Workflow Modelling of Construction Projects

Problem presented by

David Myers (Heathrow)

ESGI138 was jointly hosted by
The University of Bath
The University of Bristol

with additional financial support from
Bath Institute for Mathematical Innovation (IMI)
The Engineering and Physical Sciences Research Council (EPSRC)
GW4 Alliance
The Infrastructure Industry Innovation Platform (I3P)
Industrially Focused Mathematical Modelling CDT (InFoMM))
Innovate UK's Knowledge Transfer Network (KTN)

Report Authors and Contributors

Facilitators: Ellen Murphy[1] and Alan Champneys[2]
Contributors: Hanan Batarfi[3], Edmund Barter[2], Chris Budd[1],
Ran Dong[4], Jan Foniok[5], Kamil Kulesza[6], Andrew Lacey[7], Xiaodong Li[8],
Francisco Rodrigues[9], Alex Wendland[10], Ambrose Yim[11]

1. University of Bath, imi@bath.ac.uk. 2 University of Bristol. 3. King Abdulaziz University. 4. University of Strathclyde. 5. Manchester Metropolitan University. 6. Centre for Industrial Application of Mathematics and Systems' Engineering, Poland. 7. Heriot Watt University. 8. University of the West of England. 9. Instituto de Ciências Matemáticas e de Computação, Brasil. 10. University of Warwick. 11. University of Oxford.

Contents

1 Executive Summary

This report details the work carried out by the Study Group on workflow modelling of construction projects.

Data on the progress of about a hundred projects over a single five-year planning period were provided by Heathrow Airport (the client) and their four Tier 1 construction contractors. These data are mapped and analysed. Several unusual features are discovered. For example, most projects undergo several tens of adjustments in their scope and price such that while most projects are technically completed under budget, the price and duration is significantly higher than originally planned.

The main question addressed was whether an optimised scheduling of the project would lead to decreased costs and more rapid completion.

First, a machine learning approach is used to gain insight onto which factors are most significant in predicting the final cost and duration of each project. If more data were available, these methods could be further exploited to allow for predictions to be made on which projects are likely to over-run or go over budget and to examine connections between projects at the subcontractor level.

In addition to the data-centric approach, a complementary mathematical model was developed to gain a better understanding of the effect of resource constraints on cost and price extension due to resource competition of concurrent projects, ignoring the confounding effect of scope creep seen in the data. The model takes the form of a discrete time stochastic simulation, whose parameters are fit to the existing data. Tentative conclusions from the model indicate that better outcomes can be achieved by spreading out project start dates, and by prioritising completion of smaller projects.

While more data is needed to validate the model, the results suggested that gains can be made if more thoughtful scheduling of projects is implemented, and also if the prioritisation of projects is monitored and adjusted intelligently.

Our major recommendation to Heathrow Airport is to collect or retrieve more data, as outlined in the report, so that both models can be made more realistic and useful. This would allow Heathrow Airport and their contractors to develop and test strategies to make the system more efficient, ultimately saving time and money.

2 Introduction

Heathrow Airport commission about £1B worth of construction projects per annum. This work is split into approximately 100 separate projects per year, 80% of which are at the <£1M level. Within a framework agreement, each project is awarded to one of four preferred main contractors, with each contractor being responsible for a separate physical region of the airport. These projects are typically planned and approved during a fixed 5-year planning window, mostly independently of each other. This process therefore tends to lead to sub-optimal deployment of the resources available to each contractor. As with most construction projects, the complexities of planning and financing each project tend to lead to an approach where everything is scheduled to be delivered *as soon as possible* once the go-ahead is granted. This is unlike manufacturing sectors (such as automotive) in which planning is done *from the back* to keep factories as close to capacity at all times.

The key question is whether workflow modelling, such as would be used in bulk manufacturing, could lead to a more optimal result. The things to be optimised include costs, both to the client (the airport) and the supplier (the contractor). Other factors to be optimised include quality of delivered projects and minimisation of over-runs to enable robust planning. An optimal solution would take into account the benefits of even employment levels for the contractors and their supply chain, and minimisation of frustration among the airport's users (the airlines and their passengers).

Problem complexity

The complexity of the problem comes from the unusual array of stakeholders of the UK's flagship national airport, such as the UK Government regulators who set the airport charges, the airlines whose representatives desire the efficient running of the airport for the benefit of all the carriers, the owners of Heathrow airport who aim to provide the best airport service in the world whilst delivering consistent returns to their shareholders, and the preferred contractors together with their supply chain. Other complexities come from the planning, regulation, security and logistical constraints that are inherent in working in a 24/365 airport. The construction industry itself is unlike manufacturing in many other respects because most projects are treated as once-off bespoke *make-to-order*. Yet any owner of a large portfolio of infrastructure assets such as an airport, various public sector organisations or utility companies, might find it advantageous to take a more rational approach to scheduling its construction projects.

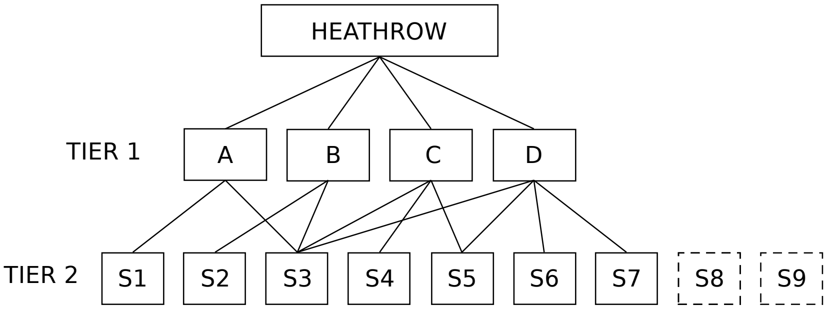


Figure 1: An illustration of how projects are undertaken. Heathrow Airport commission projects from four Tier 1 contractors. In turn, the contractors engage Tier 2 subcontractors to carry out works, of which there are more than 90.

Current practice

Heathrow Airport commission construction projects from four Tier 1 contractors, denoted contractors A, B, C and D for the purposes of the Study Group, each responsible for a different physical region of the airport. The Tier 1 contractors often engage Tier 2 subcontractors to carry out aspects of the work. This system is illustrated in Figure 1. There are more than 90 subcontractors employed by the contractors and they have various specialities, from design works to asphalt laying.

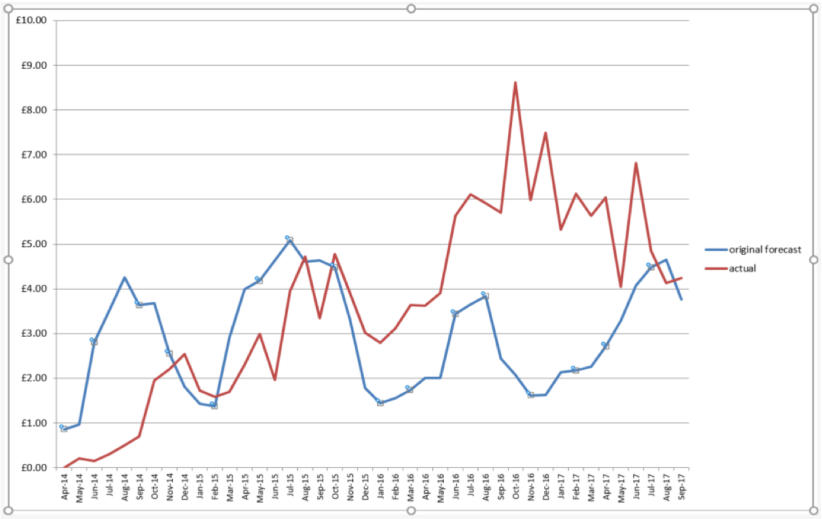


Figure 2: The forecasted (blue line) and the actual (red line) cashflow per month for Heathrow Airport construction projects.

Construction projects have two distinct phases. The first is the project definition, carried out by Heathrow Airport at an average spend rate of £6M per month. Once the investment decision has been made, projects enter the construction phase undertaken by the contractor at an average spend rate of £44M per month. Upon completion of the construction phase, the project is handed back to the airport for general operations.

