Waving into the Future

Development of a Predictive Model for the Deployment of Airport Marshallers
A Case Study at Amsterdam Airport Schiphol

Anne Ruha

Delft University of Technology

TU Delft

AMS

Waving into the Future

Development of a Predictive Model for the Deployment of Airport Marshallers
A Case Study at Amsterdam Airport Schiphol

by

Anne Ruha

to obtain the degree of Master of Science at the Delft University of Technology, to be defended publicly on Thursday January 15, 2026.

Student number:	4669835	
Project number:	2025.MME.9149	
Date:	January 8, 2026	
Thesis committee:	Ir. M. B. (Mark) Duinkerken	TU Delft, Chair and first supervisor
	Dr. A. (Alessia) Napoleone	TU Delft, Supervisor
	N.M. (Nandini) Suresh	TU Delft, Committee member
	C. (Casper) Moll, MSc.	Amsterdam Airport Schiphol, Company supervisor
Mastertrack:	Multi-Machine Engineering	
Faculty:	Faculty of Mechanical Engineering, Delft	

Preface

In front of you lies my thesis, ”*Waving into the future: Development of a predictive model for the deployment of airport marshallers*” on which I have worked for the past few months. With the completion of this thesis, I conclude my masters in Mechanical Engineering with a specialisation in Multi-Machine Engineering at Delft University of Technology. This step marks the end of my student days, a period that I look back on with great pleasure.

When I first received this assignment, I had no idea how fascinating I would find the aviation sector. Over the past months I had the opportunity of exploring this sector upclose, and I am amazed with the complexity, precision and dynamics of the operation. During my time at Schiphol I encountered countless learning opportunities and unforgettable experiences. The absolute highlight was undoubtedly the moment that i was able to marshall an aircraft myself, performing the role I have been studying and modelling for months. The enthusiasm and openness of everyone at Schiphol made this project an inspiring journey and kept me motivated throughout.

First of all, I would like to express my gratitude to Casper, for making this thesis an unforgettable experience. Your enthusiasm and knowledge about the subject continuously inspired me to push further. I appreciate your willingness to answer every question I have, your patience and the time you always made for me. I still remember a conversation we had in one of the first weeks, when I asked basic questions about operations at Schiphol. You laughed and said that you couldn’t remember the last time you spoke to someone who did not know the difference between a tow and a pushback. Looking back on this conversation this perfectly shows how much I have learned during this time.

I am also deeply grateful to my TU Delft supervisors, Mark Duinkerken and Alessia Napoleone, for their valuable feedback and guidance throughout the process. Your advice helped me reflect on my report, bring structure to my research and strengthened the quality of this work.

I would also like to thank my Schiphol colleagues Just and Jop for their support, all the coffee breaks and lunches we have had together. Your company made my time not only productive but also filled with laughter. Furthermore, I want to thank all my colleagues at APM for creating such a pleasant and learning environment, making my time even more memorable.

Lastly, I would like to thank my family, especially my parents. Without them I would never have made it this far. They have supported me through all the years of my studies, all the way up to finishing this thesis. I would also like to thank my boyfriend and friends for their endless support. All the conversations we had, kind words, home-cooked meals and coffee breaks helped me stay motivated and focused till the end. Your support meant everything to me.

Anne Ruha
Delft, January 2026

Summary

The objective of this study is to develop a predictive model that can accurately predict the marshaller task demand at an airport. Amsterdam Airport Schiphol is used as a case study for this research. Marshallers form a critical link between ground operations and airside operations, ensuring safe aircraft movements and a continuous operation. Findings indicate a misalignment between static staffing practices and dynamic operational demand. The aim is to build a data-driven model using historical data and influencing factors. ADS-B vehicle data, geofence polygons, aircraft arrival times, and engine testing records were used to determine the task demand. A process of spatial matching, task labelling, and segmentation was used to combine these datasets, after which the labelled segments were then combined into hourly counts. Resulting in an aggregated dataset with hourly task count that can be used in a machine learning model. LightGBM was implemented as a multi-output model that forecasts all task types jointly. The models were trained on nine months of data and performance was evaluated using the Mean Absolute Error, Root Mean Squared Error, Mean Error and the Coefficient of Determination. The models were validated each on their own forecasting horizons: 24 hours, 168 hours, and 2160 hours, and compared to two baselines: Seasonal naïve, and weekly hourly average. The results show that the predictive capability strongly differs between the task type. Docking is the only task that can be forecasted reliably, it follows daily patterns and has a clear link with aircraft arrivals. Docking shows stable performances over all horizons with RMSE values between 1.66 and 1.75 and MAE values between 1.26 and 1.35. The model outperformed the baselines on all prediction horizons and on all evaluation metrics. The other tasks did not show any valuable forecasting, with R2 values close to zero or negative. This indicates that these tasks are irregular or occur in low volumes. Making them not suitable to provide reliable hourly forecasts with the current data. Future work could assess if additional operational factors improve the predictive structure of the other three tasks. A longer dataset may also reveal patterns that are not visible in the current nine month data. Other resolutions such as 15-minute or 30-minute time intervals, or shift intervals may be relevant.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

Abbreviation	Definition
AAS	Amsterdam Airport Schiphol
AIBT	Actual In-Block Time
ANN	Artificial Neural Network
APC	Apron Control
ASE	Airside Support Employee
ADS-B	Automatic Dependent Surveillance–Broadcast
BC	Bird Controller
DES	Discrete Event Simulation
EASA	European Union Aviation Safety Agency
Etesting	Engine testing
EO	Exact Optimization
FOD	Foreign Object Debris
GA	Genetic Algorithms
GBM	Gradient Boosting Machine
ICAO	International Civil Aviation Organization
IP	Integer Programming
IQR	Interquartile range
LGBM	LightGBM
LP	Linear Programming
LVNL	Luchtverkeersleiding Nederland
LVO	Low Visibility Operations
MH	(Meta)Heuristics
MILP	Mixed-Integer Linear Programming
ML	Machine Learning
MAE	Mean Absolute Error
ME	Mean Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
SA	SARIMA
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
SARIMAX	Seasonal Auto-Regressive Integrated Moving Average with Exogenous Variables
SAX	SARIMAX
SHAP	Shapley Additive Explanations
TaS	Tabu Search
TS	Time Series
UTC	Coordinated Universal Time
VDGS	Visual Docking Guidance System
VOP	Vliegtuigopstelplaats (Aircraft stand)
XGB	Extreme Gradient Boosting

1 Introduction

This chapter provides an introduction to the research topic of workforce planning in airport operations with a focus on the specialisation of marshallers. Section 1.1 presents the background of the study, describing the role of airport operations, their division into landside and airside processes, and the increasing complexity caused by traffic growth and staff shortages. Section 1.2 defines the problem addressed in this research, emphasising the mismatch between variable operational demand and static workforce planning methods. Section 1.3 outlines the knowledge gap in the literature, showing that although workforce scheduling has been studied extensively in related domains such as turnaround management and crew scheduling, there is little attention to marshallers in scientific forecasting models. Section 1.4 describes the research objective and scope, including the levels of workforce planning and the limits of the study. Section 1.5 introduces the case study of Amsterdam Airport Schiphol as the environment in which the predictive model will be developed and tested. Finally, Section 1.6 presents the outline of this thesis, providing an overview of the chapters and the corresponding research subquestions.

1.1. Background

Airport operations are a vital link in the chain of an airport. Every flight ultimately depends on what happens on the ground, without the activities that take place between arrival and departure, aircraft cannot leave on time, and the whole system would experience delays and fall apart. These activities are all grouped to airport operations. Airport operations are all the processes and activities that take place continuously inside and around an airport that ensure smooth, safe, and efficient functioning for passengers, staff, and aircraft. This includes passenger services, aircraft support, flight coordination, and others. Effective airport operations are essential to maintain schedules and safety standards, but also guarantee a positive experience for travellers and staff. This requires careful planning, coordination and the integration of multiple stakeholders to handle the complexity of modern air travel.

Airport operations cover all processes happening on different parts of the airport such as the terminal, landside and airside. Processes related to this are check in of the passengers, baggage handling, security, and the coordination of the arrival and departure of aircraft. A disruption on the ground almost always leads to wider delays, and because air transport is a network, even a small disturbance at one airport can cause problems at many others. This shows why the organisation of airport operations is such an important part of aviation. [2, 16, 38]

These airport operations can be divided into airside and landside operations. The physical division is at security in the terminal, before security is landside and after is airside. A rough outline can be seen in (Figure 1.1). Airside operations are related to aircraft activities and include runways, taxiways, aprons and all tasks necessary to prepare aircraft for a flight. Access to the airside zone requires screening for both passengers and staff to ensure security [20, 67]. Landside operations, on the other hand encompass passenger terminals, cargo facilities, and other elements of the airport's land-based infrastructure [67].

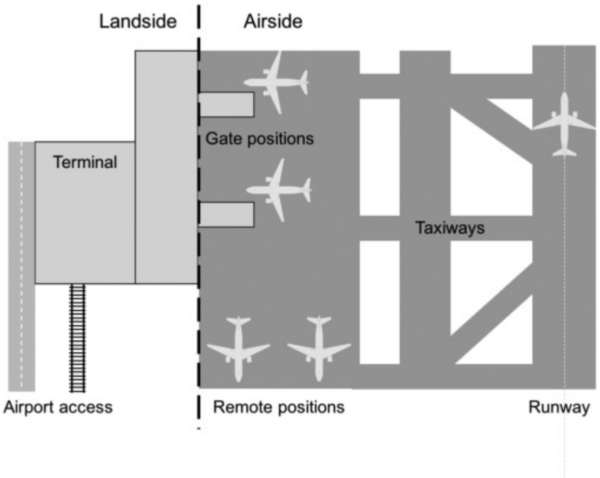


Figure 1.1: Generic division of landside and airside at an airport [57]

Regarding the activities that happen on airside outside the terminal, they can be divided into two main groups: airside operations and ground operations. Airport ground operations, also referred to as ground handling, cover those activities required by an airline between landing and take-off of the aircraft, such as marshalling of aircraft, (un)loading of baggage, refuelling, cleaning, catering, baggage handling, passenger handling, cargo handling, aircraft maintenance, and aviation security services [30].

Airside operations, on the other hand, involve responsibilities such as allocating aircraft parking and escort services, this also entails reacting to airside incidents, accidents, and emergencies. In addition, a variety of responsibilities that need to be monitored and efficiently managed through a variety of techniques include supervising runway and taxiway inspections, enforcing airside driving, contractor management, wildlife hazard management, and foreign object debris management [7]. The division of these responsibilities is not the same at every airport. Depending on the organisational structure, different parties such as the airport operator, ground handling companies, or airlines may take charge of specific tasks. Because these roles are distributed differently across locations, close coordination and cooperation among all actors is essential to keep the overall airport operation safe and efficient.

In recent years, air traffic has continued to grow, both in terms of passenger demand and the number of scheduled flights. With this increase, delays have also been rising, as more flights need to be fitted into already congested schedules [36]. At the same time, labour shortages in general and strikes in the ground handling sector, such as those at KLM in the Netherlands [40, 39], have highlighted the vulnerability of airport operations to staffing disruptions. Combined with stricter regulatory demands and the introduction of new technologies, these developments make airport operations more complex than ever. This growing complexity reinforces the need for better planning and predictive methods to ensure that ground and airside operations can continue to function reliably under increasing pressure.

1.2. Problem

Airside and ground operations are dependent on their personnel. The safe and efficient functioning depend directly on the availability of sufficient staff at the right time and place. Many of the procedures associated with the operations require human presence. If not enough staff are available, these processes stall, with direct consequences for punctuality, safety, and efficiency across the airport system.

The challenge lies in the fact that workforce demand in airside and ground operations is highly variable. Aircraft arrivals often cluster in waves, creating peak demand within a short time frame. Certain conditions can suddenly shift the demand from automated to manual procedures. Safety related tasks can appear at irregular moments with irregular time intervals. These combining factors cause a demand pattern that is difficult to anticipate with static rosters.

When a reliable prediction is absent, organisations often face two extremes: overstaffing and understaffing. Overstaffing leads to inefficiencies, higher costs, and the potential of employees without tasks risk losing concentration and motivation [8, 66]. Understaffing, on the other hand, has more immediate consequences. It increases workload pressure on present staff, reduces opportunities to take proper breaks and lowers job satisfaction [4]. Understaffing is especially problematic as it reinforces itself. Higher work pressure is the most common consequence of staff shortages [12]. In ground and airside operations, where tasks are time critical this cycle is particularly painful. Understaffing increases the work pressure, high pressure drives staff away, the staff shortage increases, and the cycle repeats. When not enough staff are available, tasks may be carried out too late or not at all, leading to operational disruption and delays. Insufficient staffing can also cause inbound aircraft to wait before docking, delaying flights and causing unnecessary emissions.

Marshallers are a good example of a group support employees in airside and ground operations that illustrate this problem. They carry out a mix of routine and ad-hoc tasks and are often required at the moment when automation fails or unusual events occur. Their workload varies, making it difficult to match supply and demand. Their demand is highly variable and operationally critical.

In current practice, workforce demand in ground and airside operations is treated with a deterministic input in planning models, or staffing models are fixed based on experience and intuition. These approaches overlook the variable task demand in these operations. The gap between how demand is modelled and how it actually behaves leads to a mismatch in supply and demand. As a result leading to inefficiencies, delays and risks to safety. Developing a data-driven predictive method that captures variability is therefore essential to prevent disruption and to sustain airport operations under growing pressure.

1.3. Literature gap

Workforce scheduling has been studied in sectors where personnel demand fluctuates and operations are continuous. Several studies illustrate how predictive models are applied to optimise workforce deployment across multiple aviation tasks. Crew scheduling and aviation maintenance are two goals within this industry. Heuristics methods and integer programming are used improve aircraft maintenance workforce planning, optimising both task allocation and spare parts forecasting [21]. Exact optimisation is combined with a worst fit decreasing heuristic to distribute maintenance tasks [18]. Within crew scheduling, various models are considered. Integer programming is used to optimise workforce scheduling while taking into account labour agreements and cost efficiency [65]. A combination of exact optimisation and metaheuristics is applied to create crew pairings, reducing crew idle time and improving workforce utilisation [64]. These approaches have shown that workforce scheduling problems can be addressed more effectively using data-driven methods than with fixed rules or intuition alone.

With regard to ground handling services, research has also been conducted. Different studies focused on baggage handling, a study focused on the workforce optimisation problem for airport ground handling operations, focussing on baggage loading and unloading [17]. Machine learning techniques are applied to predict future workload, but used for resource allocation in ground handling at Cairo International Airport [37].

Operational planning has a direct impact on emission levels. A relevant study integrated environmental factors into airline crew scheduling. Aiming to minimise fuel consumption and greenhouse gas emissions while keeping control of the staffing costs [63]. A non-dominated sorted genetic algorithm was applied in the building sector to optimise multisite workforce scheduling, accounting for project duration, costs, and greenhouse gas emissions [43]. Suggesting that workforce deployment can be used to reduce emissions without significantly increasing operational costs.

Within the literature different names are used for marshallers. They are referred to as: airport marshaller, aircraft marshaller, ground marshaller and sometimes also a variant with marshall. In the ICAO regulations, an organisation with the mission to establish standards across 193 countries, there is always referred to as (aircraft) marshaller. There is limited research available on marshallers, especially regarding their deployment. Previous studies have examined the acceptance of visual docking guidance systems and the potential impact on ground marshalling jobs [1, 41]. The use of innovative technologies, such as virtual reality, to improve marshaller training and safety [61], and the recognition of hand and body signals for gesture recognition [14, 59] have been researched. These studies focus on improving task execution but not on predicting when and how many marshallers are required.

Within the literature, a lot of research has been conducted across various sectors regarding workforce problems. However, there is currently no research that specifically focuses on the deployment of marshallers. Existing studies have explored the aviation industry regarding cabin crew and maintenance staff, but marshallers remain unexplored despite their varied workload and need for availability to ensure safe and efficient ground operations. While parallels can be drawn from other sectors, the unique variability of marshallers means that findings from other workforce researches cannot be simply applied. There is a clear gap in the literature showing how predictive modelling can forecast the required marshallers necessary in this dynamic environment.

1.4. Research objective and scope

The objective of this study is to develop a predictive model that can accurately predict the marshaller task demand at an airport. In order to obtain a realistic presentation of the situation, Amsterdam Airport Schiphol is used as a case study for this research. Accurate forecasting of task demand is essential to ensure operational continuity, while indirectly minimising inefficiencies caused by overstaffing and reducing the safety and workload risks associated with understaffing. By providing reliable predictions, the model can eventually contribute both to the short-term allocation of staff during operations and to the tactical scheduling of shifts in advance.

Workforce planning problems are commonly structured into three levels, defined by the length of the planning horizon and the type of decisions involved: The three levels of workforce planning are strategic, tactical and operational and are divided by the length of the planning horizon.

- Strategic level: long term planning. Covering multiple years and includes decisions such as workforce size, hiring policies and training strategies.
- Tactical level: mid-term planning. Usually weeks to months, focusing on creating staff schedules and aligning workforce capacity.
- Operational level: Short-term planning. Covering days, decisions are made about the exact deployment of staff to meet the immediate operational needs on a daily basis.

This study focuses on the alignment between the tactical and operational level of workforce planning. It contributes to the operational level as the demand on a daily basis will be reviewed that could support the deployment, ensuring enough staff is available each shift to match the actual workload. On the other hand, it provides input for the tactical level, as accurate predictions allow planners to create schedules for the upcoming weeks that better reflect the fluctuations in the demand.

The physical boundaries of this research are the platforms on the airside of Amsterdam Airport Schiphol. This includes the aprons, aircraft stands, taxiways, taxilanes and service roads used by marshallers vehicles. In practice, this represents the area where all vehicles on the airport are able to move to execute their specific tasks. Processes outside of the airside zone, such as terminal passenger handling, baggage processing, or air traffic control are excluded. The focus is therefore explicitly on the tasks and mobility of marshallers within the airside operational system.

The aim is to build a data-driven model using historical data and influencing factors. Different factors will be considered, including the number of arriving flights, average durations of different categorised tasks, and additional circumstances such as maintenance activities. Based on these variables, a prediction will be made of the daily marshallers' requirements. To achieve this, it is first necessary to understand the current planning and operational processes for marshallers. In addition, the relevant influencing factors need to be identified, the available data collected, and the data processed so it can be used in a prediction model. These considerations lead to the following research question:

What is the predictive capability of forecasting hourly marshaller task demand using historical operational data and influencing factors?

In order to answer this research question, the following subquestions have been drawn up:

1. What is the current operational situation regarding the deployment and responsibilities of marshallers, and what are the factors influencing their daily workload and demand at AAS?
2. Which type of predictive model is most suitable to forecast the marshallers' task demand?
3. What relevant data is available, and how can it be prepared for use in a predictive model?
4. How can a predictive model be developed using the prepared dataset to forecast the marshallers' task demand?
5. How can the developed predictive model be validated to ensure accurate and reliable forecasts of marshaller demand?

1.5. Case study Amsterdam Airport Schiphol

This research uses Amsterdam Airport Schiphol (AAS) as a case study to develop and validate a predictive model for marshaller task demand. Schiphol is the largest airport in the Netherlands and one of Europe's busiest hubs, processing more than 470.000 air transport movements and more than 66 million passengers in 2024 [46]. With more than 36 % of its passengers transferring, Schiphol functions as one of Europe's most connected hub airports [45, 46]. Creating concentrated inbound and outbound traffic, which directly reflects on the workload of ground and airside personnel.

Schiphol airport operates six runways. Depending on the availability of the runways, the wind and weather conditions and the environmental regulations for runway use, different combinations of two or three runways are used at the same time [44]. As runway use changes throughout the day, taxi routes and flow patterns also change constantly. Schiphol consists of three terminal and multiple piers and remote platforms with aircraft stands. Runways, platforms and aircraft stands are connected with an extensive taxiway network. An overview of the layout of Schiphol can be seen in Figure 1.3.

More than half of the aircraft movements at the airport is from KLM. As Schiphol acts as a main hub for KLM, it is characterised by strong inbound and outbound waves of flights, seen in Figure 1.2. These waves are designed to maximise opportunities for transfer passengers, but also create sharp peaks in relatively short time periods. Within these peaks, the workload is higher than outside these periods. This leads to a strong variability in operational demand across the day.

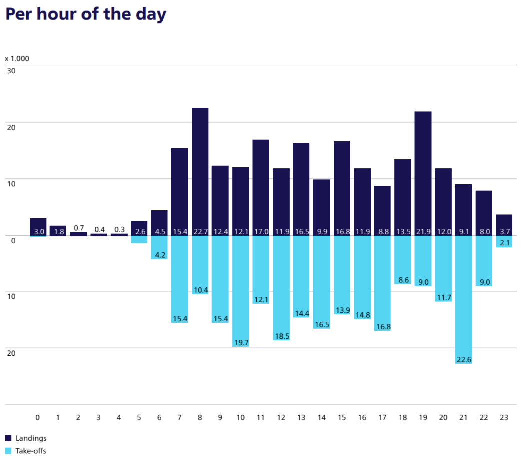


Figure 1.2: Landings and take-offs per hour of the day in 2024 at AAS [46]

Disruptions, such as delays in inbound flights or rerouting due to temporary maintenance works, can have a rolling effect across the tightly scheduled planning. Resulting in the fact that predicting the demand for staff such as marshallers cannot rely on averages only. But requires an approach that captures both the overall operations, but also the short-term fluctuations.

The combination of Schiphols high traffic volume, multiple runways, centralised layout and the role as a transfer hub creates a complex environment. These characteristics make AAS an ideal case study to develop and test a predictive model for task demand to determine the required number of marshallers.

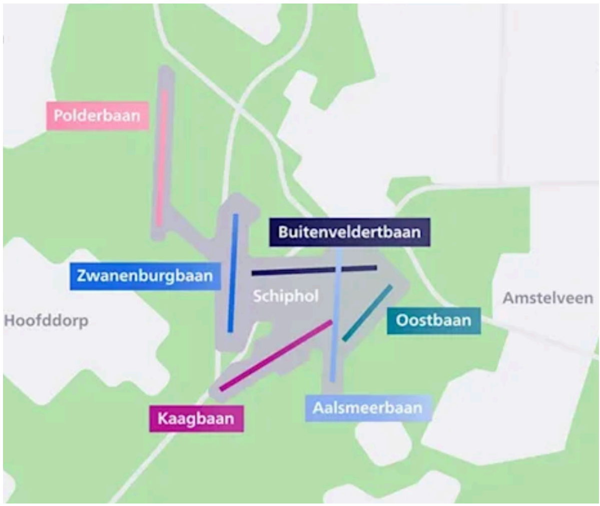


Figure 1.3: Map layout Amsterdam Airport Schiphol [49]

It can provide valuable insights into workforce planning at other airports that face similar challenges of variability in operational demand.

1.6. Outline of the report

The proposed structure of this research is visualised in Figure 1.4. Showing the chapters, which sub-questions is answered and which method is applied. The methodology combines, literature review, the collection and analysis of qualitative data, and quantitative data, and model development and validation. Chapter 2 discusses the current state at AAS, answering sub-question 1 using data analysis of quantitative data and visualising processes in a swimlane diagram. Chapter 3 provides insights from the literature review to determine which type of predictive model is most suitable, answering sub-question 2. Chapter 4 focusses on data preparation including the selection, processing and structuring of relevant datasets, answering sub-question 3. Chapter 5 describes the development of the predictive model answering sub-question 4, and Chapter 6 the validation of the model, answering sub-question 5. Finally, Chapter 7 gives the conclusion and discussion of this study.

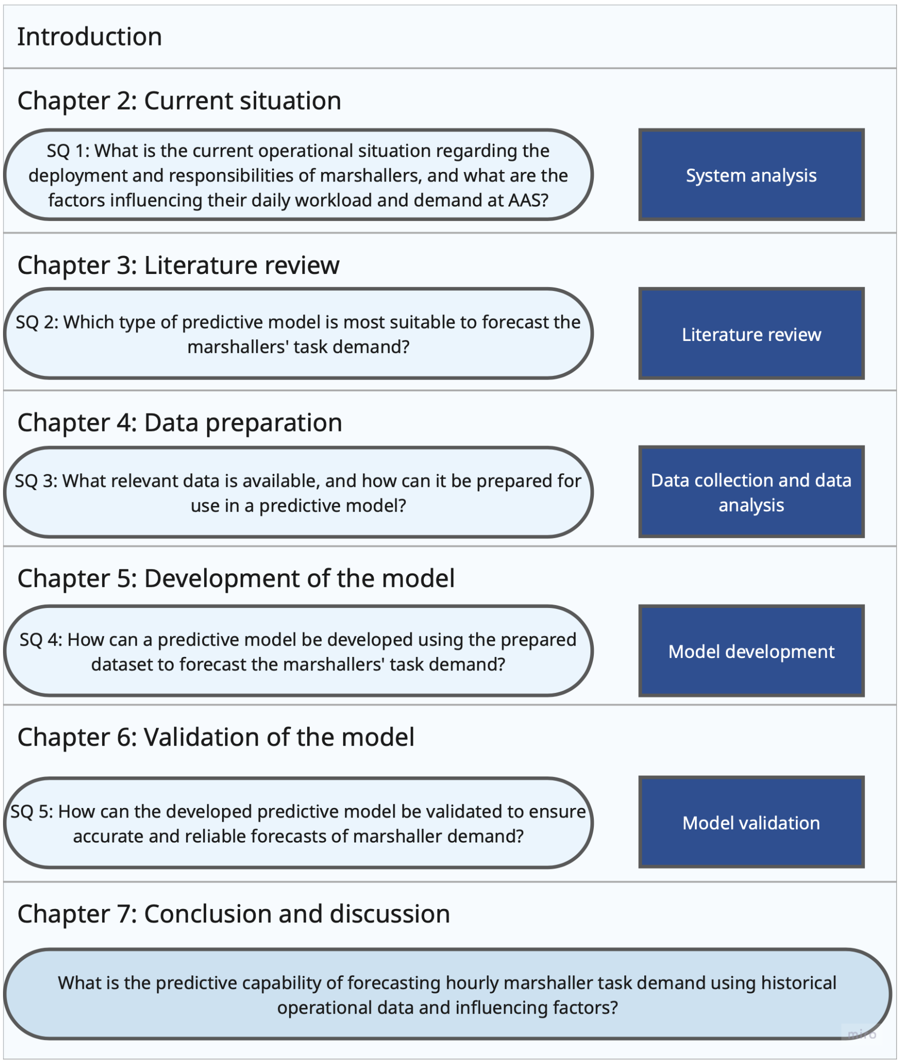

Introduction
Chapter 2: Current situation
SQ 1: What is the current operational situation regarding the deployment and responsibilities of marshallers, and what are the factors influencing their daily workload and demand at AAS?
System analysis
Chapter 3: Literature review
SQ 2: Which type of predictive model is most suitable to forecast the marshallers' task demand?
Literature review
Chapter 4: Data preparation
SQ 3: What relevant data is available, and how can it be prepared for use in a predictive model?
Data collection and data analysis
Chapter 5: Development of the model
SQ 4: How can a predictive model be developed using the prepared dataset to forecast the marshallers' task demand?
Model development
Chapter 6: Validation of the model
SQ 5: How can the developed predictive model be validated to ensure accurate and reliable forecasts of marshaller demand?
Model validation
Chapter 7: Conclusion and discussion
What is the predictive capability of forecasting hourly marshaller task demand using historical operational data and influencing factors?

Figure 1.4: Structure of the research

2 Current situation

This chapter answers the first sub-question: *What is the current operational situation regarding the deployment and responsibilities of marshallers, and what are the factors influencing their daily workload and demand?*. An overview of airport operations is given in Section 2.1, which also introduces the differences between ground and airside operations and their interdependence. Section 2.2 explains how these operations are organised at Amsterdam Airport Schiphol and identifies the marshaller's role within this structure. Section 2.3 describes the tasks of marshallers and the regulations, procedures and operational conditions that come with. The operational factors that impact marshallers' workload are covered in Section 2.4, while Section 2.5 outlines the organisational structure and communication between the actors involved. Finally, Section 2.6 examines the current staffing approach and available data, providing insight into the relationship between flight activity and marshaller demand. Section 2.7 concludes the chapter.

Airports function as a operational system that provide safe and reliable movements for aircraft, passengers and cargo. Marshallers contribute to this system by ensuring movements of aircraft within airside can be carried out safely and without delays. The system boundaries in this research are limited to the airside areas that play a direct role in these processes. These include runways, taxiways, aprons, and aircraft stands. Components of this system include infrastructure elements such as VDGS, and equipment such as marshaller vehicles. Activities and procedures on the airside are planned, coordinated, and carried out by operational actors. These actors include Air traffic control, airlines, ground handling companies, the airport operator's apron and safety services, and those in charge of monitoring, safety, and incident response. Marshallers operate at the point where actors, processes and components intersect.

2.1. Airport operations

Airport operations include all the processes and activities that take place inside and around the airport to ensure safe efficient and seamless functioning of aviation. This handles movement of aircraft and the handling of passengers, baggage, and cargo, but also handles safety, maintenance, and the coordination of tasks that enable every departure and arrival to occur smoothly. The coordination of these operations are complex due to all activities that take place in parallel. Each one performed by different companies, with their own responsibilities, but dependent on each others timing. To manage this complexity airport operations are divided into two domains: ground operations and airside operations (Figure 2.1).

2.1.1. Ground operations

Ground operations refer to all activities required to prepare an aircraft for its next flight, also known as the turnaround process. The turnaround process at airside outside the terminal starts at the so called ”on-block time”, when the aircraft reaches the parking position after landing. The process ends at the ”off-block time”, when the aircraft leaves the stand to taxi for departure [57].

Ground operations include a wide range of services: aircraft refuelling, baggage and cargo loading and unloading, passenger handling, catering, cleaning, pushback, de-icing, towing. But also setting and removing the chocks at landing and take-off, apron safety, connecting the power supply at the stand, inspection of the aircraft stands, docking of the aircraft, management of equipment on the platform, and handling of the jet bridge [36]. Some of these processes can occur parallel to each other, while others have a strict order in which they have to be executed [57].

All these processes are in general provided by more than one service provider [57]. The exact division of these responsibilities varies between airports. The main stakeholders in these operations are the airport operator, airlines, and handlers. That is, even though all airports have corresponding functions, the operational organisation behind them can differ significantly.

2.1.2. Airside operations

Airside operations refer to all activities related to aircraft operations that take place in the airside area of an airport, which includes runways, taxiways, aprons and aircraft parking stands. They cover all the operational aspects on the airside of the airport such as aircraft landing, taxiing and takeoff. These operations ensure that the flow of aircraft, vehicles, and personnel is safe and efficient [7, 27].

