Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

John Lau

Modelling D&B Tunnelling Construction Data

Master's thesis in Geotechnics and Geohazards
Supervisor: Amund Bruland
August 2020

NTNU
Norwegian University of
Science and Technology

John Lau

Modelling D&B Tunnelling Construction Data

Master's thesis in Geotechnics and Geohazards
Supervisor: Amund Bruland
August 2020

Norwegian University of Science and Technology
Faculty of Engineering
Department of Civil and Environmental Engineering

NTNU
Norwegian University of
Science and Technology

Preface

This document represents my Master's Thesis as part of my MSc in Geotechnics and Geohazards. The research was performed at the Department of Civil and Environmental Engineering at the Norwegian University of Science and Technology (NTNU), and in partnership with *Norsk Forening for Fjellsprengningsteknikk* (NFF).

The wheels were in motion, as early as October 2018: shortly after my first semester in Norway had begun. From the get-go, I already knew that I wanted to study and write about *tunnels*. So when given the chance- I, of course, jumped at the offer to partner-up with Professor Amund Bruland. Casual chit-chats between my then-lecturer (who later on became the supervisor of this Master's Thesis) were often open-ended. Though, time after time, his inclination tilted me towards big data, and towards digitisation. The theme was extremely ambiguous; had an unlimited scope - and frankly, completely out of my expertise. But it was exactly the kind of puzzle I had envisioned for myself when I first came to Norway.

My first intuition was to sign up to, and attend as many statistics and mathematics classes as I could - and so I did. There, I sought out any and all lecturers, student assistants, and classmates - whomever showed the faintest interest in my project, or my ideas. I presented to my data, and plead to them for an opinion, and for insight. Time and time again, I received the same response, "But what is your problem? What are you trying to solve?". Extremely vague, and equally frustrating. At first, I deemed their response to be apathetic - and downright unhelpful. I had envisioned someone to simply hand me a super-complex cutting edge prediction algorithm - the *magic bullet* too all my problems: neither of which came true, of course.

In hindsight, they were completely right. I had severely lacked any project understanding: which eventually nudged me towards the data science approach.

Trondheim, August 2020

John Lau

John Lau

Acknowledgment

First of all, I would like to express my sincerest gratitude to Professor Amund Bruland. During my first semester here in Norway, he was simply "just another lecturer", in "just another course". Before long, he was my greatest source of inspiration; and I owe my fascination (or perhaps craze) for Norwegian tunnelling, undoubtedly all to him. In the end- he was my supervisor, a confidant, and a kind friend. Thank you sincerely for your guidance, support, and for taking a wild chance on me.

Likewise, I would also like to thank *Norsk Forening for Fjellsprengningsteknikk* (NFF), for introducing me to the Norwegian tunnelling industry: whether it through their seminars or conferences; or with a part-time summer job. And of course, a big thanks to them for supplying to me the data used in this study. Thanks guys.

Further, I would also like to thank my very first friend in Norway, Eldbjørg: for her top-notch jokes; and for always sharing her snacks with me. I am also thank to my buddies Glowy, Śnieżynka, Paulinka and Zuko and for their never-ending encouragements and warm comfort.

To conclude, I would like to thank my dear Majka. She was always the first to offer a lending hand; and the last to complain. When I needed a proof reader, an editor, a spell checker; or when I was hungry, when I needed a couple of hugs — she was always there. Thank you thank you Majka.

J.L

Abstract

Since its inception, the Norwegian Tunnelling Contract System (NoTCoS) has confided in theoretical and empirical studies; and leaned upon industry experience and intuition, to derive its time capacity values. However- as we zip through this digital-age- it has become clear, that digitisation is not slowing down; and continues to engulf *all things*. The push- and the scramble for data-driven solutions; and for automation, is inevitable across all industries today. Norwegian drill and blast (D&B) tunnelling is no exception.

In this study, preliminary investigations were performed to identify useful modelling techniques for converting D&B construction data into reliable time capacity values. Supplementary to this, are two secondary objectives. The first, was to highlight the current obstacles (such as gaps in the research, and insufficient data) impeding successful data-based modelling. While the second was to propose realistic reform to the current data collection process: for effective predictive analysis in the future.

Overall, the study conducted here alluded to the non-negative least squares (NNLS) algorithm as a useful predictive tool for extracting D&B time capacity values from construction data. However- its success is utterly dependent on a strict set of conditions: if unmet, the consequences are quickly hampering.

Contents

List of Tables

List of Figures

Nomenclature

Construction Terms

BoQ	Bill Of Quantities
D&B	Drill & Blast
ETS	Equivalent Time System
NoTCoS	Norwegian Tunnelling Contract System
TBM	Tunnel Boring Machine

Computer-related Terms

.CSV	Comma-Separated Values file
.JPEG	Joint Photographic Experts Group file
.PDF	Portable Document Format file
.XLS	eXceL Spreadsheet file

Statistical and Mathematical Notations

β_0	Intercept, also the constant
β_k	Coefficient, also the parameter
ϵ	Error, also the noise
μ	Mean (or average)
σ	Standard deviation
d	Effect size (Cohen)
N	Population size
n	Subsample size

p	P-value
EDV	Error Distribution Value
k	Dimension
LR	Likelihood Ratio
Mdn	Median
MLR	Multiple Linear Regression
NNLS	Non-negative Least Squares
OLS	Ordinary Least Squares
RTO	Regression-Through-the-Origin
SEM	Standard Error of the Mean
x	Independent variable, also the predictor or regressor
y	Dependent variable, also the response or outcome

Chapter 1

Introduction

Since its inception, the Norwegian Tunnelling Contract System (NoTCoS) has confided in theoretical and empirical studies; and leaned upon industry experience and intuition, to derive its time capacity values. However- as we zip through this digital-age- it has become clear, that digitisation is not slowing down; and continues to engulf *all things*. The push- and the scramble for data-driven solutions; and for automation, is inevitable across all industries today. Norwegian drill and blast (D&B) tunnelling is no exception. In this study, preliminary investigations are performed to identify useful modelling techniques for converting D&B construction data into reliable time capacity values.

1.1 Beginnings

All signs were indicating that Norwegian tunnel builders were - for the most part - supportive of the NoTCoS: a scheme promoting risk sharing between owner and contractor. This had been the case for the last 30 years (if not more) (Kleivan 1989) (Grøv 2012). A not-so-surprising feat, as the original framework of this scheme was rooted amongst joint-efforts between those in the tunnelling industry and those in the academic sector. Today, the development and promotion of this contract scheme (among other agendas) is spearheaded once again by a similar ensemble: a committee consisting of members from *Norsk Forening for Fjellsprengningsteknikk* (NFF). Notable participants included experts in the field; the construction industry; and from the public road authority. Amund Bruland, a professor at the Norwegian University of Science and Technology (NTNU), was one such member. One of his duties included the coordination of scientific research pertaining to Norwegian tunnelling, and was charged with preserving the industry's state-of-the-art performance. I too, had been moonstruck by such a wild ambition: and aspired to contribute to the Norwegian tunnelling industry, as he had.

For some time now, Amund Bruland had been squatting on a collection of raw construction data. A mismatch array of bill of quantities (BoQ) to be exact. From recent and not-so-recent D&B tunnelling projects. Not exactly groundbreaking, mind you. Construction ledgers of registered works completed have existed since the beginning of construction itself. Nonetheless, this particular database (among other's like it) was ever-increasing; and its exponential growth was not slowing down (Chen et al. 2014). Amund had hypothesised that the data he had accumulated was in fact valuable. He had envisioned that perhaps this data could be analysed or even modelled - but had only lacked the time and resources to do so. A recurring theme in today's digitised world: where the rate of data capture far exceeded the capacity for meaningful analysis. As he presented this notion to me, I insisted I could be of help. I agreed to not only adopt his ideas, but promised to also tack on original ones of my own.

1.2 Research agenda

After several iterations, the principle research agenda of this report then becomes:

> **An investigation into predictive analysis techniques relevant to Norwegian drill and blast (D&B) tunnelling construction data, with considerations to the Norwegian Tunnelling Contract System (NoTCoS)**

Supplementary to this, this study aims to also:

- Highlight the current obstacles (the knowledge gap, and the insufficient data quality and quantity) impeding successful data-based modelling;
- Identify the "ideal" data required to effectively derive data-driven time capacity values; and
- Propose realistic reform to the current data collection process in order to achieve this "ideal" data quality.

This master thesis is thereby written in partnership with NTNU and the NFF committee; and is a reflection of their initiatives. The motivation behind the research is to promote and support the Norwegian D&B tunnelling industry. To provide to those involved, the tools necessary to embrace this ever-digitised environment. The ambition (perhaps too large) is to hopefully find an admissible procedure that may harness the abundant data in today's digitised-world: to transform it into reliable and logical time scheduling decision making. No- the objective is not to abolish existing models, but instead provide additional tools to facilitate risk sharing between owner and contractor. If successful, a data-driven approach would allow for the model to be "self correcting", dynamic, and adjustable according to the real-time performance at the face.

1.3 The chosen methodology and the data collected

The selected methodology used in this report consisted of two major elements: an exploratory phase and a primary analysis. In the exploratory phase, the data science workflow was applied to *real* D&B tunnelling data: where the objective was to identify the "best" modelling technique. In the primary analysis, a back-calculation of the chosen model was conducted with the aid of *simulated* data. The objective was to reveal deficiencies in the currently-available data, and to gauge the model's performance through stress testing.

1.4 Limitations

The limitations of the proposed methodology are mostly related to the model selection process. Admittedly, my grasp of data science is rather shaky; and absolutely, I still have a lot to learn. Due to my limited competences, the overall models I choose to investigate, and implement are most likely not going to be the "best". With all things considered however, a model not deemed "the best", is not immediately doomed for failure. Models discovered in this study can be still function effectively come time for real-world applications - as long as their tolerances are not breached, of course.

1.5 Report structure

Standard convention is to maintain a distinct separation between the methodology, theory and the analysis (as recommended by several Norwegian universities (NTNU 2019)). However - in foreshadowing the proposed methodology: that is, a data science approach - I believe that it is in the best interest of the reader for the thesis to be structured in a similar "cyclical" fashion. As such, this report does occasionally intertwine these elements together. Therefore, to achieve a more-fluid narrative, I have instead opted for chronological sequencing, in some sections.

The format of this document is intended to not only provide the research and findings as stated in the above section, but to also guide the reader through the back-and-forth thought-process. Overall, the structure of this report has been split into the following parts:

Part I: Introduction In this opening section, the relevant background information is presented alongside a brief exploratory literature review. The information gathered here, created the necessary framework for hypothesis development. Thereafter, the research agenda and its research methodology is proposed.

Part II: Methodology and Data Part II of this report includes a description of the research methodology; and the collected-data used to address the research agenda is described. Their respective limitations are also discussed.

Part III: Theoretical Background The third part contains a presentation of the data science approach; and the relating theoretical background required prior to application and analysis. Also inclusive in this component, is the supporting literature for statistics, machine learning, and mathematical optimisation.

Part IV: Exploratory Part IV is an application of the data science methodology to *real* D&B tunnelling data. This step-wise process provided a much-needed systematic and objective approach for selecting relevant algorithms.

Part V: Analysis As part of the primary analysis component of this thesis, a back-analysis of the selected algorithm is performed and documented in Part V. In this section, hypothetical tunnel data is simulated to isolate and identify the shortcomings of the currently-available data. Research conducted here also serves to assess the limitations and performance of the chosen-predictive algorithm.

Part VI: Discussions and Summary In Part VI, the findings from the exploratory and analysis phases are synthesised and presented for discussion. This chapters consists of; discussions and a summary of this research endeavour; comparative studies between *existing* and *new* prediction models; proposals for improving the data quality; as well as recommendations of future works. Thereafter, some closing remarks are made in the final chapter.

Work performed over the course of a year

This master thesis began well before the final and fourth semester of this master's degree. Work presented here is a combination of research conducted over the course of my summer job, and during the specialisation project course. As such, the reader should be wary to keep this in mind - especially true if assessment is required.

Special focus on the Norwegian Method of Tunnelling

A diverse selection of tunnelling philosophies are practised all around the world. Aside from the drill and blast tunnelling method, other examples include: the *New Austrian Tunnelling Method* (NATM); or perhaps the *Tunnel Boring Machine Method* (TBM) (Singh and Goel 2011). Although these are both tremendously popular, and tend to dominate the markets outside of Norway, the Norwegian Method of Tunnelling (NMT) is still the most prominent excavation method here in Norway. Therefore, unless stated otherwise, the contents of this report is in the context of the Norwegian D&B tunnelling method.

Part I

Hypothesis Development

Hypothesis development

In this report, hypothesis development was comprised principally of three components. These elements were not exclusively hierarchical, but rather cyclical. Naturally, several loops were taken before a definitive research agenda, and an eventual methodology could be established.

Research element	**Chapter**
Why is this research being conducted?	**Chapter 2** Defining the research agenda
What is the knowledge gap?	**Chapter 3** Conducting a literature review
How will the research question be addressed?	**Chapter 4** Establishing a methodology
Then finally: *Reevaluate and redefine the research agenda.*	

The provisional research question

The initial research agenda proposed to me was extremely (and most likely deliberately) vague. To restate this objective- it began crudely as:

> **An investigation into predictive modelling techniques relevant to drill and blast (D&B) tunnelling construction data**

This provisional research question provided the starting point for the following background and literature review chapters.

Chapter 2

Background

As part of hypothesis development, this chapter aims provide to the reader brief background information pertaining to the Norwegian drill and blast (D&B) tunnelling industry; and to the Norwegian Tunnelling Contract System (NoTCoS). The pre-investigative research presented here forms the groundwork necessary for a better understanding of the real-world operations and legal problems facing the D&B industry. The information here is essential to when forming the research agenda; and for revealing the motivation behind this research.

Objective

The objective of this background chapter, and of a literature review is to define **why** this research is conducted; and to define:

- What the knowledge gap is; and
- What the industry needs are.

Contents

This chapter begins by first presenting:

- The Norwegian D&B tunnelling industry background information
  - Section 2.1 - Norwegian drill and blast tunnelling
  - Section 2.2 - Norwegian Tunnelling Contract System (NoTCoS)
  - Section 2.2.1 - The "equivalent time system" (ETS)
- Section 2.3 - The motivation behind this research
- Section 2.4 - The knowledge gap so far

2.1 Norwegian drill and blast tunnelling

The construction processes involved in a typical D&B tunnelling excavation are generally considered cyclical. That is it to say, the tunnel is *constructed* segment by segment (commonly referred to as a round cycle), and can vary in cross-section, length, and shape (dictated by function, geological conditions, and cost). The principle construction activities within each round cycle are: drilling (of blast holes); charging; blasting (firing); ventilation; loading and hauling (removal of muck); and then finally, scaling and rock support. Figure 2.1 shows the processes of a typical round cycle. Depending on tunnel conditions, supplementary tasks (such as probe drilling or the installation of rock support) may also be required after each round cycle. These activities may be implemented systematically - or reactively, as geological conditions dictate. It is this very geology that governs the quantity of required works, and these decisions are made directly at the tunnel face, and in *real-time*. All in all, these operations directly influence time consumption.

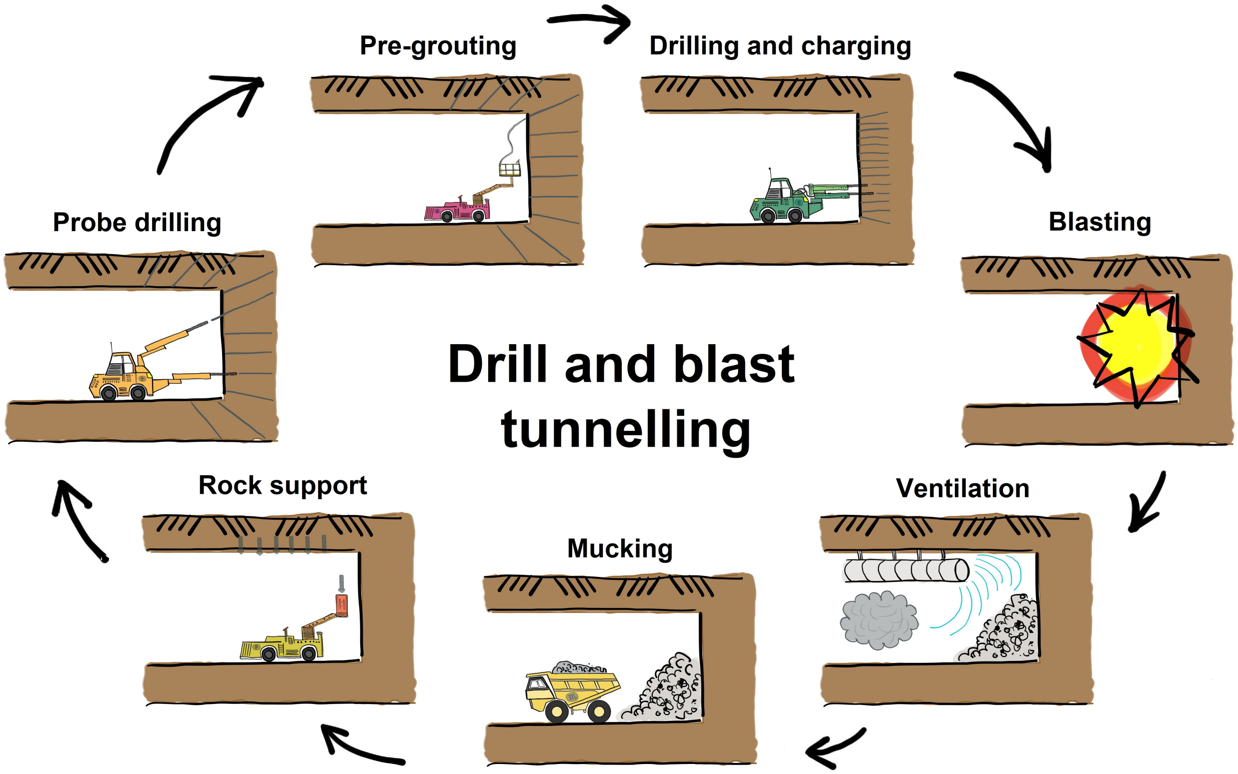


Figure 2.1: The typical construction tasks within a round cycle of drill and blast tunnelling

2.1.1 The inherent presence of uncertainty

The ever-looming existence of uncertainty is a cause for concern for all construction projects. The impact of uncertainty becomes inevitable as the degree of complexity and uniqueness increases (Samset 2010). Uncertainty may exist in a project for a myriad of reasons: such as unanticipated market changes, or extreme weather conditions. In the case of D&B tunnelling, this uncertainty generally stems from *geological* uncertainty. This is due to the fact that it is extremely difficult (and unfeasible) to map actual geological conditions with any accuracy. *Actual* ground conditions at the tunnel face may deviate substantially from the *expected*. This in turn results in construction tasks that differ from those stipulated in the initial contract. Examples may include:

- Additional quantity of work: worse-than-expected geology may warrant the need for more-than-expected quantities work (e.g., additional rock support components, grouting requirements)
- Changes to the scope and construction methodology: by the same vein, these unforeseen circumstances may deem it necessary to completely revise the excavation method or the chosen rock support measures

As illustrated in an article by (Kleivan 1989), and reports by NFF and NTNU (NFF 2019), a typical tunnel project can experience substantial differences between the *real* and *estimated* due to unforeseen geological conditions and other uncertainties. Tunnel contracts have historically tended to underestimate required construction time – more so than overestimating. Regardless of its tendency, commonplace misevaluations have resulted in unrealistic project deadlines. Depending on the contract type, one party may be struck with misfortune: and end up bearing a disproportionate amount of the consequences. In Norway however, the Norwegian Tunnelling Contract System (NoTCoS) has been developed to address this exact issue in a logical and just manner.

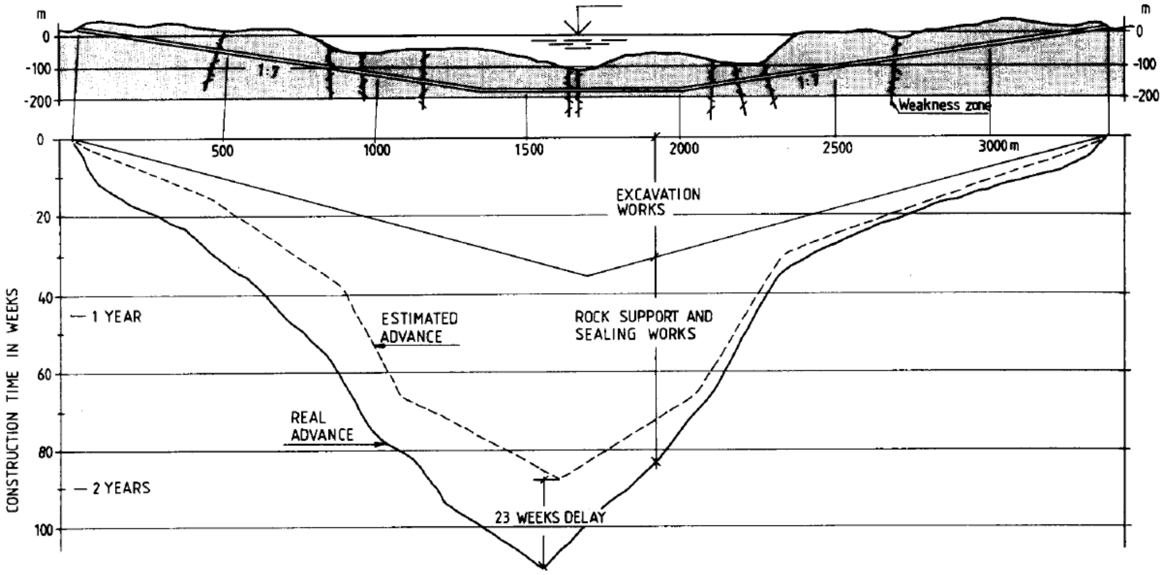


Figure 2.2: Advance diagrams for a subsea tunnel driven from both sides (Kleivan 1989)

2.2 The Norwegian Tunnelling Contract System

The concept of "drill and blast" method existed in Norway as early as the 19th century, following the introduction of the dynamite and the steel drill (Johnsen 2014). But it wasn't until the hydropower boom in the 1960s, that we saw the emergence of risk-sharing principles in contract writing (Grøv 2012). This methodology, is often referred to as the Norwegian Tunnelling Contract System (NoTCoS) (Kleivan 1989) and (Grøv 2012). The term "risk" in this context refers to the financial consequences that may occur should the ground conditions encountered at the tunnel face deviate from the anticipated.

The principles of risk sharing is intended to address the following elements of risk:

- **Ground conditions**: There is inherent aleatory and epistemic uncertainty associated with predicting the existing geology. However, it is the owner that "provides the ground" for the contractor. The owner is therefore responsible for the actual ground conditions encountered.
- **Performance**: The contractor is responsible for the construction activities. Works shall be performed in an efficient manner, and according to the technical specifications.
- **Cost**: In a scenario where the contractor bares all the risk, tender bids will naturally increase across the board, in order to account for the risk. The asking price will most definitely exceed the actual cost of the project, thus resulting in lower return-on-investment. Conversely, should the owner assume all the risk, the contractor may under-bid and -estimate the total price of the project. As a consequence. they may not be able to complete the project, should cost overruns occur.

This is achieved by having "regulation mechanisms … built into the contract" (Grøv 2012). One of these features can be described as the equivalent time system (ETS), or the unit price system.

2.2.1 The equivalent time system

As mentioned earlier, the scope of works is directly related to the geological conditions. Should these ground conditions differ from that which was originally anticipated, contractual parties must be able to implement amendments to the construction duration (and cost) without resorting to timely and costly litigation (Kleivan 1989). Nested within the NoTCoS, the equivalent time system (ETS) is described. All major operations are assigned an equivalent time in the form of time consumption (hour/metre of tunnel), time capacity (unit/hour) and unit time (hour/unit). A contractor seeking time extension, may use these values as the basis of their request.

2.3 Motivation behind this research agenda

In this section, some standout reasons for conducting this research are presented.

A flexible model is potentially more useful than an accurate one

The reduction in uncertainty, especially concerning geological characteristics, demands unfeasible amounts of resources, both in cost and time. This rings especially true during the planning stages of a project, where investigative works are bound by the confines of a predefined budget. Ironically, it is during these early phases of a project that information serves as the most beneficial time for acquirement (Samset 2010).

The research conducted here however- unashamedly concedes to the fact uncertainty exists, and it is prevalent. Rather than of eliminating (or exposing) it entirely, we attempt to conform to its erratic behaviour: by introducing *flexibility* into the model. This is what the NoTCoS was originally set out to achieve, and has continued to do so.

The current Norwegian contracting system *works*

Ever since the introduction of risk sharing contracts, disputes relating to changes in the quantity of work has been essentially non-existent in the Norwegian tunnelling industry (Kleivan 1989). Although this is difficult to prove, I can verify anecdotally of this claim. Conversations between folks on the ground: have all advocated that this system indeed *works* [1].

Continued development of the Norwegian Tunnelling Contract System

A majority of tunnels constructed use the NoTCoS (Grøv 2012). In 1989, it garnered an 80% adoption rate, and as of 2012, these figures may have increased. The continual development of this system and the values within time equivalent system may even further bolster these adaption rates.

2.4 What is the knowledge gap so far?

In this section, the more obvious knowledge gaps are first discussed. Though, subsequent to this, a formal literature review is conducted and detailed in Chapter 3.

The absence of data-driven research

These aforementioned time capacity rates, have been developed by *Norsk Forening for Fjellsprengningsteknikk* (NFF): a joint committee of experts in the field. These are often used as a starting point in time scheduling during

[1] Arne Aakre, EBA; Jarl Åge Haugan, Statens vegvesen; Stein Bjøru, Veidekke, personal communication, December 11, 2019.

the tunnel planning, tendering and contract stages. These values are however, considered “loose estimates”, and were only *approximated* by a panel of experts: such as contractors, engineers, and researchers. Their assessments were based on their collective experiences in the tunnelling industry, in combination with theoretical calculations and empirical studies.

Continual state-of-the-art improvements into tools that support risk-sharing may help redress the inequity between contractors and clients. This may come in the form of a data-driven approach. Where a data-driven model may be useful in combination with existing theoretical and empirical-based models. Such a trio will indeed strengthen the equivalent time system (ETS) against scientific inquiry. Overall, an improvement in time estimation may hopefully alleviate the number of tunnel construction legal disputes in Norway.

A digitised world

The technological advancements in data storage and data collection have created cost- and time-effective solutions to the capture of useful data (Gandomi and Haider 2015). The rate of data generation has doubled every second year (Chen et al. 2014). This "statement", although not directly relevant to Norwegian tunnelling (and not exactly true either in 2020), does still provide some insight about the direction this world is currently headed - that is towards digitisation. Naturally, it would be sensible for the Norwegian tunnelling industry to embrace "big data" in order to remain competitive.

Although the rate of data capture within the Norwegian tunnelling industry has not been prominent when compared to that of other industries, its prominence is still remarkable. Machinery data, such as positional and performance output measurements are already collected automatically. Tunnel mapping and progress records can now all be collected remotely *and* automatically. And of course, this data is currently being collected to register the bill of quantities (BoQ) in most tunnel projects.

Point being- this information already exists. Insufficient research however, has (at this time) been directed towards a coupling of the Norwegian tunnelling industry and digitisation. A recurring theme in today’s digitised world: where the rate of data capture far exceeded the capacity for meaningful analysis.

2.5 Some noteworthy definitions

In this brief section, it feels relevant to clarify some already-mentioned terms and phrases, before proceeding any further.

The ideal data

At this point, the D&B dataset is persistently appended with the term *currently-available*. This description is deliberate. It implies that the data (and the industry) is open to change: and that perhaps there is room for improvement in the current data collection procedure. Should a *working* data-driven approach indeed be realised, the supplementary goal of this research is to identify whether or not the current data is ideal for reliable modelling. And if not - identify and propose realistic changes to the data collection process.

Actionable

The term "actionable" is limited by requirements and standards set by the Norwegian tunnelling industry. These conditions are closely related to the Norwegian Tunnelling Contract System (NoTCoS), and will be elaborated further in Chapter 4 - Methodology and Data.

2.6 Chapter summary

Norwegian D&B tunnelling time scheduling so far has relied on a mixture of empirical and theoretical models; added with a splash of industry experience and intuition. However, as evident in our day-to-day lives, the digitised environment is becoming quickly the norm. Data-driven research and solutions are more and more commonplace. Continued focus towards embracing data-based solutions is therefore vital, for the Norwegian D&B tunnelling industry to remain competitive, and to uphold scientific inquiry.

The next course of action

In the next chapter, a literature review is performed to uncover the extent of existing data-driven research within the Norwegian D&B tunnelling industry; and to better-understand how other industries have been able to incorporate big data into their own unique prediction models.

Chapter 3

Literature Review

With a provisional research agenda set in the previous chapter, a literature review is now presented in this chapter. This review was intended to provide a critical analysis of the existing time scheduling techniques: specifically relating to prediction models; and of the methods already in use for the analyse of construction data. Furthermore, the knowledge gathered here made light of potential knowledge gaps and industry needs. This step was central to the eventually-selected research methodology and to defining the scope of works.

Structure

To begin, the preliminary literature review was confined to a very narrow and very specific theme: that is, existing time scheduling techniques within the drill and blast (D&B) tunnelling industry. However, it was soon evident that studies within the field were either not "data-driven", or did not satisfy the requirements of the Norwegian Tunnelling Contract System (NoTCoS). Consequently, the confines of this literature review expanded incrementally, until these requirements could be met (if at all). Figure 3.1 illustrates a progression chart of the process. Following this, each change in the research domain has been discussed in the sections below.

Contents

All in all, the literature review spanned across the following industries and fields:

- Section 3.1 - Time scheduling in the drill and blast tunnelling industry;
- Section 3.1 - Time scheduling in the general tunnelling industry;
- Section 3.2 - Time scheduling in the general construction industries;
- Section 3.3 - General predictive analysis techniques; and then,
- Section 3.4 - A return to the basics: the data science methodology

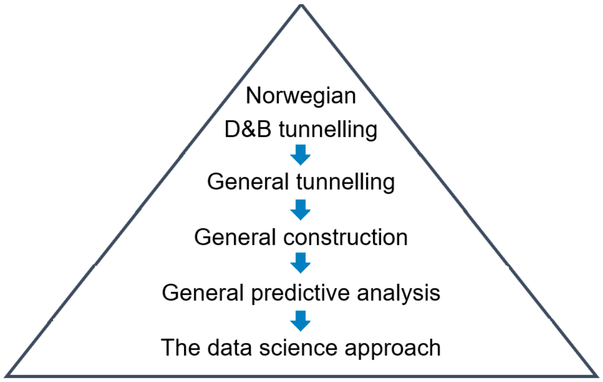


Figure 3.1: A progression chart of the research process

A systematic approach

In order to conduct a thorough assessment of scholarly literature, a systematic approach was used in this paper. The methodology, as illustrated in Figure 3.2, was designed around the guide *Reviewing the Literature* (Academic-Skills 2013). The work flow process essentially entails the following.

- Search parameter preparation (establishing relevant keywords and phrases)
- Literature retrieval (digital and physical material is obtained via various medians)
- Screening (and sorting) process (results are filtered and categorised, to identify relevant material)
- Literature evaluation (the material is evaluated systematically using the T-O-N-E principles (NTNU 2017))

The entire work flow can be considered cyclical (a recurring theme in this report): where the steps are often repeated as refinements in the scope or research agenda deem it necessary. The procedure was revisited regularly throughout the entirety of the research timeline, and not exclusive to the early investigative phases. Continual appraisal of the literature is crucial in order to reassert the research question's relevance; and to ensure that information is current.

A brief note on *time scheduling*

Before proceeding, it bares mentioning the concept of "time scheduling". In the context of D&B tunnelling, it closely resembles "time management" in a project. It can include estimating the overall construction duration; as well as, the more-specific, time influence of individual construction tasks (time capacity values). In this particular context, *time* is also considered a resource, and in turn, every resource can be converted into a cost (Bruland 2018). This procedure is concerned about decisions that result in *real* action: such as during the contract writing phase (estimated total construction duration), and post-contract phases (such as litigation and requests for time-extension).

The benefits of *accurate* time scheduling within the construction industry has been emphasised since the beginning of construction itself. Benefits are abundant, and is often pointed out as one of the key reasons for a

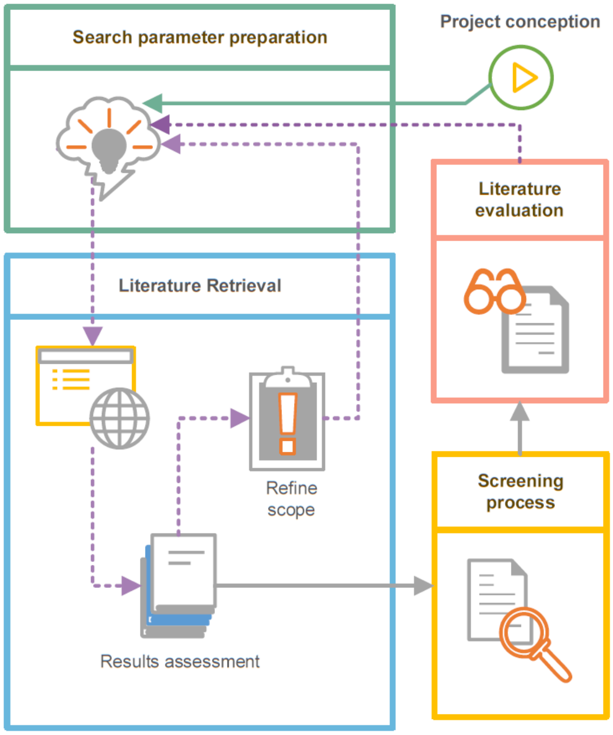


Figure 3.2: Systematic approach to conducting a literature review

project's success. It's usefulness transcends across all phases of a project's lifeline: including the planning-, inte-, and even post-stages. The study into enhancing and optimising the methods in which we conduct time scheduling is therefore not a brand-new concept.

3.1 The general tunnelling industry

To being, a literature review on time scheduling methods within Norwegian D&B tunnelling is first conducted. Following this, the investigation is expanded to include Tunnel Boring Machine (TBM) tunnelling as well. Notable observations within the literature, and their effects on the overall decision path has been detailed in the sections below.

Post-contract time extensions decisions are made based on the "equivalent time system". (Odd Johannessen 2000; NFF 2019). Current *time capacity values* are derived from a combination of empirical research, theoretical modelling, and subjective input from industry experts.

Could a data-driven approach also be relevant to today's time scheduling?

In 1975, a time scheduling method is developed by Statskraftverkene, Rasjonaliseringskontoret. [1] These time estimates are made on the basis of both, direct and indirect variables. In it, the internal machinations of each individual construction task is assigned a time capacity. Furthermore, various "non-construction tasks" such as fixed lost time, proportional operational time, and incidental lost time, are also taken into consideration. At the same time, these time capacity values have been "weighted" according to external factors such as equipment type (categorical), or cross-section dimension (ordinal). Finally, the model is linear, and assumes that all construction tasks occur independent of each other (Zare 2006; Zare and Bruland 2006; Zare and Bruland 2007; Zare 2007; Zare 2016). All in all, a very comprehensive model. Which leads to the question:

> Is it also possible to develop a data-driven model that can consider both, internal and external factors (as it is with the original model)?

Models that factor in geological conditions. D&B tunnelling advance rates are modelled with considerations to the expected condition (as characterised by the Q-system), as well as the tunnel design cross-section. (Kim and Bruland 2009). In these models, the estimates are reliant on external factors which are oftentimes extremely variable across the project's lifeline.

> Can a model be successful without relying on the precise mapping of external factors?

Decision Aids for Tunnelling (DAT). Using the Decision Aids for Tunnelling (DAT), tunnel construction time can also be estimated using a probabilistic approach, as opposed to the conventional deterministic methods. In this study, Monte Carlo simulation is used to make probabilistic time and cost predictions (Min 2008). Overall, a probabilistic approach does indeed provide the model some flexibility to account for the varying geological conditions. However, time estimates with a probability distribution do not satisfy current Norwegian tunnelling contract types.

> Is it possible to implement the concept of probability to the current time capacity values, or even to the NoTCoS?

Models with limited and only select-variables. TBM tunnelling construction time estimates here are modelled using a limited and select number of variables (Rostami 2016). This model produces a "best fit" model by only including a few select-variables, and by excluding others that "reduce" the model's performance.

> Can a modelling technique be developed to incorporate **all** construction activities?

Mixed-models that combine both quantitative and qualitative variables. In some models, their input variables are a combination of both quantitative and qualitative variables (Macias et al. 2017; Rostami 2016; Bruland 2000). These models have been largely successful. However, qualitative features (such as geological conditions) possess large variability, and may be costly to confirm (especially true during the post-contract phases).

[1] In 2006, the 6th revision was published.

> Can a model be developed using only quantitative variables?

Support vector regression models. Tunnel boring machine (TBM) penetration rates are estimated using support vector regression models in a study (Mahdevari et al. 2014). The researchers once again employ a mixture of tunnel dimensions and geological conditions to predict advance rates. Although the concern is once again directed towards the choice of input parameters, the research conducted here is a useful point of reference for future studies relating to machine learning application.

Mixed-models with a deterministic and probabilistic approach. TBM tunnelling construction time estimates, depending on the stage of implementation, are derived using deterministic and probabilistic approaches (Špačková et al. 2013). These models take into account the uncertainties of a project, and predictions are presented as probabilistic estimates.

Interestingly- the authors argue that current input variables should account for the uncertainty (instead of deterministic estimates). Furthermore, they stress that time capacities should be dependent on external factors, such as geology and geometry.

> Once again, it would be interesting to see if the concept of probability distribution can be introduced to the NoTCoS.

3.1.1 Notable observations

Initial literature review revealed that the current time scheduling within the D&B industry was mostly confined to empirical and theoretical methods. Although data-driven models did exist within other excavation methods, these models conflicted with the NoTCoS's equivalent time system (ETS) and currently-available D&B data. The most obvious knowledge gaps have been briefly summarised below:

- Models included external factors as an input variable.
  - External factors (such as geological conditions) may be fraught with large variability. Such data points are difficult to obtain with any degrees of accuracy
- Models contained only limited and select input variables.
  - Prediction models only include variables that produce a "best fit" model. To satisfy the NoTCoS, **all** construction activities must be included.

These observations prompted another step back, and the scope of the literature review then expanded merely to: the general construction industry

3.2 The general construction industry

Many researchers have developed *working* techniques to estimate the total construction time. These prediction values are generally derived using statistical and machine learning programming; and these values are oftentimes extremely valuable during the early phases of a project. However, come the construction-phase, and even the post-construction phase, their usefulness and functionality becomes diminished. This is mostly due to the fact that any variations encountered during these phases are unable to be fed back to their prediction models.

For *standard* construction projects (in this case, defined as a typical urban / above-ground project) the disregard for such unforeseen variations will not significantly effect the overall outcome of the project. This is due to the more-predictable nature of a *standard* building project: as the typical construction project is generally confined to their own *controllable* work-environment. Furthermore, the amount of standard construction projects performed completely dwarfs the amount of subsurface projects. Such an abundance usually results in more information, more data, more experience. This means that any disruptions are more readily and reliably addressed.

Models are once again selective with their input variables. In a study titled “Developing a construction-duration model based on a historical dataset for building project” by (Lin et al. 2011), researchers built several regression models with different arrangements of input variables. Thereafter, the "best-performing" model was selected. Similar to some other tunnel-themed models, the model is mostly concerned about accuracy and not interpretability.

A stepwise regression. In a case in Poland, the construction duration was estimated using a stepwise regression technique (Czarnigowska and Sobotka 2014). Similar to our Norwegian scenario, the model discounts any external factors (such as non-technical factors), and only focuses on the known deterministic data. Nonetheless, this is where the similarities end. And just like the previously reviewed prediction models, input variables required both internal and external variables, and may not be compatible with the NoTCoS.

Construction model based on regression analysis. The use of simple and multiple linear regression was used in a study to estimate the total construction time (Odabaşi 2009). Although at its conclusion, the model performances were deemed unsuccessful, it was promising to see to other researchers attempt such a feat. Nonetheless, the model parameters were once again reliant on both internal and external (in this case, cost) factors.

Probabilistic time estimates. Many of the more *complex* prediction models may incorporate probabilistic features, to estimate the construction duration as a probability distribution. Some examples discovered include:

- The Monte Carlo Method (Hofstadler 2010)
- Probabilistic Time Coupling Method (Kostrzewa and Rogalska 2019)

In these studies, researchers are able to calculate estimates to a very precise degree. This methodology is extremely useful in quantifying the amount of risk associated with the project. Especially during the planning phase.

The caveat here, is that these types of models rely heavily on subjective input variables. These variables, although contribute directly to time consumption, are *weighted* at the discretion of the user. This process is highly subjective and based on experience - which may not stand up to scientific inquiry.

3.3 General predictive analysis techniques.

After reaching, what seemed like, the extent of existing time scheduling research, I temporarily distanced myself from these scholarly research papers. Instead, I transitioned my research towards general predictive analysis techniques, and attempted to identify models relevant to D&B construction data *myself.* Over time, I developed a taste of various kinds of statistical and machine learning models. These will be discussed in Chapter 6.

3.3.1 *A return to the basics*

In hindsight, I had most likely approached this exercise in a roundabout way. The scope of my investigation had been too wide, and lacked direction. Evaluations of individual predictive analysis methods appeared aimless and, at times, random. Too long, was I simply sifting through algorithms that were only *superficially* compatible with my construction data. Yes, I was able to plug my data into a fancy program, and I was quickly rewarded with some arbitrary values - a "best fit prediction". But I began to realise (after a *long* period) that I had in fact lacked a "measuring stick": a tool to assess the relevance of my model.

I returned back to the basics, and I signed up to as many statistics and mathematics classes as I possible could. A short description of this *new* information avenue is illustrated in Table 3.1 below.

Table 3.1: New information avenue after reaching the apparent limits of existing research

Verbal communications	
Source	**Contribution**
Meetings with supervisor	- To guide and to help keep research within the scope.
Classes with lecturers	- To inform about concepts, lingo, jargon, terms, etc.
Discussions with classmates	- For inspiration and to stimulate innovative thought.
Conversations with the industry	- To be informed about how scientific knowledge is practically applied in the Norwegian tunnelling industry.
Grey Literature	
Source	**Contribution**
TMA4268 - Statistical Learning	- Information about (statistical) modelling techniques
TTK4260 - Introduction to Multivariate Data Modelling	- Information about complex (machine learning) methods to model high dimensional data
TMA4180 - Optimization 1	- Information about mathematical optimisation

3.4 The data science method

As the number of complex mathematics-based predictive algorithms grew and grew, it became apparent that a unique dataset, can in fact be (sometimes forcefully) fit to more than one model. This left the untrained often wondering: which model is then the "correct" one? This is how the discipline of data science, over time, was conceived - as data analysts required a formal method to help identify the "best" model.

In this report, I look to this exact methodology in an attempt to identify useful models relevant to the D&B tunnelling construction data, and to the NoTCoS.

3.5 Concluding remarks and the decision path

The scoping literature review undertaken in this report, brings to light the fact that D&B tunnelling construction time estimation studies have been far and few between in recent years. Instead, research on TBM tunnelling penetration rates decisively outpaces that of D&B tunnelling. Furthermore, if you take one more step back, it is clear that prediction analysis within the construction industry is mostly dominated by those in building sector.

Nonetheless, a myriad of prediction techniques relating to tunnel construction time *do* exist (Isaksson 2002). However- when concerning the NoTCoS, and to the currently-available digital data, these methods do quite fit the bill, for the following reasons:

- Prediction models are mostly empirical and theoretical based
- Their input variables demand a mixture of internal and external factors
- The models are unable to take on **all** construction tasks (only selective)
- Prediction estimates are probabilistic-theme *(not necessarily bad)*
- "Working" models are largely focused on accuracy and not on interpretability
- Research pertaining to parametric estimating values and time scheduling has been relatively underexplored.

To elaborate, current methods are incompatible for a marrying between the NoTCoS, and a data-driven approach. For the most part, empirical and theoretical models have been developed and successfully used for time scheduling. These models have relied on a combination of, both, qualitative and quantitative input variables to derive construction duration estimates (Isaksson 2002). Furthermore, when it comes to the diversity of these input variables, the models themselves are generally very restrictive. Internal machinations and quantifiable (and sometimes deterministic) construction tasks are seldom considered. Instead external influences such as geological conditions, or the time-cost relationships, are instead explored. Factors that, most obviously should influence the construction time are selectively omitted for the sake of achieving a “best fit regression line”.

> Popular forecasting techniques such as machine learning can act “as a “black box”; meaning that they can be employed to predict the value of a target based on data, but the rules or implicit patterns within the model cannot be interpreted.”
>
> - (Salimi et al. 2016).

This highlights the fundamental disconnect between the NoTCoS and a data-driven approach. For the longest time, the equivalent time system (ETS) has depended entirely on quantifying these aforementioned “implicit patterns” (to be exact, these are simply time capacity values for each construction task). At the same time, other time forecasting methods instead have fine-tuned their models towards an higher overall prediction accuracy, by sacrificing the interpretability of the model.

The research performed here aims to bridge this knowledge gap; and to hopefully develop a model that will satisfy both, the NoTCoS, and a data-driven approach. As such, from this point, there is a return to the basics. With the help of the data science methodology, perhaps a *new* model can be realised. One that does not attempt to size-up and measure the ever-present geological uncertainty - but instead be flexible enough to conform with it.

The next course of action

At the conclusion of the preliminary literature review, the research agenda receives a slight revision; and has been revised and presented in the next section. Immediately following this, the data science workflow, and the data to be analysed will be introduced formally in Chapter 4.1 - Description of the two-part analysis.

Redefining the Research Agenda

Primary objective

The primary objective of this study is to explore modelling techniques useful for extracting actionable insight from Norwegian D&B tunnelling construction data. The term actionable implies that the model predictions and inferences can be used as supplementary tools within the Norwegian Tunnelling Contract System (NoTCoS), and to develop a data-driven methodology for deriving their time equivalent system.

Secondary objective

The secondary objective of this study is to encourage realistic reform to the data collection process. In Part IV - Analysis, it was hypothesised, that the reliability and effectiveness of the model's predictions are at the mercy of the data quality. Be that as it may, the currently-available data gathered and analysed in this report, originate simply as Norwegian D&B tunnelling construction BoQs. The contents of such BoQs are currently governed by the tunnel owners' and builders' *own* discretion. This highlights a serious disconnect between the analysis, performed from behind a desk; and the data, collected from the tunnel face.

The ambition of this study is to therefore identify specific areas where the data collection methodology can be improved. This includes proposing changes to type of data collected; and begins by first appraising its quality.

Chapter 4

Methodology and Data

The process detailed here, describes the chosen methodology and the collected-data used to address the research agenda. Although it is customary to maintain a distinct separation between *procedure* and *action*, this report does occasionally intertwine the two elements together. For example, the simulated data used in the back-analysis, was only possible after the findings realised in the exploratory phase. Therefore, to achieve a more-fluid narrative, I have instead opted for chronological sequencing, in some sections.

4.1 The proposed methodology

The selected methodology used in this report consisted of two major elements: an exploratory phase and a primary analysis. In the exploratory phase, the data science workflow was applied to *real* D&B tunnelling data: where the objective was to identify the "best" modelling technique. In the primary analysis, a back-calculation of the chosen model was conducted with the aid of *simulated* data. The objective was to reveal deficiencies in the currently-available data, and to gauge the model's performance through stress testing. The two-part analysis has been summarised in Table 4.1, and thereafter elaborated in the subsequent sections.

Table 4.1: Description of the two-part analysis

Phase	**Objective**	**Process**	**Dataset**
Part: III *Exploratory phase*	Identify the "best" modelling technique	Application of the data science method	*Real* tunnel data • The Svartås-tunnel • The Kongsberg-tunnel
Part: IV *Primary analysis*	• Evaluate the model's performance • Describe the *ideal* the data	A back-analysis of the selected-model	*Simulated* tunnel data • The Kangaroo-tunnel • The Koala-tunnel

A step in the wrong direction

The initial months of this research can be described as random and aimless. Although it lead to the discovery and pilot-testing of numerous exciting predictive models, the time spent was without structure and objectivity: something I had, early on, severely lacked. Over time, it became clear that the subject-data could actually be fit to *more than one* model. Under the gaze of traditional model validation techniques, contrasting models were performing equally well. This realisation was of course a contradiction. It indicated that my earliest attempts were misguided- or perhaps, just *wrong*. And so, I took a few steps back.

4.2 Exploratory phase

The first stage of this analysis is therefore a revisit to the basics, and begins as:

> **An application of the data science workflow to D&B tunnelling construction data**

In this exploratory phase, the data science process model is applied to *real* tunnel data. This component of the report is not intended to scrutinise the data science method itself. Instead- its workflow is employed to provide a much-needed systematic and objective method for selecting the "best" models capable of extracting actionable insight from D&B tunnelling construction data. The term "best" is project-specific, and is assessed according to the data science criteria (Chapman et al. 1999; Wirth and Hipp 2000). A step-by-step guide of the data science workflow, as well as the relevant background information, is documented in Part II - Theoretical Background (Chapter 5). All in all, this stage serves a preparatory function, and forms the basis of the subsequent primary analysis component of this research.

4.2.1 The structure

For the sake of coherence, the processes in workflow have been separated into three components. The exploratory stage first begins with a clear definition the project objective and constraints. Thereafter, *real* tunnel data is analysed using the prescribed data science workflow. In the third component, results and inferences are discussed: and a modelling algorithm is finally selected for further analysis in Part IV - Analysis. Table 4.2 provides a summary of the exploratory phase.

4.3 Primary analysis

Part IV features the primary analysis element of this thesis. In this two-part study, a back-analysis of the predictive algorithm, as selected in the exploratory phase, is performed. For this exercise, *hypothetical* D&B tunnel data is analysed to address the following questions:

Table 4.2: Contents of the exploratory phase

Category	**Elements**	**Subject concerning**	**Chapter**
Problem understanding	• Objective definition • Constraints and assumptions	*The Norwegian D&B tunnelling industry*	**Chapter** 10
Data analysis	• Data preparation • Exploratory Data Analysis • Data modelling • Model validation • Results	• The Svartås-tunnel • The Kongsberg-tunnel • The other tunnels	**Chapter** 11 **Chapter** 12
Decision making	• Inferences • **Model selection**	*All data*	**Chapter** 13

- Are errors in the prediction caused by incorrect model choice, or by the data quality?
- Just how much *lost-time* can the model handle before predictions begin to falter?

Ultimately, the objective of this analysis is to determine the effectiveness of the model, under an *ideal* (yet realistic) environment. In turn, results from this study may also reveal deficiencies in the current data collection process. All in all, this study is intended to induce a change to the data collection process in the future.

4.3.1 The structure

The experimental analysis first begins with the generation of pseudo-random D&B tunnelling data. Although technically "random", this process is in fact regulated by weighted distributions, and constraints: to replicate the actualities of a *real* tunnel project. Using this method, two principle datasets are generated: each with their own unique properties and specific objectives. The first dataset (labelled: Kangaroo-tunnel) is simulated using a **constant** construction rate (NFF's time capacity value): with an increasing amount of noise (lost-time) imposed at each iteration. While the second dataset (labelled: Koala-tunnel), is generated using a **variable** construction rate: with an increasing amount of variance at each iteration. This exercise functions as limit testing of the selected-model: and is used to examine the effects of varying degrees of "unaccounted for" -time (missing data). A summary of the two-part analysis is presented in Table 4.3.

Table 4.3: Contents of the primary analysis

Category	**Elements**	**Subject concerning**	**Chapter**
Experiment setup	• Simulation process • Iteration development	Hypothetical tunnel data	**Section 14.2**
Data analysis	• Data modelling • Model validation • Error diagnostics • Results	• The Kangaroo-tunnel • The Koala-tunnel	**Section 14.3.1** **and 14.3.2** **Section 14.3.3**
Discussions	• Discussions • Recommendations	Hypothetical tunnel data	**Chapter 15**

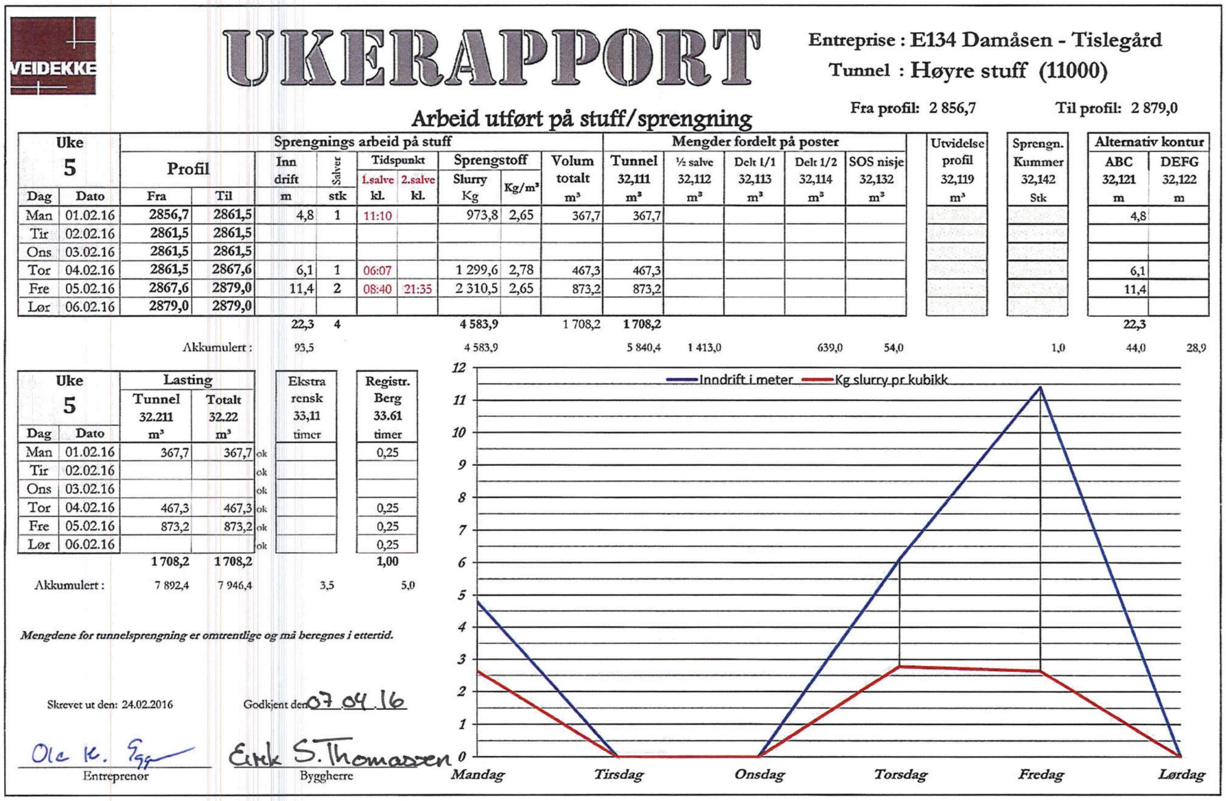
4.4 The data analysed

In this report, analysis was performed on both *real* and *simulated* D&B tunnelling construction data. The following sections is a presentation of this data.

