. One prediction, as can be seen, is negative but most of the remaining predictions are within the same range as the observations which range between one and seven tender offers. However, few predictions are at the exact number as the observation. Moreover, out of the predictions that fall within the same range as the real values there still does not seem to be any significant correlation between the predictions and the real values. Therefore, it is possible that the model is generating values within the same range rather than actually predicting the number of tenders based on the project characteristics.

For validating the model we chose RMSE to measure how far away from the observations our predictions were. This is a very exact validation measure that validate the performance of the model in a clear way. However, the real value doesn't necessarily lie in how exact the model predicts in comparison with the observation. It lies in seeing whether or not the model predicts better than the current practice in place. Hence we suggest that it would be interesting to evaluate the current intuition-based predictions in order to understand how well a model would have to perform in order to be an improvement from current practices. This could be done by presenting several projects to an experienced practitioner and let that person guess the number of tender bids. The guesses by the practitioner could then be compared to the predictions by the model to see which performs best.

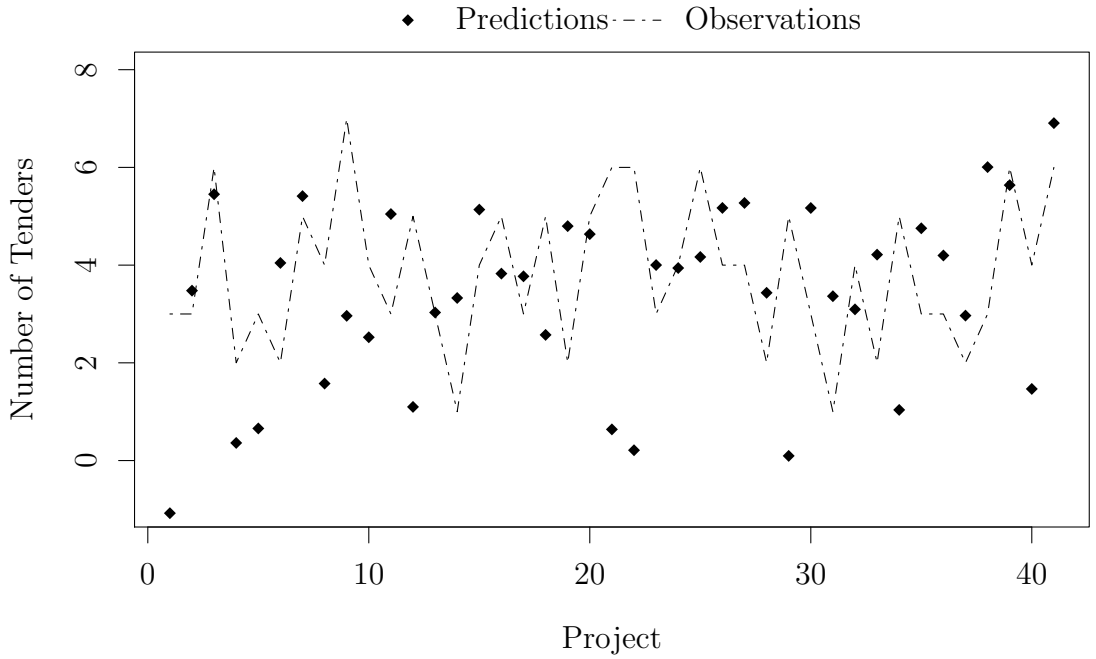


Figure 5: Predictions from the model with $k = 4$, 3 hidden layers and 10 neurons in each layer.

Figures 6 and 7 illustrates the learning curve of the model for test size chosen to be five.

