



ORbit

medlemsblad for Dansk Selskab
for Operationsanalyse og Svenska
Operationsanalys- föreningen

Redaktion:

Ansv. João Fonseca

DORS

DTU Management, bygn. 358
Danmark Tekniske Universitet
DK-2800 Kgs. Lyngby

E-mail: orbit@dorsnet.dk

Tryk:

Ecograf Gruppen

Oplag: 360

ISSN: 1601-8893

Forside:

Billede leveret af Jacob Buchhave,
under Unsplash License.

Bagside:

Billede leveret af Zachary
Shakked, under Unsplash License.

Editor

Hi everyone! Hope you had a great summer break and are back in good shape for the year ahead! It's been long since the last time I wrote you and an ORbit issue came out. If you recall the e-mail communication sent out some months ago, we in the DORS board decided to change the cadence of printed ORbit issues to once a year (or whenever there are enough articles to make sense for a printed version). The reality is that article submissions have been low in the recent past, so only now we felt like we had enough to send it to the printing plants.



In this issue we have a three excellent articles to share with you! Have you ever wondered how OR and AI can complement each other or be used in parallel? Fear not: Michael Lindahl is here to fill your curiosity and also talk about his personal use case with FindGaven. We then give the word to the 2023 DORS Prize winner, Ella Pirttikangas, to show us how she solved assignment problems considering operational uncertainty in maintenance during her master studies at DTU, which she also complemented with an excellent presentation at the General Assembly. Finally, Matthias Als is here to write about how Operations Research is being used at ECCO to improve their decision making tools. We add on top a couple of event reports and news from the community, to keep you up to speed with what is happening in the community, including the fantastic Euro 2024 held in Copenhagen!

I hope that you enjoy this issue and fingers crossed that it does not take too long until I am able to write another one of these introductions. As always, feel free to reach out if you have an idea for an article for ORbit, I'd love to discuss it with you. Until then, happy reading!

Best, João Fonseca (Editor)

Aktuelt om DORS

Medlemsskab		Sekretariat	
Kontingentsatser		DORS	
Personlige medlemmer		DTU Management	
(incl. ph.d.-studerende):		Akademivej, byg. 358	
Studerende:		2800 Kgs. Lyngby	
Firmamedlemmer:		e-mail: secretary@dorsnet.dk	
Institutmedlemmer:		Internet: www.dorsnet.dk	
Indbetales på Giro 9123865 (reg.nr. 1199)			

Svenska Operationsanalysföreningen

Autumn is here again, and most of us have returned to our daily routines of work and study. It was a fine summer that gave me many good memories, but I am also happy to once more go to the office and meet colleagues around the coffee machine. At the same time, we live in a time of war and political unrest. During the EURO conference in Leeds, there was a discussion about the political involvement of EURO (the Association of European Operational Research Societies), and this discussion continues. EURO has previously taken political stances, for example by suspending Russia's membership after the invasion of Ukraine.



I feel that the winds of change are once again blowing over us. AI and large language models have, in a short time, had an impact on research, industry, and education. These tools do not provide complete solutions, but they can greatly help us. For us in operations research, they can, for example, make it significantly easier to implement optimization methods and to quickly test new ideas. If you have not tried it, I recommend asking ChatGPT to generate code to set up, solve, and illustrate a Traveling Salesperson Problem (TSP). Of course, we can do this manually, but none of us writes the code in five seconds. We still need to understand the code, check that it is correct, and often make small changes. But the tools can save us a great deal of time. This opens the door both to more advanced projects in education and to more efficient ways of working.

In recent years, several international conferences in operations research have been organized in Sweden. CPAIOR was held in Uppsala in May 2024, and in September 2025 the Stockholm Global Optimization Workshop (STOGO) took place. STOGO was a successful workshop with 65 participants, mainly from Europe but also from North America and Asia; focusing on theory, methods, and applications of non-convex optimization.

In autumn 2026, SOAF plans to organize a Nordic operations research conference, where we hope to gather participants from our neighboring countries. More information about this will be shared later this autumn.

A new initiative within SOAF is a travel grant, and we plan to award two grants annually. Our series of online seminars also continues, and we aim to hold them more regularly. A big thank you to everyone who has contributed to this seminar series. We look forward to many inspiring talks during the autumn and winter. If anyone is interested in giving a seminar or has ideas for activities, please feel free to contact me at ordforande@soaf.se.

Jan Kronqvist, President of SOAF

Indhold

Redaktøren har ordet	2
SOAF	3
The newest tool in the OR practitioner's toolbox: Large Language Models	4
Case study: Assignment problem under operational uncertainty maintenance	10
EURO 2024 in Copenhagen, Denmark - An incredible experience	23
Helle Welling 1932-2024	25
Empowering reliable automated decision making in ECCO with DecisionOps	26
AOO 2024, Copenhagen, Denmark	30

Aktuellt om SOAF

Individuellt medlemskap: avgiften är 160 kr/år

Företagsmedlemskap:

Avgiften för företaget/institutionen beror på antalet

OR-intressenter enligt följande tariff:

Akademiska institutioner: 1600 kr/år; 1-5 intressenter: 1600 kr/år;

6-10 intressenter: 3200 kr/år; över 10 intressenter: 4800 kr/år

I företagsmedlemsskapet ingår 2-6 exemplar av ORbit, beroende på företags storlek.

Betala in på postgiro: 19 94 48-2 (Svenska

Operationsanalysföreningen)

e-mail: ordforande@soaf.se

Internet: www.soaf.se

The newest tool in the OR practitioner's toolbox: Large Language Models

by Michael Lindahl

Since the release of ChatGPT in November 2023, Large Language Models (LLMs) have seen unprecedented adoption by users. These models can work with unstructured, incomplete data and output in easily understandable text, but the downside is that they cannot reason or apply logic.

This is very opposite to Mathematical Optimization Models, which require structured and complete data but, on the other hand, can provide optimal solutions to complex and large decision problems. A challenge when using these models to build real-world applications to help an organization make better decisions is often that getting structured and complete data can be very difficult. This leads to projects never making it past the proof-of-concept stage or a decision application is built, but doesn't get the expected adoption because it is cumbersome for the users to input the required data or interpret the output.

This article explores the potential role of LLMs in Operations Research and how they can lead to faster development cycles and more intuitive decision applications. It shows how LLMs fit into the different parts of such development processes and applications. This is illustrated with examples and a case study of the application, www.findgaven.ai, which combines LLMs and simple optimization.

Operations Research and mathematical optimization are providing tons of value across many different industries. However, many organizations struggle to implement and utilize it in practice, and many projects fail in implementation due to

- Data is incomplete and of too low quality to be used.
- Users don't trust and understand the optimized solutions and therefore hesitate to put them into action.
- Users may lack the necessary skills and access to interact with and adjust the underlying model if new requirements arise that weren't considered in the development phase, rendering the solutions useless.

These issues aren't caused by the lack of efficient optimization algorithms or good modeling techniques, but happen when a real-life business problem needs to be translated into an optimization problem, and when the solution needs to be put into action.

With the recent advances in large language models (LLMs), we have a new tool to help tackle many of the problems that are preventing the adoption of OR, and I believe that the use of LLMs will lead to

- Faster development cycles to build prototypes and optimization-powered decision systems.
- Increased trust in the solutions generated by optimization models
- Higher-quality solutions, due to end-users being more empowered to control the data and parameters going into the optimization model.

Large language models

Large Language Models (LLMs) have made many significant advances over the last years, both increasing the quality of text produced and making them more steerable for users to produce the intended output. Adoption skyrocketed after ChatGPT was released in the fall of 2022, reaching 100 million users only two months after its release.

LLMs have been shown to be able to summarize texts, translate between languages, and write code. However, they also tend to make up incorrect information and fail at basic reasoning.

The capabilities around the models have also greatly increased e.g. you can.

- Get structured output to use the data in another application
- Have it use external tools to answer questions such as searching the web or writing code and executing it.
- Do a semantic search through a large amount of data to find relevant pieces of information.
- Read images and other unstructured information that greatly expand where you can get data from

In the Gartner hype cycle, generative AI such as ChatGPT is at the peak of inflated expectations. This

raises the challenge of figuring out where it works and where it doesn't.

Besides the enormous work and resources poured into building bigger and better language models, a large ecosystem of tools is also being built around these models that help integrate them with other software and make them usable in a production environment.

While autonomous agents that can navigate an organization by themselves, identify relevant data, and suggest optimal actions are still far out in the future, more tangible and narrow use cases are starting to become feasible and reliable.

The synergy between LLM's and Optimization models

It should by now already be clear that the two methods are very different and serve very different purposes.

In the book *Thinking Fast and Slow*, Daniel Kahneman discusses how humans use different methods to come up with an answer to a problem: the fast and intuitive and the slow and deliberate.

Solving a complex planning problem with a huge solution space and interconnected decisions is definitely slow. Humans have been outperformed by optimization algorithms for decades. However, many tasks that are still critical to succeed with an optimization project are more intuitive, and the practitioner needs to rely more on their intuition, years of experience, and context to connect the business and the model.

At a high level, the two methods can be distinguished where large language models offer high flexibility in input but low reliability in output, while mathematical optimization offers low flexibility in input but very high reliability in output, as outlined in Table 1.

Having these two tools in the toolbox will allow us to handle a bigger set of problems required when solving a real business problem.

The OR Process is ripe for LLM's

All problems start in the real world, where an organization has a problem with limited resources and wishes to use its resources more efficiently, aligned with the organization's strategic goals.

This leads to a new project which starts with the gathering of information and data. This gets turned into a scoped problem formulation with clear goals and what should be included and what shouldn't. This is then turned into an unambiguously mathematical model that can be used to generate solutions to the problem. These solutions need to be taken back into the real world to first be verified and evaluated. This will then lead to updates to the problem formulation, and then through the cycle to generate new and improved

solutions. When the solutions are verified, they are turned into actual actions to capture the value. This is outlined in Figure 1

One of the things LLMs have proven quite capable of is translating between languages and domains. You can quickly translate a text from English to Danish, from Danish to Pirate speak, and finally summarize it or turn it into an image in the style of your favorite artist.

The tasks in the OR process can similarly be seen as moving a problem between different domains. Let's dive into how LLMs can support each of these steps

Getting to a structured problem formulation

To create a well-structured problem formulation, we must scope what decisions to include, gather hard and soft constraints, and formulate the objective. To validate our understanding, we often analyze current solutions to ensure we capture what a solution needs to take into account. We also need to structure relevant data, often related to the resources we plan and their demand.

The data and information will often be unstructured and spread over multiple sources. Therefore, interviews, manual data entry, cleaning, and detective work will often be required to get all relevant information.

LLMs can support this process to make it much quicker. Here are a few examples

- Databases that only exist in .pdf files can be moved into a structured format that is readable.
- Unstructured data, such as a note field on a customer stating that "this customer should be visited before 11 a.m.", can be identified using an LLM and turned into a structured format that can be used for modeling.
- In many rostering problems, you will need to ensure that the solutions don't violate any union agreements or working time regulations. An LLM can scan through large documents and identify potential rules to add.

From problem to mathematical model

Mathematical modeling is often considered an art, and the way to become really good is to have seen many different kinds of problems. There are many base models for standard problems, and often the same side constraints appear in the same kind of problems.

Software developers already have tools such as Github Copilot to support them in converting a description of a piece of functionality into code. Being able to quickly add new constraints or get suggestions for different ways of modeling something will definitely speed up

	Input	Output
Large Language Models	✓ High flexibility Can read unstructured data in multiple formats such as text, tables, charts, and images. It works with incomplete data and will try to guess any missing data based on the context around it and its training data.	✗ Low reliability Non-factual. No clear quality metric. Doesn't necessarily obey the constraints it has been given.
Mathematical Optimization Models	✗ Low flexibility Requires both structured and complete data.Requires clear definitions of what everything is.	✓ High reliability The quality of the output can be measured and directly aligned with business objectives.Interpretable results and relationships.Optimality guarantees can be given.