4.4.1 The *real* data

Data collected for the exploratory phase - in its crudest form - are simply D&B tunnelling bill of quantities (BoQ), or invoicing documents. In today's Norwegian D&B tunnelling industry, it is common for the contractor to maintain a weekly record of (billable) work completed to date. These progress logs are mandatory components when submitting official invoices to the client.

These D&B construction logs were collected with the assistance of government and private organisations. In partnership with *Norsk Forening for Fjellsprengningsteknikk* (NFF), I was granted access to several D&B tunnelling construction databases. These documents were stored in Statens Vegvesen's eRoom database, via Erfaringstall tunnel: https://www.vegvesen.no/e-room/3/eRoom/AltaVest/Erfaringstalltunnel. A standard BoQ form is shown in Figure 4.1.



VEIDEKKE

UKERAPPORT

Entreprise : E134 Damåsen - Tislegård
Tunnel : Høyre stuff (11000)

Arbeid utført på stuff/sprengning

Fra profil: 2 856,7   Til profil: 2 879,0

Uke		Sprengnings arbeid på stuff									Mengder fordelt på poster					Utvidelse	Sprengn.	Alternativ kontur	
5		Profil		Inn drift	Salver	Tidspunkt		Sprengstoff		Volum totalt	Tunnel	½ salve	Delt 1/1	Delt 1/2	SOS nisje	profil	Kummer	ABC	DEFG
						1.salve	2.salve	Slurry	Kg/m³		32,111	32,112	32,113	32,114	32,132	32,119	32,142	32,121	32,122
Dag	Dato	Fra	Til	m	stk	kl.	kl.	Kg		m³	m³	m³	m³	m³	m³	m³	Stk	m	m
Man	01.02.16	2856,7	2861,5	4,8	1	11:10		973,8	2,65	367,7	367,7							4,8	
Tir	02.02.16	2861,5	2861,5																
Ons	03.02.16	2861,5	2861,5																
Tor	04.02.16	2861,5	2867,6	6,1	1	06:07		1 299,6	2,78	467,3	467,3							6,1	
Fre	05.02.16	2867,6	2879,0	11,4	2	08:40	21:35	2 310,5	2,65	873,2	873,2							11,4	
Lør	06.02.16	2879,0	2879,0																
				22,3	4			4 583,9		1 708,2	1 708,2							22,3	
			Akkumulert :	93,5				4 583,9			5 840,4	1 413,0		639,0	54,0		1,0	44,0	28,9

Uke		Lasting			Ekstra	Registr.
5		Tunnel	Totalt		rensk	Berg
		32.211	32.22		33,11	33.61
Dag	Dato	m³	m³		timer	timer
Man	01.02.16	367,7	367,7	ok		0,25
Tir	02.02.16			ok		
Ons	03.02.16			ok		
Tor	04.02.16	467,3	467,3	ok		0,25
Fre	05.02.16	873,2	873,2	ok		0,25
Lør	06.02.16			ok		0,25
		1 708,2	1 708,2			1,00
Akkumulert :		7 892,4	7 946,4		3,5	5,0

Mengdene for tunnelsprengning er omtrentlige og må beregnes i ettertid.

Skrevet ut den: 24.02.2016   Godkjent den 07.04.16

Ole K. Egg
Entreprenør

Eirik S. Thomassen
Byggherre

Figure 4.1: A typical Norwegian D&B tunnelling construction BoQ (Veidekke and VegVesen 2016)

Data description

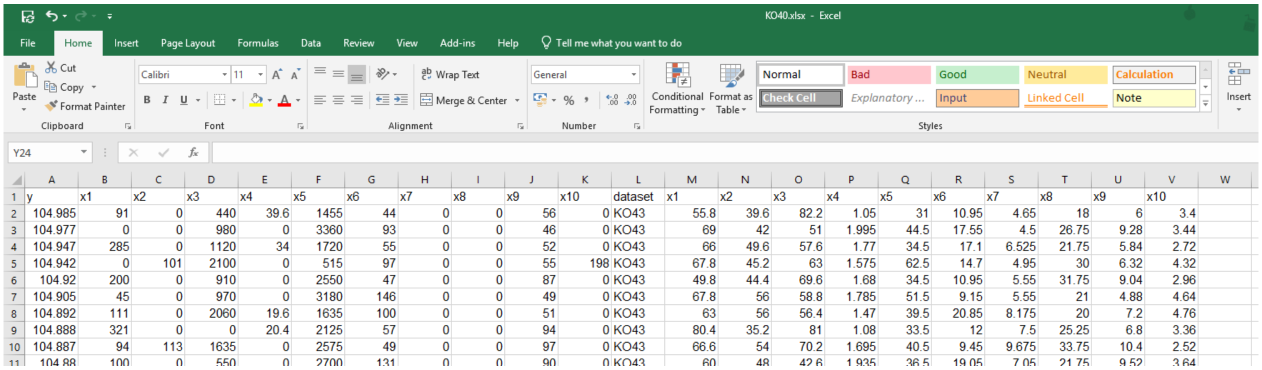
The contents of a typical tunnel BoQ will usually include the quantities of all major tunnelling elements (for example, the amount of excavated material, rockbolts and shotcrete). These values are usually presented in a daily or weekly format. In theory, these records are intended to reflect the amount of time consumed at the tunnel face. Therefore, non-construction tasks are occasionally included as well (for example, owner's half hour, or rigging times). Depending on the tunnel project, documents may also register over 80 other unique tasks. At the same time, the quality and quantity of data can also vary significantly across the available databases. As shown in Table 4.4, the currently-available data differed greatly between the tunnel projects. All in all, eight Norwegian tunnels were reviewed in this report. However, due to the combination of time restrictions and a limited dataset, select-tunnel projects were more closely analysed than others.

Table 4.4: The contents of the currently-available D&B construction logs

Registered construction activities														
	Probe-drilling	Control holes	Injection	Excav. material	Rockbolts ≤ 4m	Rockbolts > 4m	Shotcrete	Rigging times	Man. cleaning	Combi. bolts	Embedded bolts	Arches	Securing bolts	Straps
E16 Filefjells-tunnel	✓			✓	✓	✓	✓		✓	✓				✓
Kongsberg- tunnel	✓	✓	✓	✓	✓	✓	✓				✓	✓		✓
Larviks-tunnel														
Mælefjell-tunnel				✓								✓	✓	
Reinforsheia-tunnel				✓							✓	✓		✓
Røddøls-tunnel				✓							✓	✓	✓	
Svartås-tunnel	✓	✓	✓	✓								✓		
Vollås-tunnel				✓										✓

4.4.2 The *simulated* data

Data used for the back-analysis component of this report was simulated using pseudo-random number generator functions. Although technically "random", this process is in fact regulated by weighted distributions, and constraints: to replicate the actualities of a *real* tunnel project. Figure 4.2 is a screenshot of an example simulated tunnel dataset. Overall, the simulation process has been detailed further in Section 14.2.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	dataset	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	
2	104.985	91	0	440	39.6	1455	44	0	0	56	0	KO43	55.8	39.6	82.2	1.05	31	10.95	4.65	18	6	3.4	
3	104.977	0	0	980	0	3360	93	0	0	46	0	KO43	69	42	51	1.995	44.5	17.55	4.5	26.75	9.28	3.44	
4	104.947	285	0	1120	34	1720	55	0	0	52	0	KO43	66	49.6	57.6	1.77	34.5	17.1	6.525	21.75	5.84	2.72	
5	104.942	0	101	2100	0	515	97	0	0	55	198	KO43	67.8	45.2	63	1.575	62.5	14.7	4.95	30	6.32	4.32	
6	104.92	200	0	910	0	2550	47	0	0	87	0	KO43	49.8	44.4	69.6	1.68	34.5	10.95	5.55	31.75	9.04	2.96	
7	104.905	45	0	970	0	3180	146	0	0	49	0	KO43	67.8	56	58.8	1.785	51.5	9.15	5.55	21	4.88	4.64	
8	104.892	111	0	2060	19.6	1635	100	0	0	51	0	KO43	63	56	56.4	1.47	39.5	20.85	8.175	20	7.2	4.76	
9	104.888	321	0	0	20.4	2125	57	0	0	94	0	KO43	80.4	35.2	81	1.08	33.5	12	7.5	25.25	6.8	3.36	
10	104.887	94	113	1635	0	2575	49	0	0	97	0	KO43	66.6	54	70.2	1.695	40.5	9.45	9.675	33.75	10.4	2.52	
11	104.88	100	0	550	0	2700	131	0	0	90	0	KO43	60	48	42.6	1.935	36.5	19.05	7.05	21.75	9.52	3.64	

Figure 4.2: A screenshot of an example simulated tunnel dataset

4.5 Limitations

The limitations of the proposed methodology are two-fold. These concern: first, the effectiveness and reliability of the model selection process; and secondly, the collected data.

4.5.1 Limitations about the model selection process

The performance of the data science method is confined by the practitioner's awareness and knowledge of predictive algorithms in existence. Even should one follow the work-process to a T, and is able to conduct a thorough and effective assessment of subject dataset, all these nobles efforts may be dwindled come time for model selection. A practitioner unacquainted with sufficient amount of prediction models will have difficulties selecting the *right* model for their dataset; and results are potentially unreliable.

With all things considered, a model not deemed "the best", is not immediately doomed for failure. Models discovered in this study can be still function effectively come time for real-world applications - as long as their tolerances are not breached of course. These tolerances (limitations) are discussed in Chapter 15.

4.5.2 Limitations about the collected data

The caveat here is that these progress bills are merely kept for billing and quantity purposes only – rather than for research purposes. Currently, these documents are not regulated, and their contents are therefore at the discretion of the tunnel builders and owners. For this reason, typically, only *billable* activities are documented. Lost times (fixed times: such as rigging, or incidental lost times (Zare and Bruland 2006)), are not recorded. There is no burden on the contractors to precisely log construction operations on a day-to-day basis. Table 4.5 presents some factors that may result in misrepresentative data.

Table 4.5: Common factors that can cause misrepresentative data

Factors	**Causes**	**Effect**
Contractors are mostly concerned about the net quantity.	Contractors may lump works completed in one week, with another week.	Activities logged for a specific week, may in reality may not have actually been performed during the corresponding week.
Construction tasks can sometimes begin on one week, but carry over to the following week (especially for activities commenced on the weekend).	There is no discretion on the ones responsible for recording the data (the contractors), to differentiate the partition of works that were completed for a given week. (As mentioned above, the emphasis is on the net quantity of work.	Overlapping activities may inadvertently reduce/increase an activity's influence on its time capacity value; thus, skewing the overall prediction accuracy.
Bill of quantities may contain errors or data may be missing on a week to week basis.	Due to human input-error, mistakes can occur. However, when mistakes are eventually noticed, adjustments are instead made to the *following* week's bill of quantities. Subtractions or additional to the quantities are made.	Data points may result in negative or surplus (not representative) quantities. These are often difficult to identify - and can cause inaccuracies in the predicted values, or produce false-indicators of outlier data.

There is a risk that data may therefore not reflect real world circumstances – and outliers and errors become increasingly difficult to identify. Inaccuracies like this are problematic for conventional statistical approaches but may still be useful when testing prediction techniques.

Uncertainty, and the assumptions to be made about the data

At this time, there is no formal quality control or standardised regulations and guidelines in place, when it comes to D&B tunnelling completed-works record keeping. There is therefore a level of uncertainty associated with the obtained construction logs. The data appraised in this study is considered representative of the current industry practice and standards. Until the record-keeping practice of these logs can be standardised and verified format, the following assumptions are made throughout this study, and listed below:

- The BoQ is statement of construction activities performed at the tunnel face. A separate BoQ is kept for activities performed behind the tunnel face.
- All construction tasks were performed within the industry standard 101 weekly working-hours.
- Records logged for a given day/week imply that the recorded tasks were performed during the same working-day/week.
- It is assumed that all major contributors to time consumption have been accounted for, and included in the acquired BoQs.
- The construction methodology does not deviate from the norm, throughout the entire construction duration.

The above comments, by the same token, are an insinuation to the preferred level of quality and format of future record keeping practices within the D&B tunnelling industry. This is discussed further in the closing sections (Chapter 16).

A few remarks

During my encounters with data scientists and the ilk, I was often reminded by them that it is best to "learn by doing". This is particularly true within the data science discipline: where there is usually no *one* correct procedure. Instead several *ideal* methods and models are possible, and it is then up to the practitioner to decide, using experience and intuition. However- due to my lack of experience within data analysis, the selected methodology emits the impression of a "trial and error" approach - because it is!

Part II

Theoretical Background

Data Science, and the Tools in the Tool Box

The followings chapters form the theoretical components of this study. To begin, the background information, and a brief step-by-step guide of the chosen methodology is introduced. Following this, the primary theoretical background of relevant predictive algorithms is presented. Finally, model validation techniques used to assess the reliability and performance of the models are briefed.

The objective of these sections is to prepare the reader with sufficient background information prior to the exploratory and analytical phases. A summary of the contents has been outlined below.

Category	**Elements**	**Subject concerning**	**Chapter**
The data analysis methodology	Introduction to the field of data science, and descriptions of its workflow used to select the "best" model	• The data science approach • Relevant programs	**Chapter 5**
Prediction models	Theoretical background of the models examined during the model selection process	• Classification (partial) • Regression analysis • Mathematical optimisation	**Section 6.3** **Chapter 7** **Chapter 8**

Chapter 5

The Data Science Methodology

> "A jack of all trades and a master of some."
>
> — Brendan Tierney (Tierney 2012)

In this chapter, the data science methodology is introduced. It first begins with brief background information on the subject. Thereafter, its data analysis workflow is introduced in detail.

- Section 5.1 - Background information
- Section 5.2 - The data science workflow

5.1 Background

Data science is an umbrella term for describing the multi-disciplinary approach to extracting actionable insight from raw data. In recent times, it has been recognised - and with strong consensus - as the fourth paradigm in scientific discovery (following experimental, theoretical and, computational science) (Hey et al. 2009). These insights are attained by applying analytical tools such as: mathematics, statistics, and informatics.

Overall, data science incorporates three primary disciplines (as aptly illustrated in Figure 5.1 (Conway 2010)). Beginning at the green circle, the term **Math & Statistics Knowledge** describes the general mathematical and statistical theory behind the programming codes and the algorithm. Moving onto the red circle, **Hacking Skills** describes the proficiency required to operate the predictive algorithms within a programming environment (computer software). This includes the prowess to write and run the necessary programming language. Finally, in the blue circle, the term **Substantive Expertise** describes the knowledge of the field in which the data science approach is being applied. This attribute is an essential component required for effective objective understanding; model selection; and the eventual inference. Table 5.1 below, describes just how data science will influence this particular D&B-related study.

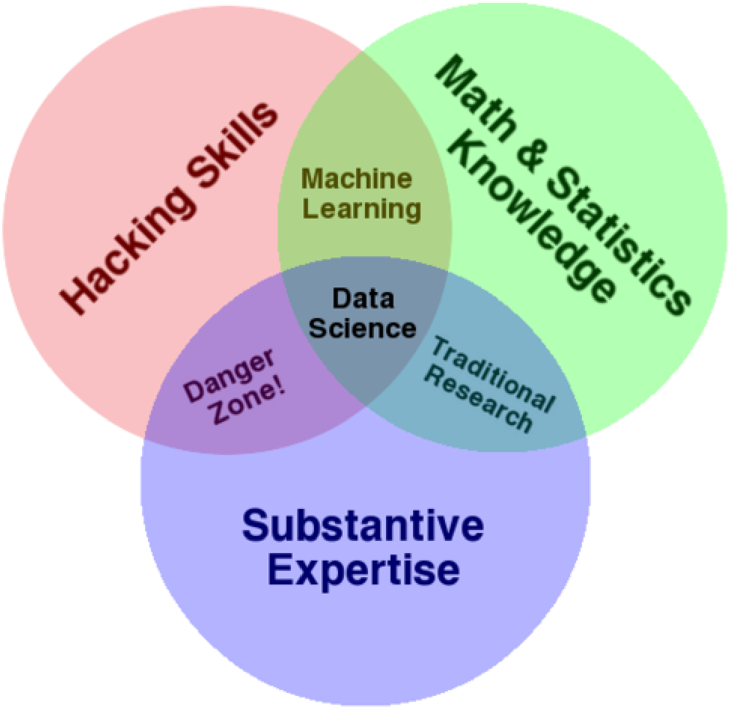


Figure 5.1: Data science depicted as a multi-layered discipline (Conway 2010)

5.1.1 Varying literature and perspective

Relevant theoretical background, and the such, as presented in this report was collected and synthesised from multiple medians (such as textbooks, articles, and lecture notes). These were oftentimes written from unique perspectives (whether it be from, a statistician's or perhaps a computer scientist's point of view). Naturally, distinctions between the terminology (as well as the jargon and language style) and even in the methodologies exist within the discipline and its literature. For example:

...in terminology and methodology: Abbreviations, symbols, and terminologies often came in a variety of flavours, depending on the writer and their preferred field of application.

...and in the workflow: The same phenomena can be seen within the preached methodology and its endorsed workflow. A particular "stage" may at times be omitted by one author - only to be emphasised by another.

Data science is still maturing

These inconsistencies may be due to the fact that *data science*, is practiced across many different sub-fields, each with their own unique set of objectives and perspectives. Furthermore, as a discipline, it is still rather *new* and therefore intrinsically dynamic. The work process is still maturing, and theories are ever-changing and continually optimising: as new tools and applications are being discovered on a regular basis.

Table 5.1: How data science affects D&B tunnelling modelling

Discipline	**Competences**	**Included in this report**
Math & Statistics Knowledge	General mathematical and statistical theory behind the programming codes and the algorithm	• Classification • Regression analysis • Mathematical optimisation
Hacking Skills	Computer software operation. Also includes writing and running the code necessary in order to implement the mathematical and statistical algorithms	• *R* studio • MATLAB • Python • Excel
Substantive Expertise	Knowledge of the field in which the data science approach is being applied	• NoTCoS: The Norwegian Tunnelling Contract System • Norwegian D&B tunneling

Synthesis according to the needs of the drill and blast industry

The Norwegian D&B tunneling industry however, has yet to experience such a boom in success compared to that of other fields such as medicine, economics, and even sports. There is currently no hard-and-fast (or *correct*) terminology within this sector. Therefore, during the amalgamation of the studied reference material, I did my best to remain consistent, and to select terminologies and methods that would be most appropriate for the D&B tunnelling industry.

5.2 The data science workflow

Practicing data science is a step-wise process, with many distinct phases: where the outcome of one step, will dictate how to proceed with another. Although the stages are officially separated into standalone procedures of a workflow, in practice however, these stages tend to overlap between one another. Several back-and-forths is almost always required. Inferences and predictions are usually not made with conviction, without first performing multiple iterations (and cycles) in the data science process.

There have been attempts to standardise the process model (Chapman et al. 1999)(Wirth and Hipp 2000). Yet, the truth of the matter is that practitioner's continue to customise their workflow depending on their field of application: that is to use one's own discretion to identify stages that are vital; and to deem others as unnecessary. Following lengthy trial and error, the steps below are deemed necessary to the research topic, and will be implemented in the analysis component of this report.

- Objective definition
- Data collection
- Data wrangling
- Exploratory Data Analysis (EDA)
- Data modelling

- Model Validation
- Inference
- Real-world assessment
- Control measures

These will be elaborated in the following sections.

5.2.1 Objective definition

This stage describes the initial step of the life cycle. In this phase, the objective and it's requirements are clearly identified. The knowledge gathered here can then be reframed into a data science problem (or an objective function). Naturally, this step creates a "point of reference" for the output model to evaluate its performance against. Furthermore, this step controls the constraints and other requirements that must be applied to the model. It is therefore commonplace that the problem definition undergoes multiple revisions (especially following post-deployment analysis) before consensus is reached. All in all, objective definition is arguably the most crucial step in the data science life cycle.

5.2.2 Data collection

Data collection is the phase which deals with how the data is created or acquired. Depending on the project, this can be collected manually, or even autonomously (due to advent and surge of internet of things devices). This step is sometimes subdivided into several very-specific and sometimes niche components: such as "data housing and architecture", and "data sensor calibration" set up requirements. However, on a practical level, the D&B data used in this report does not demand such high levels of control, and will therefore regress to more basic requirements, such as:

- Frequency of measurements;
- Precision of measurements;
- Number of measurements taken;
- Types of measurements taken; and
- How and where the data is stored.

Furthermore, the information uncovered during the post-deployment phase can alter the way the data is collected in follow-up projects.

5.2.3 Data wrangling

During the data wrangling stage, raw data is converted into a standardised form, and then adjusted depending on the algorithm requirements (compatibility). This stage is commonly conducted in conjunction with the EDA stage.

Several iterations and back-and-forths are therefore almost always required between the two stages. This report employs the following five data wrangling steps:

- Data description;
- Data import;
- Data clean up;
- Data transformation; and
- Feature engineering.

5.2.3.1 Data description

The collected data and its input variables are characterised and described in detail during this preliminary step. At this point, generalisation is kept to a minimum. Precise details enable more unambiguous inferences: as the analyst is informed exactly what the inputs are subjected to the model. Furthermore, clear descriptions allow for more effective cross-project analysis. This function rings especially true when attempting to combine datasets from different sources; or when conducting comparative analysis.

5.2.3.2 Data import

The data import step typically involves converting the raw collected data into a format more practical for data science analysis: such as .CSV or .XLS formats; or into dedicated data management software. This process can be extremely tedious and time consuming, particularly true in circumstances where the data source is not originally designed for data analysis, for example BoQ or receipts.

5.2.3.3 Data clean up

During the data clean up step, select-observations are omitted from data analysis. Typically, data clean up is required when outlier observations or missing data are present. This step is highly subjective, and can potentially result in misleading inferences. Data cleaning is therefore commonly performed in conjunction with exploratory data analysis (EDA) for assistance and validation.

5.2.3.4 Data transformation

Should the data not be compatible with the chosen model type, data transformation may be required to convert the original data into a more relevant format. For example, this step is sometimes necessary when EDA reveal that the data is non-linear, and that a linear-to-logarithmic transformation may produce in fact a better fit model.

5.2.3.5 Feature engineering

Every now and then, the existing raw data may need to be modified, collated; or perhaps even, new features need to be created using existing data. Table 5.2 below describes some examples reasons to implement feature engineering to a data set; the available techniques; as well as how the procedure can be applied to D&B tunnelling construction data.

Table 5.2: Example reasons for applying feature engineering to a dataset

Objective	Action	Example
• Reduce the number of dimensions in the dataset • Reduce some of the multi-collinearity between variables	• Transform predictor variable to response variable • Combine select variables together, and form new ones	Adjustments to the weekly working hours can be made according to already-known time consumed (such as owner’s half hour)
Improve the quality of the data	Split variables into separate variables	Probe drilling works can be separated into two categories: with and without plugging
Create a better fit model when the variables are in fact non-linear	Impose a separate constant value to specific variables	• A new variable can be created to describe the fixed-time for construction task (such as rigging) • A new variable can be created to describe the lost-time
Standardised datasets	Ensure that similar input variables between differing datasets, are of the same unit	The "tunnel advancement length" can be converted to "amount of excavated material" quantity

This process is designed to improve model results. However, tremendous caution must be taken, should one choose to exercise this step, or else the integrity of the data may be jeopardised. Excessive feature engineering may conversely result in misrepresentative data, and in turn, unrealistic inferences or predictions.

5.2.4 Exploratory data analysis

Exploratory data analysis (EDA) provides descriptive and visual aids for effective and efficient data wrangling, and model selection. EDA techniques are designed to characterise the structure of the collected data; and to reveal the inner relationships that might exist between the input variables. For example, histograms can reveal that input variables are in fact not normally distributed; or perhaps scatter plots can indicate that the input variables possess a non-linear relationship between one another. These will be discussed further within the model theory chapters.

5.2.5 Modelling

A common misconception about "modelling" data, is that it is mostly concerned about running complicated codes and algorithms. However, with the advancement of high performing computers, a hefty portion of the time and effort is actually consumed during phases preceding (and succeeding) this. This phase in the workflow consists principally of two steps:

- Model selection; and
- Data modelling.

5.2.5.1 Model selection

Model selection plays an important role in data science. The step is engaged prior to any modelling, and then revisited again after model validation. The process is multifaceted, and oftentimes subjective. It is not uncommon for the practitioner to end up *multiple high* performing models - and according to traditional model validation testing - all of which are equally credible and valid. In order to select the *correct* model, the practitioner must therefore consider more than just the final predicted outputs. Effective model selection must also take into account the following:

- the model must satisfy the project objective and its requirements;
- the modelled data is mathematically compatible with the predictive algorithm;
- the model performs highly according to model validation tests;
- the predictive algorithm is efficient to run; and
- the required dataset is cost-effective to obtain.

5.2.5.2 Data modelling

The data modelling step involves the implementation of the selected algorithm onto the ready-dataset. Although possible by hand, the utilisation of computer software and its code is almost always required for the execution of predictive algorithms. Further details of these computer programs have been included in Appendix A for reference.

5.2.6 Model validation

Following data modelling, a suite of model validation techniques can be employed to measure the models performance. Results obtained in this step allow for a more-objective assessment, and comparison between the various models in consideration - and is essential for effective model selection. For example, the R-squared test may be used to describe how much signal the model is able to encapsulate in comparison to other models. Validation techniques however, vary depending on the model type. These will instead be discussed within their corresponding model theory sections.

5.2.7 Inference

Output results from select-models are finally converted into actionable insights. As an example, should the model possess high predictive prowess, conclusions may be drawn from the final response values. Conversely, models high in interpretability may offer inferences through the estimated beta coefficients ((Bratko 1997; Plate 1999). Decisions made here form the basis of future real-world applications. Inferences must therefore not rely merely on the final output estimates, but also be made with considerations to the project objective.

5.2.8 Real-world assessment

In this practical phase, inferences made in the previous stage are applied to real-world situations. Using the Norwegian D&B tunnelling industry as an example, the time equivalent system may be amended with new suggestions to their time capacity values. The performance and consequences of such actions are thereafter assessed. This stage is not purely about gauging a model performance. It also serves as a vital tool for revealing deficiencies in the original model. Post-release comparative and sensitivity analysis may help identify:

- **Missing variables:** to reveal supplementary data that should be collected in the future;
- **Incorrect data type:** to indicate that variables require transformation; and
- **The data quality:** to justify improvements in the data collection process, to increase its truthfulness.

5.2.9 Control measures

In this *final* stage, the results from real-world application are analysed in an attempt to improve the model. Action plans may be created to collect additional data, or to make adjustments and corrections in the original model.

Information source

The theoretical background presented in this section was collected from a variety of textbooks and lecture notes. It is therefore admittedly difficult (and often disorderly) to single out a particular aspect of my learning, and attribute it *exactly* to a specific source. Instead, a general list of all literature that eventually contributed to my overall knowledge base, is presented below. However- major and pivotal ideas have been individually cited throughout this report as earnestly as possible.

- *R for data science: import, tidy, transform, visualize, and model data* by (Wickham and Grolemund 2016)
- *Introduction to Data Science: Data Analysis and Prediction Algorithms with R* by (Irizarry 2019)
- “CRISP-DM: Towards a standard process model for data mining” by (Wirth and Hipp 2000)
- “The CRISP-DM user guide” by (Chapman et al. 1999)

Chapter 6

Modelling Techniques

> "A more complicated model may fit the data better than a simpler one, but does the better fit justify the additional complexity?"
>
> - Peter G. Bryant (Bryant 1996)

6.1 Framework

In this part, I rummage through various unique data modelling methods, and highlight notable models based on a selection criteria. For each selected technique, the following descriptions and attributes are presented:

- Its theoretical background;
- A brief description of its algorithm and code;
- Related model validation techniques;
- Their Limitations; and
- Some useful definitions.

6.1.1 Purpose

As mentioned in the previous chapter, the data science approach is an umbrella term that incorporates a wide variety of analytical tools to create the eventual model. Surely, to successfully select the "best" model, one must first be familiar with the available options. The aim of the research performed here, is therefore intended to "add more tools to the tool box": that is, to expand the arsenal of algorithms I have at my disposal during the model selection phase. Inadequate knowledge of the possible options, may lead to incorrect model choice, and an unrealistic fit.

6.1.2 The model selection criteria

A model selection criteria was created to more-effectively assess a model's compatibility with the currently-available D&B data base; and to quickly gauge its relevance to the research topic. This checklist was established following formal project understanding (to be described in Chapter 10). A summary of the selection requirements and some considerations are presented in Table 6.1. Items marked with an asterisk (*) have also been clarified in the sections below.

Table 6.1: Model selection criteria

Requirements and considerations	**Fixed**	**Semi-fixed**	**Preferred**
Multiple input variables	✓		
Only quantifiable variables	✓		
High interpretability*	✓		
Positive beta coefficients		✓	
Multicollinearity*		✓	
Constant response variables		✓	
High accessibility*			✓
Low complexity*			✓
Automation potential*			✓

6.1.3 Clarification of terms

In the section below, short descriptions of notable terms are presented.

Statistics versus *machine learning*

There is without a doubt, significant overlap between the models used within the "statistics" and "machine learning" fields. In fact, it is not uncommon to see the terms "statistics" and "machine learning" used interchangeably. To add to this confusion, there is much contention amongst the "experts" and academia, on exactly where their boundaries start and end. Therefore, for the sake of keeping this thesis consistent, I would like to first allude to an article titled: *Points of significance: statistics versus machine learning* by (Bzdok et al. 2018).

They suggest that the biggest difference between these two fields lies in their project objectives. In conventional statistics, the primary goal is "inference", and the model created is "project specific". While in machine learning, the focus is more towards generalisation, and the analysis of 'wide data': where the number of input variables are far greater than the number of dimensions.

Interpretability

Interpretability describes the degree of transparency a model possesses. A model with high interpretability is able to point out the relationship between individual variables; as well as the significance of these variables. The model, in some circumstances, can indeed also be used for prediction, but in most cases, has issues generalising (when tested on unseen data). On the other hand, a model with low interpretability may perform well with overall predictions, and is able to generalise well with unseen data. However, there is a trade off, and the model may have difficulties explaining "why" and "how" these predictions are derived. Due to the sheer complexity, or large number of moving parts of the model, the relationship between individual variables becomes unclear (Schielzeth 2010).

Difference between *modelling* and *algorithm*

There is merit in clearly stating the difference between the term *model* and *algorithm*. Within the literature, these words are occasionally use interchangeably, and can lead to confusion for the reader. That said, an *algorithm* is a set of rules (or the actual code) used to solve a problem. A *model* on the other hand, is the final "formula" that is built after applying an *algorithm* to a dataset.

Accessibility

The accessibility describes how effectively large volumes of raw data can be converted into a model. The procedure should not demand intensive computer processing power, bulky equipment, or high-priced programs.

Complexity

On the same vein, a model which is high in complexity may require complex training in order to set up the dataset, and to run the algorithm.

Automation

Vegvesen's eRoom database is anticipated to progressively increase with continual updates and new information from the industry over time. Ideally, the model should be adaptable to basic automated scripts. The intention is to create an automated process for new data input; and for automated real-time updates to the model.

6.2 Models included in this report

Although a plethora of models exist, this report will only focus on machine learning, statistics, and mathematical optimisation. Due to the constraints of time; and my limited skill-set (achievable within the time frame), this thesis reports only on three principle models. The select-models are intended to represent each of the three major branches in data science. This decision was made to establish a broad but clear "starting point" to this exploratory research; and to develop an overall understanding of the data at hand. Table 6.2 below details the selected models.

Table 6.2: Documented models in this report

Discipline	**Focus**
Statistics	**Chapter 7** *Structured data* Regression analysis
Machine learning (*partial*)	**Chapter 6.3** *Unstructured data* Classification Clustering
Mathematics	**Chapter 8** *Structured data* Mathematical optimisation

The above list is not intended to be a complete list of available models whatsoever. However, it is simply a representation of some of the methods I managed to investigate, and was learned enough to document.

6.3 Machine learning

We would be remiss if we did not mention the heavily-hyped field of machine learning. This discipline is, at its core, concerned with analysing data; running algorithms to create a model; and to make predictions. As perhaps evident, tremendous overlap exists between the branches of data science (specifically statistics and machine learning). To the general public however- the very notion of "machine learning" will immediately invoke the image of artificial intelligence; facial recognition; automation; complex "black box" type prediction models; and so forth. So much so, that the very definition of machine learning, can only be described as dynamic, evolving, and continually edging itself towards this bias public perception.

Nonetheless, this report will also lump these themes together with this rendering of machine learning. The truth of the matter is however, that algorithms from this discipline are not entirely practical nor compatible with NoTCoS's agendas. For starters, classification methods are designed for "categorical" variables. However, in the case of the ETS, their time capacity values are independent of changes in geology. As such, their input variables are strictly quantitative.

At the same time, an interesting statement by (Salimi et al. 2016) spotlights the reality that traditional machine learning algorithms sacrifice interpretability for a "perceived" high prediction accuracy.

> Popular forecasting techniques such as machine learning can act “as a “black box”; meaning that they can be employed to predict the value of a target based on data, but the rules or implicit patterns within the model cannot be interpreted.”
>
> - (Salimi et al. 2016).

This very notion is contradictory to the NoTCoS; and to its equivalent time system (ETS). This judgement call, and the concept of interpretability will be deliberated further in Chapter 10.1. But for now- the dive into "machine learning" waters prematurely ends here. However, recommendations on how machine learning may be implemented in the future, are discussed in the concluding chapters.

6.4 Supporting computer software

Predictive analysis (particularly, algorithms described in this study) are generally impractical to solve by hand. It is therefore common to utilise modern computer processing power and software. A major component of this analysis was conducted using the program *R*. A presentation of all programs used and their relevant code has been included in Appendix A for reference.

6.5 Chapter summary and remarks

In hindsight, I approached this exercise in a roundabout way. Nonetheless, we got there in the end. I had not known it at the time, but the interpretability of model was fundamental to my research agenda; and to writing a scientific and sound report.

Some remarks regarding the selection criteria

In hindsight, this selection criteria was not completely in line with the eventual project objective. Had I instead first clearly defined the project objective, prior to investigating modelling techniques, I would have immediately realised that "machine learning" based models would not have been suitable for the problem at hand.

Chapter 7

Regression Analysis

In this chapter, the concept of regression analysis is explored and the background information is presented. The contents of this chapter has been summarised below.

- Section 7.1 - Background information
- Section 7.2 - Traditional linear regression
- Section 7.3 - Estimating the regression function
- Section 7.4 - Regression-through-the-origin (RTO)

7.1 Background

Regression analysis is a collection of statistical methods used to model the relationship between variables (Kutner et al. 2005). It allows users to study the changes in a *response* variable, when the *predictor* variables are adjusted. Many types of regression exist, and each of these are uniquely appropriate according to the characteristics and features of the data being analysed; and dependent on the project objective. Some examples of the more popular regression algorithms are explored in this chapter, and also listed below:

- Ordinary least squares (OLS); and
- Regression-through-the-origin (RTO).

Literature source

Regression analysis has been well-studied and practiced throughout history. Its theory and analytical techniques are continually refined and updated. It is therefore admittedly difficult to attribute the *basic* fundamental theory to a singular source. Instead, a general list of all literature that eventually contributed to my overall knowledge base, is presented below. However- major and pivotal ideas have been individually cited throughout this chapter, as earnestly as possible.

- *Regression - Models, Methods and Applications* (Fahrmeir et al. 2013)
- *Introduction to linear regression analysis* (Montgomery et al. 2012)
- *Applied linear statistical models* (Kutner et al. 2005)
- *Linear models with R* (Faraway 2016)
- *An Introduction to Statistical Learning: with Applications in R* (James et al. 2014)
- *Applied predictive modeling* (Kuhn and Johnson 2013)
- *NIST/SEMATECH e-Handbook of Statistical Methods* (NIST/SEMATECH 2012)
- *TMA4268 Statistical Learning* (Langaas and Muff 2019)

7.2 Linear regression

Linear regression is a statistics method to model the linear relationship between a scalar *response* (also known as a dependent variable) and one or more *predictor* variables (also known as independent variables). This technique aims to accomplish this by fitting a predictive (regression) line, to an *observed* dataset.

7.2.1 Simple linear regression

When only *one* predictor and *one* response component is considered, this is called **simple linear regression**. This technique can be used to estimate a quantitative response (Y) based on the single predictor (X).

7.2.2 Multiple linear regression

When the regression model demands more than one predictor, it is instead called **multiple linear regression (MLR)**. In reality, the total construction time (response) of a D&B tunnelling project will be influenced by *multiple* construction activity (predictor). MLR is therefore more equipped to handle such a scenario. Yes, there is perhaps merit in fitting a separate model for each individual predictor (in the form of scatterplots: which will be discussed later). However, MLR instead can provide direct insight into the influences of multiple predictors occurring concurrently. The theoretical background, its usefulness and its applicability are discussed in the following section.

7.2.3 The model

Mathematically, the dataset used to create the MLR model is commonly expressed as:

$$\{y_i, x_{i1}, \ldots, x_{ik}\}_{i=1}^{n} \tag{7.1}$$

where:

k = denotes the number of variables in the dataset; and

n = is the number of samples within the dataset.

Thereafter, the MLR model, which describes the relationship between the variables, can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_k X_k + \epsilon \tag{7.2}$$

where:

Y = is the estimated dependent variable (sometimes called, the *response* or the *outcome*). This variable is the estimated quantitative outcome for a given predictor value;

β_0 = is the intercept (sometimes called, the *constant*). This component is constant, and its function is to ensure that the model is *unbiased*[1]. The intercept is simply the expected average of the response, when the predictors (x) equal zero (0);

$\beta_{1...k}$ = is the slope of the model and is also a constant. Along with β_0, these are called the *coefficients* or *parameters*. The regression coefficient (β_k) is the degree of influence a specific predictor will have on the response variable ($\hat{Y}$), should there be a change in the predictor variable (X_k), and the other predictor variables are held constant;

X = is the column vector for independent variable (sometimes called, the *predictor*, or the *regressor*). This represents the component that influences the outcome of the quantitative response Y; and

ε = is a vector of errors of prediction. This is a random variable that accounts for factors that results in the model not fitting perfectly. These *inaccuracies* in the predicted outcome may be due to errors in the measurements; non-constant variable properties; variables unaccounted for; or an incorrect model choice.

7.2.4 Matrix notation

Should each data point (n) be presented as a collective, it can be expressed in a matrix notation as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{7.3}$$

where:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Note:

- $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are $n \times 1$ vectors, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector and $\boldsymbol{X}$ is a $n \times (k+1)$ matrix.
- The Gauss-Markov assumptions are: $E(\boldsymbol{\varepsilon}) = 0$, $Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$ (these will be discussed in a later section).
- These result in $E(\boldsymbol{Y}) = 0$, $Var(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$.

[1] Bias tendency will be discussed further, in the following section

7.2.5 Assumptions

In order to estimate the predictor coefficients – and in turn, the response – several assumptions are typically made. These will vary depending on the model however. Some of the major assumptions relating to linear regression are as follows:

- It is assumed that the errors are to average zero (0);
- It is assumed that the errors have an unknown variance σ^2;
- It is assumed that the errors are uncorrelated: the value of one error should not depend on the value of any other error; and
- Linear relationship between the dependent variable (y) and the k-vector of regressors (x).

7.3 Estimation of the regression function

In order to estimate the coefficients in a multiple linear regression model, several methodologies are possible. However, the procedure chosen is highly dependent on thorough pre-application assessments and preparations, prior to any actual modelling. These will be discussed in the sections below.

7.3.1 The least squares method

In order to fit the dataset to a linear regression models, the method of least squares method is typically used to estimate the model's intercept constant value (β_0), and the predictor coefficients $(\beta_1 \cdots \beta_n)$. This approach - also called *linear least squares* (LLS) - aims to achieve this by methods which minimise the sum of squared residuals (errors). The three primary formulations for LLS are:

- Ordinary least squares (OLS);
- Weighted least squares (WLS); and
- Generalised least squares (GLS).

Supplemental to this, additional techniques will also be discussed in the following sections. Furthermore, as the dataset to be studied in this report will contain exclusively multiple regressor variables, the emphasis in the subsequent sections will on applying LLS methods to multiple linear regression.

7.3.2 Ordinary least squares

The ordinary least squares (OLS) method is an optimization strategy. The method aims to assign a linear regression line which minimises the sum of the square differences between the observed (actual, y_i) and predicted (estimated, $\hat{y}$) values. The mathematical background is illustrated below as instructed in *Introduction to linear regression analysis* (Montgomery et al. 2012). The matrix notation is given as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{7.4}$$

where:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Typically, $\boldsymbol{Y}$ is an $n \times 1$ vector of the observations, $\boldsymbol{X}$ is an $n \times p$ matrix of the levels of the regressor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. The objective of OLS is to calculate the vector of least-squares estimators, $\hat{\boldsymbol{\beta}}$, that minimizes:

$$\boldsymbol{S(\beta)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\varepsilon_i^2} = \boldsymbol{\varepsilon'\varepsilon} = \boldsymbol{(y - X\beta)'(y - X\beta)} \tag{7.5}$$

which can be simplified as:

$$\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'y} \tag{7.6}$$

Equation 7.6 is the least-squares normal equation. After solving the normal equations, the least-squares estimator of $\boldsymbol{\beta}$ then becomes:

$$\boldsymbol{\hat{\beta}} = \boldsymbol{(X'X)^{-1}X'y} \tag{7.7}$$

where:

$E[\hat{\beta}] = \beta$ (unbiasedness) and $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$: where $\hat{\beta}$ is a linear function of the observations $\boldsymbol{Y}$.

7.3.3 Weighted Least Squares

Sometimes, when specific data points are recognised to have more significance than the other data points, the *Weighted Least Squares* method can be employed.

7.3.4 Gauss Markov theorem

According to the Gauss Markov theorem (credited to Carl Friedrich Gauss and Andrey Markov), the OLS methods result in the best linear unbiased estimate (BLUE), if a set of assumptions are met. These conditions are summarised below, as described in the article titled: “Gauss–Markov Theorem in Statistics” (Hallin 2014), and in the textbook: *Applied linear statistical models* (Kutner et al. 2005).

- The regression model is linear in the coefficients and the error term;
- The errors are uncorrelated with each other;
- The errors and the predicted values are uncorrelated with each other;
- The errors have equal variances (no heteroscedasticity);
- The error term has an expected mean of zero (0); and
- The error term is normally distributed *(optional)*.

The term "best" in BLUE, implies that the model will give the lowest variance of the estimate, as compared to other unbiased, linear models. This will be tested in the subsequent model validation phase. Furthermore, in Hallin’s article - as mentioned above - he suggests an amendment to the Gauss Markov theorem should be made for a more realistic approach (Harville 1976; Shaffer 1991).

The intercept and an unbiased model A model is regarded as *unbiased*, when the estimated response values are closely matched to the observed values. Conversely, a *biased* model is one where the estimated response values are largely dissimilar despite the highly accurate predictor coefficients. This is examined during the validation phase, and will be discussed in the validation section.

The primary function of the intercept is to ensure that the model is *unbiased*. It aims to *center* the regression line: to provide the model a “starting point”. Should the dataset space be unlimited or undefined, a model with an erroneous (or non-existent) intercept coefficient may result in large variances between the predicted value and the observed values. This is due to the fact that the predictor beta coefficients ($\beta_1 \dots \beta_k$) can be considered – simply, and only – as the perceived “degree of influence”, a predictor (X) will have on the estimated outcome (Y). Therefore, a biased model, although possessing highly accurate predictor beta coefficients ($\beta_1 \dots \beta_k$), may ultimately produce unrealistic results should an incorrect intercept coefficient is selected.

7.4 Regression-through-the-origin (RTO)

It is commonplace for regression models to contain an intercept coefficient (β_0). It is however, also entirely possible to manually remove this element. This type of human interjection or *inference*, is undeniably controversial and potentially dangerous in the statistical analysis industry: where improper practice may lead to a biased and inaccurate model (Casella 1983; Eisenhauer 2003).

Nonetheless, there are specific scenarios when it is indeed correct to remove the intercept: that is, with a regression-through-the-origin (RTO). This process results in a null (and omittance of the) intercept coefficient (β_0). This implies that when the predictor variable (X) equates to zero (0), then the mean response variable $\hat{Y}$ is also zero (0). There are unique circumstances and conditions that permit this type of practice. These criterion and considerations are summarised in Table 7.1, as instructed by the textbook *Applied linear statistical models* (Kutner et al. 2005); and the articles "Leverage and Regression Through the Origin" (Casella 1983), and "Regression through the Origin" (Eisenhauer 2003).

The mathematical equation for a regression-through-the-origin then transforms to:

$$Y = \beta_1 X_1 + \beta_2 X_2 \cdots + \beta_k X_k + \epsilon \tag{7.8}$$

Specific and unique models require their own unique set of validation procedures. As such, caution must be taken when removing a seemingly harmless intercept. These will be discussed in the model validation section later on.

Table 7.1: Criterion and considerations for regression-through-the-origin (RTO)

Criterion	**Considerations**	**Diagnostic devices**
RTO is permissible when there is a high level of certainty that, as the predictor variables sums to zero (0), the average response variable also equates to zero (0).	Examine the observed range of data. The values closer to the origin may in fact behave initially exponentially, but thereafter stabilise to a linear response around the data points which are most prominent.	• It is beneficial to first, plot the data onto scatter diagrams. This allows the practitioner to observe whether the region in which the majority of the data lies, is in close proximity to the origin. • Should "the data lie in a region of x space remote from the origin" (Montgomery et al. 2012), then it may be appropriate to include an intercept to more accurately align the regression line: to ensure an unbiased model.
RTO is permissible when it is possible for the predictor variables to equate to zero (0).	The dataset should be standardised prior to RTO.	• Alternatively, both iterations – with and without an intercept – can be tested. Thereafter, the quality of the fit can be compared against each other (Hahn 1977).
RTO may be permissible when the predictor variables all exhibit the same direction of influence (that is, either all negative impact or positive impact towards the response variable).	If the response variable is set as a time dimension, then all predictor input variables should in theory possess the same direction of influence.	It is possible to artificially include dummy data point within the dataset, to create a leverage point (Casella 1983).
	The practitioner must be wary when selecting the appropriate model and the eventual computer software used to run an RTO. Varying output values may occur depending on the selected software. (Prvan et al. 2002).	Test the null hypothesis $H_0 : \beta_0 = 0$, and use the t-statistic to investigate the intercept's significance.

Chapter 8

Mathematical Optimisation

In this chapter, basic concepts relating to mathematical optimisation are introduced. The contents of this chapter has been summarised below.

- Section 8.1 - Background information
- Section 8.2 - The basic principles of optimisation
- Section 8.3 - The linear system of equations
- Section 8.4 - The introduction of constraints
- Section 8.5 - Some useful definitions

8.1 Background

Mathematical optimisation, (sometimes called "mathematical programming") is a collection of mathematical tools and principles used to determine a "best" solution, for a given quantitative system, by manipulating a set of variables. Because almost all problems can be defined from a mathematical point of view, optimisation has become ubiquitous and integral to a wide range of disciplines: such as science, engineering, economics, logistics and scheduling. To serve such a diverse array of functions and users, many unique optimisation methods have developed. These techniques are generally distinguishable by their class, and are characterised by their function and system (or mathematical properties). Some notable distinctions between these classes include;

- continuous or discrete quantities;
- linear or nonlinear programming;
- deterministic or stochastic solutions;
- unconstrained or constrained problems; and
- singular, multiple, or even, the omission of an objective function

Following the preliminary review of general mathematical optimisation, this study has decided to focus on the "linear system of equations" and "constrained optimisation". The decision was made by considering its sheer relevance to the research agenda, and by taking into account time restrictions. Though, untested - but surely useful - other methods are also discussed in the concluding chapters of this report (Chapter 16 - Recommendations)

Literature source

Mathematical optimisation concepts have been well-documented and practiced throughout history. Its theory and analytical techniques are continually refined and updated. It is therefore admittedly difficult to attribute the *basic* fundamental theory to a singular source. Instead, a general list of all literature that eventually contributed to my overall knowledge base, is presented below. However- major and pivotal ideas have been individually cited throughout this chapter, as earnestly as possible.

- *Numerical optimization* (Nocedal and Wright 2006)
- *Practical methods of optimization* (Fletcher 2013)
- *Practical optimization* (Gill et al. 2019)
- TMA4180 - Optimization 1: classroom material (Bogfjellmo 2019)

8.2 The basic principles of optimisation

In an optimisation problem, the "best" solution derived with optimisation techniques is governed by the following three elements:

- The **objective function(s)**;
- the **variables**; and
- if required, the **constraints**.

In *unconstrained* optimisation, the objective function (or the domain), is a numerical quantity that needs to optimised: which either means maximising or minimising its value. While the variables (or its degrees of freedom), are the values which are manipulated, in order to achieve the objective function. However-, in the case of *constrained* optimisation, additional constraints (such equations and inequalities) are imposed: which act as restrictions for the variables and their degrees of freedom.

8.3 The linear system of equations

For some optimisation problems, the "variables" are controlled by the input data. In quadratic programming, these inputs can be described as a "linear system of equations". Mathematically, this is:

$$\boldsymbol{A}x = b \tag{8.1}$$

with:

$$\boldsymbol{A} \in \mathbb{R}^{m \times n}, \qquad x \in \mathbb{R}^{n}, \qquad b \in \mathbb{R}^{m},$$

where:

$\boldsymbol{A}$ = a matrix, where its length is $m \times n$);

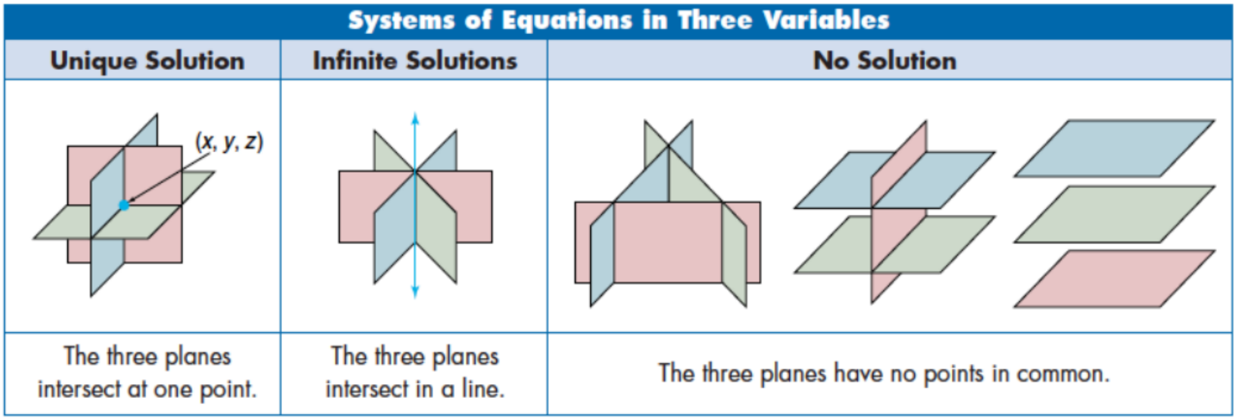
x = the unknown vector, which is a vector of length n; and

b = the total number of equations, which is of length m.

The linear system is not necessarily always a square-matrix (where $m \times n$). In the scenario where there are more unknowns (or dimensions) than the number of equations ($m < n$), the system is described as being *underdetermined*. On the contrary, for cases where there are more equations than the number of dimensions ($m > n$), the system is considered *overdetermined*.

8.3.1 The solution

Depending on the matrix, the optimisation problem may have different types of solutions. When there is only one singular solution, it is considered "consistent independent". In the case of *underdetermined* systems, these generally have an infinite number of solutions, and are labeled as "consistent dependent". These can be visualised in Figure 8.1. Oppositely, when there are no solutions, the system is deemed "inconsistent".



Systems of Equations in Three Variables		
Unique Solution	Infinite Solutions	No Solution
The three planes intersect at one point.	The three planes intersect in a line.	The three planes have no points in common.

Figure 8.1: The many different types of possible solutions for a given system of equations (School 2012)

However, there is the notion of an "approximate solution" to the system. Should a problem have more equations than unknowns, the system is considered to be *overdetermined*. Although there is usually no *exact* solution, a "closest" solution is still possible (as shown in Figure 8.2).

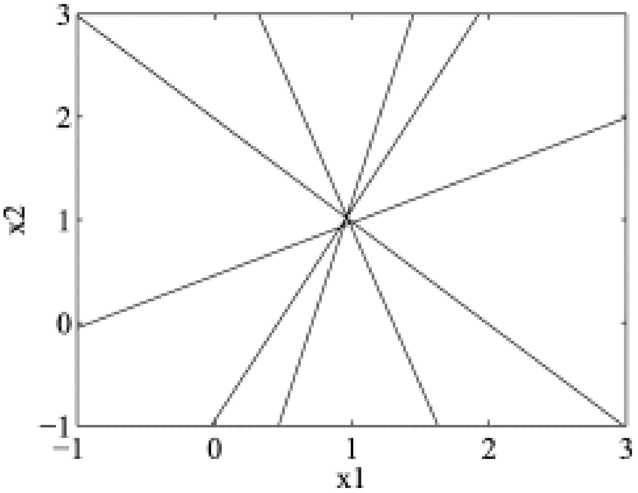


Figure 8.2: When the system is overdetermined, only an approximate solution is possible (Lindfield and Penny 2019)

Regarding an approximate solution, this is achieved by searching for a vector $x^{\star} \in \mathbb{R}^n$ that minimises the Euclidean norm of the residual vector (Equation 8.2).

$$r(x^{\star}) = \boldsymbol{A}x^{\star} - b \tag{8.2}$$

With this in mind, the mathematical formula for such a solution can be represented as:

$$\left\| r\left(x^{*}\right) \right\|_2^2 = \left\| \mathbf{A}x^{*} - b \right\|_2^2 \leq \|\mathbf{A}y - b\|_2^2 = \|r(y)\|_2^2 \tag{8.3}$$

This shall hold for every $y \in \mathbb{R}^n$. All in all, this minimiser, (if it exists) is called the least square solution of an *overdetermined* system. For an *underdetermined* system, there is an infinite amount of solutions (Shen 2015).