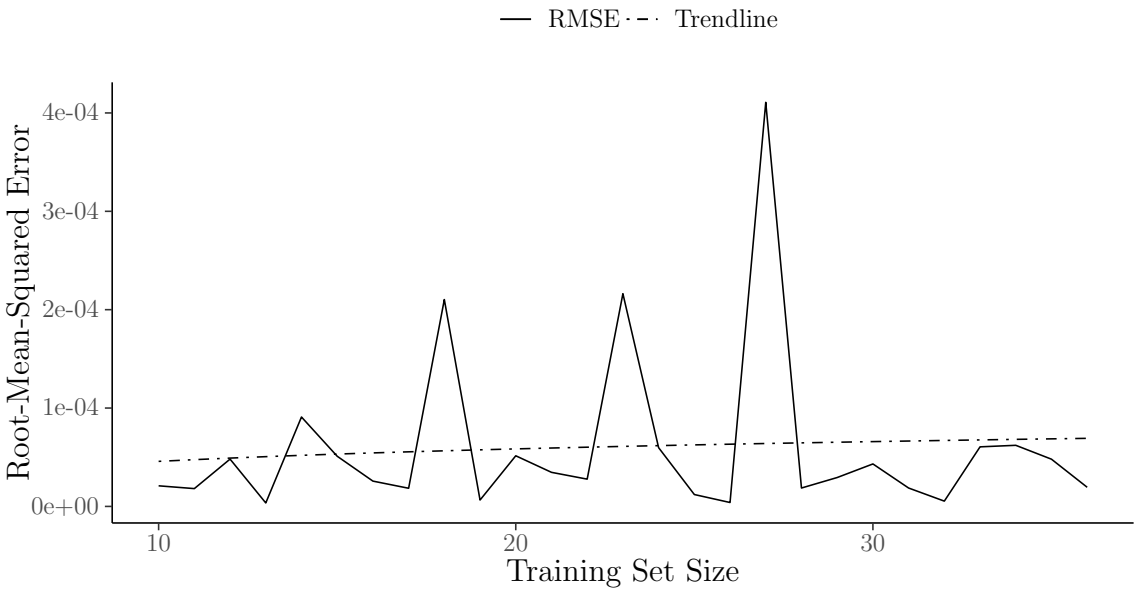


Figure 6: Learning curve on training set.

Figure 6 plot shows the learning curve for the training set which plots the RMSE for the model for each test size. The trend line in the Figure is weakly positive which is expected since the ability to create a model to exactly predict all data points becomes more difficult the more data points there are. Figure 7 illustrates the plot for test data and the trend line is negative. This trend is also expected since the RMSE should become smaller and smaller the more data the model gets to train on. However, it is worth mentioning that during different runs, the graphs in Figures 6 and 7 varied quite a bit and the positive respectively negative trend are inconsistent from one run to another. The test data learning curve in Figures 6 and 7 does not indicate a clear pattern when more training examples are used. Instead, as more training examples are added, the RMSE fluctuates with a slightly negative overall trend indicated by the dotted line in Figures 6 and 7. Outside the scope of this study, there are potential reasons for this behavior left to be investigated and eliminated.