Table 1: There are very different capabilities of Large Language Models and Mathematical Optimization Models

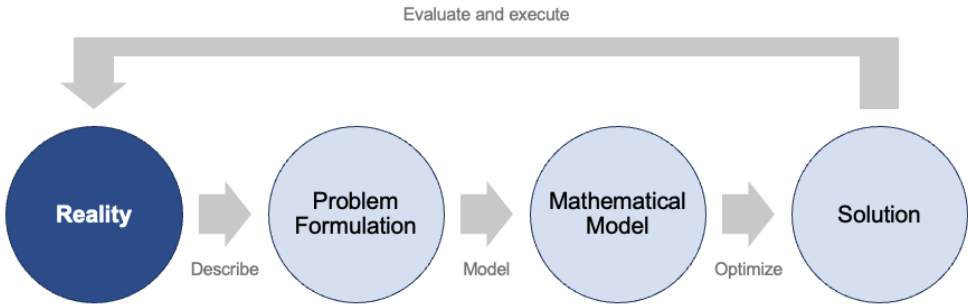


Figure 1: The OR process is about moving a problem from the real world to the model, and generating solutions that can be acted on in the real world

the development flow, also for optimization applications.

Having access to help in writing mathematical formulations, will also open up the field to more people. A few ways this can positively impact.

- Create a first version of a model based on an existing model from the literature linked to the available data.
- Adding a new side constraint to a model you haven't worked on for 2 years because a new business requirement has come in.
- Allow the end-user to add new constraints and tweak the objective using natural language.

Evaluating and turning solutions into actions

When we have generated solutions based on our model, we need to put them back into the real world. This is both to validate that we captured all relevant aspects

of the problems and, even more importantly, to turn the solutions into real actions that lead to value for the organization.

Building trust with the decision-makers is an essential aspect of turning solutions into actions. Some systems run entirely autonomously, but often, human decision-makers are in the loop and are both responsible and accountable for the final decision and the implications.

Therefore, it is important not only to output a solution but also to provide context around it so the decision-maker understands that it provides the right trade-offs and also understands any negative consequences of the solution.

LLMs can help translate the output of a mathematical model back to the users. Here are some cases this can be supported:

- Users often need a quick summary of a solution. What does it cost? Are there any broken rules the users should be aware of? An LLM can explain a solution in natural language before the users start to dive into the specifics.

- In early iterations, a problem often becomes mathematically infeasible, and no solution obeys all the stated rules. An LLM can explain why the problem is infeasible, for example, using the Irreducible Inconsistent Subsystem (IIS) and potential ways to fix it in natural language that the end users understand.
- Users often need to compare different solutions. These solutions could be baselines, optimized solutions, or solutions optimized toward different objectives. It can be valuable to give the user a high-level comparison of the pros and cons of each solution, which could be generated in an LLM based on the model's output.

Putting it all together

Speeding up the development of new proofs-of-concept and models is valuable. Still, it is undoubtedly that putting the user in the front seat to work directly with the underlying model will unlock a vast potential to analyze different what-if scenarios and increase the impact in the organization.

Having a mathematical model is an extremely powerful tool for different strategic decisions. Still, it can in practice be hard to use as it requires a highly specialized skillset to use and understand it. A decision-support system of the future might enable end-users to perform complicated what-if analyses without the need of an optimization specialist, where a workflow could look like this:

1. A company is doing rostering. They believe a new shift structure will allow them to utilize their staff better and want to understand what the cost implications will be and how it will impact employee satisfaction based on pre-defined metrics agreed with the union.
2. Users ask to see what the solutions would look like with a changed shift structure.
3. The program is able to identify what changes to data and constraints that should be added to the model
4. The new solution is generated, and the program creates a report that compares the new shift structure to the current one, ready to present to the management.

With LLMs, we have the required components to build such a solution.

Case study: Finding the “optimal” gift for your loved ones

To illustrate how this can be put into practice, this part shows an example of how a real-world LLM-powered

application works. The study is based on the website www.findgaven.ai. The site helps users find the perfect gift for their loved ones based on a simple text description of the person and searches through thousands of products across more than 16 different web shops. The site then presents the best gifts together with an explanation of why they suit the recipient. The algorithmic and mathematical complexity of the problem is simple, but the aim is to show how the LLMs are used to move between domains as described to work with large amounts of unstructured data and create a simple user experience. This is illustrated in Figure 2.

1. **From unstructured data to problem formulation.** Since the products come from different sources in different data formats, an LLM is first used to capture the important features based on the unstructured product description. The output is then a structured description with features such as whether the product suits a specific gender or requires particular interests and also detects if a minimum age is required, for example, if the product involves alcohol.
2. **From problem to mathematical model.** From the description, the features are converted into a vector with 512 elements using an LLM-based embedding model. The user query is then converted into a similar vector. This means that everything is now represented in the same vector space, and we can calculate the distance between the query and a product. Since the embedding model is trained to match pieces of text, the shorter the distance between a product and the query, the better the match the gift is.
3. **Finding the “optimal” solution.** The mathematical problem is now to select a subset of products to display to the user. This means selecting the products with the highest similarity while obeying side constraints to ensure we display a diverse set of results (e.g., not five pairs of pants in different colors).
4. **Evaluating and turning solutions into actions.** We now have a list of gifts we want to present to the user. However, if they were just presented to the user, it wouldn't necessarily be clear why they match, as e.g. presenting a distance score, wouldn't be very helpful. To help this, a LLM takes the user query and the description of the gift and writes up an explanation of why it is a good gift, and what the user should be aware of before choosing it. This builds more confidence in the results and allows the user to evaluate them quickly and make their decision.

Without the power of LLMs, the above would require extensive manual data labeling and a more compli-

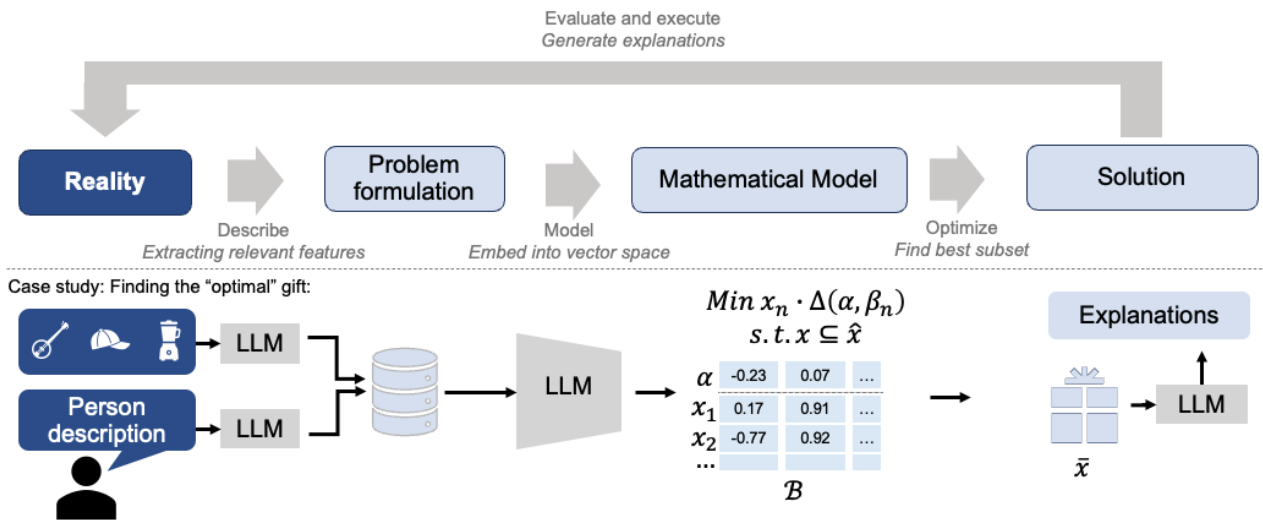


Figure 2: A case study of an AI application that uses LLMs to move from unstructured data to structured data and add additional context to the output solution.

cated user interface for inputting the person's relevant attributes.

have already decreased by almost 80% in the last year, and as competition is getting bigger it is likely prices will fall even more.

Current risks and future prospects

As mentioned earlier, the technology is still very immature and rapidly developing. There are still limited theories for knowing which problems LLMs can reliably solve and limited best practices for doing so in a way that would live up to the standard of modern system development that requires testability, scalability, and maintainability.

I expect to see a few things in the near future that will help with this.

More best practices and tooling. Besides the limitations of the models, the availability of tooling and best practices is still very limited. Even if we don't see the same step changes in LLM capabilities as previously it is safe to assume that the tooling around them will become much better and at the same time we will learn more about when to use them and when not to.

Keeping sensitive data to yourself. Many decision applications require business-critical data you don't want to share with 3rd parties. We currently see that Open Source models are slowly catching up to state-of-the-art chat models. So it is likely that we will be able to run chat models on-premise, to ensure data stays within the organization.

Lower costs. The variety of different models you can choose is increasing and each has different capabilities and prices. Some of the best LLMs

Get started with LLM's

Operations Research practitioners have all the right skills to take advantage of this new technology. We are used to working with data and business processes.

We are also used to working with stochastic algorithms, where we must have good benchmarks and conduct experiments to evaluate them.

I think it is beneficial for most practitioners to get an understanding of the basic capabilities and limitations to be able to identify when it can be used. Besides playing around with a chat tool like ChatGPT, it is also worth looking into the APIs to understand what underlying functionalities are available to you. There are plenty of places to start, but a few good places are:

- OpenAI API: <https://platform.openai.com/docs/overview>
- Langchain: <https://www.langchain.com/>
- Llamaindex: <https://www.llamaindex.ai/>

To be successful with a first use case, I suggest to keep the following things in mind

Scope a narrow use case. It is easy to overestimate the capabilities of LLMs today and think of elaborate use cases where the user chats back and forth with a Chatbot that does multiple things. Find a specific step in the process where you have a clear input and output you are looking for, and think about how you can formulate it as a prompt for a chatbot.

Create an evaluation set. What is the ideal response you want for a specific input? This allows you to create a benchmark set to validate if your algorithm is actually solving the problem and allows you to measure how reliable it is.

Keep a human in the loop. Many decisions are too critical and complex for a decision system to execute them directly. The same is true for LLMs. Ensure that it is easy for a human to validate the output to prevent polluting your final results with inaccuracies caused by the LLM.

I am excited to see how LLM's and Optimization will play together in the applications for the future!

Michael Lindahl has worked with applying Operations Research and Analytics for 12 years across various industries, such as airports, container shipping, and offshore wind farms. He has a PhD from the Technical University of Denmark. Michael was also president of DORS from 2015-2018.



Case Study: Assignment Problem under Operational Uncertainty in Maintenance

by Ella Pirttikangas

Decision-making in Maintenance Management contains different decision levels, where former levels give guidelines for the following decision levels. These former-level decisions are commonly made under uncertain outcomes at the following level. Therefore, there is a high motivation to integrate the uncertainty in the former decision-making levels.

This study focuses on the decision-making problem of an industrial maintenance provider, where the tactical decision of assigning equipment to technicians is made under the uncertainty of the operational circumstances. The uncertainties are unplanned maintenance activities that occur into the plan. These tasks additionally induce uncertainty in the maintenance tours, as this unplanned task disrupts the tour. The aim is to study how to account for this uncertainty when creating the assignments so that the negative consequences of such events are minimised.

The problem is formalised through a Stochastic Program, where two alternatives for the recourse are created. The formalised problem is solved using a metaheuristic, which is designed for solving large-scale problem instances of the case company. The algorithm uses a combination of Local Search and Large Neighbourhood Search to leverage principles of diversification and intensification. The neighbourhood functions use a graph formulation to quantify the qualitative empirical knowledge of the structure of optimal assignments. The uncertainty is modelled through simulation, where, depending on the recourse function, either just Monte-Carlo simulation is used or a combination of Monte-Carlo simulation and Simulation Optimisation.

The objective is to minimise the imbalance in workload and overtime. For the case study instances, the designed metaheuristic is able to improve the objective by an average of 90 %. The two recourse functions are studied, and it concluded that Simulation Optimisation, which modelled the recourse decision, creates more reliable results. Solving this model created more balanced workloads in terms of different workload types, had fewer negative effects on the utilisation of the technicians and generally created smaller service districts compared to the other recourse function.

Introduction

“This is the essence of intuitive heuristics: when faced with a difficult question, we often answer an easier one instead, usually without noticing the substitution.”

— Daniel Kahneman, *Thinking, Fast and Slow*

Generally, the intuitive heuristic of the human mind is successful in creating good-quality solutions to questions. Still, as addressed in Kahneman’s book 2012, when a problem becomes complex enough, we tend to simplify the problem to answer the question.