8.4 The introduction of constraints

For certain specific tasks and scenarios, it is sometimes necessary for the conventional least squares method to be reformulated with linear inequality or equality constraints. This can act as a boundary or limit for the model; which if implemented correctly, may provide a fit to the data that is more in line with the objective at hand.

8.4.1 Non-negative least squares

Should non-negative constraints be required, the non-negative least squares (NNLS) algorithm is typically applied. NNLS aims to solve the least squares problem with only non-negative coefficients.

This section presents an active-set method to solve the NNLS problem, using the procedures set out in the textbook *Solving Least Squares Problems* (Lawson and Hanson 1974, Chapter 23, p. 161), and (Lawson and Hanson 1995; Haskell and Hanson 1981)

The NNLS algorithm

Given a matrix $\boldsymbol{A} \in R^{m\times n}$ and the set of observed values given by $\boldsymbol{b} \in R^{m}$, find a non-negative vector $\boldsymbol{x} \in R^{n}$ to minimise the function $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{Ax}-\boldsymbol{b}\|^2$, i.e.

$$\min_x f(x) = \frac{1}{2}\|Ax-b\|^2 \tag{8.4}$$

$$\text{subject to } \boldsymbol{x} \geq 0.$$

Accompanying software

The statistical program *R*, and their "nnnpls" package allows for the implementation of least squares with both non-negative *and* non-positive constraints. The MATLAB variant on the other hand only permits non-negative constraints. Some relevant information, such as the code and arguments have been documented within the Appendix (Appendix A.3) for reference.

According to the Mathworks webpage (MathWorks 2019a).

> lsqnonneg uses the algorithm described in (Lawson and Hanson 1995). The algorithm starts with a set of possible basis vectors and computes the associated dual vector lambda. It then selects the basis vector corresponding to the maximum value in lambda to swap it out of the basis in exchange for another possible candidate. This continues until lambda ≤ 0.

This method is very much a “brute force” type approach, often requiring a large number of iterations, step-cycles and repetitions. Hand calculations are therefore not advised. The study “A fast non-negativity-constrained least squares algorithm” has been published in an attempt to improve the efficiency of the active-set method (Bro and De Jong 1997).

8.4.2 Statistical algorithm

Non-negative constraints can, in theory, also be applied to statistical models: such as linear regression. This is possible by rephrasing the quadratic programming algorithm using basic statistical concepts (D. Q. Wang et al.

2004). However- concerns regarding the validity arise, when such a method is applied. When the model demands such non-negative constraints, it typically signals that the problem is no longer (or was never in the first place) a statistical problem.

8.5 A useful definition

Below, a useful definition relating to mathematical optimisation is presented. This concept was quite important during the exploratory and analysis components.

Superposition principle

The superposition principle (or superposition property) describes the concept that the net response of two or more stimulus is the sum of the individual response caused by each stimulus individually. This principle forms the fundamental characteristics of a linear function. For input variables with a time dimension, the superposition principle can be expressed mathematically as:

$$f(x_1(t)) = y_1(t) \tag{8.5}$$

$$f(x_2(t)) = y_2(t) \tag{8.6}$$

$$f(x_1(t) + x_2(t)) = y_1(t) + y_2(t) \tag{8.7}$$

Broken down further, if the two properties "additivity" and "homogeneity" are satisfied, then the function is defined as a linear function.

$$\text{Additivity} \qquad f(x_1 + x_2) = f(x_1) + f(x_2) \tag{8.8}$$

$$\text{Homogeneity} \qquad f(c \cdot x_1) = c \cdot f(x_1) \tag{8.9}$$

where, c is a scalar (constant).

Chapter 9

Model Validation and Diagnostics

In this chapter, model validation and diagnostics techniques are introduced. These methods are useful for objectively assessing the model's outputs, and the internal interactions of the model; and for evaluating the model's validity and performance. The objective of this step is to measure the reliability of the output results; and its aptness to the research question. Both internal and external tests are discussed and performed in the section below. The contents of this chapter has been summarised below.

- Section 9.1 - Significance of regression
- Section 9.2 - Sufficient sample size
- Section 9.3 - Resampling methods
- Section 9.4 - Some useful definitions

9.1 Significance of regression

The significance of the regression test can be used to examine whether or not there is a linear relationship between the response and any of the predictor variables. The procedure is intended to assess the model's adequacy. The test procedure is commonly presented in an analysis of variance table such as in Table 9.1 below.

Table 9.1: Analysis of variance for significance of regression

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	F_0 **(F-statistic)**
Regression	SS_R	k	MS_R	$\frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

If $\hat{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{9.1}$$

then the variance of the data can be evaluated using three sums of squares formulas (presented below).

The total sum of squares (SS_{tot}) (proportional to the variance of the data):

$$SS_{tot} = \sum_i \left(y_i - \bar{y}\right)^2 \tag{9.2}$$

The regression sum of squares SS_{reg} (also called the explained sum of squares):

$$SS_{reg} = \sum_i \left(\hat{y}_i - \bar{y}\right)^2 \tag{9.3}$$

The sum of squares of residuals SS_{res} (also called the residual sum of squares):

$$SS_{res} = \sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2 \tag{9.4}$$

where:

y_i = the actual y value for the observation i

$\hat{y}_i$ = the predicted value for y for observation i

$\bar{y}$ = the mean of the y value

9.1.1 R-squared test

R^2 (also known as coefficient of determination): used to indicate the goodness of fit of a model and its precision (Draper and Smith 1998). The most general definition of the coefficient of determination is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{9.5}$$

This is also commonly expressed as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{9.6}$$

Therefore, values closer to 1 indicate that the relationship between variables can be explained by a large percentage of the variation in the data.

9.1.2 Adjustment for a regression-through-the-origin

If the intercept is compromised however, an adjustment to the coefficient of determination is required (Eisenhauer 2003). For models that exclude an intercept (regression-through-the origin), the no-intercept model analogue then becomes:

$$R_0^2 = 1 - \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} \tag{9.7}$$

Some limitations to this technique The R^2 value should however be used with caution. According to (Barrett 1974), it is a useful tool to measure how closely the estimated points fit the regression surface, and to inspect the steepness of the regression surface. He caveats to readers however that a steeper regression line is prone to inflate the R^2 value, and misrepresent the model's precision. Barrett, in his article "The Coefficient of Determination-Some Limitations" (Barrett 1974), suggests that the predictive precision of a regression equation may be more useful value than R^2. Furthermore, he proposes graphs and confidence intervals alongside an R^2 test for a more thorough and complete assessment of the regression model. All in all, these techniques will be performed together with R^2 analysis to help validate the model.

Adjusted and predicted r-squared When comparing models that contain a different number of predictors, it may be useful to instead use the *adjusted* r-square. Additionally, when attempting to assess how well the model will interact with new observations, the *predicted* r-square should be used.

9.2 Sufficient sample size

An appropriate sample size is required for validity, by increasing the precision of representative data, and by minimizing the margin of error (Hsieh et al. 1998; Dupont and Plummer Jr 1998). The sample size can be examined by running confidence interval and confidence level testing. The lower the confidence interval required, the higher sample size is needed.

The sample size can be assessed according to its confidence intervals. This suggests that if the population parameter can be determined, then the statistician can examine whether or not the samples lie within this confidence interval.

Sample size calculation based on the effect size

An alternative approach of calculating the sample size is effect size. Effect size is known as the difference between the sample statistics divided by the standard error. The potential benefits are discussed in more detail in "Using effect size—or why the P value is not enough" (Sullivan and Feinn 2012).

9.3 Resampling methods

Below, some resampling methods are discussed.

9.3.1 External validation

An external validation of the model can be performed using additional sample data from the same population or *similar* population.

> Here, the term "population" is a term used to describe a large set of objects or events (circumstances) that are of similar nature. A specific example of this may include: female students aged 16-21 in Bergen, or white painted houses in Trondheim. In our case however, a similar population may be a tunnel project that was constructed using the same construction methodologies, and through similar geological conditions; and with a similar tunnel cross section.

Should such supplementary samples be obtained, this new data can be used to test the model's performance. This involves applying the new sample data to the initial model, and then assessing components such as the goodness-of-fit, and for validity's sake. As described by (Harrell Jr et al. 1996), this process is referred to as an "external" validation. It is commonly agreed to be an extremely effective and unbiased test method to inspect not just the model, but also the data collection process.

9.3.2 Internal validation

Conversely, internal validation of a model can be also be performed. The most recognised and most frequently applied techniques include data-splitting, repeated "data-splitting", "jack-knife" techniques and "bootstrapping". As with external validation, these methods are similarly designed to assess a model's performance. The list of available methods is noticeably more comprehensive than external validation methods.

Nonetheless, a common feature between each method, is that they involve some sort of partitioning of the initial dataset into a "training" and "testing" subset. The difference between each method is in the manner in which they select these subsets. A short description is presented below.

> *Training* data is the data which is set aside for model development. On the other hand, the *testing* data is the remaining data (from the same population), that will be applied to the developed model, in order to examine its performance.

Data-splitting: A random portion of the initial dataset, (usually between two-thirds and three-quarters) is designated as the training set. The remaining data is thereafter set aside to be used as the testing set (Harrell Jr et al. 1996; Picard and Berk 1990; Snee 1977).

Repeated data-splitting: This method simply implies that "data-splitting" is repeated numerous times – but with iteration, a new and random portion is defined. This analysis is considered more robust and accurate than a simple "data-splitting" (Harrell Jr et al. 1996).

Jack-knife technique: This technique is very similar to data-splitting. Instead of a random set of allocated for *testing* however, only one sample is selected (at random) for *testing*. Numerous iterations are performed: typically, as many times as there are samples (Stone 1974).

Bootstrapping: Bootstrapping begins the same way, where samples are divided into a training or testing set. However, samples removed for testing, are replaced with random samples from the original dataset. Consequently, duplicate-data within the training set may arise. According to (Harrell Jr et al. 1996), this replication procedure allows the testing of a “large” sample despite the original sample being “small”. Controversially, if the original data set is erroneous (or not representative of the population), this sampling technique may not be effective. Furthermore, the results interpretation and inference are considered contentious. Consequently, it may be insincere for someone as inexperienced as me to apply this method without complete assurances. (Efron and Tibshirani 1994)

9.4 Some useful definitions

Before proceeding, it is worthwhile to define some of the terms and concepts used in this report.

The curse of dimensionality

This expression describes the phenomenon, that as you increase the number dimensions to a problem (in this case, the number of construction activities to the model), the volume space also increases- but at an exponential rate. In turn, the number of samples required for reliably prediction also increases in a similar fashion. Therefore, there exists a "sweet-spot" between a too-complex model (too many variables) and one that is "missing variables".

An interesting analogy is to examine the k-nearest neighbour algorithm. The near neighbour method usually requires 10% of the data points per neighbourhood capture. However, the nearest neighbours tend to be far away in high dimensions problems. As shown in Figure 9.1, in a single dimension problems, the radius of this neighbourhood is roughly 10% of the sample width. However, as the dimensions increase in size, the radius required increasingly significantly. As a result, the algorithm has trouble finding a suitable "neighbour".

Multicollinearity

In statistics, the term multicollinearity implies that there is a "high" degree of correlation between the predictor variables. In these scenarios, the model may have difficulties distinguishing which variables are in fact influencing a change in the response variable. In particularly sensitive datasets, the model may incorrectly distribute the beta

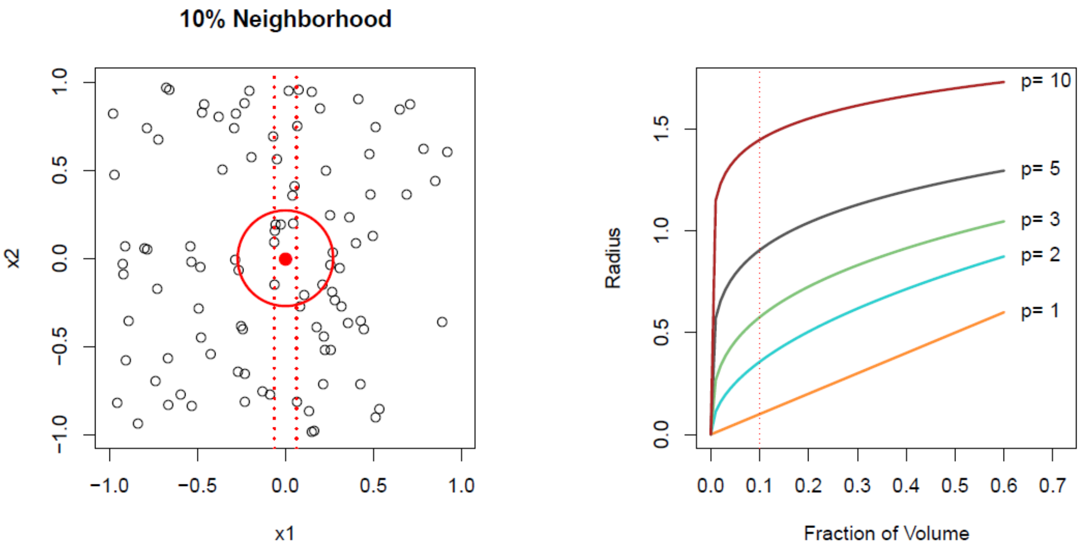


Figure 9.1: The k-nearest neighbour algorithm becomes increasingly difficult to locate a suitable point as the number of dimensions increases (Source: Unknown)

coefficients and the errors between highly correlated predictor variables. There are several techniques one may be able to employ to reduce such effects. These will be discussed throughout the thesis when relevant (Kovács et al. 2005; Farrar and Glauber 1967).

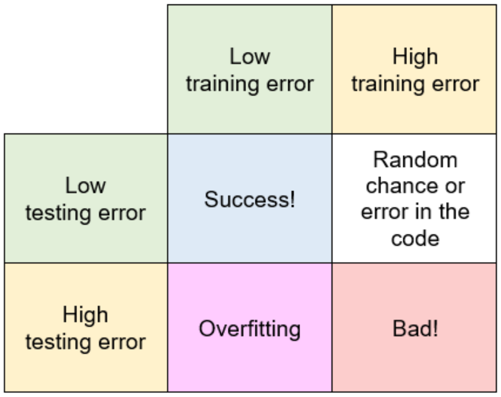
An *overfit* model

Overfitting is a phenomenon that occurs when a model or algorithm too closely models its training set (this is the sample data set initially used to develop the model). It is described as the inability of a model to *generalise* the training set. The model can typically produce estimations that match *most* closely with its training set (by returning the least amount of squared-errors, compared to any other model). However, the model breaks down or becomes unreliable, as soon as new data is introduced (Sammut and Webb 2010). It is therefore important that the data scientist selects a model that strikes a balance between achieving a low variance across the training set, *and* the testing set. The "sweet-spot" can be visualised in Figure 9.2.

Heteroscedasticity

Heteroscedasticity describes a model which possesses unequal variability between the residuals and the fitted values. It is important to assess this parameter as its existence in a model can undermine validation methods such as the analysis of variance or statistical tests of significance. Furthermore, it breaks several assumptions as described in the Gauss-Markov theorem, particularly:

- The errors are to be uncorrelated with each other;



	Low training error	High training error
Low testing error	Success!	Random chance or error in the code
High testing error	Overfitting	Bad!

Figure 9.2: Overfitting occurs when training errors are minimum, and testing errors are high

- The errors and the predicted values are to be uncorrelated with each other; and
- The errors are to have equal variances (no heteroscedasticity).

Heteroscedasticity infers that the regression model chosen may not be Best Linear Unbiased Estimator (BLUE). The inverse of this behaviour is homoscedasticity. As described by (Faraway 2016), the presence of heteroscedasticity can be assessed by first visualising the data in a residual versus fitted plot. Figure 9.3 (left) suggests a regression model with uncorrelated variables. The second plot (middle) however, illustrates non-constant variance (heteroscedasticity). Finally, the third (right) indicates some non-linearity, and should prompt some change in the structural form of the model.

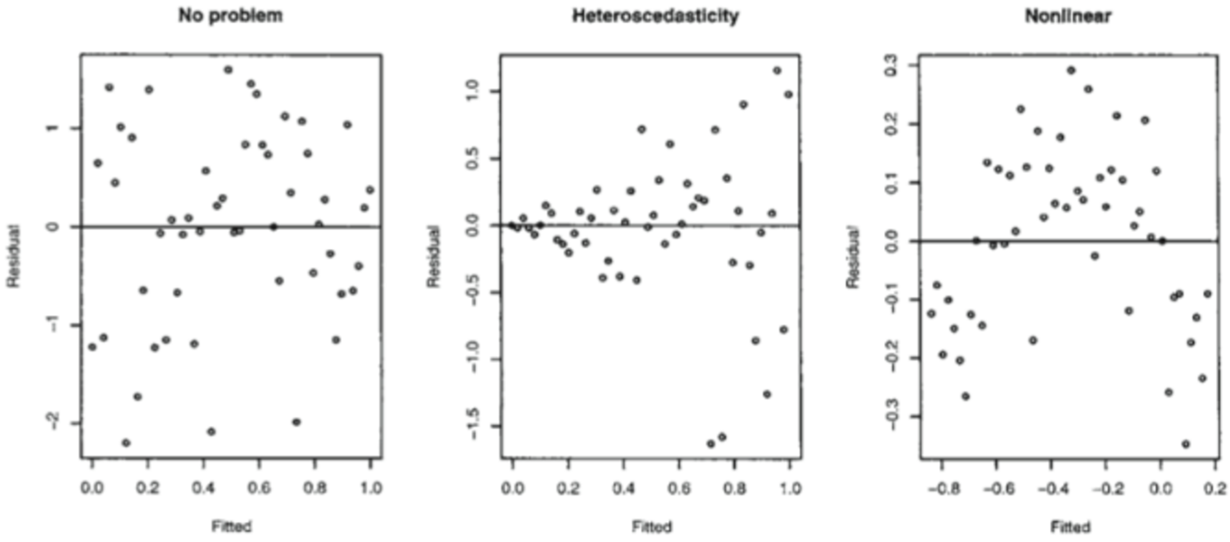


Figure 9.3: Residuals vs. fitted plots illustrating heteroscedastic behaviour (Faraway 2016)

In reality, there will be *some* correlation between the data points. The emphasis, and attention is therefore placed on its magnitude. Figure 9.4 by Daniel J. Hocking (an assistant professor in the Biology Department at

Frostburg State University) illustrates the varying types of homoscedastic and heteroscedastic behaviour.

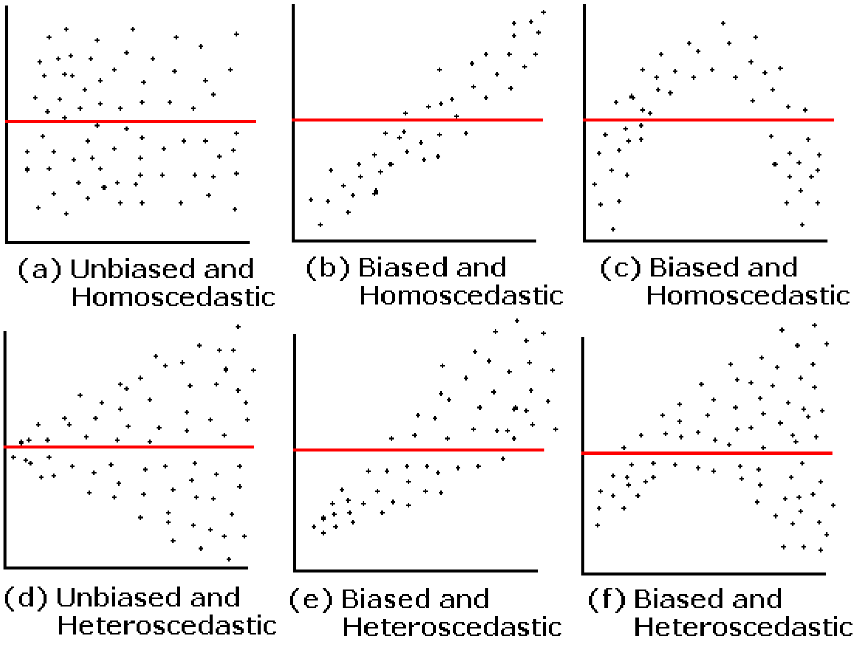


Figure 9.4: Distinguishable types of homoscedastic and heteroscedastic behaviour (Hocking 2011)

Anscombe's quartet

One must be wary when validating one's model. As pointed out in the journal entry titled: "Graphs in statistical analysis" in The American Statistician (Anscombe 1973), datasets can perform very similarly in simple descriptive statistic validation tests, despite possessing very different distributions. The statistician must therefore be careful and thorough when assessing the relevance of their model. In his article, Anscombe highlights the importance of visualising the data; as he believes that there is too much reliance on numerical calculations. Figure 9.5 below demonstrates this phenomenon.

9.5 Evaluating the performance of an optimisation solution

For mathematical optimisation problems, it may be futile to assess the performance of a model using statistical validation techniques. Take for example the R^2 test- a model validation technique used to measure how much of the data the model is able to encapsulate. Conversely, an optimisation algorithm, by design, combines both noise and signal to derive its estimated coefficients. Therefore, by its very nature optimisation models will naturally score very highly in R^2 testing, and results must not be taken wholeheartedly.

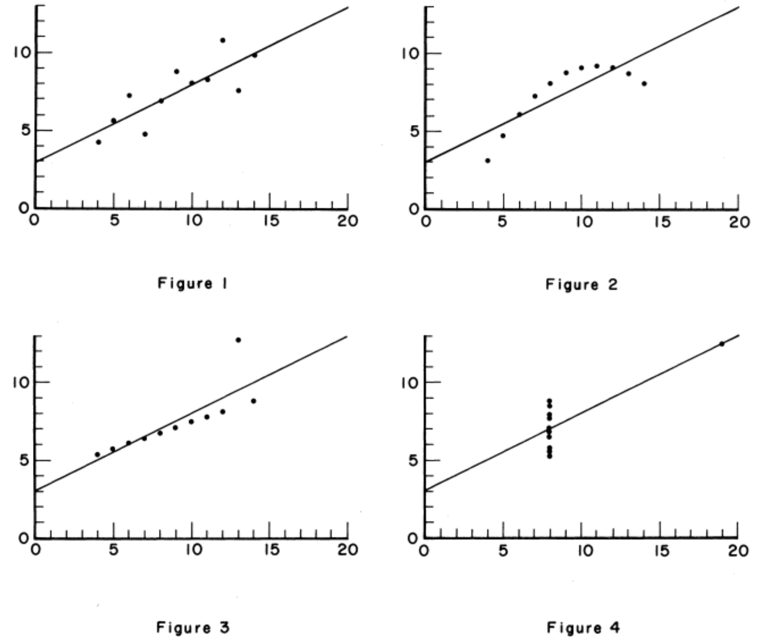


Figure 1

Figure 2

Figure 3

Figure 4

Figure 9.5: Four distinct distributions with similar simple descriptive statistics (Anscombe 1973)

Part III

Exploratory Phase

The *Right* Model, for the *Right* Job

In this exploratory phase, the data science process model is applied to *real* D&B tunnelling construction data. The aim of this exercise is to identify an *effective* and *usable* model. In short, this implies that the selected-model is one that fulfils the following criteria:

- The model satisfies the project objective and its conditions;
- The model meets all model validation testing requirements; and
- The model scores the highest in model performance testing.

For the sake of coherence, the processes in the workflow have been separated into three components. The exploratory stage first begins with project understanding. Thereafter, *real* tunnel data is analysed using a data science workflow. Finally, the results and any inferences are discussed; and a model is selected for further analysis. A summary of this exploratory part is presented below.

Category	**Elements**	**Subject concerning**	**Chapter**
Problem understanding	• Objective definition • Constraints and assumptions	*The Norwegian D&B tunnelling industry*	**Chapter 10**
Data analysis	• Data preparation • Exploratory Data Analysis • Data modelling • Model validation • Results	• The Svartås-tunnel • The Kongsberg-tunnel	**Chapter 11** **Chapter 12**
Decision making	• Inferences • Model selection	*All data*	**Chapter 13**

Chapter 10

Project Understanding

> "Far better an approximate answer to the *right* question, which is often vague, than an exact answer to the *wrong* question, which can always be made precise."
>
> - John Tukey (Tukey 1962a)

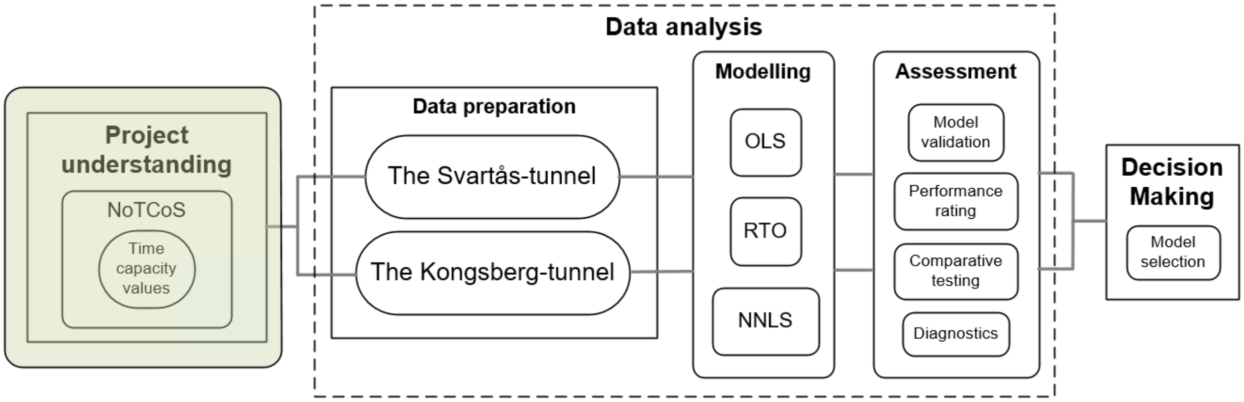


Figure 10.1: A progression chart of the data science workflow - project understanding

It might have been painfully obvious that the introductory chapters leaned more towards the lengthier side. However, effective data analysis can be only stem from a firm understanding of the project characteristics, and its requirements. In this chapter, the project objective is first clearly defined. Only then, can a target objective function be formulated. Overall, the following steps of the data science approach are explored:

- Definition of the project objective (section 10.1)
- Establishing an objective function (section 10.2)

Figure 10.1 shows a nifty progression chart illustrating which stage the chapter pertains to in the data science workflow.

10.1 Definition of the project objective

The project objective is to derive useful "time capacity values" for Norwegian D&B tunnelling construction. These performance rates are to be developed in line with the requirements set out in the Norwegian Tunnelling Contract System (NoTCoS)'s and their "equivalent time system" (ETS); as well as *Norsk Forening for Fjellsprengningsteknikk*'s (NFF) agendas. This notably implies that the model must require parametric estimating values, a methodology that has been perhaps underexplored in the current literature.

10.1.1 Characteristics of the equivalent time system

The ETS is characterised by the following assumptions and conditions, as described in Table 10.1 below.

Table 10.1: Characteristics of the time equivalent system

Characteristic	Description
Characteristic 1:	Rigging times are incorporated into the time capacity value
Characteristic 2:	The ETS assumes that only *one* construction task is being performed at the tunnel face at any given time
Characteristic 3:	Performance rates are independent of varying geology and other site-specific conditions
Characteristic 4:	The ETS assumes work to be continuous

Characteristic 1: The time capacity value is expected to be inflated according to this additional rigging element. The idea is that the rigging component, over the entire construction duration, will be self-correcting.
Characteristic 2: Logistically, construction tasks at the tunnel face cannot be performed simultaneously.
Characteristic 3: ETS's time capacity values are intended to reflect the *average* performance rate.
Characteristic 4: ETS's time capacity values do not account for any downtime between construction tasks. The repercussions of non-working time (waiting and down-time) is therefore absorbed completely by the contractor.

10.1.2 Objective of NFF's research agenda

Furthermore, the overall objective of NFF's research (NFF 2019) endeavours to develop a model that fulfills the following requirements:

- The model should be used for time scheduling within uncertainty for tunnel activities.
- The model should be applicable to other disciplines within the transport industry, in addition to tunnel activities.
- The model should be invoked for reference only: discretionary assessment should be made on individual contracts.

- The model should be tool for adjusting the deadlines for tunnel works after the contract has been signed.
- The model only accounts for "critical" construction tasks: works performed at the tunnel face (where changes to the quantity of works for activities behind the face will be addressed separately).

10.2 Establishing an objective function

In consideration to the project objective, the model must thereafter be able to:

> Estimate the *average* time consumption for individual critical construction tasks (performed at the tunnel face).

In terms of predictive modelling, the objective function is then constrained to the following requirements (Table 10.2).

Table 10.2: Objective function requirements

Requirement	**Description**
Requirement 1:	A constant time capacity value
Requirement 2:	Superposition properties
Requirement 3:	High interpretability

Requirement 1: Time capacity values are unaffected by varying geology and other site-specific conditions. NoTCoS time capacity values resemble the average performance rate over the entire construction duration.
Requirement 2: Construction tasks performed at the tunnel face, are critical: meaning they cannot occur simultaneously with each other.
Requirement 3: The time equivalent system aims to provide flexibility as opposed to accuracy, to Norwegian D&B tunnelling time scheduling. In terms of the objective function, this implies that the model needs to possess high interpretability.

Objective summary

Overall, the objective function implies that the model estimates will be representative of the *average* performance rate over the entire construction duration. To achieve this, the model will presumably focus more on the internal implicit patterns as opposed to the overall prediction accuracy.

10.3 The next course of action

In the next phase of the data science workflow, data preparation; -exploration; and -modelling is performed. In these chapters, *real* tunnel projects (the Svartås-tunnel and Kongsberg-tunnel) will be prepared and scrutinised.

Chapter 11

Data Analysis: The Svartås-tunnel

"Give me six hours to chop down a tree and I will spend the first four sharpening the axe."

- Abraham Lincoln

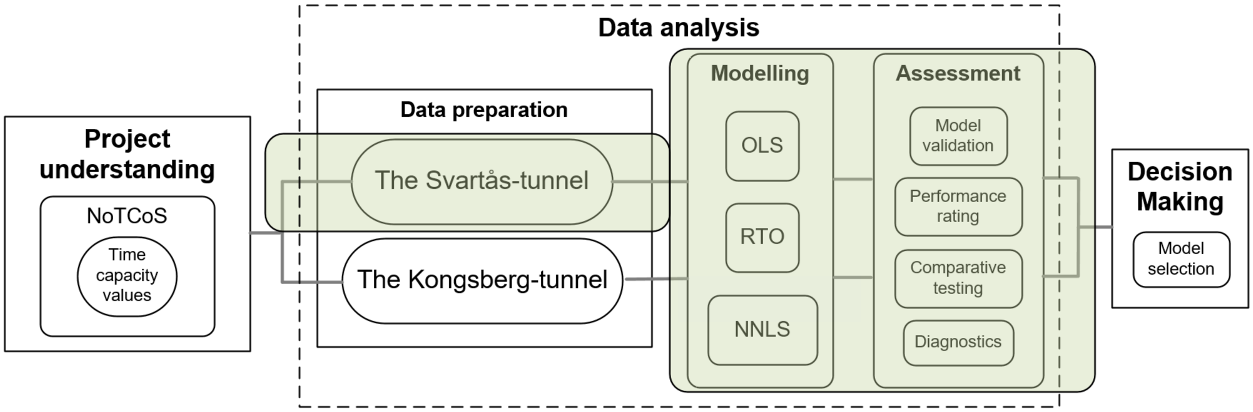


Figure 11.1: A progression chart of the data science workflow - data analysis: the Svartås-tunnel

Public perception of "prediction modelling" and the ilk, is often fixated towards their complex computerised algorithms, or large computational capabilities. Understandable- their resulting predictions and inferences are, after all, most interesting to the audience or the client. In practice however, the majority of efforts and time are actually given priority to the data preparation process (Crowdflower 2015) - sometimes, as much as 80% even (Crowdflower 2016). Either way, this process is absolutely crucial. Should this process be simply glossed over, the dataset may be not representative of reality, and predictions and inferences may be unreliable - despite scoring highly on conventional model validation techniques.

In this chapter, the data science methodology (in particular the data preparation and analysis phases) (as described in Chapter 5) is applied to the Svartås-tunnel data. However- in order to avoid repetition, subsequent tunnel databases are notably less bulky. Figure 11.1 shows a progression chart illustrating which stages the chapter pertains to in the data science workflow.

Table 11.1: Select data science work stages required for the Svartås-tunnel data

Category	**Phases**	**Action**
Project understanding	Objective definition	Unchanged
	Constraints and assumptions diagnosis	Unchanged
Data preparation and exploration	Data collection	Unchanged
	Data wrangling	**Required**
	Exploratory Data Analysis (EDA)	**Required**
Modelling	Model selection	**Required**
	Modelling	**Required**
	Model validation	**Required**
Communication	Inferences	**Required**
	Real-world application	Unchanged

The primary data analysis steps included in this chapter are:

- Data preparation and exploration;
- Modelling; and
- Communication.

11.1 Data preparation and exploration

To begin, the data preparation and exploration steps are first performed. A breakdown of the processes involved is presented below.

- Data collection
- Data wrangling
  - Data description
  - Data import
  - Data clean up
  - Data transformation
  - Feature engineering
- Exploratory Data Analysis (EDA)

Figure 11.2: The Svartås-tunnel under construction, September 2019 (Kjell Wold 2019)

11.1.1 Data collection

The data "collected" for analysis are simply Norwegian D&B tunnelling construction BoQs. A more-detailed description of the data, as well as their limitations can be referenced in Chapter 4 - Methodology and Data.

11.1.2 Data wrangling

The required steps in the data wrangling stage are as follows:

- Data description;
- Data import;
- Data clean up; and then
- Feature engineering.

Data description

The Svartås-tunnel is a two-tube road tunnel, between Sellikdalen and Trollerudmoen, in Kongsberg, Norway (Figure 11.2). With a cross-section of T9.5 (width: 9.5m), the tunnel allows for two lanes of traffic from both directions across the 1.5 km stretch. Excavation was performed using the switch-tube method, from two directions (east and west), with a total of four simultaneous faces: Sellikdalen (east and west) and Trollerudmoen (east and west). This resulted in four separate construction registries. All in all, the obtained construction data spans over approximately 60 working weeks, and involves 82 unique construction registries. A summary of the key figures has been presented in Table 11.2.

Table 11.2: The Svartås-tunnel bill of quantities description

Raw data description			
Tunnel face	Excavation method	Observations n	Dimensions k
Sellikdalen (east)	Switch-tube	38 (weeks)	82
Sellikdalen (west)	Switch-tube	28 (weeks)	82
Trollerudmoen (east)	Switch-tube	55 (weeks)	82
Trollerudmoen (west)	Switch-tube	55 (weeks)	82
Total		**176**	**82**

Data import

The Svartås-tunnel data were originally photocopy scans of construction invoices. These files were obtained as a portable document format (.PDF) file, and therefore required conversion to a digital format prior to data analysis. This process was performed using Microsoft's Excel program, and thereafter stored as an excel spreadsheet file (.XLS) and a comma-separated values file (.CSV).

Standardised format When sifting through tunnel construction logs, it was clear that their formats varied depending on the contractor in charge. For some, a weekly format was the preferred format - while others instead opted for daily records. So, to maintain consistency between the datasets, recorded quantities were organised into weekly-totals (as opposed to daily values). Each observation therefore represents the work performed in one week (Monday to Saturday). At the same time, the duration of each week is assumed to be 101 hours - the industry standard. As a result, the initial response variable (notation: Y_0) for each observation is set as 101 hours.

Data clean up

Only activities that directly influence time consumption at the tunnel face were included in the initial dataset. Some examples of non-relevant quantities that were excluded from the dataset include:

- Quantities of raw construction material; and
- Tasks performed behind the tunnel face.

Following the rejection of non-relevant data, correlation plots were created to quickly visualise any abnormalities with the dataset (included in Appendix B for reference). These plot revealed that perhaps outliers existed. An "outlier" may be identified as: an observation (working-week), in which a significant portion of the time spent had been dedicated to *non-regular* construction tasks (i.e.: those not defined as an eventual input variable $X_1 \dots X_{10}$). Because of the sheer variability in the amount of time consumed, and their infrequent occurrence rate, these tasks are often very difficult to model, and hence characterised as an outlier. Some examples included:

- Mobilisation, construction start-up, and pack-up phases; and
- Construction of secondary tunnel elements:

- Connection tunnels
- Niches (for emergency kiosks)
- Excavation for technical structures.

These outlier observations were excluded from the dataset and not included in the modelling process. From the initial 176 total observations, the final dataset was then reduced to 135.

Initial problems with a constant response variable

D&B tunnelling invoices and their quantities are typically documented on a (constant) 101-hour working week basis: which is indeed a practical feature for billing purposes. Nevertheless, when examined from a statistical analysis point of view, a constant variable is not desirable.

At the outset of this research, the original weekly working-hours were assigned as the response variable. A constant variable however, hampers conventional regression analysis techniques. Through trial, it was evident that programming software could not run such a model: resulting in error messages and eventual model breakdown.

A proposed solution was however put forward: was it possible to nullify this "constant" characteristic, by simply incrementally increasing – ever so slightly – the response value across the entire dataset? This "non-constant modifier" is then expressed mathematically as:

$$Y_i^\star = Y_i + (0.0001 \times i) \tag{11.1}$$

where:

$Y_i^\star$ = the non-constant modified response variable of the i^{th} sample
Y_i = the original response variable of the i^{th} sample
i = the sample number

An example dataset, with a non-constant modifier applied, may then look like the following (Table 11.3).

Table 11.3: Response variable adjusted with a "non-constant modifier"

Sample number (i)	**Original response variable** (Y_i)	**Non-constant modified response variable** ($Y_i^\star$)
1	101.0	101.0001
2	101.0	101.0002
3	101.0	101.0003
4	101.0	101.0004

Although equation 11.1 perhaps still fundamentally breaks the laws of regression modelling, analyse of this type is still useful when investigating the construction rates of particular tasks. Besides, such a "non-constant modifying" ultimately enabled statistical programs to function after all. Nonetheless, this hiccup may be a clue:

that the equivalent time system (ETS) and its time capacity values, may in fact not be a statistical one. Despite serious inquiry, I was unable to locate any literature regarding such a technique. I was therefore unable to validate such a bizarre "solution". Regardless, new findings uncovered using a regression approach will hopefully reveal more.

Feature Engineering

To limit the effects of multicollinearity and the curse of dimensionality, the initial 82 different construction activities were collated and eventually reduced to 10 unique variables. This was achieved by grouping similar construction tasks into singular variables. The pivotal decisions made in this step have been elaborated in the following section; and also summarised in Table 11.5. However- the exact compositions of each input variable has been included in Appendix B for reference.

Two types of probe drilling The construction task "probe drilling" involves exploratory drilling into the rock mass ahead of the tunnel face. This process is performed systematically, or as the ground conditions dictate. At its completion, the process can either transition into pre-grouting injection (X_3); or conclude with a plugging of the hole. Naturally, probe drilling with this plugging element will consume more time, than those without. A distinction between the two scenarios in the BoQ will presumably result in a higher performing model.

In the case of the Svartås-tunnel BoQ however, there is no clear-cut data to describe the plugging installation. However, this plugging feature can be inferred by examining the pre-grouting injection data immediately following the probe drilling data point. New input variables can then be created using this assumption. This description of the two probe drilling input variables has been summarised in Table 11.4.

Table 11.4: Description of the two probe drilling input variables

Notation	Description	Distinction
X_1	Probe drilling **with** plugging element	No injection is performed immediately following probe drilling
X_2	Probe drilling **without** plugging element	Injection is immediately performed following probe drilling

Effective workweeks In this experiment, only ten input variables are considered in the model. In reality however- other "time consumers" do exist. These are typically manifested in the form of lost-time (as described in Section 4.5.2). The magnitude of its occurrence - if unaccounted for - will greatly impede the model's overall performance. Therefore - when detectable - any non-construction tasks (such as owner's half over, manual cleaning, or curing times) have been deducted directly from the initial 101 hour workweek. Mathematically, this is expressed as:

$$Y' = Y_0 - Y_N \tag{11.2}$$

where:

Y' = effective workweek [h]

Y = initial workweek [h]

Y_N = non-construction time, at the tunnel face [h]

11.1.3 Limitations to this methodology

Feature engineering is limited by experienced: The data clean up and feature engineering processes, to the untrained, is a risky endeavour. The decisions made in this step are oftentimes highly subjective (especially when there is no direct lines of communication with the data's origins). Overall, the effectiveness of this step is governed by the practitioners knowledge of; and experience within the D&B tunnelling industry (as described *Substantive Expertise*, in Section 5.1). This step of the data science work-flow was therefore conducted under close guidance by a Norwegian tunnelling expert (Professor Amund Bruland).

Distinctions between the two probe drilling types are not perfect: BoQs are simply quantity values after all, and the associated time dimension is an assumption only. For example, in reality, injection may not actually occur immediately following probe drilling activities. Should the injection occur instead on the following workweek, this may result in a misdiagnosis of the true X_1 and X_2 values. Nonetheless, this limitation is a failing of the data quality (which can be improved), and not of the model itself. The concern of this exercise is to model both types of probe drilling (in accordance with standard NoTCoS practises).

Data import: In hindsight, it was admittedly futile to expend such large efforts on the data import process. The BoQ, as a dataset, was not "complete" in its original form anyway. Therefore, high data input precision is trivial in the grand scheme of things. At this stage of the research, the focus should have been on the modelling techniques and not on reproducing exact figures.

The actual time spent at the tunnel face is unknown: The truth is, the 101 weekly work-hours will not be spent entirely at the tunnel face. Occasionally, the allocation of the resources (equipment and workers) is delegated to tasks away from the tunnel face, for example connector tunnels, and the construction of niches. This is not easily (or accurately) detectable from purely bill of qualities. This indicates that the original assumption that the construction time per week is a constant 101 hours may be incorrect. This oversight however, can be rectified by analysing the magnitude of "unaccounted for time" (to be discussed in Hypothesis 2 - Section 14.1.2); and by proposing realistic changes to the data collection process. This will be explored further in the discussions chapters.

11.1.4 Summary of the final input variables

The composition of each input variable has been summarised in Table 11.5, but a more-detailed version also been included in the Appendix for reference (Appendix B). All in all, 11 input variables were selected for the primary

Table 11.5: Svartås-tunnel input variables

Predictor variables			
Notation	Description	unit	Composition
X_1	Probe drilling (without plugging)	m	Total drill-length of probe drilling (without plugging)
X_2	Probe drilling (with plugging)	m	Total drill-length of probe drilling (plugging)
X_3	Injection and control holes	m	Total drill-length of injection and controls (all lengths, includes flushing)
X_4	Pre-grouting (injection)	t	Total amount of injection (all types, includes the plugging element)
X_5	Excavated material	m^3	Total volume of material blasted, and removed (includes drilling and charging, full profile only)
X_6	Rockbolts ≤ 4m	unit	Total number of rock bolts ≤ 4 m installed (all types, conducted at the face)
X_7	Rockbolts > 4 m	unit	Total number of rock bolts > 4m installed, conducted at the face
X_8	Straps	m	Total number of rock straps installed (all types)
X_9	Shotcrete	m^3	Total volume of shotcrete applied (all types)
X_{10}	Reinforcements and arches	m	Total length of reinforcements and arches (all types)

Response variables			
Notation	Description	unit	Composition
Y_N	Non-construction tasks (performed at the tunnel face)	h	Known and quantifiable time consumers (e.g.: owner's half hour, or cleaning duration)
Y_0	Initial workweek	h	The typical total workweek contains 12 shifts (101 hours)
Y'	Effective workweek	h	Non-construction tasks are deducted from the initial workweek ($Y_0 - Y_N$)

dataset for modelling. These include:

- **One (1) response variable:** which is defined by the effective weekly work-hours spent at the tunnel face.
- **Ten (10) predictor variables:** which is defined by the quantity of major construction tasks performed at the tunnel face.

The main objective of the input variable selection process was to ensure that all major time consumers are accounted for.

11.2 Data modelling

In the data modelling stage, the Svartås-tunnel dataset is fed through various predictive algorithms. Thereafter, results are verified and assessed using traditional model validation techniques. Due to time constraints, and my limited skill-set, this report has focused only on a select-few algorithms (from each of the three major fields in data science). The intention is to develop an overall understanding of the dataset and its characteristics; as well as establish a starting point for the primary analysis. The model themes investigated in this chapter include:

- Machine learning: classification (briefly);
- Statistics: regression analysis; and then
- Mathematics: constrained optimisation.

11.2.1 Classification

As briefly mentioned in the Theoretical Background chapters - Section 6.3, the truth of the matter is that classification-themed algorithms are not entirely practical nor compatible with NoTCoS's agendas. For starters, classification methods are designed for "categorical" variables. However, in the case of the ETS, their time capacity values are independent of changes in geology. As such, their input variables are strictly quantitative, and inappropriate for classification analysis.

11.2.2 Regression analysis

The first algorithms tested were regression-themed. Crudely speaking, the process was a systematic "trial and error", and a process of elimination. The techniques tested included traditional ordinary least squares (OLS), and regression-through-the-origin (RTO) model: identified as Series SA-O and SA-R respectively. As a starting point, the effective workweek (Y') was first assigned to be the response variable: while the remaining construction tasks, $X_1 \dots X_{10}$, set as the predictor variables. Subsequent iterations thereafter either varied in the variables included; and/or the rearrangement of the equation (to manage the effects of multicollinearity). The results and discussions are presented in the next section. However, the model iterations are summarised in Table 11.6. In the first row, the

Table 11.6: Regression models: Svartås-tunnel

Regression model input variables											
Series ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA01	Y	X	X	X	X	X	X	X	X	X	X
SA02	Y	X	X	X	X	X	X	X	X	EX	X
SA03	Y	X	X	X	X	X	X	X	NFF	EX	X
SA04	Y	X	X	X	X	X	X	NFF	NFF	EX	X
SA05	Y	X	X	X	X	X	X	NFF	NFF	EX	NFF
SA06	X	X	X	X	Y	X	X	X	X	X	X

11 input parameters are shown. With each iteration, some variables are either rearranged; excluded; or converted to a real time unit. For example, in the model series SA02, the effective workweek (Y'), was designated as the response (denoted as Y); the strapping component (X_8) is converted to a real hour time, and then deducted from the response variable; and lastly, the shotcrete component (X_9) is excluded from the dataset.

- Y: Indicates that the variable is set as the response.
- X: Indicates that the variable is set as a predictor.
- EX: The construction task is excluded from the dataset (assumed to be performed outside of the standard workweek hours).
- NFF: Quantities are divided by an average rate (NFF values) and deducted from the response variable.

11.2.3 Ordinary least squares (OLS)

To kick things off, the first OLS model iteration included all variables in the dataset. Results were however immediately disappointing. Firstly, as shown in Table 11.7, estimated beta coefficients were a combination of both negative and positive numbers. Such a performance rate is unrealistic, and quickly raised alarm-bells. Secondly, the intercept was almost identical to the average weekly work-hours. Nonetheless, additional iterations were created for further analysis. Troubleshooting involved the incremental removal of variables, one at a time: in order to observe the effects at each iteration. The first variable excluded from the model was the shotcreting component (X_9). The reasoning behind this decision was to test the hypothesis that shotcreting, in practice, is occasionally performed outside of the standard working-hours. Thereafter, variables that did not contribute (on average) significantly towards the overall time consumption, and those presenting the smallest distribution in the dataset were progressively converted to a "time dimension" using NFF's (switch-tube) time capacity values (NFF 2019), and deducted directly from the effective weekly work-hours (Y'). These performance rates are detailed in Table 11.8.

Overall, the OLS model results were all similar in characteristics. As such, only partial results will be presented in this chapter when relevant. Though- comprehensive results have been included as part of the Appendix (B) for reference.

Table 11.7: Ordinary least squares summary of outputs: Svartås-tunnel data

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-O01	OLS	0.519	0.269	0.211	2.581	135
SA-O02	OLS	0.512	0.262	0.202	2.595	135
SA-O03	OLS	0.536	0.288	0.230	2.595	135
SA-O04	OLS	0.433	0.188	0.122	3.660	135
SA-O05	OLS	0.594	0.352	0.300	7.129	135
SA-O06	OLS	0.838	0.703	0.679	414.227	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-O01	-25.4005	-0.5050	0.1237	0.8028	3.4541
SA-O02	-25.5694	-0.3817	0.1018	0.7684	3.5403
SA-O03	-25.5462	-0.3939	0.1076	0.7828	3.5449
SA-O04	-25.8993	-0.4772	0.7035	1.8256	5.5031
SA-O05	-46.597	-1.283	1.410	4.006	8.868
SA-O06	-1180.58	-237.94	7.34	248.27	1023.34

Model outputs

	SA-O01	SA-O02	SA-O03	SA-O04	SA-O05	SA-O06
Intercept	100.11335	99.93141	100.00429	95.36436	84.50114	1 055.93013
X_1	-0.00325	-0.00306	-0.00316	0.00210	-0.00327	1.79673
X_2	-0.00940	-0.00866	-0.00887	-0.00349	0.00436	0.84923
X_3	-0.00213	-0.00205	-0.00207	-0.00128	0.00189	-0.17806
X_4	0.00975	0.00786	0.00785	0.02460	0.04990	-4.80309
X_5	-0.00003	-0.00034	-0.00035	0.00137	0.00597	-
X_6	0.01077	0.00866	0.00853	0.00367	-0.00108	1.15630
X_7	-0.00774	-0.02577	-0.04331	-	-	-32.28751
X_8	0.01406	0.02464	-	-	-	17.58596
X_9	-0.01213	-	-	-	-	8.95488
X_{10}	0.00549	0.00117	0.00137	0.01354	-	-7.15391
Y'	-	-	-	-	-	-0.70358

Table 11.8: Construction performance rates used to convert quantity unit to an hour unit (NFF 2019)

NFF time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m : h	m : h	m : h	t : h	m^3 : h	unit : h	unit : h	m : h	m^3 : h	m : h
Capacity	60.00	40.00	60.00	1.50	38.46	15.00	7.50	25.00	8.00	4.00
Unit time	0.017	0.025	0.017	0.667	0.026	0.067	0.133	0.040	0.125	0.250

Observations and inferences

Conventional ordinary least squares was unable to produce coefficients relevant to the project objective: Despite numerous iterations, OLS models were unable to produce coefficients relevant to the project objective. Because the dataset consisted of an almost-constant response variable (101-hours, as shown in Figure 11.3), regression with an intercept will interfere with the results. To clarify, the "best fit line" for an OLS model is produced by setting the intercept as the origin, while the other predictor variables are simply "centered" around this origin. Although, this principle will theoretically keep the residuals at a minimum (Figure 11.4), this does not automatically imply a "useful" model.

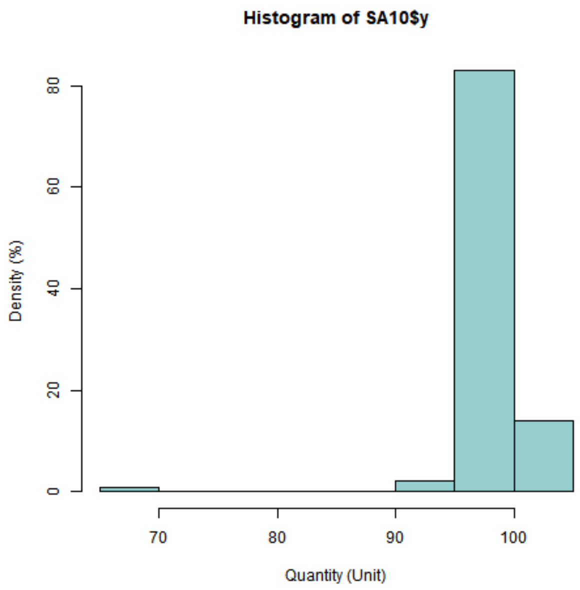


Figure 11.3: A histogram of the response variable of the Svartås-tunnel dataset

11.2.4 Regression-through-the-origin (RTO)

The removal of the intercept variable did not immediately produce relevant results. However, when outliers were removed from the equation, the estimates did improve, and appeared some-what relevant to the project objective. Comprehensive model outputs have been included in Appendix B for reference. However, snippets of this information will be presented in the following sections when relevant.