The capacity of the four contractors is flexible and can increase and decrease to meet the needs of Heathrow Airport. Currently, no significant effort is made by Heathrow to balance the work rate of the contractors over time. In addition, due to the culture of the construction industry, the end dates of projects are often set very aggressively. To facilitate this behaviour, the resources of the contractors are sized to meet peak demand. In practice, the lack of robust scheduling can result in monthly cashflows that are widely different from the original forecasts, as seen in Figure 2.

Mathematical questions

The key mathematical questions presented to the study group were the following:

1. Can we produce a model of the predicted outcome in terms of quality, cost and time over-run, and satisfaction of client, user and contractor for a given schedule of construction projects for a single contractor?
2. Can we learn parameters of such a model using machine learning and expert elicitation?
3. For a given set of projects within a given time window, can we construct a utility function for which we can seek an optimum schedule?
4. How could such an optimisation problem be validated?

3 Data and Information Provided

Each of the four contractors provided us with an anonymised list of all of the projects undertaken in current 5-year planning window. This resulted in data for 84 projects, some of which are groupings of smaller subprojects. For each project, if the data was available, we were given:

Project ID
Contract Price
Change to the Prices
Current Total of the Prices
Latest Quotations
Proposed Total of the Prices

Forecast to Completion Difference
Complete (Y/N)
Subcontractor 1
Subcontractor 2
Subcontractor 3
Starting Date

Original Completion Date
Current Completion Date
Change to Completion Date
Planned Completion Date
Actual Completion Date
Terminal Float
Assessment Date
Payment Due By Date
Provisional Change in Amount Due
Confirmed Change in Amount Due
Amount Due
Sum of Amounts Marked as Paid

3.1 Data

The variables most used during the Study Group are written in bold font: *Current Total of the Prices* is the cost of the project; *Subcontractor 1, 2 and 3* are the top three subcontractors used on a project; and *Starting Date* and *Original and Planned Completion Date* give the estimated and planned length of a project.

To visualise the projects for each contractor, they are plotted below in Figure 3. The bar charts in the lower half of the graphs show the individual projects for a given contractor. The width of the bar is proportional to the cost per month of the project. The blue section of a bar marks the original project duration and the orange section denotes the duration between the original completion date and the planned completion date. The green sections show when the projects finished before their original completion dates. The black data lines in the upper halves of the graphs show the total cost per day, calculated by dividing the total cost by the number of days in a project. The vertical black line marks the 18th of July, 2018.

To understand the typical nature of projects the distributions of the relevant project descriptors were found, as shown in Figure 4. The cost per month, shown in Figure 4a, is calculated by dividing the *Current Total of the Prices* by the length of the project in days and then multiplying by 30.42 (the average number of days in a month). Cost per month has a log-normal distribution, with $\mu = -1.3$ and $\sigma = 1$. The mean and median costs per month are £430k and £270k, respectively.

The length of the project, displayed in Figure 4b, is defined as the duration between the *Starting Date* and the *Planned Completion Date*. The project duration also displays a log-normal distribution, with $\mu = 2.9$ and $\sigma = 0.7$. The mean and median project durations are 21.5 and 19 months, respectively.

Figure 4c displays the project start dates, measured in days from the 1st of April, 2014. The start dates follow a normal distribution with mean $\mu = 724$ (equivalent to the 25th of March, 2016) and standard deviation $\sigma = 357$ (approx. 1 year). The factor by which the project over-ran can be calculated by dividing the time between the *Start Date* and the *Planned Completion Date* by the original project duration i.e., the time between the *Start Date* and

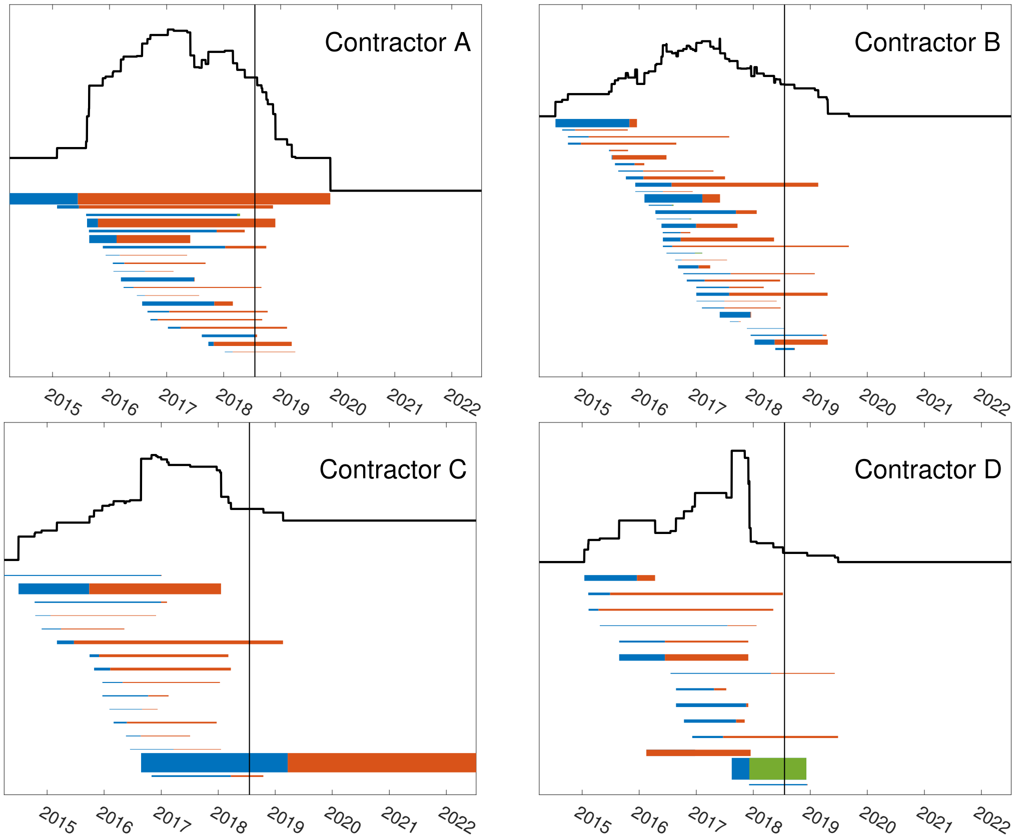
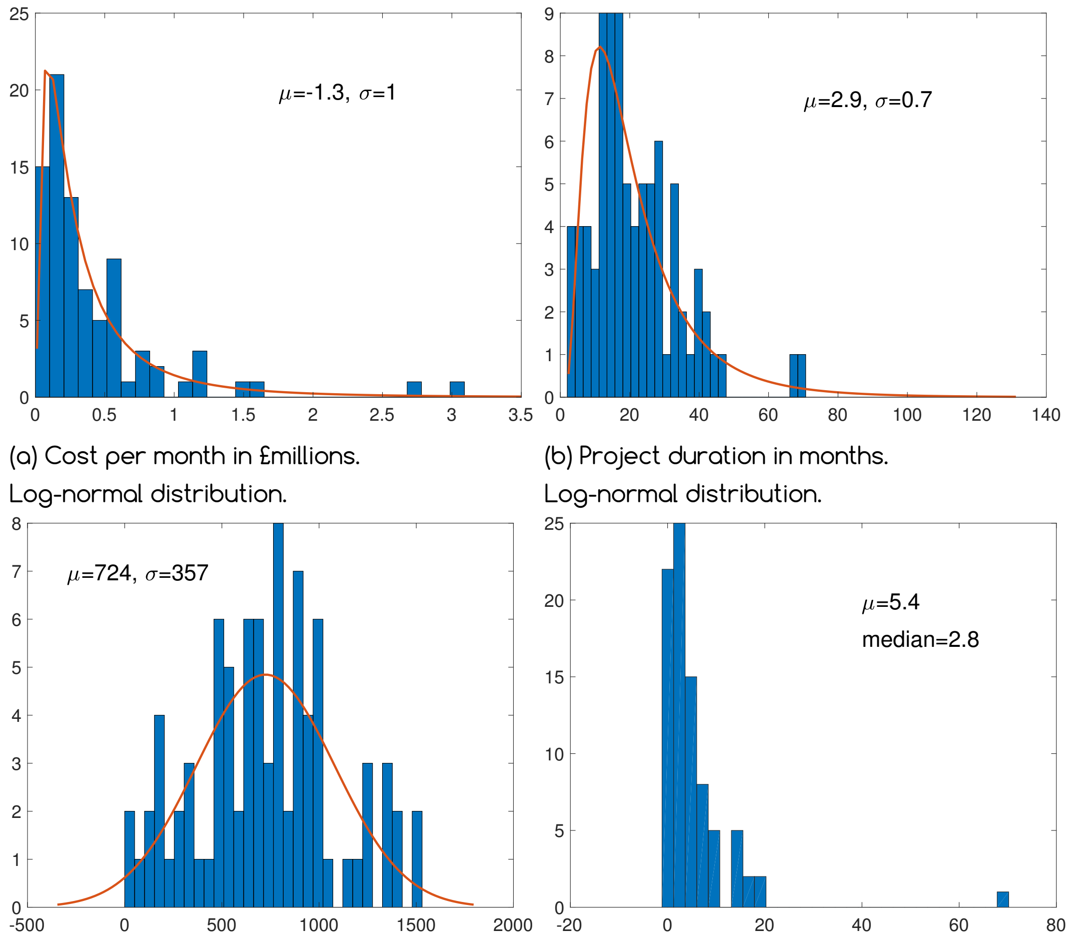