In maintenance management, the decisions to be made usually concern how to organise operations and how to utilise resources optimally. These decisions affect the employees in the maintenance fleet, and accountability for the grounds of the decisions is essential in good management practice. As maintenance organisations grow, managing these large entities becomes harder, and the ability to take in information is limited. Therefore, it is motivated to use computational tools to support these management processes. For a decision-maker to be accountable for machine-assisted decisions, they must trust the output of the algorithm and explainability has a key role in building that trust.

The maintenance decision studied here is a tactical decision, where the base of the decision lies on the operational circumstances. These circumstances are partly known when the decision is made, but some parameters are unknown beforehand. The scale of the problems can be considered to be large, which is why intuition-based decision-making is not feasible. The problem is studied as a case study of an industrial maintenance organisation which provides maintenance services for infrastructure equipment. Due to confidentiality, the concepts are described in general terms from a general maintenance management perspective. The problem described is applicable to any maintenance organisation providing service over a geographical area. Similar problems are found, e.g. in home healthcare, telecommunications, commercial refrigeration or real-estate maintenance.

Problem Statement

Optimisation and decision-making problems usually assume all data is deterministic and known. This, however, is rarely the case, especially in making strategic

and tactical decisions. The problem to be studied within the framework of the case study is how to create good equipment assignments to technicians so that the impact of disruptions from equipment breakdown is minimised and that the workload of the entire fleet is balanced, i.e. the workload is equally distributed among the technicians.

The case company is an infrastructure equipment manufacturer and supplier, and as part of their service portfolio, they offer maintenance. The maintenance offering covers, for example, inspections, scheduled maintenance and repairs, but also emergency tasks. The units lie geographically spread over an area, and some equipment can be located at the same location. The technicians in the maintenance fleet are assigned to equipment based on their capabilities, and some equipment types require certification from the technician for conducting maintenance.

The technicians' tasks contain planned maintenance, which is known beforehand and unplanned tasks that dynamically occur to the plan. The duration of these unplanned tasks is not known beforehand. The travel time between the equipment is counted into the total workload. The travel time is assumed to be deterministic, so the impact of congestion or other factors affecting the travel time are disregarded. The technicians' maintenance routes start and end in a given depot which can differ between technicians.

The aim is to assign equipment based on technicians' capabilities and capacities in terms of workload. Additionally, the assignment must enable a relatively equal workload for each technician, and the workload should be dimensioned so that the assigned technician primarily answers the ad hoc tasks. Other key performance indicators that drive the assignment problem are the overtime performed by the technicians and the utilisation of the technicians. The utilisation is the ratio in which the technicians work on equipment versus the overall work time.

It has also been found that there is an underlying value in allowing technicians to build customer relationships. Therefore, it is not appropriate to manipulate the assignment continuously. Instead, the assignments should be reconsidered if there is a change in the operational circumstances of the technicians. For example, when new pieces of equipment are introduced or removed or organisational changes require reassignments. This means that the assignment decision must be made before the operational execution of the tasks. The assignment decision gives the foundation on which the operational decisions are made.

Motivation

This thesis aims to study a large-scale tactical decision-making problem given operational uncertainty. The problem involves decisions which, both directly and

indirectly, through the assignments, affect the everyday operational circumstances of the technicians in the maintenance fleet. Therefore, making a well-founded assignment decision will have both organisational and individual value. The organisational value comes from increased productivity when the technicians use less time on non-value-creating activities, like travel time between units. Value on the employee level is created by having well-balanced assignments for the technicians, where also underlying workload due to unplanned events is accounted for. A well-balanced workload can reduce the risk of overworking and underworking the technicians. On top of individual value, a balanced workload can positively affect employee satisfaction, which naturally brings organisational value. In addition, geographically dense assignments will hopefully reduce travel time, which implicitly can decrease the environmental impact of the maintenance operations through the reduction of emissions.

The currently deployed procedures for making these tactical decisions have suffered from organisational knowledge loss, thereby decreasing trust in the procedure's output. This, on the other hand, has led to intuition-based optimisation of management decisions. The risk of this kind of decision-making is decision mechanisms, like bias and noise, which can lead the decisions to drift away from an optimum. The study will formalise the decision-making problem and facilitate principles of explainability and simplicity to solve large-scale problems with uncertain components. The aim is to test alternative procedures, principles and performance metrics and see how the decisions are affected by these.

A substantial part of the problem is preparing for the uncertain operational circumstances. So this study aims to formalise the impact of uncertain parameters on the decision-making problem and how historical data could be used to approximate the impact and integrate the impact into the decision-making procedure.

Data analysis and exploration

When creating a maintenance policy, the structure of desired services and cost structures are valuable inputs. This structure can be discovered through maintenance-related data. A brief data analysis is in place to understand the significant features and whether data supports the set motivations for the problem at hand. This section introduces the gathered data and gives overview on the available data and its limitations.

The data extracted from the ERP system (Enterprise Resource Planning system) cover a defined group with around 15 000 pieces of equipment which have scheduled maintenance for the upcoming planning period. A total of around 200 technicians belong to this maintenance fleet. These technicians are organised into approximately twenty subgroups. The aim is not to

optimise these subgroups, as it would require organisational changes of a larger scale. Thus, the study will use these subgroups as different problem instances. The network of equipment of each subgroup is unique, and the density varies depending on the geographical area covered by the subgroup. Meaning for some subgroups travel time is a bigger contribution to the total workload than for others.

The frequency and the duration of the unplanned events in the global data follow a typical distribution in maintenance, so by nature, failures are usually rare because maintenance aims to reduce the number of such events, so for both frequency and duration, the distribution is either exponential or Poisson distributed. It is to be noted that the distribution of unplanned events is not necessarily the same in the subgroups or in the assignment of the technician. So, one technician can have a subset of more unplanned events compared to their colleagues.

Figure 1 illustrates an approximation of the technicians workload. This figure approximates the work on site, excluding the time allocated for travelling between equipment. The lower part of the bar chart is the time allocated to planned maintenance actions for the upcoming planning period. This is, in other words, the total time of planned maintenance for the equipment currently allocated to each technician. The bright blue line across the lower part represents the average of the planned maintenance.

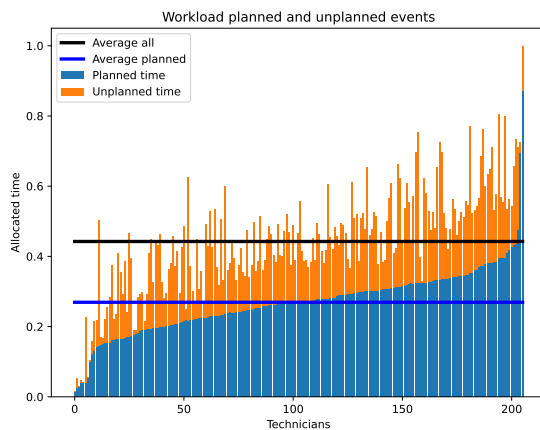


Figure 1: Planned and unplanned workload allocated to technicians, sorted by planned workload.

Suppose only the lower section bar is examined. In that case, one can see a variation among the technicians in the planned maintenance, and the workload is not equally distributed among the technicians. Of course, the workload of a technician comprises other tasks, namely unplanned tasks. The assumption is that the operations are unaware of the exact amount of unplanned maintenance for each piece of equipment over

the planning period when the assignments of equipment are made. A very simple approximation is made based on historical data to give an intuition on the amount of unplanned workload. In Figure 1 the upper orange part of the bar chart represents this. This value is obtained by averaging the historical data of each piece of equipment over three consecutive planning periods. This is then added to the assigned technician's workload. The black line across represents the average of this total approximation. This clearly is a very simple approximation, and if the actual unplanned workload for the upcoming year would be modelled, then the approximation should, on top of historical data, take into account other equipment-specific features and predictions on Remaining Useful Life (RUL) for instance. But to give an intuition to the problem, this is considered sufficient.

The interesting finding from Figure 1 is that the ratio of unplanned and planned maintenance is not necessarily equal between technicians, suggesting that units with a high amount of ad hoc activity dominate some technicians' workload. The fraction that the unplanned activities take up, on average, is over 35 % of the value-adding activities, which verifies that the unplanned activity is a considerable part of the work at the asset. Close observation of the figure shows that there are technicians (e.g. technicians 12 and 52) who lie in terms of planned maintenance below the average (the bright blue line). However, when the approximation of the unplanned maintenance is added, their workload shoot over the approximated total average (the black line). This supports the motivation of trying to integrate the effect of unplanned events into the assignment problem in other ways than simply as constant ratios, since the equipment might require different amounts of ad hoc activities. When combined with the fact that the geographical structure of different organisational subgroups requires different amounts of travel time, the actual workload profile might change substantially. So capturing and modelling the actual workload of the technicians for the assignment decision is of value.

Model Formulation

The described problem will be formalised through a mathematical formulation in this section. The modelling principle selected is a Linear Mixed Integer Stochastic Program with recourse. Stochastic Programming is selected because it is based on optimising over the expected value of possible scenarios, whereas e.g. Robust Optimisation optimises over the worst-case scenario. Since the assignment decision is a tactical decision, which is a guideline for the operational decisions later on, it is not motivated to have a high degree of risk aversion but rather to prepare for the likely outcome.

The model is built as a two-stage stochastic program with two decision stages. The first decision is the assignment of a technician to equipment. Given the assignment in the first stage and the outcome of uncertainty, the second stage decision is the routes the technicians take. A technician's total workload is assumed to be the sum of planned and unplanned visits and travel time. This model disregards workload due to internal activity like internal meetings, training or other activities.

The section is divided into two parts: the first part introduces the first stage, and the second part introduces two possible approaches to the second stage and only the verbal description is given; for exact mathematical formulation, refer to [1]. The formulation principles follow the theory in Shapiro et al. 2003 [2].

First Stage Equipment Assignment Problem

The first stage decision is assigning equipment to technicians by minimizing the total workload of the technician, the difference in technicians' total workload and the total overtime of the technician over a planning period.

The objective function minimizes the technicians' workload; given the equipment assigned to the technician, the objective is built with two terms. The first term is the planned maintenance over the planning period. Note that planned maintenance duration is not dependent on the technician; thus, this term is constant and not dependent on the assignment to the technician. The second term is the expected value of the recourse function. The recourse function describes the workload for unplanned events and the total travel time, which depends on the given assignment and the outcome of the uncertainty. This function will be described using two different formulations in the upcoming sections.

The first stage constraints ensure that a technician can only be assigned to a piece of equipment if they have the appropriate skill set for it. Lastly, each piece of equipment is to be assigned to one technician.

Second Stage using Sum of Distances

The simpler and more straightforward way to model the travel time is to assume that a tour contains only one visit to the equipment before returning to the start depot. Thus, the technician makes a series of tours for each visit they must make. This problem has no decision variable, so the technician cannot make any correcting decision after the assignment is fixed. This is because a valid route is assumed to be a visit to the equipment, thereby not allowing the technician to decide on the route. The random vector contains the total duration of unplanned events for each piece of equipment and the number of unplanned visits to each

piece of equipment. Note the first-stage assignment variable becomes a constant parameter in the second stage.

The recourse function contains a total of five terms. The first term accounts for the travel time due to planned visits over the planning period. The second and third terms contain the vectors from the random vector. The second term determines the total duration of unplanned events, and the third term describes the approximation of total travel time over the planning period for unplanned events. The stochastic parameter of the number of visits to an equipment is multiplied by the distance in time from the technician's depot to the equipment. The fourth and fifth terms are slack and auxiliary variables to measure the impact of the first-stage decisions. The slack variables are the overtime of each technician and the measure of the difference in workload between the technician with the highest workload versus the technician with the lowest workload. This is to measure the imbalance of workload between technicians.

The constraints in the given formulation are mainly for defining the auxiliary and slack variables and by observing the structure of the problem and by doing simple substitutions, the problem would only become dependent of the first stage decision, which is not dependent on the outcome of uncertainty. So, this problem de facto will not be a second-stage problem, but the total problem becomes a Linear Stochastic Program. This is proven by rewriting the function, removing the auxiliary variables and rewriting the slack variable with the constraints; for proof, please refer to [1].