Some models produced mixed results both positive and negative Early iterations of RTO models were able to produce high performing regression models (according to conventional validation methods, such as analysis of variance, and goodness of fit, etc.). However, upon closer inspection, it was obvious that some coeffi-

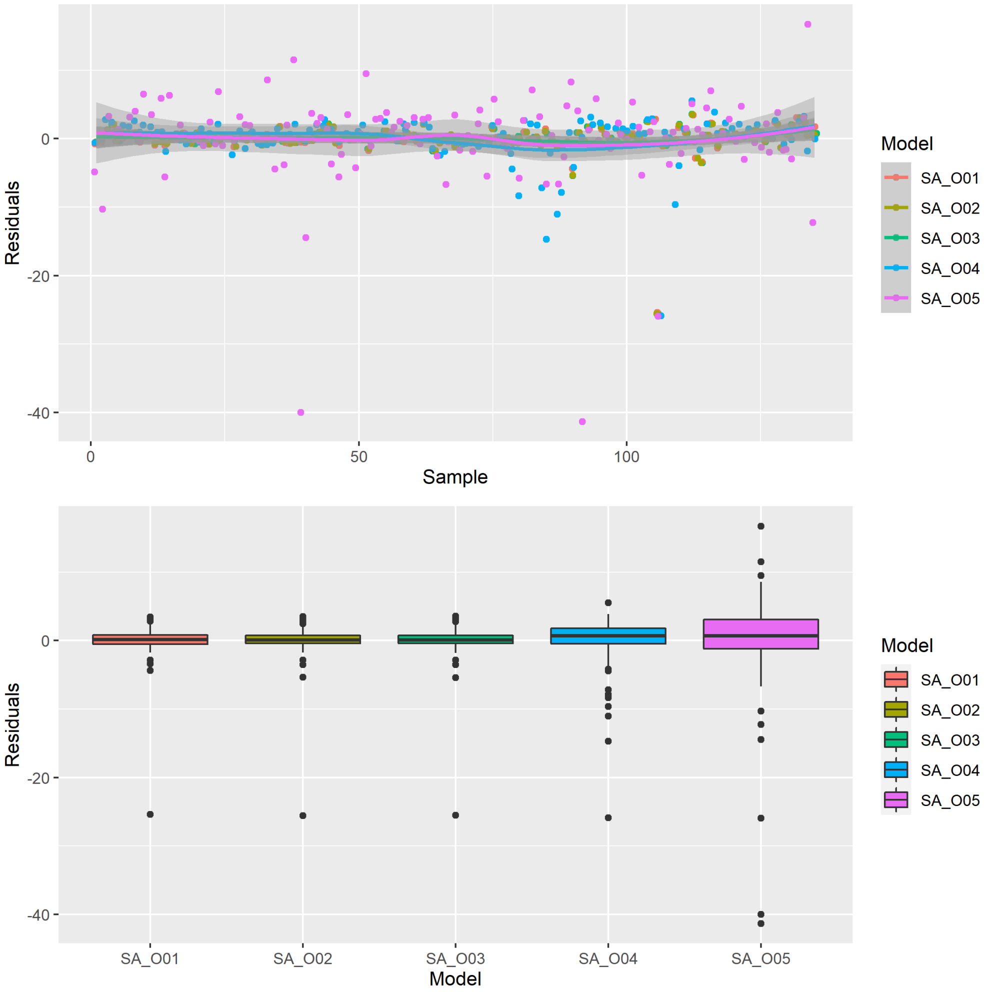


Figure 11.4: Residuals of the Svartås-tunnel ordinary least squares models

cients were negative. As shown in Table 11.9, such examples were observable in SA-R01 and SA-R02. In the case of drill and blast tunnelling construction data, such an adverse influence towards time (the response variable) is theoretically impossible. All construction activities (predictor variables) must contribute positively: thus, all coefficients must be uniform (either all positive, or all negative).

Observations and inferences

Removing outlier variables produced promising results: Regression-through-the-origin dramatically increased the statistical relevance of the model (in terms of the R-squared value). Though it was not until outliers were removed (in SA-R03, SA-R04, and SA-R05), that the results became all positive. This can be seen in Table 11.10.

The model becomes invalid when a construction task is set as the response variable: In model SA-R06, X_5 (excavated material quantity) was rearranged to become the response variable. The model and its

Table 11.9: Regression-through-the-origin: Svartås-tunnel models summary of outputs

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-R01	RTO	0.981	0.961	0.951	20.103	135
SA-R02	RTO	0.981	0.962	0.951	20.049	135
SA-R03	RTO	0.980	0.959	0.949	20.582	135
SA-R04	RTO	0.978	0.957	0.946	21.108	135
SA-R05	RTO	0.979	0.958	0.947	20.622	135
SA-R06	RTO	0.976	0.953	0.941	413.446	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-R01	-37.278	-9.327	3.708	17.010	51.569
SA-R02	-35.918	-10.298	3.826	17.426	49.749
SA-R03	-39.591	-10.068	4.206	17.931	49.292
SA-R04	-40.923	-9.735	4.434	17.692	56.916
SA-R05	-40.907	-9.725	4.453	17.684	56.918
SA-R06	-1181.560	-243.290	1.160	251.600	1025.110

Model outputs

	SA-RO1	SA-RO2	SA-RO3	SA-RO4	SA-RO5 %	SA-R06
Intercept	-	-	-	-	-	-
X_1	0.044	0.043	0.037	0.029	0.029	1.846
X_2	0.100	0.095	0.083	0.081	0.081	0.970
X_3	0.021	0.020	0.020	0.021	0.021	-0.153
X_4	0.232	0.253	0.267	0.256	0.256	-4.887
X_5	0.023	0.026	0.027	0.026	0.026	-
X_6	0.182	0.205	0.205	0.237	0.237	1.079
X_7	1.638	1.832	0.484	-	-	-32.059
X_8	-1.051	-1.167	-	-	-	17.330
X_9	0.110	-	-	-	-	9.141
X_{10}	0.177	0.219	0.249	0.248	-	-7.212
Y'	-	-	-	-	-	9.712

outputs however, did not satisfy model validation nor project objectives. Key output results are presented in Table 11.11.

- Output results from the RTO model reverted back to a mixture of positive and negative coefficients once again. These estimates are meaningless when considering the project objective.

- Perhaps more importantly, a regression model which sets X_5 as response variable, violates the conditions for running an RTO. Specifically: when the predictor variables sum to zero (0), the average response variable must also sum to zero (0) (Kutner et al. 2005). This is however not true in this iteration, as well as any iteration where a construction element is set as the response.

Table 11.10: Regression-through-the-origin: SA-R03 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R03	Y	X	X	X	X	X	X	X	NFF	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-R03	RTO	0.980	0.959	0.949	20.582	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-R03	-39.591	-10.068	4.206	17.931	49.292

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1 254 167.408	125 416.741	296.059	1.776E-81
Residual	125	52 952.599	423.621		
Total	135	1 307 120.006			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	-	-	-	-	-	-
X_1	0.037	0.027	1.387	0.168	-0.016	0.089
X_2	0.083	0.030	2.796	0.006	0.024	0.142
X_3	0.020	0.004	5.531	0.000	0.013	0.028
X_4	0.267	0.098	2.732	0.007	0.074	0.461
X_5	0.027	0.003	8.692	0.000	0.021	0.033
X_6	0.205	0.042	4.872	0.000	0.122	0.288
X_7	0.484	0.117	4.135	0.000	0.252	0.716
X_8	-	-	-	-	-	-
X_9	-	-	-	-	-	-
X_{10}	0.249	0.057	4.380	0.000	0.137	0.362

Summary: Regression analysis

Following this preliminary testing, I am inclined to believe that regression modelling with conventional OLS (with an intercept) is unreliable: or bluntly-speaking, results are meaningless for the dataset type and project objective. The problem then provokes a revisit into the model's *interpretability*. The issues with a conventional OLS model stems from the fact that the response variable is a almost-constant value (101-hours). Regression with an intercept simply misinterprets this scenario, and then assumes that all future values will also be the same. To clarify, the "best fit line" for an OLS model is produced by setting the intercept as the origin, while the other predictor variables are simply "centered" around this origin. Although, this principle will theoretically keep the residuals at a minimum, this does not automatically imply a "useful" model.

Despite removing the intercept, initial RTO models did still produce mixed estimates. Beta estimates consisting of both positive and negative values may be indicative that the correlation between the predictor values

Table 11.11: Regression-through-the-origin: SA-R06 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R06	*X*	*X*	*X*	*X*	*X*	*Y*	*X*	*X*	*X*	*X*	*X*

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
SA-R06	RTO	0.976	0.953	0.941	413.446	135

ANOVA

	df	SS	MS	F	Significance F
Regression	10	428 489 565.689	42 848 956.569	250.670	3.334E-77
Residual	125	21 367 241.311	170 937.930		
Total	135	449 856 807.000			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	-	-	-	-	-
X_1	1.846	0.515	3.582	0.000	0.826	2.867
X_2	0.970	0.631	1.537	0.127	-0.279	2.218
X_3	-0.153	0.082	-1.863	0.065	-0.315	0.010
X_4	-4.887	1.992	-2.453	0.016	-8.829	-0.944
X_6	1.079	0.966	1.117	0.266	-0.832	2.990
X_7	-32.059	10.735	-2.986	0.003	-53.305	-10.813
X_8	17.330	9.205	1.883	0.062	-0.888	35.547
X_9	9.141	1.459	6.267	0.000	6.254	12.027
X_{10}	-7.212	1.208	-5.968	0.000	-9.604	-4.820
Y'	9.712	1.623	5.985	0.000	6.501	12.924

(multicollinearity) is greater than their effect on the response variable (the true signal). This is especially evident for construction tasks which are performed less frequently (right-skewed distribution): as their (limited) signal is more-easily masked by the existence of excessive noise.

However, when outlier variables were removed or rearranged, RTO models did manage to produce all positive beta coefficients. Although difficult to confirm, this may be indicative that the data is too "noisy" (too many unaccounted for variables), and that the data quality is not sufficient at this time. Further analysis using other methods and on new datasets will reveal more information.

Although such results are inline with the project objective, and promising to see, it must be taken with caution. Further studies must be trialed in the same manner, across other new datasets, in order to test the procedure's repeatability and reliability.

11.2.5 Mathematical optimisation

Following the shortcomings of the regression-themed methods trialed so far. The attention is now fixated towards mathematics: where the problem is instead reframed as an optimisation problem. In this scenario, the construction data is regarded as a linear system of equations, and such, is classified as an *overdetermined* system (Section 8.3). Under this new assumption, the dataset is now considered *complete* and its values *true*. To elaborate: the term "complete" implies that the response variable (the weekly work-hours) is influenced **only** by the predictors variables (the included major construction tasks): and not affected by any other external (unaccounted for) factors. The term "true" implies that there are no measurement errors, and that the data is representative of actuality. Although this scenario is not entirely realistic, the basis behind this decision is intended to mirror the principles of the NoTCoS. In this contract system, their time capacity values are *constant*, and are not dependent on the geological conditions at the tunnel face. Instead, estimates are intended to reflect the average rate of performance over the course of the entire construction duration.

Without sufficient constraints, an overdetermined system can have an infinite number of solutions. The number of constraints can be increased by increasing the number of variables in the system, or by introducing boundaries and conditions to the solution. In this case, a non-negative constraint is introduced to force the model to produce positive-only beta coefficients. In this report, the constrained optimisation algorithm: non-negative least squares (NNLS) is called upon to achieve this.

For comparative purposes, the same dataset configuration from the regression analysis was reused to perform NNLS modelling. These are presented in Table 11.12.

Table 11.12: Mathematical optimisation models: Svartås-tunnel

Mathematical optimisation input variables											
Series ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-N1	Y	X	X	X	X	X	X	X	X	X	X
SA-N2	Y	X	X	X	X	X	X	X	X	EX	X
SA-N3	Y	X	X	X	X	X	X	X	NFF	EX	X
SA-N4	Y	X	X	X	X	X	X	NFF	NFF	EX	X
SA-N5	Y	X	X	X	X	X	X	NFF	NFF	EX	NFF
SA-N6	X	X	X	X	Y	X	X	X	X	X	X

Mathematical optimisation results

Presented below, in Table 11.13, are the results from the constrained optimisation modelling. Following this, the results will be discussed.

Table 11.13: Non-negative least squares: Svartås-tunnel models summary of outputs

Dataset description

Model ID	Algorithm	No intercept	Response	Observations
		R-square	Y	n
SA-N1	NNLS	0.960	Y'	135
SA-N2	NNLS	0.959	Y'	135
SA-N3	NNLS	0.959	Y'	135
SA-N4	NNLS	0.956	Y'	135
SA-N5	NNLS	0.955	Y'	135
SA-N6	NNLS	0.923	X_5	135

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
SA-N1	25.62	10.89	48.30	4.22	45.01	5.71	2.26	0	6.83	5.33
SA-N2	27.03	11.94	49.23	3.75	37.58	4.88	1.89	0	-	4.03
SA-N3	27.18	11.99	49.23	3.74	37.56	4.88	2.07	-	-	4.01
SA-N4	34.56	12.37	47.91	3.90	39.17	4.21	-	-	-	4.04
SA-N5	34.67	12.37	47.89	3.90	39.14	4.22	-	-	-	-
SA-N6	0.28	0.53	0	0	Y: 0.24	0.29	0	0	0.15	0
NFF	60.00	-	60.00	1.50	38.46	15.00	7.50	25.00	8.00	4.00

Observations and inferences

Immediately, looking at initial results, we can see that the model was unable to produce a beta estimate for the straps component X_8. Instead, a zero (0) sits in its place; and indicates that the model has opted to simply discount this input variable from the model completely. This setback however, interestingly coincides with traditional regression results. As shown in Table 11.14, for the variables where a RTO model could only produce a zero estimate, an OLS model would only produce negative values.

Table 11.14: Comparison of model results between RTO and NNLS: Svartås-tunnel

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA1	Y	X	X	X	X	X	X	X	X	X	X
SA2	Y	X	X	X	X	X	X	X	X	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-R01	RTO	0.981	0.961	0.951	20.103	135
SA-N1	NNLS	-	-	0.960	-	135
SA-R02	RTO	0.981	0.962	0.951	20.049	135
SA-N2	NNLS	-	-	0.959	-	135

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
SA-R01	22.80	10.02	48.41	4.30	43.60	5.48	0.61	-0.95	9.10	5.66
SA-N1	25.62	10.89	48.30	4.22	45.01	5.71	2.26	0	6.83	5.33
SA-R02	23.30	10.55	49.09	3.94	38.18	4.88	0.55	-0.86	-	4.56
SA-N2	27.03	11.94	49.23	3.75	37.58	4.88	1.89	0	-	4.03
NFF	60.00	-	60.00	1.50	38.46	15.00	7.50	25.00	8.00	4.00

11.3 Comparison to NFF's time capacity values

NFF's time capacity values have been largely successful and well-accepted in today's D&B tunnelling industry. In this section, their estimated performance rates are applied to the Svartås-tunnel dataset for comparative purposes, and for general insight. Note: the Svartås-tunnel was constructed using a "switch-tube" methodology, therefore a 1.3 factor was applied to the excavation rate, X_5 (marked with an asterisk). These values are summarised in Table 11.15.

Table 11.15: Industry standard time capacity values for switch-tube D&B tunnelling (NFF 2019)

NFF (switch-tube) time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m : h	m : h	m : h	t : h	m^3 : h	unit : h	unit : h	m : h	m^3 : h	m : h
Capacity	60.00	40.00	60.00	1.50	38.46*	15.00	7.50	25.00	8.00	4.00
Unit time	0.017	0.025	0.017	0.667	0.026*	0.067	0.133	0.040	0.125	0.250

When NFF's time capacity values were applied to the Svartås-tunnel, the model was only able to account for approximately 83% of the weekly-time spent (shown in Table 11.16). As such, the NFF model under-predicted the weekly work-hours by, on average, between 10 and 15 hours. In Figure 11.5, the NFF model and the NNLS model SA-N3 have been plotted to further illustrate this contrast.

Table 11.16: Comparison of model results between NFF and NNLS: Svartås-tunnel

Model input variables										
Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-N3	Y	X	X	X	X	X	X	X	NFF	EX	X

Dataset description				
Model ID	Algorithm	No intercept	Response	Observations
		R-square	Y	n
SA-N3	NNLS	0.959	Y'	135
SA-NFF	NFF	0.828	Y'	135

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
SA-N3	27.18	11.99	49.23	3.74	37.56	4.88	2.07	-	-	4.01
NFF	60.00	-	60.00	1.50	38.46	15.00	7.50	25.00	8.00	4.00

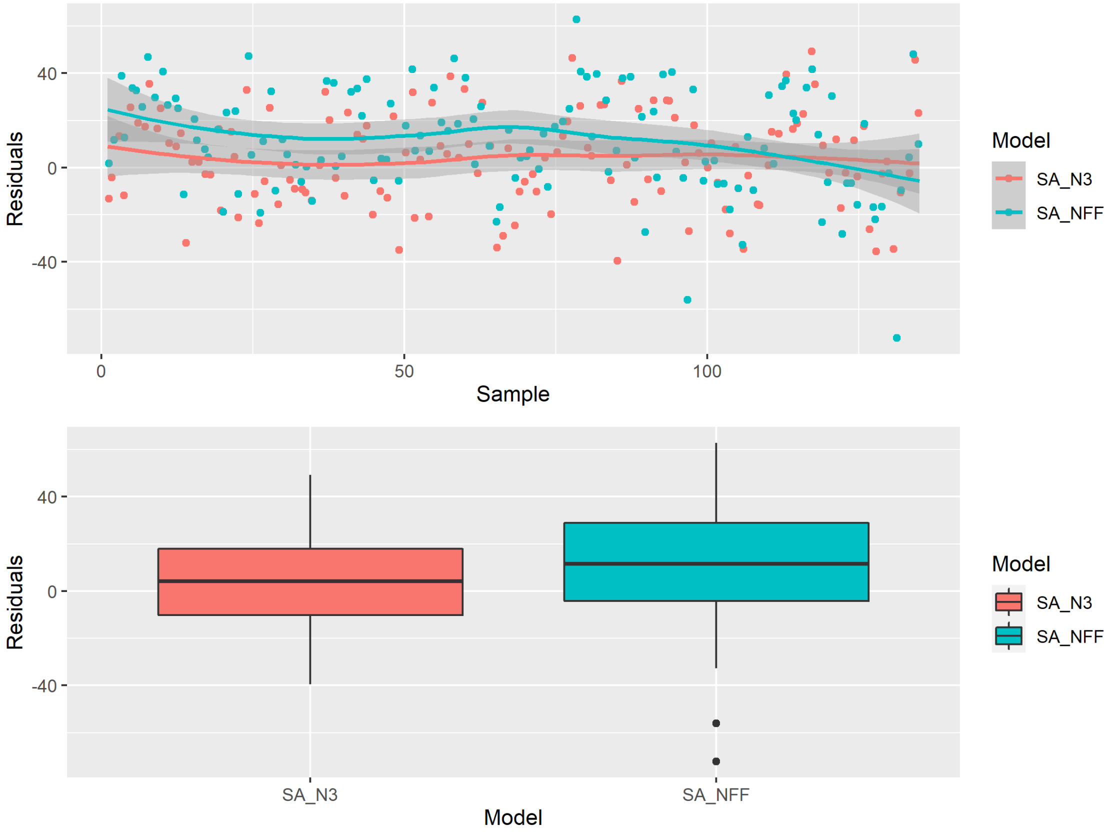


Figure 11.5: Comparison between NFF and NNLS model estimates on the Svartås-tunnel dataset. Plot A (top), is a scatter plot of the residuals verses the sample. Plot B (bottom) is a box-plot of the residuals.

11.4 Chapter summary

Concluding impressions about the NNLS least squares algorithm - at the end of the day - were rather mixed. In fact, the findings from this "opening" analysis, not only left many questions unanswered, but also stimulated *new* questions. Estimated time capacity values were realistic - but only for some variables. For the other "less-frequent" construction tasks- performance rates were *consistently* underestimated. A comparison between the NNLS models and the industry standards can be visualised in Figure 11.6. Furthermore, for variables with a more extreme right-tail distributions (low sample size), NNLS was unable to produce an estimate at all (zero). Despite this draw back, the algorithm was able to improve the estimates of the other variables, by omitting such outlier observations.

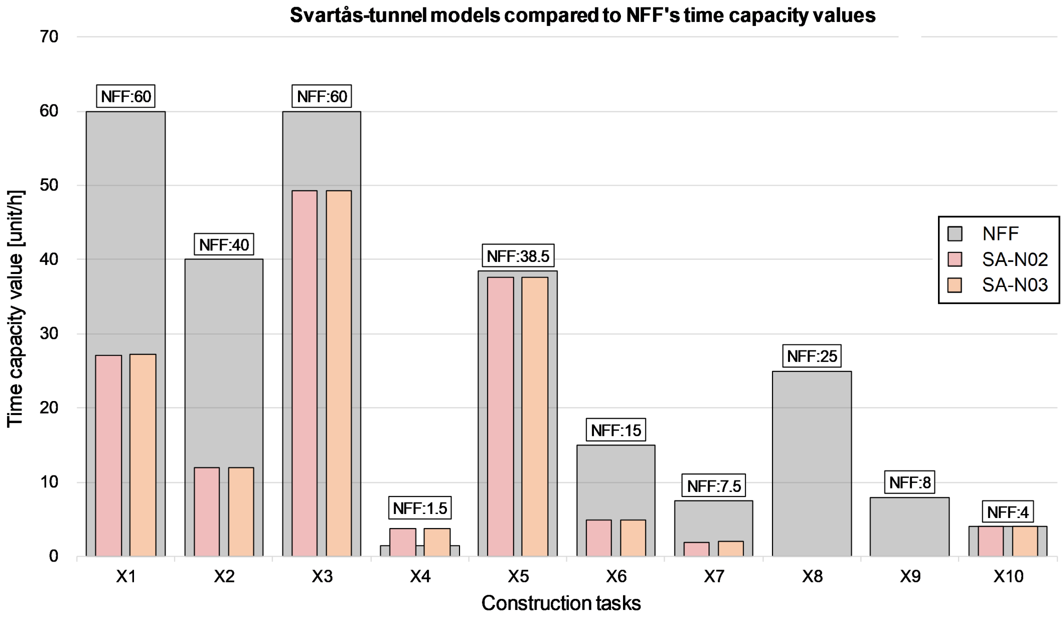


Figure 11.6: A comparison between the Svartås-tunnel models and NFF's time capacity values.

Perhaps more interestingly, the well-accepted NFF time capacity values, when tested against the Svartås-tunnel dataset, were only able to account for a little over 80% of the weekly work-hours (shown in Figure 11.5). This observation may reveal the presence of "incomplete data", or possibly expose the insufficient quality of the currently-available data. Nonetheless, further studies on additional tunnel datasets will hopefully uncover more.

Chapter 12

Data Analysis: The Kongsberg-tunnel

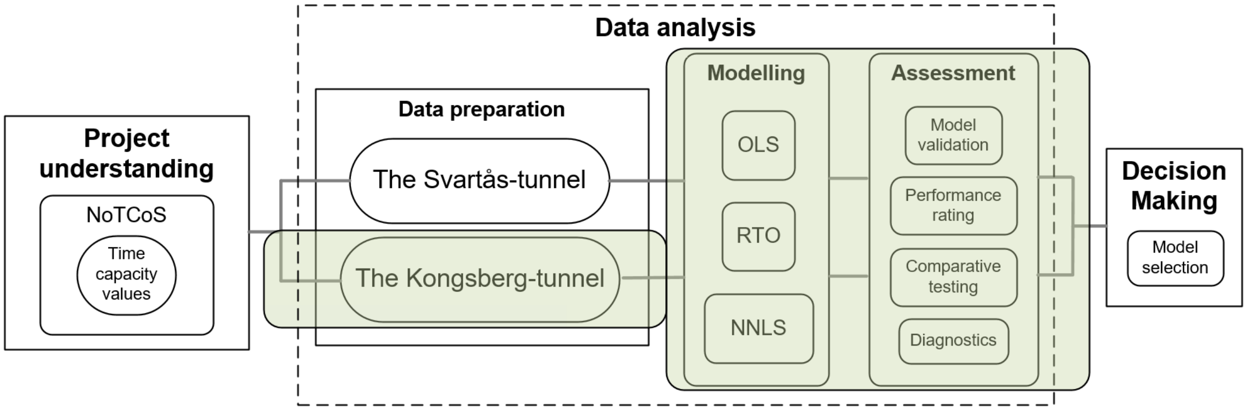


Figure 12.1: A progression chart of the data science workflow - data analysis: the Kongsberg-tunnel

Once again, the data science work-flow is applied to a *real* tunnel project. However this time, the subject-data is the Kongsberg-tunnel bill of quantities (BoQ). The analysis of a *second* tunnel and dataset, functions as a form of external model validation. The idea of this exercise is to verify that results derived from the Svartås-tunnel dataset were not coincidental. Addition to this, investigations of a second dataset may be useful for revealing unidentified characteristics of the currently-available construction data.

A cutback workflow

When compared to that of the initial Svartås-tunnel analysis, several preparatory stages of the data science work-flow were made redundant or notably trimmed. Early in this analysis, traditional regression quickly exhibited the same flaws (as it had with the Svartås-tunnel). That is why, this chapter perhaps exposes a stark pivotal moment: when the focus (or maybe bias) began to shift towards the NNLS algorithm. Hence, to reduce clutter, elements of the model selection process were omitted. Furthermore, the Svartås-tunnel and Kongsberg-tunnel dataset possessed many similarities between one another. As such, duplicate steps have not been documented in detail. The overall stages conducted in this chapter have been summarised in Table 12.1, and progression chart of the process is shown in Figure 12.1.

Table 12.1: Data science work stages required for the Kongsberg-tunnel data

Category	**Phases**	**Action**
Project understanding	Objective definition	Unchanged
	Constraints and assumptions diagnosis	Unchanged
Data preparation and exploration	Data collection	Unchanged
	Data wrangling	**Required**
	Exploratory Data Analysis (EDA)	**Required**
Modelling	Model selection	Unchanged
	Modelling	**Required**
	Model validation	**Required**
Communication	Inferences	**Required**
	Real-world application	Unchanged

A dataset of higher quality

When compared to the Svartås-tunnel, the Kongsberg-tunnel construction data is definitively of higher quality and magnitude. This allowed for other *original* forms of feature engineering and analysis. Table 12.2 is a summary of the contrasting characteristics unique to the Kongsberg-tunnel dataset; and their effects to the analysis. The contributions of these additional features will be elaborated further in the data preparation section (Section 12.1).

Table 12.2: Noteable features of the Kongsberg-tunnel bill of quantities

Feature	**Result**
"Plugging" quantities are recorded	X_1 now encompasses *all* probe drilling activities
Recorded quantities are day-specific	"Shortened" workweeks, Y_S are detectable
Secondary tunnel activities are included	The redirection of resources is verifiable
Distinction between blast round types (full-, half-, and two-part round)	Additional forms of X_5 trialed
Construction logs included billable "injection" hours	The *actual* rate of injection is used for comparative analysis and modelling

12.1 Data preparation and exploration

There is generally overlap in the data preparation process between similar projects. However, because the Kongsberg-tunnel differed in both the construction method, and in its data quality, the analysis required a renewed approach.

12.1.1 Data collection

The Kongsberg-tunnel construction data was collected in the same manner as the Svartås-tunnel. Additional documentation regarding its collection procedure is therefore unnecessary.

12.1.2 Data wrangling

The required steps in the data wrangling stage are as follows:

- Data description;
- Data import;
- Data clean up;
- Data transformation; and
- Feature engineering.

Data description

The Kongsberg-tunnel is a two-tube road tunnel, currently under construction, between Diseplass and Tislegård, in Kongsberg, Norway (Figure 12.2). With a T9.5 tunnel-profile (cross-sectional width: 9.5 m), the tunnel allows for two lanes of traffic from both directions, across the 2.2 km stretch. Excavation was performed using the single-tube method, and from two simultaneous faces: lines 11000 and 12000 (right and left respectively). All in all, the total excavation time spanned across a two-year period, and both requiring 57 unique construction tasks. A summary of the key figures are presented in Table 12.3.

Table 12.3: The Kongsberg-tunnel bill of quantities description

Raw data description			
Tunnel face	Excavation method	Observations *n*	Dimensions *k*
11000 line (right)	Single-tube	97 (weeks)	57 (unique tasks)
12000 line (left)	Single-tube	106 (weeks)	57 (unique tasks)
Total		**203**	**57**

Figure 12.2: The Kongsberg-tunnel under construction, March 2017 (Unlisted 2017)

Data import

As it was with the Svartås-tunnel, the Kongsberg-tunnel data was sourced from the same government-supplied tunnel databases. The Kongsberg-tunnel BoQs were also in .PDF format, and again required conversion to digital format prior to data analysis.

Data clean up

Similar to the Svartås-tunnel, outlier observations were excluded from the dataset and included in the modelling process. An "outlier" is identified as: a working-week where the sufficient portion of the time spent, was dedicated to *non-regular* construction tasks (i.e.: those not defined as an input variable $X_1 \ldots X_{10}$). Some examples included:

- Mobilisation, construction start-up, and pack-up phases; and
- Construction of secondary tunnel elements:
    - Connection tunnels
    - Niches (for emergency kiosks)
    - Excavation for technical structures.

Feature engineering

To limit the effects of multicollinearity and the curse of dimensionality, the 57 different construction activities were collated and eventually lumped into nine distinct variables. The composition of each input variable has been summarised in Table 12.4, but a more-detailed version also been included in Appendix B for reference.

Table 12.4: Kongsberg-tunnel input variables

Predictor variables			
Notation	Description	unit	Composition
X_1	Probe drilling (all)	m	Total drill-length of probe drilling (all lengths)
X_2	Probe drilling (with plugging)	m	No longer required
X_3	Injection and control holes	m	Total drill-length of injection and controls (all lengths, includes flushing)
X_4	Pre-grouting (injection)	t	Total amount of injection (all types, includes the plugging element)
X_5	Excavated material	m^3	Total volume of material blasted, and removed (includes drilling and charging, full profile only)
X_6	Rockbolts ≤ 4m	unit	Total number of rock bolts ≤ 4 m installed (all types, conducted at the face)
X_7	Rockbolts > 4 m	unit	Total number of rock bolts > 4m installed, conducted at the face
X_8	Straps	m	Total number of rock straps installed (all types)
X_9	Shotcrete	m^3	Total volume of shotcrete applied (all types)
X_{10}	Reinforcements and arches	m	Total length of reinforcements and arches (all types)

Response variables			
Notation	Description	unit	Composition
Y_N	Non-construction tasks (performed at the tunnel face)	h	Known and quantifiable time consumers (e.g.: owner's half hour, or cleaning duration)
Y_A	Time away from tunnel face	h	Public holidays; or a redirection of resources to other parts of the tunnel
Y_0	Initial workweek	h	The typical total workweek contains 12 shifts (101 hours)
Y'	Effective workweek	h	Non-construction tasks are deducted from the initial workweek ($Y_0 - Y_N$)
Y'_S	Shortened workweek	h	Time away from the tunnel face is deducted from the effective workweek ($Y' - Y_A$)
Y_I	Registered injection time	h	Injection time, as recorded in the BoQ

Only one plugging variable, X_1: As Kongsberg-tunnel's BoQ included a separate section for individual plugging quantities, [1] there is no longer a need to establish two separate plugging variables (both, X_1 and X_2). Unlike the Svartås-tunnel project, the "probe drilling with plugging" component (X_2) becomes obsolete, and X_1

[1] BoQ reference category: Støp 31.121 (kg)

instead encompasses *all* probe drilling activities.

The plugging component now together with injection, X_4: At the same time, this additional plugging quantity - comparable to typical injection procedures - is lumped together with other injection components, when forming X_4.

Shortened workweeks (Y_S) are discernible: "Shortened" workweeks can be derived by deducting the time away from the tunnel from the effective workweek. Mathematically, this is expressed as:

$$Y_S = Y' - Y_A \tag{12.1}$$

where:

Y_S = effective shortened workweek
Y' = effective workweek
Y_A = amount of time away from the tunnel face

Explanation: Construction records for the Kongsberg-tunnel were kept on a day-to-day basis. Therefore - under the assumption that these records can also convey a "time dimension" - it is then possible to detect "shortened" workweeks (Y_S) within the data. To clarify: it is reasonable to assume that quantities of a given task, recorded on a specific day, also imply that these task were also performed on the very same day. It is in turn, equally plausible that the *blank* days (the days left unlogged) may in fact signify the time spent *away* from the tunnel face. If there is indeed a degree of truth to this, the dataset's effective weekly work-hours (Y) should reflect this reality.

Of course, it would be careless to blindly mark any *blank* day as *non-working* time. Therefore, any decisions to label a workweek as "shortened" must be made in conjunction with support evidence. This supporting evidence can be found by crosschecking *blank* days with known causes that may deter resources from the tunnel face. Sources of this may include coinciding Norwegian public holidays, or occasionally, the relocation of resources to other parts of the tunnel. Although difficult to confirm, the latter trend is observable by comparing BoQs at the tunnel face and those of secondary tunnels (such as connection tunnels, and technical rooms).

Overall, this methodology is not foolproof. Shortened workweeks are admittedly difficult to confirm without direct affirmation from the tunnel builders. Nonetheless, this approach is only intended to improve the truthfulness of the data; and to ultimately provide a more accurate gauge of the models performance.

12.2 Data modelling

The ready-dataset was once again tested using traditional regression, and thereafter constrained optimisation.

12.2.1 Regression analysis

As it was with the Svartås-tunnel, conventional regression was the first algorithm tested. However, immediate results quickly resembled the deficiencies experienced in previous regression modelling attempts. Therefore, to reduce repetition, only a diminished subset of regression models are included in this report.

In iterations 1 and 2, the dataset (comprised of both lines) was modeled using a traditional ordinary least squares (OLS) method. Thereafter, in iterations 3 and 4, regression-through-the-origin (RTO) was performed. Addition to this, distinctions between the effective (Y') and shortened workweek (Y_S) was also made within these iterations. The model iterations are summarised in Table 12.5.

Table 12.5: Regression models: Kongsberg-tunnel

Dataset description					
Model ID	Algorithm	Observations	Response [h]	Predictors	Dimensions
		n	Y	X	k
KB-O01	OLS	172	Effective	$X_1, X_3 \dots X_{10}$	9
KB-O02	OLS	172	Shortened week	$X_1, X_3 \dots X_{10}$	9
KB-R03	RTO	172	Effective	$X_1, X_3 \dots X_{10}$	9
KB-R04	RTO	172	Shortened week	$X_1, X_3 \dots X_{10}$	9

Regression model input variables											
Series ID	Y'	Y_S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
RB-O01	Y	-	X	-	X	X	X	X	X	X	X	X
RB-O02	-	Y	X	-	X	X	X	X	X	X	X	X
RB-R03	Y	-	X	-	X	X	X	X	X	X	X	X
RB-R04	-	Y	X	-	X	X	X	X	X	X	X	X

- Models contain data from both lines 11000 and 12000.
- *OLS*: Ordinary least squares regression (standard model).
- *RTO*: MLR, Regression-Through-the-Origin (no intercept model).
- Y': Using the effective workweek as the response variable.
- Y_S: Using the effective shortened workweek as the response variable.

Table 12.6: Regression analysis: OLS models KB0 dataset description and model outputs

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
RB-O01	OLS	0.600	0.360	0.320	9.581	172
RB-O02	OLS	0.424	0.179	0.128	1.249	172

Regression model input variables

Model ID	Y'	Y_S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
RB-O01	Y	-	X	-	X	X	X	X	X	X	X	X
RB-O02	-	Y	X	-	X	X	X	X	X	X	X	X

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
KB-O01	-5.2060	-0.2906	0.2700	0.6547	2.4338
KB-O02	-37.838	-3.556	0.759	6.063	20.198

Model outputs: time capacity values

	β_0	X_1	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-O01	100.1152	-0.0021	-0.0005	0.0004	0.0002	-0.0080	0.0133	-0.0230	-0.0121	0.0393
KB-O02	65.805	0.010	0.007	0.187	0.008	-0.025	-0.016	0.430	0.064	0.445

Results

Observations and inferences

Regression analysis initially resembled that of the Svartås-tunnel. The ordinary least squares (OLS) method was once again unable to derive usable beta coefficients. However, when the intercept was removed, the regression-through-the-origin (RTO) models were much more promising.

At first, the RTO model produced a mix between positive and negative values (model KB-R03). However, when the shortened workweek was set as the response, the model instead produced positive-only results. A surprising result indeed, as positive-only results were never achieved using convention regression in the Svartås-tunnel. This will be investigated further later on.

Table 12.7: Regression analysis: RTO models KB0 dataset description and model outputs

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
RB-R03	RTO	0.976	0.953	0.945	21.891	172
RB-R03	RTO	0.984	0.968	0.960	17.269	172

Regression model input variables

Model ID	Y'	Y_S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
RB-R03	Y	-	X	-	X	X	X	X	X	X	X	X
RB-R04	-	Y	X	-	X	X	X	X	X	X	X	X

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
KB-R03	-52.825	-8.696	5.233	16.082	77.514
KB-R04	-50.493	-8.034	2.123	13.292	60.423

Model outputs: time capacity values

	X_1 m/h	X_2 m/h	X_3 t/h	X_4 m^3/h	X_5 unit/h	X_6 unit/h	X_7 m/h	X_8 m^3/h	X_9 m/h	X_{10}
KB-R03	26.01	-	78.74	1.55	68.86	6.24	1.65	2.64	3.12	-3.06
KB-R04	27.43	-	63.85	1.63	56.65	11.72	2.69	1.44	3.53	4.90
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

12.2.2 Mathematical optimisation

Five different iterations were created to test the effectiveness of the NNLS algorithm, with each iteration containing two models. The first models assigned the effective workweek (Y') to be the response variable. In the second models, the shortened workweek (Y_S) was set as the response variable. The series have been summarised in Table 12.8, with detailed descriptions immediately following.

Table 12.8: NNLS models: Kongsberg-tunnel data

Model description					
Series	NNLS	NNLS	Data source	Feature Eng.	Predictors
	Y'	Y_S	(line)		X
KB-N1	KB-N11	KB-N12	11000	Standard	$X_1, X_3 \ldots X_{10}$
KB-N2	KB-N21	KB-N22	12000	Standard	$X_1, X_3 \ldots X_{10}$
KB-N3	KB-N31	KB-N32	Both	Standard	$X_1, X_3 \ldots X_{10}$
KB-N4	KB-N41	KB-N42	Both	X_4 rearranged	$X_1, X_3, X_5 \ldots X_{10}$
KB-N5	KB-N51	KB-N52	Both	$X_1 + X_3 = X_{1+3}$	$X_{1+3} \ldots X_{10}$

- **KB-N1:** Models contain data from the 11000 (right) line only.
- **KB-N2:** Models contain data from the 12000 (left) line only.
- **KB-N3:** Models contain data from both lines.
- **KB-N4:** Raw injection times (as recorded in the BoQ) are deducted directly from the effective weekly working hours (Y' - Y_I). The injection quantity (X_4) is then no longer a predictor variable in this iteration.
- **KB-N5:** All variables comprised predominately of drilling activities are combined together into one variable: X_{1+3}.
- NNLS: Non-negative least squares algorithm
- Y': Using the effective workweek as the response variable.
- Y_S: Using the effective shortened workweek as the response variable.

Observations and inferences

The model results, together with the observations and inferences, are discussed in the following sections.

NNLS: KB1 and KB2 "separate lines"

The models that only included data from individual lines have been shown in Table 12.9 and 12.10. In most cases, the model *improved* for the shortened workweek models.

> Note: a model is interpreted as being *improved*, when estimates diverge closer to NFF's time capacity values.

Table 12.9: Kongsberg-tunnel KB-N1 dataset description and model outputs

Dataset description					
Model ID	Algorithm	Observations	Response [h]	Predictors	Dimensions
		n	Y	X	k
KB-N11	NNLS	86	Effective	$X_1, X_3 \dots X_{10}$	9
KB-N12	NNLS	86	Shortened week	$X_1, X_3 \dots X_{10}$	9

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N11	34.70	-	67.34	1.40	79.00	9.60	1.82	13.16	2.31	0.00
KB-N12	40.95	-	58.27	1.57	67.95	9.69	1.63	0.00	2.99	0.00
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 12.10: Kongsberg-tunnel KB-N2 dataset description and model outputs

Dataset description					
Model ID	Algorithm	Observations	Response [h]	Predictors	Dimensions
		n	Y	X	k
KB-N21	NNLS	86	Effective	$X_1, X_3 \dots X_{10}$	9
KB-N22	NNLS	86	Shortened week	$X_1, X_3 \dots X_{10}$	9

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N21	34.03	-	98.34	1.65	56.64	3.56	2.04	0.99	8.48	0.00
KB-N22	24.04	-	67.07	1.69	48.96	10.97	4.07	0.69	4.79	0.00
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

For example in KB-N12 and KB-N22, where the "control hole", X_3 estimate improved from 67.34 to 58.27 m/h, and 98.34 to 67.07 m/h. However, when comparing the two datasets, significant differences in their estimates are still observable. This is most notable in the "probe drilling" and "excavated material" variables X_1 and X_5, where the construction rate varied between 67.34 and 98.34 m/h; and 48.96 and 67.95 m^3 respectively. Overall, results from KB1 and KB2 are mixed. This is however not unexpected, granted that the sample size for each model is only 86 large.

NNLS: KB3 "all lines combined"

Table 12.11: Kongsberg-tunnel KB-N3 dataset description and model outputs

Dataset description

Model ID	Algorithm	Observations	Response [h]	Predictors	Dimensions
		n	Y	X	k
KB-N31	NNLS	172	Effective week	$X_1, X_3 \dots X_{10}$	9
KB-N32	NNLS	172	Shortened week	$X_1, X_3 \dots X_{10}$	9

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N31	26.26	-	79.50	1.54	68.10	6.26	1.74	2.51	3.15	0.00
KB-N32	27.43	-	63.85	1.63	56.65	11.72	2.69	1.44	3.53	4.90
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

In the KB-N3 series, models were created using data from both tunnel lines. The results from these tests are presented in Table 12.11. At a glance, the estimates immediately appear more realistic (and promising). This is especially true for KB-N32, where the shortened workweek is assigned. In almost every case, except "injection" (X_4) and perhaps "straps" (X_8), the model *improved*. Regarding the "injection" variable, X_4: the estimate only differed slightly, from 1.54 to 1.63 t/h. While the "straps" component, X_8, changed from 2.51 to 1.44 m/h. Regardless of its observed change, the "straps" predictions are well below the standard NFF rates of 25.00 m/h. This may be due to insufficient sample size, or the data distribution. This will be explored in the diagnostics sections (12.3).

NNLS: KB4 "injection times deducted"

In the fourth iteration, KB-N4, injection times specified in the BoQ, are deducted directly from the effective workweek (Y'), and the "injection" variable (X_4) is removed from the dataset. The results from these models are presented in Table 12.12. Following the exclusion of the "injection" variable (X_4), mixed results are observable. On one hand, most estimates remained unchanged (and comparable to KB-N3). On the other hand, several beta coefficients have diverged away from NFF standards. This is most evident in variables "control holes" (X_3) and (less-so) in "reinforcements and arches" (X_{10}). When comparing the estimates between KB-N32 and KB-N42 (as shown in

Table 12.12: Kongsberg-tunnel KB-N4 dataset description and model outputs

Dataset description

Model ID	Algorithm	Observations	Response [h]	Predictors	Dimensions
		n	Y	X	k
KB-N41	NNLS	172	Inj. $(X_4) \rightarrow Y$	$X_1, X_3, X_5 \dots X_{10}$	8
KB-N42	NNLS	172	Shortened week	$X_1, X_3, X_5 \dots X_{10}$	8

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N41	30.81	-	108.36	y	61.63	6.29	1.92	1.68	3.26	0.00
KB-N42	34.85	-	87.06	y	52.69	11.49	3.26	1.15	3.70	2.45
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 12.13), a change from 63.85 to 87.06 m/h, and (less significantly) from 4.90 to 2.45 m/h is observed in X_3 and X_{10} respectively.

Table 12.13: Comparison between KB-N32 and KB-N42 estimates

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N32	27.43	-	63.85	1.63	56.65	11.72	2.69	1.44	3.53	4.90
KB-N42	34.85	-	87.06	y	52.69	11.49	3.26	1.15	3.70	2.45
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

NNLS: KB5 "drilling components combined

In the final model series, KB-N5, all drilling-related activities are combined into one input variable (represented by X_{1+3}). This iteration was trialed to assess two major questions:

- Do all "drilling-related" tasks possess similar construction rates?
- Can the effects of multicollinearity be reduced, by collating similar tasks in the future?

Results from KB-N5 are presented in Table 12.14

12.3 Diagnostics

Following the analysis of two *real* D&B tunnels (The Svartås- and Kongsberg-tunnel), it was becoming apparent, that some construction tasks were consistently being underestimated, more than others. As illustrated in Figure 12.3, the issue may be related to the density distribution of the variables. This has been investigated in the primary analysis 14.5.

Table 12.14: Kongsberg-tunnel KB-N5 dataset description and model outputs

Dataset description

Model ID	Algorithm	Observations	Response [h]	Predictors	Dimensions
		n	Y	X	k
KB-N51	NNLS	172	Effective week	$X_{1+3} \ldots X_{10}$	8
KB-N52	NNLS	172	Shortened week	$X_{1+3} \ldots X_{10}$	8

Model outputs: time capacity values

	X_1	X_2	X_{1+3}	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N51	-	-	74.21	1.52	65.71	6.12	1.80	2.48	3.02	0.00
KB-N52	-	-	61.03	1.62	55.26	11.32	2.81	1.43	3.40	4.51
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

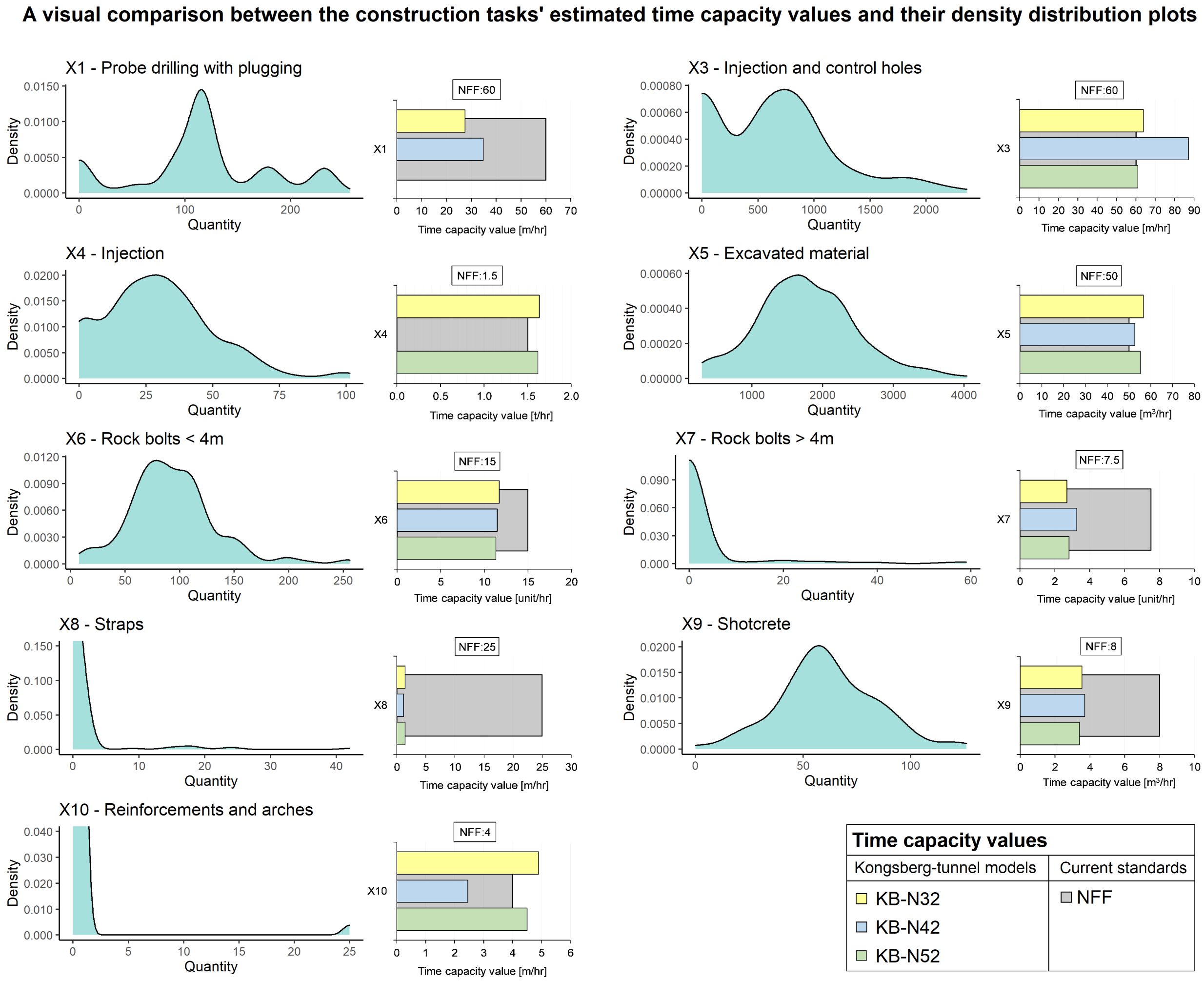


Figure 12.3: Density plot of the Kongsberg-tunnel dataset alongside the estimated time capacity values.

12.4 Comparison to NFF's time capacity values

In this section, the NFF time capacity values are applied to the Kongsberg-tunnel dataset for comparative purposes, and for general insight. Note: the Kongsberg-tunnel was constructed using a "single-tube" methodology, therefore no factor was applied to the excavation rate, X_5. These values are summarised in Table 12.15.

Table 12.15: Industry standard time capacity values for single-tube D&B tunnelling (NFF 2019)

NFF (single-tube) time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m : h	m : h	m : h	t : h	m^3 : h	unit : h	unit : h	m : h	m^3 : h	m : h
Capacity	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00
Unit time	0.017	0.025	0.017	0.667	0.020	0.067	0.133	0.040	0.125	0.250

When NFF's time capacity values were applied to the Kongsberg-tunnel, the model was only able to account for between 72 and 80% of the weekly-time spent (as detailed by the No-intercept R-square value, Table 12.16). As such, the NFF model under-predicted the weekly work-hours by, on average, roughly 15 hours. For comparative purposes, plots of the NFF model and the NNLS model Series 3 are illustrated in Figure 12.4.

Table 12.16: Comparison of model results between NFF and NNLS: Kongsberg-tunnel

Dataset description					
Model ID	Algorithm	Observations	Response [h]	Predictors	Dimensions
		n	Y	X	k
KB-N31	NNLS	172	Effective week	$X_1, X_3 \ldots X_{10}$	9
KB-N32	NNLS	172	Shortened week	$X_1, X_3 \ldots X_{10}$	9
KB-N31-NFF	NFF	172	Effective week	$X_1, X_3 \ldots X_{10}$	9
KB-N32-NFF	NFF	172	Shortened week	$X_1, X_3 \ldots X_{10}$	9

Dataset description				
Model ID	Algorithm	No intercept	Response	Observations
		R-square	Y	n
KB-N31	NNLS	0.959	Y'	172
KB-N32	NNLS	0.968	Y'	172
KB-N31-NFF	NFF	0.729	Y'	172
KB-N32-NFF	NFF	0.798	Y'	172

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N31	26.26	-	79.50	1.54	68.10	6.26	1.74	2.51	3.15	0.00
KB-N32	27.43	-	63.85	1.63	56.65	11.72	2.69	1.44	3.53	4.90
KB-N31-*NFF*	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00
KB-N32-*NFF*	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

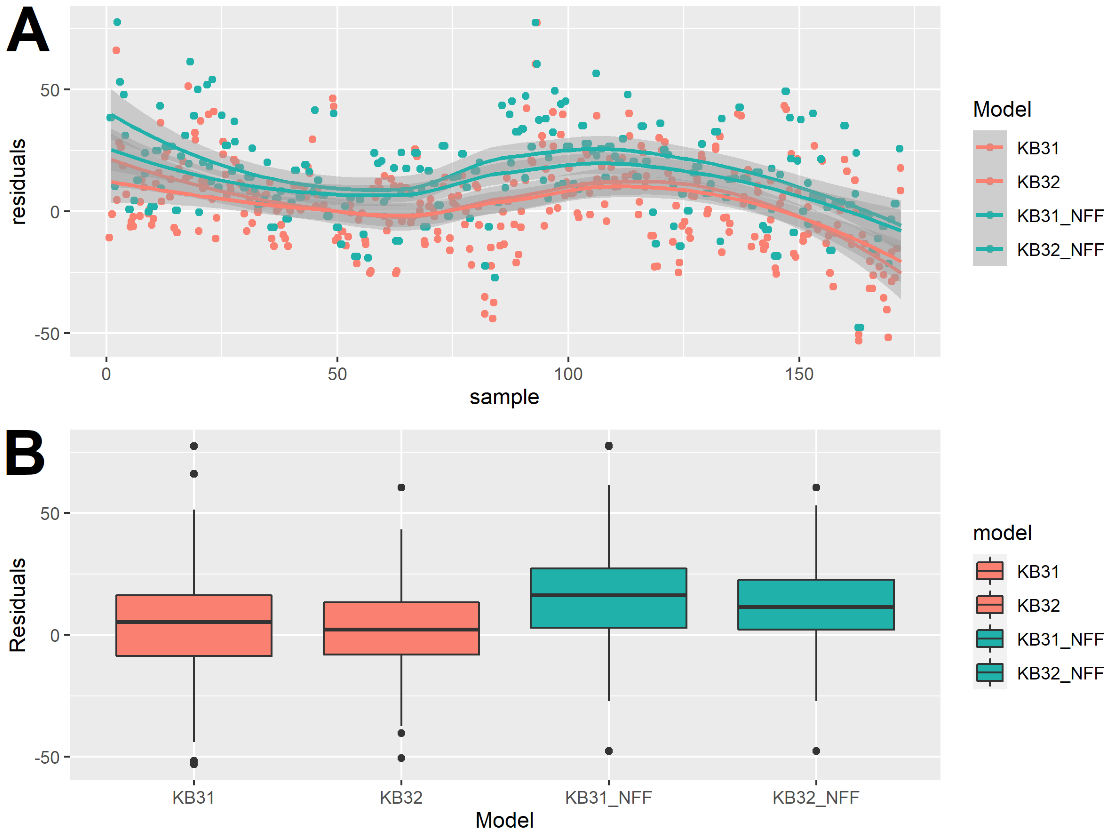


Figure 12.4: Comparison between NFF and NNLS model estimates on the Kongsberg-tunnel dataset. Plot A (top), is a scatter plot of the residuals verses the sample. Plot B (bottom) is a box-plot of the residuals

12.5 Chapter summary

In this final section, some interesting observations realised during the analysis of the Kongsberg-tunnel data (with considerations to the Svartås-tunnel analysis) are presented.

Constrained optimisation was able to produce promising results, but only for *some* variables

At a glance, the mathematical optimisation -based NNLS algorithm was able to produce "realistic" results. For example, the drilling components ($X_1 \dots X_3$), pre-grouting (X_4), excavated material (X_5), rockbolting, $< 4m$ (X_6), and even reinforcements (X_{10}) were consistently comparable to today's NFF time capacity values. But upon closer inspection, the other construction tasks were grossly underestimated. Table 12.17 is a comparative table to illustrate this.