Typical airside processes include aircraft guidance and docking, apron inspections, stand allocation, additional services such as wildlife management, marshalling, and snow removal contribute to keeping the runways and taxiways operational under all weather conditions. But also several procedures related to reacting to airside incidents such as accidents or emergencies. Liu et al. (2024) [35] categorised airside operations into four main categories: gate assignment, aircraft ground movements, runway scheduling, and stand holding. Gate assignment assigns flights to aircraft stands, runway scheduling schedules runway assignment for arrival and departure aircraft, including the sequence and schedules of the aircraft. Aircraft ground movement is the link between gate assignment and runway scheduling and navigates aircraft safely from aircraft stand to runways or the other way around. Stand holding plans each aircraft's off block time to ensure that aircraft arrive at runways in a proper order so they don't have to wait for an extended period of time. Stand holding is often paired with aircraft ground movements or runway scheduling to improve results.

The organisation of airside operations typically falls under the responsibility of the airport operator. They coordinate with air traffic control, maintenance and safety departments. However, the exact organisation differs per airport. Meaning, the way they are executed and who is responsible for each process can vary per airport.

The distinction between airside operations and ground operations within airport operations is illustrated in the Venn diagram in Figure 2.1. It shows that not all tasks can be classified exclusively as either airside operations or ground operations. The responsibility of several processes overlap such as docking, apron safety and the operation of jet bridges. This overlap not only demonstrates the dependence between both these domains but also contributes to the different organisational structures found at airports. Certain functions may fall under different departments depending on how responsibilities are divided among the airport operator, airlines, ground handlers, and all other parties related, resulting in variations in how airport operations are managed worldwide. The next section therefore focuses on how these operations are organised specifically at Amsterdam Airport Schiphol.

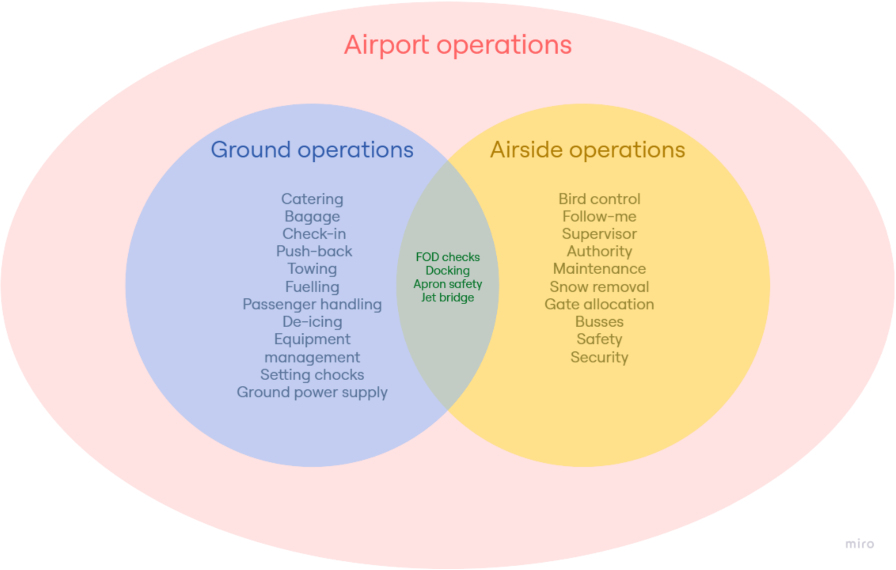


Figure 2.1: Venn diagram airport operations

2.2. Airport operations at AAS

The airport operations at AAS are organised through different stakeholders, each responsible for a different part of the process. The airport operator is AAS, and works closely together with other stakeholders to ensure safe and efficient operations on the ground. Within AAS, the airport operations department is responsible for maintaining the operational continuity.

In the tower, safe traffic flows are maintained and divided into two functional levels. The upper section is occupied by LVNL (Luchtverkeersleiding Nederland) and responsible for managing all aircraft movements take that place between leaving or entering the Dutch airspace and taxiing to or from the aircraft stand. The lower section of the tower manages all movements that take place at airside, and consist out of several departments, including Apron Control (APC), gate planning, bus coordination, and towing control. APC monitors aircraft and vehicle movements on the apron. Gate planning allocates aircraft stands and manages last minute gate changes. Bus coordination manages the transfer of passengers between terminals and remote stands. Towing control handles all aircraft towing operations across airside.

Airlines are responsible for the handling of their own aircraft, often working closely together with ground handling companies. Airlines and ground handling companies perform the turnaround process at the aircraft stand, and are responsible for the pushback and towing of aircraft. The division of responsibilities between airlines and handlers depends on the organisational model of the carrier. While some airlines handle ground operations internally, others outsource the same operational tasks to external ground handling companies to handle these duties. Both airlines and handlers work under the airport's operational and safety regulations.

The apron office acts as the central coordination point for all airside operations. They make sure that operational flows run smoothly, processes connect properly, and disruptions are handled quickly. When something goes wrong, the apron office is the first point of contact and takes the lead in restoring continuity. In case of incidents or unsafe situations, they also contact the Airport Authority. The Airport Authority monitors safety and checks whether all airside activities comply with the rules. They carry out inspections, supervise daily operations, and step in after incidents such as collisions or damage.

Within the organisation of AAS, Aircraft Support Employees (ASE) are responsible for various tasks on the airside. They escort external parties in vehicles who are not permitted to drive unaccompanied on the airside, clean taxiways and aircraft stands, and are involved in the snow removal operation. ASE has two specialisations: Bird Controllers and Marshallers, indicated by respectively ASE-BC and ASE-M. Bird Controllers scare away birds to prevent them from entering aircraft engines and chase off other animals that may interfere with the operation. Marshallers ensure the safe movement of aircraft and helicopters, either at low altitude or on the ground.

Figure 2.2 illustrates the role of the marshaller in the broader system of airport operations. Their responsibilities are within airside operations and in the cross section of ground operations and airside operations. As will be discussed in the next section, marshallers perform a wide variety of operational tasks, serving as a connection between several functional areas within airport operations.

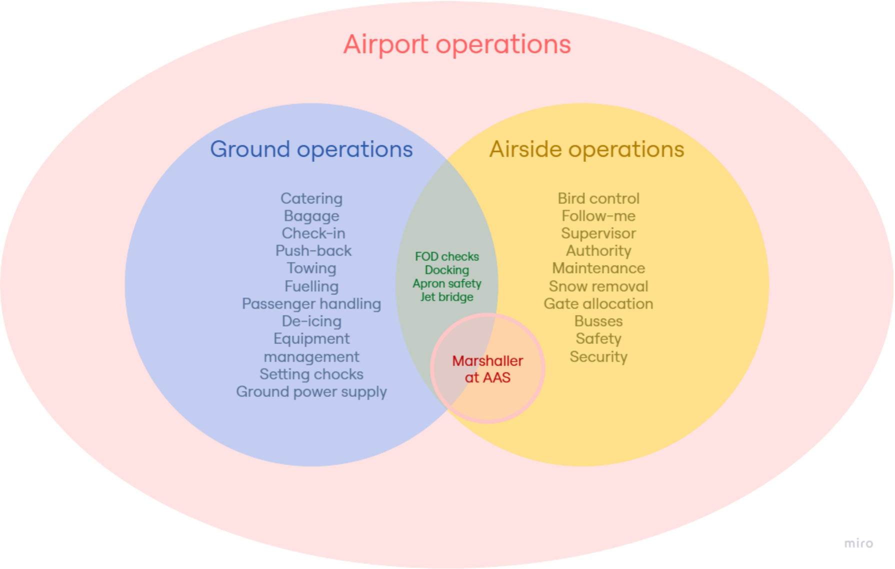


Figure 2.2: Venn diagram airport operations position marshaller

2.3. Tasks of marshallers

The specific description of the marshallers role is as follows: Marshalling is the directing of aircraft and helicopters, which are moving at low speed or low altitude at an airport, using arm and hand signals, possibly aided by optical tools, in order to give instructions to the pilot so they can manoeuvrer or land safely in a situation that is unclear or difficult to assess from their position. The main tasks of marshallers can be divided into 5 categories; follow me service, Docking, Engine testing, Controlling tasks and other tasks. They can be seen in the list beneath [50, 51, 52, 53, 54, 55, 56]:

1. Follow me service, guiding an aircraft:
   - to the correct aircraft stand when pilots are unfamiliar with the airport
   - to certain standard aircraft stands and platforms
   - under certain Low Visibility Operations (LVO)
   - that is larger than design standards allow
   - to the engine run-up area
2. Docking (Marshalling at the aircraft stand) when:
   - there is no VDGS available at the stand
   - the VDGS is malfunctioning
   - the ground handler is unavailable
   - the ground handler present is not authorized to use the VDGS
   - a dangerous situation arises during docking
   - the jet bridge is not in the parking position
   - the ground handler failed to operate the VDGS correctly

    - an aircraft is larger than standard design limits allow

3. Engine testing
    - Safeguarding engine tests on the stand
    - Safeguarding engine tests on the engine testing area
4. Controlling tasks
    - Ensuring the aircraft stand/handling area is free of Foreign Object Debris (FOD), equipment, and unauthorized personnel
    - Logging and handling VDGS malfunctions
    - Inspecting stands and verifying proper use of handling equipment
5. Other tasks
    - Following an aircraft to signal the end of a tow when anti-collision lights are broken
    - Bus coordination and assisting in the bus process

Follow me service and docking are tasks that are followed up because an aircraft arrives at the airport, while the other tasks are not explicitly related to arriving aircrafts.

2.3.1. follow me service

One of the core tasks of marshallers is to guide aircraft on the airside, either during taxi or when towed. The marshaller awaits the aircraft at the taxiway either because the marshaller notices themselves though the system or the task gets assigned by apron control (APC). Follow me service for an aircraft is required at specific stands or for certain aircraft types that exceed standard size limits.

The marshaller drives to the pick-up point and once the aircraft, already informed by LVNL about the follow me service, is positioned behind the marshaller, the vehicle activates the ’Follow-me’ lights, Indicating the start of the follow me service. The marshaller then drives the route to the correct aircraft stand or engine testing area, where the process may be followed by docking guidance or assistance during engine testing.

2.3.2. Docking

The procedure for docking an aircraft either begins when APC assigns the task or marshallers divide the responsibilities among themselves. Upon arriving at the aircraft stand the marshaller first checks whether the ground handling is present. If they are not yet on site the marshaller can communicate this, or if time allows, they will wait to communicate.

The marshaller performs a FOD inspection and checks if the area is free of other material that could obstruct safe docking. Once the stand is clear, the marshaller performs with one of the two options: either the VDGS is activated and compliance with the arriving aircraft is checked, or the marshaller takes over the docking process manually. In the second option, the marshaller indicates to the pilot the designated stand using correct hand signals.

The marshaller guides the aircraft using the correct hand signals to indicate whether the aircraft should move left, right or straight ahead. When the aircraft reaches the exact stop position for its type, the marshaller signals the stop sign. Next, the marshaller signals to the pilot to set the brakes, after which the ground handling team places the chocks in front of the wheels and, if applicable, connects the ground power unit.

Once the chocks are in place, the marshaller communicates this to the pilot and awaits in return the confirmation by the pilot. After this, the marshaller’s task at the aircraft stand is complete, and the marshaller can proceed to the next assignment. Finally, back in the vehicle, the marshaller communicates via the radio to APC the confirmation of the docking. Mentioning the aircraft registration number, aircraft stand and any special details.

2.3.3. Engine testing

Apart from these movement related tasks, they also assist in specific technical procedures like aircraft engine testing. Engine testing on the airside can take place in two locations, either at the aircraft stand or at the engine run up area. In both cases, marshallers play a key role in ensuring the procedure is carried out safely and according to protocol.

On the stand

Engine testing on the stand involves running the engines at ground idle power. A maximum of two engines may be tested at the same time, and the session can not exceed 10 minutes. An aircraft is only allowed to do this twice per day.

To carry out this procedure, several safety requirements must be met. Before the engine testing begins, the marshaller is responsible for checking the safety of the entire area. This includes the removal of any FOD and ground handling equipment, unless it is stored in assigned zones on the stand. A fire extinguisher must be available and the aircraft must be secured with wheel chocks. Furthermore, all doors and hatches must be closed and the anti-collision lights must be turned on. No people, vehicles, or obstacles may be within the danger zone of the engines. A Ground Engineer who has contact with the cockpit must also be present.

Once all these safety procedures are met, the marshaller allows the testing session to begin. During the test, the marshaller stays alert and makes sure that no disturbances or risks occur in the area. When safety is no longer guaranteed, the marshaller will stop the session immediately. The marshaller also listens to radio frequency 121.780 MHz, records the start time, and makes sure the session does not exceed the 10-minute limit. When the session ends, the marshaller reports the start and end times to APC.

Engine run up area

Before engine testing can begin at the run-up area, the aircraft must first be towed to the correct location. The marshaller receives this task from APC and is responsible for guiding the towed aircraft safely to the engine testing area. Once the marshaller arrives at the aircraft, the marshaller will get in contact with the driver of the towing device. The marshaller guides the aircraft to the run up area, positioning it in the correct wind direction, as communicated by APC.

At the run-up area, the same safety requirements apply as during engine testing on the aircraft stand. Before the engine run begins, the marshaller reports the start time to APC. Throughout the session, the marshaller monitors the situation to ensure safety is maintained. When the testing is completed, the marshallers reports the end time to APC and then guides the aircraft back to the parking position.

2.3.4. Controlling tasks & other tasks

Marshallers perform controlling tasks, which mainly involve ensuring the safety of the aircraft stands and handling area. These tasks include ensuring the area is free from any FOD, equipment or unauthorized personnel. This usually happens before the docking takes place. Furthermore they log and handle any VDGS malfunctions, either because they notice themselves though driving or because they are assigned by APC. Lastly, they inspect stands and verifying proper use of handling equipment.

In addition to these tasks they also perform other supporting tasks. These tasks almost never occur, however, a marshaller can be asked to assist. When the anti-collision lights of an aircraft are broken, a marshaller has to follow behind to signal the end of the towing procedure. When necessary they can help assist in the bus process by taking the role of bus coordination and keeping an overview.

2.3.5. EASA regulations

While the previous sections describe the tasks and operational factors affecting marshallers at AAS, it is also important to note that their activities in general are regulated by aviation safety standards. European Union Aviation Safety Agency (EASA) is an agency that is responsible for developing common safety and environmental rules at the European level in civil aviation [25]. These EASA regulations apply directly in all eu member states and cover many areas of airport operations, including how marshalling services are organized and used. In Easy Access Rules for Aerodromes Regulation (EU) No 139/2014 [24] these regulations are described.

First, marshalling is seen as one of the ways to guide aircraft during parking manoeuvres. A marshaller or an (advanced) VDGS can both be used as parking aids. Marshallers should use the official hand signals during marshalling that are set in the EASA regulation No 923/20121.

Second, there are several situations where a marshaller should be deployed. This is in cases where visual or advanced visual docking guidance systems do not exist or are unserviceable, or where guidance to aircraft parking is required to avoid a safety hazard. Furthermore, a marshaller is required to lead an aircraft that must cross a defect stop bar that is either physically disconnected or the lights of the stop bar are physically obscured. Guidance to the aircraft should be given either by marshalling hand signals or combined with radio, follow-me vehicle, or VDGS.

Third, EASA requires airports to set clear procedures for how marshalling is done. These procedures should explain the following: Ensuring the area within which the aircraft will be guided is clear of obstacles, conditions for using one or multiple marshallers or wing walkers, and actions to be taken in emergencies or incidents during marshalling.

Finally, airports must establish and ensure the implementation of a training program for the marshalling service. This programme should include fundamental knowledge and practical skills. After initial training, marshallers must also complete a proficiency check at least every 12 months. The initial training should cover at least the following aspects. Role and responsibilities, the visual hand signals, aircraft characteristics, safety, emergency, and low visibility procedures, driving at the apron, VDGS emergency stop procedures, and aircraft stand layouts. If new procedures are made or changes to existing procedures, marshallers should receive a briefing or a training.

2.3.6. Equipment

Marshallers are equipped with several tools that enable them to work safely and effectively on the airside. One of the most important elements is their own safety. They wear bright orange jackets or sweaters that contain reflective stripes, making them visible to pilots and other personnel at any given time, even in the night or adverse weather conditions. The uniform also includes dark blue trousers and safety shoes. To protect their hearing from the loud noises produced by aircraft engines, marshallers also wear ear protection while on shift.

In order to guide aircraft, they make use of two types of signalling equipment. Under normal conditions, standard batons are used to communicate with pilots (Figure 2.3a). During darkness and nights illuminated batons (Figure 2.3b) are used for better visibility, but also during heavy winds as it can be tough to work smoothly with the batons. Allowing for better communication with pilots during these circumstances.

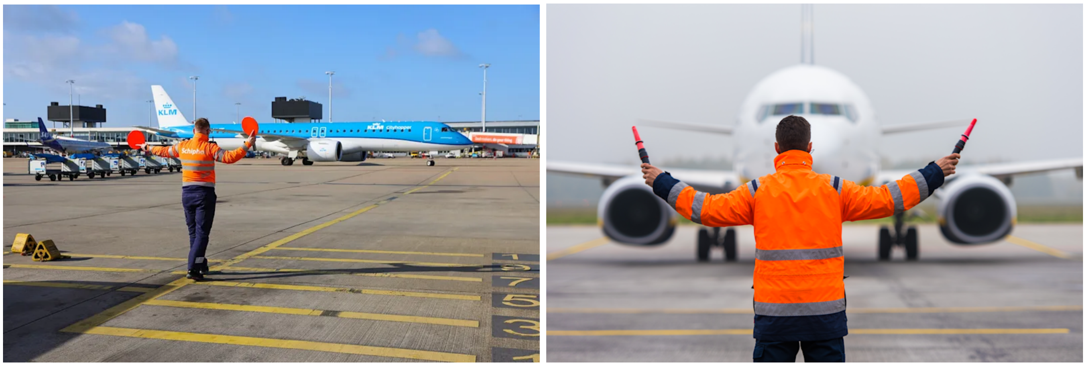


Figure 2.3: Marshaller with a) batons [48], b) illuminated batons [23]

To move across the platform, they drive around in yellow electric vehicles (Figure 2.4). Each one has a white stripe and 'airport marshaller' on it for visibility and is marked with a 'C' followed by a number for identification. Inside the vehicle, a large tablet is installed, which displays an overview of all arriving and departing flights. Displaying estimated times, designated stands, and maintenance activities. This gives them the option to handle certain tasks without orders from APC. Also, multiple radio systems are present in the vehicle. Marshallers monitor two channels constantly and simultaneously. One channel is the connection to APC, and the second is for communication among marshallers themselves.

Figure 2.4: Yellow vehicle marshallers drive around at the airside

2.3.7. Rest periods

During each shift, marshallers are assigned specific rest periods times as shown in Table 2.1. These rest periods are on fixed times to ensure regular rest periods, but also to make it not interfere with the peak periods of inbound flow. For the early and late shift, four and three rest slots are defined, each one lasting 45 minutes. During the night shift marshallers are allowed to schedule their own rest periods, as the workload during the night is lower. In addition to these planned rest periods they can take short rests, such as for coffee or restroom use. They can not take longer than 20 minutes and only allowed if they don't interfere with any ongoing operations.

To ensure operational efficiency, one marshaller at a time is allowed to take rest, leading to a situation where there is always one fewer marshaller available than in total scheduled in that shift. That is why the rest times are following up on each other, instead of overlapping, making the minimum staffing level always available.

At the beginning of each shift, C1 is responsible for determining the rest period order and communicating this to APC. This includes assigning each rest time to a corresponding marshaller. Everytime the marshaller takes a rest, this should be communicated with APC, and upon return to the field, signed in again. This ensures APC is always aware which marshallers are active in the field.

	Early shift	**Late shift**	**Night shift**
	06:30 - 14:30	14:30 - 22:30	22:30 - 06:30
Break 1	10:00 - 10:45	16:30 - 17:15	
Break 2	10:45 - 11:30	17:15 - 18:00	
Break 3	11:30 - 12:15	18:00 - 18:45	
Break 4	12:15 - 13:00		

Table 2.1: Shift schedule with corresponding rest period times

2.4. Operational influences

Besides the standard tasks described in Section 2.3, there are several other factors that can influence the workload of marshallers. One of these are the ad hoc tasks assigned by APC. Operating from the control tower, APC monitors all aircraft movements on the apron and is responsible for distributing unexpected or last-minute assignments to marshallers on duty. These ad hoc tasks include situations that are not planned in advance, such as:

1. Deploying follow me service and/or docking an aircraft to a new gate after a last-minute gate change
2. Inspecting oil spills on the airside
3. Verifying whether a broken jet bridge is positioned far enough from the aircraft to allow pushback
4. Assisting aircraft in turning around if they entered the wrong taxiway

5. Responding to and checking sudden malfunctions of the VDGS
6. Guiding aircraft with permanently broken anti-collision lights

Apart from these unexpected events construction or maintenance work on the airside can also influence the workload of marshallers. Maintenance work in certain areas can lead to extra required safety measures. Marshallers are often asked to support these situations with additional follow me service or docking tasks. For example, when a taxiway is temporarily narrowed due to construction, marshallers may need to guide aircraft through the reduced section manually. Another case occurs when the visibility of a VDGS is partially blocked by equipment. In such situations, a marshaller must manually dock the aircraft at the affected stand throughout the duration of the maintenance.

These examples illustrate that marshalling is not limited to routine operations. It also involves being able to respond quickly and safely to operational changes, ensuring the continuity and safety of the airside environment under non-standard conditions.

2.5. Organisational structure and actors

There are different actors involved in regard to the coordination and execution tasks of marshallers at the airport. Each plays a specific role to ensure ground operations. An overview of all actors can be seen in Figure 2.5 which provides a communication analysis.

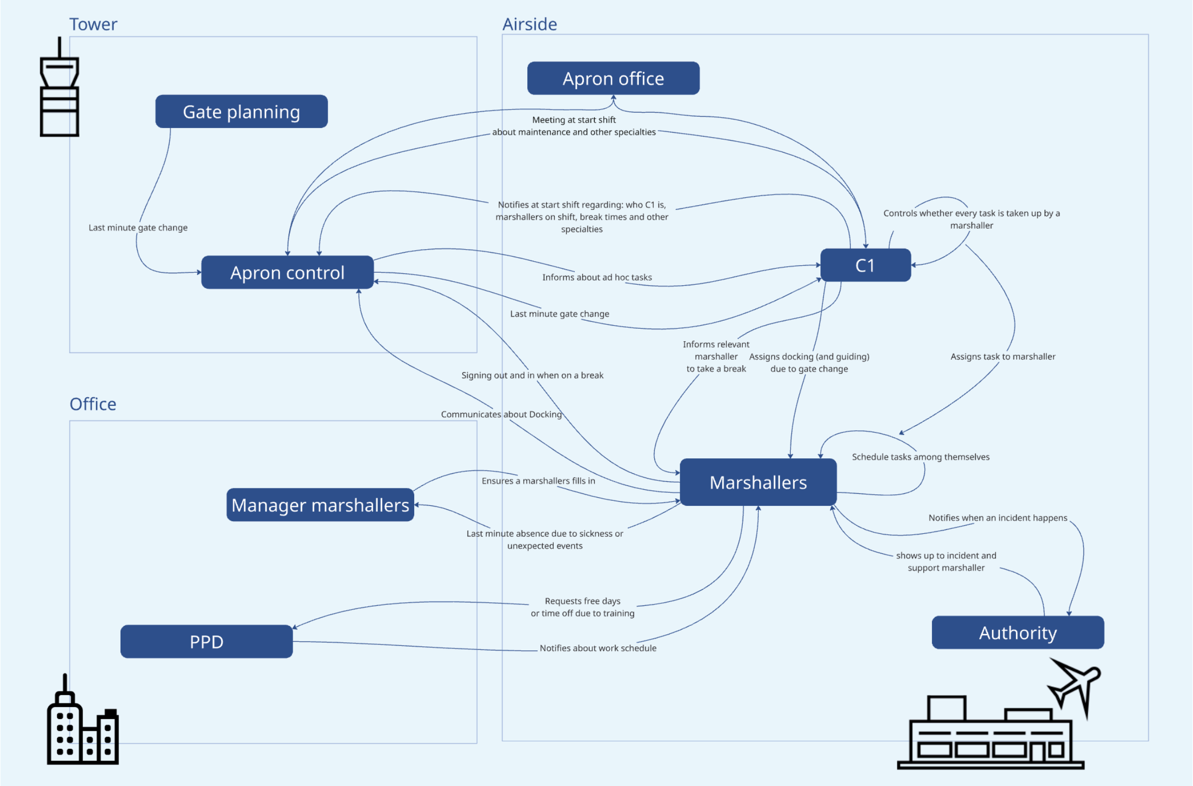

Tower
Airside
Office
Apron office
Gate planning
Meeting at start shift
about maintenance and other specialties
Notifies at start shift regarding: who C1 is, marshallers on shift, break times and other specialties
Controls whether every task is taken up by a marshaller
Last minute gate change
Apron control
C1
Informs about ad hoc tasks
Last minute gate change
Informs relevant marshaller to take a break
Assigns docking (and guiding) due to gate change
Assigns task to marshaller
Signing out and in when on a break
Communicates about Docking
Schedule tasks among themselves
Ensures a marshallers fills in
Marshallers
Manager marshallers
Last minute absence due to sickness or unexpected events
Notifies when an incident happens
shows up to incident and support marshaller
Requests free days or time off due to training
PPD
Notifies about work schedule
Authority

Figure 2.5: Communication analysis

APC operates from the control tower, which provides them with an overview of all elements on airside. This is both because they can visually see everything, but also because they have an overview on their computer screen where every vehicle on the airside is at that exact moment. They know exactly where aircraft are taxiing, where marshallers are located, which VDGS are out of order. The main control of assigning tasks lies with C1, but in special cases APC takes full control in assigning task to marshallers.

The planning department (PPD) is responsible for creating the schedules. They plan marshallers into shifts and draw up a five-week schedule. Marshallers can request days off in advance, or can let know when they have a training, and PPD adjusts the schedule accordingly.

Among the marshallers a key role is C1, who acts as the coordinating marshaller during a shift. At the start of every shift C1 checks in with APC with regard to break schedules, and which marshallers are on shift. C1 forms the point of contact for APC, Apron office and their ASE-M colleagues. C1 also attends a daily briefing with APC and the apron office to discuss operational issues and special activities for that shift.

The manager of the marshallers supervises the marshallers on a broader level and is the main point of contact for work-related questions. If there are last-minute absences, the manager is responsible for arranging a replacement.

The Apron office monitors and manages safety and operations on the airside. A marshaller has contact with the apron office if any operational disruptions occur, such as an oil leakage.

Authority is responsible for monitoring compliance with rules and permits within the restricted area and landside, and can be called upon by a marshaller to provide assistance.

Gate Planning works from the tower and prepares the gate allocation for all arriving and departing flights. They handle last-minute changes and coordinate these updates with APC.

To show these relations within a certain timeframe a swimlane diagram has been drawn up. It has been divided into two parts. The first part is scheduling, in which only marshallers, PPD and the manager of the marshallers play a role. This can be seen in Figure 2.6. The second part is what happens during a shift, in which Gate planning, APC, Apron office, C1 and marshallers play a role. This can be seen in Figure 2.7 an enlarged version can be seen in Appendix B.

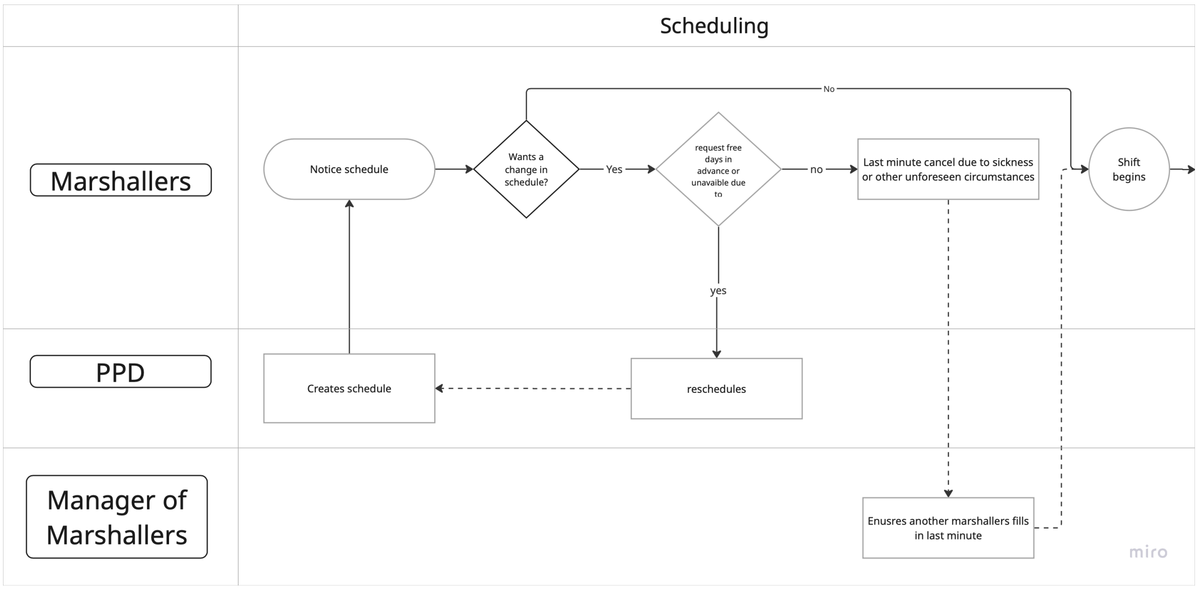

Scheduling
Marshallers
PPD
Manager of Marshallers
Notice schedule
Wants a change in schedule?
Yes
No
request free days in advance or unavaible due to
no
yes
Last minute cancel due to sickness or other unforeseen circumstances
Shift begins
Creates schedule
reschedules
Enusres another marshallers fills in last minute
miro

Figure 2.6: Swimlane diagram scheduling

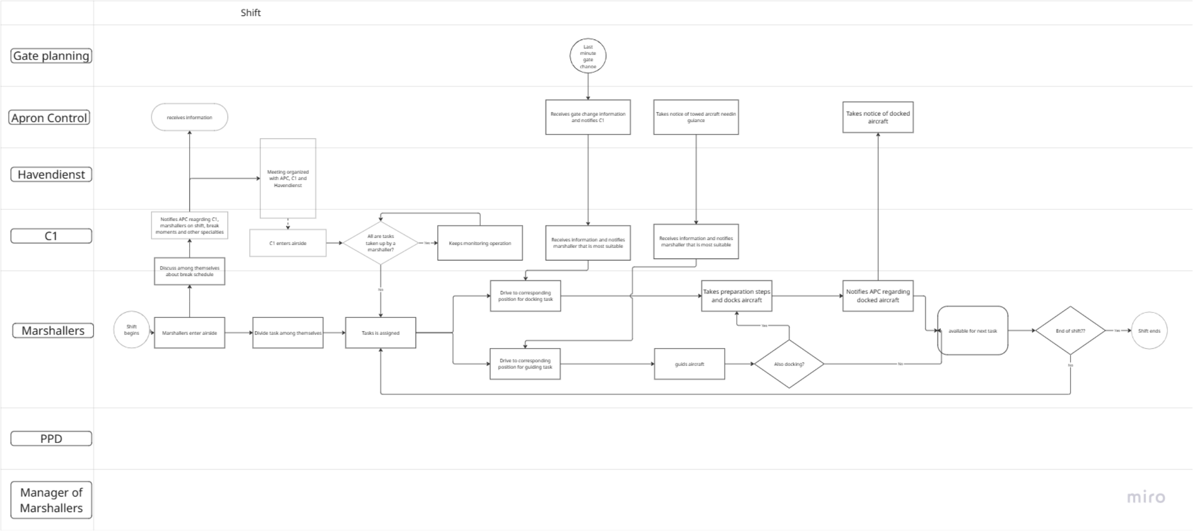

Shift
Gate planning
Apron Control
Havendienst
C1
Marshallers
PPD
Manager of Marshallers
Last minute gate chanoe
receives information
Receives gate change information and notifies C1
Takes notice of towed arcraft needin guiance
Takes notice of docked aircraft
Meeting organized with APC, C1 and Havendienst
Notifies APC reagrding C1, marshallers on shift, break moments and other specialties
C1 enters airside
All are tasks taken up by a marshaller?
Yes
Keeps monitoring operation
Receives information and notifies marshaller that is most suitable
Receives information and notifies marshaller that is most suitable
Discuss among themselves about break schedule
No
Shift begins
Marshallers enter airside
Divide task among themselves
Tasks is assigned
Drive to corresponding position for docking task
Drive to corresponding position for guiding task
Takes preparation steps and docks aircraft
Notifies APC regarding docked aircraft
guids aircraft
Also docking?
Yes
No
available for next task
End of shift??
Yes
No
Shift ends
miro

Figure 2.7: Swimlane diagram during shift

2.6. Demand and influencing factors

After describing the operational and organisational environment of the marshallers, this section goes into the current aspects of their staffing and workload, and the data availability to estimate the demand.