Second Stage using Routing

The assumption previously presented is an overestimation of the technicians' actual routes since the technician likely visits multiple pieces of equipment before returning to the depot. Alternatively, the problem can be described as a two-stage stochastic program with recourse. In this case, the recourse decision is the technician's route, given the assignment made in the first stage. Therefore, the recourse problem is modelled as a symmetric multiple Travelling Salesman (mTSP) problem where pieces of equipment can be visited multiple times.

The alternative recourse function contains, similarly to the earlier formulation, the workload of the unplanned visits and the travel time in the planning period. The difference here is that the travel time or the technicians' routes are discretised in time windows. In other words, given the assignment decision in the first stage, the travel time becomes the summed duration of the tours the technicians make in each time window over the planning period. The tours contain planned maintenance visits and unplanned visits in each time window.

The random matrix contains, as previously, the total duration of the unplanned visits over the planning period and the number of unplanned tasks. However, instead of having an aggregated number of events over the entire planning period, the number is discretised over each time window. Thus, the random vector depends on the equipment and discretised time, and the unplanned visits become a two-dimensional matrix under the uncertain outcome. This matrix of unplanned tasks, together with the matrix of planned visits, creates a set of visits for each piece of equipment to be made in the time window. Each set's cardinality depends on the outcome uncertainty of the random vector. Additionally, there are visitation sets for each technician's depot, which are deterministic, so the technician must return to the depot a given number of times. The evaluation is done for each technician, and the assignment is treated as a fixed parameter from the first stage assignment problem. The sub-tour elimination in TSP is modelled using the Miller, Tucker, Zemlin formulation [3].

This formulation requires the introduction of new decision-making variables, one set of variables being integer variables representing how often a technician traverses an edge meaning visits equipment j after equipment i . A binary decision variable determining whether visit n is executed after visit m . And lastly, an auxiliary variable for determining the visitation order is described in the Miller, Tucker, Zemlin formulation [3].

The recourse function objective is to minimise the distance travelled in each time window and the constraints have both the workload-related constraint and constraints associated to an mTSP. For further details, see [1].

In this model, the same substitutions cannot be made as in the sum of distances version, as the auxiliary variable for the total workload is not only dependent on the assignment but also on the second stage variable of the number of times equipment j is visited after equipment i in the entire planning period.

Solution Approach

The model presented previously is a Linear Mixed Integer Stochastic Programming Problem. The variables are mostly integer or binary, which implies that solving the problem using exact methods might become challenging. Literature suggested that the optimal solutions using exact methods are often limited to instance sizes below a hundred nodes for Stochastic TSPs or VRPs [4].

Briefly mentioned in the data analysis, one finding showed that there is a lack of observations per equipment and for some equipment, observations are missing; this hinders the creation of reliable distributions on equipment level. Therefore, some more general assump-

tions of the behaviour of uncertainty must be made. Hence, describing the uncertainty through rigid scenarios, as usually done in Stochastic Programs, will limit the implementation opportunities.

For these reasons, the selected solution approach is a heuristic approach, where the aim is to find improvements to the status quo and obtain feasible solutions for all instances, which is desirable from a business perspective. For the maintenance business, obtaining near-optimal solutions for all instances is preferred rather than optimal solutions for a subset of the instances and using the initial solution for the rest.

The section is structured so that first, an overview of the algorithm is given, and the main building blocks are introduced. Then, the unique contribution of this case study is presented, meaning the neighbourhood operators and the uncertainty modelling, for more details on exact implementation, refer to [1].

Overview

The overall strategy of the heuristic search algorithm is built up by two leading principles. The first principle is the combination of diversification and intensification. Intensification is incorporated by using domain- and problem-specific knowledge of what an optimal assignment could look like. The diversification is implemented by leaving the domain knowledge and adding stochastic elements to exploring solutions.

The second principle is familiar from decomposition algorithms, which divide the problem into a master problem and smaller subproblems. In this implementation, the first-stage decision could be described as the master problem, and the second stage is the subproblem. A bird's-eye view of the procedure is given in the flow chart in Figure 2.

The algorithm is built up of four larger entities. Three of these blocks, Initialisation, Local search and Large Neighbourhood Search (LNS), depend on the Evaluation block, and this block contains all processes related to evaluating the quality of the solution. The evaluation also includes the evaluation of uncertainty and recourse. The given solution is evaluated through either direct computation or Simulation Optimisation, depending on which recourse function is used.

The given problem instance is first initialised to give the initial solution, which is evaluated in the Evaluation block. The current assignment is used as the optimisation starting point, so the solution is not constructed. The initial solution is then sent to Local Search for a local optimum. The search principles in Local Search are based on problem-specific knowledge of ideal assignments to find the best possible solutions. Once the Local Search has terminated and the termination criteria are unsatisfied, the search continues in LNS, where a larger neighbourhood is explored. This

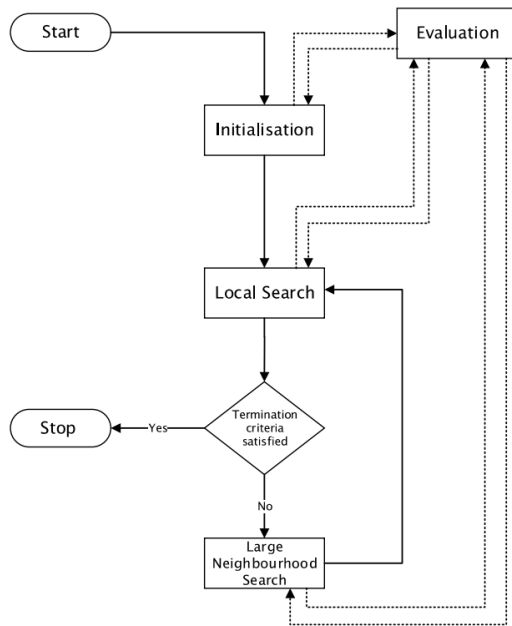


Figure 2: Overview of the algorithm.

is done by changing the search strategy to do large assignment changes without big impact on travel time and by adding some stochastic elements to the search. After a given number of iterations, the procedure then interrupts the LNS and returns the solution to Local Search, where the local search space is explored. This alternation is continued until the termination criteria are satisfied, which here is time. The LNS uses Simulated Annealing acceptance criteria, where the temperature and decay parameters are set using automatic parameter setting.

The procedure is implemented in Julia programming language and takes advantage of various calculation packages to create distributions and calculate linear algebra.

Neighbour Sets and Neighbourhood Operators

Given the first stage assignment problem, the only limiting constraint for the assignment is that the technician has the competencies needed for maintaining the given equipment, so in principle, any equipment can belong to an assignment, regardless of far geographical distances. If one would consider this problem as a graph formulation where if two equipment share an edge, the equipment can belong to the same assignment, and it is then assumed that a technician can maintain all equipment. This would produce a complete graph, meaning all vertices are connected by an edge. This becomes understandably a large solution space. For example, a graph with 100 nodes (equipment) would

produce a graph with almost 5000 edges.

Empirically, it is known that not all equipment can belong to the same assignment, despite the technician's competencies, and the reason simply is travel time. So, geographically far away equipment will not likely be in the same assignment because this would increase the travel time associated with the operations. So to reduce the search space, this empirical knowledge is leveraged using an adaption of the formulation given in Haugland et al. 2007 [5].

Adjacency graphs or matrices

Haugland et al. 2007 study and describe a delivery districting problem and describe the geographical area using what they call an adjacency graph; this graph describes how well-connected the nodes are in the area, so if equipment share an edge, they are ultimately well-connected. The study provides a pseudo-code for generating such adjacency graphs based on intersections [5]. As part of the case study, an adaption of this pseudo-code was implemented and compared with Delaunay triangulation used in automatic mesh generation in FEM (Finite Element Method) analysis [6]. The finding was that the Delaunay is more suitable as it generates fewer edges and more reliable results. Figure 3 shows one such example of Delaunay Triangulation on a case study problem instance.

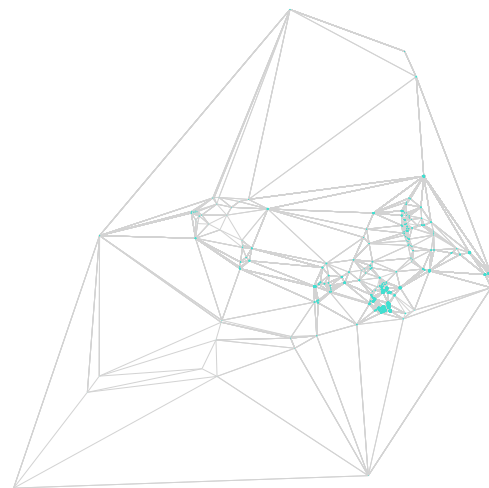


Figure 3: Example of Delaunay triangulation on one problem instance.

Based on the adjacency graph, neighbourhood operators are created for both search methods, LNS and Local Search, through so-called neighbour sets.

Neighbour sets

The idea behind this set is to compare how well an equipment e is connected to an assignment a , note the cardinality of assignments is equal to the cardinality of technicians. The set is described as a matrix of size equipment e and assignment a . The value in the matrix is dependent on the edges in the adjacency graph and the current assignment, so the value is the sum of the edges one equipment shares with equipment in a given assignment. This means the larger the value is, the more connected this equipment is to that particular assignment. A value of zero between equipment and assignment would thus suggest that the equipment is geographically distant from the assignment and would implicitly suggest that adding this equipment to the assignment would increase the travel time.

Modelling uncertainty

The uncertainty is modelled through theoretical probability distributions, whose parameters derived from historical data. Due to the lack of observations, equipment-specific distributions cannot be made. Therefore, the general assumption is made that a technician's total workload due to unplanned events over the planning period is reflected by the group of equipment that the technician is assigned to.

The procedure for creating a distribution of the historical data takes as inputs the total data for the frequency and the duration and the assignment of the technician. The equipment-specific observations are gathered by assignment and subjected to distribution fitting.

The modelled distribution for a particular assignment is used in Monte Carlo sampling to create a Sample Average approximation of the expected value for the duration and frequency. If the assignment is changed, the probability distribution used in the Sample Average approximation will also change.

Evaluation of travel time

The model formulation presented two ways of modelling the recourse, first without a decision variable (Sum of Distances) and second where the route decision was made (Route Approximation).

The evaluation through the sum of the distance from the depot is evaluated simply using the approximation of the expected value through Sample Average Approximation. The extracted Sample Average Approximation for the frequency reflected the approximation of the number of unplanned events over the planning period.

The other way to approximate the load through travelling in the network is by creating a set of possible visits over a time window and simulating the possible route the technician could take in that time window; each simulation becomes one possible scenario. Thus,

repeating the simulation a number of times will give the same number of possible scenarios for the routes over the planning period. By then, evaluating each simulation equally results in an approximation of the total workload due to travelling. This simulation consists of three parts: A sample scenario of unplanned events, aggregation of travel time over the period and travel time in a time window through the Nearest Neighbour heuristic.

The main routine repeats the approximation a given number of times and returns the Sampled Average of that route approximation. The simulation creates a sampled scenario of unplanned visits over the planning period. Then, based on that scenario, the corresponding travel time over the entire planning period is approximated.

The sampled scenario is created using the distribution of the number of unplanned events over the planning period; each piece of equipment in the assignment is assigned a sample from the distribution. Based on the Uniform distribution, the sample of unplanned events is scheduled over the planning period, hence creating the visitation matrix for that particular sample. The technician's total travel time over the planning period is calculated by running the mTSP on each time window and returning the sum of the route durations.

Tests and Results

The test instances are in total 23, of which one is a "toy" instance; the instances vary in size and network. The smallest real instance has 5 technicians and 100 pieces of equipment, whereas the largest has 14 technicians and almost 1000 pieces of equipment. The density of nodes in the instances varies from 1 piece of equipment per 100 areal units to almost 5000 pieces of equipment per 100 areal units. So the role of travel time is different depending on the instance. The algorithms have multiple parameters to be determined, but the most crucial being the number of samples needed for the Sample Average approximation. The reason why the number of samples is crucial is its impact on computation time, especially for the Route Approximation, where the construction heuristic is repeatedly run. The appropriate number of samples is estimated by simulating the accumulated mean described in Rosetti 2016 [7]. The simulation is repeated until the accumulated mean stabilises and reaches a steady state. The steady state is determined by standard error.