Table 12.17: Comparison between KB-N32 and KB-N52 estimates

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KB-N32	27.43	-	63.85	1.63	56.65	11.72	2.69	1.44	3.53	4.90
KB-N52	-	-	61.03	1.62	55.26	11.32	2.81	1.43	3.40	4.51
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Such mixed-results are curiously similar to those observed with the Svartås-tunnel dataset: where right-tail distributed variables were consistently underestimated by the NNLS algorithm. Figure 12.5 further illustrates this.

Regression analysis was able to produce positive-only results *without* a non-negative constraint

One of the most interesting discoveries during the analysis of the Kongsberg-tunnel was that regression-through-the-origin (RTO) was able to produce positive-only results, despite not having a non-negative constraint. This is a contrast to previous attempts (and to previous tunnel datasets), where RTO analysis produced mixed results, both positive and negative beta coefficients.

However- such positive-only results were only achieved for one model, KB-R04: when the dataset quality was *improved* by identifying the "time away from the tunnel" (Y_A), and by defining "shortened workweeks" (Y_S). This may reveal that, when data quality is sufficient, regression analysis may very well be a useful technique for deriving time capacity value. More research will be required to substantiate this claim however.

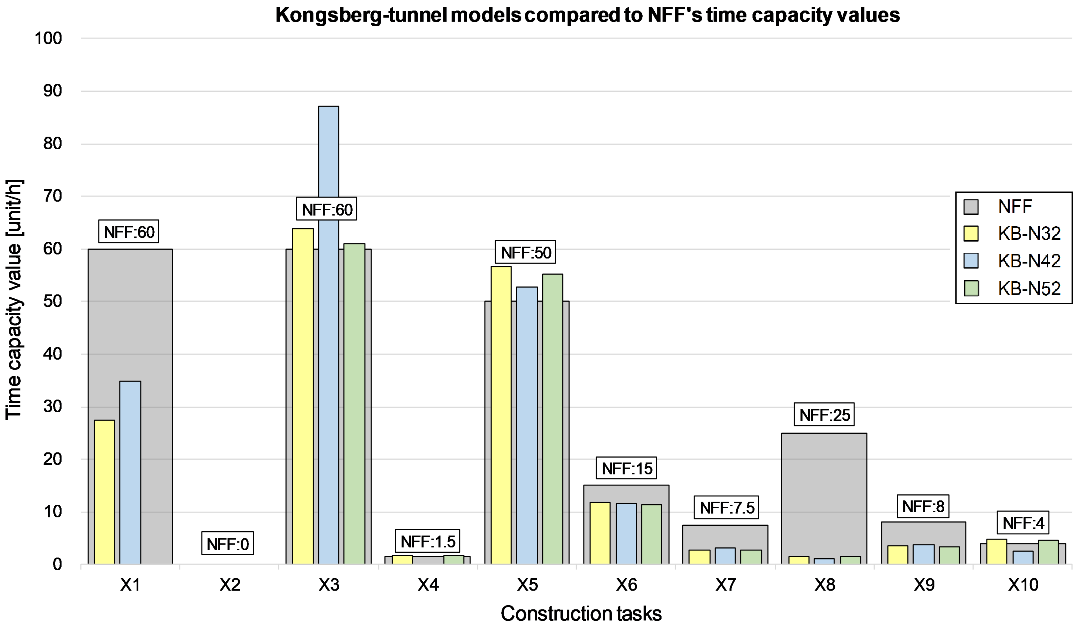


Figure 12.5: A comparison between the Kongsberg-tunnel models and NFF's time capacity values

Reduced dimensionality improved the models performance, but only slightly

The model was somewhat "improved" when the effects of high dimensionality were reduced by combining similar variables together. For example in models KB-N32 and KB-N52, almost all beta coefficient estimates moved closer to the industry standard performance rates when drilling elements were combined into one variable. Admittedly, more research and stronger evidence is needed. Nonetheless, Table 12.17 demonstrates this effect.

12.5.1 The *actual* injection rate has a large variance

Perhaps on a unrelated-note, the Kongsberg-tunnel BoQs provided insightful information to the *true* injection (X_4) performance rate. In these BoQs, the contractors had registered their "injection times" on a weekly basis. By comparing this data against their recorded injection qualities, it revealed that injection rates actually vary remarkably: and that a *constant* value may not truthfully or effectively represent the actuality.

Figure 12.6 (left) suggests that the rate of injection is not constant - instead there is a distinct increase in performance as the quantity of works (per week) increases. One contributing factor is most likely the rigging component. Although tunnel lines 11000 and 12000 are constructed in parallel, Figure 12.6 (right) shows that the average rate of injection is lower in 11000, however only ever so slightly. Nonetheless, the variance through the entire project is still significant - peaking at 2.5 t/h and dropping as low as 0.55 t/h in some cases. Fluctuating as much as 65% from the average rate. Overall, this indicates that perhaps additional factors also come into play while operating at the tunnel face, and that a *constant* time capacity value may not truthfully or effectively represent the actuality.

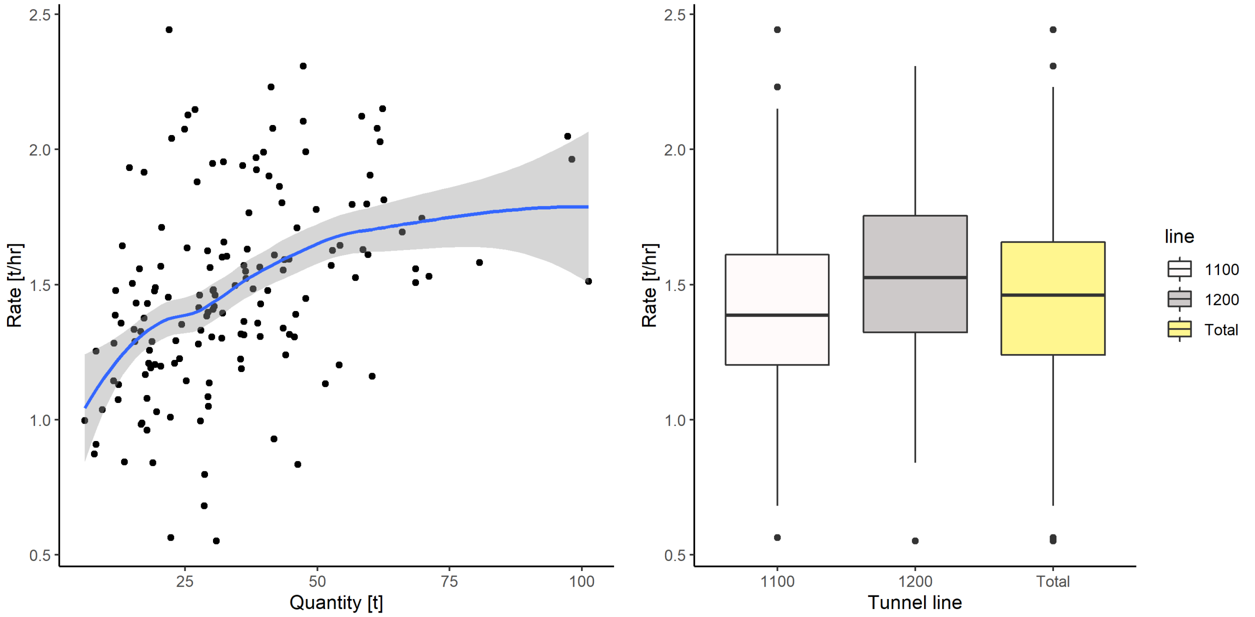


Figure 12.6: Kongsberg-tunnel injection rates. Left: Injection rate vs. quantity scatter plot. Right: Injection rate variances for tunnel line 11000 and 12000.

12.5.2 The currently-available data quality is insufficient

There is strong evidence to suggest that currently-available data is incomplete or of insufficient quality. To clarify, when the Kongsberg-tunnel dataset was modelled, prediction estimates were consistently underestimating the performance rate of the construction tasks. This observation is substantiated by the fact that even current industry standard average performance rates have been only able to account for 70 to 80% of the total construction duration.

As such, following the analysis of *real* D&B tunnel data, a new hypothesis is created. Perhaps, the data itself is not of high enough quality for reliable modelling.

> One of the biggest issues was the inability to confirm the *actual* weekly working-hours performed at the tunnel face.

Chapter 13

Decision making

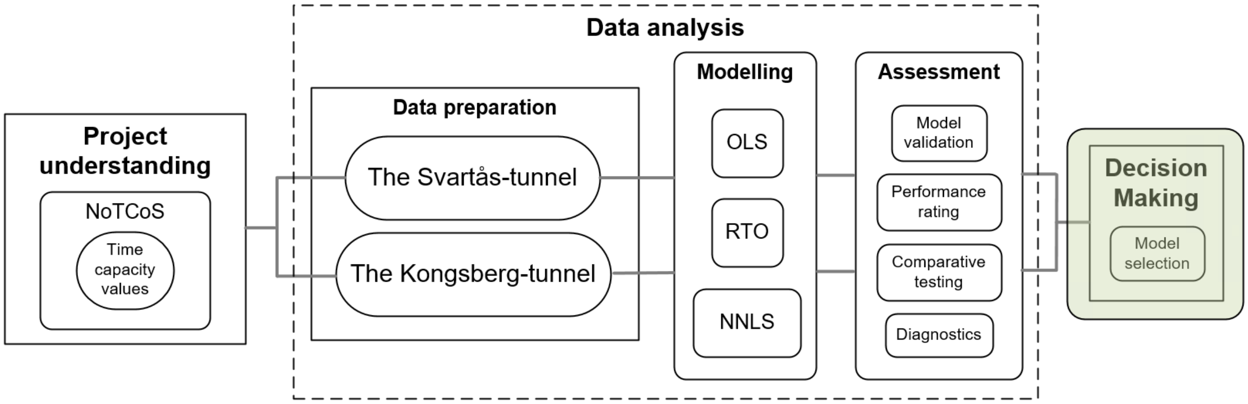


Figure 13.1: A progression chart of the data science workflow - decision making

In this final chapter of the exploratory phase, the results and observations from data analysis of *real* D&B tunnelling construction data are discussed. Following this, an algorithm is selected for primary analysis in the subsequent chapters. Figure 13.1 shows a progression chart illustrating which stage the chapter pertains to in the data science workflow. Overall, the contents of this section include:

- Summary of each of the algorithms tested;
- Discussion about why they may have been "unsuccessful";
- Some limitations to their capabilities;
- General observations; and
- Finally, a predictive algorithm is selected for further evaluation

13.1 Inferences and observations

In this section, inferences are made about the models tested. Discussions into why these models may have been "unsuccessful" are also presented.

13.1.1 Statistical inclined methods

Fundamentally, the equivalent time system may not be a pure statistical problem. When referring back to the Gauss Markov theorem, too many conditions were simply "broken". To summarise:

The parameters to be estimated are not strictly linear

The performance rate of any given construction task is variable, and not a constant rate. Furthermore, there is typically a rigging component associated with each activity. Errors in the model may arise from this fact.

The input variables possess a distinct degree of multicollinearity

Each predictor input variable, although mutually exclusive, implies a clear correlation between each other. To elaborate, when the total working-week is set as the response variable, the occurrence of one task automatically reduces the capacity of the other tasks within that observation. At the same time, some variables are even more directly correlated with each other (for example injection drilling, X_3 and pre-grouting X_4). Because of this multicollinearity, the model is at times unable to distinguish the between correlated variables reliably, when distributing the beta coefficients.

A *constant* response variable is a major red flag

In its crudest form, the bill of quantities represent a constant 101 hours at each observation. From the point of view of a statistician, this is major red flag. In fact, most statistical programs were unable to process such a dataset, when the response variable was held constant at 101 hours - at least until the non-constant modifier was applied.

Nonetheless, such a realisation implies that there is no "real" relation between the independent variables and the dependent variable. To clarify:

> At its core, *statistics* is the study of the change in the response variable, when the predictor variables are changed.

However, in the case of D&B construction logs, the response is constant (or near-constant). As such, the statistical algorithm may be unable to reliably measure the influence each predictor variable has - because the 101 hours remains the same regardless of the predictors. Which on the surface, does have some truth in it - time waits for no one, after all.

The *better* fit line is not necessarily the *better* model

Although these regression techniques oftentimes returned *better* performing estimates, it was at the cost of unconstrained coefficients. That is to say, the model included both positive and negative coefficients in order to achieve a highly performing best-fit line. There is a logic however, that all construction activities should always contribute *positively* to the weekly-working hours. Negative coefficient estimates seemed contradictory to reality, and this is what ultimately sparked my intrigue into non-negative (and positive) constraints.

13.1.2 Mathematical optimisation

From a quadratic programming point of view, the D&B tunnelling data problem can be interpreted as a mathematical optimisation problem with the following parameters:

- continuous quantities;
- linear programming;
- deterministic solutions;
- constrained problems; and
- singular objective function.

All in all, you can define the problem as:

> an *overdetermined* linear system, with non-negative constraints.

Issues with variables with a right-tail distribution

In some datasets, where some predictor variables exhibited a right-tail distribution (which meant, for most observations, a zero quantity was recorded), regression-through-the-origin (RTO) models were only able to produce a negative coefficient for these variables. Interestingly, compared to a non-negative least squares (NNLS) algorithm, the very same dataset instead produced zeros. This indicated that the NNLS model completely omitted these variables from the dataset and model.

In the NNLS scenario, the exclusion of such "outlier" variables meant that the other variables were simply inflated accordingly. When variables are completely discounted, but the response remains the same, all other construction tasks will naturally be perceived with a quicker performance rate than a standard RTO model.

At this stage, I am not entirely sure whether this implies error in the model choice, or if the data is of insufficient quality.

One possibility may be that the NNLS algorithm is better suited for datasets with smaller sample sizes. For datasets where the RTO model falters, the NNLS algorithm may be the next best thing. Nonetheless, further experiments will be performed in the primary analysis chapters.

A non-positive constraint performs equally well

In this study, non-negative constraints are the primary theme under scrutiny. In practice however- a non-positive constraint, can just as easily serve the same purpose. During my research, a non-positive constrained model was also tested. By rephrasing the system of equations with a negative response variable, the non-positive algorithm was able to produce coefficients, identical in magnitude to those derived from a "non-negative" method.

The non-positive constraints can be achieved by running *R*'s "nnpls" package and algorithm. This has been included in Appendix A for reference.

13.2 Some other notable observations

Finally, some other (perhaps unrelated) interesting findings realised in this exploratory phase are presented in this section.

The *actual* injection rate has a large variance

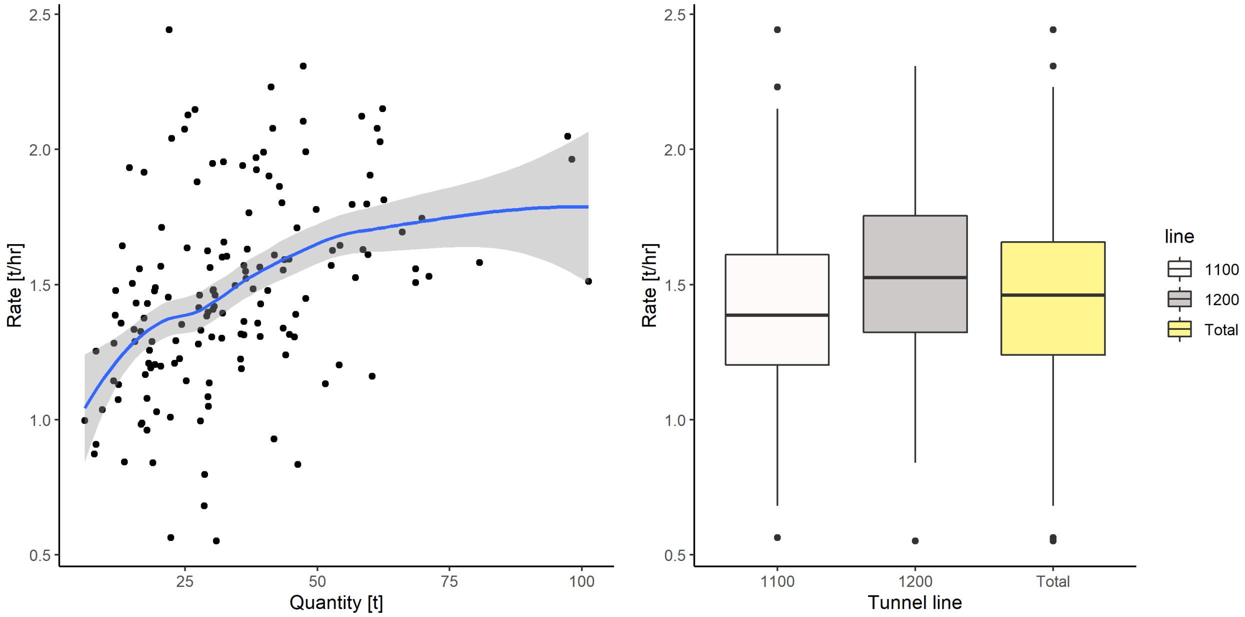


Figure 13.2: Kongsberg-tunnel injection rates. Left: Injection rate vs. quantity scatter plot. Right: Injection rate variances for tunnel line 11000 and 12000.

Figure 13.2 (left) reveals that the rate of injection is not constant. Instead, there is a distinct increase in performance as the quantity of works (per week) increases. One contributing factor is most likely the rigging component. Although tunnel lines 11000 and 12000 are constructed in parallel, Figure 13.2 (right) shows that the average rate of injection is slightly lower in 11000, however only ever so slightly. Nonetheless, the variance throughout the entire project is still significant - peaking at 2.5 t/h and dropping as low as 0.55 t/h in some cases. Overall, this indicates that perhaps additional factors also come into play while operating at the tunnel face.

Evidence that such large variances in the performance rate can occur is however noteworthy; and become instrumental for the development of Hypothesis 3 in the primary analysis. This will be discussed later in Section 14.1.3.

13.3 Limitations

Some limitations to the NNLS algorithm are discussed in this section.

Mathematical optimisation may not be able to generalise

Optimisation algorithms are, by their very nature, project-specific. That is to say, they are confined -only- within the realms of the input data, and unable to generalise. In traditional "predictive analysis", these models are therefore often degraded as simply an "overfit model": a big no-no in most circles. The model's inability to generalise implies that it will falter when confronted with "new" data (Bzdok et al. 2018). Though it is commonly argued, that with sufficient samples, such a project-specific model can still be useful for future predictions.

D&B tunnelling is not a linear system, but that's okay

The NNLS algorithm implies a linear model. However in reality, D&B tunnelling is simply a nonlinear system. For example, when referring to probe drilling with plugging (X_1). This task includes additional supporting components, such as rigging time, and of course, the plug installation. Immediately, this would suggest that the time consumed per metre drilled is in fact non-linear: and a major flaw in the model has been exposed.

However, within the NoTCoS, the equivalent time system (ETS) does directly concern itself to such "minor" supporting activities. Instead, it assumes that the time capacity value will absorb and take into account these expenditures over the entirety of the tunnel construction duration.

Although, the system is fundamentally "incorrect", the small variations are overall time estimated is considered acceptable when compared to the flexibility and transparency that the current ETS has to offer: especially advantageous when resolving time related disputes, outside of the court rooms.

13.4 Model selection

Constrained optimisation has frankly been unable to produce consistent results- but I believe the technique *is* a step in the right direction. Nonetheless, the exploratory phase revealed that perhaps a mathematical optimisation approach, using the non-negative least squares (NNLS) algorithm may indeed be a useful tool for deriving realistic time capacity values. In circumstances where there is insufficient sample size, traditional regression may, in fact, result in an infinite number of solutions. Instead, constrained optimisation techniques may allow the practitioner to "steer" the data in the right direction (so to speak), when the data quantity or quality is too scarce for traditional analysis.

The NNLS algorithm has been shown to be an interesting candidate so far, but its limitations will need thorough assessment before it's ready to contribute to any real-world decisions. As such, this report now transitions to the primary analysis phase (Part IV), where new hypotheses are tested using the chosen algorithm.

The next course of action

Following the analysis of *real* D&B tunnel data, a new hypothesis is created. That perhaps, the data itself is not of high enough quality for reliable modelling. One of the biggest issues was the inability to confirm the *actual* weekly working-hours performed at the tunnel face. Over time, this thinking became one of the major hypotheses to be examined in the next part - alongside "variable performance rates".

Part IV

Analysis

Time capacity values: perhaps an *optimisation* problem?

Part IV forms the primary analysis component of this report. In this section, a back-analysis is performed on the constrained optimisation algorithm: non-negative least squares (NNLS). The objective of this research is to determine whether the errors in the NNLS model estimates are primarily due to incorrect model choice, or to insufficient (or poor quality) data. To achieve this, hypothetical D&B tunnelling data (resembling real-world scenarios) was simulated and tested against the NNLS algorithm. Such a methodology allowed for a controlled working-space: to more-objectively measure the effects of different degrees of "poor quality" data. The description and contents of the following analytical chapters are briefed below.

Element	**Description**
Hypothesis development	Hypotheses are formed to address the question: "What are the major sources of errors in the prediction model?"
Experiment setup	Controlled simulation of D&B tunnelling construction data
Data analysis	• Data modelling • Model validation • Error diagnostics • Results
Discussions	Findings are discussed and the hypotheses is revisited

Chapter 14

A Back-Analysis of the NNLS Algorithm

> *"Garbage in, garbage out."*
>
> - Wilf Hey

Currently-available D&B tunnelling construction data is uncontrolled, incomplete and unreliable. Now- this is not to be interpreted as an accusation of negligence or the lack of foresight. The original data is, after all, not intended for analytical purposes. Be that as it may, under current circumstances, it becomes difficult to pinpoint the source of the model *noise*. The noise can be described as the difference between the predicted value and the observed value. For a typical prediction model, these are mostly attributed to the following reasons:

- Errors in the data;
- Residuals in the model itself; or
- A combination of both (Zhu and Wu 2004).

With this in mind, the process of elimination can therefore potentially unmask this error source. In this chapter, the non-negative least squares (NNLS) algorithm is once again under scrutiny - however this time - under a new lens: in the form of a "back-calculation". Instead of *real* tunnel data, the NNLS algorithm is tested with *hypothetical* tunnel data. This method creates a controlled space for analysis, and ensures that the source of errors is confined to the prescribed parameters, and not to unknown factors. Overall, the general objective is to explore the notion that:

> **The NNLS algorithm is indeed a *useful* technique; but its reliability and accuracy is dictated by the data quality and quantity.**

The contents of this analysis segment has been summarised in Table 14.1.

Table 14.1: Contents of the primary analysis

Category	Description	Output
Hypothesis development	**To address the question:** "What are the major sources of errors in the prediction model?"	**Four hypotheses to be tested:** Hypothesis 1 (Section 14.1.1) Hypothesis 2 (Section 14.1.2) Hypothesis 3 (Section 14.1.3) Hypothesis 4 (Section 14.1.4)
Experiment setup	**Section 14.2** Controlled simulation of D&B tunnelling construction data	The Kangaroo-tunnel (Section 14.3.2) • A *constant* construction rate: with an increasing degree of noise The Koala-tunnel (Section 14.3.3) • A *variable* construction rate: with an increasing degree of variance
Data analysis	• Data modelling • Model validation • Error diagnostics • Results	• Isolation of the source of error • Measurable-effects to the model performance
Communication	Findings are discussed and the hypotheses are revisited	**Chapter 15** • Discussions

14.1 Hypothesis development

In this final analytical segment, four hypotheses were first developed to observe the effects of "poor" quality data: to simulate likely circumstances arising from *real-world* data collection. These have been presented in Table 14.2.

Table 14.2: The four hypotheses to be tested using a back-analysis method

Element	Action
Hypothesis 1 In a vacuum, the NNLS algorithm is ideal for D&B tunnel construction data	**Section 14.3.1** Simulation and testing of an idealistic dataset
Hypothesis 2 Errors in the prediction are due to **lost-time**, and not because of **model error**	**Section 14.3.2** Simulation and testing of various magnitudes of "lost-time"
Hypothesis 3 A variable performance rate will decrease the model performance	**Section 14.3.3** Simulation and testing of various degrees of variance in the performance rate
Hypothesis 4 In reality, a *real* tunnel project probably contains a mixture of both "lost-time" **and** a "variable construction rate"	**Section 14.3.4** The scenarios of Hypothesis 1 and 2 are testing simultaneously.

14.1.1 Hypothesis 1: NNLS performance in a vacuum space

Hypothesis 1 suggests that the NNLS algorithm will produce the optimal time capacity value estimates should five assumptions be satisfied. These are presented in Table 14.3.

Table 14.3: Assumptions for optimum NNLS performance

Construction tasks are performed at a constant linear rate	
Assumption 1:	Construction tasks do not require rigging
Assumption 2:	Construction tasks cannot be performed simultaneously
Assumption 3:	The production rate of construction tasks is constant
Input variables are complete and account for all time consumption	
Assumption 4:	Work is continuous, with no downtime between each task
Assumption 5:	Working hours do not exceed the 101 weekly working-hours

The thought process behind the first hypothesis

There is no dispute that in practice these five assumptions are very much Utopian. However, these untruths (or their actualities) can be considered as significant or acceptable; and that their negative effects can be dampened to an extend. Regarding assumptions 1, 2 and 3 - these preconditions are largely connected to the NotCoS's structure and their stance: that the model is to be based upon the entirety of a tunnel project's duration, and not on singular points in its life cycle. The expert perception suggests that these factors are insignificant or will be self-correcting in the long-run (NFF 2019). On the other hand, contradictions to assumption 4 reveal a two-folded implication. The first, (with similarities to assumptions 1, 2 and 3) suggests that the model is an estimate of the average performance rate of each construction activity. The second overtone however, reveals perhaps a more dire oversight: that the data collected is, in fact, "incomplete" (that a variable is missing).

14.1.2 Hypothesis 2: Errors are due to "time unaccounted for"

Following the violations to assumption 4 and 5, Hypothesis 2 suggests that prediction errors are the result of missing input variables, and not because of an incorrect model choice. The Kangaroo-tunnel dataset aims to address the failings of assumption 4 by simulating isolated-noise scenarios. The objective of these tests is to:

- Measure the effects of "missing data" (unaccounted for time); and
- Isolate and verify the source of error

The proof of Hypothesis 2 aims to mitigate the distortion effects of an unfulfilled assumption 4. Should error source be identified, perhaps future data collection procedures can be updated to retrieve this missing information. Finally, concerning assumptions 4 and 5- these shortcomings can be mitigated by improving the data collection process and its quality. As such:

Hypothesis 2: may be considered a form of *reducible* error

As long as the assumptions and constraints are upheld, this ideal data can be collected, and used to derive realistic time equivalent values.

14.1.3 Hypothesis 3: A variable performance rate interferes with the model

The third hypothesis addresses assumption 3, and considers the ramifications that a variable construction rate will have on the prediction model. Assumption 3 implies that the construction rate is *constant*: as it is with NFF's time capacity rates. In practice however, these performance rate wills vary on a day-to-day basis. Just how much variance is there? - that is not known for sure. Nonetheless, as observable in the Kongsberg-tunnel BoQs (Section 12.5.1), the injection rate can fluctuate as much as 66% around the norm. Therefore, to evaluate the effects of such a scenario, a dataset using variable construction times has also been simulated. The dataset has been labeled Koala-tunnel (Model ID: KO) for reference. Conversely,

Hypothesis 3: may be considered a form of *irreducible* error

14.1.4 Hypothesis 4: *Real* tunnels contain both "lost-time" and a "variable construction rate"

In Hypothesis 2 and 3, "lost-time" and "variable performance rates", are isolated and analysed as mutually exclusive elements. In reality however, a *real* tunnel project probably features a mixture of both. In this concluding hypothesis, data is simulated using the characteristics from Hypothesis 2 and 3. The idea is to simulate how a *real* tunnel project might behave, and to observe how the NNLS algorithm will perform in such conditions.

14.2 Data simulation process

Simulation tunnel data was created using Microsoft Excel's random number generator functions. However- for the dataset to resemble characteristics of a *real* Norwegian tunnel project, several "weights" had to be imposed to the random number generation algorithm. This pseudo-random dataset - while still "random" - creates a more realistic scenario for this thought exercise. A brief description of how the weights were imposed onto the data simulation process is detailed in Table 14.4, however the entire process, along with the coding, has also been fully documented in Appendix D for reference.

14.2.1 Naming convention of generated datasets

In this chapter, a large number of model iterations are simulated and tested. Each with unique parameters and properties. Therefore, to avoid confusion, hypothetical tunnel data generated using a "constant" performance rate (weight) have been labeled "The Kangaroo-tunnel" (Model ID: KR). While data generated using a "variable" weight, has been labeled "The Koala-tunnel" (Model ID: KO).

Table 14.4: Data simulation process

Process	**Action**	**Decision basis**
Step: 1 Simulate a bill of quantities: (**quantity** of works complete)	Generate random **quantities** for each construction task $X_1 \dots X_n$	Distribution and range of random quantities are weighted (based on the histogram data of a typical Norwegian D&B tunnel)
Step: 2 Convert quantities to time consumed (hours)	Apply a *realistic* time capacity value to the generated BoQ	Applied time capacity value is based on current industry estimates (NFF 2019)
Step: 3 Compute total time consumption for each data point (weekly)	Direct summation of resultant hours	Assumption that generated time consumption is mutually exclusive of one another
Step: 4 Select data points that are representative of a typical working week duration	Select data points between 95 and 101 weekly working-hours	Range of data selected is based on the industry standard: 101-hour weekly working hours

Step 1: Simulate a bill of quantities

Histograms from the Svartås-tunnel dataset were examined and formed the basis for the chosen distribution for each construction input variable during the data simulation process. As shown in Figure 14.1, the resultant pseudo-random generated data is resemblant to that of typical Norwegian D&B tunnels, for example the Svartås-tunnel and Kongsberg-tunnel. It is noted however- that the abnormal spike in the X_1 density plot is a result of the Kongsberg-tunnel BoQ structure. This dataset did not have a X_2: instead, all its probe drilling activities were lumped together into X_1 - hence the distinct difference in quantities distribution.

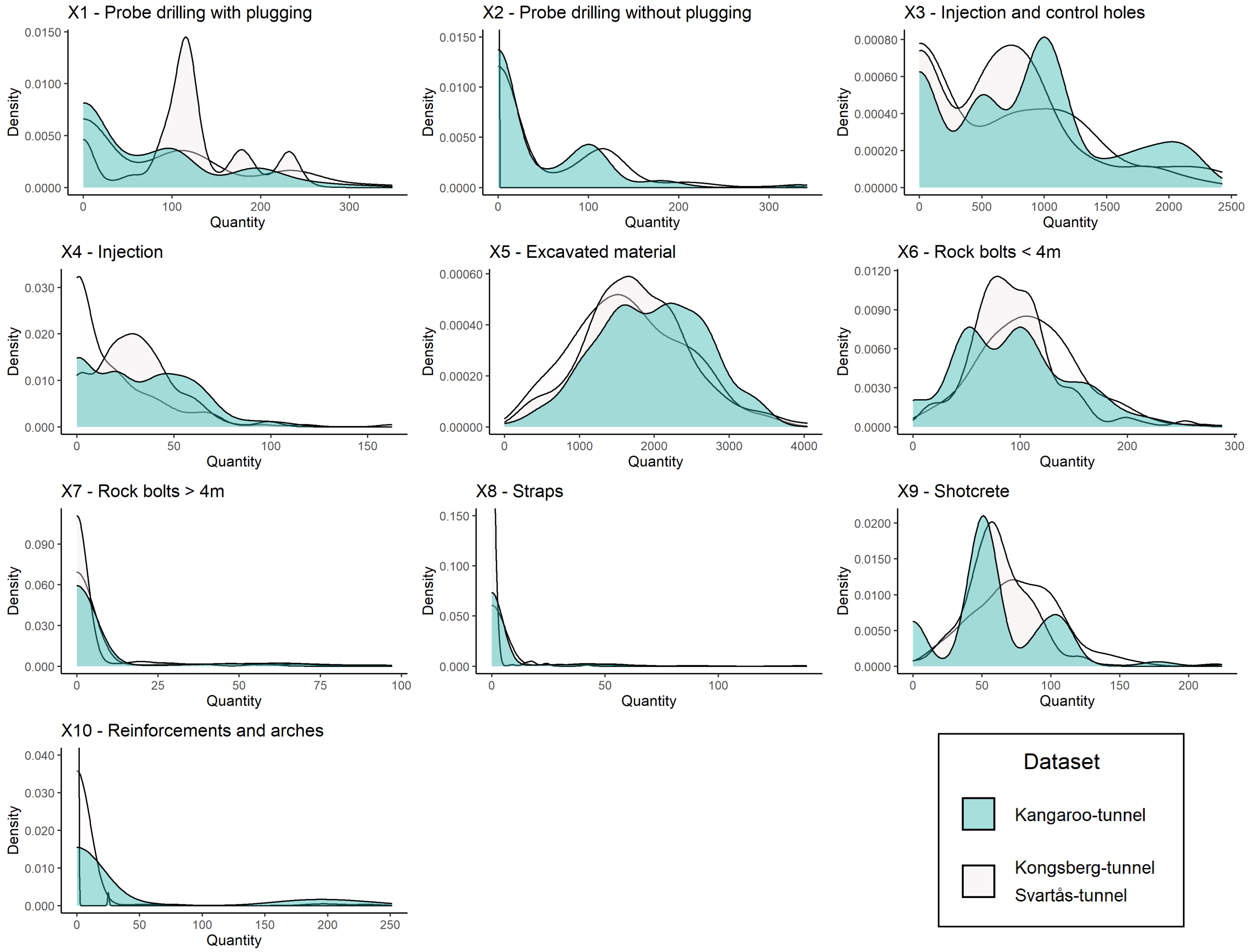


Figure 14.1: Density plots showing that the Kangaroo-tunnel's simulated variables are resemblant to that of typical Norwegian D&B tunnels

Step 2: Specify time capacity values

NFF's standard time capacity values were used to convert quantity values into time values. These rates are generally accepted by the industry, and therefore used as a starting point. These "weights" began as *constant* values. This decision is intended to be a reflection of the NoTCoS. As a reminder, the NoTCoS stipulates that the time capacity value is independent of the geological conditions at the tunnel face. This value was designed to reflect the average production rate over the course of the entire project's construction duration - but a constant value nonetheless. The selected time capacity values used to generate the Kangaroo-tunnel dataset are shown in Table 14.5.

Table 14.5: Production rates used to convert simulated quantities into time values

Standard time capacity values (for single-tube construction) (NFF 2019)										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m : h	m : h	m : h	t : h	m^3 : h	unit : h	unit : h	m : h	m^3 : h	m : h
Capacity	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00
Unit time	0.017	0.025	0.017	0.667	0.020	0.067	0.133	0.040	0.125	0.250

However, in reality, construction rates are variable. As such, a variable weight is also imposed on the data generation process to create the Koala-tunnel dataset.

Step 3: Simulate the total weekly working-hours

The total weekly working-hours (dependent variable, y) were calculated under the assumption that each construction task recorded in the bill of quantities (BoQ) is critical: and therefore cannot occur simultaneously with one another. Mathematically, the dependent variable y can therefore be derived using the superposition principle: with a straight forward summation of the time values derived in Step 2 above.

Step 4: Select data range

Following Step 3, we are left with a bountiful number of new data points! However- upon closer inspection, it was evident that many data points were ill-suited. As illustrated in Figure 14.2, the data points generated ranged between 0 and 200 hours per week. Theoretically, the results for the proceeding set of thought-exercises would not be compromised, should the dataset continue unabridged. Be that as it may, for the sake of authenticity, only data points ranging between 95-102 hours - representing that of a typical Norwegian D&B work week - was selected for modelling.

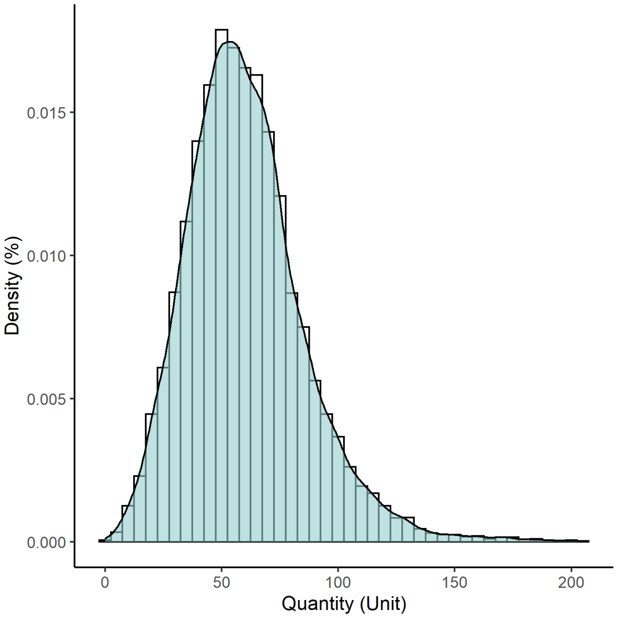


Figure 14.2: Histogram of generated total weekly working-hours

Kangaroo-tunnel prototype configuration

To recap, the configuration of this prototype Kangaroo-tunnel simulation (Model ID: KR00) has been summarised in Table 14.6. And with the tunnel data simulation complete, the analysis can finally begin!

Table 14.6: Kangaroo-tunnel initial template model configuration

Dataset description						
Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR00	0	221	95 to 102	10	$X_1 \ldots X_{10}$	NFF, constant

14.3 Data analysis

In this section, the simulated data is tested to assess the validity of Hypotheses 1, 2, 3 and 4. Hypothesis 1 and 2 can be evaluated using the Kangaroo-tunnel dataset: a hypothetical tunnel with a *constant* production rate. Hypothesis 3 and 4, on the other hand, require new parameters and will be examined with a *variable* construction rate: the Koala-tunnel data (Section 14.3.3).

14.3.1 Hypothesis 1: *Model performance in a vacuum space*

To test this preliminary hypothesis (Section 14.1.1), the idealistic dataset KR00 (as described in Table 14.6) was run through the NNLS algorithm.

Model outputs: Hypothesis 1

The results for the model formed in Hypothesis 1 are presented in Table 14.7.

Table 14.7: Kangaroo-tunnel KR00 (0 noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR00	0	221	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR00	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

14.3.2 Hypothesis 2: *Unaccounted for time*

To verify Hypothesis 2 (Section 14.1.2), an increasing magnitude of "noise" was added to the dataset in this second rendition of modelling. As a reminder, prediction errors produced by the NNLS optimisation model can be attributed to two main reasons: errors in the model, or errors in the data.

In the case of the hypothetical Kangaroo-tunnel, we know that the dataset was generated using *constant* time capacity values. Within this vacuum space, it is therefore logical to deduce that any errors in the prediction can only be attributed to errors in the data: or "missing input variables" - which brings us back to the concept of "lost-time". The idea is to recreate *real* D&B tunnelling phenomena that may result in "missing data": where it by from "lost-time", or from "time unaccounted for" (as observed in the Kongsberg-tunnel, Section 12.1.2). This method allows for this "noise" to be introduced to the data set in a supervised manner. The overall objective was to observe the effects that "missing data" will have on the model's performance.

To achieve this, a random value was generated between an elected range. This value is then added directly to the construction time (response variable, Y). This random value is intended to replicate a real tunnel scenario: in which an unknown amount of time is "unaccounted for" within current construction logs. Note: this random value is evenly distributed. So, for example in Series KB3, where the noise is between 0 and 10 hours, we can expect that, across the entire dataset, the noise will be on average somewhere in the middle: approximately 5 hours. Finally, after a steady increase in imposed noise (up to 20 hours), the probe drilling quantities are combined to reduce the effects of multicollinearity, and the same tests are run.

The eventual model iteration created are described in Table 14.8. Following this, model outputs are presented.

Table 14.8: Noise imposed models: Kangaroo-tunnel data

Dataset description					
Model series	Noise [hr]	Algorithm	Feature Eng.	Predictors	Weights
	ϵ			X	β
KR1	0 to 2	NNLS	Standard	$X_1 \dots X_{10}$	NFF, constant
KR2	0 to 5	NNLS	Standard	$X_1 \dots X_{10}$	NFF, constant
KR3	0 to 10	NNLS	Standard	$X_1 \dots X_{10}$	NFF, constant
KR4	0 to 15	NNLS	Standard	$X_1 \dots X_{10}$	NFF, constant
KR5	0 to 20	NNLS	Standard	$X_1 \dots X_{10}$	NFF, constant
KR6	0 to 20	NNLS	Drill comb.	$X_{123} \dots X_{10}$	NFF, constant

Model outputs: Hypothesis 2

Below, are the test results for the Kangaroo-tunnel, with a constant weight. The results will be further discussed in the inferences section.

Table 14.9: Kangaroo-tunnel KR10 (0 to 2 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR11	0 to 2	188	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR12	0 to 2	746	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR13	0 to 2	3844	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR11	59.32	39.84	59.02	1.49	49.57	14.97	7.26	25.35	7.91	3.97
KR12	61.18	39.54	59.25	1.49	49.40	14.93	7.55	24.33	7.95	3.96
KR13	59.60	39.28	59.33	1.49	49.42	14.79	7.42	25.04	7.91	3.96
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.10: Kangaroo-tunnel KR20 (0 to 5 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR21	0 to 5	208	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR22	0 to 5	821	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR23	0 to 5	4108	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR21	57.59	40.14	58.92	1.47	48.51	14.41	7.15	23.10	7.59	3.95
KR22	58.93	38.18	58.38	1.47	48.64	14.32	7.24	22.83	7.86	3.90
KR23	56.73	38.36	58.44	1.46	48.67	14.22	7.33	25.17	7.75	3.91
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.11: Kangaroo-tunnel KR30 (0 to 10 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR31	0 to 10	241	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR32	0 to 10	926	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR33	0 to 10	4669	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR31	61.59	45.13	57.93	1.44	47.38	13.22	6.98	23.16	7.38	3.78
KR32	57.69	36.08	57.02	1.44	47.16	13.93	7.44	23.47	7.38	3.82
KR33	54.33	35.50	57.22	1.44	47.16	13.27	7.36	23.60	7.40	3.86
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.12: Kangaroo-tunnel KR40 (0 to 15 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR41	0 to 15	286	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR42	0 to 15	1075	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR43	0 to 15	5207	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR41	47.62	50.76	54.75	1.44	45.93	13.34	5.96	19.28	7.04	3.80
KR42	50.63	32.67	56.37	1.43	45.24	12.76	7.46	19.13	7.44	3.82
KR43	49.89	33.74	55.64	1.41	45.64	12.62	6.62	22.41	7.17	3.83
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.13: Kangaroo-tunnel KR50 (0 to 20 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR51a	0 to 20	293	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR51b	0 to 20	282	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR52	0 to 20	1254	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR53	0 to 20	5767	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR51a	37.47	30.82	55.60	1.41	43.75	11.83	5.45	0	7.04	3.82
KR51b	44.60	35.90	54.88	1.41	44.68	12.38	5.17	16.05	6.22	3.95
KR52	41.04	25.89	55.10	1.41	44.79	10.51	7.37	28.70	7.24	3.82
KR53	44.95	30.24	54.94	1.40	44.68	11.28	6.85	18.85	6.56	3.77
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.14: Kangaroo-tunnel KR60 (0 to 20 hr noise) dataset description and model outputs

Dataset description						
Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR61	0 to 20	311	100 to 102	8	$X_{123} \dots X_{10}$	NFF, constant
KR62	0 to 20	1195	95 to 102	8	$X_{123} \dots X_{10}$	NFF, constant
KR63	0 to 20	5695	80 to 105	8	$X_{123} \dots X_{10}$	NFF, constant

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR61	-	-	55.56	1.42	44.21	12.34	6.35	12.05	6.65	3.84
KR62	-	-	54.14	1.39	43.87	12.29	6.85	17.29	6.97	3.84
KR63	-	-	54.55	1.40	44.20	11.31	6.72	21.74	6.69	3.83
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

14.3.3 Hypothesis 3: *A variable performance rate*

Hypothesis 3 anticipates that variances in the production rate will interfere with the overall prediction model. As such, an increasing magnitude of "variance" is applied to each model series. With each iteration step, an increasing-10% deviation from the average performance rate is imposed. For example, the deviation for model series KO3, is 30%. This means that its shotcreting component (normally 8 m^3/h) will be a random rate between 10.4 and 5.6 m^3/h). This method allows the average rate of performance to remain the mostly the same across all iterations; and this way, only the variance factor changes. Mathematically, the range in which the random numbers are restricted to, is simply:

$$
\begin{aligned}
min &= \bar{\beta} - \bar{\beta} \times d, && (14.1)\\
max &= \bar{\beta} + \bar{\beta} \times d && (14.2)
\end{aligned}
$$

where:

d = is the deviation magnitude; and

$\bar{\beta}$ = is average production rate of the construction task.

The overall Koala-tunnel model iterations are presented in Table 14.15.

Table 14.15: Variable construction rate models: Koala-tunnel data

Dataset description					
Model series	Variance [%]	Algorithm	Feature Eng.	Predictors	Weights
	d			X	β
KO1	10	NNLS	Standard	$X_1 \ldots X_{10}$	NFF, variable
KO2	20	NNLS	Standard	$X_1 \ldots X_{10}$	NFF, variable
KO3	30	NNLS	Standard	$X_1 \ldots X_{10}$	NFF, variable
KO4	40	NNLS	Standard	$X_1 \ldots X_{10}$	NFF, variable
KO5	50	NNLS	Standard	$X_1 \ldots X_{10}$	NFF, variable
KR6	0 to 20	NNLS	Drill Comb.	$X_1 \ldots X_{10}$	NFF, variable

Model outputs: Hypothesis 3

The results for the Koala-tunnel models are presented in this section. Following this, are the inferences and discussions.

Table 14.16: Koala-tunnel KO10 (10% variance) dataset description and model outputs

Dataset description						
Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	n	Y	k	X	β
KO11	10	179	100 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO12	10	715	95 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO13	10	3703	80 to 105	10	$X_1 \dots X_{10}$	NFF, variable

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO11	66.83	34.34	59.22	1.55	49.17	14.33	5.85	19.25	7.87	4.09
KO12	58.04	46.67	59.62	1.52	49.24	14.39	8.11	24.34	7.79	4.13
KO13	55.66	39.89	60.15	1.52	49.65	14.09	7.58	24.78	7.82	4.08
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.17: Koala-tunnel KO20 (20% variance) dataset description and model outputs

Dataset description						
Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	n	Y	k	X	β
KO21	20	169	100 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO22	20	762	95 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO23	20	3808	80 to 105	10	$X_1 \dots X_{10}$	NFF, variable

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO21	37.85	32.18	62.79	1.63	48.59	11.30	6.94	9.88	7.53	4.53
KO22	49.17	41.25	61.44	1.61	47.02	11.03	7.27	63.93	8.02	4.31
KO23	48.03	31.95	61.96	1.58	48.31	12.54	8.11	23.47	7.46	4.23
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.18: Koala-tunnel KO30 (30% variance) dataset description and model outputs

Dataset description

Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	n	Y	k	X	β
KO31	30	190	100 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO32	30	821	95 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO33	30	3991	80 to 105	10	$X_1 \dots X_{10}$	NFF, variable

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO31	41.58	27.22	61.71	1.68	46.40	9.36	6.44	41.27	6.58	4.25
KO32	47.51	24.10	61.16	1.67	44.44	11.04	10.25	17.15	7.60	4.44
KO33	44.70	32.76	62.43	1.64	46.14	11.13	7.61	26.44	7.31	4.41
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.19: Koala-tunnel KO40 (40% variance) dataset description and model outputs

Dataset description

Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	n	Y	k	X	β
KO41	40	209	100 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO42	40	852	95 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO43	40	4183	80 to 105	10	$X_1 \dots X_{10}$	NFF, variable
KO44	40	10530	60 to 105	10	$X_1 \dots X_{10}$	NFF, variable

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO41	105.85	32.83	62.10	1.89	41.50	8.93	6.06	8.81	6.71	5.12
KO42	27.25	23.20	65.42	1.77	42.93	8.34	9.65	17.46	7.33	4.98
KO43	36.64	29.52	64.40	1.75	44.35	9.50	8.64	18.64	6.83	5.03
KO44	41.28	32.68	61.69	1.68	48.11	10.50	8.07	24.38	6.78	4.58
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.20: Koala-tunnel KO50 (50% variance) dataset description and model outputs

Dataset description

Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	*n*	*Y*	*k*	*X*	*β*
KO51	50	254	100 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO52	50	975	95 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO53	50	4363	80 to 105	10	$X_1 \dots X_{10}$	NFF, variable
KO54	50	10603	60 to 105	10	$X_1 \dots X_{10}$	NFF, variable

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO51	44.64	22.44	67.31	1.95	42.29	6.10	13.64	96.39	4.86	5.81
KO52	22.80	26.11	64.88	1.98	42.24	7.97	8.69	13.74	5.15	5.56
KO53	33.43	23.80	64.32	1.93	44.06	7.54	9.58	15.64	6.14	5.47
KO54	34.69	25.51	62.20	1.78	48.48	8.70	7.68	21.05	6.37	5.22
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table 14.21: Koala-tunnel KO60 (50% variance) dataset description and model outputs

Dataset description

Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	*n*	*Y*	*k*	*X*	*β*
KO61	10	496	100 to 102	8	$X_{123} \dots X_{10}$	NFF, variable
KO62	10	1995	95 to 102	8	$X_{123} \dots X_{10}$	NFF, variable
KO63	10	8754	80 to 105	8	$X_{123} \dots X_{10}$	NFF, variable

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO61	-	-	56.02	1.86	41.93	7.80	8.61	12.59	5.28	5.58
KO62	-	-	59.63	2.02	40.23	7.36	9.58	19.82	6.16	5.27
KO63	-	-	61.17	1.88	43.64	7.48	7.88	15.14	5.82	5.38
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

14.3.4 Hypothesis 4: *A mixture of noise and variability*

In this concluding hypothesis, the "noise" element is introduced to the Koala-tunnel dataset: resulting in a dataset with both "noise" and "variable construction rate". The idea is to simulate how a *real* tunnel project might behave, and to observe how the NNLS algorithm will perform in such conditions.

Sensitivity testing in Hypothesis 2 and 3 was conducted using extreme magnitudes of missing data and variability ranges. In reality however, these values are most likely not so drastic. In this final hypothesis, I postulated that perhaps, a typical scenario would instead entail the following characteristics (Model ID: KO99):

- 0 to 2 hours of weekly missing data (lost-time);
- 10% variance in the performance rate; and
- All drilling elements combined (X_{123}).

Model outputs: Hypothesis 4

The results for the Koala-tunnel model KO99 are presented below.

Table 14.22: Koala-tunnel KO99 (0 to 2 hours noise and 10% variance) dataset description and model outputs

Dataset description						
Model ID	Noise [hr]	Variance [%]	Observations	Response [hr]	Predictors	Weights
	ϵ	d	n	Y	X	β
KO99	0 to 2	10	1549	95 to 102	$X_{123} \dots X_{10}$	NFF, variable

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO99	-	-	62.40	1.45	49.98	10.19	7.79	17.62	6.67	4.26
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

14.4 Inferences

In this section, the results from the Kangaroo-tunnel and the Koala-tunnel are discussed.

Evaluating the hypotheses made

All in all, the results indicated that the root causes of errors are in fact a combination of reducible errors ("missing data"), and irreducible errors ("variability in the construction rate"). These have been elaborated further below.

Hypothesis 1: *Model performance in a vacuum space*

To address Hypothesis 1, the notion of an *ideal* dataset was analysed using the NNLS algorithm. The Kangaroo-tunnel: characterised as a tunnel with a constant -unwavering- performance rate; and all time-spent-at-the-tunnel face accounted for, was one such model dataset. Unsurprisingly, as observed in the idealistic dataset KR00, the algorithm was able to produce beta coefficients identical to that of NFF's unit times, with zero errors. As such, Hypothesis 1 can be considered *valid*.

Hypothesis 2: *Unaccounted for time*

In the second hypothesis, it is theorised that the errors in the model can be explained by the magnitude of "unaccounted for time". As such, the Kangaroo-tunnel is once again called upon, but this time, an increasing level of noise is imposed to each model step. Surprisingly, the algorithm was able to produce respectable estimates despite the missing data. For example, a quick glance to Table 14.23 reveals potentially promising prediction rates. Despite the 0 to 10 hour noise range (which equates to, on average, 5 hours of "missing data" per week), the NNLS algorithm was still able to produce estimates, on average within 5% of originally prescribed time capacity values. Furthermore - up until this point - the model errors (caused by the noise), appears to be distributed quite evenly across all 10 dimensions. This can be visualised in Figure 14.3.

However, as also indicated in Figure 14.3, the reliability of the model begins to quickly falter when the noise range approaches "0 to 15 hours". Additional to this, some construction tasks seem to take on a disproportionate amount of the errors, more than others. This is especially obvious when the results are examined as a "difference from the NFF's time capacity value" plot. In Figure 14.4, appears as though X_1 and X_8 are taking on the bulk of the errors. As such, the variance of the overall predictions also appears to be greater than that of the other construction tasks. Nonetheless, an interesting observation. This will be investigated in the diagnostics section (Section 14.5).

The model prediction estimates improved slightly when dimensions were reduced In model series KR6, all drilling components (X_1, X_2 and X_3) were combined into one variable X_{123} to reduce the effects of high dimensionality. The model results improved, but only slightly.