2.6.1. Current staffing and data availability

As the airport operates 24/7, marshallers work in three 8-hour shifts per day: the early shift (06:30 - 14:30), the late shift (14:30 - 22:30), and the night shift (22:30 - 06:30). The standard staffing levels for these shifts are 4,3, and 2, respectively. As can be seen in Table 2.2. This schedule remains constant throughout the whole year, regardless of seasonal influences or flight activity. Importantly, the current staffing model is not data-driven but based on experience and intuition.

Shift Times	Shift Name	Minimum Staffing level
06:30 - 14:30	Early	4
14:30 - 22:30	Late	3
22:30 - 06:30	Night	2

Table 2.2: Shift schedule with corresponding names and minimum staffing levels.

Meaning, there is currently no complete overview of the number of tasks performed by marshallers in a day, or the time each task takes. Therefore, it is not yet possible to accurately determine how many marshallers are required at a given time. However, in order to gain more insight into their workload, an internal pilot study was conducted. They were asked to record their activities using a digital form during their shift in their vehicle. The marshallers were asked to select a task category and, when relevant, a reason for execution before submitting the entry.

Although this pilot provided valuable first insights of task patterns, the number of responses decreased over the months, leading to incomplete datasets. As a result, the collected data can not be used as a quantitative input for predictive modelling. However, it does provide meaningful qualitative insights, allowing for identification of certain trends.

It shows that approximately 79% are reported as regular operational activities, 10% are follow me service activities, 7,3% are activities on request, 2,0% relate to engine testing, 1,3% are classified as other, and 0,3% is related to the jet bridge. The reason for regular activities were often left blank, which in most cases indicates standard docking procedures. Additional explanations include follow me service task, operational disturbances with a VDGS, or certain aircraft types are not registered in the system. For activities on requests, the most frequently mentioned reasons are the absence of a ground handler, a request from APC, or a malfunction of the VDGS.

Figure 2.8 illustrates the number of recorded task per hour throughout the day. To ensure consistency only the start times were used in this figure. The figure shows clear peaks at hours 8, 13, 15, and 19.

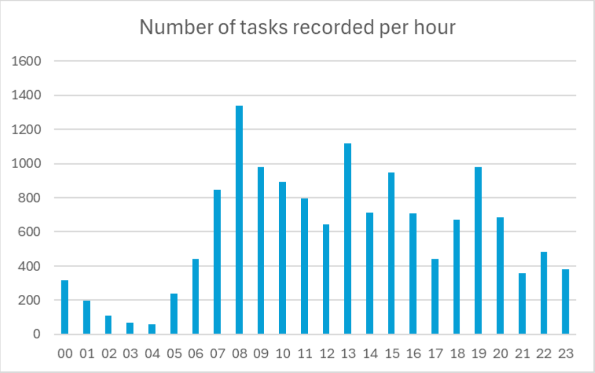
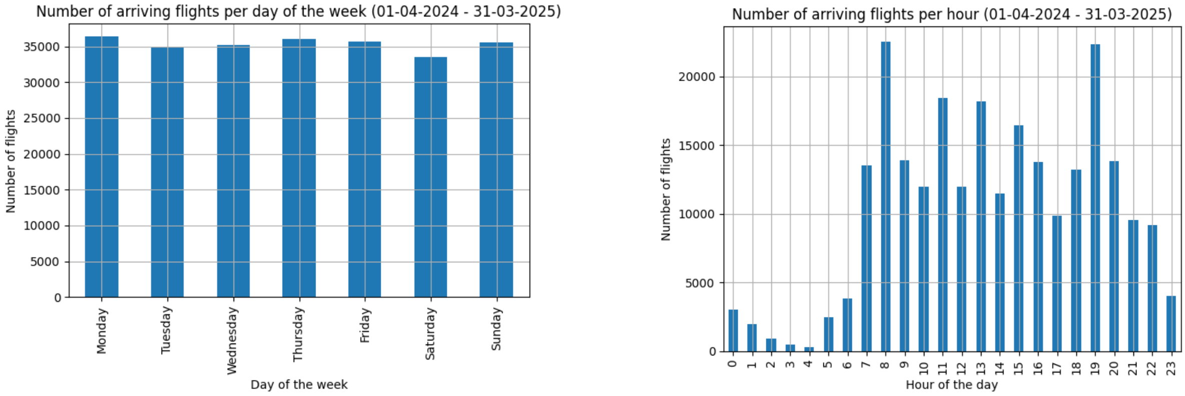


Figure 2.8: Number of marshaller tasks recorded per hour (pilot data).

2.6.2. Seasonality in flight arrivals

Operational observations suggest that the tasks performed by marshallers partly depend on the number of arriving flights. These arrivals are not the same throughout the year, variations can be observed on a monthly, weekly, and even daily basis. This is therefore an important factor to consider when developing the model. To gain insight into the extent of this variation, statistics for one year were analysed. The dataset covers the period from April 1, 2024, to March 31, 2025, and includes all arriving flights.

Figure 2.9a shows the number of arriving flights for each day of the week. It can be seen that Monday has the highest number of arrivals, while Saturday has the fewest, however the differences are small. Figure 2.9b presents the number of arriving flights per hour, where two major inbound peaks can be observed at 08:00 and 19:00, and three smaller peaks around 11:00, 13:00, and 15:00. Finally, the annual overview is shown in Figure 2.10. This figure clearly illustrates the difference between the winter season, which starts at the end of October, and the summer season, which begins at the end of March.



(a) Arrivals per day of the week

(b) Arrivals per hour

Figure 2.9: Arrivals per day of the week and per hour

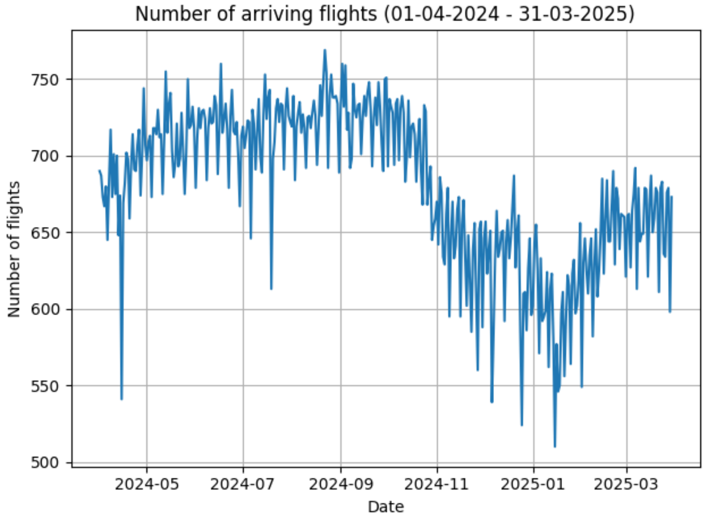


Figure 2.10: Arrivals per year

2.7. Operational drivers influencing marshaller workload

The marshaller workload in the current situation is influenced by predictable operational patterns, and irregular ad-hoc events. These influences can vary across hours, days and seasons. The main operational drivers are identified from the previous sections based on desk research, shadowing of a marshaller, informal interviews with AAS employees and literature. The drivers are summarised below:

1. Aircraft arrivals
   Many marshalling tasks follow directly from arriving aircraft, in particular follow me service and docking activities (Section 2.3). Aircraft arrivals show hourly peaks and variation in patterns (Section 2.6).
2. Gate changes
   Last minute gate changes can create extra tasks as marshallers may have to handle these ad-hoc adjustments (Section 2.4). These changes are managed through the coordination between Gate Planning, Apron Control and marshallers (Sections 2.2 and 2.5).
3. Ground handling availability
   If ground handlers are delayed or not present at the aircraft stand, marshallers take over the docking process to ensure safe aircraft parking (Section 2.3). These situations are unpredictable, but influence the workload.
4. Technical malfunctions
   Failures of the VDGS or other stand related issues require marshallers to dock at the stand or perform additional safety checks (Sections 2.3 and 2.4). These malfunctions are irregular, but influence task demand when they occur.
5. Maintenance activities
   Maintenance activities on airside can require extra follow me service or other safety-related tasks by marshallers (Section 2.4). Even when maintenance is planned in advance, its impact on marshaller workload can vary per hour.
6. Special and ad-hoc tasks
   Marshallers occasionally provide supervision with engine testing, and perform other tasks such as following aircraft with broken anti-collision lights, or responding to urgent requests from APC (Sections 2.3 and 2.4). These tasks can occur unexpectedly and increase the workload and depend on the coordination between APC and the marshallers (Section 2.5).
7. Calendar and seasonal effects
   Public holidays and the distinction between summer and winter season influence aircraft flight

arrivals and indirectly influence the marshalling workload (Section 2.6).

These factors describe the operational drivers influencing the marshaller workload. Not all drivers are measurable or available to use for a predictive model. The drivers used for this study will be assessed in Chapter 4.

2.8. Conclusion

In this Chapter an answer is given to the sub-question: *What is the current operational situation regarding the deployment and responsibilities of marshallers, and what are the factors influencing their daily workload and demand?*. Marshallers perform five categories of tasks: follow me service, docking, engine testing, controlling and other activities. They form a critical link between ground operations and airside operations, ensuring safe aircraft movements and a continuous operation. Analysis suggests that the demand for marshallers is primarily driven by the arrival pattern of flights, while additional influences involve operational factors such as maintenance activities, and system malfunctions.

The organisational structure of marshallers at Amsterdam Airport Schiphol depend on a close coordination between multiple actors: gate planning, apron control, Apron office, and planning department. They influence when and where marshallers are assigned. Even though the organisation is structured, the current staffing levels are static and based on experience rather than data. The internal pilot study shows that the task demand fluctuates significantly across hours, but with uncertainty across days and months. However, no quantitative system yet exists to predict these variations.

These findings indicate a clear misalignment between static staffing practices and dynamic operational demand. Demonstrating the need for a predictive model that is capable of identifying patterns in the marshallers workload and forecasting task demand. The next chapter therefore investigates which type of predictive model is most suitable to capture these variations and to provide a data-driven basis for workforce planning.

How effectively marshallers can support the airside process in this operational context depends on a number of factors. These include keeping apron activities safe, ensuring aircraft can be guided and docked on time, and using marshallers capacity effectively through the entire day. Engine idle time can be indirectly decreased by fewer waiting aircraft and smoother flows, which can lead to more sustainable ground operations. The predictive model in this study does not directly impact these factors. The model predicts when and how many marshalling activities will occur, which can serve as a useful tool for future staffing decisions. Any impact on safety, punctuality, workload or sustainability depend on the implementations of these forecasts.

3 Model type

This chapter answers the second sub-question: *Which type of predictive model is most suitable to forecast the demand?*. Section 3.1 defines the functional and non-functional requirements for the predictive model. In Section 3.2, an overview of the predictive models are given through a literature review. Section 3.3 compares the models against the requirements and identifies the most suitable model type for forecasting marshaller demand. Section 3.4 contains the conclusion to this chapter.

The analysis in Chapter 2 showed that marshaller workload fluctuates strongly across hours, days, and seasons, and is influenced by both predictable flight patterns and irregular operational events. These fluctuations do not follow a fixed or linear structure, and simple averages or static planning assumptions may not capture these patterns. A forecasting approach is therefore required that can learn from historical behaviour and provide predictions for different planning horizons.

3.1. Design requirements

Before selecting a predictive model, the design requirements must be defined. The functional and non-functional requirements provide design specifications for the model. It defines what the model must be able to achieve and under which conditions it must operate. These requirements are derived from the operational context in Chapter 2 and the forecasting problem.

3.1.1. Functional requirements

The predictive model must be able to:

1. Predict the number of marshaller tasks per hour, including a distinction between different tasks types.
2. Capture seasonal, weekly and daily variations.
3. Use historical flight arrivals data as input.
4. Use historical tasks activity per hour as input, including distinction between different tasks types.
5. Integrate exogenous factors into the model.
6. Give predictions on a short term and mid term planning.
7. Give reliable predictions based on the available nine months of historical data.
8. Preferably forecast multiple task types within a single modelling framework

3.1.2. Non-functional requirements

The predictive model must be:

1. Scalable, so it can be used on other airports or different situations without big changes.
2. Efficient in computation time, producing forecasts quickly enough to support operational planning.

3. Robust, giving stable predictions even with unexpected variations in data.
4. Accurate, producing reliable forecasts of marshaller tasks.

Although marshallers contribute to broader operational objectives, the forecasting model developed in this study does not optimise or evaluate these performance indicators directly. Its purpose is to predict hourly task demand based on operational drivers that are available. The forecasting output should therefore be seen as an input that can support future staffing or planning decisions. Any potential improvements in safety, punctuality, workload or sustainability would depend entirely on how these predictions are implemented. Based on these functional and non-functional requirements, different predictive models can be evaluated in the literature to determine which is most suitable for forecasting marshaller demand.

3.2. Overview of predictive models

Several studies illustrate how predictive models are applied to optimise workforce deployment across multiple aviation tasks. Crew scheduling and aviation maintenance are two main goals within this industry. Heuristics methods and integer programming are used to improve aircraft maintenance workforce planning, optimising both task allocation and spare parts forecasting [21]. [18] combined exact optimisation with a worst fit decreasing heuristic to distribute maintenance tasks. Their model efficiently distributes maintenance tasks across available time slots, ensuring that critical operations are completed within operational deadlines. Within crew scheduling, various models are considered. Integer programming is used to optimise workforce scheduling while taking into account labour agreements and cost efficiency [65]. A combination of exact optimisation and metaheuristics is applied to create crew pairings, reducing crew idle time and improving workforce utilisation [64].

Different predictive models are applied for the optimization of workforce scheduling. Within machine learning, Gradient Boosting Machines (GBM) refers to an algorithm commonly used. Two implementations of GBM are LightGBM and XGBoost, both showing strong performance in workforce demand planning. A data-driven approach is developed combining LightGBM, XGBoost, and Random Forest to predict delivery positions and optimise workforce planning. Using GBM models reduce forecast deviations compared to manual predictions [22]. Similarly, LightGBM was applied in a two-stage predictive model to manage employee workload in railway control rooms, outperforming other methods such as Random Forest and XGBoost across different operational scenarios. To add to the interpretability of the results, Shapley Additive exPlanations (SHAP) were used to show how specific features contributed to predictions [58]. In the healthcare sector, LightGBM was evaluated with deep learning and traditional statistical models for forecasting long-term nurse staff capacity. LightGBM proved to be a robust and efficient alternative, especially when incorporating external variables.

Exact optimisation models such as Integer Programming (IP), Linear Programming (LP), and Mixed-Integer Linear Programming (MILP) are particularly used for operational situations with constraints and regulations [31, 33, 62, 65]. Exact optimisation models are designed to optimise decisions such as shift schedules, task allocation, or staffing levels, assuming that workload or demand is already known or given. In the reviewed studies, EO models are used to generate feasible and optimal schedules under constraints such as labour agreements, capacity limits, or regulatory requirements [31, 33, 62, 65]. (Meta)Heuristics, such as Genetic Algorithms (GA) and Tabu Search (TaS), offer near-optimal solutions for complex planning problems [28, 42, 64]. (Meta)heuristic methods are mainly applied to search for scheduling or allocation solutions when exact optimisation becomes computationally expensive or infeasible. In the reviewed literature, MH approaches iteratively improve rosters, assignments or task distributions, often in combination with exact optimisation or other decision models [28, 42, 64].

Simulation models like Discrete Event Simulation (DES) are effective for evaluating how different scenarios perform under uncertain or varying conditions [19, 68, 60]. In the reviewed studies, simulation is applied to test how staffing decisions perform when demand, task durations, or system conditions vary, allowing organisations to compare scenarios [68, 19, 60].

Several studies show that hybrid approaches often perform better than a single method on its own. For example, a combination of DES with latent survival analysis is used to evaluate different workforce strategies in nursing homes, reducing labour costs while ensuring service quality [60]. A non-linear machine learning model, CatBoost, was applied to predict task time in warehouse operations, and a

GA was used to optimise batch assignments while accounting for human factors as learning and fatigue [26].

Looking at the seasonality trend of this problem there are several studies that use other models than in the workforce problems mentioned above. Research was conducted on accurate rainfall forecasting, which is tailored to the unique seasonal rainfall patterns of that region. They showed that the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model consistently outperformed alternatives in capturing seasonal trends while moderating the influence of anomalies [6]. SARIMA model with exogenous factors (SARIMAX) outperformed its competitors in terms of forecasting accuracy, overfitting, redundancy elimination, training time, and testing execution time, proving that it has remarkable performance [3]. When forecasting the daily hotel demand SARIMAX outperformed the various models in six out of seven scenarios in which Artificial Neural Network (ANN) showed more robust results in one of those scenarios [5].

The suitability of each model type strongly depends on the objective of the study. As this study focuses on the forecasting the number of marshalling tasks per hour, the different model types are evaluated based on their ability to produce workload forecasts rather than optimised schedules or scenario evaluations.

3.3. Suitability of model types

To assess which predictive model is most suitable for this study, the models reviewed in Section 3.2 are evaluated against the functional and non-functional requirements (Section 3.1). Table 3.1 provides an overview of the results. A black checkmark (✓) indicates that the model fully meets the requirement, a grey checkmark (✓) shows that the model can meet it but with limitations or additional adjustments, and a cross (✗) means the model is not able to meet the requirement.

Exact optimisation (EO) models such as MILP or IP are used for the optimisation of problems, and therefore not suitable for forecasting workforce demand. EO models are used to generate feasible and optimal schedules rather than to predict future demand. This makes EO models suitable for structured scheduling problems, but not for forecasting the number of tasks over time. As the objective of this study is to forecast marshalling tasks per hour, EO models do not match this objective and are therefore excluded.

(Meta)heuristics (MH) such as Genetic Algorithms and Tabu Search are primarily used to search for near-optimal solutions in complex planning problems. MH approaches are used to refine rosters, task assignments, or workforce distributions, often in combination with exact optimisation or other decision models. As these methods do not focus on predicting future task demand, they are not suitable as forecasting models for this study.

Simulation (Sim) models such as Discrete Event Simulation are useful to compare different scenarios with varying conditions. Simulation is applied to test how staffing decisions perform in practice, for example by analysing waiting times, or service levels under different scenarios. Because simulation models evaluate given decisions rather than generating direct forecasts of task demand, they are excluded as primary models in this study. As predicting the task per hour is the output of the model, EO, MH and Sim do not provide a suitable option for this problem and are therefore not taken into account furthermore.

In contrast to EO, MH, and simulation models, Machine Learning and Time Series models directly learn patterns from historical data and can produce explicit forecasts of task demand. Machine Learning (ML) and Time Series (TS) models are suitable for this problem, as they both can directly forecast the number of tasks per hour. SARIMA (SA) and SARIMAX (SAX) are TS models specifically designed to capture seasonality. They can handle daily, weekly, and yearly patterns effectively. LightGBM (LGBM) and XGBoost (XGB) can incorporate different data sources, including historical tasks and external factors.

A key difference in both SA and SAX is that they are univariate models, they predict only one task type at a time. This means that to forecast multiple task categories, separate models need to be build for each task. SAX has the ability to incorporate external factors, where SA is not able to. ML models can handle multiple inputs and outputs more flexible in a single framework. All four models can make short-term and mid-term predictions with nine months of data. The ML models are scalable to other airports and contexts. However, SA and SAX are less scalable as their parameters have to be re-estimated for each new context. In terms of computation time, LGBM is most efficient among the models.

Considering all requirements, LightGBM is the most suitable option. It outperforms SARIMA and SARIMAX on the ability to predict all tasks in one model, and outperforms XGBoost on the computation time. It is able to forecast tasks accurately and is scalable. making it possible to apply the model not only in the current context but also in other operational settings with minimal adjustments.

Table 3.1: Overview of requirements and model suitability

	Machine Learning		**Time Series**		**EO**	**MH**	**Sim**
Requirements	**LGBM**	**XGB**	**SA**	**SAX**	**MILP/IP**	**GA/TaS**	**DES**
Functional requirements							
Predict tasks per hour	✓	✓	✓	✓	✗	✗	✗
Seasonalities	✓	✓	✓	✓			
Historical flight arrivals	✓	✓	✓	✓			
Historical tasks per hour	✓	✓	✓	✓			
Integrate exogenous factors	✓	✓	✗	✓			
Short-/mid-term forecasting	✓	✓	✓	✓			
Work with 9 months of data	✓	✓	✓	✓			
Forecast multiple task types	✓	✓	✗	✗			
Non-functional requirements							
Scalable to other airports	✓	✓	✓	✓			
Computation time	✓	✓	✓	✓			
Robust to data variation	✓	✓	✓	✓			
Accurate predictions	✓	✓	✓	✓			

3.4. Conclusion

In this Chapter an answer is given to the sub-question: *Which type of predictive model is most suitable to forecast the demand?*. Different predictive models were evaluated against functional and non-functional requirements. Exact optimisation, metaheuristics and simulation models are not suitable because they are mainly used for scenario testing or optimisation, not for forecasting tasks per hour. Time series models such as SARIMA and SARIMAX can capture seasonality but are less flexible, as they can only predict one task at a time. Machine learning models such as LightGBM and XGBoost are able to use multiple input sources, integrate external factors and scale to other situations, LightGBM performs better on computation time. LightGBM also satisfies the requirement to forecast multiple task types within one modelling framework, whereas SARIMA and SARIMAX would require a separate model for each task category. This makes LightGBM more consistent and easier to maintain for operational use. Therefore, LightGBM is the most suitable model type for forecasting marshaller task demand.

4 Data preparation

This chapter answers the third sub-question: *What relevant data is available, and how can it be processed and prepared for use in a predictive model?*. Section 4.1 discusses the opertional drivers used in this study, including a black box model. Section 4.2 outlines the system and data requirements. Section 4.3 describes the datasets used in this study, including ADS-B data from marshaller vehicles, the airside geofence map, aircraft arrival times and engine testing records. In section 4.4 outlines how the ADS-B positions are matched to airside locations, forming the basis for understanding vehicle movements. Section 4.5 explains how these matched points are grouped into meaningful activity segments. Section 4.6 provides the logic into task labelling of each segmen, using both spatial information and external datasets. Section 4.7 describes how the labelled segments are aggregated into hourly task counts and durations, creating a consistent time series that forms the input for the forecasting model in Chapter 5. Section 4.8 presents an analysis of the aggregated dataset, showing the distribution and frequency of the four primary task categories. Finally, in section 4.9 the conclusion to this chapter is given.

4.1. Operational Drivers

Based on the seven operational drives that influence the marshaller workload as defined in Chapter 2, the relevant drivers for this study are assessed. Not every operational driver is measurable, documented or available in a structured dataset. This section explains which drivers are included in this study.

4.1.1. Included and excluded drivers

Aircraft arrivals are recorded consistently, contain precise timestamps and trigger marshalling tasks. Their patterns in hours, weeks and seasons are well structured in data (Section 2.6.2). Therefore this aircraft arrival information forms a reliable basis for forecasting demand. Calendar related events such as day of the week or public holidays are also included. These factors influence flight activity and therefore indirectly affect marshallers workload. These factors are fully known for historical and future periods, making them suitable as exogenous variables. Seasonality and calendar related events are commonly used within workforce scheduling problems [18, 34, 47, 65, 68].

The historical workload itself also carries valuable information. It does not only show the actual tasks carried out, but it also contains indirect influences from other drivers that are not recorded separately. The effect of gate changes, handler delays, and VDGS or other technical issues are captured within these tasks load (based on informal interviews and shadowing with marshallers). The consequences of these are visible in the observed workload. Therefore, the historical workload is essential information for capturing operational behaviour. In related problems across different sectors, a direct demand category is commonly used an input variable [13, 60, 64].

Even though some drivers are operationally relevant, they can not be used as explicit inputs because the data does not exist in a structured format. Information on delayed or unavailable ground handlers that require marshallers to step is not recorded to use in modelling. The same applies for technical malfunctions, such as VDGS failures or the direct effect of maintenance activities. This exact information can not be included as it is not available in a usable format (Informal interviews). These drivers may not be modelled directly, but their effects are still present in the historical task counts. The forecasting model can indirectly learn from their impact through the past workload observations.

4.1.2. Black box model

Combining these considerations leads to three main categories of information that can be used as input in the forecasting model.

1. Flight arrivals
2. Historical workload patterns
3. Exogenous factors

Based on these categories, the black box of the forecasting system (Figure 4.1) can be defined. This black box describes the inputs and outputs of the system without specifying the internal working of the model. The inputs of the model are historical data on flight arrivals, historical data on tasks per hour, including different types of tasks, and exogenous factors such as holidays. The output of the model is the predicted number of tasks per hour, with a distinction between different task type, for a planning horizon of 1 day ahead up to 3 months ahead. The internal working of the black box depends on the chosen model in Chapter 3 and will be explained in Chapter 5. The next sections describe how the datasets corresponding to these input categories are collected and processed.

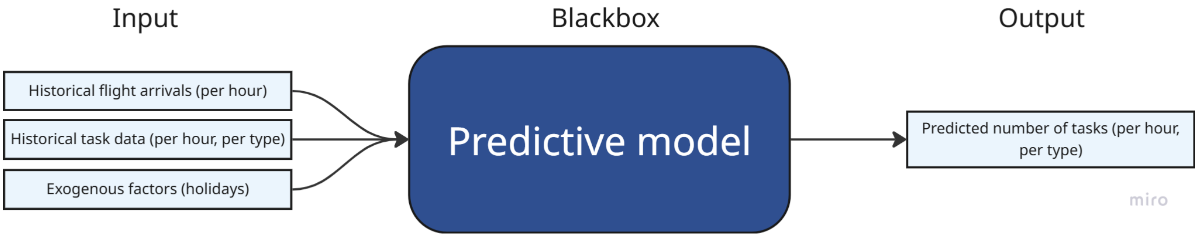

Input
Blackbox
Output
Historical flight arrivals (per hour)
Historical task data (per hour, per type)
Exogenous factors (holidays)
Predictive model
Predicted number of tasks (per hour, per type)
miro

Figure 4.1: Black box model

4.2. System and data requirements

The marshalling process is part of an operational chain that starts with the aircraft approaching the airport and ends with the aircraft leaving the stand again. Each airport has its own procedures, but the general structure of this system is similar. The aircraft arrive at the airport, taxi to their assigned stand, and are guided during the final part of this movement. Marshallers play several roles in this process as they direct the aircraft into the correct position. Other than that they supervise engine tests, and perform various other activities around the stand. Each of these operations generates a moment when the workload can be seen in the operational data.

The most significant factor is the arrival of the aircraft because that creates docking and follow me service tasks directly. Other events like gate changes or technical checks can cause work to be at irregular times. Throughout the day these triggers cause fluctuating pattern in the the marshallers' activity, where busy hours can be followed up by quieter periods. Because these patterns depend on operational conditions that change throughout the day and week, the resulting workload varies significantly over time. The variation is also influenced by the regular structures such as time of the day, day of the week, or differences between working days and holidays.

Several types of data are needed to analyse and predict this workload. To create a generic marshalling system information on vehicle movements, aircraft arrivals and calendar related events is needed. Vehicle movements are needed to see where marshallers drove on the platform. Aircraft arrivals indicate a docking or another related tasks to arrivals. Calendar information, such as hour of the day helps to explain patterns and provides structure. These types of data are available at most airports where workload is still unknown.

As for this study, the same logic applies. The data used must capture when movements happened, and which operational task might be carried out. By structuring this data according to these movements, the dataset becomes suitable to identify tasks and forecast future workload. The overview of how this data is processed into the final hourly dataset and analysis can be in in Figure 4.2. The next sections explain each data source and processing step in detail, and describe how they are combined to create the final dataset used for model development.

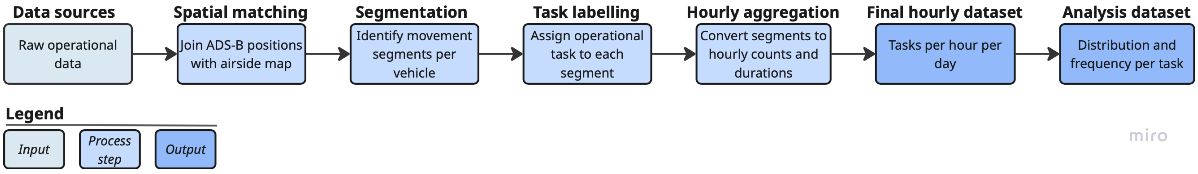

Data sources
Raw operational data
Spatial matching
Join ADS-B positions with airside map
Segmentation
Identify movement segments per vehicle
Task labelling
Assign operational task to each segment
Hourly aggregation
Convert segments to hourly counts and durations
Final hourly dataset
Tasks per hour per day
Analysis dataset
Distribution and frequency per task
Legend
Input
Process step
Output
miro

Figure 4.2: Overview data processing

4.3. Data sources

The analysis in this study relies on multiple datasets to label marshallers tasks based on their location. The datasets used are: ADS-B data, Geofence polygon map, Actual In Block Times, and Engine testing on the stand.