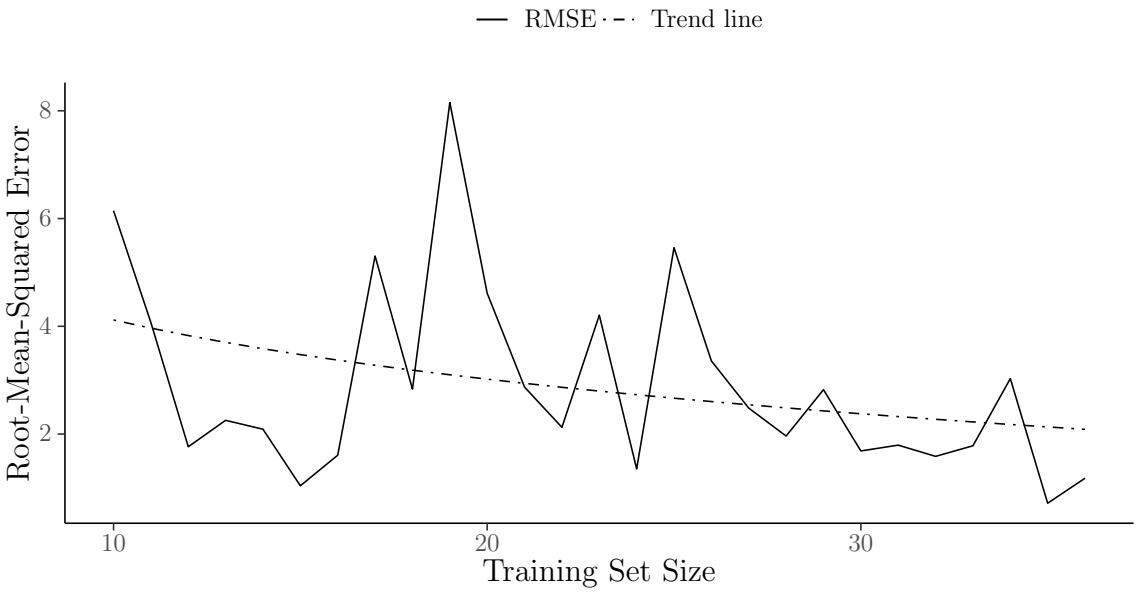


Figure 7: Learning curve on test set.

6 Discussion

The following Section contains a discussion of different aspects of the study.

6.1 Accessibility and Usefulness of Tender Data from Trafikverket

In order to deploy a model based on tender data in practice, it is likely that a larger data set than the one used in this study is required, which would require considerably more time to collect from Trafikverket's archives. Therefore, for future attempts it would be beneficial to limit the document collection to only post-tender documents and content specifications, but for a larger number of projects. This approach would reduce the amount of documents for each project, yet provide many of the important project information types mentioned in the interview. Moreover, as public institutions become more digitized, it is possible that the data will become available without having to request documents from Trafikverket. An example of this is Tender Electronics Daily, which publishes some information about European public procurements in csv-format. Such initiatives could mean that the importance of data driven decisions in the tender process will increase as tender data becomes more accessible.

A growing interest in data-driven decisions could increase the pressure on Trafikverket as companies become more interested in using tender data. Moreover, if tender data can be used to increase the share of tenders won for a contractor, there is less need to compensate for lost tenders in the bid price. This would result in lower bid prices (Ngai et al., 2002), and thus it would be in the interest of the procurer to make the data available to contractors. Therefore, making data from past projects available digitally could benefit Trafikverket in two ways. First, if the interest for this kind of data grows, it is likely to be the most efficient way to comply with the principle of public access. Second, according to Ngai et al. (2002), it is likely to result in lower prices for future projects.

The document assessment in Tables 5 and 6 revealed that the number of tender bids and their value are included in the data from Trafikverket. The three metrics that the interviewee deemed useful for supporting decisions in the tender process are the following: expected number of tender bids, spread between tender bids, and company performance on different project types. These are included in or can be calculated using the tender documents from Trafikverket. Moreover, several variables, such as the possible presence of a maximum penalty, the size of the project, and the type of contract used, that affect the number of tender bids were also included in the tender documents. Thus, judging by the available data, the documents retrieved from Trafikverket show potential for being used as support for decisions in the tender process. However, no measure of project complexity was used in this study as this would require a method for quantifying the complexity of a project based on information in the tender documents. The size of the projects were also measured as the value of the winning bid, which is only known after a tender is complete and therefore, can not be used in practice to predict the number of tender offers on a project. Therefore, we suggest that an important next step in using tender documents to guide decision making in the tender process is to use the content specification to represent the contents of the projects numerically. Since the content specification lists all items included in a project using a standardized reference system, it is reasonable to assume that it can be used to represent the

differentiating features of a project numerically for input into a model.

While reviewing the structure of the data from Trafikverket, it is easy to believe that it will be less cumbersome to transfer unstructured data into structured data in the future as software reading PDF increases in its performance. A large part of this study was identifying, finding, and structuring data in order to assess it properly. It is highly recommended to map the ratio of documents received to documents requested, in a similar manner as in Tables 7, 8, 9, and 10. This process is fairly time consuming; nevertheless, it is crucial while reiterating to include new parameters and to assess the quality of the data regarding completeness. Using the mapping of represented documents, one can easily assess which documents are rather complete over the data set and by that investigate what possible parameters can be extracted from those documents. Thus, an interesting parameter found in one project is quite useless if that specific document only exists in a small portion of the projects in the data set. The mapping can also be useful to see what part of the documents are missing and request this specific information from Trafikverket. If many projects have fragmented documentation, it is important to note since that would imply Trafikverket does not fulfill their duty as a government authority.