As the algorithms have stochastic elements, they are tested using different random seeds, and it can be concluded that the patterns in the search do not differ substantially with different random seeds.

Running tests on both evaluation methods shows that the route approximation using the mTSP can execute fewer iterations with the given computation time

compared to the Sum of Distances, which is expected as the Route Approximation requires repeatedly solving the Nearest Neighbour heuristic. However, it is observed that using the route approximation, the actual travel time is more realistic, whereas, in the Sum of Distances, the travel time overpowers the other workload types, which impacts how the workloads are balanced. So, the Route Approximation balances the workload between technicians in terms of planned and unplanned workload. Both evaluation methods can improve the objective of minimising the workload imbalance and overtime by around 90 % for both evaluation methods.

Comparing the recourse functions

The sum of distances metric was significantly larger compared to the other workload types, which led the optimisation to target mainly the travel time. However, it is able to run a significant number of iterations with only a tenth of the time given to the route approximation.

Even though delta evaluation is used in the Local Search to determine the quality of the moves, the computational load, due to the simulation optimisation, is significant compared to the sum of distances. This then naturally causes the LNS to execute fewer iterations. However, the impact on the average utilisation is less severe for the route approximation. The average over the instances is -0.35 %, where the same value for the sum of distances is -3.21 %. So both decrease the utilisation, but they both successfully balance the workload, and the change for the route approximation is marginal compared to the improvement in the objective. Additionally, the route approximation reaches a fairly good solution with fewer changes to the assignment, so the Hamming distance between the initial solution is smaller, which is positive in terms of customer relation management.

The two recourse functions evaluate travel time relatively differently. The sum of distance looks at the equipment's proximity to the tour's starting point. In contrast, the route approximation creates a series of possible routes, thereby using information on the proximity between the equipment. So, the sum of distances metric does not evaluate the equipment clusters. In other words, an assignment where the units lie in the same location would get the same value in the sum of distances as an assignment where the units are evenly spread over a circle with an equal distance to the starting point. Thus, the sum of distance does not account for the district or service area.

Therefore the solutions are evaluated in terms of the geographical area by looking at each assignment and measuring the service area by creating a polygon over the assignments. The sum of distance evaluation tends to create larger service areas. Figure 4 shows the average district area for each instance. Out of the 22

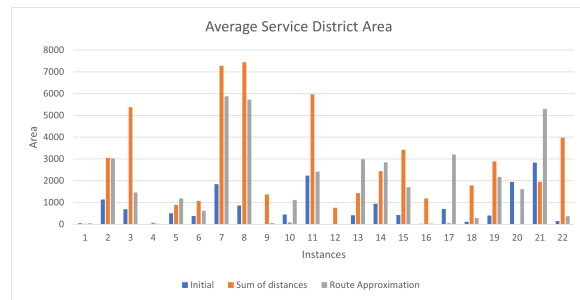


Figure 4: Average service district area.

case-study instances, the sum of distances created in 14 cases larger district areas. The initial solution has, in some cases, smaller district areas than both solutions; however, the workload is also less equally distributed, so the increase in the service area can be evaluated as the cost of having a more equally distributed workload.

Discussion

Methods and Limitations

The problem statement and the available data have largely affected the selection of the methods used to solve this problem. All of the decisions described in the problem are most naturally described as either binary or integer decisions, and considering the size of the instance, the most preferable method is a metaheuristic approach. Also, the primary problem is a tactical decision, which is a framework for the upcoming operational decision, and hence, using methods which integrate the likely outcome of uncertainty is preferable compared to risk-averting modelling methods — furthermore, the lack of equipment-specific data conditions the methods to describe the uncertainty in more general terms. Simulation optimisation is selected to enable more generalised assumptions of the historical data.

Uncertainty Modelling and Suggested Improvements

Probability distributions are used to describe the uncertainty of the frequency and duration of unplanned events. The distributions are created using historical data, but the data is partly incomplete, and the number of observations makes it unreliable to describe each piece of equipment by a distribution. Using the same total historical data to describe each unit would result in the same expected outcome for all equipment, which is known not to reflect the overall data. Therefore, the observations are aggregated over the assignment. So, the assumption is that the joint distribution of a particular equipment group describes the unplanned workload of that particular assignment.

Suppose a technician is assigned to a group of equipment which all have had one unplanned event in the three last planning periods; hence, it is also fair to assume that the technician will in the next period have one unplanned event at each equipment in the upcoming planning period. If the technician is assigned to a piece of equipment with significantly more unplanned events, say, twenty events in a planning period, the aggregated distribution for the equipment becomes skewed towards higher frequencies. The assumption here is that the aggregated distribution reflects a total load of unplanned events for that specific assignment of units over the planning period. So, when the unplanned workload is aggregated over the entire planning period, the sum of unplanned activity is approximately the same when using individual or aggregated distributions.

However, the problem which emerges with this assumption is associated with travel time. Let's say that the units with only one unplanned event over the period are very closely positioned, but the problematic equipment with twenty events per period is far away from the assignment cluster. Then using the total assignment's average outcome will underestimate the load of the technician travelling to that far-off location, as the aggregated distribution will give values lower than the actual value of the problematic equipment. This assumption is to be considered when analysing the results and their impact.

Additionally, the theoretical distribution depends on the subset of the historical data, and the best-fitting theoretical probability distribution is not necessarily the same after reassignments are done. Suppose the distributions fit poorly to the data. In that case, the uncertainty will be described poorly, which, in the worst case, could misguide the search and reduce the credibility of the proposed solution.

By describing uncertainty differently, the limitations described could be eliminated. Possible methods could include creating equipment-specific models to analyse common equipment features. The common features could be, for instance, the equipment type and model, the age of the equipment and when recent upgrades or rebuilds have been made. This would enable the modelling of possible unplanned events over the planning period and the creation of representative scenarios to be used in the evaluation part of the procedure. However, this requires a lot of research to describe the equipment in a reliable way, as each scenario implies a decision chain with different associated probabilities. Another possible way to integrate uncertainty is using a Robust Optimisation framework, where the observations would be used directly in a box uncertainty set, where the outcome can vary between the maximum and minimum values depending on the worst case. Using Robust Optimisation makes the model more risk-averting, but the problem considering the distributions is eliminated. Us-

ing Robust Optimisation will require some assumptions concerning the data as it is partly incomplete.

Algorithm Design Choices

In the recourse function based on route approximation, the mTSP is modelled through a construction heuristic. The advantage of generating a solution through a construction heuristic is quick computation, which is preferable in a simulation setting. Also, since the target problem to be solved is not the actual route, it is not necessary to search for any improving solutions but rather simulate the decision process.

It is identified that mTSP formulation had a limitation, constraining consecutive visits to the same equipment but allowing the solution to return directly after it has visited another equipment. This return to the equipment was hindered in the algorithm by a tabu list which allowed a revisit only after a certain amount of other visits. There are, however, cases where the routes become unrealistic as the tabu list is simply insufficient. Such cases appear especially if many units are in the same location, so all visits in the location are executed consecutively. This could be averted by increasing the number of time windows in the planning period so that there is only one visit per equipment in the time window. This, however, increases the computational load, and additionally, the received data does not support it, as visits are currently aggregated to multiple visits in a time window. So, implementing more time windows would require additional assumptions about the data. On the other hand, the trade-off between more exact modelling and generating a solution for the tactical problem should be kept in mind. What is the cost of an exact model and the benefit for the tactical problem to be solved?

The search algorithm uses problem-specific traits to find good neighbouring solutions. Describing the geographical area through these adjacency graphs enabled both the Local Search and the Large Neighbourhood Search to use the regional connections between the assignments through the so-called Neighbour Sets, which provided good neighbours. The Local Search uses a candidate list for testing reassigned equipment. The list is created through different criteria. The first criterion is to only have the equipment in the list with adjacent equipment in the target assignment, ensuring that the equipment will be regionally connected after reassignment. Then among those units, the first units to be evaluated are units which are geographically disconnected from their current assignment. Testing those first will hopefully result in more connected assignments. The second sorting criterion is the number of assignments in the equipment's neighbour set, where the equipment with the least assignments in the neighbour sets is tested first. The idea here is that equipment with many assignments in their neighbour sets will

likely be centrally located overall and, therefore, easy to access. The fewer neighbours the equipment has, the farther away from the central areas the equipment is, and the more reassigning that unit will affect the source technician's workload. This was proven in the tests to be a good result as the procedure found good solutions in the proximity of the initial solution for all of the instances.

The approach in the LNS is the opposite, where operators target centrally located equipment to provide alternative solutions. By doing so, it moves the solution to a completely another area in the solution space. At the same time, the travel time of the technicians is not changed drastically as all of the moved units lie centrally in the border regions. Combining the two approaches gave good results as the approach uses proven principles of diversification and intensification.

Using Simulated Annealing acceptance criteria supports the principles of diversification and intensification. Implementing an automatic parameter setting method is beneficial because it gives intuition as to what is a good value for the parameters, especially in cases where iterations are executed in a reasonable time, which is the case in the sum of distances recourse. Even though the automatic parameter setting is not run for the route-approximation metric, using the parameter setting for the other metric gives intuition for the parameters for route approximation.

Adjacency Graph and Limitations

The strength of the neighbourhood operators in both Local Search and Large Neighbourhood Search depends on the preprocessed Adjacency Graph. The graph describes the geographical area and how the units are located. Two units can only belong to the same assignment if they have a shared edge or share adjacent equipment. Therefore, it is important that the generated adjacency matrix is symmetrical and that all units have a number of edges. Additionally, the graph should represent the different densities of units in the geographical area. So centrally located equipment should have more edges compared to more rural equipment. This is because centrally located units are more easily accessible than rural equipment. The distances between the central equipment are smaller, so reassigning central equipment has less effect on the travel time in the assignments. Using the graph formulation converted abstract problem-specific traits to quantifiable measures, enabling the creation of good neighbourhoods for the search algorithms.

The algorithm used to create these graphs was inspired by the work of Haugland et al. in 2007, where the edges are filtered away based on whether they intersect each other or not [5]. The algorithm shown in Haugland et al. seemed not to be well suited for the large instances in this study, and the structure of the instances resulted

in unexpected cavities in the graphs, likely due to node density. The algorithm is modified to go through edges only once, which makes the algorithm more dependent on the order in which the edges are observed. So, the units observed first will have more edges than the last units. Therefore, additional filtering is made based on instance size, where the edges are compared once more. This approach is very case-sensitive since each instance's size and geographical structure differ. In some instances, the preprocessing creates sufficient results, whereas, in other instances, the approach leads to graphs that do not present the geographical area well.

Other methods for generating the graphs could be explored to reduce the case dependency of the Adjacency Graphs. One method could be leveraging the k -nearest neighbour method, where edges connect the equipment to the k -nearest units. The number of neighbours, k , is a user parameter, so this approach does not remove this limitation, and k -nearest would not take the different densities of the geographical area into account because each piece of equipment would have a given number of edges. Another possible method to consider the density is creating a set of Minimum Spanning Trees and then connecting the missing edges to create a graph.

Methods inspired by mesh generation for FEM analysis could provide good results. One automatic mesh generation algorithm is Delaunay Triangulation, which takes a set of points and creates non-overlapping triangles [6].

Managerial Insights

Accountability in management decisions was stressed in the introduction, and explainability was lifted as a means for increased trust and accountability in machine-assisted decisions. Therefore, one of the guiding principles of the study is to use methods which are grounded on problem-specific principles and logical intuition. At the same time, the methods leverage well-known principles and algorithms to solve large-scale optimisation problems, which produce, on average, 90 % improvements in the objective compared to the initial solution. It has also become clear that unplanned maintenance activity is a considerable part, on average, a third of the total workload and should be considered in the assignment problem.

Algorithmic design decisions facilitate explainability, for instance, through the modular modelling principles used, where the workload types are discretised to planned and unplanned actions and travel time. The modular structure enables the decision-maker to validate the procedure's outcome against other data and thereby gain trust in the suggested assignment. Additionally, the described variables are of the same units (time), which makes comparing the variables eas-

ier and reduces the need to correct using weighting factors, which usually reduces the explainability. If other variables of different units, e.g. utilisation of technicians or ratios between workload types, should be integrated, then some conversion factor, for instance, monetary measurements, should be integrated. Determining some monetary value of abstract concepts is not straightforward. Still, decision support tools like multi-criteria decision analysis can assist when dealing with qualitative abstract variables [8].