4.3.1. ADS-B data

ADS-B data is collected from marshaller vehicles operating on the airside. This dataset holds records for the positions, speed, and timestamps of all marshaller vehicles during their shifts. The ADS-B data serves as the primary data source to analyse vehicle movements, task execution, and interactions among marshallers. An example can be seen in Table 4.1. The callsign corresponds to the vehicle, the Casper ID is a unique number that represents the callsign on a specific day. The local timestamp is given and the coordinates, latitude smooth and longitude smooth can be seen. Lastly the speed in knots is also available. This does not directly show which task was being performed. This raw data must therefore be linked to the layout of the airport before tasks can be identified.

Callsign	Casper ID	Timestamp Local	Latitude Smooth	Longitude Smooth	Speed (knots)
CX	8888888	2025-04-26 06:17:30	52.319763	4.753529	—
CX	8888888	2025-04-26 06:17:31	52.319748	4.753559	5.05
CX	8888888	2025-04-26 06:17:32	52.319710	4.753588	9.14
CX	8888888	2025-04-26 06:17:33	52.319698	4.753649	8.07
CX	8888888	2025-04-26 06:17:34	52.319674	4.753697	8.22
CX	8888888	2025-04-26 06:17:35	52.319650	4.753750	8.83
CX	8888888	2025-04-26 06:17:36	52.319625	4.753804	8.89
CX	8888888	2025-04-26 06:17:37	52.319601	4.753861	9.02
CX	8888888	2025-04-26 06:17:38	52.319574	4.753916	9.16

Table 4.1: ADS-B data example for callsign CX (anonymised).

4.3.2. Geofence polygon map

To interpret the ADS-B data, these points must be matched with the physical airside infrastructure. A geofence polygon map is used, made by the airport operator. This data contains polygons for all important locations, providing the infrastructure network. It consists of aircraft stands, taxiways, exits, taxilanes, holdings, and bays. Each polygon has a location type, location name and coordinates. The coordinates are represented by the longitude and latitude of each corner of the polygon. An overview can be seen in Table 4.2

The polygons can be visualized using Kepler.gl, providing a clear overview, as shown in Figure 4.3.

properties_loctype	properties_locat	geometry_coordinates
VOP	M81	[4.793788791428247, 52.30449341579768]
EXIT	V1	[4.711289406706462, 52.34363785476059]
TAXIWAY	W2	[4.741190672187478, 52.32938278396526]
TAXIWAY	B-W7-z	[4.740455747427581, 52.30984630466635]
TAXIWAY	C-C2	[4.743239879786556, 52.32701250284305]
TAXIWAY	A-A5	[4.769804477854149, 52.304352368730854]
TAXIWAY	B-A8	[4.775243998060318, 52.30523800488354]
TAXIWAY	B-A8-E1	[4.775217175064292, 52.306625464910184]
TAXIWAY	A-A8	[4.774251579892509, 52.30691738268924]
EXIT	E1/W	[4.776155949688168, 52.305815297772476]

Table 4.2: Overview of airside locations with their geometry coordinates.

This map is particularly useful for illustrating the precise locations where a marshaller has operated, helping to analyse their movements and activities across the airside network. By combining this map with the ADS-B data, every point can be assigned to a specific location.

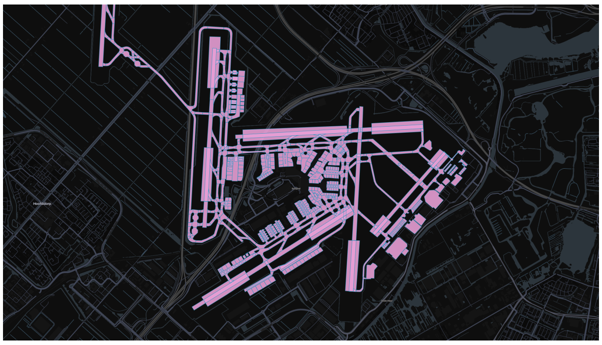


Figure 4.3: Geofence polygons of AAS visualized via Kepler.gl

Example combined trajectory

By combining both the Casper data and the geofence polygon map, the data points can be plotted on the existing polygons. Creating an overview of the trajectory where the vehicle has driven. This visualisation assists in checking the spatial matching and provides an intuitive insight into the vehicle's movement during the shift. Within Kepler.gl these datasets are combined for the example dataset of CX on April 26th. This can be seen in Figure 4.4

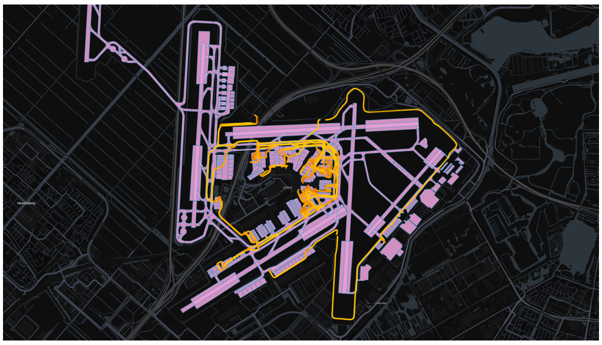


Figure 4.4: Route driven by CX on 26th of April

4.3.3. Actual in-block times

To identify docking tasks, the ADS-B trajectories must be linked aircraft arrivals. The dataset with the Actual In Block Times (AIBT) represents the moment an aircraft is positioned at the stand, and the blocks are placed. This is representative of when an aircraft is considered ’docked’. The data set shows FlightID, AIBT UTC, Date local, Time local, and RAMP@AIBT. An example of this dataset can be seen in Table 4.3. By linking the AIBT data with the vehicles presence on a polygon, the model is able to determine if a segment is related to a docking activity.

FlightID	AIBT UTC	Date local	Time local	RAMP@AIBT
HV6458	31/03/2024 22:49	01/04/2024	00:49:24	D07
HV6338	31/03/2024 22:53	01/04/2024	00:53:28	D51
HV6886	31/03/2024 23:10	01/04/2024	01:10:14	D03
HV5754	31/03/2024 23:10	01/04/2024	01:10:22	E07
HV6734	31/03/2024 23:11	01/04/2024	01:11:56	C10
HV6506	31/03/2024 23:31	01/04/2024	01:31:30	E09
HV558	31/03/2024 23:33	01/04/2024	01:33:08	E05
HV5226	31/03/2024 23:46	01/04/2024	01:46:23	E08
HV6146	01/04/2024 00:07	01/04/2024	02:07:54	D54
KL0588	01/04/2024 03:50	01/04/2024	05:50:28	E07

Table 4.3: Selected flight data showing arrival times and ramp information.

4.3.4. Engine testing dataset

Engine testing involves running aircraft engines while at the stand for maintenance checks or operational verifications. Marshallers are sometimes required to supervise these events, making this dataset relevant for task labelling. This dataset consists of the following points, date, start time, end time, aircraft stand and duration. An example can be seen in Table 4.4, showing the irregularity this task happens. By linking this dataset with the Casper vehicle data and the geofence polygons, it is possible to identify when marshallers were involved in engine testing tasks and distinguish these from other operational activities.

Date	Start time	End time	VOP	Duration (min)
01-04-2024	11:16	11:22	P16	6
02-04-2024	00:52	01:00	C07	8
02-04-2024	02:12	02:19	C07	7
03-04-2024	23:54	23:58	J87	4
04-04-2024	02:21	02:31	C07	10
04-04-2024	04:34	04:38	C07	4
05-04-2024	00:42	00:50	E04	8
05-04-2024	02:17	02:25	J84	8
05-04-2024	03:12	03:18	J84	6
05-04-2024	03:52	03:57	D56	5

Table 4.4: Example of engine testing dataset.

4.3.5. Summary and quality of data sources

The data sources used in this study are summarised in Table 4.5 and additional explanation of the quality of the dataset is given below. When combined they provide an overview of how marshallers move across the airside and which tasks they perform. The table shows the role of each dataset and how it contributes in this study. These data sources allow to interpret raw ADS-B trajectories, identify operational tasks, and prepare a structured dataset that can be used for forecasting.

The ADS-B vehicle data provides continuous GPS positions, timestamps speeds and vehicle identifiers, allowing to identify the raw movements of marshallers. This data is fully automatically generated from internal systems and provides one location point per second. In general this data is continuous, but in some cases individual seconds are missing. The exact number of missing observations can not be quantified, but inspections of fifteen selected days spread across the nine month dataset indicate that such occurrences are rare. These short data gaps are addressed in later processing steps by merging segments. Through visual inspections an area on airside is located that has reduced signal quality. In later processing steps this problem has been solved by making an exception in the task labelling. The last quality points of this data is that only marshaller vehicles have been analysed, anytime another vehicle has been used this shift will not be used for this study. Through interviews with operational staff such occurrences are considered infrequent.

The geofence polygon map is created based on the operational layout of the map. It helps to connect each GPS point to an operational location. Manual adjustments are made by enlarging the polygons and adding service roads to improve the detection of tasks. Predefined priority rules have been applied for overlapping polygons. Vehicles operating close to polygon boundaries may be assigned to a different area, which can affect task labelling.

Both AIBT and engine testing record are derived from internal regsitration systems. These datasets allow for docking and engine testing activities to be identified, and the qialuty is considered reliable. However, uncertainty mainly relates to timing. The recorded timestamps of these datasets do not exactly match the start and end time of marshaller activtities. Nevertheless, these records provide an indication of when such tasks take place and are therefore suitable for identifying task occurrence at an aggregated level.

Dataset	What it contains	How it is used in this study
ADS-B data	Continuous GPS positions, timestamps, speed, and vehicle identifiers for all marshaller vehicles.	Provides the raw movement traces of the marshallers, which form the basis for identifying where and when activity occurred.
Geofence polygon map	Spatial polygons for aircraft stands, taxiways, service roads, buildings, and other airside locations.	Links each GPS point to an operational location, allowing movements to be interpreted within the airport layout.
Actual In Block Times (AIBT)	The exact time an aircraft reached its stand and the blocks were placed.	Identifies docking activities by matching vehicle presence at the stand with the aircraft's arrival moment.
Engine testing records	Dates, times, stands, and durations of engine tests that require marshaller supervision.	Labels engine testing tasks and separates them from regular driving or standing behaviour.

Table 4.5: Overview of datasets and their role in the analysis.

4.4. Spatial matching ADS-B data

The first step in preparing the data for the forecasting system is creating a spatial match with the ADS-B data. The aim of this step is to match every ADS-B point to the correct polygon from the geofence map. The outcome of this step is a table with labelled locations, which can be used for later segmentation and task identification.

The geofence polygon map consists of all different types of polygons that represent the airside infrastructure network. It consists of different elements such as aircraft stands and taxiways. To match a point in the ADS-B data, the points are compared to the coordinate of all polygons to determine which polygon it matches. In the initial inspection it became clear that the original geofence polygon map was not complete. In order to complete this dataset buildings and service roads are manually added. This ensures movements on the service roads and around buildings are recognised correctly.

A second adjustment is necessary for the aircraft stands. Some ADS-B points at the edge of the stand polygons were not matched, even though the vehicle was clearly operating at that stand. This happens because marshallers often park their vehicle close to the aircraft, but at the outer edge of the stand polygon to ensure safety. To make sure these points are still identified as stand activity, the stand polygons were manually enlarged in the direction of the driving lane. The enlargement was done in such a way that stands do not overlap with each other, but only with nearby taxiways or service roads.

Adding these extra polygons and enlarging the stand increases the chance of overlapping polygons. When an ADS-B points fall into two polygons at the same time, it must be assigned to only one. To avoid incorrect matches, a priority rule is applied. This ensures that tasks will be able to correctly identified, otherwise it will be labelled as standing or driving (Section 4.6).

A small number of points may fall outside any polygon, leading to no match available for the specific data point. This can happen due to location inaccuracies or points on the boundary of two polygons. These points are labelled as NULL.

The output of this first step is a table for each vehicle on each day. A sequence of data points with coordinates has been transformed into locations that describe how the marshallers moves across the airside. Each row contains the callsign, a timestamp, coordinates, matched polygon location and location type. This spatial matching introduces some uncertainty at individual point level, it provides the robust structure needed to create segments (Section 4.5) and to identify tasks (Section 4.6).

4.5. Segmentation

Once every ADS-B point has been assigned to a specific location on the airside, the next step is to group these points into segments. A segment represents a continuous period in which a vehicle stays within the same operational context. Segmenting this data ensures that blocks of activity are identified which can be labelled later on (Section 4.6).

The segmentation is necessary to provide a clear view of all locations. As the frequency of the ADS-B is high, roughly one per second. Without segmentation this would give a too detailed overview and make task labelling complex.

A new segment begins whenever the vehicle moves from one polygon to another. As long as the vehicle stays in the same matched polygon and the timestamps follow each other in time, the points belong to the same segment. Because the sampling frequency is fixed, the number of points in a segment can be used as an estimate for the duration of that segment in seconds.

Segments are also split based on speed. This is done because sometimes there are large polygons, which could lead to misinterpreting results. Stationary periods should not be grouped together with driving periods, even though they happen at the same location. Therefore the split should be made based on driving or standing. A new segment is created when the location stays the same, but the speed changes from standing (<1 knot) to moving (>= 1 knot) or the other way around.

The ADS-B data points are not always stable. Sometimes a single points shortly falls outside any polygon or the spatial matching assigns a second incorrectly to a neighbouring polygon. If this would left, it would lead to a break in these activities leading to multiple segments for 1 task.

In order to avoid this two corrections are applied. If a single point is labelled as NULL but the points before and after belong to the same location and the timestamps are one second apart, the NULL point is merged into that segment. Sometimes a one second segment appears as the same location as the next segment, but with a speed of zero. If these points follow up on each other within a second, this point is merged into that second.

After the segmentation the resulting tables show per vehicle per day the callsign, the location and type, the first and last timestamp of the segment, the number of points in the segment, and the average, minimum, and maximum speeds during the segment. An example can be seen in Table 4.6.

4.6. Task labelling

After the ADS-B points have been grouped into segments, these can be assigned to tasks. The segments are combined with other datasets and constraints to create the task labelling.

The labelling is based on two main principles. If a segment matches a clearly defined event it receives that corresponding label. Otherwise, if no event matches based on the speeds of that segment the label

Seg ID	Callsign	Loc	Loc type	Start time	End time	Points
1	C1	OTG	Building	2025-04-26 06:15:00	2025-04-26 06:15:45	46
2	C1	Dienstweg	Service road	2025-04-26 06:15:46	2025-04-26 06:16:10	25
3	C1	D07	Stand	2025-04-26 06:16:30	2025-04-26 06:17:20	51

Table 4.6: Example of segments created from spatially matched ADS-B points.

driving or standing is given. The labels that can be given are as follows: Docking, Follow me service, Engine testing, Shift change periods, Havendienst, Meeting, Driving and standing. Their individual logic will be explain in the following subsections.

4.6.1. Docking

Docking happens when an aircraft arrives at a stand and the marshaller guides the final metres of the taxiing process. To identify a segment as docking, each segment is compared with the AIBT dataset. A segment is labelled as docking when the following two conditions apply:

- The location of the segment is the same as the aircraft stand the aircraft arrives, and
- The AIBT of the aircraft falls between the start and end time of that segment.

For some location the transponder doesn't submit a continuous signal. The aircraft stand associated with these location therefore have an exception. The stands G79, G76, G73 are allowed an extra five-minute window before the AIBT to ensure docking is still labelled in these events. If these conditions are met, the entire segment receives the label Docking.

4.6.2. Follow me service

The follow me service on an airport occurs when a marshaller drives in front of an aircraft to indicate the route during taxiing. A follow me service task is always followed on by docking. Therefore the docking segments are checked. If the prior segments of the docking match a follow me service path defined for that stand, the prior segments are labelled as follow me service.

4.6.3. Engine testing

Marshallers also supervise engine test to ensure safety during the session. This can happen at the engine run-up area or directly at the aircraft stand. These two cases are treated separately.

For engine testing on the aircraft stand a comparison is made with the engine testing dataset. The segment is labelled Engine testing if the start and end time of the segment matches the time window of a known engine test at that stand. However, due to the fact that the marshallers have to supervise the safety, the vehicle is usually parked around the aircraft stand. Therefore, a 10 minute margin is used. Each test session can only be used once.

For engine testing on the run-up area, a segment that corresponds tot hat polygon automatically receives the label engine testing. this is due to the fact that is location is not used for any other purposes.

4.6.4. Shift change periods

Shift changes always take place within specific hours of the day: between 06–07h, 14–15h, and 22–23h, at the building location. Three dlabels may apply: start shift, end shift, and start & end shift. This indicates whether a vehicle is being passed on, or started to use in the shift. Segments located at the building in this time period are labelled accordingly.

4.6.5. Apron office

The vehicle assigned to C1 must check in at the apron office at every start of the shift. These take place between 07-08h, 15-16h, and 23-00h. The apron office building is located near two aircraft stands, which means that ADS-B points from vehicles parked nearby are often matched to these stands instead of the building itself. A segment is labelled apron office if the following conditions apply:

- The location is apron office, or the other two aircraft stands,

- No docking takes place at that stand, and
- the time falls within one of the known check-in hours

4.6.6. Meeting

Meetings take place in specific buildings. When a segment is located at one of these buildings and does not fall in a shift-change window, it receives the label Meeting. This logic ensures that meeting behaviour is separated from shift changes and other activities at the same buildings.

4.6.7. Driving and standing

If a segment has not been assigned any label so far, it is classified using its average speed. Standing is assigned when the average speed is below 1 knot, and driving is assigned when the average speed is equal to or above 1 knot. This ensures every segment receives a task label.

4.6.8. Output task labelling

The result of the task labelling step is a table per vehicle per day in which every segment has been assigned exactly one operational task. This dataset contains the following information for each segment. The callsign, the location and location type, the start and end time, the number of points, the assigned task label, and the average, minimum and maximum speed.

4.7. Hourly aggregration

The final step in the data preparation is to convert these tables per day per vehicle into hourly workload values. The forecasting model requires a consistent time series, this means that all activity must be summarised per hour and per day.

Every segment is assigned to the hour in which it begins. This creates a table with 24 rows per day. For each hour, all segments are counted per task type, and multiple vehicles are summurised per day.

In addition to counting how many times a task occurred, the total time spent on several tasks is calculated. This makes it possible to report, per hour, how many seconds were spent per task category. These durations help to understand not only how often something happens, but also how long marshallers are occupied with different tasks.

The aggregation also records which vehicles were active. A vehicle is considered active in an hour if it has at least one segment that starts in that hour. All active callsigns are combined in a single column. This gives an overview of how many marshallers were present and contributing to the workload at any moment.

Each day results in a complete table of 24 rows, containing the task counts, durations and active vehicles. An example of this structure is shown below. The aggregated dataset forms the final output of the data preparation process. It represents a complete and structured overview of the operational workload and serves as the input for the forecasting model developed in Chapter 5.

Table 4.7: Aggregated tasks per hour for a full 24-hour day.

Hour	Docking	Driving	End Shift & Start Shift	Follow me	Havendienst	Meeting	Standing	Start Shift	End Shift	Engine Testing	Active_Cs	Total Tasks
0	4	137	0	0	0	3	6	0	0	0	2	150
1	3	70	0	0	0	0	26	0	0	0	2	99
2	0	90	0	0	0	1	91	0	0	0	2	182
3	0	68	0	0	0	1	6	0	0	0	2	75
4	0	103	0	0	0	2	1	0	0	0	2	106
5	3	56	0	2	0	0	3	0	0	0	2	64
6	1	149	2	0	0	0	10	2	0	0	4	164
7	8	223	0	0	4	0	22	0	0	0	4	259
8	10	206	0	4	0	6	13	0	0	1	4	241
9	5	169	0	4	0	2	30	0	0	0	4	210
10	2	59	0	1	0	2	11	0	0	0	4	75
11	6	168	0	2	0	2	12	0	0	0	3	193
12	7	149	0	1	0	1	16	0	0	0	3	174
13	6	128	0	0	0	0	15	0	0	0	3	149
14	5	222	2	2	0	0	9	0	1	0	3	241
15	4	127	0	2	1	0	9	0	0	0	2	154
16	1	113	0	0	0	0	24	0	0	1	2	139
17	1	123	0	0	0	0	10	0	0	0	2	134
18	2	66	0	0	0	0	2	0	0	0	2	70
19	2	33	0	0	0	0	4	0	0	0	1	39
20	2	121	0	0	0	0	3	0	0	0	1	126
21	0	21	0	0	0	0	14	0	0	0	1	35
22	0	27	0	0	0	0	5	0	1	0	1	33
23	0	0	0	0	0	0	0	0	0	0	1	0

4.8. Analysis of aggregated dataset

This section gives an analysis of the aggregated dataset as described in the previous steps. the four task categories will be discussed: docking, follow me service, meeting and engine testing. For each task category, two figures are provided. The first a box and whiskers graph with the distribution of the task duration per hour. The second is a bar plot that shows how many times the task occurred per hour over the full dataset. Both plots are based on data from events that took place between 1 december 2024 and 31 august 2025.

Each boxplot corresponds to one hour of the day (0-23) and visualises the variability within that hour. The blue box represents the interquartile range (IQR), while the horizontal line shows the median and the cross indicates the mean value. The whiskers extend to 1,5 x IQR, giving the minimum and maximum value. Individual points that fall outside that range can be seen as outliers. These descriptive results provide insights into how frequent and how variable the different tasks are.

4.8.1. Docking results

Figure 4.5 shows the distribution of docking times per hour. This is based on 5470 recorded dockings. The docking durations are relatively stable throughout the day. The average value of a docking duration is 9,80 minutes. Outliers occur throughout the entire day, but particularly high values can be observed at hour 4. The IQR at this hour was much wider than at other hours, and two outliers went up to 90 minutes. Hour 3 and 5 both show outliers at around 60 minutes. After hour 6 the distributions become more compact and the central values had smaller spreads. Lower whiskers are shorter than the upper whiskers, suggesting that a few high values stretch the distribution.

Figure 4.6 shows the number of dockings per hour over the same period of time. Between hour 1 and hour 6 the docking activity was low, but from hour 7 there was an increase in the number of dockings until hour 20. Peaks can be observed at hour 8, 9, 13, 15 and 19. From hour 20 until hour 0 the activity remains relatively steady.

The two figures together show that while the frequency of the docking activity varies throughout the day, the duration of these events remain relatively constant.

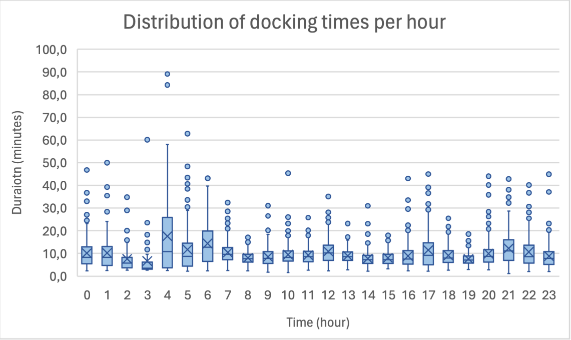


Figure 4.5: Distribution of docking times per hour

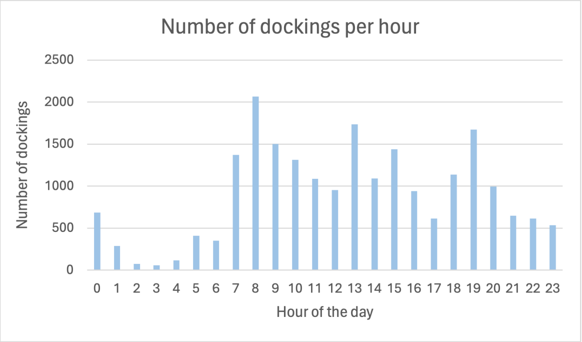


Figure 4.6: Number of dockings per hour

4.8.2. Follow me service results

Figure 4.7 shows the distribution of follow me service events per hour. This is based on 1204 recorded follow me service events. The average value of a follow me service event is 71,2 seconds. The median values remain below 90 seconds throughout the entire day, with relatively narrow IQR for most hours. Outliers occur throughout the entire day but especially between hours 9 and 18. Hours 1 and 2 show minimal variability because only one or two events were recorded in those hours. The upper whiskers are generally longer than lower whiskers, indicating that occasional longer follow me service events broaden the upper range.

Figure 4.8 shows the number of follow me service events per hour. follow me service activity is very low during the evening and night hours (20 - 6), and increases sharply from 7 with peaks at 8 and 9. A second period of increased activity occurs around hours 14 to 16.

Overall, the results indicate that the follow me service task are relatively short and show strong peak in frequency during the daytime.

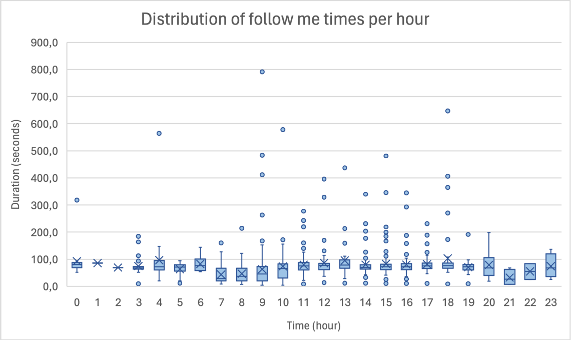


Figure 4.7: Distribution of follow me service times per hour

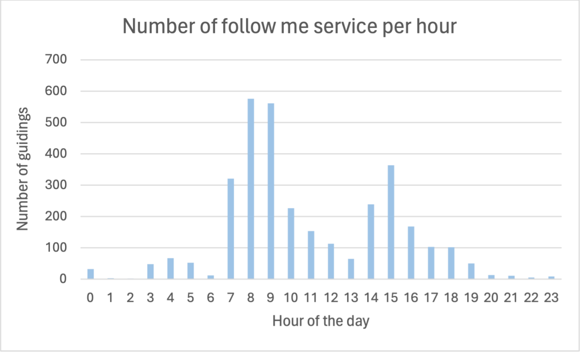


Figure 4.8: Number of follow me service events per hour

4.8.3. Meeting results

Figure 4.9 shows the distribution of meeting durations per hour, based on 6500 recorded meetings. The average meeting duration is 14,6 minutes. Most meeting durations fall between 0 and 60 minutes, with median values that are relatively stable throughout the day. However, several extreme outliers occur: durations above 200 minutes at hour 0, 8, 13, 15 and 16.

Figure 4.10 shows the number of meetings per hour. Meeting activity is more evenly distributed across the day than other task types. Activity gradually increases from hours between 7 and 12, with a peak around hour 11. Other peaks can be seen at hour 16 and 3.

Together, these figures show that meetings occur consistently throughout the day with average frequency, and while most meetings are short, a small number last significantly longer.

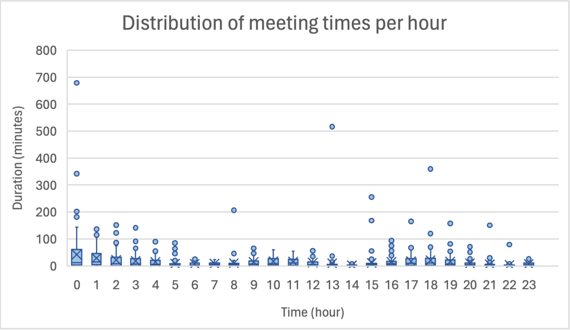


Figure 4.9: Distribution of meeting times per hour

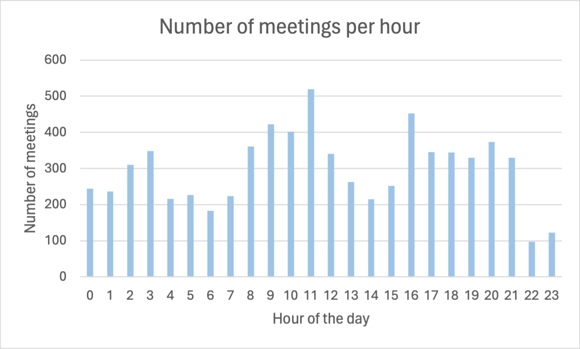


Figure 4.10: Number of meeting events per hour

4.8.4. Engine testing results

Figure 4.11 shows the distribution of engine testing durations per hour, based on 710 recorded events. The average duration is 6,63 minutes. The distributions are highly consistent across all hours, with median values between 6 and 8 minutes. Only one outlier occurs, one above 15 minutes at hour 12. The lower whisker are longer than the upper whiskers.

Figure 4.12 presents the number of engine testing events per hour. A period with high activity can be seen between hours 0 and 4, with peaks at hours 1 till 3. a peak in activity can be seen in the hours 9 till 12, and at hour 17.

These results show that engine testing is relatively rare, and mostly happens during the nighttime, it has stable durations.

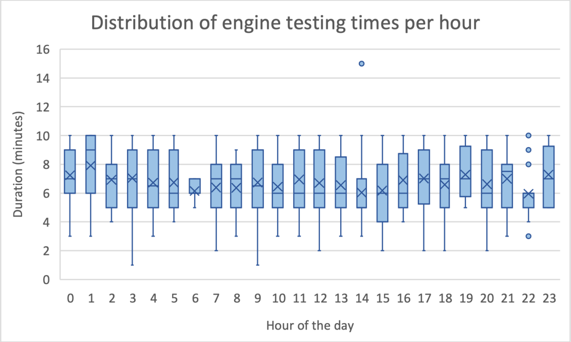


Figure 4.11: Distribution of engine testing times per hour

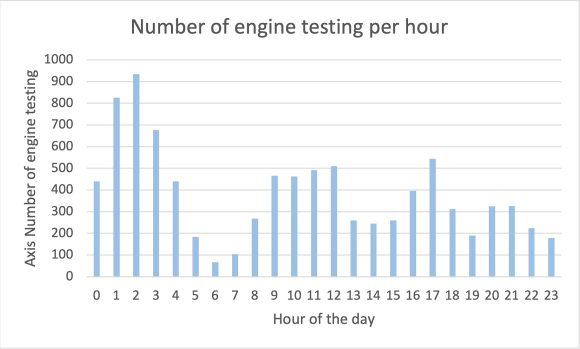


Figure 4.12: Number of engine testing events per hour

4.9. Conclusion

This chapter answers the third sub-question: *What relevant data is available, and how can it be processed and prepared for use in a predictive model?*. The analysis shows that several operational datasets are needed to describe the work of marshallers in a structured and measurable way. The ADS-B data from the vehicles provides the movement patterns, while the geofence polygons translate these movements into meaningful airside locations. The Actual In Block Times link the vehicle trajectories to aircraft arrivals, making it possible to identify docking activities, and the engine testing dataset allows supervision tasks to be recognised. Together, these data sources capture the key operational triggers that influence marshaller workload.