One important aspect of this study is the small amount of data it is based upon. Due to cumbersome access from Trafikverket and time restrictions for the thesis, the data set is smaller than ideal. Since the study aims to find patterns and identify features in the data to support the tendering decision making process, it would have been better with a larger amount of data to have each parameter represented in multiple cases. One example where the lack of representation in the data becomes clear is in the geographical regions where "Uppsala", "Blekinge" and "Värmlands" county only appear once. When using k-fold cross validation, during every fold, each observation is used either for training the neural network model or for testing how well it generalizes. Thus, the geographical regions only included in one observation is either used for training with no possibility for testing generalization, or for testing generalization without training the network with observations from that region. Moreover, in order to learn the interdependencies between input variables and output variables, it is likely that many observations on each variable are required.

Due to the rigidity in Trafikverket's archives and the the time-consuming nature of collecting large quantities of historical data, this approach is inefficient. Another way to collect data would be to systematically collect all the projects that a construction company tender on and through that, over time, build up a solid database of projects. If each tender practitioner in a company were to collect all documents on each project in their field, a database could be compiled and quickly reach the size necessary to build a representative model. In this way, construction companies can make use of a larger set of data which arguably would speed up the development of practices that utilize data. Assuming all companies utilize data from Trafikverket, it would give a competitive advantage to gather additional data from projects as early as possible.

6.2 Representative Input to The Model

One of the greatest challenges encountered when modelling projects with at least one bridge was finding input parameters that capture the heterogeneity among projects. For example,

the projects used in this study varied greatly in terms of size and content; some of the projects' contents were specific to an area such as moving or removing existing things on the site. When projects are this diverse, it is challenging to find a set of input variables that can fully capture the unique properties of each individual project. This is a possible explanation for why the learning curve in this study does not show a strong negative trend when more examples are used to train the model. In order to model the type of projects used in this study, it is likely that more parameters, that capture the characteristics of each project in more detail, are needed. The parameters proposed by authors such as Drew and Skitmore (1997) and (Minli & Shanshan, 2012) provide some guidance, but are often abstract measures such as risk and complexity. In order to use such parameters as input in a model, it is first necessary to develop a method of quantification. Quantifying parameters such as risk and complexity requires knowledge about the type of project in question. Therefore, we argue that domain knowledge is a prerequisite for future work in finding input parameters to capture the characteristics of a project. In the following paragraphs, some important parameters for modelling the number of tender offers are discussed. These are based on the contents of the inquiry documents from Trafikverket, the findings from the interview in Section 4.1, the parameters used in previous studies presented in Section 2.5, and the assumptions from the authors themselves.

Looking at possible factors affecting the number of tender bids for a project, it has been implied both in the interview in Section 4.1 and in Section 2.5.2 that the level of economic activity might affect tendering behaviour. In Section 2.5.2 it is presented by Ngai et al. (2002) that when the economy is on a high level of the business cycle, there will be more possibilities, and companies will be utilizing their capacity more. Conversely, when the economy is low in the business cycle, the competition might increase due to more slack in capacity utilization and, therefore, more companies wish to contract each project. However, assuming that, in a low business cycle, one of the counteractions that governments use to stimulate the economy is initiating infrastructure projects, it could result in the opposite effect on competition in tendering compared to what Ngai et al. (2002) argued. Moreover, in this study, the data only contained projects performed in a period with a positive economic trend; hence, the model only trains with optimistic business cycle indicators. Over time however, and with a larger data set, the business cycle indicator could arguably be used as one factor to predict competition, price levels and number of bids. On the other hand, given the arguments above, the business cycle indicator could result in both higher and lower competition and given it is not consistent, the parameter could be contradicting itself. Thus, the correlation between bidding competition and business cycle indicators must be established before it can be factored in the model. Arguably domain knowledge of how the business cycle usually affects the industry would be beneficial in establishing this correlation.