Figure 3: The black lines in the upper halves of the graphs show the total cost per day, calculated by dividing the total cost by the number of days in a project, for each of the four contractors. Below the black lines, the bars show the individual projects for each contractor, plotted in order of starting date. The blue section marks the original duration and the orange denotes the duration between the original completion date and the planned completion date. The green sections are when the projects finished before their original completion dates. The width of the bars is proportional to the cost per day of the project.The vertical black line marks the 18th of July, 2018.



(a) Cost per month in £millions.
Log-normal distribution.

(b) Project duration in months.
Log-normal distribution.

(c) Starting date measured in days from the 4th of April, 2014.
Normal distribution.

(d) The planned project duration divided by the original project duration i.e., the factor by which the project over-ran.

Figure 4: Histograms, in blue, and the fitted distributions, in orange, of key attributes of the project data set.

the *Original Completion Date*. This distribution is displayed in Figure 4d. The mean over-run factor of a project is 5.4, while the median is 2.8, with the majority of projects taking at least twice the original estimate.

In addition to the data described above, we were also given more detailed data on three linked projects. However, because this data arrived late in the week of the Study Group, it was not possible to perform on it any significant analysis.

3.2 Qualitative Information

Before and during the week of the Study Group, representatives from Heathrow Airport and the four contractors were available to answer questions and give context to the quantitative data provided. The information gleaned from these discussions played a significant role in the development of the model and so will be described here.

The most striking feature of the individual project data in Figure 3 is that for most projects, the actually duration of the projects is significantly longer than the original, on average over-running by 50%. Visual inspection of the data in this figure shows that this is most significant for the longer and larger projects. Another surprising feature from the data was that almost no project appears to come in over budget.

After consultation with the contractors, it was discovered that projects are repeatedly revised throughout their duration. On average, a project may have of the order of 50 price adjustments throughout its duration. The chief reasons given for such adjustments are changes in the scope of a project, or unforeseen circumstances. Most of these adjustments are agreed between contractor and client, through their framework agreement so that the *Current Total of the Prices* on any project typically bears no relation to the original *Contract Price*. In fact, for some projects, a contract is issued before the final scope is understood and the initial contract price contains just the design costs. The final price can therefore be several orders of magnitude larger. There is also evidence with one contractor that a single large contract was issued, and additional smaller projects were continually added to this same contract rather than issue new contracts.

For these reasons, it is hard to gain direct evidence from the data on the extent to which project costs increase due to 'unforeseen circumstances' that could have been eliminated through greater operational efficiency, due to genuine unforeseen circumstances, or due to changed scope. It is also hard to gain direct evidence on how a price may have increased due to inefficiency, because generally all price changes are agreed between the Tier 1 contractor and Heathrow Airport. Therefore we relied on anecdotal and expert elicitation evidence from Tier 1 contractors to gain an understanding of the effect of inefficiencies.

While Heathrow Airport have visibility of the current and planned projects undertaken by the four Tier 1 contractors, they do not have visibility of the activity of the lower tier contractors. As a result, it falls to the Tier 1 contractors and subcontractors to determine the scheduling of shared resources. Occasionally, Tier 1 contractors will discuss and plan a programme of use of shared resources at the outset of projects.

Tier 2 and 3 contractors can differ substantially in how the operate. Some subcontractors must be used by all Tier 1 contractors. While not a subcontractor, some Heathrow resources can be considered to fall in to this category as the Tier 1 contractors have no choice in their use, and so they must be shared amongst all Tier 1 contractors. In contrast, some subcontractors work exclusively for a particular Tier 1 contractor, while others compete for work.

The contractors agreed that there was an optimal level of activity and that inefficiencies resulted from working either above or below this level. When working below the optimal level of activity, there are inefficiencies due to a lack of resource sharing, e.g. employing a health and safety officer with capacity to oversee three projects but working on fewer. On the other hand, working over the optimal level can result in poor-decision making at the manager level, potentially causing delays due to rework.

Another common cause of delays is unforeseen complexities in a project. In general, delays result in additional costs, most often shared between Heathrow and the contractor.

4 Machine Learning Approach

We can analyze the Heathrow data from a machine learning perspective. In this case, the goal is to predict a target variable from other measures. Initially, we are interested in studying how projects are connected. We assume that two projects are connected if they share at least one subcontractor. For the considered database, Figure 5 presents the network of projects.

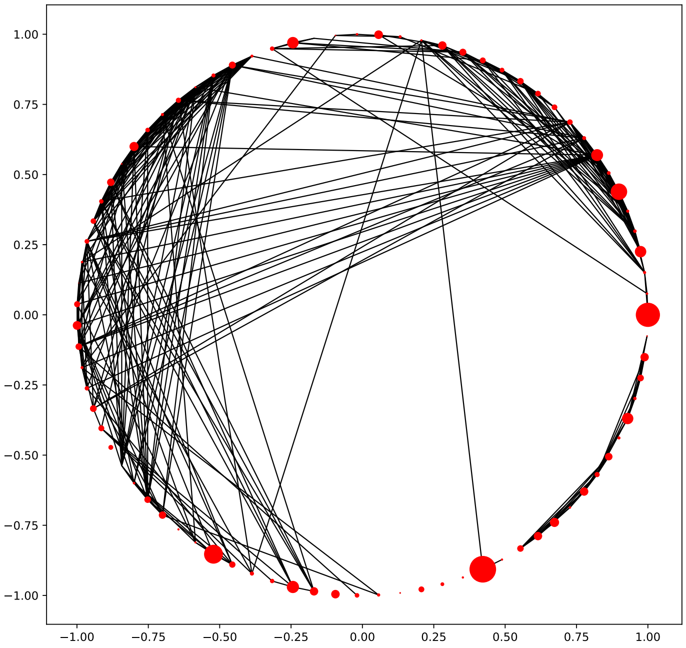


Figure 5: The network of projects. Each node represent a project and two projects are connected when they share at least one subcontractor. The size of the node is related to the cost of the project.

From the network of projects, we can verify whether there is a tendency that projects of similar costs are connected. In this case, we construct an assortative mixing matrix as shown in Figure 6. The Pearson correlation coefficient between the price of connected projects is $r = 0.19$, which indicates that there is a low tendency that projects developed by the same subcontractors tend to present similar costs. This behavior may be related to the number of workers and the specialization of a given subcontractor. For instance, some subcontractors may build rooms, while other ones are more specialized in performing the maintenance of airport runway. These two tasks involve different amount of resources and,

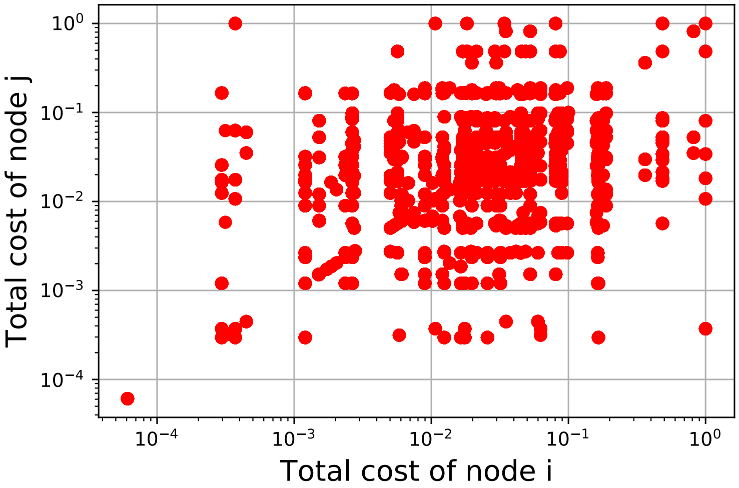


Figure 6: Assortative mixing matrix obtained from connected projects.

therefore, it is expected that they do not share projects.