To prepare these datasets for forecasting, a multi step process was developed. First, each ADS-B point was matched to the correct polygon on the airside map. This required several improvements to the geofence data, including the addition of missing service roads and buildings and the extension of aircraft stand polygons. These adjustments ensured that vehicle positions could be interpreted reliably. Next, the points were grouped into segments that represent stable periods of activity. This step removed noise, corrected short interruptions and produced meaningful time blocks. These segments were then labelled using a combination of spatial location, time windows, external datasets and operational rules. Each segment was assigned one task, such as docking, follow me service, meeting, engine testing, shift change, standing or driving. Finally, these labelled segments were aggregated per hour, producing a complete and consistent time series of the workload. The dataset shows clear differences in frequency patterns and duration variability between the task types.

The resulting dataset forms a complete and structured representation of marshaller activity. It captures the moments when tasks occur, how often they happen, how long they last and how these patterns vary throughout the day. This makes the dataset suitable for use in a predictive model, which requires regular time steps and clearly defined target variables. The next chapter therefore focuses on developing such a forecasting model and explains how the processed data will be used to predict future task demand.

5 Model development

This chapter answers the sub-question: *”How can a predictive model be developed using the prepared dataset to forecast the marshallers’ task demand?”*. The focus is on translating the processed data and selected model type into a working forecasting approach. Section 5.1 introduces the model scope. Section 5.2 presents the inputs and outputs of the model. Section 5.3 explains the working of the model LightGBM. Section 5.4 explains the implementation and parameter settings, including the training and testing structure. Finally, Section 5.5 discusses the applied evaluation metrics and explainability methods.

5.1. Model scope

The predictive model aims to forecast the hourly task demand of marshallers. The prediction is how many marshaller tasks are expected to occur for different time horizons. Three prediction horizons are implemented and each one is implemented in its own LightGBM model. The three models correspond to the following prediction hours: 24 hours, 168 hours and 2160 hours into the future, representing one day, one week and approximately three months. The prediction targets are the hourly counts of four main task categories:

1. Docking
2. Follow me service
3. Meeting
4. Engine testing

These tasks represent the core operational activities of marshallers at Amsterdam Airport Schiphol identified in earlier chapters. However, the approach is general and can be applied to any other type of task that marshallers perform, as long as historical data for these tasks are available. This makes the framework suitable for different airport contexts or for expanding the model with new task categories in the future.

The model uses a set of operational inputs that are available before the prediction time. This includes calendar variables, flight arrivals and time based features that describe recent task activity.

5.2. Model inputs and outputs

Feature engineering forms an essential step in developing the predictive model, as it determines which operational factors are translated into model inputs. The selected features describe both temporal patterns and operational drivers that influence the workload of marshallers.

Several calendar based variables are included to capture daily, weekly, and seasonal patterns. These consist of hour of the day, day of the week, month, weekend indicator, summer/winter indicator, and a holiday indicator. The number of incoming flights is also included. In addition to the hourly flight count, a 24 hour and 168 hour rolling mean is added. This feature helps the model identify broader fluctuations in the arrival pattern that do not depend only on the most recent hour. In addition, two lag features are included. The first reflects the number of flights at the same hour on the day (lag 24), and the second reflects the value one week earlier (lag 168). Together, they provide information about daily and weekly trends in the arrival pattern.

In addition to these variables, several temporal features were created to include recent workload information. This is done by using lagged and rolling mean features. Lagged features represent the value of a variable in the previous time step. For example, *Docking_lag1* stores the number of docking tasks one hour earlier, allowing the model to learn short-term temporal dependencies. Rolling features represent the average value value over a given time window. For example, *Docking_roll3* stores the average number of docking tasks over the last three hours. These lag windows cover 1, 3, 6, 12, 24 and 168 hours, which represent hourly, half-day, daily and weekly effects. Rolling averages over the same windows are included to smooth fluctuations and reflect the recent workload trend.

Furthermore, trend-based features compare each task category with its value one day earlier (24 hours) and one week earlier (168 hours). These differences help the model recognise upward or downward movements in task demand. These temporal features are especially important for capturing regular operational rhythms and sudden changes over time.

An overview of all used model features can be seen in Table 5.1. These features provide a representation of both temporal and operational effects on marshaller task demand. They allow the models to learn not only from structural patterns such as time of day or weekday, but also from short-term variations in workload, while remaining applicable for real forecasting situations.

The outputs of the model are the expected hourly counts of the four marshaller tasks at the chosen prediction horizon. For each horizon, the model returns one set of continuous values for docking, follow me service, meeting and engine testing. This setup allows the model to capture shared patterns, since all four predictions are based on the same inputs.

Table 5.1: Complete overview of features used in the predictive model

Feature	Type	Description
Calendar and seasonal features		
hour	Numerical	Hour of the day (0 – 23)
day_of_week	Numerical	Day of the week (0=Mon, 6=Sun)
is_weekend	Binary	1 if Saturday or Sunday
month	Numerical	Month of the year (1 – 12)
is_summer	Binary	April – October = 1
is_holiday	Binary	Public holidays + school vacations
Flight-related features		
incoming_flights	Numerical	Number of flight arrivals in that hour
incoming_roll24	Numerical	24-hour rolling mean of arrivals
incoming_roll168	Numerical	168-hour (weekly) rolling mean of arrivals
incoming_lag24	Numerical	Value 24 hours earlier
incoming_lag168	Numerical	Value 168 hours earlier
Lag features (per task: Docking, follow me service, Meeting, Etesting)		
task_lag1	Numerical	Value 1 hour earlier
task_lag3	Numerical	Value 3 hours earlier
task_lag6	Numerical	Value 6 hours earlier
task_lag12	Numerical	Value 12 hours earlier
task_lag24	Numerical	Value 24 hours earlier
task_lag168	Numerical	Value 168 hours earlier (one week)
Rolling mean features (per task)		
task_roll1	Numerical	Rolling mean over 1 hour
task_roll3	Numerical	Rolling mean over 3 hours
task_roll6	Numerical	Rolling mean over 6 hours
task_roll12	Numerical	Rolling mean over 12 hours
task_roll24	Numerical	Rolling mean over 24 hours
task_roll168	Numerical	Rolling mean over 168 hours
Trend features (per task)		
task_diff_day	Numerical	Difference with value 24 hours earlier
task_diff_week	Numerical	Difference with value 168 hours earlier

5.3. Working of LightGBM

Following the comparison in Chapter 3, LightGBM is selected as the forecasting approach because it efficiently trains gradient-boosted decision trees and can utilise multiple structured variables as inputs. This enables the model to capture non-linear effects and interactions that are relevant for forecasting hourly task demand. LightGBM was created to increase predictive efficiency, manage large datasets, and shorten training times, and generally used when working with large tabular datasets [11].

LightGBM is a type of gradient boosting decision tree (GBDT), also called gradient boosting machine (GBM), algorithms, introduced by ke et al (2017) [29]. Boosting is an ensemble learning method, where multiple weak learners, decision trees, are combined to create a strong learner [10, 15]. By merging these decision trees, complex relationships between features can be efficiently recorded [15]. Mathematically this can be written as:

$$f(x) = \sum_{m=1}^{M} f_m(x)$$

where $f(x)$ is the final model (the sum of all weak learners), M is the total number of trees, and $f_m(x)$ represents the individual m-th decision tree. Each new tree is trained to correct the residual errors made by the previous model.

Given a loss function $L(y, f(x))$, the goal of gradient boost is to approximate a function that minimizes the total loss, which can be represented as:

$$L(y, f(x)) = \sum_{i=1}^{N} L(y_i, f(x_i))$$

where y_i are the true observed values, $f(x_i)$ are the predicted values from the model, and N is the total number of training samples.

The choice for the loss function depends on the type of problem. In this research the model aims to predict target values, the number of marshallers tasks. Therefore, the Mean Squared Error (MSE) is used as the loss function. It is the average squared difference between the estimated value and the true value. MSE as a loss function is defined as:

$$L(y, f(x)) = \frac{1}{N} \sum_{i=1}^{N} \big(y_i - f(x_i)\big)^2$$

The loss function is optimised using gradient descent:

$$F_m(x) = F_{m-1}(x) + \rho_m f_m(x)$$

where $F_m(x)$ is the combined model after m iterations, ρ_m is the weight of the newly fitted tree $f_m(x)$. Therefore, each new tree corrects the residuals of the previous model. This process continues until the maximum number of trees is reached.

A key difference of LightGBM compared to other boosting algorithms is the way trees grow. Most algorithms use a level-wise growth, where LightGBM uses a leaf-wise growth (Figure 5.1). Level-wise growth is where all leaves at a given depth are expanded simultaneously. However, in the leaf-wise growth, the algorithm finds the leaf with the largest reduction in loss and splits it. Resulting in trees that are effective at minimizing loss and leading to increased accuracy [11]. However, leaf-wise growth can also lead to overfitting if the process is not properly regularised. LightGBM has parameters to prevent this such as the maximum depth of the trees or how many instances are required to split a node [9].

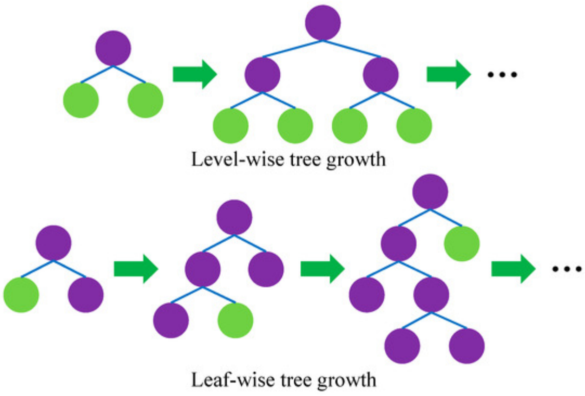


Figure 5.1: Level-wise vs leaf-wise tree growth [32]

The next section explains how the LightGBM model is implemented, including parameters setting, hyperparameter optimisation and training procedures.

5.4. LightGBM implementation and parameter setting

The predictive models are implemented in Python, using a combination of open-source libraries. The lightGBM package provides the gradient boosting tree algorithm and the Scikit-learn library allows the use of multi-output regressor. The Optuna framework is used for hyperparameter optimisation, as this allows a flexible and efficient way to automate the search for optimal hyperparameters. This reduces the need for trial and error and helps to improve the models generalisation.

The forecasting system is trained for multiple prediction horizons. Three horizons are included:

- 24 hours ahead (1 day)
- 168 hours ahead (1 week)
- 2160 hours ahead (3 months)

For each horizon, new target columns are created by shifting the task counts forward in time. Missing values at the end of the time series are removed so that every training sample has valid future targets. Each horizon has its own model. Within a single horizon, four task types (Docking, follow me service, Meeting, Engine testing) are predicted simultaneously using a MultiOutputRegressor. This wrapper trains one LightGBM regressor per task type but keeps them grouped under one model object.

The LightGBM models use regression as objective. Several parameters influence how the trees grow, such as the number of boosting iterations, the depth of the trees and the minimum number of samples required to create a split. These parameters are not fixed in advance. Instead, they are selected through hyperparameter optimisation with Optuna. The optimisation runs within predefined ranges and searches for the combination of parameters that minimises the loss for each horizon. Table 5.2 shows the search ranges and the final selected values. These values form the basis of the final model for each prediction horizon.

The number of estimators controls how many trees are added during boosting. The learning rate sets the step size in updating the model after each tree. The number of leaves and the maximum depth determine how complex each tree may become. The parameter for minimum child samples sets the minimum number of observations required to create a split. The subsample and colsample_bytree settings influence how many rows and features are used when constructing trees.

The models are trained in a deterministic to ensure consistent behaviour. LightGBM has a deterministic training mode, but that setting alone is not enough to provide a full deterministic training. Random variation can still happen from row sampling, feature sampling, bagging and the initialisation of the optimisation procedure. These effects are removed by fixing the random seed, disabling sampling by setting the subsample and colsample_bytree parameters to one and by turning off bagging. The Optuna sampler also uses a fixed seed. Resulting in the fact that the same data always leads to the same model.

Table 5.2: Hyperparameter search ranges and final configurations per prediction horizon

Parameter	Search range	t+24	t+168	t+2160
n_estimators	200 - 800	341	645	433
learning_rate	0.01 - 0.20	0.01136	0.01338	0.01023
num_leaves	20 - 120	52	20	32
max_depth	3 - 15	5	4	13
min_child_samples	5 - 60	40	9	8
subsample	fixed at 1.0	1.0	1.0	1.0
colsample_bytree	fixed at 1.0	1.0	1.0	1.0

5.5. Evaluation and explainability

The performance of the predictive models is evaluated using three metrics: Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), Mean error (ME), and the coefficient of determination (R^2). The results are given in Chapter 6, but this section explains how the metrics are defined and what they measure.

The MAE measures the average size of the prediction error. It expresses how many tasks the model is off on average, with lower values indicating better accuracy. It can be expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Where y_i represents the observed values, y_i the predicted value, and N is the total number of samples.

The RMSE measures the average magnitude of the prediction error and is expressed in the same unit as the target variable. The lower the RMSE the higher the predictive accuracy is. It can be expressed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Where y_i represents the observed values, y_i the predicted value, and N is the total number of samples.

The Mean Error (ME) measures the average systematic difference between the observed value and the predicted value. It shows if the model systematically overestimates or underestimates the task demand. A positive ME indicates overestimation and a negative ME indicates underestimation. It can be expressed as:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$$

where y_i represents the observed values, $\hat{y}_i$ the predicted value, and N is the total number of samples.

The R^2 shows to which extent the model explains the variation in the observed data. It has a value between 0 and 1, with values closer to 1 indicating greater significance. It can be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

In this study, MAE, RMSE, ME and R^2 are calculated for each of the four marshaller task categories separately to identify which task types are most predictable.

Apart from measuring accuracy, it is also important to understand the reasoning behind the models' predictions. Explainability methods provide insight into which input features drive the outputs, and how strongly they contribute to each prediction. A widely used method is SHapley Additive exPlanations (SHAP). SHAP can highlight whether variables such as the number of incoming flights, the time of day, or the day of the week are the main drivers of predicted marshaller task demand.

5.6. Conclusion

This chapter answers the subquestion: *"How can a predictive model be developed using the prepared dataset to forecast the marshallers' task demand?"* The scope of the model is presented with the four marshallers task categories and three prediction horizons. The model inputs are constructed from calendar information, flight arrivals and time based features. The outputs consist of hourly task counts for each horizon. The working principle of gradient boosting and leaf wise tree growth is explained. The implementation is set to deterministic training so that the same data always leads to the same model. Hyperparameters are selected with Optuna for each prediction horizon. The evaluation metrics and explainability methods used to analyse the model are defined in this chapter, and the results of these evaluations are presented in the next chapter.

6 Model results and validation

This chapter answers the sub-question *How can the developed predictive model be validated to ensure accurate and reliable forecasts of marshaller demand?*. Section 6.1 presents the experimental plan. Section 6.2 provides the model's performance across the three horizons. Section 6.3 compares the model against two other baseline models. Section 6.4 reports a feature selection experiment and section 6.5 provides the conclusion of this chapter.

6.1. Experimental plan

The goal of the validation is to test how well the model can predict the hourly task demand for three planning horizons, and to see what the influence of the chosen features is on the performance of the model. The total dataset covers nine months of data, of which an analysis can be seen in Section 4.8. An 80/20 train-split is applied, which means the first 80% of the data is used for training, and the most recent 20% is used for testing. As the data is time relevant, the split is chronological. This ensures that future data is not used to predict past events, creating a realistic operational forecasting set-up. However, using the last part of the dataset to test, means that the evaluation is focused on one specific time window.

This is appropriate for assessing how well the model performs when predicting the most recent period, but it also makes the validation narrow, as earlier parts of the dataset are not used for testing. This limitation is considered in the interpretation of the results.

The models are evaluated each on their specific prediction horizon:

1. 24 hours ahead (1 day)
2. 168 hours ahead (1 week)
3. 2160 hours ahead (3 months)

For each horizon the corresponding model predicts four tasks: Docking, Follow me service, Meeting and Engine testing. The performance of the models is evaluated using four metrics: the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), The Mean Error (ME), and the coefficient of determination (R^2). The MAE provides the average absolute deviation between predicted and observed task counts, while the ME indicates if the model systemically underestimates or overestimates the task demand. The RMSE shows the average error prediction in task per hour, and R^2 indicates how much of the variation in the task counts is explained by the model. Together, these metrics indicate both the reliability of the predictions and the amount of structure the model captures. The performances of the models and the interpretation are given in Section 6.2.

As there currently is no baseline model to compare the performance of the LightGBM models with, the results are first presented on their own. However, to see the value of the LightGBM model two baseline model have been constructed. A seasonal naïve model and a weekly hourly average model have been chosen as they reflect simple and realistic forecasting approaches that an organisation could implement in practice. Both models capture basic recurring patterns in the data, without relying on complex assumptions, making them good reference models. The results can be seen in Section 6.3.

In addition to the main model, a reduced-feature model is tested (Section 6.4. This second version uses only the most important features that are identified through a SHAP analysis. By comparing the main model to the feature selection model, it can be evaluated whether all features are necessary, or whether the model performs similarly with a smaller and more interpretable feature set.

6.2. Model performance across prediction horizons

This section presents the performance results of the predictive models for each of the three forecasting horizons: 24, 168, and 2160 hours ahead. For each horizon, a model generates four predictions: Docking, Follow me service, Meeting and Engine testing. Performance is evaluated using the MAE, RMSE, ME and R^2.

6.2.1. Performance for the 24-hour horizon (t+24)

The 24-hour horizon represents short-term forecasting. Table 6.1 shows the results for each of the four task types. Docking has the highest performance with, with an RMSE of 1.746 and R^2 of 0.428. Meaning the model captures a clear part of the daily pattern. The MAE and ME are the highest for docking with values of 1.360 and 0.464 respectively. Showing that the model is off by about one to one and a half task, and it mildly overestimates on average. The other tasks have R^2 values close to zero, indicating that the model can not explain much variation for these categories. The ME values are close to zero, for Follow me service and Meeting the model slightly overestimates and for Engine testing the model slightly underestimates.

Table 6.1: Model performance for the 24-hour prediction horizon

Task	MAE	RMSE	ME	R^2
Docking	1.360	1.746	0.464	0.428
Follow me service	0.315	0.668	0.124	0.078
Meeting	0.904	2.002	0.054	0.007
Etesting	0.181	0.344	-0.007	-0.007

6.2.2. Performance for the 168-hour horizon (t+168)

The 168-hour horizon corresponds to weekly forecasting. Table 6.2 gives the results for each of the four task types. Docking performs best with an RMSE of 1.702 and an R^2 of 0.452, suggesting that the weekly structure is captured well in the data. The MAE and ME are 1.309 and 0.328, again showing that the model is off by about one to one and a half task, and it mildly overestimates on average. Follow me service, Meeting, and Engine testing show limited predictive performance, with R^2 values close to zero. The MAE for these tasks are lower compared to docking. For Follow me service and Meeting the model slightly overestimates, and for Engine testing the model slightly underestimates.

Table 6.2: Model performance for the 168-hour prediction horizon

Task	MAE	RMSE	ME	R^2
Docking	1.309	1.702	0.328	0.452
Follow me service	0.327	0.667	0.123	0.076
Meeting	0.892	2.028	0.015	-0.001
Etesting	0.184	0.341	-0.002	0.009

6.2.3. Performance for the 2160-hour horizon (t+2160)

The 2160-hour horizon represents mid-term forecasting. Table 6.3 presents the corresponding performance metrics. Docking still performs best with an RMSE of 1.662 and R^2 of 0.437. The MAE and ME for Docking are 1.315 and 0.565, showing that the model is again off by about one to one and a half task and that it overestimates on average. Follow me service shows a small increase in the R^2 compared to other horizons, leading to a value of 0.120. Meeting and Engine testing show negative R^2 values, meaning that a simple average would perform slightly better. The MAE values for these three tasks remain lower than for Docking, and the ME values are close to zero, slightly underestimating.

Table 6.3: Model performance for the 2160-hour prediction horizon

Task	MAE	RMSE	ME	R^2
Docking	1.315	1.662	0.565	0.437
Follow me service	0.245	0.687	-0.026	0.120
Meeting	0.876	2.272	-0.078	-0.003
Etesting	0.185	0.354	-0.008	-0.009

6.2.4. Interpretation of forecasting performance

Figures 6.1, 6.2, 6.3, and 6.4 show the overall performance of the model across the three prediction horizons.

Docking is the only task type that has a consistent and meaningful performance. Its R^2 values range from 0.43 to 0.45 across the horizons. This means the model captures a substantial part of the variation in docking demand. The RMSE ranges between 1.66 and 1.75, and the MAE remains stable between 1.30 and 1.36 tasks per hour, showing that the prediction error is relatively similar on all horizons. The ME values are positive for the three horizons, indicating that the model overestimates the docking demand. However, the magnitude of this bias is limited by 0.57 on the longest horizon. Together these metrics show that the model is able to learn the Docking structure and provide meaningful predictions.

Follow me service shows a much lower predictability than Docking. The RMSE is relatively low (0.67 - 0.69) and so is the MAE (0.25 - 0.33). The R^2 values remain close to zero, with an outlier above the 0.12 on the longest horizon. This indicates that the model captures little of the underlying variation. The ME is positive on the hour and week horizon and negative for the 3 month horizon, meaning the model first overestimates and lastly underestimates the demand. The low RMSE, MAE and ME values for follow me service are mainly a result of the low number of tasks per hour. As the hourly counts are often close to zero or zero, the possible error magnitude is limited. The low overall volume of follow me service tasks contribute to this, as the hourly tasks stay small. This leaves little structure for the model to learn resulting in a low R^2 value. The metrics suggest that the low scale of follow me service data in the dataset and the limited structure in the hourly patterns make it difficult to predict this task accurately.

Meeting is the most irregular task type. It has the highest RMSE values of all task types (2.01 - 2.27) and relatively high MAE values (0.88 - 0.90). The R^2 values are close to zero and becomes negative at the longest horizon. A negative R^2 means that a simple mean value would outperform the model. The ME values are positive and close to zero on the first two horizons, and negative on the longest horizon. The high error mistakes and close zero or negative R^2 values suggest that Meeting tasks occur in highly variable patterns and difficult for the model to generalise.

Engine testing has the smallest RMSE values of all task types, and the MAE values stay between 0.18 and 0.19. However, similar to follow me service, this has low task volumes. These low values are influenced by the small magnitude of the hourly task counts. The R^2 values are close to zero or negative across the horizons, indicating that little of the variation is explained by the model. The ME values are close to zero. These metrics suggest that Engine testing is difficult to forecast with the current data characteristics.

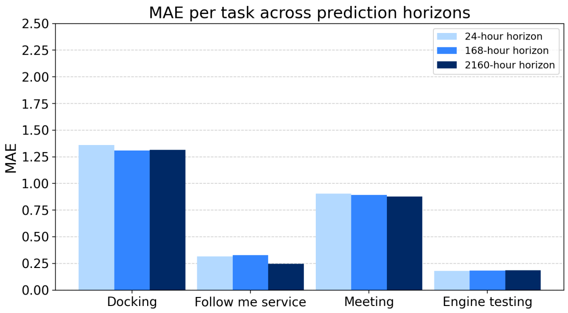


Figure 6.1: MAE values per task across prediction horizons

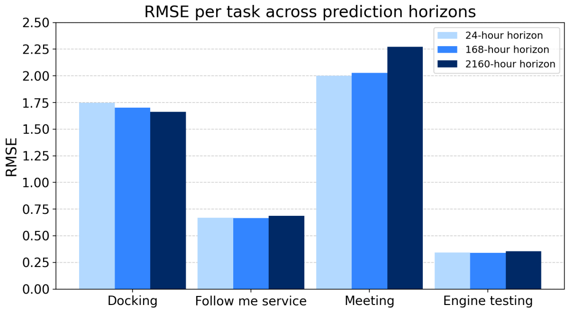


Figure 6.2: RMSE values per task across prediction horizons

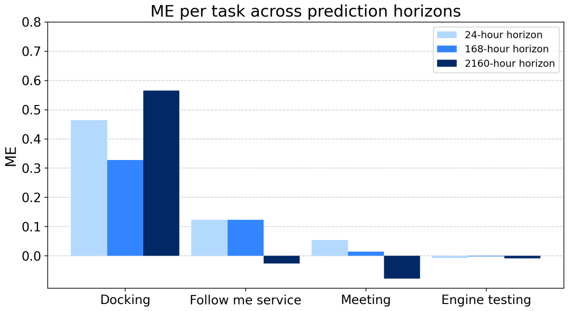


Figure 6.3: ME values per task across prediction horizons

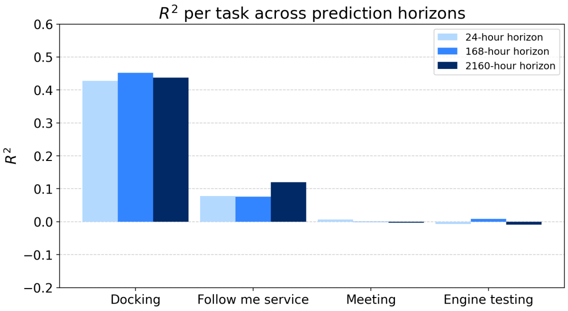


Figure 6.4: R^2 values per task across prediction horizons

6.2.5. Interpretation of mean error and tasks

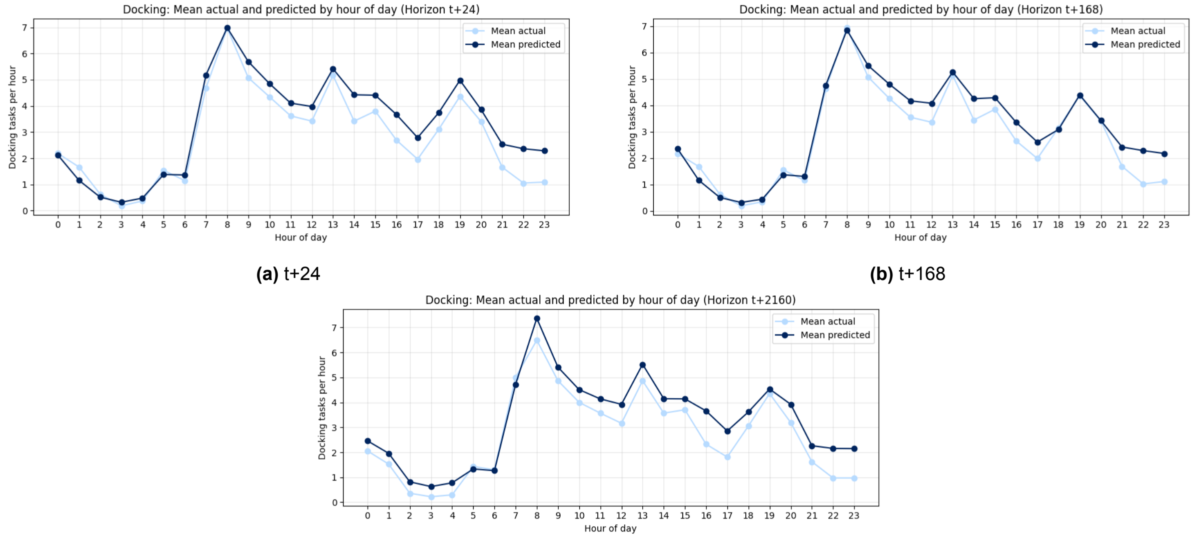
The Mean error indicates whether the model systemically over- or underestimates, however to provide more context this should be compared to the task volumes. The average observed and predicted tasks are reported for each task type and horizon. Furthermore, the mean error is given, but also expressed as a percentage of the mean observed tasks value. Lastly, a bias per day is given to show the influence on a full day. The results are given in Table 6.4.

Table 6.4: Average error analysis for different prediction horizons

Horizon	Task	Mean actual	Mean predicted	ME	ME (%)	Bias/day
24	Docking	2.818	3.282	0.464	16.4	11.1
24	Follow me service	0.140	0.263	0.124	88.5	3.0
24	Meeting	0.686	0.740	0.054	7.9	1.3
24	Etesting	0.104	0.097	−0.007	−6.4	−0.2
168	Docking	2.803	3.132	0.328	11.7	7.9
168	Follow me service	0.138	0.261	0.123	89.1	3.0
168	Meeting	0.686	0.701	0.015	2.2	0.4
168	Etesting	0.105	0.102	−0.002	−2.3	−0.0
2160	Docking	2.704	3.269	0.565	20.9	13.6
2160	Follow me service	0.153	0.127	−0.026	−17.1	−0.6
2160	Meeting	0.699	0.621	−0.078	−11.1	−1.9
2160	Etesting	0.106	0.098	−0.008	−7.7	−0.2

Docking shows a higher average task volume than the other tasks, with values between 2.70 and 2.82 task per hour across all horizons. Taking into account, as found in Section 4.8, that the average Docking time is 9.80 minutes, this leads to about 27 minutes of docking tasks per hour on average. The ME is between 0.33 and 0.56 which corresponds to a relative bias of approximately 12% to 21%. This leads an overestimation of around 8 to 14 tasks per day. Translating to a roughly additional 78 - 133 minutes of Docking tasks per day. On the other hand, Follow me service, Meeting and Engine testing occur at much lower hourly rates. Low ME values for follow me service lead to high bias percentage. This absolute influence however is still small as the average follow me service takes 71.2 seconds.

To interpret the Docking results Figures 6.5a, 6.5b and 6.5c are plotted showing the average predicted and actual values per prediction horizon given for every hour of the day. The predicted and observed values follow a similar pattern across all three horizons. For most hours of the day the model slightly overestimates. The underestimation occurs mainly during the early morning hours: 1, 2, 5, 6, and 7.



(a) t+24

(b) t+168

(c) t+2160

Figure 6.5: Mean observed and predicted Docking volume per hour of day for the three forecasting horizons.

Equivalent hour of the day figures for the other three tasks are given in Appendix D. Follow me service is the only task that shows a somewhat similar pattern for the actual and predicted results. However, this is mainly driven by the peak around the hours 7 till 10. Outside this peak the tasks volume is very low. For Meeting and Engine testing no clear similar patterns can be identified across all horizons.

6.3. Comparison with baseline

To evaluate whether the LightGBM model adds predictive value a comparison is made with two other baseline models. The first is a seasonal naive baseline, which predicts the same value as exactly one week earlier. This method reflects the weekly cycle in the data. The second baseline is a weekly hourly average. For each hour of the week, it uses the average value observed during the training period. This approach smooths fluctuations by aggregating several weeks. Each baseline uses one forecasting rule that is applied across all horizons. The rule itself remains unchanged, but each horizon evaluates the forecasts for a different set of future time steps.