In order to characterise unique features of the project and to define the content of the project, complexity must be included as an input parameter. The complexity can be defined in different ways and can be measured based on different factors. One possibility is to define complexity by the number of specified variables in the content specification found in the tender enquiry documents from Trafikverket. Here it can be assumed that the more defined the content specification is, the easier the execution of the project and the less risk for the contractor. This would lead to higher competition in projects that are well defined and subsequently more similarly valued tender bids. The complexity of a project could

also be defined by including the combination of project components such as roads, water infrastructure, bridges and railways. Projects that are more complex and of higher risk are more likely to have lower bidding competition as fewer companies will be willing to take them on. This can be illustrated with the following example. A company can choose to bid on two projects. One is the construction of a bridge over a railway, and the other is the construction of a bridge over a body of water that does not have any traffic. The first project is more complicated and has a higher risk as the traffic can be shut down for only a short time, whereas the second does not have this limitation. Therefore, more companies will bid on the second project.

Furthermore, the geographical area of a project can define the complexity. From the interview it became clear that even though one can generalize between geographical regions, there exists geographical differences within the regions making an easy project in one place significantly more difficult 200 meters further away. Based on this, an idea for further development of the model would be to specify the geographical parameter more narrowly, possibly using the coordinates of the project rather than the county. However, it can be assumed that in order to fully add value to this data, the geographical differences must be weighted and evaluated by someone with domain knowledge.

These input parameters would help the model to predict what we already know, possibly in a more rapid and accurate way. However, the real value would lie in getting the model to predict similarly to the predictions based on the intuition of an experienced tenderer. By capturing the essence of intuition based on experience we would argue that the predictions would be even better than the tenderer and, thus, would be able to support the tenderer in making their decisions. Having this in mind, in the interview it was stated that there is no easy way to define the complexity of a project. The geographical aspect, the size of the project and its components might all affect the complexity. Both domain knowledge used to identify what is usually taken into account and a more extensive correlation plot to identify relationships between factors are required.

Reviewing the model's sustainability and usefulness over time, we must analyze the possibility to add parameters that may not be useful today, but may be crucial in the future. Ecological sustainability is one such parameter. Sustainable choices seem to be more and more important for companies in all industries. Based on the data in Section 4.1 there is an emergence of monetary bonuses for projects performed in accordance with sustainability goals specified by Trafikverket. However, this is not something that currently is accounted for in the calculations of a tender, but rather something that can be viewed as a bonus after the project. Green initiatives have existed in Sweden, but Trafikverket almost exclusively chooses the lowest tender due to the simplicity in the selection process. Supposing that in the future the maximum amount of CO_2 or other similar resource consuming measurements can be included in the enquiry documents, eco-requirements, such as minimum thresholds or excesses weighted as additional fees, can be added to projects. It is possible to add a parameter for ecological sustainablity (e.g. CO_2 consumption) into the model, enabling it to continuously support decision making in tendering over time, even as prerequisites change. This possibility is important enough to highlight the value of proposing a universal process such as the one in Section 5.1.

6.3 Evaluating the Process

Based on the foundation of the DIKW hierachy and the KDD process, the intention with this study was to propose a detailed process of how to transform data into useful information. In the DIKW hierachy the definition of data is, in essence, observations without meaning or structure, and information is data that has been structured to give meaning and value to a recipient. This is the spirit from which this study has been performed. However, the definitions from the DIKW hierachy merely provide a way of thinking but not guidance for how to transform the data into information. Thus, it is too vague to be solely satisfactory for constructing a data transformation process. On the other hand, the KDD process has been of great importance at guiding the construction of the different steps. In the KDD process each step is displayed for transforming data into knowledge. The proposed process in this study is similar to the nine step KDD process, but is intended to be more focused on the settings of the construction industry.

The suggested process from this study includes the steps "Document Collection", "Document Assessment" and "Data Transfer", which the KDD process does not. While the KDD process assumes that users start with a structured data set, the construction industry requires three additional steps. Due to the fragmented data practices within the industry, the KDD process is meant to be universal while this study aims to be more detailed, yet still adaptable to different tender processes. Thus, it is common for a construction company to use different data, modelling techniques and/or decision situations than those considered in this study. In this study a single path for the process was chosen, and other alternatives were considered but not tested. For the process to be applied in other situations, the decisions in each step should be adapted to its specific requirements.