By considering the database available, we can also try to predict the cost associated to each project from other features. Our goal is to predict the output variable Y, representing the total cost (variable *Current Total of the Prices* in the data set), from a set of input features $\mathbf{X} = \{X_1, X_2, ..., X_d\}$, where X_i is a feature considered in our analysis, i.e., (i) contract price, (ii) forecast to completion difference, (iii) difference between the starting and original dates, (iv) change to completion date, and (v) amount due. We assume a statistical learning model as:

$$Y \approx f(\mathbf{X}) + \epsilon \tag{1}$$

where ϵ is a random error independent of $\mathbf{X}$ with mean zero and variance σ. The goal here is to estimate the function $f : \mathbb{R}^d \to \mathbb{R}$, where d is the number of features in the vector $\mathbf{X}$.

To estimate f, we use a supervised learning method known as a random forest. Random forests are based on an ensemble of decision trees. Decision trees construct a tree-like structure by recursively partitioning the data into subsets according to a target variable. Random forests construct multiple decision trees and merge them together to get an accurate prediction. The algorithm considers the bagging ensemble learning method — successive trees that do not depend on earlier trees are generated and a simple majority vote is taken for prediction. Unlike linear regression models that calculate the coefficients associated with the variables, tree regression models quantify the relative importance of variables. As well as being accurate, random forests have few parameters to tune, like the number of trees and the minimum sample leaf size.

Figure 7 shows the prediction of the cost by using the regression model in equation 1. As

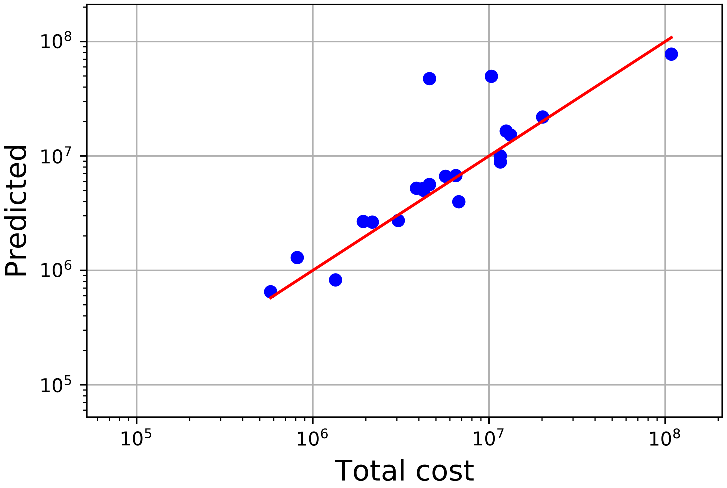


Figure 7: Prediction of the cost by the random forest algorithm.

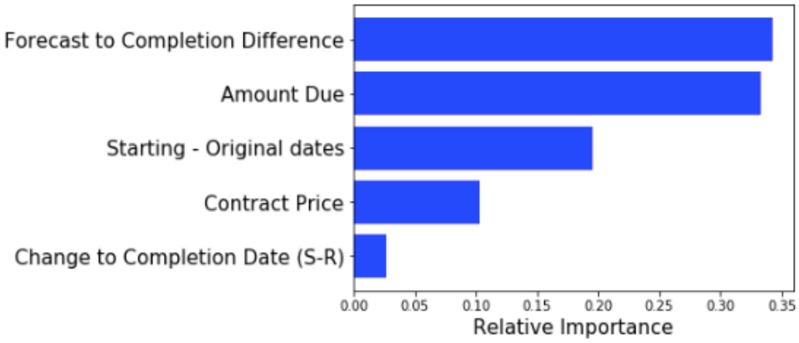
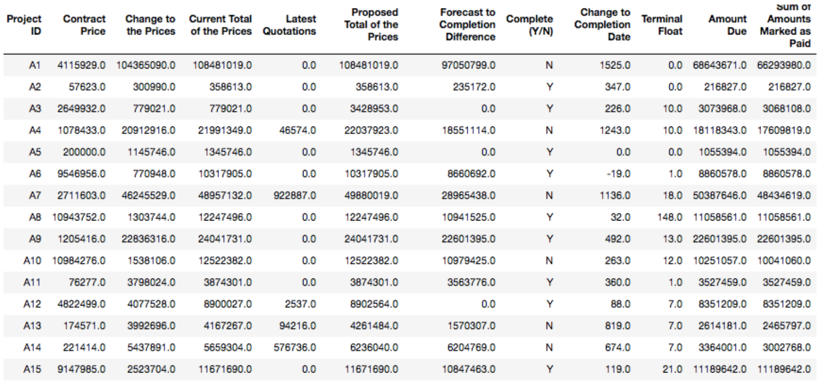


Figure 8: Feature importance on the prediction of the prices, obtained from the random forest algorithm.

we can see, we can predict the cost associated to a project from the other features. Moreover, the random forest algorithm enables us to quantify the importance of each variable on the prediction. The results are presented in Figure 8. As we can see, the *Forecast to Completion Difference* and the *Amount Due* are the variables that most influence the prediction of the final cost. On the other hand, the *Initial Contract Price* is not strongly related to the total cost.

This regression analysis shows that we can predict the cost of each project from variables associated with it. If we had other relevant information, for example the total number of workers, the amount of the material used in each project and the type of the project, then we could construct a more accurate model. Moreover, we would be able to quantify the importance of each feature on the cost and, therefore, be able to change the costs associated with each project by changing its associated variables. This study would enable us to



Project ID	Contract Price	Change to the Prices	Current Total of the Prices	Latest Quotations	Proposed Total of the Prices	Forecast to Completion Difference	Complete (Y/N)	Change to Completion Date	Terminal Float	Amount Due	Sum of Amounts Marked as Paid
A1	4115929.0	104365090.0	108481019.0	0.0	108481019.0	97050799.0	N	1525.0	0.0	68643671.0	66293980.0
A2	57623.0	300990.0	358613.0	0.0	358613.0	235172.0	Y	347.0	0.0	216827.0	216827.0
A3	2649932.0	779021.0	779021.0	0.0	3428953.0	0.0	Y	226.0	10.0	3073968.0	3068108.0
A4	1078433.0	20912916.0	21991349.0	46574.0	22037923.0	18551114.0	N	1243.0	10.0	18118343.0	17609819.0
A5	200000.0	1145746.0	1345746.0	0.0	1345746.0	0.0	Y	0.0	0.0	1055394.0	1055394.0
A6	9546956.0	770948.0	10317905.0	0.0	10317905.0	8660692.0	Y	-19.0	1.0	8860578.0	8860578.0
A7	2711603.0	46245529.0	48957132.0	922887.0	49880019.0	28965438.0	N	1136.0	18.0	50387646.0	48434619.0
A8	10943752.0	1303744.0	12247496.0	0.0	12247496.0	10941525.0	Y	32.0	148.0	11058561.0	11058561.0
A9	1205416.0	22836316.0	24041731.0	0.0	24041731.0	22601395.0	Y	492.0	13.0	22601395.0	22601395.0
A10	10984276.0	1538106.0	12522382.0	0.0	12522382.0	10979425.0	N	263.0	12.0	10251057.0	10041060.0
A11	76277.0	3798024.0	3874301.0	0.0	3874301.0	3563776.0	Y	360.0	1.0	3527459.0	3527459.0
A12	4822499.0	4077528.0	8900027.0	2537.0	8902564.0	0.0	Y	88.0	7.0	8351209.0	8351209.0
A13	174571.0	3992696.0	4167267.0	94216.0	4261484.0	1570307.0	N	819.0	7.0	2614181.0	2465797.0
A14	221414.0	5437891.0	5659304.0	576736.0	6236040.0	6204769.0	N	674.0	7.0	3364001.0	3002768.0
A15	9147985.0	2523704.0	11671690.0	0.0	11671690.0	10847463.0	Y	119.0	21.0	11189642.0	11189642.0

Figure 9: First 14 rows and variables used to predict if a project is finished in a given time.