In Table 6.5 a summary of the performance of both baselines and the LightGBM model for Docking can be seen. Docking is the only task type that shows meaningful predictive structure, and therefore is only chosen. The results show that LightGBM consistently outperforms both baseline methods across all horizons. The RMSE and MAE values for the seasonal naive baseline is almost the same for all three horizons. Compared to this baseline, the RMSE decreases from approximately 2.16 to between 1.66 and 1.75, and the MAE decreases from around 1.55 - 1.60 to about 1.31 - 1.36. The R^2 values are much higher for the Lightgbm on all three horizons. The weekly average baseline performs better than the seasonal naive baseline, but the LightGBM model still improves the RMSE and achieves higher R^2 values across all horizons. Seasonal naive model performs best with regard to the ME, with negative values close to zero. Weekly hourly average model has negative values for the ME higher than the LightGBM model.

For the other task types (Follow me service, Meeting and Engine testing), the differences between the model and the baselines are much smaller. All methods achieve low or negative R^2 values, and the baseline errors are close to the model errors. This confirms that these tasks contain little predictable structure, and that adding model complexity does not improve forecasting accuracy. The summarised table for these results per task type can be seen in Appendix E.

Overall, the baseline comparison shows that the LightGBM model adds predictive value for Docking, while for the other tasks the baseline methods already perform at a similar level. This supports the conclusion that meaningful forecasting is only feasible for Docking. The other tasks are too unpredictable and need to be handled via other methods.

Table 6.5: Performance comparison for Docking across baseline methods and the LightGBM model

Horizon	Model	MAE	RMSE	ME	R^2
24h	Seasonal naive	1.598	2.162	-0.084	0.103
	Weekly avg.	1.395	1.811	-0.607	0.384
	LightGBM	1.360	1.746	0.464	0.428
168h	Seasonal naive	1.592	2.162	-0.088	0.094
	Weekly avg.	1.397	1.808	-0.632	0.381
	LightGBM	1.309	1.702	0.328	0.452
2160h	Seasonal naive	1.546	2.102	-0.147	0.045
	Weekly avg.	1.432	1.830	-0.871	0.317
	LightGBM	1.315	1.662	0.565	0.437

6.4. Feature selection model

Feature selection modelling evaluates whether the full feature set used in the LightGBM model is necessary, or whether a smaller and more interpretable subset of features can achieve comparable forecasting performance. The analysis is a combination of SHAP feature importance, cumulative contribution patterns, and reduced-model experiments. The goal is to identify the essential influences of the Docking predictions and to validate whether these drivers are sufficient for forecasting. As earlier results showed that Docking is the task with predictable structure this is the only task chosen in this experiment.

For each prediction horizon, SHAP values were computed to determine how much each feature contributes to the forecast. SHAP provides both the magnitude and direction of influence, which makes it possible to identify the features the model relies on. The most influential features across all horizons are: Hours, Docking_lag24, Docking_lag168, Docking_roll1, Docking_roll24, seasonal and weekly indicators, and for the 2160-hour horizon, flight-pattern indicators such as incoming_lag168.

To select the most relevant features, the SHAP importances were sorted and cumulatively summed for each horizon. An overview of the 15 most influential per horizon can be seen in Appendix F. As the goal of this experiment is to see what the impact is of a reduced featured model on the performance, the six most influential feature are used to construct a reduced set per horizon. The overview of the selected features can be seen in Table 6.6. These features are used in a reduced-feature model. For the 24-hour horizon, the six highest-ranking features together explain 75.01% of the cumulative SHAP impact. These include Hour, Docking_lag168, Docking_lag24, Docking_roll1, Docking_roll24, and day_of_week. For the 168-hour horizon, the six most influential features account for 77.71% of the SHAP importance and consist of Hour, Docking_roll1, Docking_lag24, Docking_lag168, Followme_roll6, and Followme_roll168. For the 2160-hour horizon, the first six features capture 68.11% of the cumulative importance and include Hour, incoming_lag168, month, Docking_lag24, Docking_roll1, and Docking_lag168.

Table 6.6: Top six features selected per prediction horizon and their cumulative SHAP contribution

Rank	t+24	t+168	t+2160
	75.01% SHAP contribution	*77.71% SHAP contribution*	*68.11% SHAP contribution*
1	Hour	Hour	Hour
2	Docking_lag168	Docking_roll1	Month
3	Docking_lag24	Docking_lag24	Docking_lag168
4	Docking_roll1	Docking_lag168	Incoming_lag168
5	Docking_roll24	Followme_roll6	Docking_lag24
6	Day_of_week	Followme_roll168	Docking_roll1

The reduced-feature models are evaluated against the full model. The results show that the reduced model performs nearly identically to the full model (Table 6.7). On the 24-hour horizon the reduced model achieved an MAE of 1.360, RMSE of 1.742 and an R^2 of 0.430. Which is a difference of 0.006, 0.004 and 0.002 respectively indicating they are nearly the same. For the 168-hour horizon The MAE went down with 0.045 to a value of 1.264, RMSE went up with 0.020 to a value of 1.722 and the R^2 went down 0.014 to a value of 0.438. The same goes for the longest horizon, where the reduced model again achieves a lower MAE and RMSE, and a higher R^2.

Table 6.7: Comparison between full and reduced model performance for Docking across metrics

Horizon	MAE			RMSE			ME			R^2		
	Full	Red.	Δ	Full	Red.	Δ	Full	Red.	Δ	Full	Red.	Δ
t+24	1.360	1.354	-0.006	1.746	1.742	-0.004	0.464	0.444	-0.020	0.428	0.430	+0.002
t+168	1.309	1.264	-0.045	1.702	1.671	-0.031	0.204	0.184	-0.020	0.452	0.471	+0.019
t+2160	1.315	1.265	-0.050	1.662	1.601	-0.061	0.556	0.516	-0.040	0.437	0.477	+0.040

6.5. Conclusion

This chapter answers the sub-question: *How can the developed predictive model be validated to ensure accurate and reliable forecasts of marshaller demand?*. The validation shows that the forecasting performance strongly depends on the task type. Docking is the only task type with predictable structure as can be seen in the consistently positive R^2 values and stable RMSE and MAE scores across all horizons. The model outperforms both baseline models at every prediction horizon. Showing that LightGBM provides added predictive value for short-term, and midterm forecasting of Docking activities.

For the other tasks, Meeting, Follow me service and Engine testing the validation results show limited or no predictive structure. The R^2 values are close to zero or negative and the models performance is similar to the baseline performance. This indicates that these tasks are irregular or occur in low volumes. Making them not suitable to provide reliable hourly forecasts with the current data.

The baseline comparison support these conclusions. For Docking, the LightGBM models have a lower prediction error and a higher capture of the variation for every horizon. For the other tasks this improvement can not be observed, confirming that the structure needed for forecasting is not strong enough in those categories. The feature selection experiment provides the result that a reduced set of Docking features achieves almost the same performance as the full model. This shows that the main pattern is captured by a small number of temporal features, such as the hour of the day and the daily and weekly lags.

The validation shows that the developed model can only provide reliable forecasts for the task Docking. The other tasks do not show enough structure in the available nine months of data to accurately forecast hourly predictions. The next chapter discusses the conclusion and discussion of this study.

7 Conclusion and discussion

This chapter provides the main outcomes of this study. In Section 7.1, the conclusions to this study is given by first answering the main research question and the sub-questions. Section 7.2 gives the discussion and future work recommendations are given in Section 7.3.

7.1. Conclusion

This section summarises the key findings of the study and presents the answers to the research questions. The goal of this study was to create a predictive model that uses historical data and relevant operational factors to estimate the hourly task demand of airport marshallers while using Amsterdam Airport Schiphol as a case study. The previous chapters described the operational context, data preparation, model selection and the forecasting model development and validation. Based on these findings the following conclusions can be drawn.

7.1.1. Answer to the main research question

The main research question is as follows:

What is the predictive capability of forecasting hourly marshaller task demand using historical operational data and influencing factors?

The predictive capability was assessed by developing three forecasting models for three prediction horizons based on the hourly task demand. ADS-B vehicle data, geofence polygons, aircraft arrival times, and engine testing records were used determine the task demand. A process of spatial matching, task labelling, and segmentation was used to combine these datasets, after which the labelled segments were then combined into hourly counts. Resulting in an aggregated dataset with hourly task count that can be used in a machine learning model.

The LightGBM models were trained on this dataset and evaluated on three prediction horizons using the Mean Absolute Error, Root Mean Squared Error, Mean Error and the Coefficient of Determination (R^2). The models were compared against two baselines, to asses whether the model adds predictive value.

The results show that the predicative capability strongly differs between the task type. Docking is the only task that can be forecasted reliably with the current dataset. Docking follows daily patterns and has a clear link with the aircraft arrivals. This can be seen in the historical workload data and can therefore be captured by the model. Docking shows stable performances over all horizons with RMSE values between 1.66 and 1.75 and MAE values between 1.26 and 1.35. The R^2 values range between 0.43 and 0.45, indicating that a substantial part of the hourly variation is captured. The ME values lie between 0.33 and 0.56, showing that the model slightly overestimates the hourly demand on average. Expressed relatively to the average Dockings of around 2.7-2.8 tasks per hour, this corresponds to an average overestimation of roughly 12-21%. When translated to a daily total, the model predicts about 8-14 extra Docking tasks per day, depending on the horizon. Given an average docking duration of 9.8

minutes, this corresponds to roughly 1.3–2.2 marshaller-hours per day. The maximum average task count is 7 and happens at hour 8.

The comparison with the two baselines models shows that the LightGBM models add predictive value for Docking. For the three models the Docking RMSE reduces by approximately 19 - 21 % compared to the seasonal naïve, and 3 - 9 % compared to the weekly hourly baseline. The R^2 improves by 0.33 - 0.39 for the seasonal naïve baseline and 0.04 - 1.2 for weekly hourly average.

The predictive capability of the other task types is limited. The R^2 values are close to zero showing not much variation can be captured in these patterns. Follow me service and engine testing have a low error, but this can be explained by the low number of tasks in this dataset. Engine testing also shows low average hourly tasks, with a maximum of 0.3 task per hour. The current inputs and dataset are not able to support hourly forecasting.

Based on these results, the predictive capability of the models is restricted to one task type. With the current data available only Docking can be forecasted reliably.

7.1.2. Answer to Sub-question 1

What is the current operational situation regarding the deployment and responsibilities of marshallers, and what are the factors influencing their daily workload and demand?

Marshallers perform five categories of tasks: follow me service, Docking, engine testing, controlling activities and other supporting tasks. Their workload is influenced by predictable and irregular operational factors. Predictable factors include the daily and weekly arrival pattern of flights, which creates clear peaks in task demand. Irregular factors include gate changes, ground handling delays, technical malfunctions such as VDGS outages, maintenance work and ad-hoc assignments from Apron Control. The current staffing levels at Amsterdam Airport Schiphol are static and do not vary across seasons. Creating a mismatch between demand and capacity and motivating the need for a predictive model.

7.1.3. Answer to Sub-question 2

Which type of predictive model is most suitable to forecast the marshallers' task demand?

Several models were compared using functional and non functional requirements. Exact optimisation and simulation models were found unsuitable in this study as they do not generate direct forecasts of task counts. Time-series models SARIMA and SARIMAX can capture seasonal behaviour but predict only one task type at a time. Machine learning models can integrate multiple inputs and predict several task types simultaneously. LightGBM was selected as the most suitable model as it allows multi output forecasting and performs efficiently in computation time. Therefore meeting the requirements defined in this study.

7.1.4. Answer to Sub-question 3

What relevant data is available, and how can it be prepared for use in a predictive model?

The available data used is ADS-B positions of marshaller vehicles, geofence polygons of the airside infrastructure, Actual In-Block Times of arriving aircraft and records of engine testing. These datasets were combined through several processing steps. ADS-B points were matched to polygons representing stands, taxiways and service roads. The matched points were grouped into segments based on location and speed. These segments were labelled to tasks by linking them to arrival times, engine test intervals and other operational rules. Finally, all labelled segments were aggregated per hour, resulting in a structured dataset that forms the input for the forecasting model.

7.1.5. Answer to Sub-question 4

How can a predictive model be developed using the prepared dataset to forecast the marshallers' task demand?

The models were developed by defining the prediction targets (Docking, follow me service, meeting and engine testing) and creating features that describe temporal patterns and operational drivers. These features include hour of the day, day of the week, upcoming holidays, flight arrivals, lagged task values and rolling averages. LightGBM was implemented as a multi-output model that forecasts all task types jointly. The models were trained on nine months of data and performance will be evaluated using separate test sets for each prediction horizon.

7.1.6. Answer to Sub-question 5

How can the developed predictive model be validated to ensure accurate and reliable forecasts of marshaller demand?

The models were validated on three forecasting horizons: 24 hours, 168 hours, and 2160 hours. The performance was measured using four evaluation metrics: Mean Absolute Error, Mean Error, Root Mean Squared Error, and the Coefficient of determination (R^2), and compared to two baselines. Docking showed a higher accuracy than the other task types. The model outperformed the baselines on all prediction horizons and on all evaluation metrics. The other tasks did not show any valuable forecasting, with R^2 values close to zero or negative. A feature selection experiment showed that a small set of features provides the same results for Docking. The validation showed that only Docking can be forecasted reliably with the current dataset.

7.2. Discussion

The results of this study show that only one task category can be forecasted reliably with the current dataset. Docking follows a clear pattern that is driven by the aircraft arrivals. This pattern is visible in the historic workload data and can therefore be learned by the model. The bias analysis shows that Docking models have a positive mean error, this is especially visible in the operational hours from 8 - 23 in which most of the Docking events occur. Docking represents a substantial amount of the workload as indicated by the pilot data, by the average task volume per hour, and by the average duration of 9.80 minutes.

Looking at these results in operational context, the predicted task demand can be seen as an indication of marshaller workload. These predictions provide an indication of peak hours, with a maximum of 7 events on average. As the average Docking duration includes the activities performing a FOD inspection and waiting for aircraft to arrive, translating the task demand to time gives a first indication of workload at an hourly level. However, the forecasting results can not directly be translated into the number of marshallers needed. Factors such as driving time, idle time, and additional tasks that are not included in the dataset set are not explicitly modelled.

The other tasks do not follow such a structure or have a low number of tasks represented in this dataset. The occurrence could depend more on irregular events that are not captured directly in the available dataset. Follow me service shows a sharp peak around hours 7-10, but outside these hours the task volume is low. The model appears to capture the timing of the peak, but limited information is available to learn from for the other hours, reducing the reliability of these forecasts. The operational impact of the event is low as the average duration is 72.2 seconds and the maximum mean tasks lies around 1.4 per hour. Meeting and Engine testing show no alignment in the pattern of the actual and observed values. For all three horizons the direction of the prediction varies, alternating over- and underestimation periods throughout the day.

The models are evaluated using a chronological 80/20 split. This ensures that future data is not used to predict past events, creating a realistic operational forecasting set-up. A limitation however is that the evaluation is based on the most recent time window. As a result, the performance reflects the conditions present in that specific window, and it is unknown how stable the model would be across other months or seasons. The evaluation may overestimate or underestimate on the true performance. The model performs well on the final window for Docking, but the performance across other periods remains unknown.

The feature selection model experiment showed that a small number of features influence the Docking prediction. Across all horizons, hour is the most influential factor, indicating a day pattern. Lagged Docking values for 24 and 168 hours also appear in the most important features for all horizons. These values reflect daily and weekly repetition in the workload. For the longest horizon the feature month becomes relevant, suggesting that seasonal variation plays a role when the forecast window increases. Showing that Docking follows follows a predictable pattern across days, weeks and months, and explain why it can be forecasted reliably with the available dataset.

When comparing the three prediction horizons for Docking, differences can be observed across the evaluation metrics. The 168-hour horizon shows the lowest ME and MAE, while maintaining a high R^2 value. The 2160-hour horizon performs slightly better in terms of RMSE, suggesting a better handling of larger deviations, although the ME is higher compared to other horizons. The 24-hour horizon shows slightly weaker performance across all metrics compared to the 168-hour horizon. Suggesting that the weekly horizon is the most balanced option for forecasting the Docking demand within this study.

The findings suggest that forecasting can support marshaller planning, but only for part of the workload. As discussed above, Docking can be interpreted in terms of workload and peak periods, making it useful for planners to identify moments with increased operational activity. However, for the other tasks, accurate forecasting is not possible with the current data. Meaning that the models only cover a subset of the total workload, which limits their operational impact.

This study has some limitations that could affect the results. The dataset in this study only covers a period of nine months. Seasonal effects are only partly represented as not a whole summer and winter season cycle is included. As a result some long-term trends may be missing in the sets. In addition, some operational drivers were also not available. These include handler delays, VDGS malfunctions and other disruptions that influence the workload of marshallers. Because these factors were not directly included as features, but indirectly in the historical workload, the model cannot fully capture their effect on the task demand.

The dataset contains a low number of Follow me service and Engine testing tasks. These tasks provide limited information for the model to learn from and therefore limits the predictive capability of these tasks. There are also smaller tasks that marshallers perform, but for which no data was available. These tasks were therefore not included in the dataset. However, it is likely that they would show a similar effect as Engine Testing and Meeting. Because they also occur infrequently and not often as for example Docking. The model therefore assumes that the four selected task types represent the workload sufficiently, and that excluded smaller tasks do not contribute severely to additional predictable structure.

Several modelling assumptions are made that can have an effect on the results. Each prediction horizon is modelled using a separate LightGBM model. This means that patterns per horizon are learned independently, and no information is shared across horizons. Secondly, the workload is aggregated into hourly task counts. The task assigned to each hour are based on their starting time, even when they only fall partially in that hour. A different rule for assigning the task could have placed that task in the next hour when most of the duration happens there. Lastly, a limited set of drivers is implemented is the models, as some drivers are not available in structured form. Indirectly assuming that their influence is either small or expressed within the historial workload.

The usage of ADS-B data also introduces some limitations. Some raw datapoints are missing, which means that certain activities may not have been recorded. The task labelling depends on the accuracy of the geofence polygons. The polygons have been manually enlarged, and even though this was done with care, it may still affect the accuracy of the matching. At least one area on the airside field exists where the ADS-B signal is not correctly transmitted. Trough visualisation of the trajectories at different days this are was identified. However, it is possible a similar area exists that is not discovered. Resulting in that some movements may be missing in the data. Small deviations in the accuracy of the ADS-B data itself can happen, meaning that the task labelling is not fully accurate and some tasks are missing in the aggregated dataset. Furthermore, even though the labelling rules have been made up with precision, and several days of data were manually checked by comparing trajectories with their segmented task tables, it remains possible that some inaccuracies were not detected.

7.3. Future work

This section gives two types of recommendations. The first part gives suggestions that can be applied directly at Amsterdam Airport Schiphol. The second part outlines directions for future research.

7.3.1. Recommendations for Amsterdam Airport Schiphol

The first step is to extend the dataset with different factors. The current dataset covers nine months. A longer dataset that includes multiple summer and winter seasons would allow more stable patterns to be identified and fully validate longer prediction horizons. In addition, operational drivers should be collected and included, such as VDGS malfunctions, handler delays, and maintenance activities. Including these factors could improve predictability. Furthermore, it is also recommended to record smaller marshaller tasks that are currently not documented. These tasks may show similar behaviour to the low-frequency tasks in this study and should be included in future datasets.

Furthermore, an analysis of the transponder system should be carried out to see if there are any additional areas where the ADS-B data is not submitted correctly. Identifying these areas would help to improve the accuracy of task labels. It may also be useful to compare vehicle transponder data with aircraft transponder data to extract additional follow me service activities.

The link between forecasting and planning should be explored. The predicted Docking demand gives an indication of peak periods and could be used to support planning decisions at Schiphol. As task types and their average duration have been identified, this can be used in a simulation model. These simulation models could explore the travel time between tasks, and idle time, especially when additional marshaller tasks are identified and included. In combination with the forecasting model this would not only asses how many marshallers are required under different conditions, but could also provide an analysis of different allocation strategies or the sensitivity of the system to operational disruptions. The effects of certain maintenance activities can be explored, but also assigning marshallers to a specific operational area within a shift.

7.3.2. Recommendations for future research

The method proposed in this study should be tested on other airports, to test the generality of the system. This would show which parts of the approach can be directly transferred. It also shows if the difficulties found in forecasting the follow me service, Meeting and Engine Testing tasks also appear elsewhere. Such tests would clarify whether the observed limitations are specific to Schiphol or reflect structural characteristics of marshaller operations. This could also indicate to what extent operational drivers differ per airport.

Future research should also examine if alternative forecasting methods can improve the performance for the task types that could not be predicted reliably in this study. These tasks show irregular patterns and occur with low frequency. Methods that are designed for limited event data or methods that predict the probability a task occurs, may give better results. Time series cross validation could also be applied to evaluate the model across multiple time windows in addition to the 80/20 split.

It could also be investigated if hourly forecasting is the most suitable approach. Other resolutions such as 15-minute or 30-minute time intervals, or shift intervals may be relevant. As different rules for assigning tasks to time windows leads to different workload patterns, this option could also be examined to see the influence on the dataset. Finally, the current study uses three separate models, one for each prediction horizon. Future research could explore if the horizons can be combined into one forecasting model.

References

[1] Adetayo Olaniyi Adeniran and Olufunto Adedotun Kanyio. *Artificial intelligence in aircraft docking: the fear of reducing ground marshalling jobs to robots and Way-Out*. Tech. rep. 2018. URL: www.cribfb.com/journal/index.php/aijmsr.

[2] Aeroclass. *Airport Operations*. Accessed: 2025-10-01. 2021. URL: https://www.aeroclass.org/airport-operations/.

[3] Fahad Radhi Alharbi and Denes Csala. "A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach". In: *Inventions* 7.4 (2022). ISSN: 2411-5134. DOI: 10.3390/inventions7040094. URL: https://www.mdpi.com/2411-5134/7/4/94.

[4] Sobia Ali and Yasir Farooqi. "Effect of Work Overload on Job Satisfaction, Effect of Job Satisfaction on Employee Performance and Employee Engagement (A Case of Public Sector University of Gujranwala Division)". In: (Apr. 2014).

[5] Apostolos Ampountolas. "Modeling and Forecasting Daily Hotel Demand: A Comparison Based on SARIMAX, Neural Networks, and GARCH Models". In: *Forecasting* 3.3 (2021), pp. 580–595. ISSN: 2571-9394. DOI: 10.3390/forecast3030037. URL: https://www.mdpi.com/2571-9394/3/3/37.

[6] Francis Ayiah-Mensah et al. "Advancements in seasonal rainfall forecasting: A seasonal autoregressive integrated moving average model with outlier adjustments for Ghana's Western Region". In: *Scientific African* 28 (2025), e02632. ISSN: 2468-2276. DOI: 10.1016/j.sciaf.2025.e02632.

[7] Tisiyanah Abd Baki et al. "Digitalization of Airside Operations Process to Improve Airport Operations For The Case of Malaysia Airports". In: *2022 4th International Conference on Smart Sensors and Application (ICSSA)*. 2022, pp. 130–134. DOI: 10.1109/ICSSA54161.2022.9870954.

[8] Casher Belinda, Shimul Melwani, and Chaitali Kapadia. "Breaking Boredom: Interrupting the Residual Effect of State Boredom on Future Productivity". In: *Journal of Applied Psychology* 109.6 (2024), pp. 829–849. DOI: 10.1037/apl0001161. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85188520892&doi=10.1037%2fapl0001161&partnerID=40&md5=2e3880a0260c9163683acedb568e8730.

[9] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms". In: *Artificial Intelligence Review* 54.3 (2021), pp. 1937–1967. ISSN: 1573-7462. DOI: 10.1007/s10462-020-09896-5. URL: https://doi.org/10.1007/s10462-020-09896-5.

[10] Emily Brown et al. *Comparative Analysis of LightGBM with Traditional Credit Assessment Methods*. Nov. 2024. DOI: 10.13140/RG.2.2.29039.65444.

[11] Quang-Thanh Bui et al. "Gradient Boosting Machine and Object-Based CNN for Land Cover Classification". In: *Remote Sensing* 13 (July 2021), p. 2709. DOI: 10.3390/rs13142709.

[12] Centraal Bureau voor de Statistiek (CBS). *Hogere werkdruk belangrijkste gevolg personeelstekort volgens ondernemers*. Accessed: 2025-10-01. 2024. URL: https://www.cbs.nl/nl-nl/nieuws/2024/35/hogere-werkdruk-belangrijkste-gevolg-personeelstekort-volgens-ondernemers.

[13] Chien-Ming Chen and Howard Hao-Chun Chuang. "Time to shift the shift: Performance effects of within-day cumulative service encounters in retail stores". In: *Omega (United Kingdom)* 119 (2023). DOI: 10.1016/j.omega.2023.102892. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159635648&doi=10.1016%2fj.omega.2023.102892&partnerID=40&md5=a58c89d06790a8fed599fe916d93be08.

[14] Cheolmin Choi, Jung-Ho Ahn, and Hyeran Byun. “Visual recognition of aircraft marshalling signals using gesture phase analysis”. In: *2008 IEEE Intelligent Vehicles Symposium*. 2008, pp. 853–858. DOI: 10.1109/IVS.2008.4621186.

[15] Bryan Clark and Fangfang Lee. *What is Gradient Boosting?* Oct. 2025. URL: https://www.ibm.com/think/topics/gradient-boosting.

[16] Copenhagen Optimization. *Airport Operations 101*. Accessed: 2025-10-01. 2025. URL: https://copenhagenoptimization.com/blog/airport-operations-101.

[17] Giacomo Dall’Olio and Rainer Kolisch. “Formation and Routing of Worker Teams for Airport Ground Handling Operations: A Branch-and-Price-and-Check Approach”. In: *Transportation Science* 57.5 (2023), pp. 1231–1251. DOI: 10.1287/trsc.2022.0110. URL: https://doi.org/10.1287/trsc.2022.0110.

[18] Qichen Deng, Bruno F. Santos, and Wim J.C. Verhagen. “A novel decision support system for optimizing aircraft maintenance check schedule and task allocation”. In: *Decision Support Systems* 146 (2021). DOI: 10.1016/j.dss.2021.113545. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102879361&doi=10.1016%2fj.dss.2021.113545&partnerID=40&md5=136eed0b53b45e41365fb459ad07748d.

[19] Vincent Derkinderen, Jessa Bekker, and Pieter Smet. “Optimizing workforce allocation under uncertain activity duration”. In: *Computers and Industrial Engineering* 179 (2023). DOI: 10.1016/j.cie.2023.109228. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85152096762&doi=10.1016%2fj.cie.2023.109228&partnerID=40&md5=76ff660e821a852991b0420dfec7db8e.

[20] Dimitrios Dimitriou, Maria Sartzetaki, and Aristi Karagkouni. “Airport Landside Area Planning: An Activity-Based Methodology for Seasonal Airports”. In: *Transportation Research Procedia* 82 (2025). World Conference on Transport Research - WCTR 2023 Montreal 17-21 July 2023, pp. 1167–1184. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2024.12.119. URL: https://www.sciencedirect.com/science/article/pii/S2352146524004186.

[21] Duarte Dinis and Ana Paula Barbosa-Póvoa. “On the Optimization of Aircraft Maintenance Management”. In: *Studies in Big Data* 15 (2015), pp. 49–57. DOI: 10.1007/978-3-319-24154-8_7. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85132872165&doi=10.1007%2f978-3-319-24154-8_7&partnerID=40&md5=e5441b3385e53e49508a0e7b43326719.

[22] Patrick Eichenseer, Lukas Hans, and Herwig Winkler. “A data-driven machine learning model for forecasting delivery positions in logistics for workforce planning”. In: *Supply Chain Analytics* 9 (2025). DOI: 10.1016/j.sca.2024.100099. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85213570297&doi=10.1016%2fj.sca.2024.100099&partnerID=40&md5=45ff1ae21b71d4cab07e68cff95af45f.

[23] Eindhoven Airport. *Beroepen op de luchthaven*. 2025. URL: https://www.eindhovenairport.nl/nl/beroepen-op-de-luchthaven.

[24] European Union Aviation Safety Agency (EASA). *Easy Access Rules for Aerodromes Regulation (EU) No 139/2014*. Geraadpleegd op 3 juni 2025. 2024. URL: https://www.easa.europa.eu/en/document-library/easy-access-rules/easy-access-rules-aerodromes.

[25] European Union Aviation Safety Agency (EASA). *European Union Aviation Safety Agency (EASA) Homepage*. Geraadpleegd op 3 juni 2025. 2024. URL: https://www.easa.europa.eu/en/home.

[26] Matteo Gabellini et al. “A hybrid approach integrating genetic algorithm and machine learning to solve the order picking batch assignment problem considering learning and fatigue of pickers”. In: *Computers and Industrial Engineering* 191 (2024). DOI: 10.1016/j.cie.2024.110175. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85191660844&doi=10.1016%2fj.cie.2024.110175&partnerID=40&md5=5c90d51c73d0e3cdec1b7f1ee03850e9.

[27] GlobeAir. *Airside Operations*. https://www.globeair.com/g/airside-operations. Accessed: 2025-10-01. n.d.

[28] Marcin Jurczak, Grzegorz Miebs, and Rafał A. Bachorz. “Multi-criteria human resources planning optimisation using genetic algorithms enhanced with MCDA”. In: *Operations Research and Decisions* 32.4 (2022), pp. 57–74. DOI: 10.37190/ord220404. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168451174&doi=10.37190%2ford220404&partnerID=40&md5=53a8381956676862678ba192edb73b72.

[29] Guolin Ke et al. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. Dec. 2017.

[30] Ivan Kovynyov and Ralf Mikut. “Digital technologies in airport ground operations”. In: *NETNOMICS: Economic Research and Electronic Networking* 20.1 (2019), pp. 1–30. ISSN: 1573-7071. DOI: 10.1007/s11066-019-09132-5. URL: https://doi.org/10.1007/s11066-019-09132-5.

[31] Mariel S. Lavieri and Martin L. Puterman. “Optimizing nursing human resource planning in British Columbia”. In: *Health Care Management Science* 12.2 (2009), pp. 119–128. DOI: 10.1007/s10729-008-9097-0. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-64149092201&doi=10.1007%2fs10729-008-9097-0&partnerID=40&md5=835a308a5936f90239da246630a4dece.

[32] Weizhang Liang et al. “Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms”. In: *Mathematics* 8.5 (2020). ISSN: 2227-7390. DOI: 10.3390/math8050765. URL: https://www.mdpi.com/2227-7390/8/5/765.

[33] Tomas Lidén, Christiane Schmidt, and Rabii Zahir. “Improving attractiveness of working shifts for train dispatchers”. In: *Transportmetrica B* 12.1 (2024). DOI: 10.1080/21680566.2024.2380912. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85199979329&doi=10.1080%2f21680566.2024.2380912&partnerID=40&md5=16d9788657ce666b9a19dbcbe850805e.