Modularity is also implemented in the algorithms, making the procedure easy to modify by exchanging blocks. So, if more well-suited methods are found, these could be easily plugged into the procedure. For example, the current distribution-based uncertainty modelling could be exchanged for more sophisticated methods using predictive maintenance models and prediction confidence intervals. Modifying the procedure requires fewer resources, and changes are easier to implement.

The assumption of the capability of the technicians assumes that the current plan represents the complete capabilities of the technicians; however, this might not always be the case. For example, when a technician has hidden capabilities and is not assigned to equipment of that type. Therefore, it is highly motivated to conduct a capability mapping in the organisation and use the actual data directly to leave the assumption made here. Due to the modular design of the procedure, exchanging the data is simply exchanging the input matrix.

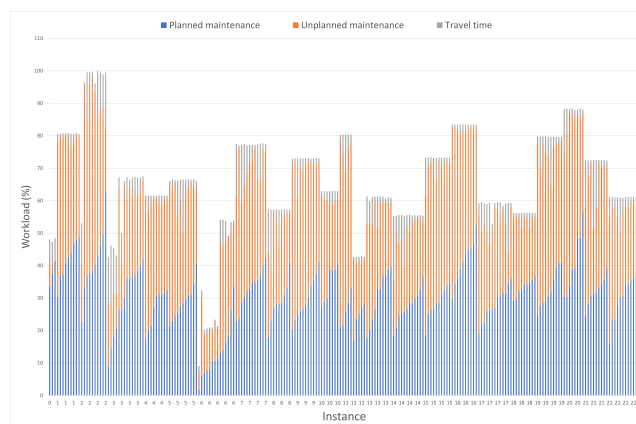


Figure 5: Route approximated workload for all technicians in all instances.

The procedure did not optimise over the entire geographical area, but the optimisation results for each subgroup can be analysed together to troubleshoot the entire geographical area. One such analysis is shown in Figure 5, where all technicians estimated workload is plotted. The bar plot creates additional visual segments for the technicians are grouped by subgroups. Figure 5 shows that the workload is not evenly distributed over the subgroups. If a certain target workload is defined,

the management can see which subgroups have too many or too few resources and can react based on the result. Say a target workload would be 80 %, then management can see there are six subgroups in need of more resources. Additionally, the benefit of approximating the uncertain workload is that the decision-maker can see if there is an imbalance in the workload types. If a technician has a high amount of travel time compared to the work on the equipment, then organisational changes in the subgroups could be considered so that the technician's travel time is reduced, for instance, by reassigning the technician to a new subgroup.

Also, if a technician has an alarmingly large number of unplanned maintenance hours in their assignment, the results of the optimisation could serve for troubleshooting and facilitate sales opportunity discovery. Suppose the ratio of unplanned maintenance is high. In that case, there is likely equipment with some chronic issues in the assignment causing a significant workload and offering some upgrade or rebuild could open new sales opportunities.

Additionally, the procedure can be used to test the sensitivity of upcoming planned maintenance plans because the procedure adds an estimation of the unplanned workload. If the solution contains high overtime, the planners are alerted that the proposed plan might be unrealistic. Thereby, either new resources can be allocated, or the plan can be re-calibrated.

Much of the presented literature divide units into so-called locations, which contain multiple units close to each other. For instance, Zhong et al. 2007 call these operational cells, and when generating the route, these are used instead of individual visits [9]. In their case, all visits are equal, and executing certain visits does not require specific capabilities. This is not the case in this study, where locations can have multiple equipment of different types. Aggregating the units to locations would decrease the solution space, as the entire location will have more requirements on capabilities, and furthermore, assigning technicians to locations reduces the possibility of sharing the workload. Suppose a location has ten units of three different equipment types, and the total workload of the location exceeds the allowed capacity. The technician must be capable of maintaining all these equipment types, and it must be accepted that they will be at overcapacity. By not creating these subsets by location and assigning them to technicians, the workload can be more easily shared and the equipment more flexibly assigned.

The different recourse functions were tested and compared. Even though the computation time of the Route Approximation is longer, the results showed that simulating the decision produced lower travel times, which enabled the algorithm to balance all workload types rather than just focusing on reducing the travel time as occurs using the Sum of distances recourse. Therefore, considering the overall workload, it is more

beneficial to try to incorporate more accurate travel time modelling as it balances the workload types better. Additionally, the solutions using the route approximation recourse produced in smaller district areas, which would indicate shorter travel times and could positively affect the overall utilisation.

Finally, the procedure's cost function only used travel time imbalance and overtime to evaluate the quality of the solutions. Still, other cost functions could be explored to see how the solutions change—for instance, minimising travel time, maximising utilisation, etc.

Conclusion

The assignment problem of equipment to technicians is studied as a tactical decision-making problem under operational uncertainties. The assignment quality is evaluated on the capacity, capability and balance in workload for the maintenance fleet. Because the workload depends on uncertain parameters, such as whether unplanned events occur and the duration of the events, the problem was modelled as a Stochastic Program, where two different recourse functions were introduced. Depending on the recourse function, the total problem is either described as a Linear Stochastic Program (LSP) or as a Stochastic Program with Recourse (SPR). The recourse action in the SPR is modelled as a multiple Travelling Salesman Problem.

The model is implemented and solved using a metaheuristic approach, combining local search (LS) and large neighbourhood search (LNS). The model was tested in a real company case on instances of varying size and geographical structure, all of the instances can be considered to be large optimisation instances. The results for the particular case study showed that the generated solutions minimised the objective of overtime and workload imbalance by an average of 90 % for both recourse functions. The combination of the two search methods provided good results as they facilitated principles of diversification and intensification, which have been shown in previous research to provide good solutions for Stochastic Combinatorial optimisation problems. The metaheuristic's search is based on the generated Adjacency Graphs, which describe the geographical area and the relation between the units. By generating these Adjacency Graphs, problem-specific knowledge can be used to reduce the solution space to a subset of potential solutions, which supports the empirical knowledge. So, in other words, the Adjacency Graph enables quantifying the qualitative intuition of a good assignment and thus creates high-quality solution candidates.

The simpler recourse function, which was described as the LSP, used a Sample Average Approximation for the unplanned visits and summed the distances

between the assigned equipment and the technician's depot. This was identified as an overestimation, and in the metaheuristic, the approximation overpowered the other workload types, leading to the optimization focusing on travel time. Additionally, the LSP tended to create larger service districts for the technicians compared to the SPR. The SPR model used Simulation Optimisation to model the recourse of the assignment decision by generating numerous scenarios and simulating the nearest neighbour heuristic to simulate the technician's tour. The computational load due to the simulation is large as the number of samples has to be significant to stabilise the Sample Average Approximation. However, the model approximates the travel time to be smaller, which allows other workload types to be balanced, and the negative effects on technician utilisation are smaller than for the LSP model.

The utilisation of the technicians is identified to be an important indicator of productivity. When reassigning the equipment, there is an underlying trade-off between utilisation and a balanced workload due to the reassignments. So even though the provided solutions did not significantly increase the utilisation, they could minimise the workload imbalance without really affecting the utilisation. So, the imbalance is practically removed, while the utilisation is almost the same as initially. Implicitly, the same applies to the environmental impact of the operations, as the trade-off between workload types applies to travel time.

The modularity in the model enables the decision-maker to validate the proposed assignment against some other data, as the workload is discretised by type. Since the search algorithms start from the current assignment and use problem-specific knowledge, the explainability of the procedure increases. The variables are of the same unit when formulating the models and algorithms. Thereby, they can be directly compared, so the use of scaling and corrective parameters is minimised.

Future research efforts should focus on describing the uncertainty, generating the Adjacent Graph, and validating the suggested assignment to other data. Describing the uncertainty more reliably could be valuable, as some of the assumptions that needed to be made were not supported in the technicians' tours. This would give falsified values for travel time and could mislead the search. The remedy suggested for this is equipment-specific models. Using other algorithms to generate the result could remove the user parameters and case sensitivity of the generated Adjacency Graphs.

The described problem is a common problem in organisations which use subgrouping of units to support the operational planning of activities. The case study is an assignment problem of an infrastructure equipment manufacturer offering maintenance services over a geographic area. However, any organisation with the same workload structure could utilise the principles

presented in this report to optimise their assignment. Examples of such sectors could be the maintenance of commercial cooling equipment, real estate services or the assignment of nurses for a home care unit. The applicability of the problem statement to various sectors increases the motivation to model and solve such problems.

The report contributes to the research field by formalising the problem and solving a particular case study using relevant methods and principles to generate solutions for large problem instances. The formulations and designed solver take two sources of uncertainty into account: frequency and duration of unplanned events. Based on the literature research, the complexity of similar problems usually limits the number of uncertain parameters to a maximum of one or two parameters. Additionally, studies on VRP problems with stochastic visits are sparse compared to other forms of stochastic VRPs. The report also discusses the industrial domain of the case study, and based on the literature review, there are only a few studies made of stochastic VRP or TSP problems in a maintenance management setting. The work summarised in this report will hopefully contribute to the specific research field.

Finally, the selected methods succeeded in balancing the workload for the subgroups and the selected methods facilitated principles of explainability, where the decision-maker can validate and analyse the suggested solutions. These approaches will hopefully increase trust and traceability in automated decision-making procedures. These procedures will hopefully be a valuable addition to the decision-makers' toolkit as they can simulate the mechanisms and uncertainties in the operations and thereby provide new insights. These automated decision-making algorithms should be considered as decision-support tools to enable the decision-maker to manage larger entities and use quantified data-based procedures to support intuition-based decisions. Thereby creating more balanced maintenance fleets, which still create value for the organisation and the customers.

References

- [1] Ella Amanda Pirttikangas. Case study: Assignment problem under operational uncertainty in maintenance, 2023.
- [2] Andrzej Ruszczyński and Alexander Shapiro. *Handbooks in Operations Research and Management Science, Volume 10, Stochastic Programming*.
- [3] C. E. Miller, A. W. Tucker, and R. A. Zemlin. Integer programming formulation of traveling salesman problems. *J. ACM*, 7(4):326–329, Oct. 1960.
- [4] Jorge Oyola, Halvard Arntzen, and David L. Woodruff. The stochastic vehicle routing problem, a literature review, part i: models. *EURO Journal on Transportation and Logistics, Volume 6, Issue 4*.
- [5] Dag Haugland, Sin C. Ho, and Gilbert Laporte. Designing delivery districts for the vehicle routing problem with stochastic demands. *European Journal of Operational Research, Volume 180, Issue 3*.
- [6] Simena Dinas and Hector J. Martínez. *Delaunay Triangulation*, pages 1–6. Springer International Publishing, Cham, 2020.
- [7] Manuel D. Rossetti. *Simulation modeling and Arena*. Wiley, 2016.
- [8] Benoît Le Maux and Jean-Michel Josselin. Multi-criteria decision analysis. *Statistical Tools for Program Evaluation*, pages 385–416, 2017.
- [9] Hongsheng Zhong, Randolph W. Hall, and Maged Dessouky. Territory planning and vehicle dispatching with driver learning. *Transportation Science*, 41(1):74–89, 2007.



Ella Pirttikangas is the recipient of the Prize 2023 for her Master's thesis conducted at the Technical University of Denmark in the Department of Technology, Management and Economics as part of her examination in M.Sc. in Business Analytics.

lytics.

Ella has a background in Process Engineering specifically in the industrial maintenance and service sector, which inspired her to learn more on technical methods to leverage industrial data and the decision-making processes in the specific sector. After graduation, Ella joined DecisionBrain as a Business Analyst, where she works with maintenance providers in creating decision support solutions and continues formalising complex maintenance problems for added operational and business value.

EURO 2024 in Copenhagen, Denmark - An incredible experience

by Marta Szachniuk and Dario Pacino

What a conference! With 3029 registered participants from 70 countries across the globe and 2458 presentations, *EURO 2024*, held in Copenhagen, Denmark, from 30th June to 3rd July, 2024, was the second largest *EURO-k* conference ever held.