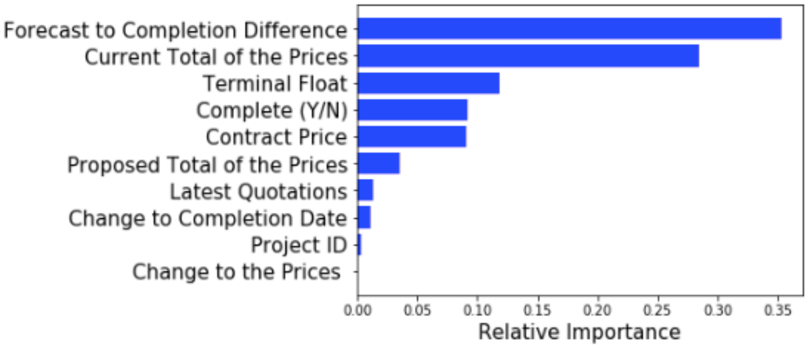


Figure 10: Feature importance for classification of the state of the projects (finished and incomplete).

control the cost by changing a small number of attributes associated to each project and a more suitable allocation of resources in order to avoid the waste of resources.

By using machine learning, we can also predict if a project is completed (or not) in a give time from a set of variables associated to it. Figure 9 shows the data we consider in this analysis. The variable associated to the project finalization can assume two values, zero or one, where one indicates if a project is finished. Thus, we have a binary classification and we can use the random forest algorithm again. Indeed, the statistical learning model we consider is the same as shown in equation 1, but now $Y = \{0, 1\}$, instead of a continuous variable, as in the cost prediction.

The use of the random forest algorithm provides an accuracy of 75% in the prediction, showing that we can predict the state of a project from other variables. Moreover, in Figure 10 we show the variables most related to the project state. As we can see, the *Forecast to Completion Difference* and the *Current Total of Prices* are the variables that most influence the state of the project.

The analysis presented here show that we can predict target variables associated to project and detect the features that most influence these target variables. The analysis presented here is just an initial analysis and further studies can be performed with additional data and statistical learning models. The use of the methods can help Heathrow to get a better allocation of subcontractors, increase its efficiency and reduce the waste of resources.

5 Mathematical Model

We now describe a simple, proof-of-concept mathematical model, developed during the week to gain a better understanding of the system.

5.1 Overview

The aim of this simulation model is to estimate the duration and cost of each project planned and executed within a given fixed time period. We fundamentally want to understand whether a more planned approach to project scheduling can lead to better outcomes. For this purpose we want to separate delays and over-runs due to scope increase from those that occur due to genuine operational over-runs.

The central tenets of the model are:

- Cost and price are essentially the same thing, because most changes in cost are shared between contractor and client;
- It is possible to know at the outset of each project the total amount to be done, of its total amount of work to be done amount of work to be done work to be done for each project, which can be expressed in terms of
  - the project's *budget at completion* W (in pounds), and
  - the total *planned duration* T of the project (in weeks).
- The influences causing delays to projects are of two kinds:
  - *competition for resources* among projects, leading to unavailability of certain resources (workers, subcontractor project management, specialist equipment, etc.) for some projects;
  - and *inefficiencies caused by outside influences*, such as weather, regulators, operations, etc.

5.2 Assumptions

For this simple model, we assume that:

1. There are a finite number of projects that are to be started within a finite given time window.

2. There are no changes to projects.
   The scope (extent) of a project is fixed throughout. (For example, if a project is originally planned as the replacement of all windows in a building, we do not allow it to be extended to another building.) Similarly, the planned duration of the project is fixed, and, in particular, not adjusted depending on the actual progress made.

3. Uniform planned progress.
   We use a discrete time model, with time intervals representing weeks. We assume that the work done is planned to increase linearly, by W/T in any given week. That is, the planned 'intensity' of work is constant throughout the duration of the project. In reality, for many projects the initial and final phases of its execution will see a lower intensity of work.

4. Over-run increases cost.
   We assume that over-runs of projects increase the total cost. Cost is measured in terms of resources R_j used in week j. These resources may be broken down into labour costs, depreciation, raw material, subcontractor costs, etc. but we shall lump these into a single variable R_j which we shall refer to as the *Manpower* (or womanpower) used in by a project in a given week.

5. Increase in work intensity for projects behind schedule.
   When a project starts falling behind schedule, project managers will try to keep the originally planned completion date by increasing work intensity proportionately. However, there is a cap on this increase, so they will never want to increase the work done by more than cW/T for a relatively small constant c (in our simulations, we use $c = 1.5$).

6. Projects compete for resources.
   When several projects are carried out at the same point in time, competition for resources may frustrate their progress. The more work-intensive the concurrent projects are, the more severe the delays will tend to be.

7. Work is not always carried out fully efficiently.
   Even with the assigned resources, efficiencies may be influenced by environment conditions such as weather, regulators, operations, etc. So efficiency is modelled as a random variable. We assume the efficiency in a given week is positively correlated with the efficiency in the previous week.

5.3 Formulation

5.3.1 Working progress

The progress of work on a project is tracked by the variable π_j, for $j = 0, 1, 2, \ldots$, which represents the proportion of the work has been done by week j.

$$\pi_j = \frac{E_j}{W},$$

where E_j is the *work done* on the project j weeks after its start. For example, $\pi_4 = 0.2$ says that four weeks after the project started, 20% of all the work has been done.

Put $\pi_0 = 0$. We will iteratively compute π_j for $j = 1, 2, 3, \ldots$ as follows.

1. The *requested amount of resources* is
$$R_j = W \cdot \min\Big\{\frac{1-\pi_j}{T-j}, \frac{c}{T}\Big\}, \tag{2}$$
where c is a suitably chosen constant.
2. The *available amount of resources* is G_j, which is a random variable depending on the total amount of resources requested by all the projects in a given week; the computation is given below.
3. The *progress made* in week $j+1$ is
$$\Delta\pi_j = e_j \cdot \frac{G_j}{W},$$
where e_j is the *efficiency factor*, which is a random variable further explained below.
4. Put $\pi_{j+1} = \pi_j + \Delta\pi_j$.

This iterative process is repeated until $\pi_j \geq 1$, when the project is completed. Then we can get the *actual duration* of project p as

$$D_p = \min\{j : \pi_j \geq 1\}. \tag{3}$$

In the following five subsections, we provide the details of the above model.

5.3.2 Resources requested by a project

Note that j weeks after the project has started, the work done is $E_j = \pi_j W$, and thus the work that remains to be done is $W - E_j$. This work is to be done in $T - j$ weeks, so we expect

$$\frac{W - E_j}{T-j} = W \cdot \frac{1-\pi_j}{T-j}$$

to be completed each week, which is the first term in the minimum in (2).

The second term in (2) implements our assumption that weekly progress will never be more than c times the originally planned progress.

5.3.3 Resources made available to a project

The calculation of available resources for the project takes into account the total requested resources by all projects as follows. For any given week t, let P_t be the set of all projects that are being carried out. For each $p \in P_t$, the requested amount of resources $R_{p,t-s_p}$ is known; here, s_p is the week in which p started. Let

$$R_t^{\text{tot}} = \sum_{p \in P_t} R_{p,t-s_p}$$

be the total amount of resources requested by all projects in week t.

The total amount of resources available for all projects in week t is given by

$$G_t^{\text{tot}} = \beta \cdot \tanh\left(\frac{R_t^{\text{tot}}}{\beta}\right) \tag{4}$$

for a suitable constant β. This relationship, illustrated in Figure 11, simulates the competition for resources among projects. When many projects request a large amount of resource, not all of these requests will be satisfied. The use of $\tanh$ is not the only possible choice, but it gives us a realistically looking shape for how we imagine the influence of the competition for resources.

Finally, it remains to determine the allocation of the available resources G_t^{tot} among the individual projects. In reality, the allocation will be determined by project managers of the contractors and their subcontractors. Available resources will not always be allocated in proportion to the share of the projects' resource requests; prioritisation and deal-making will typically occur. From the point of view of an external observer not involved in the project management process, however, this allocation will appear random; that is the point of view we take for our model.