[34] Lei Lin et al. “Quantifying uncertainty in short-term traffic prediction and its application to optimal staffing plan development”. In: *Transportation Research Part C: Emerging Technologies* 92 (2018), pp. 323–348. DOI: 10.1016/j.trc.2018.05.012. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047176727&doi=10.1016%2fj.trc.2018.05.012&partnerID=40&md5=23e375b72d125a885a1e135a7b9710b1.

[35] Jiaming Liu et al. “A non-hierarchical approach to integrate airport airside operations using adaptive large neighbourhood search”. In: *Transportation Research Part C: Emerging Technologies* 171 (2025), p. 104948. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2024.104948. URL: https://www.sciencedirect.com/science/article/pii/S0968090X24004698.

[36] Caterina Malandri, Luca Mantecchini, and Vasco Reis. “Aircraft turnaround and industrial actions: How ground handlers’ strikes affect airport airside operational efficiency”. In: *Journal of Air Transport Management* 78 (2019), pp. 23–32. ISSN: 0969-6997. DOI: https://doi.org/10.1016/j.jairtraman.2019.04.007. URL: https://www.sciencedirect.com/science/article/pii/S0969699719300183.

[37] Maged Mamdouh, Mostafa Ezzat, and Hesham Hefny. “Optimized Planning of Resources Demand Curve in Ground Handling based on Machine Learning Prediction”. In: *I.J. Intelligent Systems and Applications* 13.1 (2021), pp. 1–16. DOI: 10.5815/ijisa.2021.01.01. URL: https://www.mecs-press.org/ijisa/ijisa-v13-n1/IJISA-V13-N1-1.pdf.

[38] Mappedin. *Airport Operations Management: Definition, Types, and How to Improve*. Accessed: 2025-10-01. 2025. URL: https://www.mappedin.com/resources/blog/airport-operations-management-definition-types-and-how-to-improve/.

[39] NOS. *Opnieuw meer dan 100 vluchten geannuleerd door KLM vanwege staking*. Accessed: 2025-10-01. 2025. URL: https://nos.nl/artikel/2583607-opnieuw-meer-dan-100-vluchten-geannuleerd-door-klm-vanwege-staking.

[40] NOS. *Vliegverkeer twee keer gestaakt op Eindhoven Airport, vluchten vertraagd*. Accessed: 2025-10-01. 2025. URL: https://nos.nl/artikel/2578549-vliegverkeer-twee-keer-gestaakt-op-eindhoven-airport-vluchten-vertraagd.

[41] Adeniran Adetayo O and Akinsehinwa Feyisola O. “Acceptance of visual docking guidance system by ground marshallers in Nigerias’ airport”. In: *International Journal of Advanced Networking and Applications* 13.01 (Jan. 2021), pp. 4845–4854. DOI: 10.35444/ijana.2021.13107. URL: https://doi.org/10.35444/ijana.2021.13107.

[42] Alena Otto and Armin Scholl. “Reducing ergonomic risks by job rotation scheduling”. In: *OR Spectrum* 35.3 (2013), pp. 711–733. DOI: 10.1007/s00291-012-0291-6. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84879212526&doi=10.1007%2fs00291-012-0291-6&partnerID=40&md5=d400da858b3fab14d9ed016f2aeed37a.

[43] Rahmat Rabet, Seyed Mojtaba Sajadi, and Mahshid Tootoonchy. “A hybrid metaheuristic and simulation approach towards green project scheduling”. In: *Annals of Operations Research* (2024). DOI: 10.1007/s10479-024-06291-z. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85208584292&doi=10.1007%2fs10479-024-06291-z&partnerID=40&md5=ee0d7db43e8057a932f10c5118356d24.

[44] Royal Schiphol Group. *A look ahead at air traffic*. https://www.schiphol.nl/en/schiphol-as-a-neighbour/look-ahead-at-air-traffic/. Accessed: 2025-10-02. 2025.

[45] Royal Schiphol Group. *Airport Industry Connectivity Report*. https://www.schiphol.nl/en/advertising/news/airport-industry-connectivity-report/. Accessed: 2025-10-02. 2025.

[46] Royal Schiphol Group. *Annual Traffic Review 2024*. https://assets.ctfassets.net/biom0eqyyi6b/4v1daDHAKUyLAPzQMpDegS/cc3ffaf282654440c0de380a2553c038/Annual_Traffic_Review_2024_C.pdf. Accessed: 2025-10-02. 2024.

[47] Emily Schiller et al. “No Data Left Behind: Exogenous Variables in Long-Term Forecasting of Nursing Staff Capacity”. In: 2024. DOI: 10.1109/DSAA61799.2024.10722806. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85209404049&doi=10.1109%2fDSAA61799.2024.10722806&partnerID=40&md5=86f3446eee4a2d34d77e626d8ec6ce78.

[48] Schiphol. *Een dag in het leven van een marshaller*. Geraadpleegd op 30 april 2025. 2024. URL: https://www.schiphol.nl/nl/blog/een-dag-in-het-leven-van-een-marshaller/.

[49] Schiphol. *Flight paths and runway use*. https://www.schiphol.nl/en/schiphol-as-a-neighbour/flight-paths-and-runway-use/. Accessed: 2025-12-11. 2025.

[50] Amsterdam Airport Schiphol. *AO.ASE.n.3.4: vliegtuigen begeleiden en docken*. 2019.

[51] Amsterdam Airport Schiphol. *ASE-M.302: Vliegtuigdocking d.m.v. marshalling*. 2023.

[52] Amsterdam Airport Schiphol. *ASE-M.308 Begeleiden naar alternatieve proefdraaiplaats*. 2023.

[53] Amsterdam Airport Schiphol. *ASE-M.309 Toezicht houden op staat en gebruik VOPs*. 2023.

[54] Amsterdam Airport Schiphol. *ASE-M.311 Buscoördinatie*. 2023.

[55] Amsterdam Airport Schiphol. *Basis.3005 VDGS storing aanmaken en afhandelen*. 2024.

[56] Amsterdam Airport Schiphol. *Basis.3025 Begeleiden aanpikken*. 2025.

[57] Michael Schmidt. “A review of aircraft turnaround operations and simulations”. In: *Progress in Aerospace Sciences* 92 (2017), pp. 25–38. ISSN: 0376-0421. DOI: https://doi.org/10.1016/j.paerosci.2017.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0376042117300039.

[58] Léon Sobrie, Marijn Verschelde, and Bart Roets. “Explainable real-time predictive analytics on employee workload in digital railway control rooms”. In: *European Journal of Operational Research* 317.2 (2024), pp. 437–448. DOI: 10.1016/j.ejor.2023.09.016. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85173074876&doi=10.1016%2fj.ejor.2023.09.016&partnerID=40&md5=e2e3c0c3c7080272f170bad6fcfe3c4a.

[59] Yale Song, David Demirdjian, and Randall Davis. “Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database”. In: *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*. 2011, pp. 500–506. DOI: 10.1109/FG.2011.5771448.

[60] Xuxue Sun et al. “A latent survival model integrated computer simulation-based evaluation for nursing home staffing”. In: *Computers and Industrial Engineering* 177 (2023). DOI: 10.1016/j.cie.2023.109074. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85147843355&doi=10.1016%2fj.cie.2023.109074&partnerID=40&md5=43b3382faafaeabce6596394a2afd2ec.

[61] Patrik Šváb et al. “Innovations in the field of aircraft ground handling”. In: Nov. 2021, pp. 159–162. DOI: 10.1109/NTAD54074.2021.9746151.

[62] Hamish Thorburn et al. “A time-expanded network design model for staff allocation in mail centres”. In: *Journal of the Operational Research Society* 75.10 (2024), pp. 1949–1964. DOI: 10.1080/01605682.2023.2287613. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179740922&doi=10.1080%2f01605682.2023.2287613&partnerID=40&md5=54562657111a7f909d8e0c3410a4381d.

[63] Xin Wen et al. “Airline crew scheduling with sustainability enhancement by data analytics under circular economy”. In: *Annals of Operations Research* 342.1 (2024), pp. 959–985. DOI: 10.1007/s10479-023-05312-7. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85158099233&doi=10.1007%2fs10479-023-05312-7&partnerID=40&md5=5a6b68f8446df50151b45bb3d5ec1e46.

[64] Xin Wen et al. “Individual scheduling approach for multi-class airline cabin crew with manpower requirement heterogeneity”. In: *Transportation Research Part E: Logistics and Transportation Review* 163 (2022). DOI: 10.1016/j.tre.2022.102763. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85131405716&doi=10.1016%2fj.tre.2022.102763&partnerID=40&md5=e826f9a25a499bfc79159f5ae90f7bad.

[65] Cheng-Lung Wu and Shao Xuan Lim. “Effects of enterprise bargaining and agreement clauses on the operating cost of airline ground crew scheduling”. In: *Journal of Air Transport Management* 91 (2021). DOI: 10.1016/j.jairtraman.2020.101972. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096000602&doi=10.1016%2fj.jairtraman.2020.101972&partnerID=40&md5=0032df26fc22741f5c1839b4b9ac87ff.

[66] Mark S. Young. “In Search of the Redline: Perspectives on Mental Workload and the ‘Underload Problem’”. In: *Human Mental Workload: Models and Applications*. Ed. by Luca Longo and Maria Chiara Leva. Cham: Springer International Publishing, 2021, pp. 3–10.

[67] Baocheng Zhang et al. “Evaluating the operational performance of airside and landside at Chinese airports with novel inputs”. In: *Transportation Planning and Technology* 41.8 (2018), pp. 878–900. DOI: 10.1080/03081060.2018.1526966. eprint: https://doi.org/10.1080/03081060.2018.1526966. URL: https://doi.org/10.1080/03081060.2018.1526966.

[68] Samantha L. Zimmerman et al. “Optimising nurse schedules at a community health centre”. In: *Operations Research for Health Care* 30 (2021). DOI: 10.1016/j.orhc.2021.100308. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110680973&doi=10.1016%2fj.orhc.2021.100308&partnerID=40&md5=29567d7517318edff5719d6e6af775c9.

A

Scientific paper

See next page

Waving into the Future
Development of a Predictive Model for the Deployment of Airport Marshallers
A Case Study at Amsterdam Airport Schiphol

A.I. Ruha, Ir. M.B. Duinkerken, Dr. A. Napoleone, C. Moll Msc.

Abstract— **This study develops a method to forecast hourly task demand of airport marshallers. Amsterdam Airport Schiphol is used as a case study. The methods combines ADS-B vehicle data, Geofence polygon map, AIBT data and engine test logs to label marshallers tasks. LightGBM is used to predict four task types at three prediction horizons (24, 168 and 2160 hours ahead). Temporal features, lagged workload, and flight arrivals are used as input. The performance of the model is tested using MAE, RMSE, ME and R^2, and compared with a seasonal naïve and weekly hourly baseline. The results show that only Docking can be forecasted reliably with the available dataset. Docking has stable predictions across all horizons, with MAE values between 1.30 and 1.36 and RMSE values between 1.66 and 1.75. The R^2 values lie between 0.43 and 0.45. The Docking model also outperforms both baselines on these metrics. The other task types show no meaningful predictive structure, mainly due to low frequency and irregular occurrence.**

I. Introduction

Airport operations are a vital link in the chain of an airport. Every flight ultimately depends on what happens on the ground, without the activities that take place between arrival and departure, aircraft cannot leave on time, and the whole system would experience delays and fall apart. Airport operations are all the processes and activities that take place continuously inside and around an airport that ensure smooth, safe, and efficient functioning for passengers, staff, and aircraft. This includes passenger services, aircraft support, and flight coordination [1], [2], [3]. Effective airport operations require careful planning, coordination and the integration of multiple stakeholders to handle the complexity of modern air travel.

Airport operations can be divided into landside and airside operations. Airside operations are related to aircraft activities and include runways, taxiways, aprons and all tasks necessary to prepare aircraft for a flight. Access to the airside zone requires screening for both passengers and staff to ensure security [4], [5]. Landside operations encompass passenger terminals, cargo facilities, and other elements of the airport's land-based infrastructure [5]. Ground operations, also referred to as ground handling, cover services required by an airline between landing and take-off, such as marshalling, (un)loading of baggage, refuelling, cleaning, catering, baggage handling, passenger handling, cargo handling, aircraft maintenance, and aviation security services [6]. Airside operations involve responsibilities such as allocating aircraft parking and escort services, reacting to airside incidents, wildlife hazard management, and foreign object debris management, and supervising inspections of runways and taxiways [7]. These responsibilities differ per airport, and close coordination among all actors is essential to keep the overall airport operation safe and efficient.

In recent years, air traffic has continued to grow, both in terms of passenger demand and the number of scheduled flights. With this increase, delays have also been rising, as more operations need to be fitted into already congested schedules [8]. Labour shortages and strikes in the ground handling sector, such as those at KLM in the Netherlands [9], [10], have highlighted the vulnerability of airport operations to staffing disruptions. These developments reinforce the need for better planning and predictive methods to ensure that ground and airside operations can function reliably under increasing pressure.

Airside and ground operations are dependent on their personnel. The safe and efficient functioning depend directly on the availability of sufficient staff at the right time and place. Workforce demand in these operations is highly variable. Flight arrivals often cluster in waves, creating peaks in workload within short periods. Certain conditions can suddenly shift the demand from automated to manual procedures, and safety related tasks can appear at irregular moments. These factors cause a demand pattern that is difficult to anticipate with static rosters. When a reliable prediction is absent, organisations face overstaffing or understaffing. Overstaffing leads to inefficiencies, higher costs, and the potential of employees without tasks risk losing concentration and motivation [11], [12]. Understaffing increases workload pressure, reduces opportunities to take proper breaks, and lowers job satisfaction [13]. Higher work pressure is the most common consequence of staff shortages [14].

Marshallers are a good example of a group support employees in airside and ground operations that illustrate this problem. They carry out a mix of routine and ad-hoc tasks and are often required at the moment when automation fails or unusual events occur. Their workload varies, making it difficult to match supply and demand. In current practice, workforce demand in ground and airside operations is treated with a deterministic input in planning models, or staffing models are fixed based on experience and intuition. These approaches overlook the variable task demand in these operations. The gap between how demand is modelled and how it actually behaves leads to a mismatch in supply and demand. As a result leading to inefficiencies, delays and risks to safety. Developing a data-driven predictive method that captures variability is therefore essential to prevent disruption and to sustain airport operations under growing pressure.

Workforce scheduling has been studied in sectors where personnel demand fluctuates and operations are continuous. Several studies illustrate how predictive models optimise workforce deployment across aviation tasks. In aircraft maintenance, heuristics and integer programming improve workforce planning [15], [16]. Within crew scheduling, integer programming and metaheuristics are used to optimise scheduling while taking into account labour agreements and utilisation [17], [18]. Research in ground handling has focused on baggage operations [19] and machine learning for workload prediction at Cairo International Airport [20]. Studies have also shown that operational planning affects emission levels [21], [22].

However, there is limited research available on marshallers, especially regarding their deployment. Existing studies focus on acceptance of visual docking guidance systems [23], [24], training and safety [25], or gesture recognition [26], [27]. These studies do not predict when and how many marshallers are required. There is a clear gap in the literature showing how predictive modelling can forecast the required marshallers in this dynamic environment.

The objective of this study is to develop a predictive model that can accurately predict the marshaller task demand at an airport. Amsterdam Airport Schiphol is used as a case study. Accurate forecasting of task demand is essential to ensure operational continuity, minimise inefficiencies caused by overstaffing, and reduce the safety and workload risks associated with understaffing. The model contributes to both the operational and tactical level of workforce planning, by providing reliable predictions

The aim is to build a data-driven model using historical data and influencing factors. Different factors are considered, such as the number of arriving flights. Based on these variables, a prediction is made of the hourly task demand of marshallers. This leads to the following research question:

What is the predictive capability of forecasting hourly marshaller task demand using historical operational data and influencing factors?

II. Methodology

In order to describe the methodology first the Machine Learning model is chosen based on multiple requirements. the functional requirements:

1) Predict the number of marshaller tasks per hour, including a distinction between different tasks types.
2) Capture seasonal, weekly and daily variations.
3) Use historical flight arrivals data as input.
4) Use historical tasks activity per hour as input, including distinction between different tasks types.
5) Integrate exogenous factors into the model.
6) Give predictions on a short term and mid term planning.
7) Give reliable predictions based on the available nine months of historical data.
8) Preferably forecast multiple task types within a single modelling framework

The non-functional requirements:

1) Scalable, so it can be used on other airports or different situations without big changes.
2) Efficient in computation time, producing forecasts quickly enough to support operational planning.
3) Robust, giving stable predictions even with unexpected variations in data.
4) Accurate, producing reliable forecasts of marshaller tasks.

Different predictive models were evaluated against functional and non-functional requirements. Exact optimisation, metaheuristics and simulation models are not suitable because they are mainly used for scenario testing or optimization, not for forecasting tasks per hour. Time series models such as SARIMA and SARIMAX can capture seasonality but are less flexible, as they can only predict one task at a time. Machine learning models such as LightGBM and XGBoost are able to use multiple input sources, integrate external factors and scale to other situations, LightGBM performs better on computation time. LightGBM also satisfies the requirement to forecast multiple task types within one modelling framework, whereas SARIMA and SARIMAX would require a separate model for each task category. This makes LightGBM more consistent and easier to maintain for operational use. Therefore, LightGBM is the most suitable model type for forecasting marshaller task demand.

The methodology of this study follows the steps that transforms raw operational data into marshaller task demand, by using processing steps and a LightGBM machine learning model. The approach is structured into data collection, data processing, task labelling, feature engineering and model development. The first three steps are divided into smaller steps and can be seen in Figure 1. The datasets used in this research consists of ADS-B vehicle data from marshaller vehicles, geofence polygons describing the airside layout, Actual In-Block Times (AIBT), engine test logs. These datasets together represent the movements of marshallers within the airside zone and which tasks they perform. These data sources allow us to interpret raw ADS-B trajectories, identify operational tasks, and prepare a structured dataset that can be used for forecasting.

A. Data processing

Every ADS-B point from the dataset is matched to the correct polygon from the geofence map. These points are

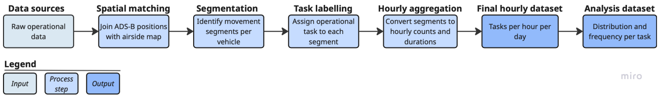

Data sources
Raw operational data
Spatial matching
Join ADS-B positions with airside map
Segmentation
Identify movement segments per vehicle
Task labelling
Assign operational task to each segment
Hourly aggregation
Convert segments to hourly counts and durations
Final hourly dataset
Tasks per hour per day
Analysis dataset
Distribution and frequency per task
Legend
Input
Process step
Output
miro

Fig. 1. Overview data processing steps

grouped into segments. A segment represents a continuous period in which a vehicle stays within the same operational context. After the ADS-B points have been grouped into segments, these can be assigned to tasks. The segments are combined with other datasets and constraints to create the task labelling. the labelling is based on two main principles. If a segment matches a clearly defined event it receives that corresponding label. Otherwise, if no event matches based on the speeds of that segment the label driving or standing is given. The labels that can be given are as follows: Docking, Follow me service, Engine testing, Shift change periods, Apron office, Meeting, Driving and standing.

The final step in the data preparation is to convert these tables per day per vehicle into hourly workload values. The forecasting model requires a consistent time series, this means that all activity must be summarised per hour and per day. Every segment is assigned to the hour in which it begins. Each day results in a complete table of 24 rows, containing the task counts. An The aggregated dataset forms the final output of the data preparation process. It represents a complete and structured overview of the operational workload and serves as the input for the forecasting model.

B. Model scope

The predictive model aims to forecast the hourly task demand of marshallers. The prediction is how many marshaller tasks are expected to occur for different time horizons. Three prediction horizons are implemented and each one is implemented in its own LightGBM model. The three models correspond to the following prediction hours: 24 hours, 168 hours and 2160 hours into the future, representing one day, one week and approximately three months. The prediction targets are the hourly counts of four main task categories:

1) Docking
2) Follow me service
3) Meeting
4) Engine testing

C. Feature engineering

Feature engineering forms an essential step in developing the predictive model, as it determines which operational factors are translated into model inputs. The selected features describe both temporal patterns and operational drivers that influence the workload of marshallers. Several calendar based variables are included to capture daily, weekly, and seasonal patterns. These consist of hour of the day, day of the week, month, weekend indicator, summer/winter indicator, and a holiday indicator. The number of incoming flights is also included and a 24 hour rolling mean is added. This feature helps the model identify broader fluctuations in the arrival pattern that do not depend only on the most recent hour.

Lagged features represent the value of a variable in the previous time step. For example, *Docking_lag1* stores the number of Docking tasks one hour earlier. Rolling features represent the average value value over a given time window. For example, *Docking_roll3* stores the average number of Docking tasks over the last three hours. These lag and roll windows cover 1, 3, 6, 12, 24 and 168 hours, which represent hourly, half-day, daily and weekly effects. Furthermore, trend-based features compare each task category with its value one day earlier (24 hours) and one week earlier (168 hours). These differences help the model recognise upward or downward movements in task demand.

The outputs of the model are the expected hourly counts of the four marshaller tasks at the chosen prediction horizon.

D. Model implementation

The predictive models are implemented in Python, using a combination of open-source libraries. The lightGBM package provides the gradient boosting tree algorithm and the Scikit-learn library allows the use of multi-output regressor. The Optuna framework is used for hyperparameter optimisation, as this allows a flexible and efficient way to automate the search for optimal hyperparameters. The LightGBM models use regression as objective.

The number of estimators controls how many trees are added during boosting. The learning rate sets the step size in updating the model after each tree. The number of leaves and the maximum depth determine how complex each tree may become. The parameter for minimum child samples sets the minimum number of observations required to create a split. The subsample and colsample_bytree settings influence how many rows and features are used when constructing trees.

The models are trained in a deterministic to ensure consistent behaviour. LightGBM has a deterministic training mode, but that setting alone is not enough to provide a full deterministic training. Random variation can still happen from row sampling, feature sampling, bagging and the initialisation of the optimisation procedure. These effects are removed by fixing the random seed, disabling sampling by setting the subsample and colsample_bytree parameters to one and by turning off bagging. The Optuna sampler also uses a fixed seed. Resulting in the fact that the same data always leads to the same model. The search ranges and the final configurations are seen in Table I.

TABLE I
HYPERPARAMETER SEARCH RANGES AND FINAL CONFIGURATIONS PER PREDICTION HORIZON

Parameter	**Search range**	**t+24**	**t+168**	**t+2160**
n_estimators	200 - 800	341	645	433
learning_rate	0.01 - 0.20	0.01136	0.01338	0.01023
num_leaves	20 - 120	52	20	32
max_depth	3 - 15	5	4	13
min_child_samples	5 - 60	40	9	8
subsample	fixed at 1.0	1.0	1.0	1.0
colsample_bytree	fixed at 1.0	1.0	1.0	1.0

E. Evaluation

The performance of the predictive models is evaluated using three metrics: Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), Mean error (ME), and the coefficient of determination (R^2).

The MAE measures the average size of the prediction error. It expresses how many tasks the model is off on average, with lower values indicating better accuracy. It can be expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$$

Where y_i represents the observed values, y_i the predicted value, and N is the total number of samples.

The RMSE measures the average magnitude of the prediction error and is expressed in the same unit as the target variable. The lower the RMSE the higher the predictive accuracy is. It can be expressed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Where y_i represents the observed values, y_i the predicted value, and N is the total number of samples.

The Mean Error (ME) measures the average systematic difference between the observed value and the predicted value. It shows if the model systematically overestimates or underestimates the task demand. A positive ME indicates overestimation and a negative ME indicates underestimation. It can be expressed as:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)$$

where y_i represents the observed values, $\hat{y}_i$ the predicted value, and N is the total number of samples.

The R^2 shows to which extent the model explains the variation in the observed data. It has a value between 0 and 1, with values closer to 1 indicating greater significance. It can be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

In this study, MAE, RMSE, ME and R² are calculated for each of the four marshaller task categories separately to identify which task types are most predictable.

Apart from measuring accuracy, explainability methods provide insight into which input features drive the outputs, and how strongly they contribute to each prediction. A widely used method is SHapley Additive exPlanations (SHAP). SHAP can highlight which variables are the main drivers of predicted marshaller task demand.

TABLE II
FORECASTING PERFORMANCE FOR DOCKING ACROSS HORIZONS

Horizon	**MAE**	**RMSE**	**ME**	R^2
24h	1.360	1.746	0.464	0.428
168h	1.309	1.702	0.328	0.452
2160h	1.315	1.662	0.565	0.437

III. RESULTS

Each model predicts the hourly counts of four task categories: Docking, Follow me service, Meeting, and Engine Testing per forecasting horizon (24, 168, and 2160 hours ahead) Performance is evaluated using the MAE, RMSE, ME, and R^2.

A. Docking performance

Docking is the only task type that shows meaningful predictive structure. The model achieves stable MAE, RMSE, and R^2 values across all three horizons, indicating that patterns in Docking demand can be learned reliably. Table II summarises the results for Docking.

The R^2 values lie between 0.43 and 0.45 for all horizons, showing that a substantial part of the variation in Docking demand is captured. The RMSE ranges between 1.66 and 1.75. The MAE values range from 1.30 to 1.36, meaning that the model is typically off by around one to one and a half tasks per hour. The ME values are positive, indicating a mild overestimation, but the magnitude remains limited.

B. Performance for other tasks types

The predictive capability for Follow me service, Meeting, and Engine Testing is limited. Their R^2 values are close to zero or negative across all horizons, indicating that little of the variation in the hourly counts is explained by the model. The MAE and RMSE values for these tasks are lower than for Docking, but these low values primarily reflect the small absolute number of tasks rather than strong predictive accuracy. The ME values remain close to zero, showing no consistent over- or underestimation. These metrics indicate that the available dataset does not contain sufficient structure to support meaningful hourly forecasting for these task types.

Figure 2 highlights that Docking is the only task type with consistently positive R^2 values across all horizons, while the other tasks are close to zero.

C. Comparison with baseline models

To assess the added value of the LightGBM models, the results were compared against a seasonal naïve baseline and a weekly hourly average baseline. For Docking, the LightGBM model consistently outperforms both baselines across all horizons. The RMSE decreases by approximately 19 - 21% compared to the seasonal naïve baseline, and the R^2 values are substantially higher. For the other task types, the performance of the LightGBM model is similar to or only slightly better than the baselines.

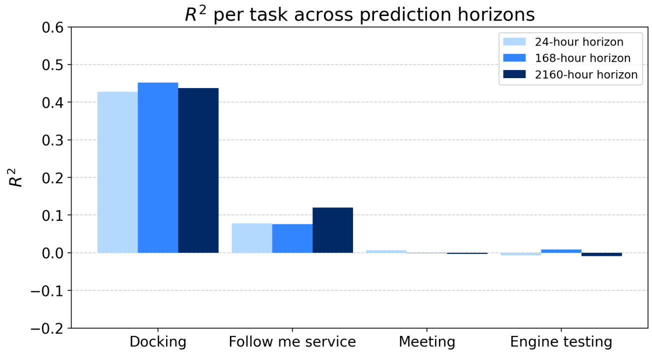


Fig. 2. R^2 values per task across prediction horizons.

D. Feature selection experiment

A reduced-feature model was tested using only the six most influential features per model identified through SHAP analysis. The reduced model perform nearly identically to the full model for Docking across all horizons. This indicates that a small number of temporal features, such as hour of the day and lagged values at 24 and 168 hours, already capture most of the predictive structure.

IV. Discussion

A. Conclusion

The results show that the predicative capability strongly differs between the task type. Docking is the only task that can be forecasted reliably. Docking follows daily patterns and has a clear link with the aircraft arrivals. This can be seen in the historical workload data and can therefore be captured by the model. Docking shows stable performances over all horizons with RMSE values between 1.66 and 1.75 and MAE values between 1.26 and 1.35. The R^2 values range between 0.43 and 0.45, indicating that a substantial part of the hourly variation is captured. The ME values lie between 0.33 and 0.56, showing that the model slightly overestimates the hourly demand on average.

The comparison with the two baselines models shows that the LightGBM models add predictive value for Docking. For the three models the Docking RMSE reduces by approximately 19 - 21% compared to the seasonal naïve, and 3 - 9% compared to the weekly hourly baseline. The R^2 improves by 0.33 - 0.39 for the seasonal naïve baseline and 0.04 - 1.2 for weekly hourly average.

The predictive capability of the other task types is limited. The R^2 values are close to zero showing not much variation can be captured in these patterns. Follow me service and Engine testing have a low error, but this can be explained by the low number of tasks in this dataset. The current inputs and dataset are not able to support hourly forecasting.

Based on these results, the predictive capability of the models is restricted to one task type. With the current data available only Docking can be forecasted reliably.

B. Discussion

The results of this study show that only one task category can be forecasted reliably with the current dataset. Docking follows a clear pattern that is driven by the aircraft arrivals. This pattern is visible in the historic workload data and can therefore be learned by the model. These other tasks do not follow such a structure or have a low number of tasks represented in this dataset. The occurrence could depend more on irregular events that are not captured directly in the available dataset.

The models are evaluated using a chronological 80/20 split. This ensures that future data is not used to predict past events, creating a realistic operational forecasting set-up. A limitation however is that the evaluation is based on the most recent time window. As a result, the performance reflects the conditions present in that specific window, and it is unknown how stable the model would be across other months or seasons. The evaluation may overestimate or underestimate on the true performance. The model performs well on the final window for Docking, but the performance across other periods remains unknown.

The feature selection model experiment showed that a small number of features influence the Docking prediction. Across all horizons, hour is the most influential factor, indicating a day pattern. Lagged Docking values for 24 and 168 hours also appear in the most important features for all horizons. These values reflect daily and weekly repetition in the workload. For the longest horizon month becomes relevant, suggesting that seasonal variation plays a role when the forecast window increases. Showing that Docking follows follows a predictable pattern across days, weeks and months, and explain why it can be forecasted reliably with the available dataset.

The findings suggest that forecasting can support marshaller planning, but only for part of the workload. Docking is predictable and follows clear patterns. This information could help planners identify periods that have a higher expected task demand. However, for the other tasks, accurately forecasting is not possible with the current data. This means that the models only cover a subset of the total workload. Meaning that the operational impact of the models is limited.

This study has some limitations that could affect the results. The dataset in this study only covers a period of nine months. Seasonal effects are only partly represented as not a whole summer and winter season cycle is included. As a result some long-term trends may be missing in the sets. In addition, some operational drivers were also not available. These include handler delays, VDGS malfunctions and other disruptions that influence the workload of marshallers. Because these factors were not directly included as features, but indirectly in the historical workload, the model cannot fully capture their effect on the task demand.