As in tradition, the program began with an *opening ceremony* on 30th June, 2024, Sunday afternoon with an opening speech by Rasmus Larsen, Provost, Technical University of Denmark, followed by a welcome speech by EURO President, Anita Schöbel, followed by the presentation of the prestigious EURO awards. Prof. Lidija Zadnik Stirn received the *EURO Distinguished Service Award* and Prof. M. Grazia Speranza was awarded the *EURO Golden Medal* this year.

The conference's official opening was celebrated with a welcome party in which a marching band of children, *Tivoli Guards*, dressed in traditional royal military uniforms escorted the participants to the party venue.

EURO 2024 showcased a diverse program, ranging from classical operational research (OR) topics to modern integrations with AI technologies, which was evident in its invited talks. The conference featured three *plenary speakers* including industry speaker Berit Brouer (Maersk), IFORS Distinguished speaker Prof. Charles Corbett (UCLA), and Prof. Pascal Van Hentenryck (Georgia Tech). They delivered outstanding presentations, offering insights into OR's impact on life, its practical applications, and the integration of AI into OR.

Additionally, the conference included 12 *keynote talks* by Javier Alonso-Mora (University of Technology, Netherlands), Pedro Amorim (University of Porto, Portugal), Maria Besiou (Kühne Logistics University, Germany), Péter Biró (Corvinus University of Budapest, Hungary), Immanuel M. Bomze (University of Vienna, Austria), Emma Frejinger (Université de Montréal, Canada), Bahar Yetiş Kara (Bilkent University, Turkey), Miloš Kopa (Charles University of Prague, Czech Republic), Veronica Piccialli (Sapienza University, Italy), Graham Rand (Lancaster University, UK), Rubén Ruiz (Universitat Politècnica de València, Spain), and Pascale Zarate (Toulouse Capitole University, France). Their presentations encompassed all aspects of OR represented at the conference.

The program featured special events such as award sessions, including the *EURO Doctoral Dissertation Award* (EDDA), the *EURO Excellence in Practice Award* (EEPA), and the *EURO Prize for OR for the Common Good* (EPOCG). Additionally, it included EURO forums like *Making an Impact - the EURO Practitioners' Forum* and *The EURO WISDOM Forum - Women in Society Doing Operational Research and Management Science*. There was also a memorial session dedicated to the late Prof. Dr. Jakob Krarup.

In the conference dinner, which took place at the Langelinie Pavilion, situated in the heart of Copenhagen, the participants were entertained by a female choir performing traditional Danish songs. After dinner, the participants hit the dance floor to music played by the EURO unofficial DJ, Prof. Bernard Fortz (a member of the EURO council). Kudos to Bernard for his outstanding performance!

Given the size of the event, a second gala dinner venue was also set up at Kayak Bar in the Copenhagen harbor area. Prof. David Pisinger from the organizing committee, entertained the crowd present here by making a grand entrance at the venue by sea by arriving on his own kayak!

The conference ended with a “bang”, a spectacular *farewell party* in which the *Copenhagen Drummers* delivered a mesmerizing percussion performance using drums, trash cans, ladders, water, and fire. Participants enjoyed local beer and the lively atmosphere that invariably accompanies the conclusion of a successful conference.



Figure 1: EURO 2024 (top-down, left right): Tivoli Gard marching towards the welcome party; EURO President Anita Schöbel Ceremonies: left: EURO Distinguished Service Award, and right: EURO Golden Medal.



Figure 2: EURO 2024: Impressions from the Gala Dinner.



Figure 3: EURO 2024: Impressions from the Farewell Party.

Helle Welling 1932-2024

by Graham Rand

The IFORS Statutes used to state: 'The duties of the Secretary are to act as administrative servant of the Federation'. Helle Welling, IFORS secretary for 21 years, who has died at the age of 92, was a marvellous servant to the International Federation of Operational Research Societies (IFORS). In her case, perhaps 'mother' would be a better word than 'servant', because she said that 'to function as the IFORS secretary is in many ways like running a big family. To be in charge of the organisation of a General Meeting, a President's dinner, and to cooperate with various organizing committees of a Triennial Conference is like organising a big party at home'. She went on to comment that 'to me it was like having pen friends all over the world. I sat in the middle trying to meet demands from the member societies and from the IFORS officers, answering questions, organising, planning meetings, and solving problems. And - as in a real family - when you give, you receive. I will never cease to marvel how much you receive in return for trying to be there for the member societies and the IFORS' officers'. For many operational researchers in the final quarter of the last century, Helle was simply Mrs. IFORS: she was the embodiment of IFORS, serving as IFORS Secretary from 1976 until 1997, and playing a valuable role at the Triennial Conference in Beijing in August 1999.

Helle had first become involved in IFORS in 1970, when working for the then IFORS President- Elect, Arne Jensen, Professor of Mathematical Statistics and OR at the Technical University of Denmark. He assumed that Helle would know what IFORS was all about, but eventually he realised that this was not the case, and suggested that she went to London to visit the IFORS' secretary, Margaret Kinnaird. Helle commented that 'after half a day spent with Margaret, things started falling into place'.

In 1972 Helle went with Arne Jensen to the IFORS Conference in Dublin. She recalled how she sat listening to the General Meeting of the Board of Representatives, with English being spoken in about 30 different accents. Helle asked Margaret how on earth she was able to minute all these statements/outbursts/questions. Margaret's answer was: 'I write what they think they said'. Helle said 'later on, when I myself became in charge of the minutes of the General Meeting, I always had Margaret's words in the back of my mind'.

In 1973 Arne Jensen, Roger Collcutt from the UK, then the IFORS Treasurer, Margaret and Helle went to Japan to look at conference facilities for the 1975 Conference. They were entertained in the most magnificent way. Apparently, Margaret and Helle became so absorbed in the Japanese way of entertaining that one night when they returned to the hotel, their heads filled with the day's discussions, they decided to have their own private tea ceremony. Helle recalled that 'there we were at 12 o'clock at night all dressed up in our version of the Japanese kimonos, trying to copy

the Japanese way of relaxing. I am not sure we met all the requirements for a Japanese tea ceremony, but I do recall that we put the work behind us and had a good night's sleep, although disturbed by a minor earthquake'. Helle much appreciated the enrichment of her life through her service to IFORS.

Helle played a key role in the conferences held in Toronto (1978), Hamburg (1981), Washington DC (1984), Buenos Aires (1987), Athens (1990), Lisbon (1993) and Vancouver (1996). She worked with IFORS' Presidents Takehiko (Bill) Matsuda (Japan, 1974-76), David Hertz (USA, 1977-79), Roger Collcutt (UK, 1980-82), Heiner Müller-Merbach (Germany, 1983-85), Jacques Lesourne (France, 1986-88), William P. Pierskalla (USA, 1989-91), Brian Haley (UK, 1992-94) and Peter Bell (Canada, 1995-97). With these Presidents and their Administrative Committees, she served OR Societies and operational researchers all over the world.

In 1976, the IFORS Bulletin, a non-scientific publication carrying information on IFORS and OR activities worldwide, was initiated by Heiner Müller-Merbach, who was then a Vice-President. The Bulletin was distributed to all member societies and to individuals interested in OR. Helle took over the editorship in 1979, a role she carried out for two decades, and developed it into a useful means of transmitting information on what was happening in the OR world, in particular to smaller societies. But its main role, in a pre-internet age, was to establish a link among the various member societies. In the Bulletin, IFORS member societies and individual OR people could publish anything concerning the OR world. Such information reached other OR Societies, and a great number of lone OR workers throughout the world. Helle became an IFORS Fellow in 2022. In her nomination documents, one former president said that 'she published a regular newsletter with all information related to IFORS, which was distributed to all Societies. In a time when all communications were by mail, she kept all our societies in touch'. *Andres Weintraub (President 1998-2000)*

Another said that 'of all of the great persons that I have been privileged to know in my years with and for IFORS I have known none with the dedication and capabilities she has demonstrated in care and work for our Societies'. *Bill Pierskalla (President 1989-91)*

Helle attended a gathering at the EURO conference, held in Copenhagen in July, and met some of those who hold her in a great deal of respect and affection, including IFORS President, Janny Leung, and three former presidents. With the news of her death a month later, we realise how fortunate we were to have that opportunity.

Graham Rand

Empowering reliable automated decision making in ECCO with DecisionOps

by Matthias Als

This article presents the concept of DecisionOps, a systematic approach for developing and maintaining Operations Research (OR) algorithm software, ensuring transparency and reliability throughout the development process. Utilizing their experiences from ECCO, a global leather goods retailer with a comprehensive supply chain, the author describes how ECCO has leveraged this approach to optimize their supply chain operations through the development of the Intelligent Auto Replenishment and Omni-Channel Fulfillment solutions. The article provides a detailed walk-through of the DecisionOps development cycle, from the initial research and development stages through to the productionalized solution, and finally, the crucial stage of continuous feedback. Looking ahead, the author shares ECCO's future plans for OR projects leveraging DecisionOps. By sharing ECCO's experiences and future plans, this article offers inspiration for businesses seeking to enhance their OR algorithm software development processes.

Introduction

Operations research (OR) has been since the middle of the 20th century, and still is, an active research area in academia. Recently, OR methods have received more attention in the private industry in the wake of the popularization of artificial intelligence (AI). The big communities of machine learning (ML) and generative AI, have for some time established good practices for development processes such as MLOps and LLMOps. Good practices for development processes are not as widely documented in OR. Therefore, in this article, we will introduce DecisionOps, and how it can help scale and systematize the development of OR algorithm software and ensure transparency and reliability throughout the development process.

Our description is based on our application and experience in ECCO. ECCO is a global retailer of leather shoes and leather goods with a headquarters in Bredebro, Denmark. ECCO owns its entire supply chain from raw cow hides to finished products sold to consumers, as shown in Figure 4. This gives the company unique opportunities to control and optimize the supply chain to increase overall profitability, ensure customer satisfaction, and reduce environmental impact using OR algorithms. DecisionOps has helped ECCO scale the development of fully automated OR algorithms both to a global retail network and to a larger development team while ensuring reliability at the same time.

The Current State of Operations Research in ECCO

The ECCO Data & AI department was kickstarted in 2016 to be the powerhouse of data engineering, data infrastructure, ML, AI, and OR. To this day, they have developed and fully productionalized two ML and OR-based solutions to drive automated decision making in ECCO. The global extend of the OR solutions is shown in Figure 5. Both solutions are centered around the use of state-of-the-art ML forecasting models to predict the expected customer demand for products in the supply chain. The ML models are helping to account for historical sales, seasonality, product type, shoe size, and even visual similarity.

Intelligent Auto Replenishment

First launched in Europe in 2019, Intelligent Auto Replenishment (IAR) is today used globally across 29 countries to optimize the inventory levels in the retail stores and distribution centers (DC) based on expected future customer demand, shipping lead times, current inventory levels, prices, and more. Optimal inventory levels are computed multiple times a week and are used to place product replenishment orders to be sent from regional DCs to retail stores, but also from upstream production DCs to regional DCs. IAR is finding the balance between not losing potential sales and costs related to transporting, handling, and storing products. At the same time, constraints such as location capacities, inventory availability, minimum order quantities, bundling logic, and more are satisfied. This inherently makes the IAR MIP model a large-scale optimization problem. Every month, IAR makes more than 15M decisions for ECCO's global supply chain.

Omni-Channel Fulfillment

The second productionalized solution was the Omni-Channel Fulfillment (OCF) algorithm, also introduced in 2019. OCF is used globally to determine pick-and-pack locations for e-commerce orders and products ordered by the customer in-store. Pick-and-pack locations can be both stores and DCs. OCF is based on a very similar mathematical concept as IAR, but decisions are made real-time within a few seconds after the order has been placed to ensure timely processing and delivery of orders. In addition to previously mentioned objectives and constraints, OCF also safeguards store personnel by limiting fulfillment orders per location. At the same time, it balances not stocking out the local stores and minimizing the shipping lead time for the customer. In total, OCF handles and fulfills all in-store and e-commerce orders in 28 countries around the world.



Figure 4: ECCO's supply chain

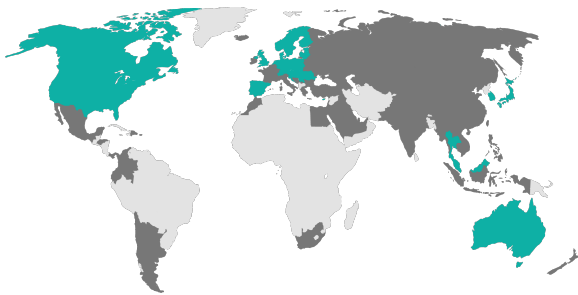


Figure 5: Grey = ECCO markets. Turquoise = IAR and/or OCF algorithms used.