So the amount of resources $G_{p,t-s_p}$ allocated to project $p \in P_t$ will be normally distributed around the 'fair' amount:

$$G_{p,t-s_p} \sim N\Big(\frac{R_{p,t-s_p}}{R_t^{\text{tot}}} \cdot G_t^{\text{tot}},\ \sigma^2\Big). \tag{5}$$

We have also implemented variants of the model where the allocation of resources is prioritised depending on the planned duration of the project; we have considered both the case where short projects are prioritised and the case where long projects are prioritised. The results are discussed in Section 6.2.

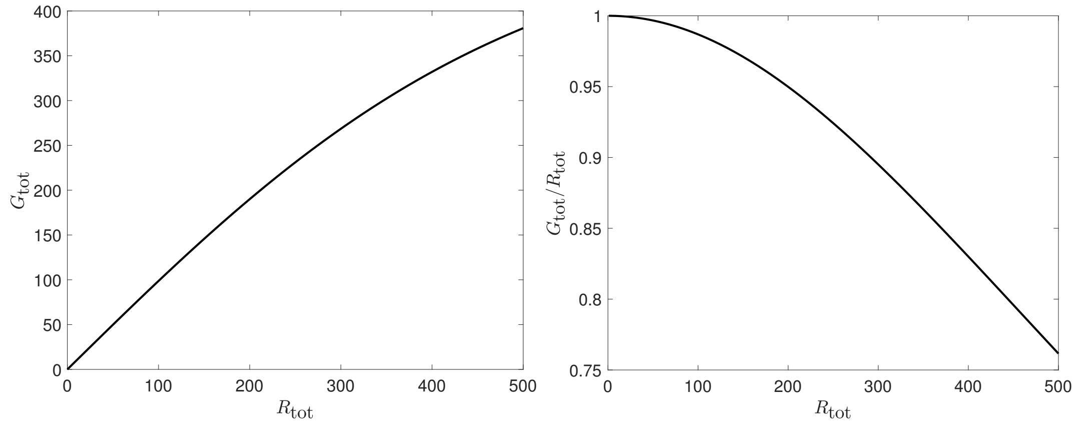


Figure 11: Illustration of resources available as a function of total requested resources (left) and as a proportion of the request (right) for $\beta = 500$.

5.3.4 Technical efficiency factor

The technical efficiency factor models the efficiency of use of the resources allocated to a project. We employ a variation on a theme of the technical efficiency introduced by [1]. Following [2], the technical efficiency would be a random variable

$$e = \exp(-U),$$

where U is a non-negative random variable. Often U is assumed to have an exponential or half-normal distribution. We opt for half-normal, but allow technical efficiency to be larger than 1, although not larger than $\alpha = 1.1$. (The right value of α should be determined by a closer analysis of the available data.) Recall that we assume that technical efficiency in a given week is positively correlated with that in the previous week. Hence, technical efficiency in week j is

$$e_j = \frac{3e_{j-1} + \alpha \exp(-U)}{4}, \tag{6}$$

where $U = |X|$ for normally distributed $X \sim N(0, \sigma'^2)$. Typical behaviour of the efficiency factor in the course of a project is captured in Figure 12.

5.3.5 Starting time of a project

The starting times s_p of all the projects are the decision variables in this model. We may use this model either as a simulation model or as an optimisation model.

In a simulation model, we try various choices of starting times for the projects and compare the results of the simulation. Later we show the results in various scenarios: clustering all

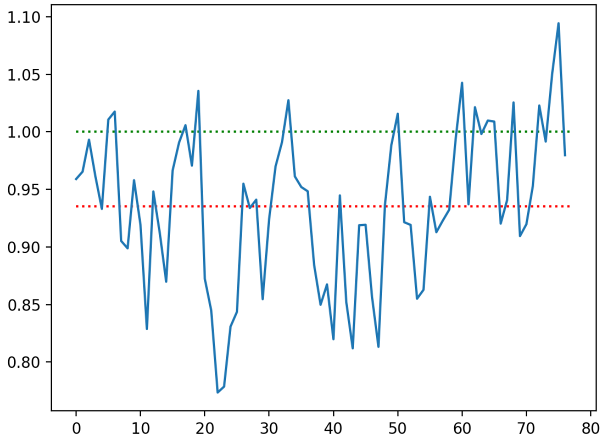


Figure 12: Simulation of the efficiency of a project as a random walk. The green dotted line indicates technical efficiency 1, whereas the red dotted line marks the mean efficiency in the period of 80 weeks.

starting times towards the beginning of a period, spreading them uniformly across a period, or choosing random starting times according to various probability distributions.

In an optimisation model, we would like to minimise the total duration of all projects, that is, to

$$\text{minimise} \quad \sum_{p \in P} D_p,$$

where P is the set of all projects to execute and D_p is the duration of p defined in (3). (We assume that the total cost is closely related to over-runs of projects, so this objective is a good proxy for minimising the total cost as well.)

We have not implemented an optimisation algorithm to solve this problem, but it is a topic for potential future work.

5.3.6 Parameters

The model has the following parameters whose value has to be determined experimentally when validating the model:

- the cap c on required resources of a project behind schedule, in equation (2);
- the denominator β in equation (4);
- the variance of the normal distribution in equation (5);

- α and σ' in the efficiency calculation (6), as well as initial efficiency e_0 for each project, and the weight given to previous week's efficiency in (6).

5.4 Algorithm

The algorithm, described in pseudocode, is in Algorithm 1.

```
Data: time horizon $T$; a set $P$ of projects; for each project $p \in P$, we have its budget at
        completion $W_p$, its planned duration $T_p$, and its starting time $s_p \leq T$
Result: the duration $D_p$ of each project $p \in P$
begin
    $t \leftarrow 0$;
    $P_{\text{unf}} \leftarrow P$;                       // project that haven't yet finished
    $P_{\text{curr}} \leftarrow \emptyset$;                  // projects currently in progress
    forall $p \in P$ do
        $\pi_p \leftarrow 0$
    end
    while $t \leq T$ or $P_{\text{unf}} \neq \emptyset$ do
        forall $p \in P_{\text{curr}} : \pi_p \geq 1$ do          // projects that have finished
            remove $p$ from $P_{\text{curr}}$;
            remove $p$ from $P_{\text{unf}}$;
            $D_p \leftarrow t - s_p$;                     // duration of project $p$
        end
        forall $p \in P : s_p = t$ do                 // projects that are starting
            add $p$ to $P_{\text{curr}}$;
            $e_p \leftarrow 0.9$;                         // initial efficiency
        end
        forall $p \in P_{\text{curr}}$ do
            $R_p \leftarrow W \cdot \min\{\frac{1-\pi_p}{T_p-t+s_p}, \frac{c}{T_p}\}$ ;   // requested resources
        end
        $R^{\text{tot}} \leftarrow \sum_{p \in P_{\text{curr}}} R_p$;
        $G^{\text{tot}} \leftarrow \beta \cdot \tanh(R^{\text{tot}}/\beta)$;
        forall $p \in P_{\text{curr}}$ do
            calculate $G_p$ according to (5);
            calculate $e_p$ according to (6);
            $\pi_p \leftarrow \pi_p + e_p \cdot \frac{G_p}{W_p}$;
        end
    end
    return $(D_p : p \in P)$;
end
```

Algorithm 1: Simulation algorithm to determine project durations

5.5 Illustrations

Here we demonstrate the output of a simulation where we included only five projects, which have 5 different colours.

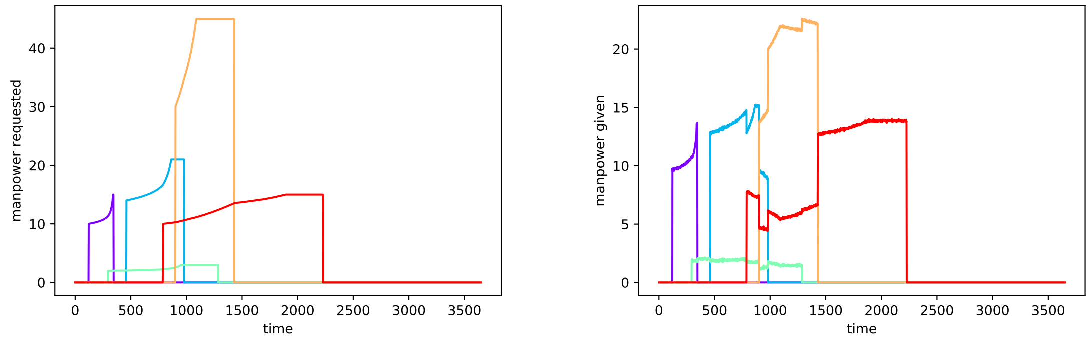


Figure 13: Manpower data of simulation ran with 5 projects.