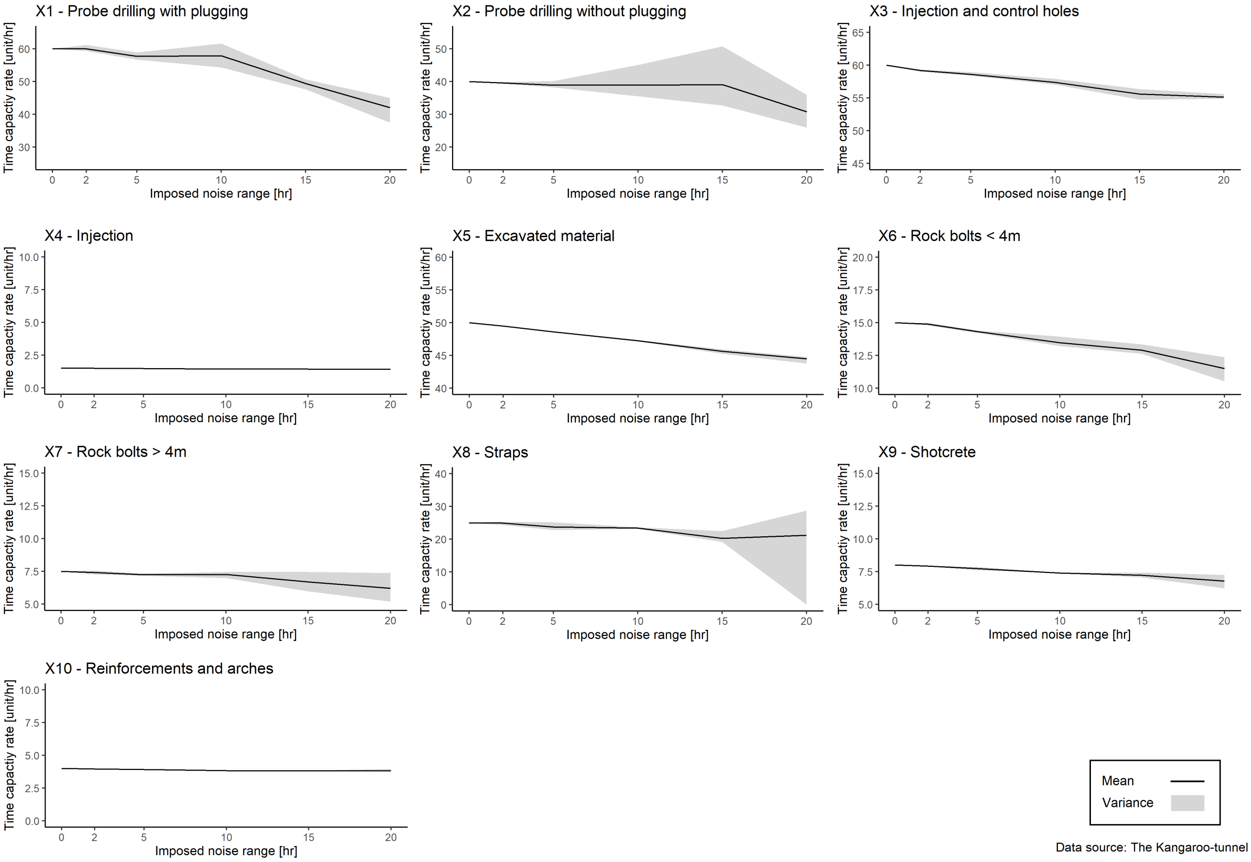
Table 14.23: Kangaroo-tunnel KR30 (0 to 10 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR31	0 to 10	241	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR32	0 to 10	926	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR33	0 to 10	4669	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR31	61.59	45.13	57.93	1.44	47.38	13.22	6.98	23.16	7.38	3.78
KR32	57.69	36.08	57.02	1.44	47.16	13.93	7.44	23.47	7.38	3.82
KR33	54.33	35.50	57.22	1.44	47.16	13.27	7.36	23.60	7.40	3.86
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00



Data source: The Kangaroo-tunnel

Figure 14.3: Estimated time capacity values against the magnitude of imposed noise plot

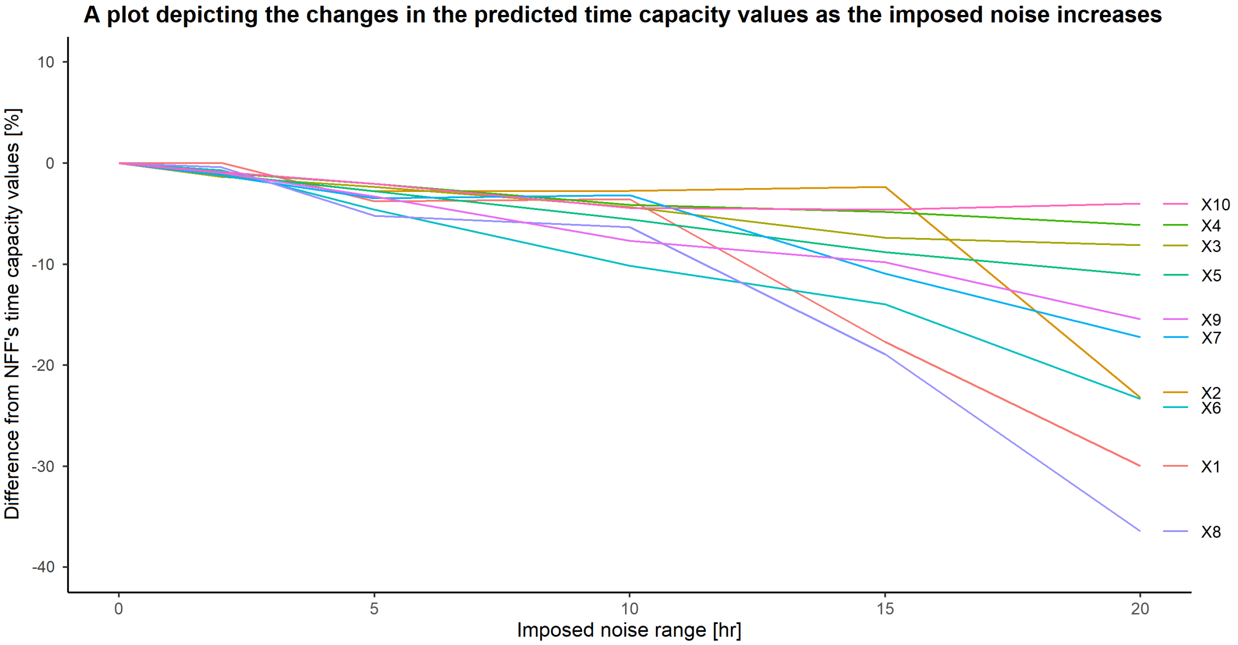


Figure 14.4: Difference in the predicted time capacity values against the magnitude of imposed noise plot

Hypothesis 3: *A variable performance rate*

The third hypothesis was surmised to investigate the effects a variable performance rate will have on the NNLS model performance. To address this, a tunnel was simulated using variable construction rates: the Koala-tunnel. The objective is to assess the impact of the irreducible error.

In the first model hypothetical scenario, the Koala-tunnel experienced up to 10% variance in its construction rate. Despite such fluctuations, the model was able to produce reasonable estimates. As shown in Table 14.24, predicted time capacity values, on average, only strayed approximately 3% from the *real* (NFF) performance rates.

Table 14.24: Koala-tunnel KO10 (10% variance) dataset description and model outputs

Dataset description						
Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	n	Y	k	X	β
KO11	10	179	100 to 102	10	$X_1 \ldots X_{10}$	NFF, variable
KO12	10	715	95 to 102	10	$X_1 \ldots X_{10}$	NFF, variable
KO13	10	3703	80 to 105	10	$X_1 \ldots X_{10}$	NFF, variable

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO11	66.83	34.34	59.22	1.55	49.17	14.33	5.85	19.25	7.87	4.09
KO12	58.04	46.67	59.62	1.52	49.24	14.39	8.11	24.34	7.79	4.13
KO13	55.66	39.89	60.15	1.52	49.65	14.09	7.58	24.78	7.82	4.08
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

However, as soon as the variance exceeded 10%, the model performance drastically decreased. As illustrated in Figure 14.5 and Figure 14.6, increased magnitudes of variance quickly pushed the beta coefficients away from their *actual* values. Furthermore, the variance in the prediction estimates became more volatile as the performance rate variance increased. Specifically notable in the probe drilling (X_1 and X_2) and straps X_8 components, where the estimates became unstable and frankly unusable.

This observation is a undoubtedly a wrench in the works. Although it may be implausible for such high levels of variance to occur in the real-world - and across no-less than each of the ten construction tasks simultaneously - *and* consistently over the entire duration of the project, the NNLS algorithm's inability to reliably model such a scenario is telling. This obstacle - if unsolved - may very well be the demise of an optimisation-based solution. But it is difficult to tell at this point.

Reduced dimensionality improved the results slightly- again As it was with the Kangaroo-tunnel dataset, the number of dimensions were reduced by melding all the drilling components (X_1, X_2 and X_3) into one variable X_{123} for the model series KO6. Again, the model results improved, albeit very slightly.

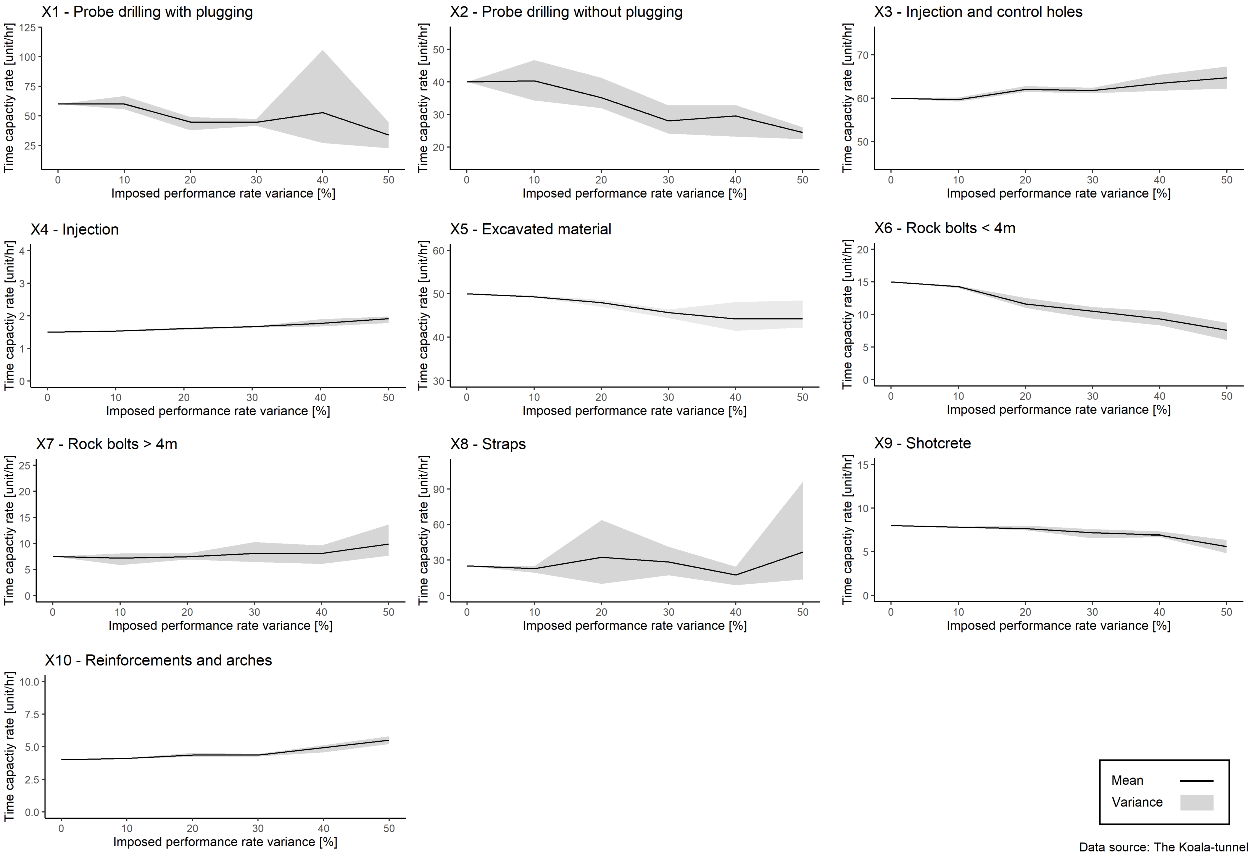


Figure 14.5: Estimated time capacity values against the magnitude of performance rate variance

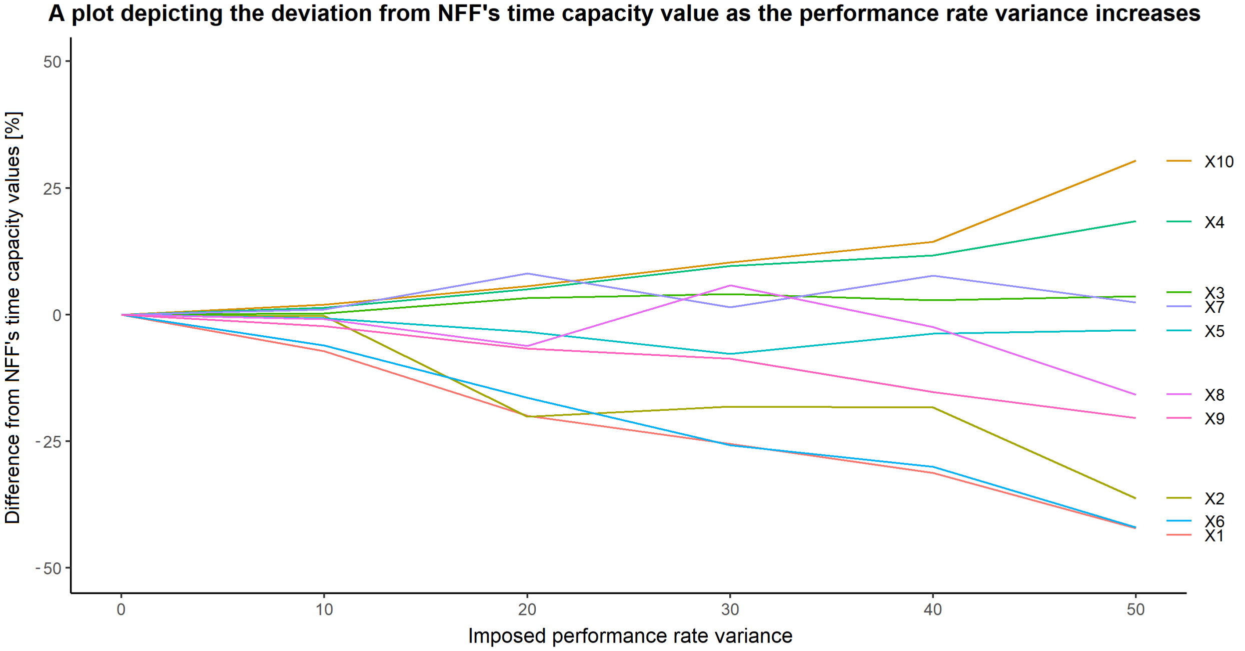


Figure 14.6: Deviations from NFF's time capacity values as the performance rate variance increases

Hypothesis 4: *A mixture of noise and variability*

In the final hypothesis, the simulated tunnel data was comprised of both "noise" and "variable construction rates". Despite these obstacles, the NNLS algorithm was able to produce promising results. Yes, some construction tasks were notably underestimated. But all in all, the predicted performance rates did not stray too far from NFF's time capacity values. Overall prediction error differed on average by roughly 10%. Figure 14.7 further illustrates this point. The research conducted here is anything but concrete or conclusive. But such results are a step in the right direction. However due to time constraints, I was unable to pursue this idea extensively. This predicament and suggestions for further work will be discussed in the concluding chapters.

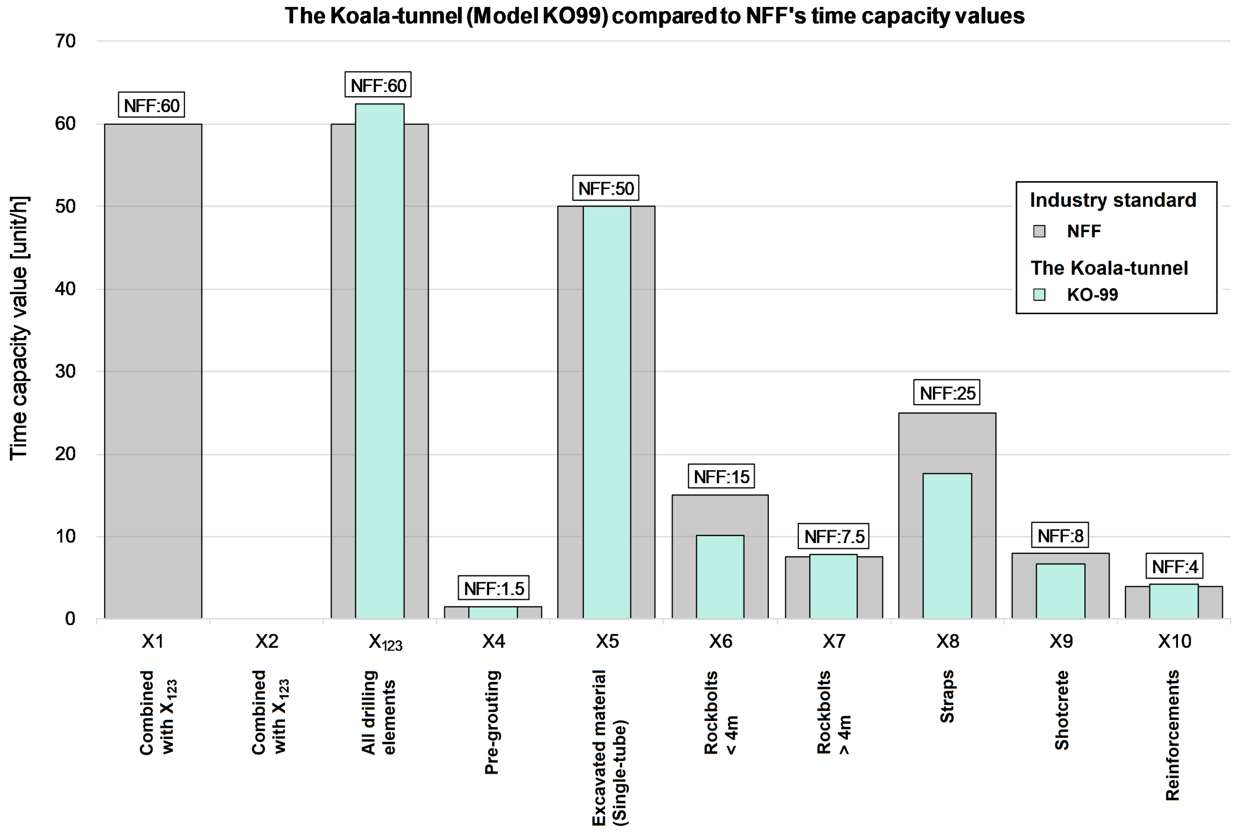


Figure 14.7: Koala-tunnel (model KO99) compared to NFF's time capacity values

14.5 Diagnostics

In the modelling process of simulated data, it was revealed that some variables appealed to inherit a disproportionate amount of the errors, compared to the other variables. This observation was also seen in *real* tunnel projects (the Svartås and Kongsberg-tunnel). As a reminder, in those studies, it was presumed that right-tail skewed construction tasks (such as probe drilling and straps) were seemingly more difficult to model than the other variables. Nonetheless, in this diagnostics section, I attempt to explain why.

14.5.1 How are the errors and the beta coefficients distributed?

The test-subject chosen for this exercise is the Kangaroo-tunnel dataset KR53: where 0 to 20 hours of noise is imposed randomly across each observation. This dataset was selected because it was the "worst-case-scenario". Back to the point- as the Kangaroo- and Koala-tunnel were both simulated using histogram information from the Svartås-tunnel, I had the suspicion that the "distribution density" may have something to do with the apparent disproportionate errors.

Step 1 I first examined the average quantity of each construction task across the entire construction duration (Table 14.25).

Table 14.25: The average quantity of each variable per week

Average quantity per week										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m	m	m	t	m^3	unit	unit	m	m^3	m
KR53	129.3	104.3	1069.8	34.8	1926.0	95.2	44.8	50.3	68.7	133.9

Step 2 Thereafter, I divided this quantity by the *actual* performance rate (NFF's values), to find the average time effect (time consumed) each construction task would in *actuality*, per each observation, **when they occur** (Table 14.26).

Table 14.26: The average time consumed by each variable per week

Average time consumed per week										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	h	h	h	h	h	h	h	h	h	h
KR53	2.2	2.6	17.8	23.2	38.5	6.3	6.0	2.0	8.6	33.5

If you interpret Table 14.26, you will notice that some construction tasks - when they occur - will take up a large portion of the entire week. For example, reinforcements (X_{10}), consumes, on average, 33.5 hours, when the task is initiated. This seems logical, as this construction activity is generally required on demand, and not systematic.

Step 3 I assumed that the amount of "noise" or error per week has been distributed evenly across all ten variables. So, in this step, I noted down the amount of noise present at each of these observations. Note: But **only** for observations where the variable actually occurred. Remember, some construction tasks do not occur at all for the most part. Finally, the average of this noise was divided by the number of dimensions. As shown in Table 14.27, the average noise experienced by the construction tasks was noticeably unique, but nonetheless floated between around 1 hour, per variable, per week, for the most part.

Table 14.27: The average noise experienced by each variable per week

Average noise per week										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	h	h	h	h	h	h	h	h	h	h
KR53	1.14	1.12	1.11	1.07	1.14	1.14	1.09	1.12	1.13	0.97

Across the entire dataset, the average noise was approximately 11.36 hours. Perhaps logically, this means that each construction task is expected to be inflated by approximately 1.14 hours each week.

Step 4 If you then add this average amount of noise experienced, per variable, per week, to each average time consumed (as calculated in Step 2, Table 14.27), you will return a *new* average time consumed, per variable, per week. This has been summarised in Table 14.28.

Table 14.28: The *new* average time consumed by each variable per week

***New* average time consumed per week**										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	h	h	h	h	h	h	h	h	h	h
KR53	3.3	3.7	18.9	24.2	39.7	7.5	7.1	3.1	9.7	34.4

Step 5 If you divide the average quantity (Step 1, Table 14.25) by this *new* expected time consumed, per week (Step 4, Table 14.28), you will derive the following time capacity values (Table 14.29). Interestingly enough, if you compare these results with the *actual* NNLS model outputs of KR53, you will notice that they are eerily similar (Table 14.30).

Step 6 Finally, if we return back to Step 3, where we looked at the average noise experienced by each variable, we can calculate the ratio of this noise compared to the expected time consumed, per variable, per week (Table 14.31).

An immediate reaction, is that the ratio between signal and noise is grossly disproportionate for some variables. For example, the probe drilling and straps variables prediction estimates are, in theory, comprised of approximately

Table 14.29: Back-calculated time capacity values

Time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR53-(Back calc.)	39.25	28.00	56.47	1.43	48.57	12.72	6.34	16.06	7.07	3.89

Table 14.30: Comparison between back-calculated and NNLS time capacity values

Time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR53 (Back calc.)	39.25	28.00	56.47	1.43	48.57	12.72	6.34	16.06	7.07	3.89
KR53 (Actual)	44.95	30.24	54.94	1.40	44.68	11.28	6.85	18.85	6.56	3.77

50% of pure noise! However, you may also notice that some variables with "low-ratios" are still performing poorly (for example the rock bolts $> 4m$, X_6). An additional factor must be in play.

Step 7 To examine this further, the ratios are then multiplied by their *actual* performance rates. For this thought-exercise, I have labeled this value as a "error distribution value" (EDV) for future reference. Table 14.32 summarises the resultant EDVs.

At a glance, the magnitude of the EDV is almost directly correlated with the "prediction error" (the difference between the *actual* and *predicted* time capacity values). A large EDV coincides with a large difference. For example, X_1, X_2 and X_8 all scored the highest EDV - and by a significant margin - and their prediction estimates deviated by as much as 25% from the *actual* performance rate. Oppositely, X_4 and X_{10} scored the lowest EDV, and their overall performance was only compromised by 5 or 6%.

More interestingly, when these EDVs are plotted against their "prediction errors", in a logarithmic scale, the correlation becomes even more profound. Figure 14.8 illustrates this point.

Table 14.31: The ratio between the noise and the average time consumed for each variable per week

Ratio between the noise and the average time consumed										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	ratio	ratio	ratio	ratio	ratio	ratio	ratio	ratio	ratio	ratio
KR53	0.53	0.43	0.06	0.05	0.03	0.18	0.18	0.56	0.13	0.03

Table 14.32: The "error distribution value" of each construction variable

"Error distribution value" (EDV)										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	EDV	EDV	EDV	EDV	EDV	EDV	EDV	EDV	EDV	EDV
KR53	31.73	17.14	3.75	0.07	1.47	2.68	1.37	13.93	1.06	0.12

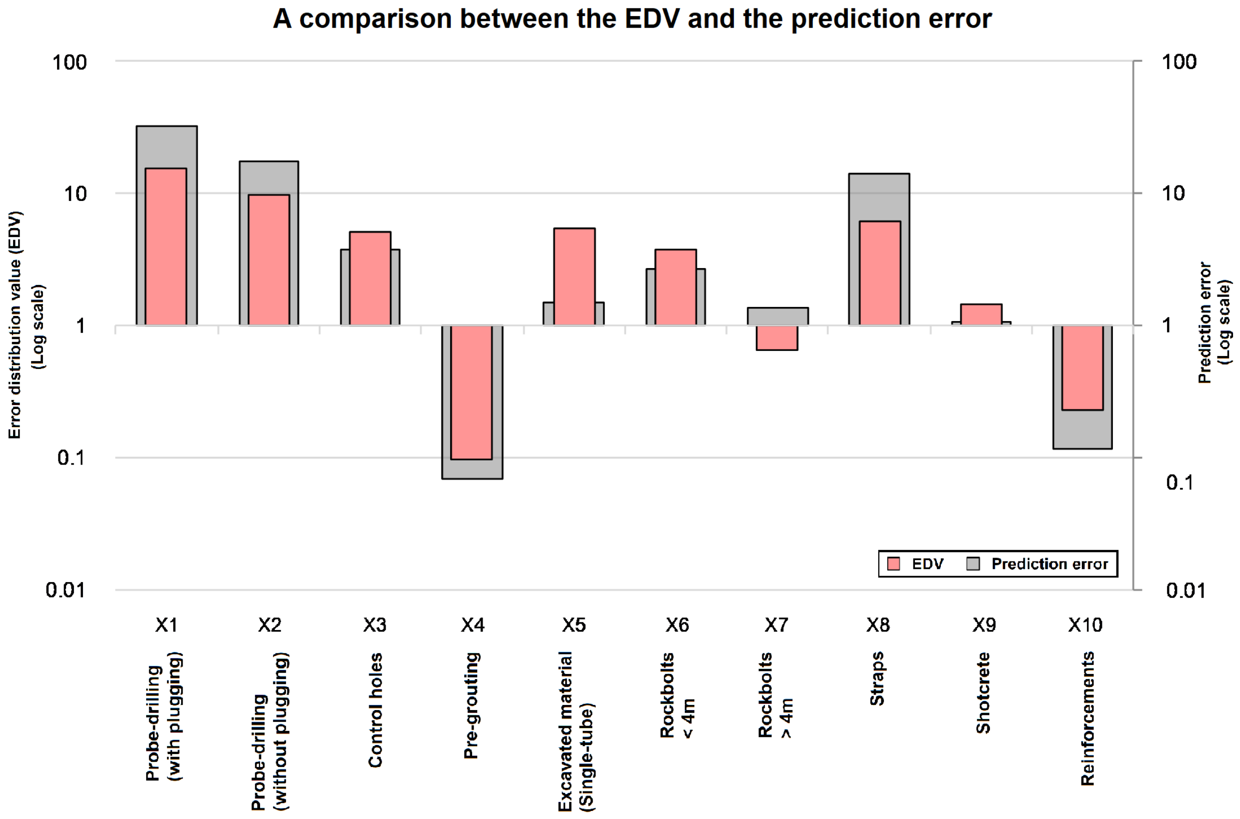


Figure 14.8: A logarithmic comparison between the EDV and the prediction error

14.6 Chapter summary

In this closing section, findings discovered in this "back-analysis" of the NNLS algorithm are briefly summarised.

Model estimates were held stable - up until a point

Despite the introduction of noise, the estimated time capacity values did not deviate substantially from the prescribed NFF values. This was notable in model series KR3 (Table 14.33), where the NNLS algorithm was able to produce estimates, on average within 5% of originally prescribed time capacity values.

Table 14.33: Kangaroo-tunnel KR30 (0 to 10 hr noise) dataset description and model outputs

Dataset description						
Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR31	0 to 10	241	100 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR32	0 to 10	926	95 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR33	0 to 10	4669	80 to 105	10	$X_1 \ldots X_{10}$	NFF, constant

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR31	61.59	45.13	57.93	1.44	47.38	13.22	6.98	23.16	7.38	3.78
KR32	57.69	36.08	57.02	1.44	47.16	13.93	7.44	23.47	7.38	3.82
KR33	54.33	35.50	57.22	1.44	47.16	13.27	7.36	23.60	7.40	3.86
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Large fluctuations in the construction rate require an even larger sample size

As the variance in the *actual* construction rate increases, the model requires exponentially more samples to achieve a reliable prediction estimate. This could be observed in this modelling exercise, as the model performance improved significantly as the number of observations increased.

The distribution of errors is not evenly spread across all dimensions

As perhaps demonstrated in the diagnostics section, when the data quality or quantity is insufficient, the distribution of errors will not be spread evenly across all dimensions. I have no concrete proof or any theoretical backing to support this, but initial investigations reveal that variables with the following characteristics are the most vulnerable to the model's errors:

- Variables with right-tail skewed distributions; and
- Variables with naturally low beta coefficients (a high performance rate).

Reaching the limits of my research

Be that as it may, we have probably reached the extent of this master thesis. Due to the time constraints imposed on me, I cannot proceed further from this point on- this is where the research and expertise ends. Nonetheless, this will be discussed further and perhaps promoted as a future topic of research in the closing chapters (Chapter 16).

Part V

Discussions and Summary

Chapter 15

Discussions

Findings from the exploratory and primary analysis were alluding to the non-negative least squares (NNLS) algorithm as a useful predictive tool for extracting time capacity values from construction data. Its success, however, is heavily dependent on a strict set of conditions. And if unmet, the consequences are quickly hampering.

Not intending to sound defeated- but perhaps these conditions will never be fully realised or feasible. Despite this, I still strongly believe the studies performed here, have been a step in the right direction.

Contents

In this discussions chapter, eight major themes are addressed. These are listed below.

15.1 So, why did the other models falter?

Statistical and *machine learning* approaches are generally interested in the inferences from the particular to the general population (Tukey 1962b). Their models are intended to make accurate predictions and inferences from data untested, unseen, and of the *same* population. The biggest culprit here is this "*same* population" condition. When concerning Norwegian D&B tunnelling, such a feat is (currently) not technologically feasible. Data from the *same* population has been impractical to define and obtain, due to the inherent variability of the ground conditions. Because tunnels are constructed within their own unique arrangement of geological features, a "working" statistical- or machine learning -themed model will definitely demand non-linear parameters; inter-variable relationships defined; and external factors considered. Such a scenario not *yet* realistic, as these measurements are not yet captured by the currently-available data.

A flexible model is more useful than an accurate one

Conversely, the Norwegian Tunnel Contract System (NoTCoS) has recognised that such levels of accuracy is not economical, nor even completely necessary (Kleivan 1989). Rather than attempt to create an accurate and precise model, the NoTCoS has relied instead on a *flexible* system for their decision-making: the equivalent time system (ETS) (Grøv 2012). This flexibility - by design - is largely attributed to the ETS's transparency, and simplicity, and linear behaviour.

> The equivalent time system (ETS) sacrifices predictive power for interpretability and flexibility.

In hindsight, this realisation may have been the root of all my unrest, and my inability to unify the D&B data with traditional statistical methodologies. At its core, the ETS shares many of the same problems faced by optimisation problems.

15.2 Time capacity values, perhaps an optimisation problem?

Within the fields of regression and machine learning, the term "overfitting" is sure to raise eyebrows: and is immediately akin to "bad" and "unreliable" models. But in mathematical optimisation? - this is simply "by design".

To clarify- a conventional data-driven model is constructed using the combination of "signal" and "noise". The signal is the true effect of each variable. While the noise is mostly a result of undefined (missing) variables, or measurement errors. For traditional forecasting models (such as regression and machine learning), the sweet spot is somewhere in-between. Accordingly, an *overfit* model might be described as a one that incorporates a high degree of noise. Convention therefore implies, that an overfit-model may be unreliable when tested on new data - as this integrated "noise" is only specific to the training data.

However - returning to the NoTCoS - this melding of signal and noise, is already an integral part of its ETS. The very principles behind the ETS consider the lost-time factor as simply an extension of the overall construction process. "Lost-time" is after all, entirely dependent on the major construction tasks (Zare and Bruland 2006). Although this factor will ultimately inflate the estimated time capacity values, the NoTCoS assumes this to be self-correcting over time.

Fundamentally, regression-based models incorrectly label this "lost-time" as noise, and attribute the errors to it. Mathematical optimisation, on the other hand, does not make such a distinction - which in the end, may be advantageous to solving NoTCoS's time capacity value problem. In this master thesis, a hypothesis was made:

that the NNLS prediction errors were mostly caused by an "unaccounted-for variable": to be exact, the "lost-time" component. If this notion is indeed well-founded, a model that combines signal ("the true performance rate") with noise ("lost-time") is therefore potentially useful for estimating time capacity values for the ETS.

Mathematical optimisation is unable to generalise, but that's okay

In mathematical optimisation, a model is created to explain how a unique system of equations behaves. With such a method, the system is considered *complete*, and therefore, the algorithm is unable to distinguish between noise and signal. These elements are instead treated as one and the same, and both deemed "true effects" (signals). As a result, the solution is entirely project-specific. Such a method is able to create a precise fit to the training data, but may falter when exposed to unseen data (Bzdok et al. 2018).

However, the inability to generalise does not automatically render a model ineffective. So long as the data quantity is sufficient, the model can still function as a useful time scheduling tool for the NoTCoS.

15.3 What are the major sources of error in a NNLS model?

In this report, two major sources of error, "reducible" and "irreducible" errors, were simulated and studied at varying magnitudes. Initial findings suggest that overall, the source of error is due to a combination of two factors:

- **Reducible errors:** the productive time spent at the tunnel face is unknown (missing data)
- **Irreducible errors:** The *actual* construction rate is variable (non-constant)

This will be discussed in the sections below.

15.3.1 Reducible errors

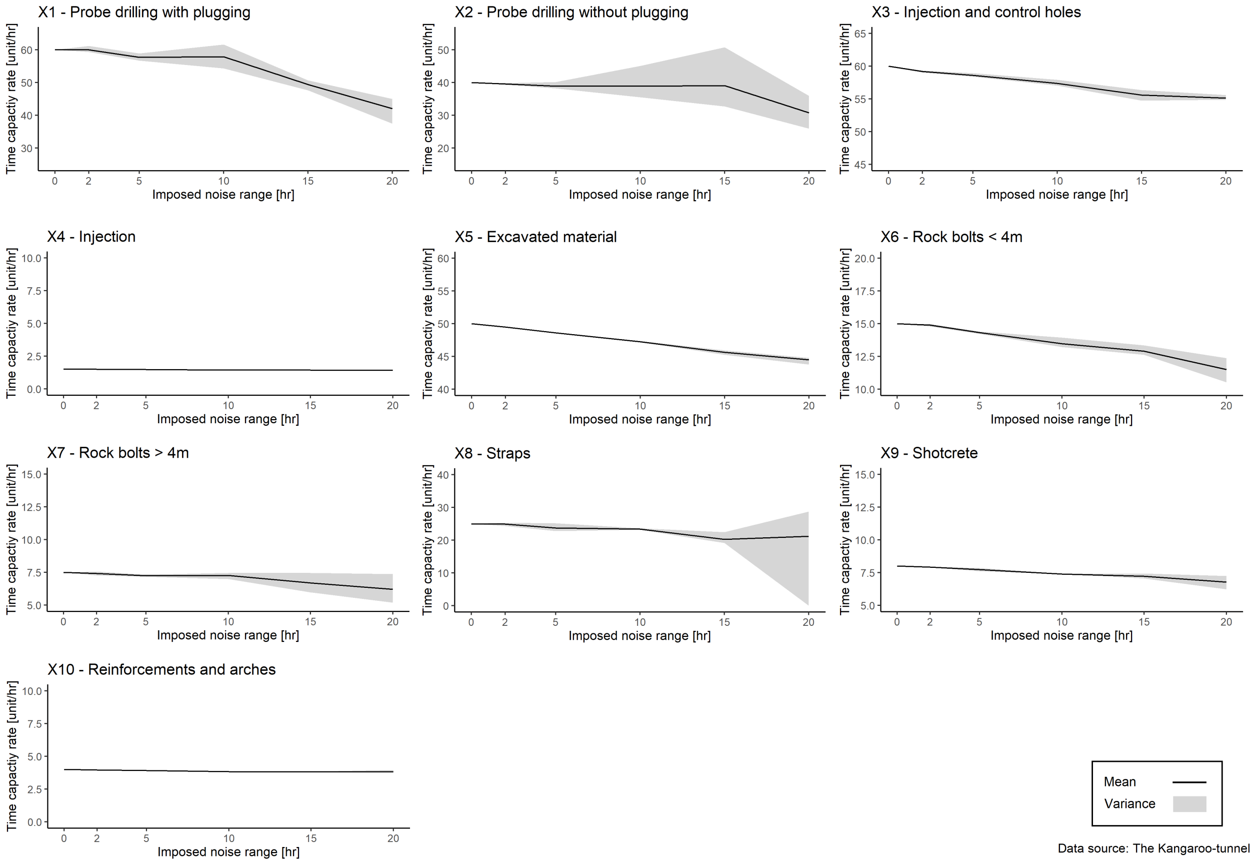
The first major source of errors appears to stem from the fact that not all time consumption variables are accounted for. The two main drivers for this is:

- The existence of lost-time (all forms)
- The exact time spent at the tunnel face is unknown (Y')

Thankfully, because this type of error is considered a "reducible" error, its distortion effects can be mitigated by improving the data quality.

The existence of lost-time

Although I am not entirely sure just how much lost-time actually exists in a *real* tunnel project, we can see in Figure 15.1, that the reliability of the model begins to wane as the noise range approaches "0 to 15 hours". In those models (and beyond), not only are the overall predictions *under*-estimated, but the variance of these predictions also appear to increase significantly. This will be discussed in-depth further below (Section 15.4).



Data source: The Kangaroo-tunnel

Figure 15.1: Estimated time capacity values against the magnitude of imposed noise plot

The exact time spent at the tunnel face is unknown

There is strong evidence to suggest that the currently-available construction logs (such as BoQs) do not provide a complete picture of all time consumption occurring at the tunnel face. The reasons are discussed below.

The weekly work-hours is not, as once-believed, a *constant* 101 hours. During the analysis of the Kongsberg-tunnel dataset, it was hypothesised that substantial "time spent away from the tunnel face" was occurring on a weekly-basis. (Section 12.1.2). Sources of this may have included coinciding-Norwegian public holidays, or occasionally, the relocation of resources to other parts of the tunnel. Although difficult to confirm, the latter trend was observable by comparing BoQs at the tunnel face and those of secondary tunnels (such as connection tunnels, and technical rooms). All in all, when such distinction were made, the model, unsurprisingly, improved in prediction prowess across the board.

Should the project experience, on average, a one day deficit per week, this can quickly amount to 15-20 hours of "noise" (or lost-time). This number may appear extreme initially, but once you factor in holidays, and unforeseen logistic hiccups, the scenario appears more and more realistic. Regardless of the tendency:

> The total time spent at the tunnel face is not clearly discernible from basic BoQs.

As a consequence, tested-models have so far consistently *under*-estimated the performance rate of several construction tasks. Figure 15.2 further supports this claim.

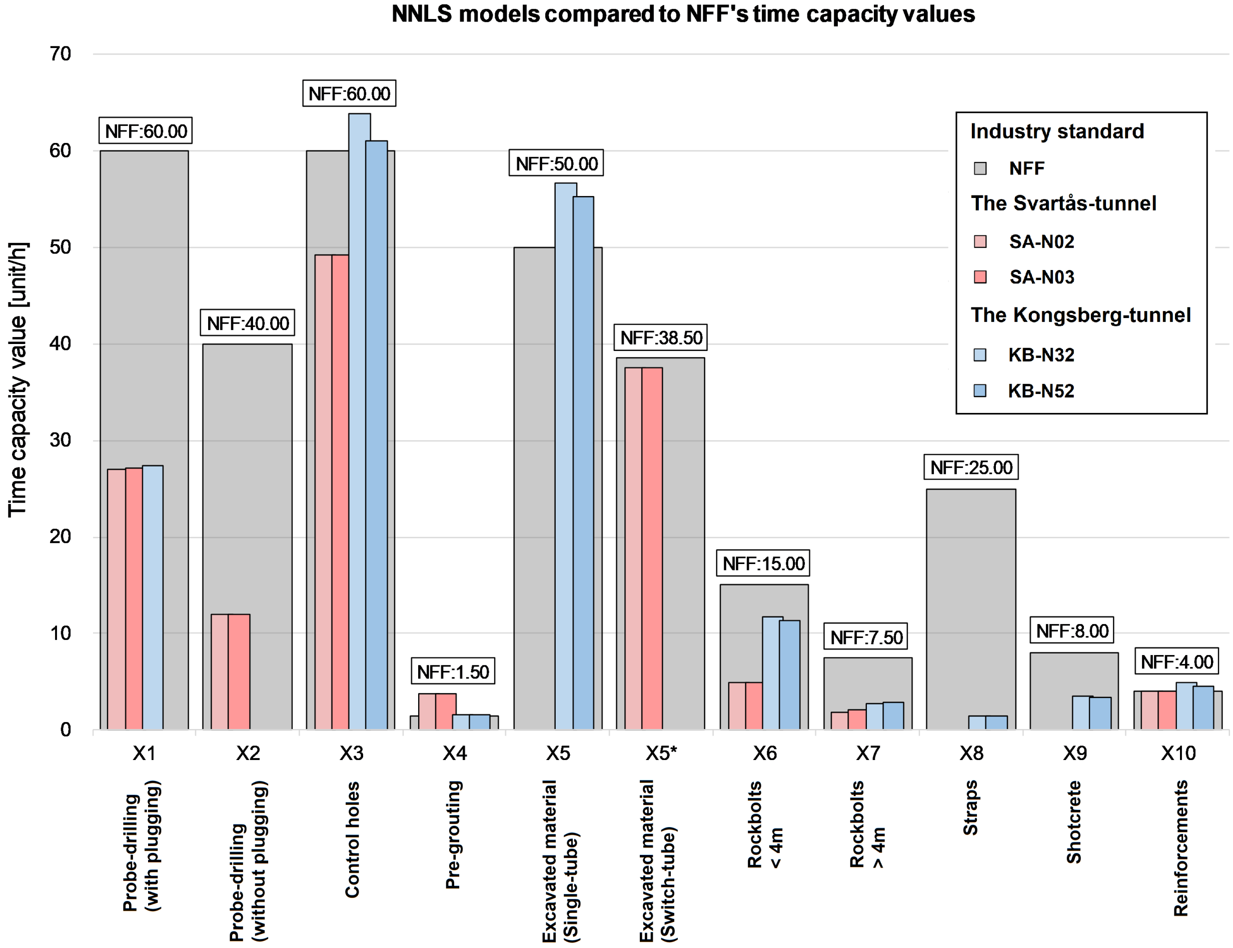


Figure 15.2: NNLS models compared to NFF's time capacity value

Nonetheless- this setback can be rectified should mitigating efforts be made to the data collection process in the future. For example, a straightforward "time at the tunnel face" measurement could suffice. This will be echoed in the following recommendations component of this report (Chapter 16). This notion is further supported by comparing current industry standards to the *real* tunnel databases.

Current NFF time capacity values only account for 70 to 80% of the workweek During the analysis of the Svartås- and Kongsberg-tunnel data (11.3 and 12.4), time capacity values from NFF, were also against these *real* datasets. Interestingly, these estimates were only able to explain, on average, 70 to 80% of the construction duration, despite being well-regarded as the industry standard amongst the general Norwegian tunnelling industry (NFF 2019). Table 15.1 and Figure 15.3 further illustrate this point.

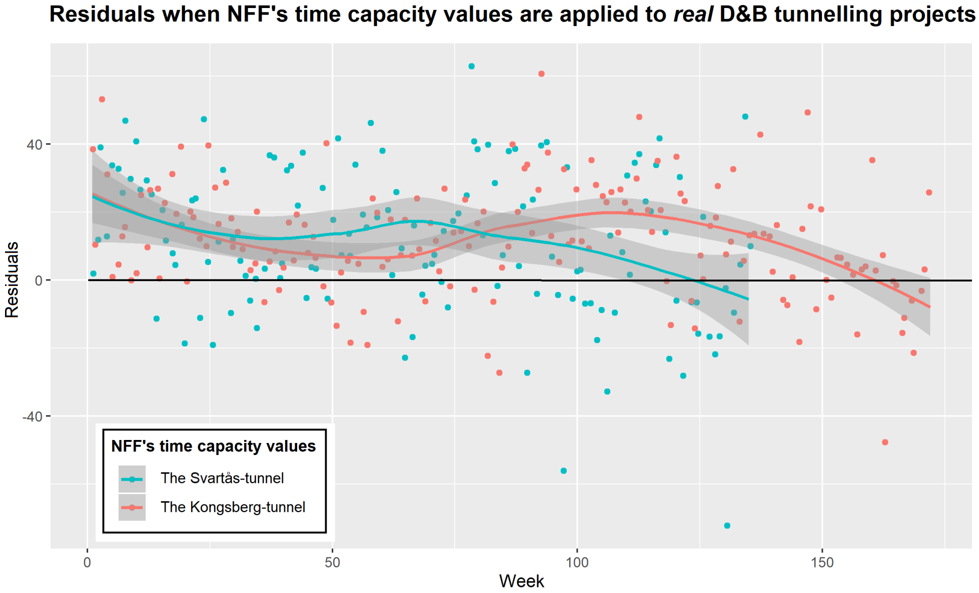


Figure 15.3: NFF's time capacity values applied to *real* tunnel projects

Table 15.1: NFF time capacity values applied to real D&B tunnel projects

Dataset description										
Model ID		Model		workweek Accounted for			Response Y			Observations n
SA-N3-NFF		NFF		0.828			Y'			135
KB-N31-NFF		NFF		0.729			Y'			172
KB-N32-NFF		NFF		0.798			Y'			172
Model outputs: time capacity values										
	X_1 m/h	X_2 m/h	X_3 m/h	X_4 t/h	X_5 m^3/h	X_6 unit/h	X_7 unit/h	X_8 m/h	X_9 m^3/h	X_{10} m/h
SA-N3-NFF	60.00	-	60.00	1.50	38.46	15.00	7.50	25.00	8.00	4.00
KB-N31-*NFF*	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00
KB-N32-*NFF*	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

15.3.2 Irreducible errors

Irreducible errors are caused by the algorithm and the model itself; and are not influenced by the data quality. As such, these errors exist in almost all models. In the case of D&B tunnelling:

> Irreducible errors are mostly attributed to the fact that the *actual* performance rate is *non-linear* (or constant), and instead *variable*.

Large variances in performance rate interferes with the model As the variance range of the performance rate becomes larger, it becomes increasingly difficult to obtain a realistic prediction estimate.

To explain why: a good place to start, is to first *visualise* the effects of "variance". However- to paint all ten dimensions of the Koala-tunnel in a 2D space is quite fruitless. A lower-dimensional version instead has been simulated for this demonstration. Figure 15.4 represents a 3-dimensional version of the Koala-tunnel. In this graphic, we can observe that as we increase the possible deviation range of the performance rate, we increase the degrees of freedom as well. As such, "outlier" observations become more and more common in the dataset. Coupled with a small sample size, the algorithm quickly (and incorrectly) classifies these "outliers" as the norm: since it is easily swayed by these "outlier" observations. Finally, without a sufficient sample size, the model is unable to correct itself back to actuality. More often than not, this results in an unreliable model.

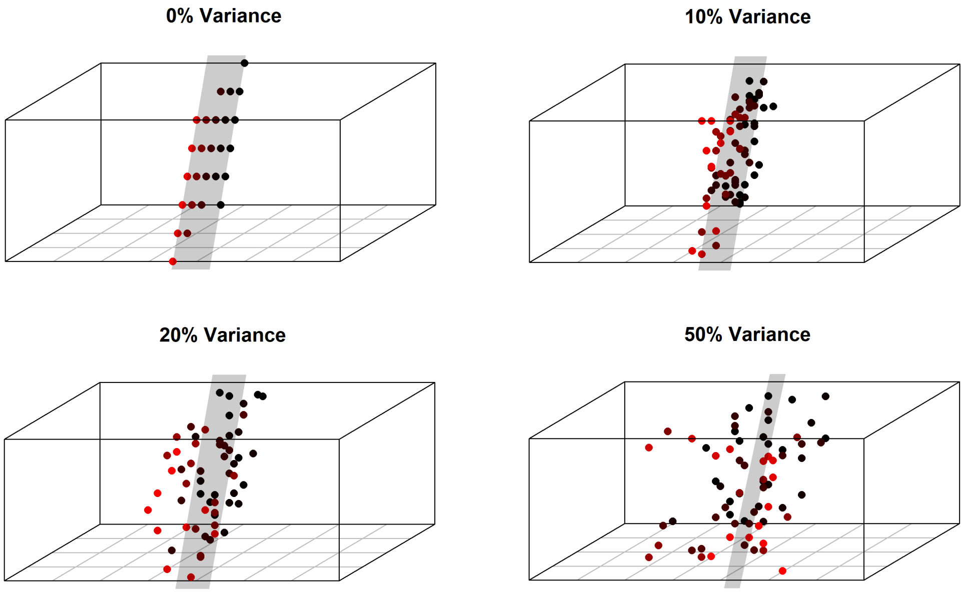


Figure 15.4: Variability in the performance rate greatly effects model estimates

The non-negative least squares algorithm (like most predictive analysis techniques) is extremely sensitive to outlier observations. This is especially undesirable when the sample size is insufficient compared to the number of dimensions.

15.4 Overall, how well does the NNLS model perform?

In the previous sections, we reiterated on the major sources of errors to be attributed to a combination of "missing data" and "non-linear performance rates". However, a back-analysis of the NNLS (Chapter 2) perhaps reveals that, despite the presence of these errors, a NNLS-based model may in fact still produce acceptable time capacity values.

15.4.1 Examining the effects of "missing data"

In the first portion of the primary analysis, a hypothetical (The Kangaroo-tunnel) tunnel was simulated with varying degrees of lost-time (unaccounted for time). Initial results suggested that, even in the presence of "missing data", the NNLS algorithm was still able to produce *acceptable* time capacity values. Table 15.2 illustrates this point. In model series KR3, the dataset experienced 0 to 10 hours of imposed "noise" (which equates to, on average, 5 hours of "missing data" per week). Despite this, the model's prediction only deviated by (on average) 5% from the *actual* performance rate.

Furthermore - up until this point - the model errors (caused by the noise), appear to be distributed quite evenly across all 10 dimensions. This was visualised in Figure 15.1 above.

Table 15.2: Kangaroo-tunnel KR30 (0 to 10 hr noise) dataset description and model outputs

Dataset description						
Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR31	0 to 10	241	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR32	0 to 10	926	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR33	0 to 10	4669	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Model outputs: time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR31	61.59	45.13	57.93	1.44	47.38	13.22	6.98	23.16	7.38	3.78
KR32	57.69	36.08	57.02	1.44	47.16	13.93	7.44	23.47	7.38	3.82
KR33	54.33	35.50	57.22	1.44	47.16	13.27	7.36	23.60	7.40	3.86
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

15.4.2 Examining the effects of "a variable construction rate"

In the primary analysis, a hypothesis was surmised to investigate the effects a variable performance rate will have on the NNLS model performance. To address this, a tunnel was simulated using variable construction rates: the Koala-tunnel. The objective was to assess the impact of the irreducible error.

In the first model hypothetical scenario, the Koala-tunnel experienced up to 10% variance in its construction rate. Despite such fluctuations, the model was able to produce reasonable estimates. As shown in Table 15.3, predicted time capacity values, on average, only strayed approximately 3% from the *real* (NFF) performance rates.

Table 15.3: Koala-tunnel KO10 (10% variance) dataset description and model outputs

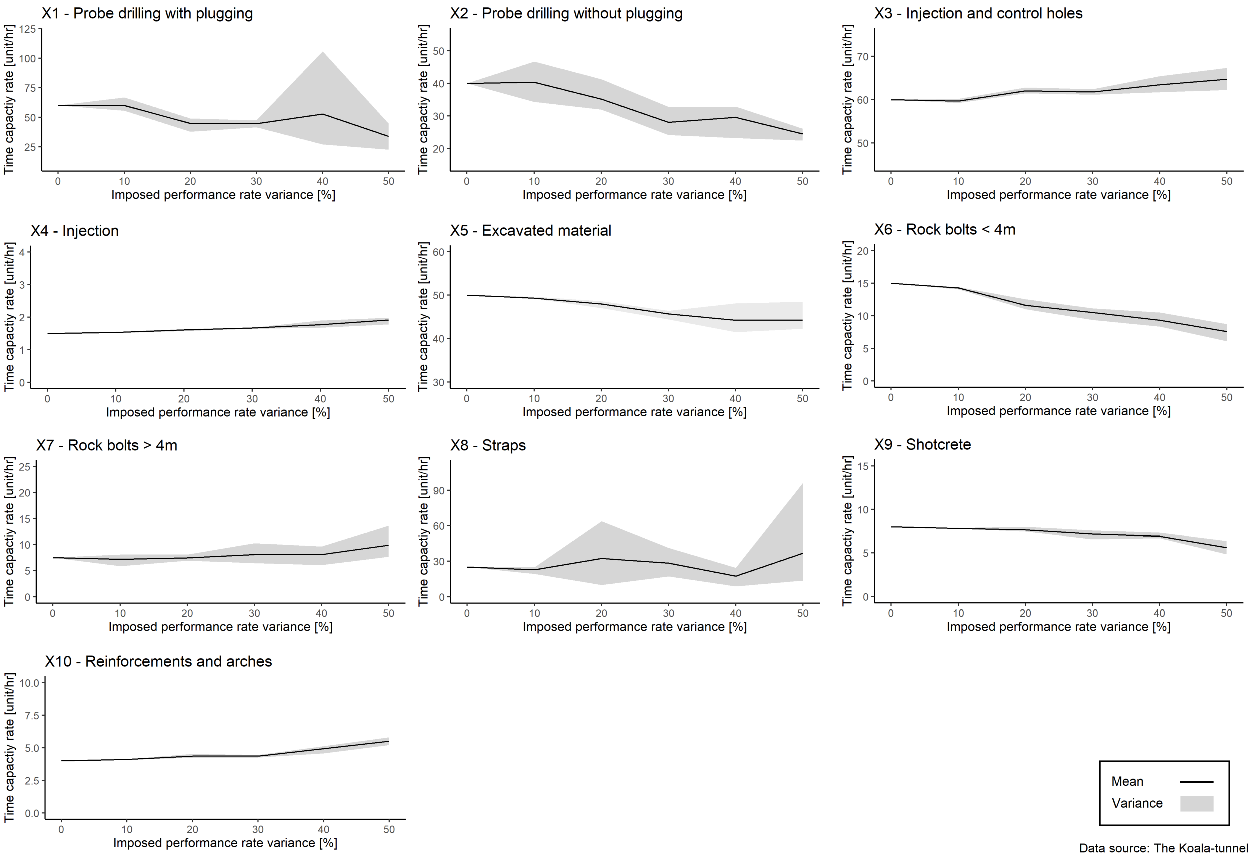
Dataset description

Model ID	Variance [%]	Observations	Response [hr]	Dimensions	Predictors	Weights
	d	n	Y	k	X	β
KO11	10	179	100 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO12	10	715	95 to 102	10	$X_1 \dots X_{10}$	NFF, variable
KO13	10	3703	80 to 105	10	$X_1 \dots X_{10}$	NFF, variable

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO11	66.83	34.34	59.22	1.55	49.17	14.33	5.85	19.25	7.87	4.09
KO12	58.04	46.67	59.62	1.52	49.24	14.39	8.11	24.34	7.79	4.13
KO13	55.66	39.89	60.15	1.52	49.65	14.09	7.58	24.78	7.82	4.08
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

However, as soon as the variance exceeded 10%, the model performance drastically decreased. As illustrated in Figure 15.5, increased magnitudes of variance quickly pushed the beta coefficients away from their *actual* values. Furthermore, the variance in the prediction estimates became more volatile as the performance rate variance increased. Specifically notable in the probe drilling (X_1 and X_2) and straps X_8 components, where the estimates became unstable - and frankly unusable.