Figure 6: Visual representation of the DecisionOps development cycle

DecisionOps in ECCO

When you have an automated algorithm software tied into business processes on a global level, such as IAR and OCF, it is very important that you have processes in place to ensure reliability and transparency throughout the software lifecycle. In this article, we call this DecisionOps and it is visualized in Figure 6. DecisionOps is closely related to DevOps and MLOps, but with a few considerations specific for algorithm software and Operations Research. In each of the sub-sections below, we have described the steps a piece of OR algorithm software is subject to in ECCO Data & AI. Each step serves as a guideline rather than a rigid rule, but each step is shown for completeness.

Research and Development

First and foremost, when you observe a business problem that can be solved by OR methods, you should estimate whether its impact on the business will outweigh the resources needed for development and maintenance. Then, all new development starts with a brainstorm and napkin math – can we actually solve the problem the way we imagined? Once a promising solution is found, both the math and algorithm are documented thoroughly. Like having a convention for good coding practices, you should also have a convention for math. The mathematical optimization model, just like code, should be clear, readable, and easy to understand by the team while being performant. The documentation is then reviewed by peers in the team, dependencies are mapped out, and development is planned. Finally, development commences in the team towards a minimal viable product that can be tested by the stakeholders or end-users.

Testing

Anything that is to be run in the production environment goes through various types of testing. Testing helps ensuring reliability and stability in the algorithm. Sometimes, even simple changes to the logic can have unexpected effects on other parts of the code. A comprehensive test suite can help identify such unexpected effects. The most common types of tests used in ECCO are outlined below from an OR perspective:

- **Unit testing** – Each component/function of the software and algorithm are tested in isolation. In terms of OR algorithms, each constraint and objective term are seen as separate components. Therefore, each constraint and objective term are tested separately on the bare minimum of the model.
- **Sensitivity testing** – The OR algorithm is run with different combinations of the parameters values relevant for the new development. This can give insight into whether the algorithm behaves as expected in both edge cases and regular cases.
- **Integration testing** – Multiple components/functions are tested together. Specifically for OR algorithms, constraints and objective terms that could potentially

interfere or counteract each other can be tested in combination.

- **End-to-end testing** – A full vertical slice of the algorithm software is tested. All components needed (such as the model itself, data integrations, supporting software, etc.) are deployed and run in the test environment, which is a replication of the production environment.

Continuous Integration

When new features of the algorithm are developed, they are immediately integrated into the existing code base, even if it is not used right away, to avoid stale code branches, increase code quality, and encourage iterative development. Before merging finished features into existing codebase, the new feature must adhere to the development guidelines, which in ECCO is all unit and integration tests must pass, as well as approval by the code linter, code formatter, and at least one peer reviewing the code. Additionally, end-to-end testing and sensitivity testing must be performed if applicable. This sounds like a lot of checks and a hindrance to rapid development, but large portions of the aforementioned checks and approvals can be fully automated in tools such as GitHub.

Deployment

When one or more new features have been merged into the existing code base, the code base is tagged with a version and ready to be deployed to the production environment. Using infrastructure-as-code (IaC) tools is highly useful for rapid deployment of versions and configurations, but also for easily duplicating the production environment to establish an identical but isolated test environment. IaC can even be used to manage your algorithm configuration, which eases collaboration in even large teams.

Operational Monitoring

Once the new features, or even new OR algorithms, are in production, it is important to set up and enable operational monitoring of the algorithm. Operational monitoring includes anything that is critical to assessing the health of the algorithm software running in production, especially if it is an automated process. Important metrics include:

- **The infrastructure supporting the algorithm** - Data read/write times, container start-up times, algorithm request rate, etc.
- **Algorithm performance** - Algorithm run time, memory consumption, MIP gap, etc.
- **Inputs to the algorithm** - Processed data volumes, data quality, data freshness, etc.
- **Outputs from the algorithm** - Distribution of decisions, outlier detection of outputs, sparsity, etc.
- **Run time exceptions** - Code exceptions, out-of-memory errors, model infeasibility, etc.

This monitoring should be processed and reviewed in real time by a monitoring service such as Grafana or Datadog to ensure fast recovery in the event of unexpected

incidents. Additionally, in-line logging/printing is useful for debugging. All information allowing for fast debugging and recovery that is not a metric can be logged there and collected by a logging aggregator platform such as Elasticsearch or Datadog. An example from one of ECCO's operational dashboards is shown Figure 7.

Strategic Monitoring

Now that we have ensured that the algorithm is running smoothly, we also need to assess its impact on the business and whether it aligns with the strategic vision. Strategic monitoring is often done as part of, or supporting, the business intelligence routines in the company. Before commencing development of a new algorithm software, you have probably established some KPIs to measure its success – these need to be reviewed regularly as well. Strategic metrics like sales and inventory numbers, marketing performance, or other business-specific KPIs are not important to measure daily, such as operational metrics, but rather monthly or quarterly to see if the strategy works and if the algorithm has the expected impact. Extra attention should be given to whether the OR algorithm is part of establishing entirely new processes, supporting existing processes. For new processes, you may need to come up with new KPIs as well to support the new process. For existing processes, measuring the estimated difference in costs associated with the process with/without the algorithm can be beneficial to continuously prove its *raison d'être*.

Continuous Feedback

Last but not least, it is crucial to get feedback directly from the stakeholders of your algorithm software to ensure continuous improvement and value creation. Not all performance can easily be measured with metrics and numbers. One example is usability. Recurrent touchpoint meetings have proven to work well in ECCO to allow stakeholders to have a forum to voice their concerns, provide feedback, and request new developments. To this end, an agile development framework can be beneficial in systematizing the way stakeholders and developers interact and in establishing a common understanding of expectations of the deliveries. Any feedback that results in development activities will loop back into the first subsection, 'Research and Development'.

The Future of OR in ECCO

There is no doubt; if your business does not utilize their data to make well-informed (or even automated) decisions, the competitors who do will take over. Therefore, ECCO will continue to leverage OR to enhance operational excellence. ECCO's suite of OR software products is also expected to increase, and the below three projects are currently being developed and will become productionalized within the foreseeable future by the same DecisionOps principles as IAR and OCF.

- **Markdown Optimization (MDO)** - Based on demand elasticity as a function of the product price, our MDO model is able to provide recommendations on end-of-season sale discounts which may vary over the

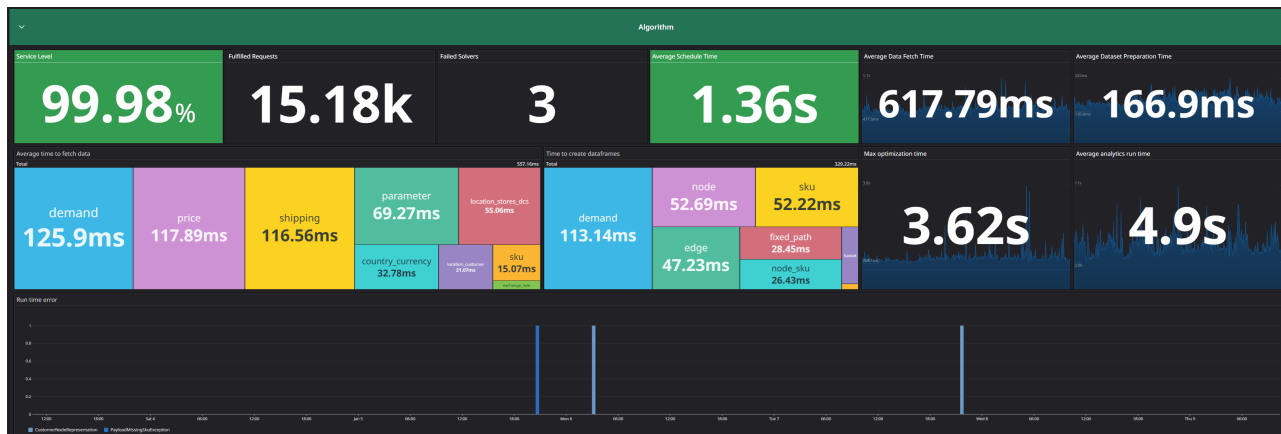


Figure 7: Part of the operational monitoring of the OCF algorithm

sales period. The MDO algorithm is currently being implemented in selected European countries after which further expansion will be assessed.

- **Return Optimization (RO)** - Like in any other fashion company, the return of e-commerce due to misfit is inevitable. These returns can be used as a supplement to replenishment to satisfy customer demand, because most returned products are in mint condition. Based on the same principles as OCF, the RO algorithm will in real-time pick the return location that is most likely to turn the returned product into profit.
- **Stock Redistribution (SR)** – Customer demand is non-deterministic, which requires retail stores to redistribute products among them during the season to rebalance their inventory levels. Redistributions are needed to close gaps in e.g. the shoe size ranges to ensure consistent availability across sizes. Based on similar principles as IAR, the SR algorithm will provide recommendations of how to redistribute inventory in the retail network. The algorithm is currently being tested in North America.

On top of that, both IAR and OCF will be subject to further expansion both in terms of covered markets, decision parameters, and the set of features.

Acknowledgements

I would like to thank my colleagues across all of ECCO, especially ECCO Data & AI. It is great to be part of an organization that fosters innovation and that allows us to build state-of-the-art operationally critical solutions like IAR and OCF. Special thanks to the Operations Research team in ECCO for supporting me and each other and, of course, for proofreading this article. Finally, thank you to DORS and its volunteers for building a community around Operations Research that enables us to get inspired and learn from each other.

Matthias Als is a Lead Data Scientist in ECCO. For more than 3 years in ECCO, Matthias has built automated decision systems especially for inventory management problems. Matthias holds a MSc degree in computer science from the IT University of Copenhagen. Since April 2024, Matthias is also on the board of DORS.



AOO 2024, Copenhagen, Denmark

by Alexandru Serbanescu



The annual workshop organized by the Danish Operation Research Society, Applications of Optimization 2024, held at the Industriens Hus, on May 6th, 2024, enjoyed this year a special edition with exciting presentations from 6 speakers, covering a broad spectrum of modern topics in Operations Research.



The official dinner of the event with presenters, sponsors, and organizers took place in an elegant and inspiring ambient at the planetarium Restaurant on the Shore of the Sankt Jorgen Lake in Copenhagen.

After the pitch presentation of the main sponsors **MOSEK** and **ØRSTED** an elegant buffet in the inspiring main hall of the Industriens Hus, gave the participants the opportunity to meet colleagues from different research fields in industry and academia.



The program started with two academic presentations from **Fredrik Odegaard**, Associate Professor with the Management Science department at Ivey Business School Canada, on a comprehensive and quantitative analysis of the effectiveness of Covid measures in Canada, followed by **Kevin Tierney**, Professor with Decision and Operation Technologies department, Bielefeld University, Germany on a benchmark comparison of deep reinforcement learning techniques for solving combinatorial optimization problems.

The program continued with a co-presentation from **Filippo Focacci**, PhD, Co-founder and CEO at Decision Brain and **Ella Pirttikangas**, Business Analyst at Decision Brain on applications of Mathematical Optimization of Maintenance and Workforce Planning, and a presentation from **Rikke Kristiansen**, Senior Simulation Modeling Engineer from Shoreline on Optimizing maintenance of off-shore wind farms.



After the coffee break **Mathias Hintze** and **Frederik Lehn**, Analysts at Mærsk Mc-Kinney Møller Centre for Zero Carbon Shipping presented different perspectives on Maritime Decarbonization and Simulations of maritime industry's transition to net-zero emissions.



Hope to see you at AOO 2025!

Finally, **Rafael Barfknecht**, PhD in theoretical Physics and Quantum Engineer at Kvantify, made an extraordinary presentation on the current practical and theoretical status in Quantum computing, including a performance analysis, titled "Prospects of quantum computing in optimization and related tasks".

The last event of the day was the conference bar where all guests had the opportunity to exchange impressions and enjoy a nice evening. Overall, there was a general enthusiasm from the interest generated by the exciting or engaging presentations on very modern concrete topics in the Operational Research coming from both academic and industrial worlds.