The dataset contains a low number of Follow me service and Engine testing tasks. These tasks provide limited information for the model to learn from and therefore limits the predictive capability of these tasks. There are also smaller tasks that marshallers perform, but for which no data was available. These tasks were therefore not included in the dataset. However, it is likely that they would show a similar effect as Engine Testing and Meeting. Because they also occur infrequently and not often as for example Docking. The model therefore assumes that the four selected task types represent the workload sufficiently, and that excluded smaller tasks do not contribute severely to additional predictable structure.

Several modelling assumptions are made that can have an effect on the results. Each prediction horizon is modelled using a separate LightGBM model. This means that patterns per horizon are learned independently, and no information is shared across horizons. Secondly, the workload is aggregated into hourly task counts. The task assigned to each hour are based on their starting time, even when they only fall partially in that hour. A different rule for assigning the task could have placed that task in the next hour when most of the duration happens there. Lastly, a limited set of drivers is implemented is the models, as some drivers are not available in structured form. Indirectly assuming that their influence is either small or expressed within the historial workload.

The usage of ADS-B data also introduces some limitations. Some raw datapoints are missing, which means that certain activities may not have been recorded. The task labelling depends on the accuracy of the geofence polygons. The polygons have been manually enlarged, and even though this was done with care, it may still affect the accuracy of the matching. At least one area on the airside field exists where the ADS-B signal is not correctly transmitted. Trough visualisation of the trajectories at different days this are was identified. However, it is possible a similar area exists that is not discovered. Resulting in that some movements may be missing in the data. Small deviations in the accuracy of the ADS-B data itself can happen, meaning that the task labelling is not fully accurate and some tasks are missing in the aggregated dataset. Furthermore, even though the labelling rules have been made up with precision, and several days of data were manually checked by comparing trajectories with their segmented task tables, it remains possible that some inaccuracies were not detected.

C. Recommendations for Amsterdam Airport Schiphol

The first step is to extend the dataset with different factors. The current dataset covers nine months. A longer dataset that includes multiple summer and winter seasons would allow more stable patterns to be identified and fully validate longer prediction horizons. In addition, operational drivers should be collected and included, such as VDGS malfunctions, handler delays, and maintenance activities. Including these factors could improve predictability. Furthermore, it is also recommended to record smaller marshaller tasks that are currently not documented. These tasks may show similar behaviour to the low-frequency tasks in this study and should be included in future datasets.

Furthermore, an analysis of the transponder system should be carried out to see if there are any additional areas where the ADS-B data is not submitted correctly. Identifying these areas would help to improve the accuracy of the task labels. It may also be useful to compare vehicle transponder with aircraft transponder data to extract additional Follow me service activities.

The link between forecasting and planning should be explored. The predicted Docking demand could be used to support planning decisions at Schiphol. As the tasks and the duration of tasks have been identified, this can be used in a simulation model. These simulation models could explore the effect of the availability and allocation of marshallers. Centralised and decentralised shifts can be compared and analysed and it can be explored how sensitive the system is to operational disruptions.

D. Recommendations for future research

The method proposed in this study should be tested on other airports, to test the generality of the system. This would show which parts of the approach can be directly transferred. It also shows if the difficulties found in forecasting the Follow me service, Meeting and Engine Testing tasks also appear elsewhere. Such tests would clarify whether the observed limitations are specific to Schiphol or reflect structural characteristics of marshaller operations. This could also indicate to what extent operational drivers differ per airport.

Future research should also examine if alternative forecasting methods can improve the performance for the task types that could not be predicted reliably in this study. These tasks show irregular patterns and occur with low frequency. Methods that are designed for limited event data or methods that predict the probability a task occurs, may give better results. Time series cross validation could also be applied to evaluate the model across multiple time windows in addition to the 80/20 split.

It could also be investigated if hourly forecasting is the most suitable approach. Other resolutions such as 15-minute or 30-minute time intervals, or shift intervals may be relevant. As different rules for assigning tasks to time windows leads to different workload patterns, this option could also be examined to see the influence on the dataset. Finally, the current study uses three separate models, one for each prediction horizon. Future research could explore if the horizons can be combined into one forecasting model.

References

[1] Aeroclass, “Airport operations,” 2021, accessed: 2025-10-01. [Online]. Available: https://www.aeroclass.org/airport-operations/

[2] Copenhagen Optimization, “Airport operations 101,” 2025, accessed: 2025-10-01. [Online]. Available: https://copenhagenoptimization.com/blog/airport-operations-101

[3] Mappedin, “Airport operations management: Definition, types, and how to improve,” 2025, accessed: 2025-10-01. [Online]. Available: https://www.mappedin.com/resources/blog/airport-operations-management-definition-types-and-how-to-improve/

[4] D. Dimitriou, M. Sartzetaki, and A. Karagkouni, “Airport landside area planning: An activity-based methodology for seasonal airports,” *Transportation Research Procedia*, vol. 82, pp. 1167–1184, 2025, world Conference on Transport Research - WCTR 2023 Montreal 17-21 July 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146524004186

[5] B. Zhang, L. Wang, Z. Ye, J. Wang, and W. Zhai, “Evaluating the operational performance of airside and landside at chinese airports with novel inputs,” *Transportation Planning and Technology*, vol. 41, no. 8, pp. 878–900, 2018. [Online]. Available: https://doi.org/10.1080/03081060.2018.1526966

[6] I. Kovynyov and R. Mikut, “Digital technologies in airport ground operations,” *NETNOMICS: Economic Research and Electronic Networking*, vol. 20, no. 1, pp. 1–30, 2019. [Online]. Available: https://doi.org/10.1007/s11066-019-09132-5

[7] T. A. Baki, B. Noordin, N. Mohamed, S. M. Idrus, and S. Z. A. Rasid, “Digitalization of airside operations process to improve airport operations for the case of malaysia airports,” in *2022 4th International Conference on Smart Sensors and Application (ICSSA)*, 2022, pp. 130–134.

[8] C. Malandri, L. Mantecchini, and V. Reis, “Aircraft turnaround and industrial actions: How ground handlers’ strikes affect airport airside operational efficiency,” *Journal of Air Transport Management*, vol. 78, pp. 23–32, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0969699719300183

[9] NOS. (2025) Vliegverkeer twee keer gestaakt op eindhoven airport, vluchten vertraagd. Accessed: 2025-10-01. [Online]. Available: https://nos.nl/artikel/2578549-vliegverkeer-twee-keer-gestaakt-op-eindhoven-airport-vluchten-vertraagd

[10] ——. (2025) Opnieuw meer dan 100 vluchten geannuleerd door klm vanwege staking. Accessed: 2025-10-01. [Online]. Available: https://nos.nl/artikel/2583607-opnieuw-meer-dan-100-vluchten-geannuleerd-door-klm-vanwege-staking

[11] C. Belinda, S. Melwani, and C. Kapadia, “Breaking boredom: Interrupting the residual effect of state boredom on future productivity,” *Journal of Applied Psychology*, vol. 109, no. 6, p. 829 – 849, 2024. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85188520892&doi=10.1037%2fapl0001161&partnerID=40&md5=2e3880a0260c9163683acedb568e8730

[12] M. S. Young, “In search of the redline: Perspectives on mental workload and the ‘underload problem’,” in *Human Mental Workload: Models and Applications*, L. Longo and M. C. Leva, Eds. Cham: Springer International Publishing, 2021, pp. 3–10.

[13] S. Ali and Y. Farooqi, “Effect of work overload on job satisfaction, effect of job satisfaction on employee performance and employee engagement (a case of public sector university of gujranwala division),” 04 2014.

[14] Centraal Bureau voor de Statistiek (CBS). (2024) Hogere werkdruk belangrijkste gevolg personeelstekort volgens ondernemers. Accessed: 2025-10-01. [Online]. Available: https://www.cbs.nl/nl-nl/nieuws/2024/35/hogere-werkdruk-belangrijkste-gevolg-personeelstekort-volgens-ondernemers

[15] D. Dinis and A. P. Barbosa-Póvoa, “On the optimization of aircraft maintenance management,” *Studies in Big Data*, vol. 15, p. 49 – 57, 2015. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85132872165&doi=10.1007%2f978-3-319-24154-8_7&partnerID=40&md5=e5441b3385e53e49508a0e7b43326719

[16] Q. Deng, B. F. Santos, and W. J. Verhagen, “A novel decision support system for optimizing aircraft maintenance check schedule and task allocation,” *Decision Support Systems*, vol. 146, 2021. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102879361&doi=10.1016%2fj.dss.2021.113545&partnerID=40&md5=136eed0b53b45e41365fb459ad07748d

[17] C.-L. Wu and S. X. Lim, “Effects of enterprise bargaining and agreement clauses on the operating cost of airline ground crew scheduling,” *Journal of Air Transport Management*, vol. 91, 2021. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096000602&doi=10.1016%2fj.jairtraman.2020.101972&partnerID=40&md5=0032df26fc22741f5c1839b4b9ac87ff

[18] X. Wen, S.-H. Chung, P. Ji, and J.-B. Sheu, “Individual scheduling approach for multi-class airline cabin crew with manpower requirement heterogeneity,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 163, 2022. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85131405716&doi=10.1016%2fj.tre.2022.102763&partnerID=40&md5=e826f9a25a499bfc79159f5ae90f7bad

[19] G. Dall’Olio and R. Kolisch, “Formation and routing of worker teams for airport ground handling operations: A branch-and-price-and-check approach,” *Transportation Science*, vol. 57, no. 5, pp. 1231–1251, 2023. [Online]. Available: https://doi.org/10.1287/trsc.2022.0110

[20] M. Mamdouh, M. Ezzat, and H. Hefny, “Optimized planning of resources demand curve in ground handling based on machine learning prediction,” *I.J. Intelligent Systems and Applications*, vol. 13, no. 1, pp. 1–16, 2021. [Online]. Available: https://www.mecs-press.org/ijisa/ijisa-v13-n1/IJISA-V13-N1-1.pdf

[21] X. Wen, S.-H. Chung, H.-L. Ma, and W. A. Khan, “Airline crew scheduling with sustainability enhancement by data analytics under circular economy,” *Annals of Operations Research*, vol. 342, no. 1, p. 959 – 985, 2024. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85158099233&doi=10.1007%2fs10479-023-05312-7&partnerID=40&md5=5a6b68f8446df50151b45bb3d5ec1e46

[22] R. Rabet, S. M. Sajadi, and M. Tootoonchy, “A hybrid metaheuristic and simulation approach towards green project scheduling,” *Annals of Operations Research*, 2024. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85208584292&doi=10.1007%2fs10479-024-06291-z&partnerID=40&md5=ee0d7db43e8057a932f10c5118356d24

[23] A. O. Adeniran and O. A. Kanyio, “Artificial intelligence in aircraft docking: the fear of reducing ground marshalling jobs to robots and Way-Out,” Tech. Rep., 2018. [Online]. Available: www.cribfb.com/journal/index.php/aijmsr

[24] A. A. O and A. F. O, “Acceptance of visual docking guidance system by ground marshallers in Nigerias’ airport,” *International Journal of Advanced Networking and Applications*, vol. 13, no. 01, pp. 4845–4854, 1 2021. [Online]. Available: https://doi.org/10.35444/ijana.2021.13107

[25] P. Šváb, P. Korba, S. Al-Rabeei, M. Tirpáková, and J. Hura, “Innovations in the field of aircraft ground handling,” 11 2021, pp. 159–162.

[26] C. Choi, J.-H. Ahn, and H. Byun, “Visual recognition of aircraft marshalling signals using gesture phase analysis,” in *2008 IEEE Intelligent Vehicles Symposium*, 2008, pp. 853–858.

[27] Y. Song, D. Demirdjian, and R. Davis, “Tracking body and hands for gesture recognition: Natops aircraft handling signals database,” in *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2011, pp. 500–506.

B

Swimlane diagram

See next page

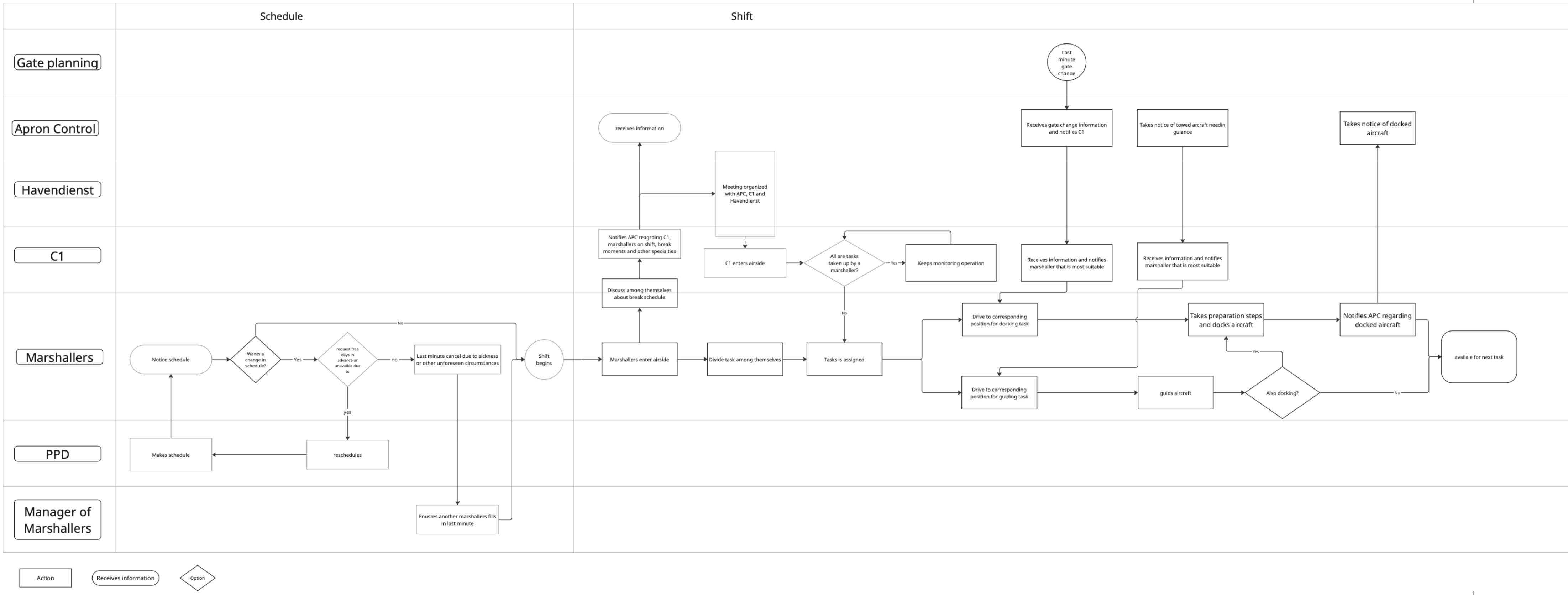

Schedule
Shift
Gate planning
Apron Control
Havendienst
C1
Marshallers
PPD
Manager of Marshallers
Notice schedule
Wants a change in schedule?
Yes
No
request free days in advance or unavaible due to
no
yes
Last minute cancel due to sickness or other unforeseen circumstances
Shift begins
Makes schedule
reschedules
Enusres another marshallers fills in last minute
Marshallers enter airside
Discuss among themselves about break schedule
Notifies APC reagrding C1, marshallers on shift, break moments and other specialties
receives information
Meeting organized with APC, C1 and Havendienst
C1 enters airside
All are tasks taken up by a marshaller?
Yes
Keeps monitoring operation
No
Divide task among themselves
Tasks is assigned
Last minute gate change
Receives gate change information and notifies C1
Receives information and notifies marshaller that is most suitable
Takes notice of towed arcraft needin guiance
Receives information and notifies marshaller that is most suitable
Drive to corresponding position for docking task
Drive to corresponding position for guiding task
Takes preparation steps and docks aircraft
guids aircraft
Also docking?
Yes
No
Notifies APC regarding docked aircraft
Takes notice of docked aircraft
availale for next task
Action
Receives information
Option

Figure B.1: Swimlane diagram

C

List of holiays used in the LightGBM models

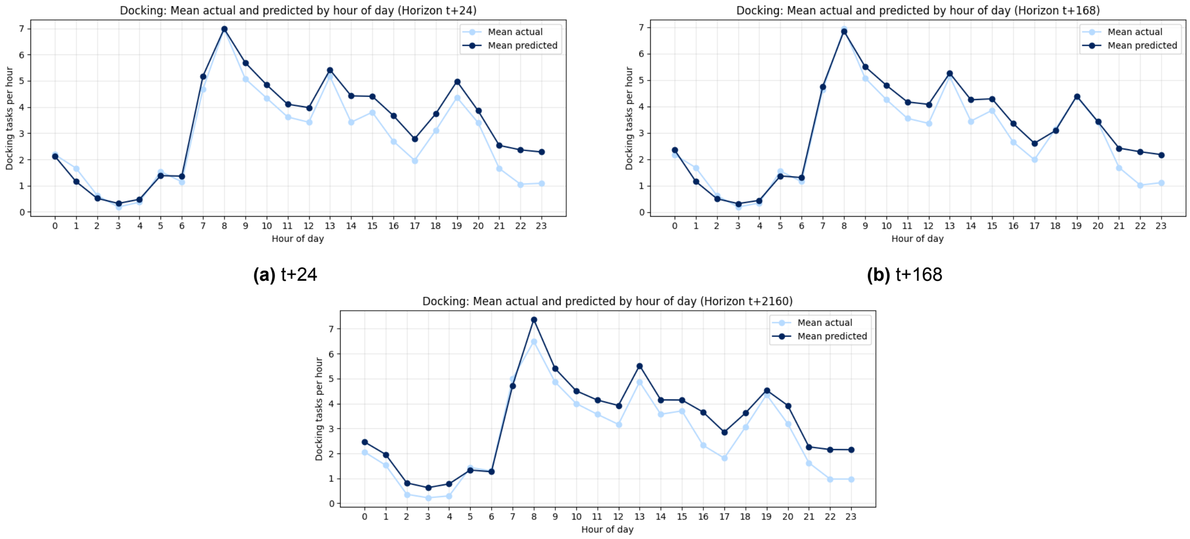
Table C.1: Overview of Dutch public holidays and school vacations December 2024 - January 2024

Date	Holiday / Vacation
25 December 2024	First Day of Christmas
26 December 2024	Second Day of Christmas
1 January 2025	New Year’s Day
18 April 2025	Good Friday
20 April 2025	First Day of Easter
21 April 2025	Second Day of Easter
26 April 2025	King’s Day celebrated
5 May 2025	Liberation Day
29 May 2025	Ascension Day
8 June 2025	First Day of Pentecost
9 June 2025	Second Day of Pentecost
21 December 2024 - 5 January 2025	Christmas Holiday 2024 - 2025
15 February - 23 February 2025	Spring Holiday (South)
22 February - 2 March 2025	Spring Holiday (North/Mid)
26 April - 11 May 2025	May Holiday
5 July - 31 August 2025	Summer Holiday

D

Mean observed and predicted volume tasks graphs

D.1. Docking tasks



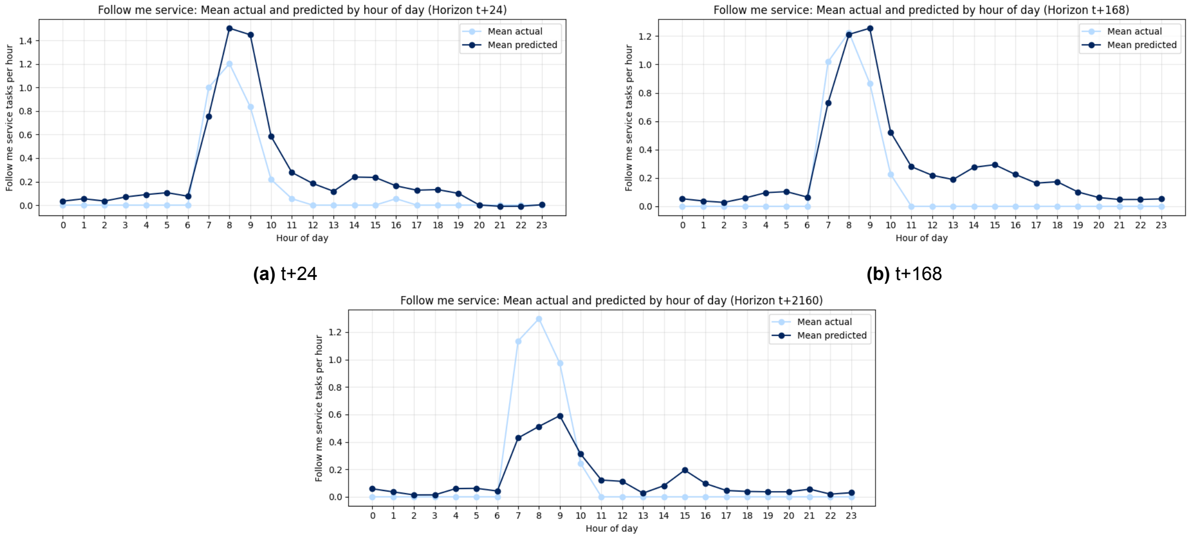
(a) t+24

(b) t+168

(c) t+2160

Figure D.1: Mean observed and predicted docking volume per hour of day for the three forecasting horizons.

D.2. Follow me service tasks



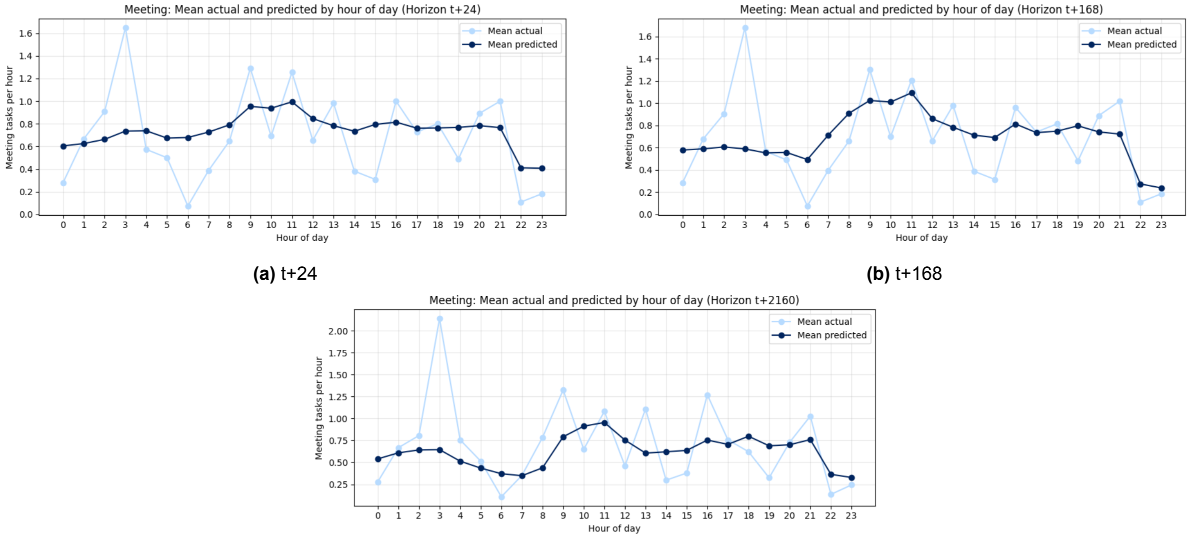
(a) t+24

(b) t+168

(c) t+2160

Figure D.2: Mean observed and predicted Follow me service volume per hour of day for the three forecasting horizons.

D.3. Meeting tasks



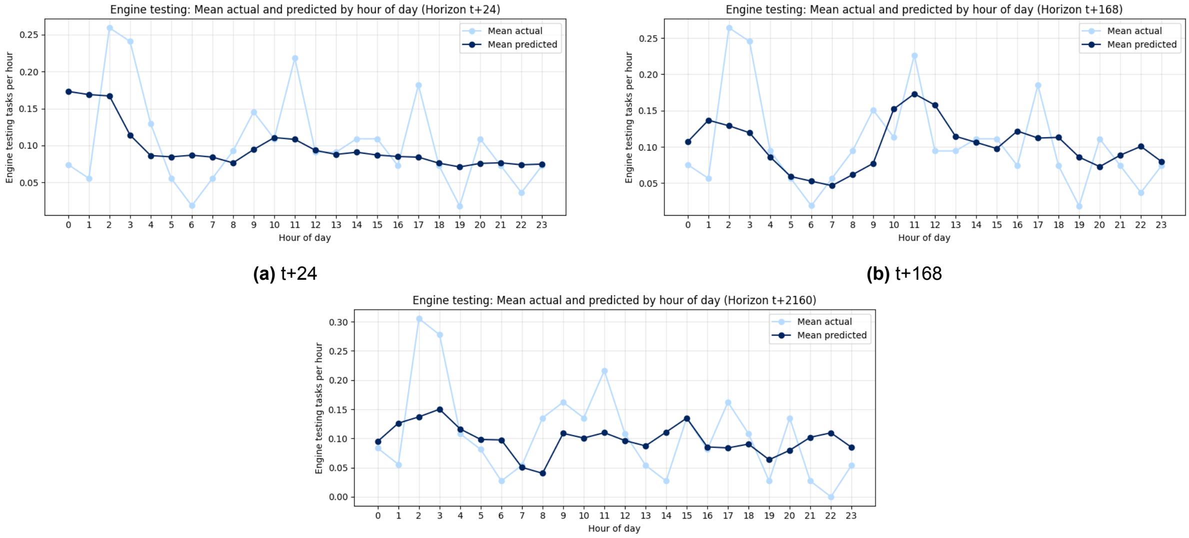
(a) t+24

(b) t+168

(c) t+2160

Figure D.3: Mean observed and predicted Meeting volume per hour of day for the three forecasting horizons.

D.4. Engine testing tasks



(a) t+24

(b) t+168

(c) t+2160

Figure D.4: Mean observed and predicted Engine testing volume per hour of day for the three forecasting horizons.

E

Validation results baseline

Table E.1: Baseline comparison for Follow me service across prediction horizons

Horizon	**Method**	**MAE**	**RMSE**	**ME**	R^2
24h	Seasonal naive	0.223	0.894	0.008	-0.571
	Weekly avg.	0.575	0.920	-0.455	-0.750
	LightGBM	0.315	0.668	0.124	0.078
168h	Seasonal naive	0.215	0.882	0.005	-0.563
	Weekly avg.	0.571	0.916	-0.455	-0.744
	LightGBM	0.327	0.667	0.123	0.076
2160h	Seasonal naive	0.218	0.883	-0.017	-0.695
	Weekly avg.	0.551	0.912	-0.406	-0.551
	LightGBM	0.245	0.687	-0.026	0.120

Table E.2: Baseline comparison for Meeting across prediction horizons

Horizon	**Method**	**MAE**	**RMSE**	**ME**	R^2
24h	Seasonal naive	1.064	2.986	-0.010	-1.024
	Weekly avg.	1.152	2.114	-0.499	-0.108
	LightGBM	0.904	2.002	0.054	0.007
168h	Seasonal naive	1.072	3.018	-0.007	-1.024
	Weekly avg.	1.155	2.134	-0.490	-0.108
	LightGBM	0.892	2.028	0.015	-0.001
2160h	Seasonal naive	1.151	3.491	0.008	-1.072
	Weekly avg.	0.973	2.297	-0.150	-0.025
	LightGBM	0.876	2.272	-0.078	-0.003

Table E.3: Baseline comparison for Engine testing across prediction horizons

Horizon	Method	MAE	RMSE	ME	R^2
24h	Seasonal naive	0.188	0.470	0.010	-0.836
	Weekly avg.	0.182	0.345	0.003	-0.017
	LightGBM	0.181	0.344	-0.007	-0.007
168h	Seasonal naive	0.188	0.469	0.008	-0.824
	Weekly avg.	0.182	0.344	0.004	-0.013
	LightGBM	0.184	0.341	-0.002	0.009
2160h	Seasonal naive	0.198	0.487	0.017	-0.760
	Weekly avg.	0.184	0.360	0.009	-0.040
	LightGBM	0.185	0.354	-0.008	-0.009

F

Feature importance

Table F.1: Feature importances for the 24-hour prediction horizon

Rank	Feature	Importance	Cum. Importance	Cum. Percent (%)
1	Hour	0.9627	0.9627	34.68
2	Docking_lag168	0.3371	1.2998	46.83
3	Docking_lag24	0.3031	1.6029	57.75
4	Docking_roll1	0.2939	1.8967	68.34
5	Docking_roll24	0.0958	1.9925	71.79
6	Day_of_week	0.0893	2.0819	75.01
7	Followme_roll168	0.0721	2.1540	77.61
8	Docking_roll3	0.0635	2.2174	79.89
9	Docking_lag1	0.0554	2.2728	81.89
10	Incoming_lag168	0.0335	2.3063	83.09
11	Followme_roll6	0.0285	2.3348	84.12
12	Docking_roll168	0.0280	2.3627	85.13
13	Docking_lag12	0.0249	2.3876	86.02
14	Meeting_roll24	0.0243	2.4119	86.90
15	Meeting_lag3	0.0225	2.4345	87.71

Table F.2: Feature importances for the 168-hour prediction horizon

Rank	Feature	Importance	Cum. Importance	Cum. Percent (%)
1	Hour	0.5873	0.5873	21.85
2	Docking_roll1	0.5071	1.0944	40.72
3	Docking_lag24	0.3904	1.4848	55.25
4	Docking_lag168	0.3465	1.8312	68.14
5	Followme_roll6	0.2009	2.0321	75.62
6	Followme_roll168	0.0561	2.0883	77.71
7	Docking_lag6	0.0436	2.1318	79.33
8	Docking_roll3	0.0387	2.1705	80.77
9	Docking_roll6	0.0360	2.2066	82.11
10	Docking_roll168	0.0297	2.2362	83.21
11	Meeting_roll168	0.0272	2.2634	84.23
12	Docking_lag1	0.0240	2.2874	85.12
13	Incoming_lag168	0.0239	2.3113	86.01
14	Docking_roll24	0.0237	2.3350	86.89
15	Incoming_roll168	0.0214	2.3564	87.68

Table F.3: Feature importances for the 2160-hour prediction horizon

Rank	Feature	Importance	Cum. Importance	Cum. Percent (%)
1	Hour	1.3418	1.3418	47.16
2	Incoming_lag168	0.1501	1.4919	52.43
3	Month	0.1334	1.6253	57.12
4	Docking_lag24	0.1244	1.7497	61.49
5	Docking_roll1	0.1165	1.8662	65.59
6	Docking_lag168	0.0717	1.9379	68.11
7	Incoming_lag24	0.0583	1.9962	70.16
8	Docking_roll6	0.0557	2.0519	72.11
9	Followme_roll168	0.0525	2.1044	73.96
10	Docking_lag1	0.0438	2.1482	75.50
11	Day_of_week	0.0378	2.1860	76.83
12	Docking_roll168	0.0364	2.2223	78.10
13	Incoming_roll168	0.0357	2.2580	79.36
14	Incoming_roll24	0.0339	2.2920	80.55
15	Docking_roll3	0.0332	2.3251	81.72