Data source: The Koala-tunnel

Figure 15.5: Estimated time capacity values against the magnitude of performance rate variance

15.4.3 A *real* tunnel project features both, lost-time *and* variability

In the previous scenarios, "lost-time" and "variable performance rates", were isolated and analysed as mutually exclusive elements. In reality however, a *real* tunnel project probably features by a mixture of both. Because it is difficult to determine the exact magnitude of its presence, I postulated my own estimates. For this thought-exercise, perhaps, an average scenario would entail the following characteristics (Model ID: KO99):

- 0 to 2 hours of weekly missing data (lost-time)
- 10% variance in the performance rate
- All drilling elements combined (X_{123})

Despite these obstacles, the Koala-tunnel model KO99 was surprisingly able to produce quite promising results. These are shown in Table 15.4)

Table 15.4: Koala-tunnel KO99 (0 to 2 hours noise and 10% variance) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Variance [%]	Observations	Response [hr]	Predictors	Weights
	ϵ	d	n	Y	X	β
KO99	0 to 2	10	1549	95 to 102	$X_{123} \ldots X_{10}$	NFF, variable

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KO99	-	-	62.40	1.45	49.98	10.19	7.79	17.62	6.67	4.26
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Yes, some construction tasks were notably underestimated. But all in all, the predicted performance rates did not stray too far from NFF's time capacity values. Overall prediction error differed on average by roughly 10%. Figure 15.6 further hones in on this point.

The research conducted here is anything but concrete or conclusive. But such results are a step in the right direction. However due to time constraints, I was unable to pursue this idea extensively. This predicament and suggestions for further work will be discussed in the concluding chapters.

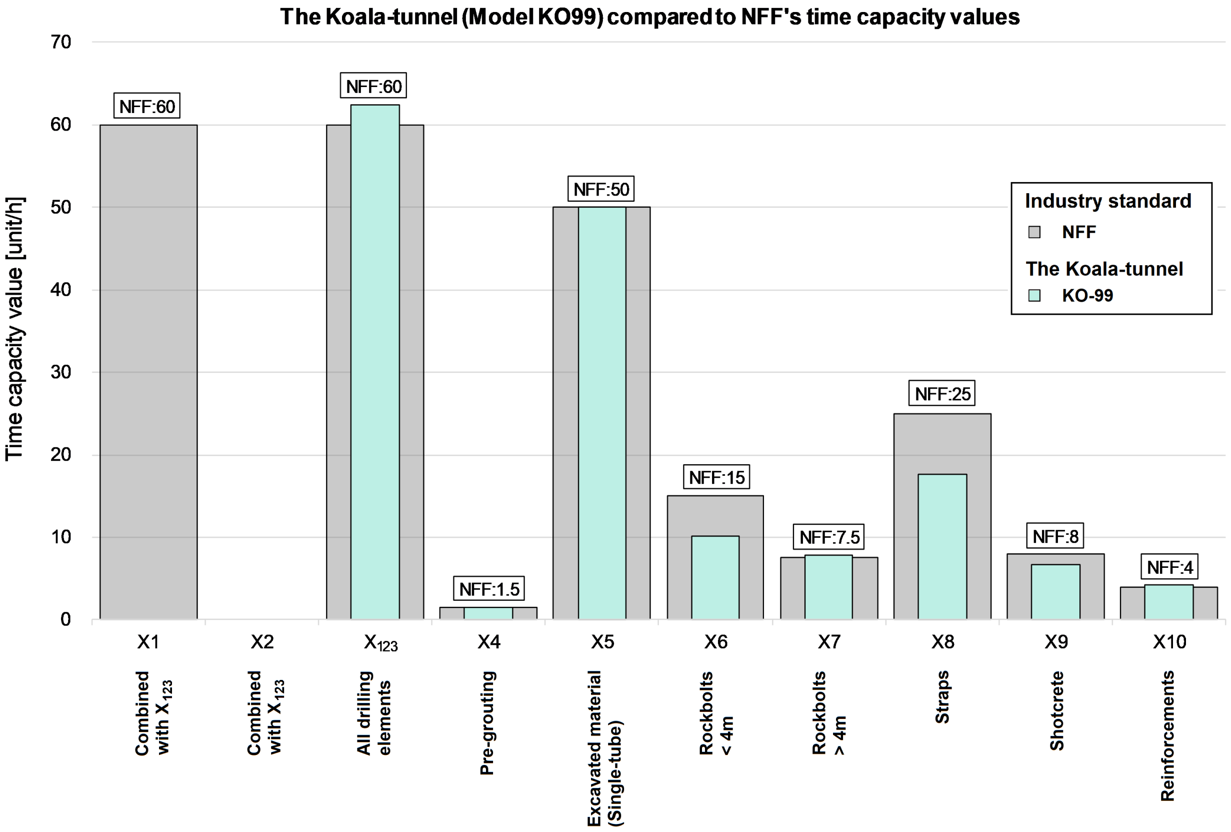


Figure 15.6: Koala-tunnel (model KO99) compared to NFF's time capacity values

15.5 Problems left unsolved

Whether due to time constraints; or the perhaps simply, a lack of competence- the problems left unsolved have been compiled and presented below. Thereafter, potential-pathways to their solutions, are promoted as a future topic of research in the next chapter (Chapter 16 - Recommendations).

Noise is evenly distributed, but overall prediction error is not

In the diagnostics (Section 14.5), we examined ***how*** the errors and beta coefficients were being distributed when employing a NNLS algorithm. Initial impressions - absolutely based on whim and intuition - is that these prediction errors are strongly correlated to the "error distribution value" (EDV). This idea is illustrated in Figure 15.7.

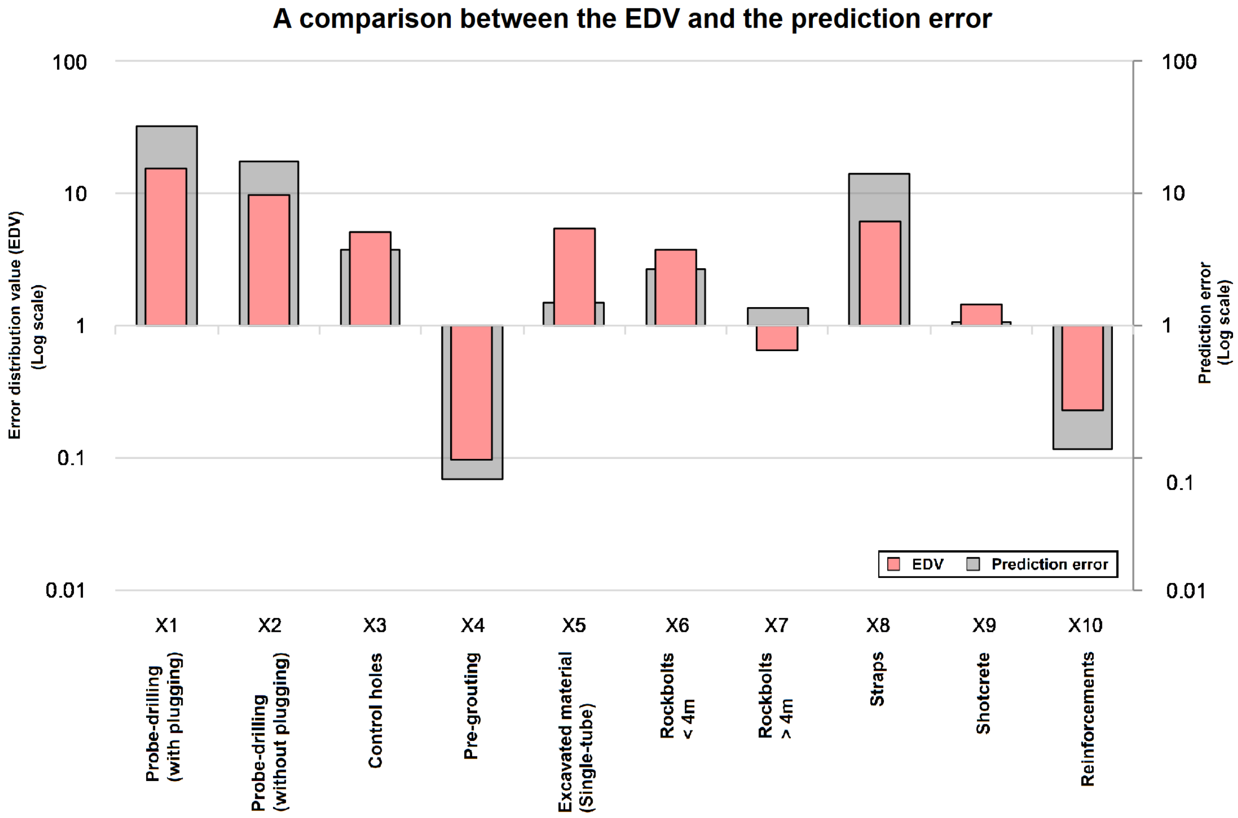


Figure 15.7: A comparison between the EDV and the prediction error

As a reminder, the EDV is closely connected to the variable's average time consumed (when it occurs), and the amount of noise in the dataset. Overall, the average time consumed is based on the variable's:

- Sample distribution and its skewness; and
- *Actual* influence (performance rate)

To clarify: when there is "noise" present in the dataset, the NNLS algorithm must inflate the input variables to account for this "additional time". However, for construction tasks that do not contribute significantly to the total weekly-time, large amounts of noise can quickly overshadow the real signal.

Table 15.5 shows the time contribution of each construction task (based on *real* tunnel data). Here, the probe drilling and straps variables (X_1, X_2 and X_8) contribute, on average, only sub-2 hours, per week. Therefore, even 1 hour of excess "noise", per week, will ultimately inflate the "true" performance rate by as much as 50%. Hence why some variables appear to disproportionately take on more of the prediction error than others.

Table 15.5: The average time consumed by each variable per week

Average time consumed per week										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	h	h	h	h	h	h	h	h	h	h
KR53	2.2	2.6	17.8	23.2	38.5	6.3	6.0	2.0	8.6	33.5

Although there is (not yet) any scientific backing to this claim, the investigation performed in the diagnostics section was intended to provide a "starting point" for future research. Nonetheless, I believe these initial findings do pass general inquest and scrutiny.

15.6 All in all, how can the model be improved?

Now that the origins of the prediction error; as well as their biggest culprits, have been singled out, the focus can now shift to more positive topics. How do we mitigate these errors, and how do we improve the performance and reliability of the model?

Reducible errors

To minimise the reducible errors in a model, the weekly "productive time spent at the tunnel face" must be accurately measured.

> A sweeping "constant 101 hours workweek" will just not cut it - and is insufficient for reliable modelling in the future.

In this report, some of the biggest reasons for "time spent away from the tunnel face" were identified and documented. The information collected here, thereby coincides precisely with the areas which future data collection can be improved. These are discussed in the next chapter (Chapter 16 - Recommendations).

Irreducible errors

In the case of D&B tunnelling data, irreducible errors are caused by heavily skewed distributions and a variable construction rate. However, these elements, and their disruptive effects, each require their own unique remedies. These have been summarised in Table 15.6 and discussed below.

Table 15.6: Methods to address irreducible errors

Irreducible error source	**Solution**
Construction tasks with a heavily skew distribution	Data transformation
A highly variable construction rate	• Increasing the sample size • Selecting a different model

These realisations are undoubtedly a wrench in the works. Although it may be implausible for such high levels of variance to occur in the real-world - and across no-less than each of the ten construction tasks simultaneously - *and* consistently over the entire duration of the project, the NNLS algorithm's inability to reliably model such a scenario is telling. This obstacle - if unsolved - may very well be the demise of an optimisation-based solution. But it is difficult to tell at this point. Nonetheless, these will be elaborated below, but future actions will be recommended in the subsequent chapter (Chapter 16).

Large variability in the *actual* performance rate requires an even larger sample size. To visualise such an effect, I call upon the sport of baseball for inspiration: perhaps the most statistical-savvy sport to ever exist. Due to the sheer amount of variance present in this sport, analysts do not consider the results "true" until sufficient samples have been acquired. Figure 15.8 demonstrates how the Cronbach's Alpha or similar methods can be used to assess the critical point when the data has "stabilised". "Stabilisation" can be described as the moment when the "observed" values represent the "real" values. Sadly, techniques like this have not explored in full, and instead, are included as part of the recommendations chapter.

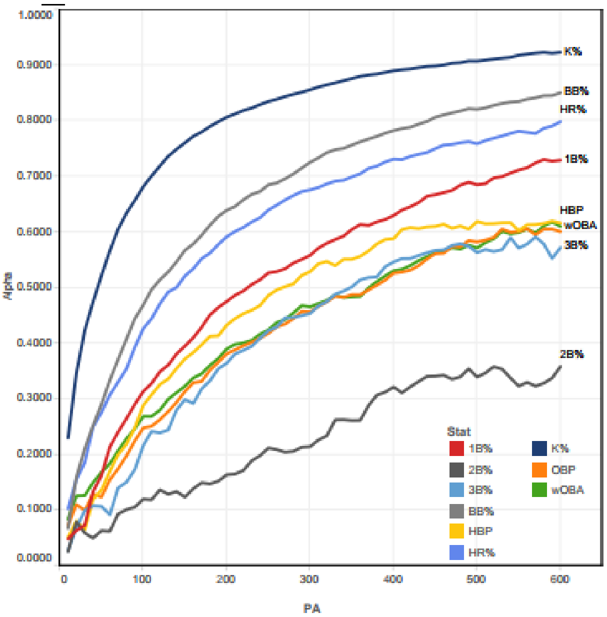


Figure 15.8: In baseball analytics, the Cronbach's Alpha is commonly employed to assess the truthfulness of a player's highly variable performance (Pemstein and Dolinar 2015)

15.7 Risk involved with this method

As with all decisions we make, the consequences and risks must be considered in tandem. In this section, I discuss how a data-driven model may shift the risk entirely to one party.

A data-driven ETS may cause an imbalance to risk sharing

As I delved deeper into this fascinating topic, I began to realise that time capacity values derived purely from construction logs may cause an imbalance in the risk sharing.

To clarify: in the previous section, there was a dialogue discussing how the presence of inevitable "noise" (such as lost-time), will inflate the overall predicted time capacity values. This time still exists after all, and must go *somewhere*. In any case, because "lost-time" is undetectable (or immeasurable) within the standard construction logs:

> The model's estimated performance rate will **always** be lower than the *actual* rate.

It is the contractor who almost-exclusively controls the extent of this "lost-time". Aside from the major contributors mentioned in Section 15.2, these can also arise from less-obvious sources: such as bouts of low productivity, poor resource management or logistical decisions; and other unforeseen delays. These examples are not directly correlated to the major construction tasks, and are therefore true representations of "noise" and perhaps a source of irreducible error within the model. Which brings me to my point:

> In its purest form, a data-driven model is a reflection of the contractor's performance rate **and** their productivity.

As such, if one were to make time scheduling decisions (such as litigation and disputes) purely based on a data-driven model, the owner may be awarding poor performance and productivity with more-than-necessary time extensions. Conversely, the very same system may punish highly productive tunnel builder's with tighter time-extensions.

Overall, these observations are just a reminder to the inherent flaws of this system; and emphasise that data-based models should not be depended on entirely and carelessly. As with all models today, they should be used alongside other established models; and with those experienced.

15.8 Some remarks and notes to end

To conclude this chapter, I babble a little about some of the obstacles I encountered: to hopefully reveal some added-insight into this research.

The absence of other parametric-themed models in the industry

I finally appreciate why there has been a lack of "unit time"- or parametric-based contract systems and prediction models in the tunnelling industry (in fact, in the general construction industry). The NoTCoS has been uniquely a *Norwegian* practice. As of yet, such a scheme has not been documented outside of Norway (Kleivan 1989). It is no wonder- that initial search returns left me empty-handed and fumbling.

Chapter 16

Recommendations

Sadly, we have now probably reached the extent of this master thesis. Due to the constraints that time has imposed on me - on us all - I no longer have the capacity to proceed any further. This is where the research and my expertise regretfully ends. Nonetheless, in this chapter, some of the major shortcomings encountered - and left unsolved - are echoed once again. However- this time, recommendations are also proposed, in the hope that future researchers can untangle the problem I could not. A summary of the recommended works has been presented in Table 16.1.

Table 16.1: Recommendations for future work

Theme	**Recommended works**
Research-related	• Data transformation • Sufficient sample size
Data collection-related	• Additional data required • Standardising the process

16.1 Improving the reliability and performance of the model

In this section, I make some suggestions to future research that may improve the reliability and effectiveness of a data-driven method.

16.1.1 More research into "data transformation" is sorely needed

Issues relating to the disproportionate distribution of prediction errors were encountered and presented in this report. To reiterate: when the "noise" or the "variation in the performance rate" becomes too large, input variables with non-normal distributions will be the first to falter in a data-driven model. This realisation has become a critical element of the time capacity value problem. Going forward, this is most likely the biggest barrier impeding a successful data-based prediction model.

Therefore, given another opportunity, the next course of action would be to first investigate and apply data transformation techniques to variables with a high "error distribution value" (EDV)[1] (such as probe drilling and straps). Preliminary research suggests that "bootstrapping" (Kulesa et al. 2015)) would be great place to begin. Supplementary to this, "inverse transformation" may be useful for the more extreme positive skewed distributions (Emery and Ortiz 2005; H. Wang et al. 2016), as shown in Figure 16.1.

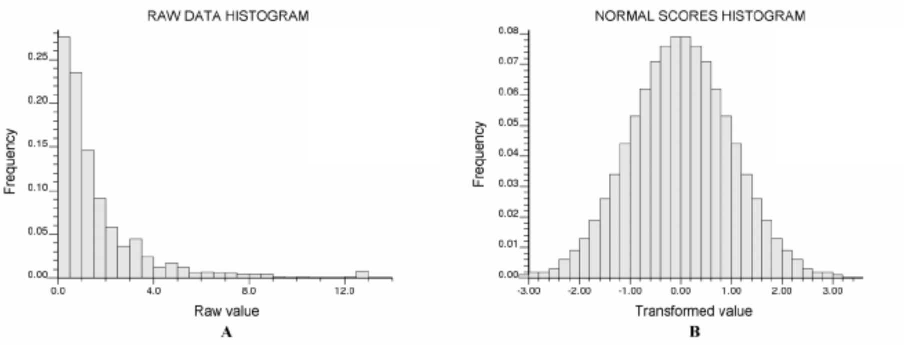


Figure 16.1: An example "positively skewed distribution" dataset (Left, A), is normalised (Right, B) by inverse transformation functions (Figure adopted from (Emery and Ortiz 2005)

16.1.2 What is the minimum sample size needed before predictions are meaningful?

Throughout this research, I investigated many different tunnel projects. All with varying sample sizes and input parameters. Each time, I was left with unique prediction estimates. Which begs the question:

> *How do we distinguish between a meaningful estimate, and one that is not?*

At the end of the day, we are looking to determine whether the samples we have are meaningful, or not. Therefore, in future studies, it will be interesting to investigate the "critical" sample size required, before predictions can be trusted as a realistic representation of the true underlying effects. A good place to begin include:

- Cronbach's Alpha
- Confidence Intervals

These are both designed to assess the reliability of the predictions: to assign the likelihood that they are a reflection of reality. Oftentimes, when the sample size is low, the "true" signal is not given sufficient observations for it to stabilise.

[1] The "error distribution value" (EDV) is described in Section 14.5

16.2 Proposed changes to the data collection

All in all, research shows that it is indeed viable to derive time capacity values using a data-driven approach. However, the effectiveness of this methodology will be greatly dependent on the quality and completeness of data collected. In this section, the deficiencies of the currently-available construction data (bill of quantities (BoQs)) are outlined. Thereafter, some changes are proposed to the data collection process. The items discussed include:

- The "completeness" of the data collected;
- Improving the quality of the data collected; and
- Standardising the process.

16.2.1 The weekly "time spent at the tunnel face" is unknown

In this report, it was suggested that the weekly "time spent at the tunnel face" is currently unknown. In its current state, this element is not easily or reliably discernible from traditional BoQs (Section 15.3.1). If unchecked, its presence is indeed a major source of prediction error. However- this shortcoming is reparable. As a "reducible error", its effects can be diminished by simply improving the data quality and completeness.

I am unsure how to implement such a measurement in real-world scenarios, or to derive its value reliably. Perhaps a rough-"stop-watch"-estimate of the "productive time spent at the tunnel face" will be suffice. But then again, such a meager proposal may quickly spur on another (cheeky) suggestions:

> *"Why not just measure the duration of each individual task then?"*

Well- they are not wrong. Though such a solution, admittedly, renders the "data-driven" approach slightly moot. Nonetheless, the harmful effects of this "missing data" is real- but its bearing is just as easily reducible should additional data be collected in the future.

16.2.2 Discerning probe drilling tasks with- and without- plugging

During the analysis of *real* tunnel projects, it was often difficult to identify probe drilling tasks *with* a plugging component, and those *without*. When speaking with the experts, it was often stated, that the performance rate between the two tasks is not trivial: and can differ by as much as 33%. Therefore, collecting additional information to better-differentiate between the two unique tasks will greatly enhance the overall performance of a data-driven model.

16.2.3 Standardising the data collected

As of now, BoQs possess:

- A lack of standardisation: inconsistency with the tasks recorded, units used, and naming convention
- Limited accessibility

Standardised format The process of converting raw construction logs (BoQ) into a usable dataset was without a doubt extremely time-consuming; and easily prone to input-error (due to the monotonous nature of the task). In the future, if there is an ambition to continue this data-themed research, it would be highly beneficial to establish a standardised records keeping process. This includes consistency between the types of items recorded; units used; and naming conventions. This will greatly reduce the bottleneck caused by excessive data cleaning and wrangling.

Data Accessibility At the same time, it would be convenient if future data-analysts be privy to more user-friendly formats of the collected data. Currently, D&B construction logs are provided to researchers in either .PDF or .JPEG (image) format. Instead, .XLS (excel) or .CSV (text) files, are undoubtedly preferred.

Chapter 17

Conclusion

The outcome of this study was ultimately a mixed bag: both objectively and emotionally - and described only as turbulent. The forward momentum gained in *one* model was quickly rendered moot, when failures from *another* had been exposed. The apparent successes experienced *one* day, oftentimes could not be replicated the very *next* day. And round like that it went.

Nonetheless, preliminary research conducted here alludes to the non-negative least squares (NNLS) algorithm as a useful predictive tool for extracting D&B time capacity values from construction data. However- its success is utterly dependent on a strict set of conditions: if unmet, the consequences are quickly hampering.

Condition 1: The *actual* time spent at the primary tunnel face must be accounted for.

The first condition is concerned with the "reducible errors": where its source is largely due to missing or incomplete data. As of now, the total weekly time spent at the tunnel face is not (reliably) discernible from basic construction logs. Yes- the current assumption, that a full working-week is comprised of 101 hours, still holds true. But such a notion does not capture the "time *away* from the primary tunnel face". Factors such as the construction of secondary tunnels; waiting-time; and even days-off, all inevitably reduce the effective working time. Without a clear depiction of the total construction time, a data-driven model will incorrectly inflate its overall estimated time capacity values. However- I trust that future endeavours will look to rectify this "missing data".

Condition 2: Large variances in the construction rate demand an even larger sample size.

The second condition pertains to the "irreducible errors": remedied only by data transformation; by collecting more data; or by changing the prediction model. In the case of D&B tunnelling data, these errors stem from the fact that the *actual* construction rate is *variable* on week-to-week basis.

Although it is difficult to gauge just how much variance exists in a *real* tunnel project, its magnitude is directly related to the minimum samples required, before a NNLS-based model is able to capture the "true" performance rate. In this study, the NNLS algorithm was demonstrated to derive realistic time capacity values, despite a 10% variance. However- the model quickly falters when deviations exceeded 20% (though estimates were corrected, when exponentially-more data samples were introduced).

Perhaps these conditions will never be fully realised or even feasible. Despite this, the study performed here, has been a step in the right direction. In this report, the deficiencies of the currently-available data were exposed: information invaluable for future data collection. Furthermore, sensitivity analysis revealed just *how* much variance is *too* much variance, before the model performance begins to wane. Lastly, findings reveal that the prediction errors are disproportionately distributed in a NNLS-based model; and that they are controlled by the "error prediction value" (EDV). Understandably, these findings do not provide a direct solution to the time capacity problem. Though I suppose they will instead shepherd future researchers towards the precise obstacles that still stand in the way.

All in all, we are on a very good way to coupling data-based solutions with time scheduling decisions in the Norwegian tunnel industry. (Or perhaps even closer, should the variance not be as extreme as I had presumed). Before closing, I would certainly be remiss if I did not mention George Box's famous statement at least once— that: *"All models are wrong, but some are useful"*.

A very candid way to wrap things up, if you ask me.

Bibliography

Academic-Skills (2013). *Reviewing the Literature*. Unpublished Work. Melbourne, Australia. URL: https://unimelb.libguides.com/lit_reviews.

Anscombe, Francis J (1973). "Graphs in statistical analysis". In: *The American Statistician* 27.1, pp. 17–21. ISSN: 0003-1305.

Barrett, James P. (1974). "The Coefficient of Determination-Some Limitations". In: *The American Statistician* 28.1, pp. 19–20. ISSN: 00031305. DOI: 10.2307/2683523. URL: www.jstor.org/stable/2683523.

Bogfjellmo, Geir (2019). *TMA4180 Optimization I*. Lecture notes. URL: https://wiki.math.ntnu.no/tma4180/2020v/start.

Bratko, Ivan (1997). "Machine learning: Between accuracy and interpretability". In: *Learning, networks and statistics*. Springer, pp. 163–177.

Bro, Rasmus and Sijmen De Jong (1997). "A fast non-negativity-constrained least squares algorithm". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 11.5, pp. 393–401.

Bruland, Amund (2000). *Hard rock tunnel boring*. Fakultet for ingeniørvitenskap og teknologi.

— (2018). *Lecuture notes: Hard Rock Tunnelling 2018 3 Time Consumption*.

Bryant, Peter G. (1996). "MDL for Mixtures of Normal Distributions". In: *Data Analysis and Information Systems*. Ed. by Hans-Hermann Bock and Wolfgang Polasek. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 3–14. ISBN: 978-3-642-80098-6.

Bzdok, Danilo, Naomi Altman, and Martin Krzywinski (2018). *Points of significance: statistics versus machine learning*.

Casella, George (1983). "Leverage and Regression Through the Origin". In: *American Statistician - AMER STATIST* 37, pp. 147–152. DOI: 10.1080/00031305.1983.10482728.

Chapman, Pete et al. (1999). "The CRISP-DM user guide". In: *4th CRISP-DM SIG Workshop in Brussels in March*. Vol. 1999.

Chen, Min, Shiwen Mao, and Yunhao Liu (2014). "Big data: A survey". In: *Mobile networks and applications* 19.2, pp. 171–209.

Conway, Drew (2010). *The data science venn diagram*. URL: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram (visited on 09/30/2010).

Crowdflower (2015). *Crowdflower Data Scientist Report 2015*. Report, p. 7. URL: https://visit.figure-eight.com/rs/416-ZBE-142/images/Crowdflower_Data_Scientist_Survey2015.pdf.

— (2016). *Crowdflower Data Scientist Report 2016*. Report, p. 6. URL: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf.

Czarnigowska, Agata and Anna Sobotka (2014). "Estimating Construction Duration for Public Roads During the Preplanning Phase". In: *Journal of Engineering, Project, and Production Management* 4.1, pp. 26–35. ISSN: 2223-8379. URL: http://www.AiritiLibrary.com/Publication/Index/22238379-201401-201502140006-201502140006-26-35.

Draper, Norman R and Harry Smith (1998). *Applied regression analysis*. Vol. 326. John Wiley & Sons, pp. 245–246.

Dupont, William D and Walton D Plummer Jr (1998). "Power and sample size calculations for studies involving linear regression". In: *Controlled clinical trials* 19.6, pp. 589–601. ISSN: 0197-2456.

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press. ISBN: 0412042312.

Eisenhauer, Joseph G. (2003). "Regression through the Origin". In: *Teaching Statistics* 25.3, pp. 76–80. ISSN: 0141-982X. DOI: 10.1111/1467-9639.00136. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9639.00136.

Emery, Xavier and Julián Ortiz (Feb. 2005). "Histogram and variogram inference in the multigaussian model". In: *Stochastic Environmental Research and Risk Assessment* 19, pp. 48–58. DOI: 10.1007/s00477-004-0205-5.

Fahrmeir, Ludwig et al. (2013). *Regression - Models, Methods and Applications*.

Faraway, Julian J (2016). *Linear models with R*. Chapman and Hall/CRC. ISBN: 1439887349.

Farrar, Donald E. and Robert R. Glauber (1967). "Multicollinearity in Regression Analysis: The Problem Revisited". In: *The Review of Economics and Statistics* 49.1, pp. 92–107. ISSN: 00346535, 15309142. DOI: 10.2307/1937887. URL: http://www.jstor.org/stable/1937887.

Fletcher, Roger (2013). *Practical methods of optimization*. John Wiley & Sons.

Gandomi, Amir and Murtaza Haider (2015). "Beyond the hype: Big data concepts, methods, and analytics". In: *International journal of information management* 35.2, pp. 137–144.

Gill, Philip E, Walter Murray, and Margaret H Wright (2019). *Practical optimization*. SIAM.

Grolemund, Garrett (2014). *Hands-On Programming with R*.

Grøv, Eivind (2012). "Contract Philosophy in Norwegian Tunnelling". In: *Norsk Forening for Fjellsprengningsteknikk (NFF)* Publication 21 - Contracts in Norwegian Tunnelling.21, pp. 15–20.

Hahn, Gerald J (1977). "Fitting regression models with no intercept term". In: *Journal of Quality Technology* 9.2, pp. 56–61. ISSN: 0022-4065.

Hallin, Marc (2014). "Gauss–Markov Theorem in Statistics". In: DOI: 10.1002/9781118445112.stat07536.

Harrell Jr, Frank E, Kerry L Lee, and Daniel B Mark (1996). "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Statistics in medicine* 15.4, pp. 361–387. ISSN: 0277-6715.

Harville, David (1976). “Extension of the Gauss-Markov theorem to include the estimation of random effects”. In: *The Annals of Statistics* 4.2, pp. 384–395. ISSN: 0090-5364.

Haskell, Karen H and Richard J Hanson (1981). “An algorithm for linear least squares problems with equality and nonnegativity constraints”. In: *Mathematical Programming* 21.1, pp. 98–118.

Hey, Tony, Stewart Tansley, Kristin Tolle, et al. (2009). *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA.

Hocking, Daniel J. (2011). Blog. URL: https://danieljhocking.wordpress.com/.

Hofstadler, C (2010). “Calculation of Construction Time for Building Projects–Application of the Monte Carlo Method to Determine the Period Required for Shell Construction Works”. In: *TG65 & W065-Special Track 18th CIB World Building Congress May 2010 Salford, United Kingdom*, p. 214.

Hsieh, Fushing Y, Daniel A Bloch, and Michael D Larsen (1998). “A simple method of sample size calculation for linear and logistic regression”. In: *Statistics in medicine* 17.14, pp. 1623–1634. ISSN: 0277-6715.

Irizarry, Rafael A (2019). *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*. CRC Press.

Isaksson, Therese (2002). “Model for estimation of time and cost based on risk evaluation applied on tunnel projects”. Thesis.

James, Gareth et al. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated, p. 430. ISBN: 9781461471370.

Johnsen, Morten G (2014). “History and Development”. In: *Norsk Forening for Fjellsprengningsteknikk (NFF)* Publication 5 - Tunnelling Today.5.

Kim, Yangkyun and Amund Bruland (2009). “Effect of rock mass quality on construction time in a road tunnel”. In: *Tunnelling and Underground Space Technology* 24.5, pp. 584–591. ISSN: 0886-7798. DOI: https://doi.org/10.1016/j.tust.2009.02.004. URL: http://www.sciencedirect.com/science/article/pii/S0886779809000303.

Kjell Wold (2019). *Fra Svartåstunnelen mot Trollerudmoen skal det jobbes natt neste uke*. [Online; accessed July 6, 2020]. URL: https://www.laagendalsposten.no/e134/bygg-og-anlegg/vei/mer-nattarbeid-pa-e134-anlegget-neste-uke/s/5-64-740917.

Kleivan, Erland (1989). “NoTCoS: The Norweigan tunnelling contract system”. In: *Tunnelling and Underground Space Technology* 4.1, pp. 43–45. ISSN: 0886-7798. DOI: https://doi.org/10.1016/0886-7798(89)90031-X. URL: http://www.sciencedirect.com/science/article/pii/088677988990031X.

Kostrzewa, Paulina and Magdalena Rogalska (2019). “Scheduling Construction Processes Using the Probabilistic Time Coupling Method III”. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 471. 11. IOP Publishing, p. 112072.

Kovács, Péter, Tibor Petres, and László Tóth (2005). “A new measure of multicollinearity in linear regression models”. In: *International Statistical Review* 73.3, pp. 405–412. ISSN: 0306-7734.

Kuhn, Max and Kjell Johnson (2013). *Applied predictive modeling*. Vol. 26. Springer.

Kulesa, Anthony et al. (2015). *Points of significance: sampling distributions and the bootstrap*.

Kutner, Michael H et al. (2005). *Applied linear statistical models*. Vol. 5. McGraw-Hill Irwin Boston.

Langaas, Mette and Stephanie Muff (2019). *TMA4268 Statistical Learning*. Lecture notes. URL: https://wiki.math.ntnu.no/tma4268/2020v/start.

Lawson, Charles L and Richard J Hanson (1974). *Solving Least Squares Problems*.

— (1995). *Solving least squares problems*. Vol. 15. Siam. ISBN: 0898713560.

Lin, Ming-Chiao et al. (2011). "Developing a construction-duration model based on a historical dataset for building project". In: *Journal of Civil Engineering and Management* 17.4, pp. 529–539.

Lindfield, George and John Penny (2019). "Chapter 2 - Linear Equations and Eigensystems". In: *Numerical Methods (Fourth Edition)*. Ed. by George Lindfield and John Penny. Fourth Edition. Academic Press, pp. 73–156. ISBN: 978-0-12-812256-3. DOI: https://doi.org/10.1016/B978-0-12-812256-3.00011-7. URL: http://www.sciencedirect.com/science/article/pii/B9780128122563000117.

Macias, Javier et al. (2017). *Drillability Assessments in Hard Rock*.

Mahdevari, Satar et al. (2014). "A support vector regression model for predicting tunnel boring machine penetration rates". In: *International Journal of Rock Mechanics and Mining Sciences* 72, pp. 214–229. ISSN: 1365-1609. DOI: https://doi.org/10.1016/j.ijrmms.2014.09.012. URL: http://www.sciencedirect.com/science/article/pii/S1365160914002536.

MathWorks (2019a). Web Page. URL: https://www.mathworks.com/help/matlab/.

— (2019b). Web Page. URL: https://nl.mathworks.com/matlabcentral/fileexchange/.

Min, Sangyoon (2008). "Development of the resource model for the decision aids for tunneling (DAT)". PhD thesis. Massachusetts Institute of Technology.

Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining (2012). *Introduction to linear regression analysis*. Vol. 821. John Wiley & Sons. ISBN: 0470542810.

Mullen, Katharine M. and Ivo H. M. van Stokkum (2015). *The Lawson-Hanson algorithm for non-negative least squares (NNLS)*. Manual. URL: https://cran.r-project.org/web/packages/nnls/nnls.pdf.

NFF (2019). *Forslag til enhetstider og kapasiteter mv for konvensjonell tunneldrift (26.08.2019)*.

NIST/SEMATECH (2012). *NIST/SEMATECH e-Handbook of Statistical Methods*. DOI: https://doi.org/10.18434/M32189A.

Nocedal, Jorge and Stephen Wright (2006). *Numerical optimization*. Springer Science & Business Media.

NTNU (2017). Web Page. URL: https://innsida.ntnu.no/wiki/-/wiki/Norsk/Hjelp+til+litteraturs%C3%B8k.

— (2019). *IMRaD - How to structure your text*. URL: https://www.ntnu.edu/sekom/imrad.

Odabaşi, E (2009). "Models for estimating construction duration: An application for selected buildings on the Metu campus". Thesis.

Odd Johannessen Randi Hermann, Baroline Log (Mar. 2000). *Tunneldrift - Enhetstidsystem for driving, sikring, og innredning*. Report. Norwegian University of Science and Technology (NTNU).

Paradis, Emmanuel (2002). *R for Beginners*.

Pemstein, Jonah and Sean Dolinar (2015). *A New Way to Look at Sample Size*. URL: https://blogs.fangraphs.com/a-new-way-to-look-at-sample-size/.

Picard, Richard R. and Kenneth N. Berk (1990). "Data Splitting". In: *The American Statistician* 44.2, pp. 140–147. ISSN: 0003-1305. DOI: 10.1080/00031305.1990.10475704. URL: https://www.tandfonline.com/doi/abs/10.1080/00031305.1990.10475704.

Plate, Tony A (1999). "Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using gaussian process models". In: *Behaviormetrika* 26.1, pp. 29–50.

Prvan, Tania, Anna Reid, and Peter Petocz (2002). "Statistical laboratories using Minitab, SPSS and Excel: A practical comparison". In: *Teaching Statistics* 24.2, pp. 68–75. ISSN: 0141-982X.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Rostami, Jamal (2016). "Performance prediction of hard rock Tunnel Boring Machines (TBMs) in difficult ground". In: *Tunnelling and Underground Space Technology* 57, pp. 173–182. ISSN: 0886-7798. DOI: https://doi.org/10.1016/j.tust.2016.01.009. URL: http://www.sciencedirect.com/science/article/pii/S0886779815301991.

Salimi, Alireza et al. (Nov. 2016). "TBM performance estimation using a classification and regression tree (CART) technique". In: *Bulletin of Engineering Geology and the Environment*. DOI: 10.1007/s10064-016-0969-0.

Sammut, Claude and Geoffrey I. Webb (2010). *Encyclopedia of Machine Learning*. ISBN: 978-0-387-30768-8. DOI: 10.1007/978-0-387-30164-8.

Samset, Knut (2010). *Early project appraisal: making the initial choices*. Springer.

Schielzeth, Holger (2010). "Simple means to improve the interpretability of regression coefficients". In: *Methods in Ecology and Evolution* 1.2, pp. 103–113. ISSN: 2041-210X. DOI: 10.1111/j.2041-210X.2010.00012.x. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2010.00012.x.

School, Ottawa Hills Local (2012). Web Page. URL: https://www.ohschools.k12.oh.us/userfiles/223/Classes/29788/10-24%5C%20notes%5C%20per%5C%201.pdf%5C?id=20813.

Shaffer, Juliet Popper (1991). "The Gauss—Markov Theorem and Random Regressors". In: *The American Statistician* 45.4, pp. 269–273. ISSN: 0003-1305.

Shen, Wen (2015). *An Introduction to Numerical Computation*. WORLD SCIENTIFIC. DOI: 10.1142/9844. eprint: https://www.worldscientific.com/doi/pdf/10.1142/9844. URL: https://www.worldscientific.com/doi/abs/10.1142/9844.

Singh, Bhawani and R.K. Goel (2011). "Chapter 8 - Rock Mass Quality Q-System". In: *Engineering Rock Mass Classification*. Ed. by Bhawani Singh and R.K. Goel. Boston: Butterworth-Heinemann, pp. 85–118. ISBN: 978-0-12-385878-8. DOI: https://doi.org/10.1016/B978-0-12-385878-8.00008-2. URL: http://www.sciencedirect.com/science/article/pii/B9780123858788000082.

Snee, Ronald D. (1977). "Validation of Regression Models: Methods and Examples". In: *Technometrics* 19.4, pp. 415–428. ISSN: 00401706. DOI: 10.2307/1267881. URL: www.jstor.org/stable/1267881.

Špačková, O, J Sejnoha, and D Straub (2013). “Tunnel construction time and costs estimates: from deterministic to probabilistic approaches”. In: *12th international conference underground construction*.

Stone, Mervyn (1974). “Cross-validatory choice and assessment of statistical predictions”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 111–133. ISSN: 0035-9246.

Sullivan, Gail M and Richard Feinn (2012). “Using effect size—or why the P value is not enough”. In: *Journal of graduate medical education* 4.3, pp. 279–282. ISSN: 1949-8357.

Tierney, Brendan (2012). *Data Science Is Multidisciplinary*. Web Page. URL: https://oralytics.com/2012/06/13/data-science-is-multidisciplinary/.

Tukey, John W (1962a). “The future of data analysis”. In: *The annals of mathematical statistics* 33.1.

— (1962b). “The future of data analysis”. In: *The annals of mathematical statistics* 33.1, pp. 1–67.

Unlisted (2017). *Kongsbergtunnelen, østre portaler.* [Online; accessed July 6, 2020]. URL: https://www.wikidata.org/wiki/Q22079482.

Veidekke and VegVesen (2016). *2016 Uke 05 Ukerapport h¢yre l¢p signert, Kongsbergtunnelen.*

Wang, Dong Qian, Stefanka Chukova, and CD Lai (2004). “On the relationship between regression analysis and mathematical programming”. In: *Journal of Applied Mathematics & Decision Sciences* 8.2, pp. 131–140.

Wang, Hao, Jun Yang, and Songhua Hao (2016). “Two Inverse Normalizing Transformation methods for the process capability analysis of non-normal process data”. In: *Computers & Industrial Engineering* 102, pp. 88–98. ISSN: 0360-8352. DOI: https://doi.org/10.1016/j.cie.2016.10.014. URL: http://www.sciencedirect.com/science/article/pii/S0360835216303837.

Wickham, Hadley (2019). *Advanced r*. CRC press.

Wickham, Hadley and Garrett Grolemund (2016). *R for data science: import, tidy, transform, visualize, and model data.* " O’Reilly Media, Inc.".

Wirth, Rüdiger and Jochen Hipp (2000). “CRISP-DM: Towards a standard process model for data mining”. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Springer-Verlag London, UK, pp. 29–39.

Zare, Shokrollah (2006). “Report 2B-05 Drill and Blast Tunnelling - Advance Rate”. Thesis.

— (2007). “Prediction Model and Simulation Tool for Time and Cost of Drill and Blast Tunnelling”. Thesis.

— (2016). “Evaluating D&B and TBM tunnelling using NTNU prediction models”. In: *Tunnelling and Underground Space Technology* 59, pp. 55–64. DOI: 10.1016/j.tust.2016.06.012.

Zare, Shokrollah and Amund Bruland (2006). “Estimation model for advance rate in drill and blast tunnelling”. In: *Intern. symp. on utilization of underground space in urban areas, 6–7 November 2006.*

— (2007). “Progress of drill and blast tunnelling efficiency with relation to excavation time and cost”. In: *33rd ITA World Tunnel Congress*, pp. 805–809.

Zhu, Xingquan and Xindong Wu (2004). “Class noise vs. attribute noise: A quantitative study”. In: *Artificial intelligence review* 22.3, pp. 177–210.

Zumel, Nina and John Mount (2014). *Practical data science with R*. Manning Publications Co.

Appendices

The following appendices are included:

Appendix A

Programs

This chapter presents the predictive analysis software and packages employed in this study. In this appendix, a brief summary of their capabilities and limitations are also included.

A.1 Software capabilities

In order to adequately and efficiently apply the proposed predictive analysis techniques, various computer programs, platforms and languages were investigated and trialed during this study. Table A.1 below, outlines the software names and their respective capabilities and limitations. Where: "✓" indicates available capabilities; "X" indicates no capabilities; and "U" indicates that the capabilities are unsure; and L = Limited. However, further training is admittedly required to obtain a better understanding of their full potential.

Following preliminary testing, this report has elected to focus on MATLAB, R, and Excel to carry out the data science in this report. This decision was made on the basis of several constraints: such as the overall software capability, ease of use, adequate training and competence required (within my time frame).

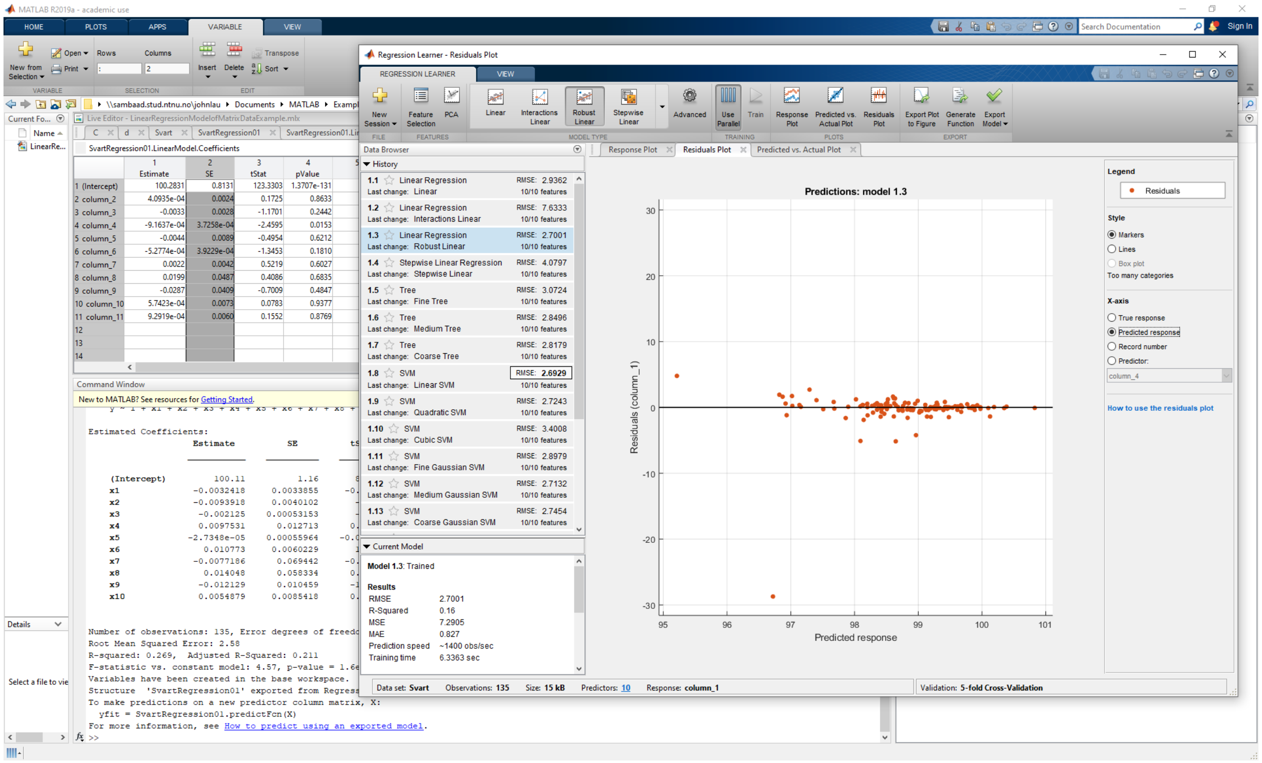
Table A.1: Computer software capabilities

	Software					
	Matlab	R-Studio	Python	Minitab	Excel	SPSS
Regression analysis	✓	✓	✓	✓	✓	✓
Constrained conditions	✓	✓	✓	X	X	X
Optimisation	✓	✓	✓	L	L	✓
Visualisation	✓	✓	✓	L	L	✓
Step-Wise Analysis	✓	✓	✓	✓	X	✓
Variance Inflation	U	U	✓	U	X	U
Time Series	L	U	✓	U	X	U
Linearization	✓	✓	✓	L	L	✓
Machine Learning	L	X	✓	X	X	X

A.2 MATLAB

MATLAB is computing environment and programming language developed by MathWorks. Its primary function is for numerical computation, but additional packages also allow for various specific and niche applications. All in all, the software is very comprehensive and possesses very high customisation: due to its ability to introduce personal/modified codes and add-ons, as well as access to various community-developed add-on packages. Conversely, due to the sheer size of its capabilities, the program can require large setup time, and operations can appear bloated at times.

In-depth and visual guides are all available free on the MATLAB website. The documentation there, provided concise insight into operating the software, and basic instructions into the application of popular regression models. Additionally, their community support webpages offered tutorials and often answered customer questions (MathWorks 2019a). This facilitated greatly throughout the semester as I attempted to better understand the ins-and-outs of program and its language. Figure A.1, is a screen-grab of typical regression analysis in action in MATLAB.



SvartRegression01.LinearModel.Coefficients

	1 Estimate	2 SE	3 tStat	4 pValue
1 (Intercept)	100.2831	0.8131	123.3303	1.3707e-131
2 column_2	4.0935e-04	0.0024	0.1725	0.8633
3 column_3	-0.0033	0.0028	-1.1701	0.2442
4 column_4	-9.1637e-04	3.7258e-04	-2.4595	0.0153
5 column_5	-0.0044	0.0089	-0.4954	0.6212
6 column_6	-5.2774e-04	3.9229e-04	-1.3453	0.1810
7 column_7	0.0022	0.0042	0.5219	0.6027
8 column_8	0.0199	0.0487	0.4086	0.6835
9 column_9	-0.0287	0.0409	-0.7009	0.4847
10 column_10	5.7423e-04	0.0073	0.0783	0.9377
11 column_11	9.2919e-04	0.0060	0.1552	0.8769
12				
13				
14				

Command Window

New to MATLAB? See resources for Getting Started.

```
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 +

Estimated Coefficients:
                   Estimate        SE         tS
                   ________     ________     ___

    (Intercept)      100.11         1.16       8
    x1           -0.0032418    0.0033855     -0.
    x2           -0.0093918    0.0040102      -
    x3            -0.002125   0.00053153      -
    x4            0.0097531     0.012713      0.
    x5          -2.7348e-05   0.00055964    -0.0
    x6             0.010773    0.0060229       1
    x7           -0.0077186     0.069442     -0.
    x8             0.014048     0.058334      0.
    x9            -0.012129     0.010459      -1
    x10           0.0054879    0.0085418      0.


Number of observations: 135, Error degrees of freedo
Root Mean Squared Error: 2.58
R-squared: 0.269,  Adjusted R-Squared: 0.211
F-statistic vs. constant model: 4.57, p-value = 1.6e
Variables have been created in the base workspace.
Structure 'SvartRegression01' exported from Regress
To make predictions on a new predictor column matrix, X:
  yfit = SvartRegression01.predictFcn(X)
For more information, see How to predict using an exported model.
>>
```

Regression Learner - Residuals Plot

History

		RMSE	
1.1	Linear Regression Last change: Linear	RMSE: 2.9362	10/10 features
1.2	Linear Regression Last change: Interactions Linear	RMSE: 7.6333	10/10 features
1.3	Linear Regression Last change: Robust Linear	RMSE: 2.7001	10/10 features
1.4	Stepwise Linear Regression Last change: Stepwise Linear	RMSE: 4.0797	10/10 features
1.5	Tree Last change: Fine Tree	RMSE: 3.0724	10/10 features
1.6	Tree Last change: Medium Tree	RMSE: 2.8496	10/10 features
1.7	Tree Last change: Coarse Tree	RMSE: 2.8179	10/10 features
1.8	SVM Last change: Linear SVM	RMSE: 2.6929	10/10 features
1.9	SVM Last change: Quadratic SVM	RMSE: 2.7243	10/10 features
1.10	SVM Last change: Cubic SVM	RMSE: 3.4008	10/10 features
1.11	SVM Last change: Fine Gaussian SVM	RMSE: 2.8979	10/10 features
1.12	SVM Last change: Medium Gaussian SVM	RMSE: 2.7132	10/10 features
1.13	SVM Last change: Coarse Gaussian SVM	RMSE: 2.7454	10/10 features

Current Model

Model 1.3: Trained

Results

RMSE	2.7001
R-Squared	0.16
MSE	7.2905
MAE	0.827
Prediction speed	~1400 obs/sec
Training time	6.3363 sec

Data set: Svart Observations: 135 Size: 15 kB Predictors: 10 Response: column_1 Validation: 5-fold Cross-Validation

Style: ● Markers ○ Lines ○ Box plot — Too many categories

X-axis: ○ True response ● Predicted response ○ Record number ○ Predictor: column_4

Figure A.1: A typical MATLAB user-interface

Input commands

In order to fit a linear regression line, using ordinary least squares, the following commands are called upon in MATLAB.