The graphs in Figure 13 display the data for manpower requested (on the left) and manpower given (on the right). On the left we see as projects fall behind they gradually request more manpower, however this is limited by the hard cap where the requested curves flatten off. In the algorithm, this will be our R_p. On the right we see that when there is no competition projects get what they have requested. Later on we see the red project get severely underfunded when it has to compete with 2 or 3 other projects. In the algorithm, this will be our G_p.

The graphs in Figure 14 display the data for progress made per day (on the left) and cumulative progress (on the right). Here we see how this closely follows the manpower given curves however there is some noise introduced by the efficiency of each project. This is e_p within the algorithm. The figures illustrate that while projects do monotonically increase

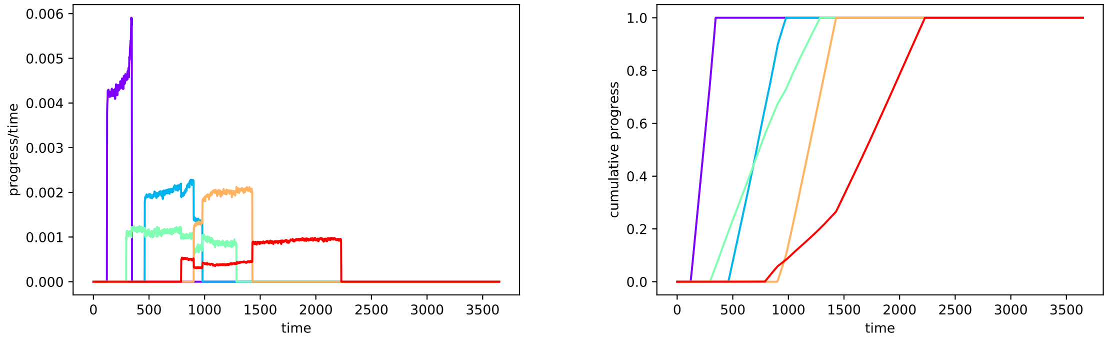


Figure 14: Progress data of simulation ran with 5 projects.

their progress, when there is competition this is at a reduced rate.

5.6 Discussion

Section 5.3.6 lists the many parameters of the model whose values need to be determined when validating the model. This can be done by first gathering more data on actual projects, their starting times, durations, and record of their progress in time; then running the simulation model with various values of its parameters and comparing the simulated results with the actual values. However, it is likely that more detailed data about projects will be needed for this validation to be successful.

Perhaps the least realistic among our assumptions is that of uniform planned progress, that is, that any project is planned to progress equally quickly in its initial phase, in the middle phases, as well as towards the end as it nears completion. In reality, however, a typical project will start slowly, with fast progress in the middle, and finish slowly again. But it is not clear whether our assumption invalidates the results: maybe the simulated durations of the projects will mirror those observed in reality, in spite of making unrealistic assumptions.

In any case, actual data will have to be compared to computational results obtained by running the algorithm, and we provide some initial results in the next section.

6 Simulations and Results

We break this section into two parts. The first deals with the choice of variables within our model and its computational validation against the data we were given. The second discusses different strategies and the results from the model.

6.1 Validation

Before we discuss the validation we want to restate some assumptions and simplifications that we have made for the model.

1. We only simulate one Tier 1 contractor and we don't directly simulate Tier 2/3 contractors.
2. We simulate one project life cycle (5 years).
3. On average, the Tier 1 contractor has 20 projects running in this time period.
4. Projects have no mission creep, they only do what they set out to do.
5. There is competition for resources (this in some way resolves the effect of Tier 2/3 contractors).
6. The duration and costs of the projects are log normally distributed, as evidenced above.
7. In the real data the start times are normally distributed, as above in Section 3.

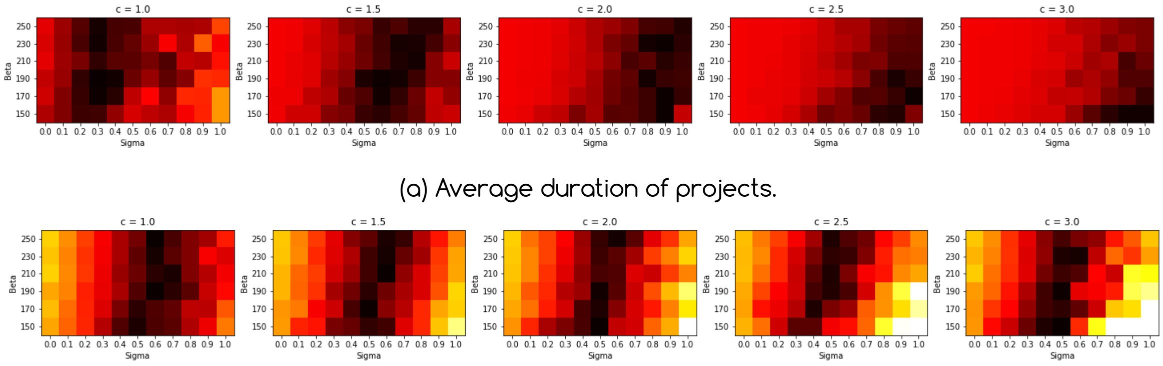
Within the model there are six main variables:

- T - the amount of time steps in the simulation.
- c - the ratio of effort that can be put into a project if it falls behind to the predicted effort.
- $init\_eff$ - The initial efficiency a project starts at.
- β - The competition variable and max global work output, where high values means low competition and vice versa.
- σ_{man} - The variance in randomness of manpower given.
- σ_{eff} - The variance in efficiency.

Here, we fix some variables and test the values of others for best fit against the data we were provided. We fix the time step to be one day, therefore for our simulations $T = 365 \times 5 = 1825$. From our discussions with the contractors and Heathrow, we decided that the initial efficiency of the projects is slightly lower than optimum so we set $init\_eff = 0.8$. We decided that the randomness in manpower output would be hard to test therefore we set $\sigma_{man} = 0.1$, however we suggest that if this model gets taken forward this variable be tested.

This leaves β, c and σ_{eff}. Here, we want to test different values of these against the data provided. The challenge with this comes from the simplifying assumption 4, i.e., no mission creep. To take this into account, we sifted the datatthat was provided assuming that if a project overspent (the ratio of *Current Total of The Prices* against *Contract Price*) by more than a factor of five, this was due to mission creep. We recommend these calculations are made again when it is clearer which projects did experience mission creep.

After the sifting was completed we were left with 52 projects. We then calculated the average project over-run (the ratio of planned duration of project against original duration) and the average over-spend (the ratio of contract price against total current spending). We found that on average for these projects, the actual duration of a project is 1.3 times the predicted length and the final spend is usually 1.7 times the predicted. Next we ran our model with a predicted real schedule at different values of β, c and σ_{eff}, then took the readings of actual over predicted for time and spend (for each set of values we ran the model 20 times and took averages). The differences are displayed below in Figure 15.



(a) Average duration of projects.

(b) Average over-spend of projects.

Figure 15: Heat map of differences between actual and predicted values, see text for details.

In these figures, the darker patches are close to 0 (i.e., the simulation was very close to the real data) and the lighter the patch, the closer to 1 (i.e., the simulation was far from the real data). From this it is hard to tell which values of the variables are closest to the real data,

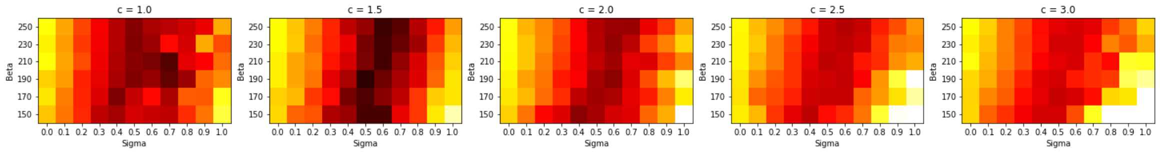


Figure 16: Heat map of combined differences.

however combining the two results we get the heat maps displayed in Figure 16.

The results have been combined using a Euclidean norm. They suggest that the parameter values that give the most realistic outcomes are

$$\beta = 190, \qquad c = 1.5 \qquad \sigma_{eff} = 0.5.$$

We will use these values in the following section to get data-driven results from the model. Given more data, it would be more meaningful to run a finer grid on these variables, potentially comparing project life cycle curves to real data rather than blunt averages.