During this study, modelling the number of tender bids using a neural network model was chosen in order to evaluate the feasibility of using an advanced algorithm on tender data. This model was chosen to illustrate one possible way to perform the process suggested in Section 5.1. However, other model types can be applied to tender data, and the process is intended to be universal, including multiple choices along the different steps. For example, if collecting tender data using the process described in this study proves to be insufficient for the amount of data required by a neural network model, other methods could be explored. Some interesting examples could be correlation studies or by visualising the data in order to bring out patterns. Correlation studies could be used in order to understand if there is a correlation between, for example, project characteristics and tender bid performance for different companies. Further, simple visualisations are interesting to find patterns. Similar to the findings of Drew and Skitmore (1997), which plotted tender offer performance as a function of project size, other patterns could be explored by visualising the data. Contrarily, given that the data set in the future would be large enough, regression analysis could be used for predictions as an alternative to a neural network model. While regression analysis has been argued to be less suitable than neural network models for construction projects (Kim et al., 2004), it allows for a more detailed analysis of the effect of each input parameter on the output parameter. Such information could help practitioners develop a better understanding of what factors contribute most to the environment they work within.

In order to fulfill the overall purpose of the study, to utilize data to guide the tender process instead of using intuition, the subjective choices made in the process' different steps must be

acknowledged and systematized. As suggested when quantifying the input parameters into the model, a certain level of domain knowledge must exist. However, the domain knowledge is primarily of value for identifying parameters, thereafter it is important to evaluate the assumptions objectively. In order for the process to give a more objective decision, all choices and assumptions should be identified and properly validated. The challenge is to reduce the level of invalidated assumptions as far as possible. For example, reviewing the max penalty parameter in our model, it may be negligible if there is a 3% or a 10% maximum penalty. However, if there is none, a contractor might not bid on the project, which means a "1" or a "0" can be sufficient. Or reversely, numbers differentiating between different maximum penalties must be used. This is a choice that is made during the data processing step. To determine the most representative way to quantify a parameter, use existing practices from experienced tenders in combination with optimizing the model through repeated iterations with variations of parameters. Additionally, in a case with large data sets, a correlation study can be performed to support the choices. This aims to make the steps in the process, in this case the data processing, as objectively representative as possible.

When reaching the last steps in the process, "Model Calibration", "Validation" and "Evaluation", with the presented data, an unsatisfactory conclusion of the results from the neural network model was reached. However, and more importantly, these steps reveal the uncertainties that must be evaluated. Additionally, they give a direction for possible interactions between steps of the process. For illustration, when viewing the neural network model's performance in the study, two initial probable reasons can be assumed for its relatively poor result. Firstly, the input parameters are not sufficient to predict the number of tender offers, and secondly, a larger data set is required to improve the performance. As presented in Section 2.5.3, there is reason to believe that a larger data set would reduce the error of the predictions. Hence, the model could be improved and before the areas of improvements are fully examined, the model can be argued to show potential due to predictions within the correct range. Thus there is potential of increasing its accuracy by iterating back in the process. In this case, going from step "Evaluation" back to either "Document Collection" to gather more data or to "Choice of Model and Parameters" to better explain the project's content in the model input are all good options for reiteration.

Overall, it can be assumed that the first set of choices used in the process from data to information, is based largely on assumptions. By reviewing the choices, the assumptions, and the performance and then iterating back to previous steps, the process will become more and more data driven and validated. The better the input, the higher the quality of the output and the value of the information will be. With information of high quality, the tender process can be supported and guided with data which will provide a more optimized tender process, which in turn will reduce the price of all projects. Moreover, when evaluating whether or not the output is accurate enough, it is important to compare the potential value of even more accurate predictions against the cost of reiterating the process further. Since the process adds value if it predicts better than the current practice, a result with decent output can from a business perspective be worth more than more precise and accurate predictions.

7 Conclusions

The aim of this study was three-fold. First, to evaluate tender data from Trafikverket as a data source for guiding decisions in the tender process. Second, to assess the suitability of using a neural network model to predict the amount of tender bids on bridge projects. Third, to propose a process for the steps of turning tender documents into information that can guide decisions in the tender process.