Listing A.1: Ordinary least squares regression fit

```
%Code with Intercept
 fitlm(X,y)
```

For non-negative least squares in MATLAB, the following algorithm is applied, (according to the Mathworks webpage (MathWorks 2019a)):

```
**Algorithm** lsqnonneg
Input: $\boldsymbol{A} \in \boldsymbol{R}^{m\times n}, \boldsymbol{b} \in \boldsymbol{R}^m$
Output: $x^* \geq 0$ such that $x^* = \arg\min \| Ax - b \|^2$
Initialization: $P = \emptyset, R = \{1, 2, \ldots, n\}, x = \mathbf{0}, \boldsymbol{w} = \boldsymbol{A}^T(\boldsymbol{b} - \boldsymbol{Ax})$
repeat
    1. Proceed if $R \neq \emptyset \wedge [\max_{i\in R}(w_i) >$ tolerance ]
    2. $j = \arg\max_{i\in R}(w_i)$
    3. Include the index $j$ in $P$ and remove it from $R$
    4. $s^P = \left[\left(A^P\right)^T A^P\right]^{-1} \left(A^P\right)^T b$
        4.1. Proceed if $\min\left(s^P\right) \leq 0$
        4.2. $\alpha = -\min_{i\in P}[x_i / (x_i - s_i)]$
        4.3. $x := x + \alpha(s - x)$
        4.4. Update $R$ and $P$
        4.5. $s^P = \left[\left(\boldsymbol{A^P}\right)^{\boldsymbol{T}} \boldsymbol{A^P}\right]^{-\mathbf{1}} \left(\boldsymbol{A^P}\right)^{\boldsymbol{T}} \boldsymbol{b}$
        4.6. $s^R = 0$
    5. $\boldsymbol{x} = \boldsymbol{s}$
    6. $\boldsymbol{w} = \boldsymbol{A}^T(\boldsymbol{b} - \boldsymbol{Ax})$
```

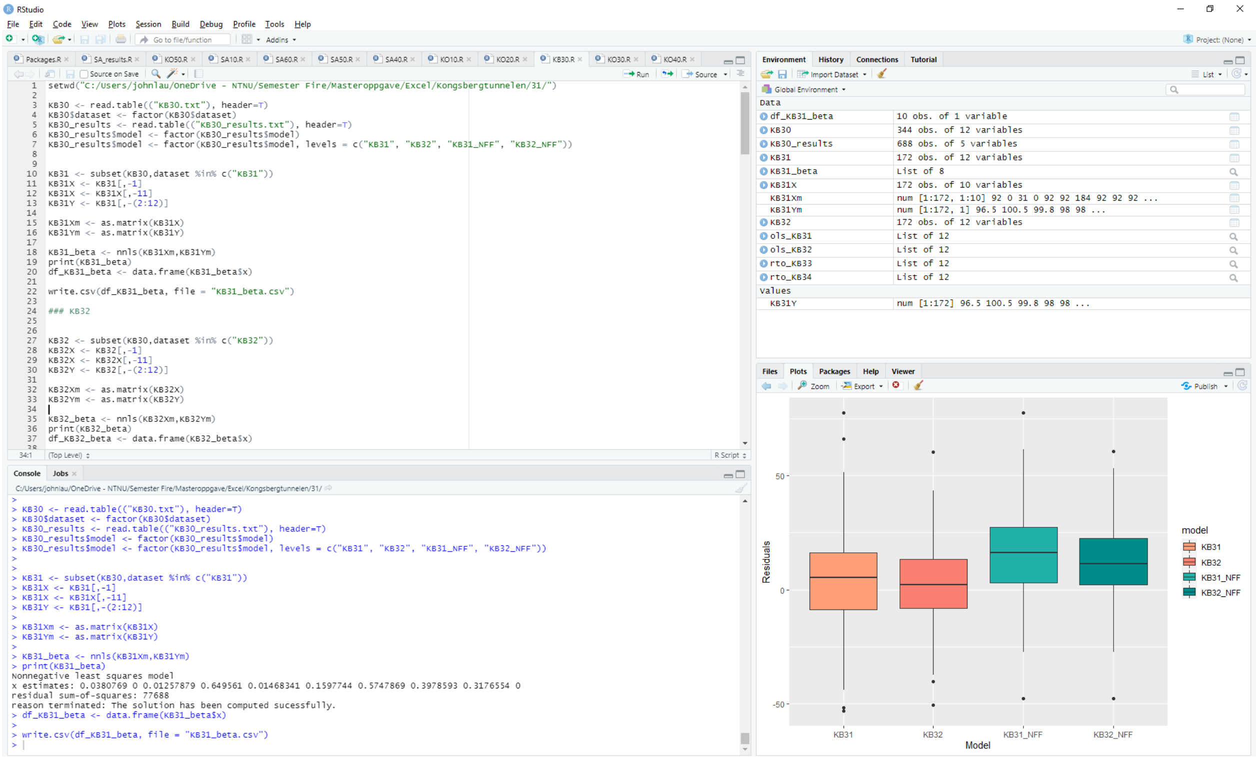
The matrix A^P is a matrix associated with only the variables currently in the passive set P.

Information source

Relevant information was drawn from the MATLAB homepage, as well as from their community subsections (such as "File Exchange" and from their forums) (MathWorks 2019b).

A.3 *R*

R is a language and environment for statistical and mathematical computing; and visual presentation (R Core Team 2013). Figure A.2 is a typical interface of the program.



```
setwd("C:/Users/johnlau/OneDrive - NTNU/Semester Fire/Masteroppgave/Excel/Kongsbergtunnelen/31/")

KB30 <- read.table(("KB30.txt"), header=T)
KB30$dataset <- factor(KB30$dataset)
KB30_results <- read.table(("KB30_results.txt"), header=T)
KB30_results$model <- factor(KB30_results$model)
KB30_results$model <- factor(KB30_results$model, levels = c("KB31", "KB32", "KB31_NFF", "KB32_NFF"))


KB31 <- subset(KB30,dataset %in% c("KB31"))
KB31X <- KB31[,-1]
KB31X <- KB31X[,-11]
KB31Y <- KB31[,-(2:12)]

KB31Xm <- as.matrix(KB31X)
KB31Ym <- as.matrix(KB31Y)

KB31_beta <- nnls(KB31Xm,KB31Ym)
print(KB31_beta)
df_KB31_beta <- data.frame(KB31_beta$x)

write.csv(df_KB31_beta, file = "KB31_beta.csv")

### KB32


KB32 <- subset(KB30,dataset %in% c("KB32"))
KB32X <- KB32[,-1]
KB32X <- KB32X[,-11]
KB32Y <- KB32[,-(2:12)]

KB32Xm <- as.matrix(KB32X)
KB32Ym <- as.matrix(KB32Y)

KB32_beta <- nnls(KB32Xm,KB32Ym)
print(KB32_beta)
df_KB32_beta <- data.frame(KB32_beta$x)
```

C:/Users/johnlau/OneDrive - NTNU/Semester Fire/Masteroppgave/Excel/Kongsbergtunnelen/31/

```
>
> KB30 <- read.table(("KB30.txt"), header=T)
> KB30$dataset <- factor(KB30$dataset)
> KB30_results <- read.table(("KB30_results.txt"), header=T)
> KB30_results$model <- factor(KB30_results$model)
> KB30_results$model <- factor(KB30_results$model, levels = c("KB31", "KB32", "KB31_NFF", "KB32_NFF"))
>
>
> KB31 <- subset(KB30,dataset %in% c("KB31"))
> KB31X <- KB31[,-1]
> KB31X <- KB31X[,-11]
> KB31Y <- KB31[,-(2:12)]
>
> KB31Xm <- as.matrix(KB31X)
> KB31Ym <- as.matrix(KB31Y)
>
> KB31_beta <- nnls(KB31Xm,KB31Ym)
> print(KB31_beta)
Nonnegative least squares model
x estimates: 0.0380769 0 0.01257879 0.649561 0.01468341 0.1597744 0.5747869 0.3978593 0.3176554 0
residual sum-of-squares: 77688
reason terminated: The solution has been computed sucessfully.
> df_KB31_beta <- data.frame(KB31_beta$x)
>
> write.csv(df_KB31_beta, file = "KB31_beta.csv")
> |
```

```
Data
df_KB31_beta    10 obs. of 1 variable
KB30            344 obs. of 12 variables
KB30_results    688 obs. of 5 variables
KB31            172 obs. of 12 variables
KB31_beta       List of 8
KB31X           172 obs. of 10 variables
KB31Xm          num [1:172, 1:10] 92 0 31 0 92 92 184 92 92 92 ...
KB31Ym          num [1:172, 1] 96.5 100.5 99.8 98 98 ...
KB32            172 obs. of 12 variables
ols_KB31        List of 12
ols_KB32        List of 12
rto_KB33        List of 12
rto_KB34        List of 12
Values
KB31Y           num [1:172] 96.5 100.5 99.8 98 98 ...
```

Figure A.2: A typical *R* user-interface

Information source

Information regarding it's operations and features was learned from lecture notes and text books, as well as internet forums. It is often difficult to pin point exactly which literature contributed to my overall training. Below is therefore a list of all the sources that contributed to my development of the program: R

- *R for Beginners* (Paradis 2002)
- *Hands-On Programming with R* (Grolemund 2014)
- *Advanced r* (Wickham 2019)
- *Practical data science with R* (Zumel and Mount 2014)
- *Linear models with R* (Faraway 2016)

Least squares with non-negative and non-positive constraints

An R interface to the Lawson-Hanson implementation of an algorithm for non-negative least squares (NNLS) and non-positive least squares (NPLS) can be achieved by calling up on the *nnnpls* package (Mullen and Stokkum 2015). Fundamentally, the code solves $\min \| Ax - b \|_2$, with the constraint $\boldsymbol{x} \geq 0$, where $\boldsymbol{x} \in R^n$, $\boldsymbol{b} \in R^m$, and $\boldsymbol{A} \in R^{m \times n}$.

A.4 Excel

Excel is a spreadsheet application from the Microsoft corporation. Although this program's primary function is to serve as a data management tool, one of its secondary features allows for basic statistical analysis: using the add-on package "Analysis ToolPak". Excel is subjectively the most user friendly, albeit most basic of the trialed programs. It was thwart with limitations in functionality and customisability. However, due to its ease of access, and minimal setup requirements, the program ended up becoming a highly valuable in the writing of this report. It provided a practical starting point for each database: with an ability to quickly gain an overview of the data, before the implementation of more in-depth analysis and modelling.

The Analysis ToolPak capabilities and commands used in this study included:

- Anova
- Correlation
- Histogram
- Regression (using the least-squares method)
- t-Test

Figure A.3 shows a typical Microsoft Excel data analysis user-interface.

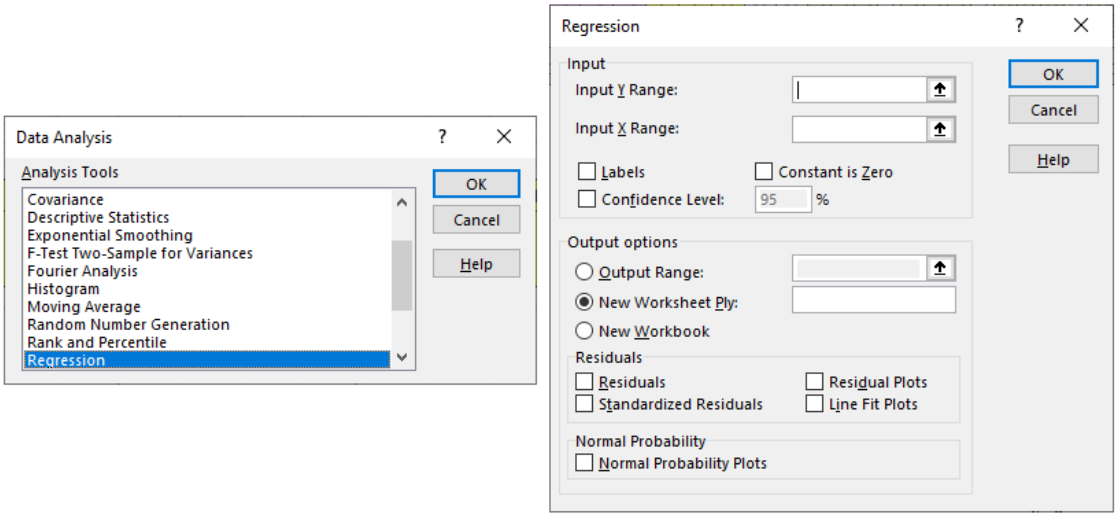


Figure A.3: A Typical Microsoft Excel data analysis user-interface

Appendix B

The Svartås-tunnel

A collection of additional information and results from the Svartås-tunnel data analysis has been included in this appendix.

B.1 Composition of each input variable

The make-up of each input variable is detailed in the Tables below.

The composition of probe drilling elements (*X_1* and *X_2*)

Contributing construction task	Size	Process code (R761)	Unit
Sonderboring ved stuff	0-12m	31.111	m
Sonderboring ved stuff	12-24m	31.112	m
Sonderboring ved stuff	24-36m	31.113	m

The composition of "control and injection holes" (*X_3*)

Contributing construction task	Size	Process code (R761)	Unit
Boring og spyling av inksjons og kontroll hull	12-18m	31.51	m
Boring og spyling av inksjons og kontroll hull	18-21m	31.52	m
Boring og spyling av inksjons og kontroll hull	21-24m	31.53	m

The composition of pre-grouting (injection) (X_4)

Contributing construction task	Size	Process code	Unit
		(R761)	
Standard injeksjonssement		31.631	kg
Mikrosement	-	31.632	kg
Ultrafin sement	-	31.633	kg
Spesialsement	-	31.634	kg
Injeksjon (Polyuretan)	-	31.635	kg
Silicaslurry	-	31.6391	kg
Superplastiserende stoff	-	31.6392	kg
Styrt herding med alkalifri akselerator	-	31.6393	kg

The composition of excavated material (X_5)

Contributing construction task	Size	Process code	Unit
		(R761)	
Fullt tverrsnitt – normal salvelengde	-	32.111	m^3
Fullt tverrsnitt – halv salvelengde	-	32.112	m^3
Todelt tverrsnitt – normal salvelengde	-	32.113	m^3
Todelt tverrsnitt – halv salvelengde	-	32.114	m^3

The composition of rock bolts ≤ 4m (X_6)

Contributing construction task	Size	Process code	Unit
		(R761)	
Fullt innstøpte, lengde 2,4 m, Ø 20 mm	2,4m	33.2211	stk
Fullt innstøpte, lengde 3,0 m, Ø 20 mm	3m	33.2212	stk
Fullt innstøpte, lengde 4,0 m, Ø 20 mm	4m	33.2213	stk
Endeforankrede, lengde 2,4 m, Ø 20 mm	2,4m	33.2221	stk
Endeforankrede, lengde 3,0 m, Ø 20 mm	3m	33.2222	stk
Endeforankrede, lengde 4,0 m, Ø 20 mm	4m	33.2223	stk
Kombinasjonsbolter, lengde 2,4 m, Ø 20 mm	2,4m	33.2231	stk
Kombinasjonsbolter, lengde 3,0 m, Ø 20 mm	3m	33.2232	stk
Kombinasjonsbolter, lengde 4,0 m, Ø 20 mm	4m	33.2233	stk
Med pakker ved stuff, lengde 3,0 m, Ø 20 mm	3m	33.2241	stk
Med pakker ved stuff, lengde 4,0 m, Ø 20 mm	4m	33.2242	stk
Endeforankrede (polyester), lengde 2,4 m, Ø 20 mm	2,4m	33.231	stk
Endeforankrede (polyester), lengde 3,0 m, Ø 20 mm	3m	33.232	stk
Endeforankrede (polyester), lengde 4,0 m, Ø 20 mm	4m	33.233	stk
Injiserbar bolt	3m	33.2295	stk
Sikringsbolter B/Stuff, ullt innstøpte	3m	33.232	stk

The composition of rockbolts ≥ 4m (X_7)

Contributing construction task	Size	Process code	Unit
		(R761)	
Fullt innstøpte, lengde 6,0 m, Ø 32 mm	6m	33.211	stk
Fullt innstøpte, lengde 8,0 m, Ø 32 mm	8m	33.212	stk
Fullt innstøpte, lengde 5,0 m, Ø 20 mm	5m	33.2214	stk
Endeforankrede, lengde 5,0 m, Ø 20 mm	5m	33.2224	stk
Kombinasjonsbolter, lengde 5,0 m, Ø 20 mm	5m	33.2234	stk
Endeforankrede (polyester), lengde 5,0 m, Ø 20 mm	5m	33.234	stk

The composition of straps (X_8)

Contributing construction task	Size	Process code (R761)	Unit
Bånd ved stuff	-	33.311	m

The composition of shotcrete (X_9)

Contributing construction task	Size	Process code (R761)	Unit
Uten tilsetting av fiber, B35 M45	M45	33.4111	m^3
Uten tilsetting av fiber, B35 M40	M40	33.4112	m^3
Med tilsetting av fiber, B35 M45 E500	M45 E500	33.4121	m^3
Med tilsetting av fiber, B35 M45 E700	M45 E700	33.4122	m^3
Med tilsetting av fiber, B35 M45 E1000	M45 E1000	33.4123	m^3
Med tilsetting av fiber, B35 M40 E500	M40 E500	33.4124	m^3
Med tilsetting av fiber, B35 M40 E700	M40 E700	33.4125	m^3
Med tilsetting av fiber, B35 M40 E1000	M40 E1000	33.4126	m^3

The composition of reinforcements and arches (X_{10})

Contributing construction task	Size	Process code (R761)	Unit
Buer med kamstål	-	33.44221	m
Buer med gitterdragere	-	33.44222	m

B.2 Additional model details and results

To avoid clutter, several Svartås-tunnel model summary and results were omitted in the main report. Instead, these have been included in this section of the Appendix for reference.

B.2.1 Regression analysis

Table B.1: Ordinary least squares: SA-O01 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-O01	Y	X	X	X	X	X	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-O01	OLS	0.519	0.269	0.211	2.581	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-O01	-25.4005	-0.5050	0.1237	0.8028	3.4541

ANOVA

	df	SS	MS	F	Significance F
Regression	10	304.793	30.479	4.574	1.595E-05
Residual	124	826.266	6.663		
Total	134	1 131.059			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	100.11335	1.16003	86.30248	0.00000	97.81733	102.40938
X_1	-0.00325	0.00339	-0.95840	0.33973	-0.00995	0.00346
X_2	-0.00940	0.00401	-2.34273	0.02074	-0.01734	-0.00146
X_3	-0.00213	0.00053	-3.99816	0.00011	-0.00318	-0.00107
X_4	0.00975	0.01271	0.76705	0.44451	-0.01541	0.03491
X_5	-0.00003	0.00056	-0.04882	0.96114	-0.00113	0.00108
X_6	0.01077	0.00602	1.78876	0.07609	-0.00115	0.02269
X_7	-0.00774	0.06944	-0.11141	0.91147	-0.14518	0.12971
X_8	0.01406	0.05833	0.24106	0.80991	-0.10140	0.12952
X_9	-0.01213	0.01046	-1.15991	0.24831	-0.03283	0.00857
X_{10}	0.00549	0.00854	0.64252	0.52172	-0.01142	0.02239

Table B.2: Ordinary least squares: SA-O02 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-O02	Y	X	X	X	X	X	X	X	X	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-O02	OLS	0.512	0.262	0.202	2.595	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-O02	-25.5694	-0.3817	0.1018	0.7684	3.5403

ANOVA

	df	SS	MS	F	Significance F
Regression	10.0	296.1	29.6	4.4	2.7E-05
Residual	124.0	834.9	6.7		
Total	134.0	1 131.1			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	99.9314	1.1554	86.4927	0.0000	97.6446	102.2182
X_1	-0.0031	0.0034	-0.8994	0.3702	-0.0098	0.0037
X_2	-0.0087	0.0040	-2.1749	0.0315	-0.0165	-0.0008
X_3	-0.0021	0.0005	-3.8686	0.0002	-0.0031	-0.0010
X_4	0.0079	0.0127	0.6202	0.5363	-0.0172	0.0329
X_5	-0.0003	0.0005	-0.6860	0.4940	-0.0013	0.0006
X_6	0.0087	0.0058	1.5011	0.1359	-0.0028	0.0201
X_7	-0.0258	0.0680	-0.3788	0.7055	-0.1604	0.1089
X_8	0.0246	0.0579	0.4255	0.6712	-0.0900	0.1393
X_9	-	-	-	-	-	-
X_{10}	0.0012	0.0077	0.1517	0.0000	-0.0141	0.0165

Table B.3: Ordinary least squares: SA-O03 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-O03	Y	X	X	X	X	X	X	X	NFF	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-O03	OLS	0.536	0.288	0.230	2.595	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-O03	-25.5462	-0.3939	0.1076	0.7828	3.5449

ANOVA

	df	SS	MS	F	Significance F
Regression	10	337.479	33.748	5.010	4.241E-06
Residual	124	835.299	6.736		
Total	134	1 172.778			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	100.0043	1.1225	89.0928	0.0000	97.7826	102.2260
X_1	-0.0032	0.0034	-0.9372	0.3505	-0.0098	0.0035
X_2	-0.0089	0.0039	-2.2743	0.0247	-0.0166	-0.0012
X_3	-0.0021	0.0005	-3.9289	0.0001	-0.0031	-0.0010
X_4	0.0078	0.0127	0.6190	0.5370	-0.0172	0.0329
X_5	-0.0004	0.0005	-0.7187	0.4737	-0.0013	0.0006
X_6	0.0085	0.0057	1.4828	0.1407	-0.0029	0.0199
X_7	-0.0433	0.0159	-2.7230	0.0074	-0.0748	-0.0118
X_8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X_9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X_{10}	0.0014	0.0077	0.1782	0.0000	-0.0139	0.0166

Table B.4: Ordinary least squares: SA-O04 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-O04	Y	X	X	X	X	X	X	NFF	NFF	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
SA-O04	OLS	0.433	0.188	0.122	3.660	135

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
SA-O04	-25.8993	-0.4772	0.7035	1.8256	5.5031

ANOVA

	df	SS	MS	F	Significance F
Regression	10	383.617	38.362	2.863	3.042E-03
Residual	124	1 661.295	13.398		
Total	134	2 044.912			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	95.364	1.469	64.907	0.000	92.456	98.272
X_1	0.002	0.005	0.445	0.657	-0.007	0.011
X_2	-0.003	0.005	-0.640	0.523	-0.014	0.007
X_3	-0.001	0.001	-1.733	0.086	-0.003	0.000
X_4	0.025	0.018	1.386	0.168	-0.011	0.060
X_5	0.001	0.001	2.093	0.038	0.000	0.003
X_6	0.004	0.008	0.454	0.651	-0.012	0.020
X_7	-	-	-	-	-	-
X_8	-	-	-	-	-	-
X_9	-	-	-	-	-	-
X_{10}	0.014	0.011	1.260	0.000	-0.008	0.035

Table B.5: Ordinary least squares: SA-O05 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-O05	Y	X	X	X	X	X	X	NFF	NFF	EX	NFF

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
SA-O05	OLS	0.413	0.171	0.104	8.067	135

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
SA-O05	-46.597	-1.283	1.410	4.006	8.868

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1 660.021	166.002	2.551	7.790E-03
Residual	124	8 068.872	65.072		
Total	134	9 728.892			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	84.501	3.050	27.706	0.000	78.464	90.538
X_1	-0.003	0.010	-0.315	0.753	-0.024	0.017
X_2	0.004	0.012	0.363	0.717	-0.019	0.028
X_3	0.002	0.002	1.188	0.237	-0.001	0.005
X_4	0.050	0.039	1.278	0.204	-0.027	0.127
X_5	0.006	0.001	4.348	0.000	0.003	0.009
X_6	-0.001	0.018	-0.061	0.952	-0.036	0.034
X_7	-	-	-	-	-	-
X_8	-	-	-	-	-	-
X_9	-	-	-	-	-	-
X_{10}	-	-	-	-	-	-

Table B.6: Ordinary least squares: SA-O06 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-O06	X	X	X	X	X	Y	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-O06	OLS	0.838	0.703	0.679	414.227	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-O06	-1180.58	-237.94	7.34	248.27	1023.34

ANOVA

	df	SS	MS	F	Significance F
Regression	10	50 350 105.595	5 035 010.560	29.344	3.735E-28
Residual	124	21 276 441.486	171 584.206		
Total	134	71 626 547.081			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	1 055.930	1 451.548	0.727	0.468	-1 817.090	3 928.950
X_1	1.797	0.521	3.449	0.001	0.766	2.828
X_2	0.849	0.653	1.300	0.196	-0.444	2.142
X_3	-0.178	0.089	-1.996	0.048	-0.355	-0.002
X_4	-4.803	1.999	-2.403	0.018	-8.759	-0.847
X_5	-	-	-	-	-	-
X_6	1.156	0.973	1.188	0.237	-0.770	3.083
X_7	-32.288	10.760	-3.001	0.003	-53.584	-10.991
X_8	17.586	9.229	1.906	0.059	-0.680	35.852
X_9	8.955	1.483	6.036	0.000	6.019	11.891
X_{10}	-7.154	1.213	-5.896	0.000	-9.556	-4.752
Y'	-0.704	14.410	-0.049	0.961	-29.226	27.819

Table B.7: Regression-through-the-origin: SA-R01 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R01	Y	X	X	X	X	X	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-R01	RTO	0.981	0.961	0.951	20.103	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-R01	-37.278	-9.327	3.708	17.010	51.569

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1 261 465.490	126 146.549	312.514	7.141E-83
Residual	125	50 456.340	403.651		
Total	135	1 311 921.830			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	0	-	-	-	-	-
X_1	0.044	0.026	1.676	0.096	-0.008	0.095
X_2	0.099	0.030	3.355	0.001	0.041	0.158
X_3	0.021	0.004	5.742	0.000	0.014	0.028
X_4	0.233	0.097	2.408	0.018	0.042	0.425
X_5	0.023	0.004	5.982	0.000	0.015	0.031
X_6	0.182	0.044	4.119	0.000	0.095	0.270
X_7	1.638	0.520	3.149	0.002	0.608	2.667
X_8	-1.050	0.444	-2.366	0.020	-1.929	-0.172
X_9	0.110	0.081	1.362	0.176	-0.050	0.270
X_{10}	0.177	0.065	2.733	0.007	0.049	0.305

Table B.8: Regression-through-the-origin: SA-R02 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R02	Y	X	X	X	X	X	X	X	X	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
SA-R02	RTO	0.981	0.962	0.951	20.049	135

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
SA-R02	-35.918	-10.298	3.826	17.426	49.749

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1261676.4	126167.6	313.9	5.5E-83
Residual	125	50245.4	402.0		
Total	135	1311921.8			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	-	-	-	-	-	-
X_1	0.043	0.026	1.654	0.101	-0.008	0.094
X_2	0.095	0.029	3.232	0.002	0.037	0.153
X_3	0.020	0.004	5.695	0.000	0.013	0.027
X_4	0.253	0.095	2.656	0.009	0.065	0.442
X_5	0.026	0.003	8.766	0.000	0.020	0.032
X_6	0.205	0.041	4.992	0.000	0.124	0.286
X_7	1.832	0.499	3.674	0.000	0.845	2.819
X_8	-1.167	0.435	-2.685	0.008	-2.027	-0.307
X_9	-	-	-	-	-	-
X_{10}	0.219	0.056	3.887	0.000	0.108	0.331

Table B.9: Regression-through-the-origin: SA-R03 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R03	Y	X	X	X	X	X	X	X	NFF	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
SA-R03	RTO	0.980	0.959	0.949	20.582	135

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
SA-R03	-39.591	-10.068	4.206	17.931	49.292

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1 254 167.408	125 416.741	296.059	1.776E-81
Residual	125	52 952.599	423.621		
Total	135	1 307 120.006			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-	-	-	-	-	-
X_1	0.037	0.027	1.387	0.168	-0.016	0.089
X_2	0.083	0.030	2.796	0.006	0.024	0.142
X_3	0.020	0.004	5.531	0.000	0.013	0.028
X_4	0.267	0.098	2.732	0.007	0.074	0.461
X_5	0.027	0.003	8.692	0.000	0.021	0.033
X_6	0.205	0.042	4.872	0.000	0.122	0.288
X_7	0.484	0.117	4.135	0.000	0.252	0.716
X_8	-	-	-	-	-	-
X_9	-	-	-	-	-	-
X_{10}	0.249	0.057	4.380	0.000	0.137	0.362

Table B.10: Regression-through-the-origin: SA-R04 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R04	Y	X	X	X	X	X	X	NFF	NFF	EX	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
SA-R04	RTO	0.978	0.957	0.946	21.108	135

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
SA-R04	-40.923	-9.735	4.434	17.692	56.916

ANOVA

	df	SS	MS	F	Significance F
Regression	10	383.617	38.362	2.863	3.042E-03
Residual	124	1 661.295	13.398		
Total	134	2 044.912			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.000	-	-	-	-	-
X_1	0.029	0.027	1.068	0.287	-0.025	0.083
X_2	0.081	0.031	2.645	0.009	0.020	0.141
X_3	0.021	0.004	5.551	0.000	0.013	0.028
X_4	0.256	0.100	2.558	0.012	0.058	0.455
X_5	0.026	0.003	8.185	0.000	0.019	0.032
X_6	0.237	0.042	5.685	0.000	0.155	0.320
X_7	-	-	-	-	-	-
X_8	-	-	-	-	-	-
X_9	-	-	-	-	-	-
X_{10}	0.248	0.058	4.245	0.000	0.132	0.363

Table B.11: Regression-through-the-origin: SA-R05 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R05	Y	X	X	X	X	X	X	NFF	NFF	EX	NFF

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
SA-R05	RTO	0.978	0.956	0.945	21.065	135

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
SA-R05	-40.907	-9.725	4.453	17.684	56.918

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1 208 226.726	120 822.673	272.276	2.535E-79
Residual	125	55 468.803	443.750		
Total	135	1 263 695.529			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-	-	-	-	-	-
X_1	0.029	0.027	1.071	0.286	-0.024	0.082
X_2	0.081	0.031	2.650	0.009	0.020	0.141
X_3	0.021	0.004	5.570	0.000	0.013	0.028
X_4	0.256	0.100	2.563	0.012	0.058	0.455
X_5	0.026	0.003	8.321	0.000	0.019	0.032
X_6	0.237	0.041	5.817	0.000	0.156	0.318
X_7	-	-	-	-	-	-
X_8	-	-	-	-	-	-
X_9	-	-	-	-	-	-
X_{10}	-	-	-	-	-	-

Table B.12: Regression-through-the-origin: SA-R06 summary of outputs

Model input variables

Model ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA-R06	X	X	X	X	X	Y	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
SA-R06	RTO	0.976	0.953	0.941	413.446	135

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
SA-R06	-1181.56	-243.29	1.16	251.60	1025.11

ANOVA

	df	SS	MS	F	Significance F
Regression	10	428 489 565.689	42 848 956.569	250.670	3.334E-77
Residual	125	21 367 241.311	170 937.930		
Total	135	449 856 807.000			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	-	-	-	-	-
X_1	1.846	0.515	3.582	0.000	0.826	2.867
X_2	0.970	0.631	1.537	0.127	-0.279	2.218
X_3	-0.153	0.082	-1.863	0.065	-0.315	0.010
X_4	-4.887	1.992	-2.453	0.016	-8.829	-0.944
X_6	1.079	0.966	1.117	0.266	-0.832	2.990
X_7	-32.059	10.735	-2.986	0.003	-53.305	-10.813
X_8	17.330	9.205	1.883	0.062	-0.888	35.547
X_9	9.141	1.459	6.267	0.000	6.254	12.027
X_{10}	-7.212	1.208	-5.968	0.000	-9.604	-4.820
Y'	9.712	1.623	5.985	0.000	6.501	12.924

B.2.2 Mathematical optimisation

Table B.13: Non-negative least squares: Svartås-tunnel summary of outputs

Model input variables

Series ID	Y'	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
SA01	Y	X	X	X	X	X	X	X	X	X	X
SA02	Y	X	X	X	X	X	X	X	X	EX	X
SA03	Y	X	X	X	X	X	X	X	NFF	EX	X
SA04	Y	X	X	X	X	X	X	NFF	NFF	EX	X
SA05	Y	X	X	X	X	X	X	NFF	NFF	EX	NFF
SA06	X	X	X	X	Y	X	X	X	X	X	X

Dataset description

Model ID	Algorithm	No intercept	Response	Observations
		R-square	Y	n
SA-N1	*NNLS*	0.960	Y'	135
SA-N2	*NNLS*	0.959	Y'	135
SA-N3	*NNLS*	0.959	Y'	135
SA-N4	*NNLS*	0.956	Y'	135
SA-N5	*NNLS*	0.955	Y'	135
SA-N6	*NNLS*	0.923	X_5	135

Model outputs: time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
SA-N1	25.62	10.89	48.30	4.22	45.01	5.71	2.26	0	6.83	5.33
SA-N2	27.03	11.94	49.23	3.75	37.58	4.88	1.89	0	-	4.03
SA-N3	27.18	11.99	49.23	3.74	37.56	4.88	2.07	-	-	4.01
SA-N4	34.56	12.37	47.91	3.90	39.17	4.21	-	-	-	4.04
SA-N5	34.67	12.37	47.89	3.90	39.14	4.22	-	-	-	-
	X_1	X_2	X_3	X_4	Y'	X_6	X_7	X_8	X_9	X_{10}
SA-N6	0.28	0.53	0	0	0.24	0.29	0	0	0.15	0
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
NFF	60.00	-	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

B.2.3 Correlation matrix

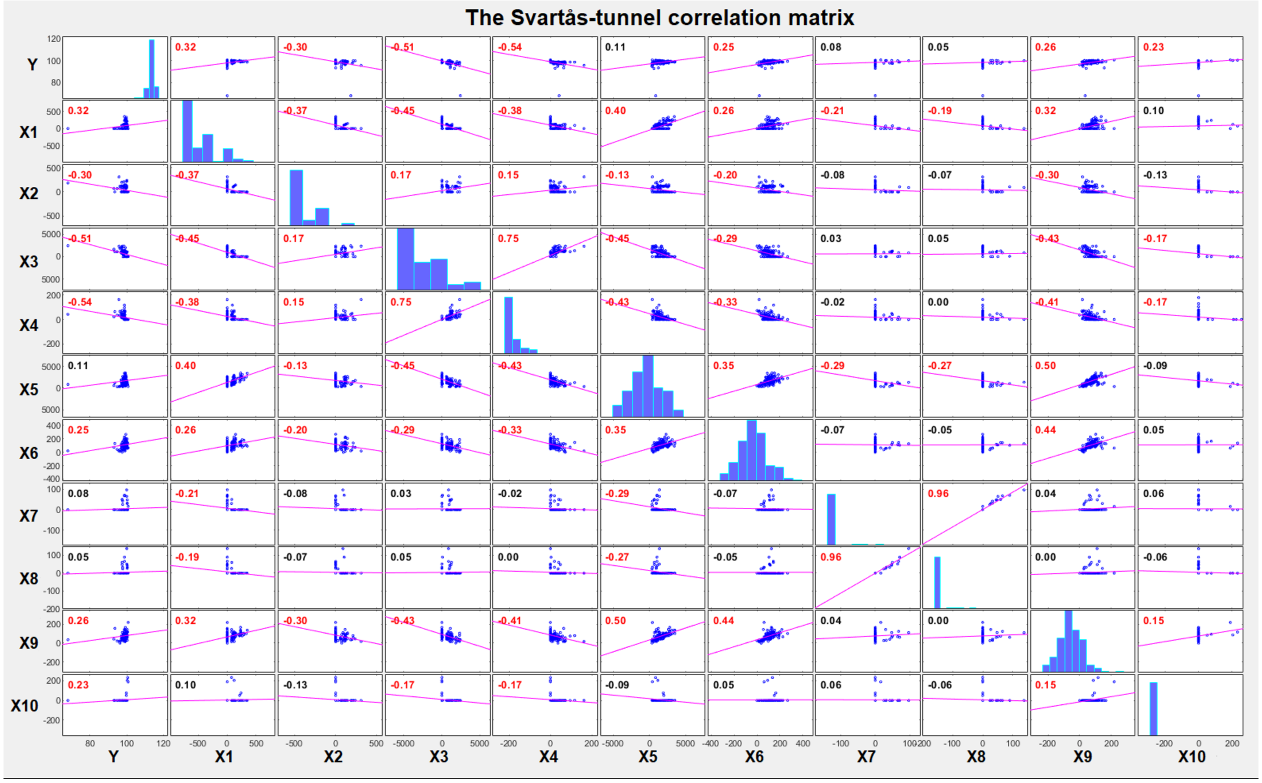


Figure B.1: The Svartås-tunnel correlation matrix

Appendix C

The Kongsberg-tunnel

A collection of additional information and results from the Kongsberg-tunnel data analysis has been included in this appendix.

C.1 Additional model details and results

To avoid clutter, several Kongsberg-tunnel model summary and results were omitted in the main report. Instead, these have been included in this section of the Appendix for reference.

C.1.1 Regression analysis

Table C.1: Regression models: OLS KB-O01

Regression model input variables

Series ID	Y'	Y_S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KB-O01	Y	-	X	-	X	X	X	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
KB-O01	OLS	0.424	0.179	0.128	1.249	172

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
KB-O01	-5.2060	-0.2906	0.2700	0.6547	2.4338

ANOVA

	df	SS	MS	F	Significance F
Regression	10.000	54.924	5.492	3.520	3.170E-04
Residual	161.000	251.216	1.560		
Total	171.000	306.141			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	100.1152	0.4496	222.6957	0.0000	99.2274	101.0030
X_1	-0.0021	0.0016	-1.3226	0.1878	-0.0053	0.0010
X_2	-	-	-	-	-	-
X_3	-0.0005	0.0002	-2.3192	0.0000	-0.0009	-0.0001
X_4	0.0004	0.0055	0.0802	0.9362	-0.0104	0.0112
X_5	0.0002	0.0002	0.8317	0.4068	-0.0002	0.0006
X_6	-0.0080	0.0032	-2.4927	0.0137	-0.0143	-0.0017
X_7	0.0133	0.0145	0.9185	0.3598	-0.0153	0.0419
X_8	-0.0230	0.0288	-0.7978	0.4262	-0.0799	0.0339
X_9	-0.0121	0.0064	-1.9009	0.0591	-0.0247	0.0005
X_{10}	0.0393	0.0557	0.7055	0.4815	-0.0707	0.1494

Table C.2: Regression models: OLS KB-O02

Regression model input variables

Series ID	Y'	Y_S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KB-O02	-	Y	X	-	X	X	X	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted	Standard	Observations
				R-square	Error	n
KB-O02	OLS	0.600	0.360	0.320	9.581	172

Residuals

	Min.	First quartile	Medium	Third quartile	Max.
		Q_1	$\tilde{x}$	Q_3	
KB-O02	-37.838	-3.556	0.759	6.063	20.198

ANOVA

	df	SS	MS	F	Significance F
Regression	10	8 299.291	829.929	9.041	9.470E-12
Residual	161	14 778.700	91.793		
Total	171	23 077.991			

Model outputs

	Coefficients	Standard	t Stat	P-value	Lower 95%	Upper 95%
		Error				
Intercept	65.805	3.448	19.084	0.000	58.996	72.615
X_1	0.010	0.012	0.797	0.427	-0.014	0.034
X_2	-	-	-	-	-	-
X_3	0.007	0.002	4.607	0	0.004	0.010
X_4	0.187	0.042	4.472	0.000	0.105	0.270
X_5	0.008	0.002	5.058	0.000	0.005	0.011
X_6	-0.025	0.025	-1.028	0.306	-0.074	0.023
X_7	-0.016	0.111	-0.148	0.882	-0.236	0.203
X_8	0.430	0.221	1.947	0.053	-0.006	0.867
X_9	0.064	0.049	1.310	0.192	-0.033	0.161
X_{10}	0.445	0.427	1.041	0.299	-0.399	1.289

Table C.3: Regression models: RTO KB-R03

Regression model input variables

Series ID	Y'	Y_S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KB-R03	Y	-	X	-	X	X	X	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
KB-R03	RTO	0.976	0.953	0.945	21.891	172

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
KB-R03	-52.825	-8.696	5.233	16.082	77.514

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1589344.928	158934.493	331.663	6.177E-102
Residual	162	77631.134	479.205		
Total	172	1666976.063			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-	-	-	-	-	-
X_1	0.038	0.028	1.378	0.170	-0.017	0.094
X_2	-	-	-	-	-	-
X_3	0.013	0.003	3.826	0.000	0.006	0.019
X_4	0.647	0.081	7.971	0.000	0.487	0.807
X_5	0.015	0.004	4.104	0.000	0.008	0.022
X_6	0.160	0.055	2.934	0.004	0.052	0.268
X_7	0.605	0.250	2.420	0.017	0.111	1.098
X_8	0.378	0.504	0.751	0.454	-0.617	1.374
X_9	0.321	0.109	2.949	0.004	0.106	0.535
X_{10}	-0.327	0.976	-0.335	0.738	-2.255	1.600

Table C.4: Regression models: RTO KB-R04

Regression model input variables

Series ID	Y'	Y_S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KB-R04	-	Y	X	-	X	X	X	X	X	X	X	X

Regression statistics

Model ID	Algorithm	Multiple R	R-square	Adjusted R-square	Standard Error	Observations n
KB-R04	RTO	0.984	0.968	0.960	17.269	172

Residuals

	Min.	First quartile Q_1	Medium $\tilde{x}$	Third quartile Q_3	Max.
KB-R03	-50.493	-8.034	2.123	13.292	60.423

ANOVA

	df	SS	MS	F	Significance F
Regression	10	1475129.349	147512.935	494.623	2.408E-115
Residual	162	48313.789	298.233		
Total	172	1523443.139			

Model outputs

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-	-	-	-	-	-
X_1	0.036	0.022	1.656	0.100	-0.007	0.080
X_2	-	-	-	-	-	-
X_3	0.016	0.003	5.981	0.000	0.010	0.021
X_4	0.612	0.064	9.565	0.000	0.486	0.739
X_5	0.018	0.003	6.323	0.000	0.012	0.023
X_6	0.085	0.043	1.979	0.049	0.000	0.171
X_7	0.372	0.197	1.888	0.061	-0.017	0.761
X_8	0.694	0.398	1.746	0.083	-0.091	1.479
X_9	0.283	0.086	3.298	0.001	0.113	0.452
X_{10}	0.204	0.770	0.265	0.791	-1.317	1.725

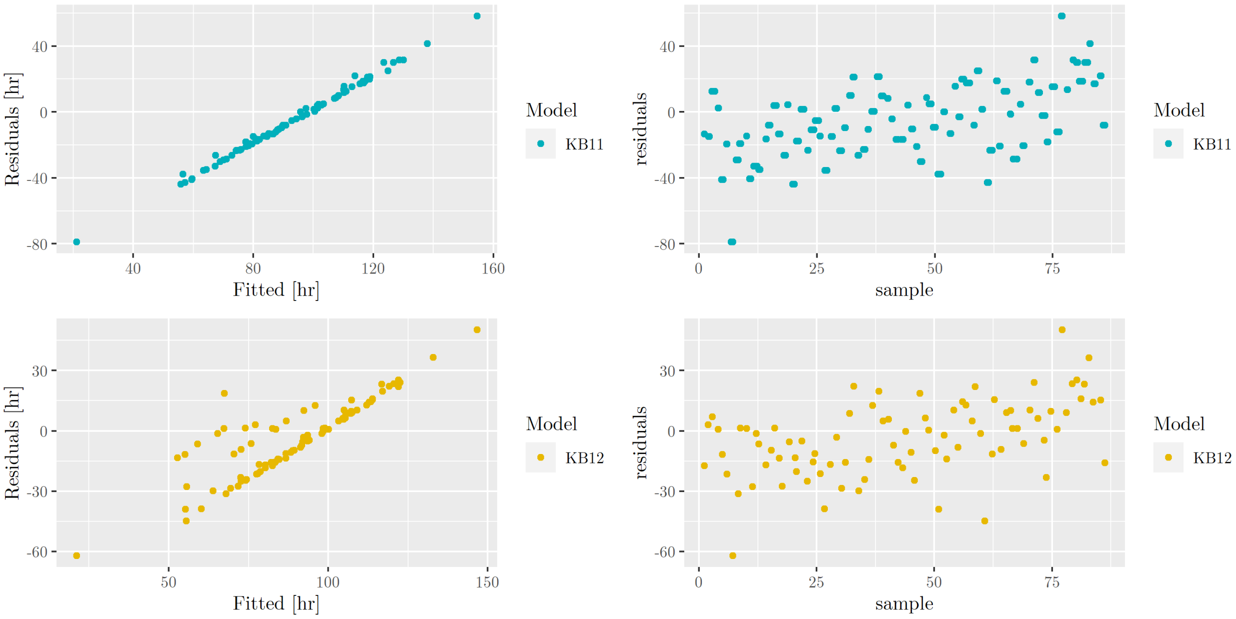


Figure C.1: Residual vs fitted plot - Kongsberg-tunnel

C.1.2 Correlation matrix

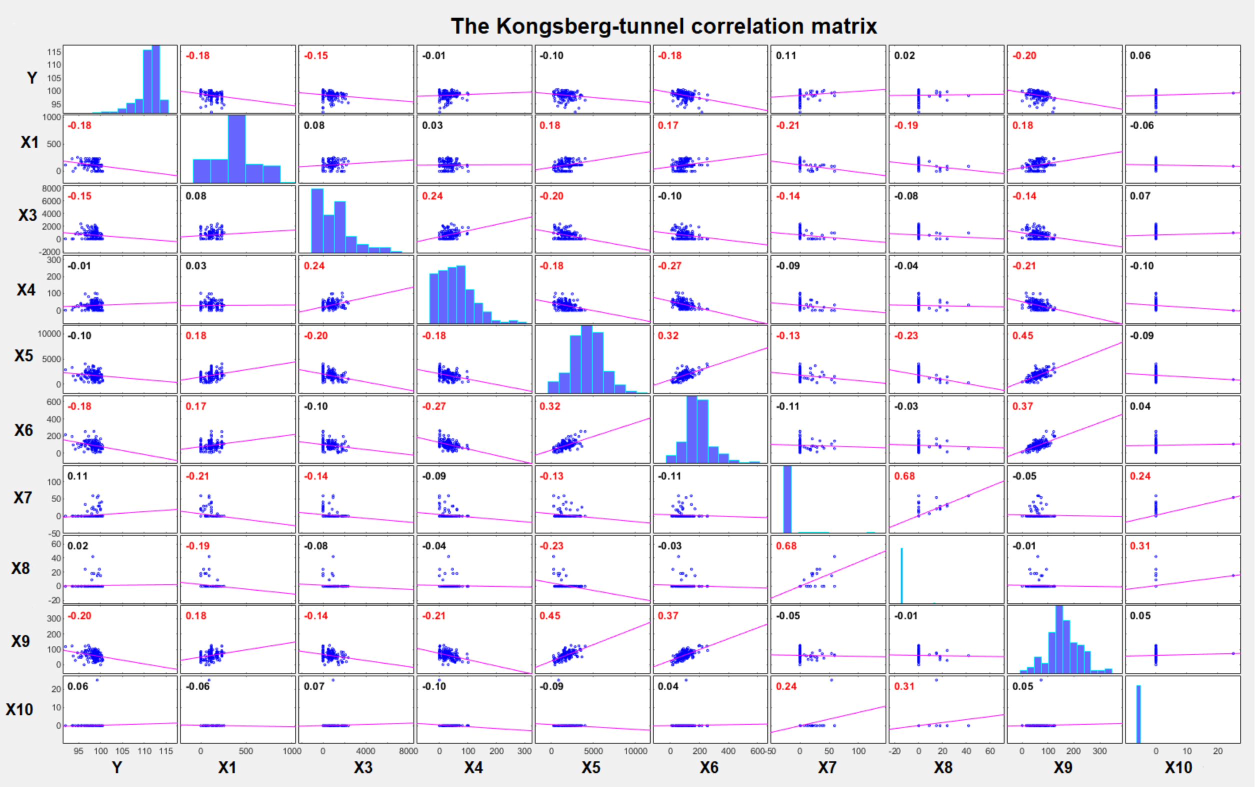


Figure C.2: The Kongsberg-tunnel correlation matrix

Appendix D

The Kangaroo-tunnel

A collection of additional information and results from the Kangaroo-tunnel data analysis has been included in this appendix.

D.1 Data simulation process

In this section, the data simulation process is described.

Table D.1: Description of how weights are imposed to create the pseudo-random Kangaroo-tunnel dataset

Characteristic	**Solution**	**Example**
Each construction task has its own unique range of possible values	The average range of quantities is to resemble that of existing (typical) tunnels	When considering X_5 (excavated material per week), it would be unrealistic to generate random values between 0 and 10 m^3, or even an arbitrary value between 0 and 100,000 m^3
The quantity for each construction task has its own unique density (frequency)	Examine the histograms (frequency distributions) of existing tunnel projects, and impose a weight to the randomly generated values accordingly	When considering X_{10} (construction of arches), it would be unrealistic to generate random values at an evenly or even normally distributed pace. The erection of arches is uncommon and therefore, values generated should be weighted closer to 0 for the most part.

D.2 NNLS model results: Kangaroo-tunnel

Table D.2: Kangaroo-tunnel KR10 (0 to 2 hr noise) dataset description and model outputs

Dataset description						
Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR11	0 to 2	188	100 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR12	0 to 2	746	95 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR13	0 to 2	3844	80 to 105	10	$X_1 \ldots X_{10}$	NFF, constant

Regression coefficients										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KR11	0.017	0.025	0.017	0.673	0.020	0.067	0.138	0.039	0.126	0.252
KR12	0.016	0.025	0.017	0.673	0.020	0.067	0.133	0.041	0.126	0.252
KR13	0.017	0.025	0.017	0.673	0.020	0.068	0.135	0.040	0.126	0.252

Time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR11	59.32	39.84	59.02	1.49	49.57	14.97	7.26	25.35	7.91	3.97
KR12	61.18	39.54	59.25	1.49	49.40	14.93	7.55	24.33	7.95	3.96
KR13	59.60	39.28	59.33	1.49	49.42	14.79	7.42	25.04	7.91	3.96
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table D.3: Kangaroo-tunnel KR20 (0 to 5 hr noise) dataset description and model outputs

Dataset description						
Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR21	0 to 5	208	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR22	0 to 5	821	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR23	0 to 5	4108	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Regression coefficients										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KR21	0.017	0.025	0.017	0.678	0.021	0.069	0.140	0.043	0.132	0.253
KR22	0.017	0.026	0.017	0.681	0.021	0.070	0.138	0.044	0.127	0.257
KR23	0.018	0.026	0.017	0.683	0.021	0.070	0.136	0.040	0.129	0.256

Time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR21	57.59	40.14	58.92	1.47	48.51	14.41	7.15	23.10	7.59	3.95
KR22	58.93	38.18	58.38	1.47	48.64	14.32	7.24	22.83	7.86	3.90
KR23	56.73	38.36	58.44	1.46	48.67	14.22	7.33	25.17	7.75	3.91
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table D.4: Kangaroo-tunnel KR30 (0 to 10 hr noise) dataset description and model outputs

Dataset description						
Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR31	0 to 10	241	100 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR32	0 to 10	926	95 to 102	10	$X_1 \dots X_{10}$	NFF, constant
KR33	0 to 10	4669	80 to 105	10	$X_1 \dots X_{10}$	NFF, constant

Regression coefficients										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KR31	0.016	0.022	0.017	0.696	0.021	0.076	0.143	0.043	0.136	0.264
KR32	0.017	0.028	0.018	0.695	0.021	0.072	0.134	0.043	0.135	0.261
KR33	0.018	0.028	0.017	0.694	0.021	0.075	0.136	0.042	0.135	0.259

Time capacity values										
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR31	61.59	45.13	57.93	1.44	47.38	13.22	6.98	23.16	7.38	3.78
KR32	57.69	36.08	57.02	1.44	47.16	13.93	7.44	23.47	7.38	3.82
KR33	54.33	35.50	57.22	1.44	47.16	13.27	7.36	23.60	7.40	3.86
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table D.5: Kangaroo-tunnel KR40 (0 to 15 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR41	0 to 15	286	100 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR42	0 to 15	1075	95 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR43	0 to 15	5207	80 to 105	10	$X_1 \ldots X_{10}$	NFF, constant

Regression coefficients

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KR41	0.021	0.020	0.018	0.694	0.022	0.075	0.168	0.052	0.142	0.263
KR42	0.020	0.031	0.018	0.701	0.022	0.078	0.134	0.052	0.134	0.262
KR43	0.020	0.030	0.018	0.707	0.022	0.079	0.151	0.045	0.139	0.261

Time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR41	47.62	50.76	54.75	1.44	45.93	13.34	5.96	19.28	7.04	3.80
KR42	50.63	32.67	56.37	1.43	45.24	12.76	7.46	19.13	7.44	3.82
KR43	49.89	33.74	55.64	1.41	45.64	12.62	6.62	22.41	7.17	3.83
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

Table D.6: Kangaroo-tunnel KR50 (0 to 20 hr noise) dataset description and model outputs

Dataset description

Model ID	Noise [hr]	Observations	Response [hr]	Dimensions	Predictors	Weights
	ϵ	n	Y	k	X	β
KR51a	0 to 20	293	100 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR51b	0 to 20	282	100 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR52	0 to 20	1254	95 to 102	10	$X_1 \ldots X_{10}$	NFF, constant
KR53	0 to 20	5767	80 to 105	10	$X_1 \ldots X_{10}$	NFF, constant

Regression coefficients

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
KR51a	0.027	0.032	0.018	0.708	0.023	0.085	0.184	0.000	0.142	0.262
KR51b	0.022	0.028	0.018	0.712	0.022	0.081	0.193	0.062	0.161	0.253
KR52	0.024	0.039	0.018	0.707	0.022	0.095	0.136	0.035	0.138	0.262
KR53	0.022	0.033	0.018	0.714	0.022	0.089	0.146	0.053	0.152	0.265

Time capacity values

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
	m/h	m/h	m/h	t/h	m^3/h	unit/h	unit/h	m/h	m^3/h	m/h
KR51a	37.47	30.82	55.60	1.41	43.75	11.83	5.45	0	7.04	3.82
KR51b	44.60	35.90	54.88	1.41	44.68	12.38	5.17	16.05	6.22	3.95
KR52	41.04	25.89	55.10	1.41	44.79	10.51	7.37	28.70	7.24	3.82
KR53	44.95	30.24	54.94	1.40	44.68	11.28	6.85	18.85	6.56	3.77
NFF	60.00	40.00	60.00	1.50	50.00	15.00	7.50	25.00	8.00	4.00

John Lau

Modelling D&B Tunnelling Construction Data

NTNU

Norwegian University of
Science and Technology