6.2 Results

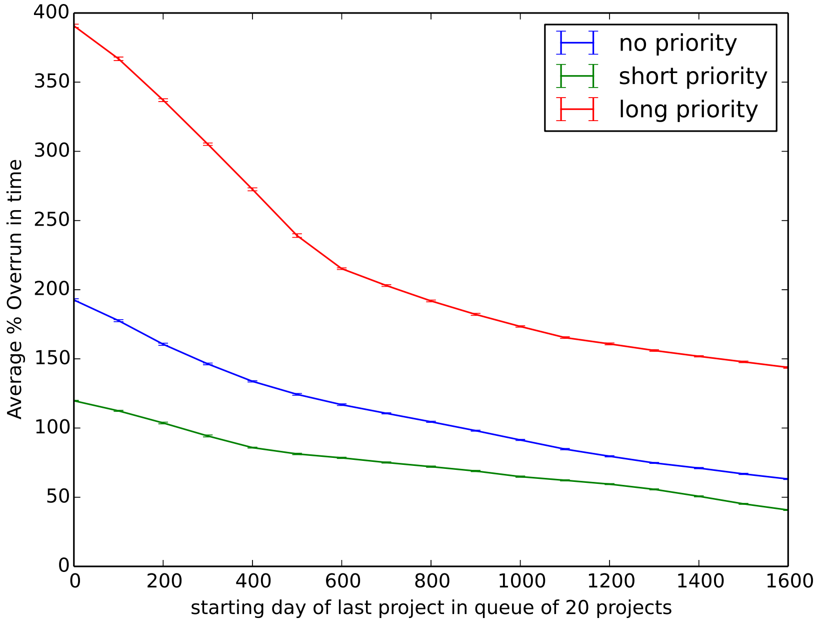


Figure 17: Comparing the impact of different strategies on the % increase of total cost.

We implement different scheduling and resource allocation strategies to gain a better understanding of the system. To demonstrate the importance of optimising the scheduling of

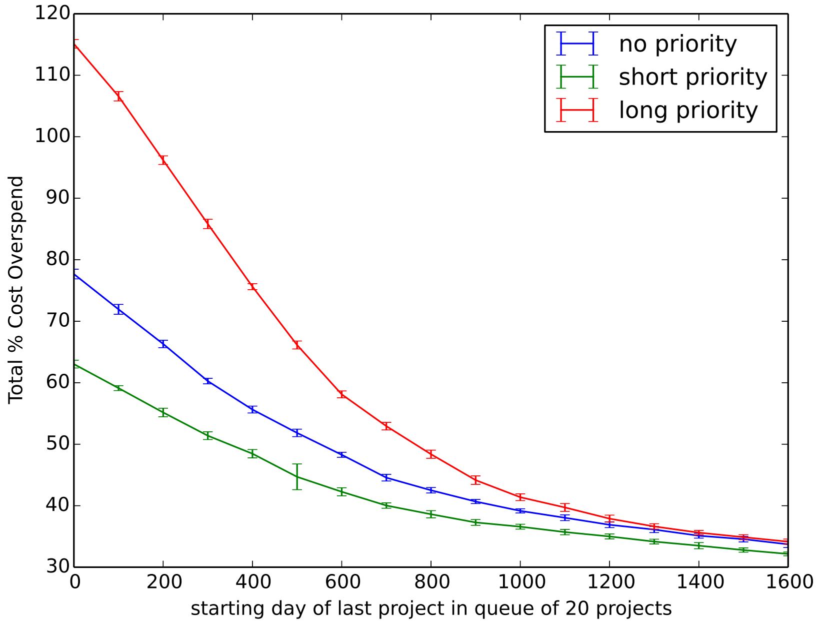


Figure 18: Comparing the impact of different strategies on the % increase of average project time delay.

projects, we feed a schedule of 20 projects distributed evenly across T days into our model, and increase T from 0 to 1600 days to examine the effect of spreading out projects on cost and time delay. The model parameters are fixed to $\beta = 190$, $c = 1.5$ and $\sigma_{eff} = 0.5$. We also investigate different resource allocation strategies. In the default 'no priority' strategy, each project gets a share of the total given manpower that is proportional to the requested manpower; in the 'long priority' strategy, the share is proportional to the requested manpower *multiplied* by the expected duration of the project, thus giving longer projects priority in obtaining resources; and in the 'short priority' strategy the share is proportional to the requested manpower *divided* by the expected duration of the project, giving shorter projects more manpower to accelerate their completion.

We measure the success of our strategies with two metrics. The first one is the ratio of total requested manpower vs the planned workload of the schedule of project, which we take to be a proxy for the cost of the project. The second is the percentage time delay averaged across all projects in the schedule, another quantity that is desirable to minimise. We see in Figures 17 and 18 that spreading out the starting dates of successive projects does lead to reductions in cost and delay. We also find that prioritising the allocation of manpower to long projects increases both cost and delay, whereas devoting more resources to accelerate the completion of short term projects tends to improve on the default strategy that is agnostic to the duration of projects. We remark that the difference between such strategies is less pronounced if the starting dates of projects are further apart. From these observations we deduce that attempts to optimise the schedule of projects may provide significant reductions in cost and delays, and that the separation of starting dates of successive projects and the allocation of resources in response to the duration of projects are significant factors that ought to be considered in optimising project schedules.

7 Future Steps and Recommendations

The sections above describe the work carried out in the week of the Study Group. This work consisted of proof-of-concept mathematical modelling and machine learning, and was limited due to the lack of relevant and easily available data. Therefore, the first recommendation make is to record (or gather from the appropriate sources) more data at the project level.

Recommendation 1. For each project, it would be useful to have:

- More information on the project type, from a construction point of view and its complexity, urgency, flexibility etc.
- Details of all contractors and subcontractors working on the project, in terms of subcontractor type, number of labourers, the weekly or monthly spend throughout the project and any other relevant information.
- Information on delays, scope changes, cost updates etc. Any change to the project should be categorised appropriately.
- If possible, information on how subcontractors price their work, including whether there are penalties for ordering work within a short time frame, or if there is a lot of demand for their services.
- Details of internal Heathrow services and how they are allocated and managed.

Recommendation 2. More generally, it would be helpful to have more qualitative information so that the assumptions are as reasonable as possible. For instance, how work gets allocated and prioritised; how competition for subcontractors is managed; why delays and increases in budgets occur; how capacity of a subcontractor is measured. It would also be useful to have some way to measure progress of a project, separately to the spend or duration. If this was available, it would be easier to understand if projects were delayed due to scope creep or due to a fundamental issue with the project.

Recommendation 3. If this information were collected and collated into an easily accessible database, it could form the basis of a machine learning tool that could be used to make predictions on the likelihood of a project over-running or increasing in cost. For example, it could allow one to predict that a project of type A, using a subcontractor of type B, at a certain time of the year is likely to over-run by 30% with a confidence of 70%. This would allow for much more accurate planning and scheduling.

Recommendation 4. These machine-learning outcomes could also be used to design a more realistic mathematical model, so that the true system could be modelled, analysed and ultimately, optimised. Different scenarios could in principle be simulated in real time and used to inform strategy.

Recommendation 5. Even without these extra data and understandings, it would be useful to see if the tentative conclusions from the mathematical model could be validated by experts. In particular, the model suggests that spreading start dates of successive projects at the same Tier 1 contractor leads to reduction in cost and delay. Also, prioritisation of resources to complete shorter/smaller projects on schedule leads to reduction in cost and delay compared to a strategy that prioritises bigger/longer projects.

Recommendation 6. The problem considered here appears to be novel, due to the nature of unexpected delays and distribution of project start times. Yet, greater effort needs to put into comparing it to other constrained resource optimisation problems in the literature.

Recommendation 7. The approach considered here that combines machine learning with a discrete-time simulation algorithm is likely to be applicable to other clients where there is a continual flow of a large number of concurrent construction projects of variable size and duration.

References

[1] S. N. Afriat. Efficiency estimation of production functions. *International Economic Review*, 13(3):568–598, 1972.

[2] G. E. Battese. Frontier production functions and technical efficiency: a survey of empirical applications in agricultural economics. *Agricultural Economics*, 7:185–208, 1992.