The data received from Trafikverket was cumbersome to access and requires a huge amount of time to gather large data sets. The data retrieved from Trafikverket is not sufficiently structured for modeling. It is also not efficient unless Trafikverket changes their method of archiving tender documents or natural language processing techniques dramatically increase in their performance. Given the prerequisites today, continuous collection of tender documents on projects internally would be advisable.

The neural network model used for predicting the number of tender bids on a project is limited by the number of observations and input variables used in this study and is, therefore, lacking in its predictions. However, similar studies have been able to improve performance significantly by adding more observations. Therefore, we conclude that a neural network model is a suitable way of predicting the number of tender bids on a project, considering the Root-Mean-Squared Error of 2.45 given only 41 observations to train on. Moreover, the projects used in this study were unexpectedly heterogeneous, and therefore more work on defining input variables that fully capture the contents of the projects is likely to improve the predictive power of the neural network model.

This study proposes an eight-step transformation process from data to information. Whereas the DIKW hierarchy inspired the process, the KDD process was used as a guide during the construction of the process. The biggest difference between our suggested process and the KDD-process is the addition of three steps at the beginning of the process. Other processes assume data to be initially in a database, while the fragmented data practices in the construction industry require additional steps.

We recommend future researchers to focus on identifying input variables that capture the heterogeneity of bridge projects. Moreover, future studies should ensure that a sufficiently large amount of observations are used so that insufficient data set size can be ruled out as a cause of low predictive power. Finally, by studying how data-driven decisions in the tender process affect tender price levels, contractors and procurers would better understand the value of making tender data more accessible.

References

Abdullah, M. F., & Ahmad, K. (2013). The mapping process of unstructured data to structured data. In *2013 international conference on research and innovation in information systems (icriis)* (pp. 151–155).

Ballesteros-Pérez, P., González-Cruz, M. C., Fernández-Diego, M., & Pellicer, E. (2014). Estimating future bidding performance of competitor bidders in capped tenders. *Journal of civil engineering and management*, *20*(5), 702–713.

Baskarada, S., & Koronios, A. (2013). Data, information, knowledge, wisdom (dikw): a semiotic theoretical and empirical exploration of the hierarchy and its quality dimension. *Australasian Journal of Information Systems*, *18*(1).

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, *41*(3), 1–52.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., ... Pasha, M. (2016). Big data in the construction industry: A review of present status, opportunities, and future trends. *Advanced engineering informatics*, *30*(3), 500–521.

Bocij, P., Greasley, A., & Hickie, S. (2008). *Business information systems: Technology, development and management*. Pearson education.

Carr, P. G. (2005). Investigation of bid price competition measured through prebid project estimates, actual bid prices, and number of bidders. *Journal of Construction Engineering and Management*, *131*(11), 1165–1172.

Cheng, M.-Y., Hsiang, C.-C., Tsai, H.-C., & Do, H.-L. (2011). Bidding decision making for construction company using a multi-criteria prospect model. *Journal of Civil Engineering and Management*, *17*(3), 424–436.

Curtis, G., & Cobham, D. (2008). *Business information systems: Analysis, design and practice*. Pearson Education.

Drew, D., & Skitmore, M. (1997). The effect of contract type and size on competitiveness in bidding. *Construction Management & Economics*, *15*(5), 469–489.

Duff, R., Emsley, M., Gregory, M., Lowe, D., & Masterman, J. (1986). Development of a model of total building procurement costs for construction clients. *Management*, *1*, 210–8.

Elhag, T., & Boussabaine, A. (1999). Tender price estimation: neural networks vs. regression analysis..

Emsley, M. W., Lowe, D. J., Duff, A. R., Harding, A., & Hickson, A. (2002). Data modelling and the application of a neural network approach to the prediction of total construction costs. *Construction Management & Economics*, *20*(6), 465–472.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37–37.

Flanagan, R., & Norman, G. (1982). An examination of the tendering pattern of individual building contractors. *Building Technology and Management*.

Galhardas, H., Florescu, D., Shasha, D., Simon, E., & Saita, C. (2001). Declarative data cleaning: Language, model, and algorithms.

Ganti, V., & Sarma, A. D. (2013). Data cleaning: A practical perspective. *Synthesis Lectures on Data Management*, *5*(3), 1–85.

Government Offices of Sweden. (2020). *The principle of public access to official records.* Retrieved 2020-02-24, from `https://www.government.se/how-sweden-is-governed/the-principle-of-public-access-to-official-documents/`

Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R journal*, *2*(1), 30–38.

Ismail, S. A., Bandi, S., & Maaz, Z. N. (2018). An appraisal into the potential application of big data in the construction industry. *International Journal of Built Environment and Sustainability*, *5*(2).

Jin, Z., Anderson, M. R., Cafarella, M., & Jagadish, H. (2017). Foofah: Transforming data by example. In *Proceedings of the 2017 acm international conference on management of data* (pp. 683–698).

Kim, G.-H., Yoon, J.-E., An, S.-H., Cho, H.-H., & Kang, K.-I. (2004). Neural network model incorporating a genetic algorithm in estimating construction costs. *Building and Environment*, *39*(11), 1333–1340.

Konjunkturinstitutet. (2020). *Statistikdatabasen.* Retrieved 2020-04-02, from `http://statistik.konj.se/PXWeb/pxweb/sv/KonjBar/KonjBar__indikatorer/Indikatorm.px/table/tableViewLayout1/?rxid=e72114d1-42ab-4548-a3bb-19c835cdb477`

Kriesel, D. (2007). A brief introduction on neural networks.

Lee, T. B. (2019, February). *How neural networks work—and why they've become a big business.* Retrieved 2020-04-29, from `https://arstechnica.com/science/2019/12/how-neural-networks-work-and-why-theyve-become-a-big-business/`

Li, G., Ooi, B. C., Feng, J., Wang, J., & Zhou, L. (2008). Ease: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *Proceedings of the 2008 acm sigmod international conference on management of data* (pp. 903–914).

McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. *Queue*, *3*(9), 48–57.

Minli, Z., & Shanshan, Q. (2012). Research on the application of artificial neural networks in tender offer for construction projects. *Physics Procedia*, *24*, 1781–1788.

Ng, S. T., Cheung, S. O., Skitmore, M., & Wong, T. C. (2004). An integrated regression analysis and time series model for construction tender price index forecasting. *Construction Management and Economics*, *22*(5), 483–493.

Ngai, S. C., Drew, D. S., Lo, H. P., & Skitmore, M. (2002). A theoretical framework for determining the minimum number of bidders in construction bidding competitions. *Construction Management & Economics*, *20*(6), 473–482.

Pigott, T. D. (2001). A review of methods for missing data. *Educational research and evaluation*, *7*(4), 353–383.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3–13.

Rowley, J. (2007). The wisdom hierarchy: representations of the dikw hierarchy. *Journal of information science*, *33*(2), 163–180.

Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M., & Marinescu, V. (2013). Converting unstructured and semi-structured data into knowledge. In *2013 11th roedunet international conference* (pp. 1–4).

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning.* Springer Science & Business Media.

Scannapieco, M., & Catarci, T. (2002). Data quality under a computer science perspective.

Archivi & Computer, *2*, 1–15.

Skitmore, M. (2002). Raftery curve construction for tender price forecasts. *Construction Management & Economics*, *20*(1), 83–89.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the ieee international conference on computer vision* (pp. 843–852).

The R Foundation. (2020). *What is r?* Retrieved 2020-05-04, from `https://www.r-project.org/about.html`

Trafikverket. (2015). *Which authority does what within transportation?* [text]. Retrieved 2020-01-28, from `https://www.trafikverket.se/en/startpage/about-us/Trafikverket/Which-authority-does-what-within-transportation/`

Trafikverket. (2020a). *Information about dealing with personal data and data protection officer.* Retrieved 2020-01-28, from `https://www.trafikverket.se/contentassets/ad412af548db4de9a20760812885c764/dealing_with_personal_data.pdf`

Trafikverket. (2020b). *Trafikverket.* Retrieved 2020-02-24, from `https://www.trafikverket.se/en/startpage/about-us/Trafikverket/`

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5–33.

Wickham, H., et al. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 ieee 6th international conference on advanced computing (iacc)* (p. 78-83).

Department of Technology Management and Economics
Division of Innovation and R&D Management
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY