Integrating Predictive and Optimization Model for Intelligent Schedule Management based on Real Time ETA Information

A Concept of Machine Learning and Exact Solution in Petrochemical Loading Facility

Master Thesis
Emanuel F Prakoso

Integrating Predictive Model and Optimization Model for Intelligent Schedule Management based on Real Time ETA Information

A Concept of Machine Learning and Exact Solution in Petrochemical Loading Facility

by

Emanuel Febrianto Prakoso

to obtain the degree of

Master of Science in Transport, Infrastructure, and Logistics

at the Delft University of Technology,

to be defended publicly on 6 October 2021

Student number	:	5093651
Project duration	:	12 April 2021 – 6 October 2021
Thesis Committee	:	

Prof.dr.ir. L.A. (Lóránt) Tavasszy	–	Chairman, Delft University of Technology
Dr. M.Y. (Yousef) Maknoon	–	Supervisor, Delft University of Technology
Dr.ir. A.J. (Adam) Pel	–	Supervisor, Delft University of Technology
Dr. R. (Ratnaji) Vanga	–	Mentor, Delft University of Technology

An electronic version of this thesis is publicly available at https://repository.tudelft.nl/.

Table of Contents

Preface

This thesis is the culmination of almost 6 months of intense work and roughly 2 years of constant struggle to obtain my Master of Science degree in Transport, Infrastructure, and Logistics at the Delft University of Technology. Fortunately, I had the chance to work on a topic that thoroughly embodies my vision and interest in integrating advanced technology to solve problems in the transport and logistic sector. When I decided on this topic, it was quite a gamble for me because I didn't possess all the required knowledge and skill to effortlessly achieve the goal, but in the end, I could safely say that hard work and enthusiasm do make up for a host of deficiencies.

I'd like to express my gratitude to my thesis committee for getting on board from the very beginning and guiding me through the end. Firstly, to Prof. Lori, for introducing this interesting topic, and providing clear solutions for all my problems with a remarkable response time. His presence represents the calmness that I truly need to navigate through this chaotic period of time. Secondly, to Dr. Yousef who has helped me establish a strong theoretical foundation to develop the models. I am grateful because he has always been able to directly point out the potential issues and then efficiently suggest better ways to formulate the solutions. Moreover, for Dr. Adam's pivotal role, especially in writing academic reports that I lack experience with. Because of him, I am able to compose a more coherent structure of report and a better flow of thinking to present the result. Last, but not least, to my mentor, Dr. Ratnaji, who has always made time for the weekly brainstorming sessions. His support in explaining the practical problem, discussing technical aspects and potential solutions, finding data, and checking words could not be overstated. Without all of them, the outcomes of this thesis would not have been the same.

Lastly, I'd like to extend my gratitude and appreciation to anyone who has played a part – regardless of the magnitude – in my memorable journey here. Particularly, I wish to deeply thank God, my mother, father, all the family, Stefani, and all my friends in Indonesia and in the Netherlands for all the unconditional prayers, supports and loves, for all the motivations that keep me believing that I could finish what I started, and for all priceless moments that render life worth living. You all made it possible for me.

Delft, September 2021

Emanuel

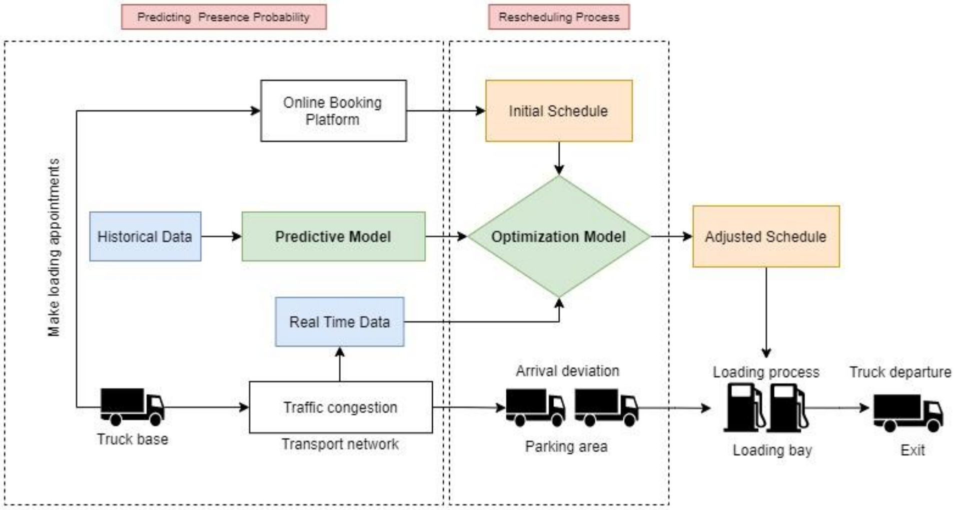
Executive Summary

Introduction

Digitalization is a key facilitator in the shift to a more sustainable and efficient logistics/freight transportation sector. One important element is the integration between logistics and traffic systems. Although the two are seemed as two different domains, in reality, they might benefit from a more integrated strategy. Essentially, both sectors face the congestion and time uncertainty issue on a regular basis. Nowadays, given unprecedented growth in logistic demand and significant shift in traffic behavior, it is imperative to unlock the potential of an integrated system. The potential extends far beyond the existing practice of providing traffic information to logistic providers. The rapid development of ICT (Information and Communication Technology) enables more intelligent utilization such as arrival/travel time prediction, dynamic routing, or adaptive planning.

In the perspective of a petrochemical loading facility that risks a major disruption of its system due to traffic congestion, the real-time information of arriving trucks could be utilized to optimize the operation. Especially in Antwerp region where there is an ongoing construction work on the ring road, truck arrival time is becoming more uncertain. In the absence of ICT infrastructure that connects the facility planning and traffic information, detrimental consequences are inevitable. Therefore, a solution based on ICT environment that integrates traffic system and loading operation is necessary to mitigate the uncertainty variables.

This research aims to investigate the impact of incorporating real-time information technology to improve truck schedule management. This potentially enables intelligent rescheduling as reactive mitigation against the real-time disturbance ahead of time. The outcome will be assessed with the perspective of the current circumstances, in which ICT is unavailable. Furthermore, another goal is to observe the added value of integrating a predictive model and optimization model that considers stochastic variables. To do so, this research is divided into 3 main parts, namely, designing a conceptual framework of advanced schedule management, developing a predictive model, and developing an optimization model. A loading facility owned and operated by [redacted] located in Antwerp is chosen as the use-case sample of this study, although the outcome is not intended as a tailor-fit solution for their particular problems.



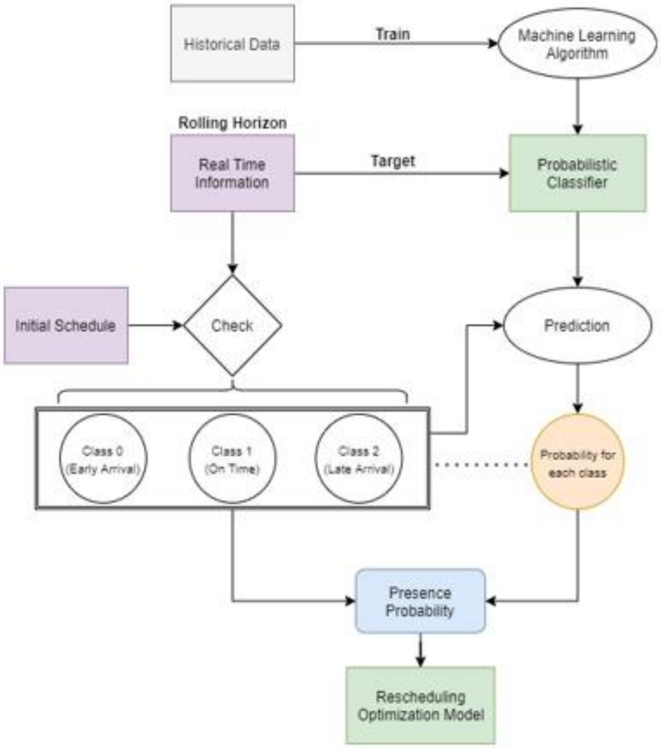
Knowledge Gap

To author's best knowledge, there are no prior studies that incorporate a predictive model based on ML (Machine Learning) algorithms in the domain of a loading facility. Furthermore, while numerous researches have been done to integrate stochastic arrival time in optimal schedule management, there are few that attempt to solve it analytically. As a result, this research will explore the potential of schedule management that combines a prediction powered by ML algorithm as well as a probabilistic optimization model with exact solution.

Conceptual Framework

The uncertainty of truck arrival time is not fully eliminated by real-time information. Therefore, to acknowledge this issue, a presence probability value that denotes the likelihood of trucks being early, on-time, and late is assigned to the corresponding Estimated Time Arrival (ETA). Given the unpredictability of the data, a predictive model based on ML algorithm is an excellent choice for making predictions. The ML algorithm is capable of detecting subtle patterns/trends in historical data as well as comprehending non-linear and complicated relationships between parameters. According to the prediction target, the classification algorithm will be used as the base of predictive model. Because there is no exact guideline for identifying the optimal ML algorithm, three alternative classifier algorithms will be tested and evaluated: Gaussian Naïve Bayesian (GBN), Logistic Regression (LR), and ANN (Artificial Neural Network).

Based on all obtained information, a conceptual framework of rescheduling system in which the main components are the integration between predictive (probabilistic classifier powered by ML algorithm) and optimization model (exact solution for rescheduling) is proposed. Historical data and real-time information are the two major inputs. Based on the real-time ETA, this conceptual framework allows determining if a truck will be able to adhere to its initial schedule, and therefore whether a truck will be categorized as early, on-time, or late. The prediction model would assign a presence probability value for that specific arrival class based on the pattern in the historical data, which would then be the input to the optimization model.



Predictive Model

Data exploration is the first phase in developing a predictive model, in which a huge unstructured and synthetic historical dataset is thoroughly analyzed for hidden patterns, distinctive traits, and significant points of interest. This procedure achieves its objectives, which are to give useful insights and to verify the validity of synthetic data in terms of logical sense and existing literature. As a result, it may be stated that synthetic data is adequate as a basis for developing a predictive model.

It was chosen to investigate three distinct algorithms: GBN, LR, and ANN. The assessment findings reveal that the ANN is the best algorithm for fitting the input historical data, with an overall F1 score of 71 %, according to the standardized KPI. In all KPI measures, the ANN model beats the LR and GBN models by an average of 5% and 20%, respectively, when compared to other evaluated models. To benchmark the result of the predictive models, a comparison with other predictive models found in similar studies is conducted. The **ANN's accuracy** of **70%** is sufficient and competitive compared to the predictive models proposed in similar studies whose accuracy approximately ranges from 40% - 80%. Hence, it can be determined the most optimal predictive model to use in this study is the ANN model, which serves the first aim of this chapter.

Two scenarios are developed to assess the extent to which real-time information improves the prediction model. The first depicts a situation in which historical data only contains conventional information, and the second depicts a situation in which a parameter of real-time information is included. The findings clearly show that adding a feature provided by the **inclusion of real-time information** improves the prediction model by about **20%**. Therefore, it emphasizes the importance of ICT infrastructure to unlock intelligent logistic planning.

Optimization Model

The P-SRP (Probabilistic Slot Rescheduling Problem) model is built on the concept of MIQP (Mixed Integer Quadratic Programming) and expected value as the way to accommodate incorporating stochastic variables in a simpler manner. As an implication, the model applies expected presence probability rather than binary presence when assigning trucks to specific loading slots. In other words, this model recognizes the uncertainty of arrival time by assigning a presence probability value to each arrival information. Consequently, as the main benefit, the system allows simultaneous assignment based on the accumulated presence probability of trucks in a given period. This approach might be employed to reduce the additional interference produced by trucks that are unable to fulfill their initial schedule. This notion is implemented in an objective function that minimizes the expected cost that is formulated as the gap between the adjusted and initial schedule multiplied by the generalized cost of rescheduling.

Verification

The proposed rescheduling system is completely applied and verified on a synthetic use-case sample. The utilization of ETA yields a significant advantage, whereby it allows rescheduling to be done ahead of time. To measure the benefit of the P-SRP model, a baseline model that considers ETA as a deterministic variable is introduced for comparison. The main feature of the P-SRP is a more intelligent rescheduling, in which it allows simultaneous assignment based on the value of presence probability. The results clearly highlight the shortcoming of the baseline model in terms of preventing the occurrence of idle slots and unnecessary schedule shifts. Consequently, the P-SRP results on **expected cost** that is **42% lower** than the baseline model.

Lastly, the proposed rescheduling system is tested on a dynamic experiment based on a rolling horizon to simulate the actual implementation and continuous update of real-time ETA data. Due to the limitation in this research, only 2 iterations with a period of 2 hours are simulated. The results showcased the superiority of the P-SRP model relative to baseline model. On top of the higher operational efficiency, the analysis indicates that P-SRP provides robustness against uncertain arrival time, whereby the adjusted schedule could withstand and anticipate the fluctuation of actual arrival of incoming trucks. Hence, it can be concluded that the main advantage of the utilization of the predictive model and P-SRP proposed in this research is a more intelligent rescheduling process, in which operational efficiency is improved without radically compromising schedule robustness. However, the results might be case-specific considering the limited data and experiments.

Numerical Experiment

Numerical experiments that consists of comparative study and sensitivity analysis are conducted to investigate the added values of P-SRP over the baseline model. To this end, a set of scenarios are developed to analyze the performance. Each scenario corresponds to different contexts relevant to the operation of the loading facility, namely congestion level in transport network, specification of loading infrastructure, scheduling policy, and the maximum permissible probability. It implies that the performance will be examined in such a way that the benefits and drawbacks of a certain technique will be highlighted in a given context.

Scenarios of varied congestion levels indicate that P-SRP has the most significant impact on medium congestion level, shown by reduced expected cost by averagely **52.07%** and reduced risk of idled by averagely **52.23%**. Interestingly, it can be deduced applying probabilistic in medium congestion level yields the best result in mitigating the potential of idled slot, because it demands more intensity of rescheduling (due to early and late arrival) than severe congestion level, not precisely because of the parameter congestion level itself. The scenario of varied configuration and specification points out that P-SRP has the most significant edge over the baseline model in the largest scale of operation (100% utilization rate and 4 loading bays), whereby it reduces the expected cost by averagely **54.84%**. Lastly, in the scenario of varied rescheduling strategy, the P-SRP model offers the best performance in situations without stand-in loading bays and the priority is to maintain the initial schedule, in which it decreases the expected cost by **51.39%** compared to the baseline. However, in the strategy of minimizing waiting time and there is stand-in loading bay, the P-SRP is not impactful, whereby it only reduces the expected cost by **15.22%** compared to the baseline model.

In addition, trade-off analysis is conducted to explore the behavior between operational efficiency and schedule robustness. Although the impact is not significant, the P-SRP model could balance the trade-off between schedule robustness and operational efficiency by altering the maximum value of permissible overlap. It is discovered that increasing the value from **5% to 20%** results in the most substantial increase in operational efficiency. The most optimum trade-off, on the other hand, cannot be established because it is dependent on decision-maker's preferences.

Conclusion

This study shows that it is conceptually possible to improve the operational efficiency of a petrochemical loading facility by having an integrated system between logistic operation and traffic system. It proposes advanced schedule management that allows intelligent rescheduling. The main elements of the rescheduling system are the predictive and optimization model that integrates the real-time information and historical data of truck arrival. The implementation potentially provides valuable managerial insights for operational planning in terms of mitigating the propagation effect of arrival uncertainty and manpower/facility resources allocation.

This research contributes to filling the stated knowledge gaps. Regarding the first knowledge gap which is the lack of exact method for capturing uncertainty in the operation of loading facility, this study shows that considering a stochastic arrival time in the form of discrete probability could improve the operational efficiency of loading facility. Concerning the second gap, which is the lack of predictive model based on ML algorithm, this study shows that predicting arrival class (early, on-time, late) is possible and beneficial in maintaining a high-performance operation. The study is conducted in a more general sense; therefore the proposed conceptual framework is not supposed to tackle a tailor-fit problem. Given the benefit of the generality of this study, the models and solutions presented in this paper are adaptable for other similar operational problems. Specifically, in a situation where the assignment is object-specific, unlike the ones in existing literature where the assignment process is in universal sense.

Limitation & Recommendation

This research is established on a set of assumptions and simplification which might lead to imperfect outcomes. Nevertheless, this does not disprove the fact several principal insights might still be valid and relevant. The limited-time results on inability to obtain real data and to conduct real observation at the loading facility, enforcing the proposed conceptual framework to be tested on synthetic data which might be slightly less representative of a real-world problem. In regard to the development of predictive model, computational power is a crucial element. This factor directly corresponds to the capability of having ANN with more layers that could potentially generate a more valid and reliable prediction. Regarding the development of the optimization model, a lack of specific facility data and access to the managerial personnel are the main concern. It results in inability to incorporate full variables relevant to the operation of the loading facility, for instance, the generalized cost is limited to only the deviation between the initial and adjusted schedule because there is no additional information to determine the weight/penalty of other factors that are essential in efficiency measure.

According to the main findings and acknowledged limitations of this research, a set of recommendations could be inferred. For instance, applying other assessment methods (CBA or MCDA) to find the tangible trade-off, considering operational and departure as stochastic variables to allow dynamic or adaptive length of loading slots, adding more features of parameters in predictive model (weather condition, type of road disruption, occurrence of accidents, routes taken by trucks, time record at series of GPS coordinate, etc.), conducting hyperparameter-tuning to get a more precise prediction, and exploring the potential of stochastic programming that includes multiple recourse actions and set of scenarios to capture uncertainty.

List of Abbreviation

AI	=	Artificial Intelligence
ANN	=	Artificial Neural Network
DL	=	Deep Learning
DSS	=	Decision Support System
EET	=	Estimated Elapsed Time
ETA	=	Estimated Time of Arrival
ETD	=	Estimated Time of Departure
GBN	=	Gaussian Naïve Bayesian
ICT	=	Information and Communication Technology
k-NN	=	k- Nearest Neighborhood
LR	=	Logistic Regression
MIP	=	Mixed Integer Programming
MILP	=	Mixed Integer Linear Programming
MIQP	=	Mixed Integer Quadratic Programming
ML	=	Machine Learning
P-SRP	=	Probabilistic Slot Rescheduling Problem
ReLU	=	Rectified Linear Activation
RF	=	Random Forest
SAP	=	Slot Assignment Problem
SP	=	Stochastic Programming
STA	=	Scheduled Time of Arrival
STD	=	Scheduled Time of Departure
SVR	=	Support Vector Regression
TAP	=	Truck Appointment System

List of Figures

List of Tables

1. Introduction

This chapter aims to provide justification and motivation of conducting this research. A general explanation on the inextricable relation between supply chain and transport network, the inherent problems that must be addressed, and the importance of having improvements in this domain considering the subsequent impacts in the bigger picture will be presented. Then, it will be followed by a more specific description on the problem case of this problem. Lastly, a research definition that includes research objectives, scope, limitation, research questions, and methodology will be described.

1.1. Background

1.1.1. Transportation Role in Supply Chain

Supply chain is highly connected network, whereby irregularities occurring in a certain point could cause disruptions on downstream events [1]. Transportation are one of the most common factors of disruption in the supply chain that contribute to almost one-third of total operational cost [2]. The significant number indicates the importance of transportation and logistics factors in the whole supply chain; hence it highlights the urgency for optimization. In petrochemical business, efficient and robust downstream supply chain system are important factors to achieve competitive edge [3]. One of the most important actors is the loading facility because of its role as a linkage between production and distribution [4]. Although its importance, loading facility is highly vulnerable to transport disruption.

In the past few years, freight traffic in Antwerp region has exponentially increased due to its strategic location, and the proliferation of freight traffic in the past few years resulted on average congestion level of 25% in 2020 [5], [6]. According to data observation of [redacted] loading facility in Antwerp, the efficiency of loading operation has been enduring negative impact of increasing congestion. This situation might be exacerbated because of the government's constructing plan of ring road in Antwerp that likely results on the significantly higher traffic congestion during the construction period [4].

1.1.2. Importance of Schedule Management

In operational level of loading facility, the performance of loading facility is dependent on truck arrival time. The main concern is that, under stochastic environment, the trucks arrive at the loading facility with time uncertainty as the result of complex transport network problem, therefore deviation in forms of early or late arrival is inevitable [7]. Consequently, due to this disruption, operational efficiency of loading facility will be reduced, leading to various severe problems in different perspective. In terms of business, it increases total operation cost. Moreover, it also limits the space, resulting on higher risk of accidents and lower productivity of the site. In environmental perspective, excessive queued trucks emit pollutant gas that could endanger the surrounding livelihood [8].

According to [8], [9], and [10] the standard approach to mitigate the negative impacts of uncertainty in arrival time is implementing Truck Appointment System (TAS). Basically, TAS is defined as a platform for truck companies to request appointment before the actual arrival to streamline the arrival flow of incoming trucks. However, basic TAS system is not entirely sufficient at eliminating inefficiency in petrochemical loading facility, especially when rescheduling as reaction of unexpected events is required [11]. Hence, this shortfall emphasizes the urgency of having a better-optimized schedule management in loading facility that considers stochastic variables. However, as observed in [12], literatures in truck rescheduling is limited in which most of the studies disregard the stochastic nature, thus only solve the problem to a limited extent in practice

1.1.3. Stochastic Arrival Time

Rescheduling heavily relies on the information of truck arrival time. To a certain extent, gradual growth in ICT (Information and Communication Technology) that incorporates continuous update and real-time elements could enable more accurate ETA (Estimated Time of Arrival) data [13]. Utilization of this technology results on ability to adjust the schedule ahead of time to anticipate the potential disordering. Nonetheless, it still could not provide a perfect information of the truck arrival time because to many variables could affect the estimation.

Recently, predictive model powered by ML (Machine Learning) algorithm has been a popular topic of research in the transportation and logistics domain, namely in airport operation [14] [15] [16] [17] [18], railway operation [19], and bus operation [20]. All the stated studies have proved that applying Machine Learning algorithm to address the uncertainty in arrival time yield promising result to a certain extent. This approach has potential to significantly enhance the arrival prediction due to its ability to identify hidden pattern/structure embedded in historical data.

1.2.Specific Problem

In a typical petrochemical loading facility owned and operated by [redacted] in Antwerp, the main flow of loading process can be described according to [11] and illustrated in Figure 1.1. There are two main components that can be considered as a series of queue system which are the parking zone and facility. The purpose of parking area is where the incoming trucks will be waiting until their scheduled loading slot is available. In between the parking area and facility area, there is a gate that acts as regulator in determining which trucks to enter the facility area. A decision-maker controls the inflow of trucks into the facility area based on the initial schedule or the adjusted schedule as the result of stochastic arrival. Upon checking process, the trucks proceeds to the loading bays inside the facility area the where the trucks will queue according to its assigned loading bay.

Currently, there is obvious gap in the truck schedule management applied in the specific loading facility. Classic approach of TAS (Truck Appointment System) is applied in which the trucks could book loading slots via online platform. Moreover, in any unexpected situation where truck cannot comply with the initial schedule because of arrival deviation, then rescheduling is done on the spot. As the result, replacement for the initial loading slots cannot be guaranteed. This practice indicates huge lag in flow of information and inefficient rescheduling strategy against disturbance such as early or late arrival.

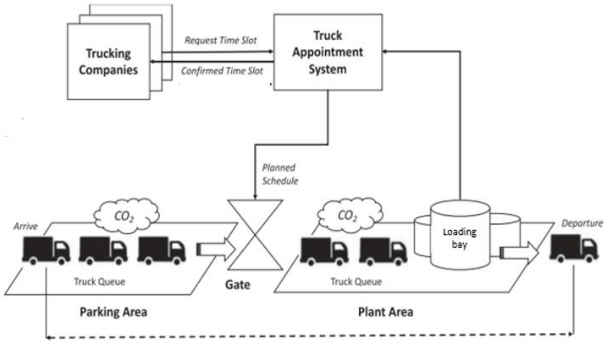


Figure 1.1 Flow in Typical Loading Facility, adapted from [11]

In addition, assigning truck to upon arrival would eliminate the necessity of rescheduling, yet this is not simply the case. The loading facility constitutes of 6 loading bays that provide different types of products, thus the predetermined product assigned to the truck could possibly be loaded only at specific loading bay [4]. Due to this requirement of a specific assignment of loading bay, risk of not getting a slot at its appropriate loading bay when trucks arrive without initial schedule is uncontrollable. It emphasizes the significance of schedule management to avoid unfavorable consequences for both parties.

Consequently, the deviation between initial schedule and actual truck arrival causes reduced operational efficiency in the sense of underutilization of loading infrastructure, congested operation, and unserved trucks. Referring to the data provided in [4], the operational efficiency score is roughly 60% of its maximum. According [21], implementation of more advanced schedule management that considers stochastic arrival time would potentially provide an improvement in operational efficiency by roughly 24%. Given the significant adjustment, it is intriguing to investigate the impact of more advanced schedule management in this case.

1.3.Research Definition

This section will provide the detail context of the study and the specific direction of this research, including the research objective, scope, limitation, research questions and methodology.

1.3.1.Research Objective

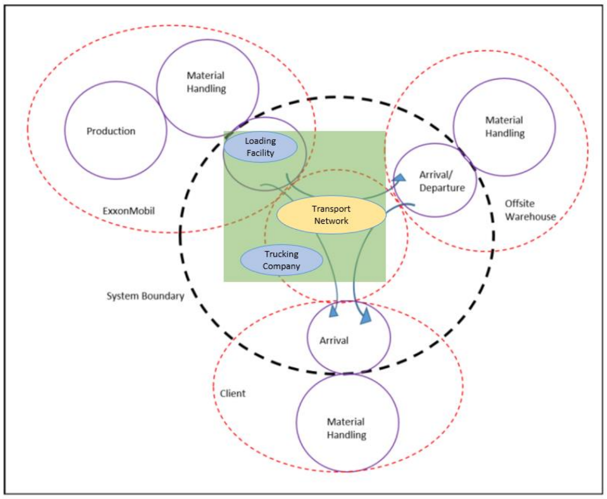
Digitalization is a key facilitator in the shift to a more sustainable and efficient logistics/freight transportation sector. One important element is the integration between logistic and traffic system. Although the two are seemed as two different domains, in reality, they might benefit from a more integrated strategy. Essentially, both sectors face the congestion and time uncertainty issue on a regular basis. Nowadays, given unprecedented growth in logistic demand and significant shift in traffic behavior, it is imperative to unlock the potential of an integrated system. In the perspective of petrochemical loading facility, the real-time flow of information of arriving truck could be utilized to optimize the operation.

The first objective is to develop an intelligent schedule management system for petrochemical loading facility that could mitigate the detrimental impact of uncertain truck arrival time, and therefore would ensure higher operational efficiency. Moreover, since this research focuses on utilizing real-time ETA information to enable rescheduling ahead of time, the second objective is to observe the added value of the technology in operation of loading facility. The result will be evaluated relative to the current situation where the real-time information is not available. Lastly, the third objective is to observe the added value of integrating a predictive model and optimization model that considers stochastic variable. The result will be assessed relative to the standard approach where the predictive model is not utilized, and the optimization model assumes all values as deterministic.

A loading facility owned and operated by [redacted] located in Antwerp is chosen as the use-case sample of this study, although the outcome is not intended as a tailor-fit solution for their particular problems.

1.3.2.Scope & Limitation

To provide a clear overview on the distribution system considered in this research, a system boundary that describes a complete operation of downstream distribution of petrochemical products is provided as it is shown in Figure 1.2.



`

Figure 1.2 Scope of Research, adapted from [4]

Due to several limitations, this research focuses only on the fractional parts of the distribution system that includes two actors, namely loading facility and trucking company whereby the operation is centered on the road transport network, as it is illustrated by the green square. Moreover, for the transportation network, this research only considers the road transportation system, hence, the possibility of using water and rail transportation as alternative solution is out of the scope. To this end, only two actors are directly involved, namely the [redacted]'s loading facility located in Antwerp and the trucks/trucking companies originated from varied locations in Europe.

Furthermore, given that scheduling is highly likely to happen multiple times in daily/hourly basis, it is inferred that this research contributes to the domain of operational decision level in truck schedule management.

With respect to the modeling, only truck arrival time is considered as stochastic variables that would be represented by probability that follows the form of certain distribution, whereas all other parameters/variables are assumed deterministic. For the predictive model, this research would only observe the varied techniques belongs to ML concept. In developing an optimization strategy, set of solutions that are based on heuristic/meta-heuristic approach are excluded, thus this research only focuses on the exact approach. The optimization model would adapt a rolling horizon approach that would be run/solved repeatedly in which the planning interval is moved forward during each solution step.

With respect to the data limitation, a synthetic dataset will be necessarily generated to fulfill the requirement of modeling and analysis. The synthetic dataset will be derived from the publicly available database that resembles the case of this research, namely the airport operation or the port operation. This means that the synthetic data must consist of the important characteristic or parameters that is deemed essential, namely the location, destination location, scheduled departure and arrival, actual departure and arrival, distance, weather, etc.

1.3.3.Research Questions

The combination of problem statement, objectives, and scope imply the following research questions:

"To what extent an integrated system of logistics and traffic would improve the schedule management of petrochemical loading facility considering the stochastic nature in transportation?"

To answers the main questions, sub-questions are defined, as follow:

Sub Research Question 1: *Is it conceptually possible to design a schedule management that is more efficient and robust against real-time disturbance?*

The goal of this research question is to develop a DSS (Decision Support System) to improve operation in loading facility. The DSS would focus on the advantage of integrating the real-time ETA and historical data. The process would require relevant theories with respect to the context of this research, which would be obtained from a literature review. The result would lead to factors that might influence the performance of loading facility, consequently, it would help identifying the precise way to model the uncertainty in arrival time. Finally, the acquired knowledge would be fundamental in determining the most appropriate conceptual framework to integrate the predictive model and the optimization model.

Sub Research Question 2: *How to improve prediction of truck arrival time based on real-time information?*

This research question aims to build a predictive model to alleviate problem stemming from the uncertainty in truck travel time by implementing ML algorithm. As the ML approach requires a large amount of data to learn the pattern, a historical dataset consisting of relevant parameters is synthetically generated. Subsequently, the synthetic dataset would be explored to prove its validity and to highlights potential insights that can be derived from the historical data. The dataset would be pre-processed to ensure that ensure that it is sufficient as the input to the predictive model. In building a predictive model, varied ML algorithm would be applied and evaluated, accordingly.

Sub Research Question 3: *What are the benefits of optimization model that considers stochastic arrival time in improving the loading operation?*

This research question aims to build an optimization model for schedule management that could increase the efficiency of slot utilization in loading facility. The optimization model would consider truck arrival time as the stochastic variable that would be integrate into the basic mathematical model. To test the model, a set of scenarios would be generated to verify the performance of the proposed solution. Finally, the result of optimization model would be assessed with regard to relevant KPI.

Sub Research Question 4: *How much added value does the proposed solution offer in dealing with uncertain arrival time?*

This research questions focuses on analyzing the result of the proposed schedule system compared to a baseline model that disregards the stochastic nature of truck arrival time. Moreover, managerial insights with respect to the implementation of this proposed system would be further explored.

1.3.4.Methodology and Deliverables

To answer the research and to attain the research objective, a research methodology is formulated. The flowchart of methodology as it is illustrated in Table 1.1 is proposed. This flowchart would also serve as the guidance to structure this research. Brief explanation of each step is provided, as following:

a. Establishing Conceptual Framework

To provide sufficient theoretical insights and to determine the relevant contexts of this research as a part in answering the Research Question 1, a literature review will be done. It would include varied domain of knowledge that constitute a sound foundation to this study. The purpose is to explore the previous study and the state-of-the-art, thus the exact way of how this research can be placed in filling the knowledge gap can be determined. Moreover, theories and concept as the foundation to achieve the objective of this research will be established. Based on the obtained information, a conceptual framework that provides solution to the defined problems will be proposed. The main principle of it is the integration of real-time ETA information and schedule optimization in loading facility. The creation of this conceptual framework would answer the main point of Research Question 1 in which a more efficient strategy of rescheduling to work against real-time disturbances is required.

b. Developing Predictive Model

To answer the Research Question 2, a sufficient dataset will be synthetically generated to accommodate the next steps of the research since the availability of a real-life data is impossible at this current state. Consequently, to ensure that the synthetic data will deliver a valid result to a certain degree, an explanation on what kind of important characteristic/parameters that should be in the synthetic dataset will be provided. Lastly, as the big data would play a vital role, especially for the predictive model, data exploration and analysis will be conducted to show the potential of using big data in deriving valuable insights related to the operational of loading facility. As the final parts to answer the Research Question 2, a predictive model based on Machine Learning algorithm will be designed in this section. Several algorithms and loss functions will be implemented to test their capability in addressing the specified challenges. Each algorithm will be benchmarked using a set of standardized evaluation metrics, subsequently, the best performing algorithm will be used in integration with the optimization model .

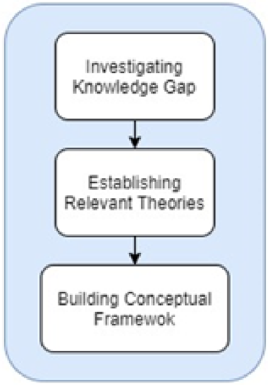
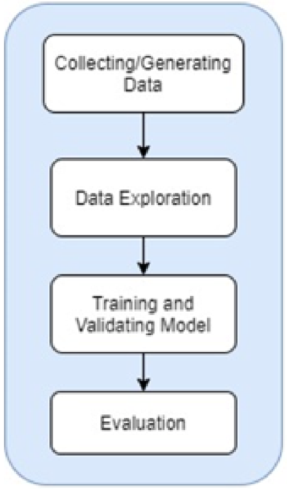
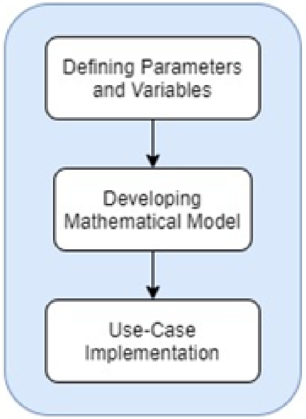
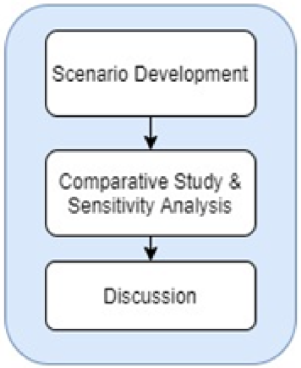
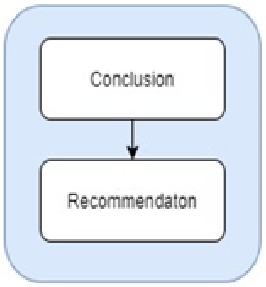
c. Developing Optimization Model

To answer the Research Question 3, an optimization model based on exact method will be proposed and integrated with the predictive model. This model will consider a stochastic arrival time; thus, this approach will be able to generate adjusted schedule to mitigate the detrimental effect of uncertain arrival time. In order to verify whether this optimization model works as intended and to showcase the potential benefit of this model when being applied in practice, a use-case analysis will be conducted. The analysis will simulate the implementation of this proposed system in the dynamic environment.

d. Numerical Experiment

To answer the Research Question 4, numerical experiment that features comparative study and sensitivity analysis will be conducted. The goal is to investigate the added value of this proposed system and to understand in which particular condition this proposed system would optimally work, respectively. To realize this, some scenarios in regard to variation of important factors or characteristics relevant to the operation of loading facility will be composed. Hence, it ensures that the result would be representative to resolve the main problem in the operation of loading facility. Since the core of this study is to grasp the added values of the proposed solution that considers uncertain arrival time, the result will be compared to the baseline model that assumes all variables as deterministic.

Table 1.1 Research Methodology

Research Question	**RQ1**	**RQ2**	**RQ3**	**RQ4**	
Methods	Literature Review	Prediction Model & Benchmarking	Optimization Model & Benchmarking	Numerical Experiment	
Process	 Investigating Knowledge Gap Establishing Relevant Theories Building Conceptual Framewok	 Collecting/Generating Data Data Exploration Training and Validating Model Evaluation	 Defining Parameters and Variables Developing Mathematical Model Use-Case Implementation	 Scenario Development Comparative Study & Sensitivity Analysis Discussion	
Deliverables	Conceptual Integrated Framework	Arrival Prediction	Optimized Schedule	Remark of added values	
Related Chapters	2 & 3	4 & 5	6 & 7	8	9

2. Literature Review

Following the previously defined methodology, the first next step is to identify knowledge gaps in this area of research. Therefore, the justification of conducting this research and its potential contribution to the existing body of knowledge will be clearly defined.

In this chapter, literatures related to schedule management are examined. Due to the similar concept in scheduling and rescheduling, it is assumed that those terms refer to same problem. There will be two main focuses which are to investigate the precedent researches in regard to prediction of arrival time, and to explore the possible optimization concepts that are possibly applied to truck schedule management in loading facility.

In general, the all the literatures are obtained in publicly accessible database, namely Scopus and Google Scholar. The keywords used are “real time ETA”, “delay arrival time”, “prediction model”, and ”stochastic model”. To get more specific information, Boolean Search is applied, for instance “operational AND scheduling”, “machine learning AND delay” and “ probabilistic distribution OR scenario-based ”. Moreover, ‘backward snowballing’ is performed in a few cases when precedent information about a specific intriguing topic is needed. To ensure clarity, it is assumed that ‘planning’, ‘assignment’ and ‘scheduling’ refer to the same problem and use them interchangeably. Same assumption also applied to ‘deviation’, ‘delay’, and ‘unpunctuality’. Lastly, all the literatures reviewed are written in English.

2.1.Previous Research

An overview of precedent researches of truck arrival prediction in the truck schedule management domain and other closely related domains. Most of the current studies are done in the domain of air transport and railway management. The literatures reviewed in this section are considered having relevance to a certain extent, therefore, to cope with the limitation, the reviewed methodology must be adjusted, and applied to solve truck rescheduling problem. Furthermore, the methods of predictive model based on simulation and queueing model are excluded because they are out of scope.

2.1.1.Predictive Model

To address the issue of inherent uncertainty in transport domain, a predictive model is commonly proposed as the solution, thus extensive reviews on predictive model is conducted. In general, predictive model in the context of predicting arrival time is based on historical data, and the real-time data. Both approaches have been studied in varied transportation domain.

- **Based on Historical Data**

This approach embodies the concept of event’s occurrence as a probability that could be estimated from historical data. There are several methods in calculating the probability distribution. Boswell & Evans [22] expressed delay as discrete probability function in which the delay propagation is used to determine cancellation analysis. More advanced approach using probability density function done in [23] explored type of distribution to best model the delay, the results verified that Poisson distribution fitted better to describe delay in arrival time.

- **Based on Real-time Data**

Exponential growth on transmission of information technology prompted researchers to utilize the real-time ETA as the basis of to derive delay arrival prediction in which all researches proved that real-time ETA provide lesser degree of stochasticity, thus could result on more accurate prediction. Larbi et al. [13] showcased that value of information in comparative study including scenario of no information, partial information, and full information, leading to conclusion that distant information yields small contribution in improving the schedule optimization model. Zhoue et al. [24] developed a model that can predict arrival time based on several related parameters according to frequency distribution and regression analysis. Another research by Xu & Ying [25] proved that combining dynamic of road networks generates a more accurate prediction of arrival time. that reflects the stochastic nature of road in reality.

2.1.2.Optimization Model

In general, optimization model has been extensively used to address the problem in schedule management. In earlier period, most of the studies [26], [27], and [28] rely on queuing theory as the tool to model the truck scheduling. However, Huiyun et al. [12] claimed that queuing model is not sufficient in providing improvement because the concept is only able to analyze the appointment data yet could not offer proper decision-making assistance in the schedule management.

Furthermore, most of recent researches on truck schedule management utilized MIP (Mixed Integer Programming) model, whereby this approach allows better quantitative evaluation. This optimization method has been proposed in many studies, in which the problem is solved either using exact algorithm or heuristic/semi-heuristic method.

- **Exact Algorithm**

In the domain of truck scheduling in loading facility, various studies of deterministic MILP model have been conducted with different objectives. Some researches constructed model with single objective, namely, to minimize emission cost prompted from idle trucks [29], to minimize penalty cost due to delay in arrival time [30], and to minimize turn over time of trucks [31], and to produce truck schedule that maximize slot efficiency considering sufficient service for all trucks [32][8], [33]. In all those studies, regardless the variation of measurement indicator being applied, the optimization model is potential approach that could enable better operation. Since in reality most of the problems are not constituted by a single factor, more recent researches incorporated multi objective or joint optimization to improve the performance of loading facility. To account for complexity of stakeholders, a deterministic MILP model with 3 different objectives corresponding to the primary interest of involved actors in facility is incorporated in determining the most optimal schedule for truck arrival [11]. Moreover, by adding emission cost as additional performance indicator, Schulte et al. [34] studied the benefit of slot assignment supported by deterministic MILP optimization model that simultaneously address generalized cost and truck congestion to optimize the schedule by enabling collaborative planning.

The majority of studies in the specific domain of truck schedule management assumed that arrival time is deterministic variable, therefore it only solves the real problems to a limited extent. It is indeed necessary to explore the study that incorporates the arrival time as stochastic variable.

- **Heuristic/Metaheuristic Algorithm**

As the uncertainty in arrival time must be taken into account to produce a more accurate scheduling, most of studies in truck rescheduling domain apply heuristic or metaheuristic method. To solve the stochastic natures in truck schedule management, studies that are underlined by simulation methods have been conducted. Huynh & Walton [35] combined MILP programming and simulation to enhance the slot assignment based on appointment requested by trucks. The goal is to determine maximum trucks that can sufficiently be served in certain period. Moreover, Huynh et al [36] expand the similar concept by varying the appointment strategy, resulting on knowledge on factors having most influence in the schedule management. Azab et al [37] created simulation model based on the concept of slot assignment system according to requested time in which the objective is to find optimal schedule based on turnaround time and delay in arrival time. Although heuristic or metaheuristic allow generating set of feasible solutions with reasonable quality under relatively short computation, they could not guarantee optimal solution. On other hand, exact algorithm could compensate this weakness and thus should be prioritized. Especially since the problem size (time horizon, parameters and variables size) of this research is not extremely massive, therefore it is justified to focus only on exact algorithm.

To that end, various studies in effort to implement exact algorithm in stochastic environment have been made to solve scheduling problem in transport sector which will be explained in the following section, yet the ones in the domain of truck scheduling in loading facility does not exist yet to author's best knowledge.

2.2.State of The Art

In this section, some existing literatures are reviewed to provide context to the present works and the state-of-the-art in the domain of schedule management. Specifically, this section focuses on researches that implement advanced method in predicting arrival time in the domain of transportation and logistics. In addition, researches that aim to incorporate the stochastic element in optimizing schedule management would also be explored as well.

2.2.1.Machine Learning for Prediction

As the technology rapidly advances, number of studies involving machine learning as a supporting tool to predict delay has exponentially increased, yet most of the studies are done in the domain of airport operation and railway management. Machine learning is deemed powerful to provide accurate prediction in which the concept requires learning process based on historical data [38]. The justification of using machine to predict arrival time stems from its ability to consider the dependency between variables which are tended to be ignored in common statistical predictive analysis [39].

Various algorithm could be applied to make prediction, however for different set of data or conditions, it is not certain which algorithm would generate the best result. Therefore, many studies that compare different algorithm under particular case have been conducted.

Lee, Malik & Jung [40] and Rebollo & Balakrishnan [41] applied RF (Random Forest) algorithm to predict delay time in airport operation, they both concluded that these methods have potential in generating better accuracy in delay prediction, depending on the data complexity. In both studies, the algorithm was trained on a combination of various parameters relevant to the operation that, to some extent, contributed to the flight delay of hundreds of data set.

Pongnumkul et al. [42] recommended the use of k-NN (k-Nearest Neighbor) to predict delays. The former applied in case of delay propagation in airport, whereby delay prediction is based on historical data recognized as similar situation. The latter conducted comparative study to highlight the benefit of k-NN algorithm in which the k-NN algorithm is to predict the current arrival time on the same station based on the last known data of arrival time at a certain station is being used as the parameter.

The capability of SVR (Support Vector Regression) algorithm in arrival time prediction has also been explored. Markovic et al. [43] analyzed the fit between the train delay and several characteristics of railway systems. Moreover, same techniques is utilized by Barbour et al. [44] to solve the ETA prediction based on two years historical data. The data incorporate both train features and network features. This results on average 14% - 21% of improvement for ETA prediction.

Recent studies by Vorage [45] and Dutrieux [46] have explored the possibility of using machine learning technique to predict discrete probability distribution for individual flight, and the studies have shown the competence of machine learning in doing so. Both studies compared several ML algorithms, and subsequently, the prediction results were evaluated based on standardized KPIs. Vorage [45] concluded that arrival delay is best modelled using ANN (Artificial Neural Network) algorithm, but in general all ML algorithm outperformed the normal statistical techniques on all KPIs. However, due to different circumstance of experiments, according to Dutrieux [46], the best fitted delay prediction was generated by the RF algorithm, followed closely by the Artificial Neural Network algorithm.

2.2.2.Capturing Stochastic Elements

Since almost all real-world problems always involve unknown parameters, the deterministic model cannot accommodate the uncertainty factors, therefore stochastic optimization model has been favored to increase the validity of results. Stochastic programming considers uncertain events as probability distributions that can be estimated via historical data or other advanced methods. Consequently, this model is beneficial to solve problem under circumstance in which decisions are made repeatedly in substantially similar situation, namely scheduling problem that requires several recourses/iterations due to uncertainty elements [47].

Most of studies build stochastic optimization model based on the concept of mathematical programming, but the method used in integrating the stochastic elements are varied. There are two main SP (Stochastic Programming) approaches that are suitable and have been normally used in the study of schedule management, which are two-stage stochastic and probability constraints [48]. Therefore the literature review on the application of stochastic model focus on those specific approaches.

Since most of the schedule management models are based on the deterministic approach, parts of model have to be adapted when encounter stochastic variables. Several studies assumed stochastic elements as random variables. Seker & Noyan [49] incorporated disruptions such as delay in arrival time, idle time, and buffer time as random variables which is integrated in the constraints of MILP optimization model. The objectives quantified total conflicting slots and idle slots as robustness measures.

Stochastic model that directly considers the uncertainty variables as discrete probabilistic distribution is also an alternative approach in dealing with stochastic variables. In railway schedule management study conducted by Meng & Zhou [50], many attributes associated with high degree of uncertainty, such as travel time, arrival time, and departure time are considered as stochastic variable, thus individual probability distribution is assigned. However, since in the distributions used in these experiments are not the empirical distributions from real statistical data over certain periods, the benefit of this model is limited.

Furthermore, Visser & van Schaijk [21] transformed deterministic slot assignment method to be able to consider stochastic elements by replacing the binary scheduled arrival time constraint with presence probability to capture the inherent stochastic delay. This approach would assign arrival to a certain location according to the maximum permissible overlap probability. Higher permissible value would result in a more compact schedule, therefore lower assignment cost. However, this research does not incorporate specific assignment, in which arrival can be universally assign to any available slots. This approach could not be directly applied here because, in this research, since specific product type could only be loaded in a certain slot, hence it rises the complexity of the problem.

2.3.Knowledge Gap

The discussed papers in previous section have pointed out that most optimization models have been designed in a deterministic fashion. Few papers have addressed the uncertain arrival time by applying exact method, yet, to author's best knowledge, none has been done in the context of truck schedule management.

Moreover, almost of all the researches in the domain of truck rescheduling have not incorporated predictive model powered by ML algorithm to capture the uncertainty of arrival time. Since this approach proved beneficial in similar cases in other domain, applying it in this research might be a promising direction.

As a first research gap, it is interesting to see the extent to which an exact optimization model that considers stochastic variable could increase the performance of loading facility with respect to operational efficiency Furthermore, as studies in the truck schedule management domain are limited to the use of simple heuristic to simulate the deviation in arrival time, as the second gap, it is compelling to explore the feasibility of utilizing ML algorithm as supporting tool to predict truck arrival time.

2.4.Chapter Overview

This chapter aims to find knowledge gaps as the basis to design a conceptual framework of advanced schedule management. Literature reviews in exact domain of truck schedule management and other similar domains of knowledge, namely airport operation, railway management, and port planning are conducted. The results indicate huge knowledge gaps, especially in the technological progress, relative to other similar domains of knowledge. Hence, this research will try to shorten the gap by adapting advanced concepts that are proven successful and observe the impacts if being applied in truck schedule management.

To clearly illustrates where the exact position of this respect with respect to the knowledge gap, a literature review matrix is composed and shown in Table 2.1. This matrix is based on several important characteristic that constitute the context of this research. In the context of loading facility, precedent studies that include a predictive model based on ML algorithm is still not existed to author's best knowledge. Moreover, many studies have been done to include stochastic arrival time in optimizing schedule management, but ones that attempt to solve it in exact/analytical manner are hardly to be found.

Therefore, the truck schedule management that will be proposed in this study will utilize the predictive powered by ML algorithm and the probabilistic mathematical optimization model based on exact algorithm

Table 2.1 Literature Review Matrix

Literatures	Truck Schedule Management	Stochastic Arrival	Predictive Model	Optimization Model with Exact Solution	Specific Assignment
[22] [23]	✗	✓	✓ (Discrete Probability Distribution)	✗	✗
[7], [11], [30]–[35]	✓	✗	✗	✓ (MILP)	✓
[35]–[37]	✓	✓	✗	✗	✓
[39]	✓	✗	✓ (Machine Learning)	✗	✓
[40], [41]	✗	✓	✓ (Machine Learning)	✗	✗
[42]–[44]	✗	✓	✓ (Machine Learning)	✗	✗
[45], [46]	✗	✓	✓ (Machine Learning)	✓ (Probabilistic MILP)	✗
[49]	✗	✓	✗	✓ (Two Stage SP)	✗
[50]	✗	✓	✓ (Discrete Probability Distribution)	✓ (Scenario-based SP)	✗
[21]	✗	✓	✓ (Regression)	✓ (Probabilistic MILP)	✗
This research	✓	✓	✓	✓	✓

3. Integrating Predictive and Optimization Model

In the previous sections, knowledge gaps are defined. They clearly illustrates that research in this domain significantly lags compared from the ones in another similar domain, especially in addressing stochastic variable and implementation of more advanced predictive model such as ML algorithm.

Regarding the main finding in literature review and as part of answering the Research Question 1, this chapter provides relevant and in-depth theories related to decision support system in supply chain, uncertainty in arrival time, rescheduling approach, and basic concept of machine learning. Therefore, this section aims to establish fundamental knowledge that would serve as solid base in designing conceptual framework to solve the stated problems.

3.1.Relevant Theories

3.1.1.Supply Chain in Petrochemical Industry

There is no official statement on the definition of supply chain, generally in most literatures, the core of the definition refers to the integration of entities and activities over the network to ensure a performance of business [51].

Petrochemical industry is heavily associated with vertically integrated supply chain whereby the system could be considered as the most complex, dynamic, and advanced system in the world. One of the reasons is due to size of the operation in which the supply chain system covers the production, refinement, transportation, and distribution to the consumer's market. Commonly, the supply chain of petrochemical industry is divided into 3 different stages which are the upstream, midstream, and the downstream as it is illustrated in Figure 3.1. The upstream segment includes all the functions related to the exploration, production and crude oil transportation to the refinery facility. Subsequently, the midstream segment is related to all refinement activities to convert the crude oil into refined products. Lastly, the downstream segment consists of mainly distribution operation including primary and secondary distribution, including storing and transporting the final product, and also the responsibility of marketing into customer's market [52].

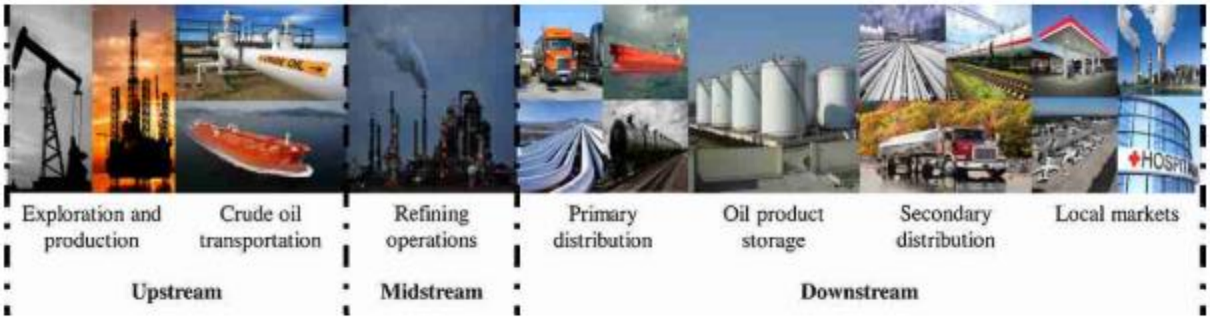


Figure 3.1 Stages in Petrochemical Supply Chain [52]

Intensive capital involvement, geopolitical upheaval, competitive environment exacerbated by the price volatility, and its social or economic impacts to the surrounding highlight the sensitivity of the petrochemical supply – chain system to disturbances and uncertainties [52]. To cope with this vulnerability, a decision support system (DSS) to guarantee profitable, sustainable, robust and efficient supply chain is necessary to be established [53].

3.1.2.Decision Support System

In general term, DSS is defined as system/technology that helps decision-making process by putting correct knowledge to the right person at the precise time and cost [54]. According to [55], transportation sector contributes 10-12% of European GDP, therefore the urgency to develop advanced decision support tool to derive reliable and effective decisions could not be overestimated. In this research, DSS in the form of schedule management is designed to improve more efficient operation. There are two common strategies of rescheduling in operational level which are using buffer time based on historical data and ad hoc planning based on real time information [4].

Modifying the schedule in a way to accommodate buffer time could enable robustness of the system because the delay probability of scheduled trucks could be compensated by the planned buffer time. However, this approach risks increasing occurrence of idled slots, in other words, this strategy thrives at the expense of the efficiency of slot utilization. In addition, this strategy also lacks robustness against the real time interference or unexpected events, for instance, vehicle breakdown, natural disaster, abnormal congestion, traffic accidents, etc. Due to described drawbacks, the online ad hoc planning is preferred as strategy solution in this research.

Unlike the buffer slot strategy, the online ad hoc planning incorporate dynamic attribute in the scheduling process. To increase the efficiency of the schedule, advanced technology, namely real-time communication to transmit the most recent update is utilized to cope with the inherent uncertainties. Online ad hoc planning would be conducted after the initial scheduling is set. This concept relies on dynamic communication between the inbound trucks and the schedule manager. The exchanged information involved the continuously update real events, for instance congestion level in transport network, internal problem of the trucks, unexpected weather condition, unforeseen external disturbance, etc. Subsequently, this collected information could be processed to generate the most possible accurate of truck ETA. Based on the continuous flow of ETA information, decision making on whether to re-assign that particular trucks in case of delay in the arrival could be made, therefore the new adjusted schedule for that truck is confirmed and communicated back to the truck's drivers.

As a decision support system, the online ad hoc planning require a set of sub-functions to constitute the framework of the system. According to [56]a general structure of real-time DSS is illustrated on Figure 3.2 in which the structure of functions is described in Table 3.1.

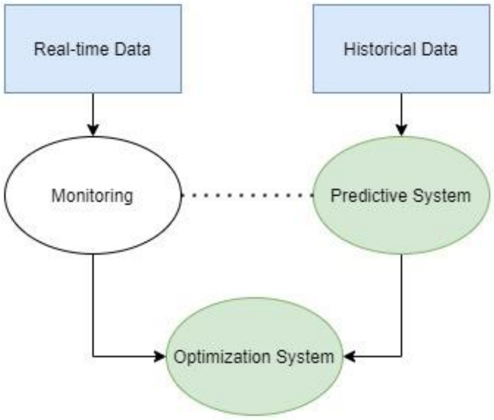


Figure 3.2 Intelligence DSS, adapted from[56]

Table 3.1 Description of DSS function

Function	Description
Real-Time Information	This function constantly transmits and updates information that might affect the truck arrival time at the loading facility. It includes communication for real time traffic condition for instance, accidents, bad weather.
Historical Data	This function provides information of past data that includes recurring incidents and records of elements that could indicate the trend or pattern of truck arrival time.
Monitoring	This function checks the state of truck arrival based on the rea-time data. The output of this process is information whether a truck could follow its initial schedule.
Predictive System	This function predicts the likelihood of a certain truck to arrive early, on-time or late with respect to its initial schedule. The basis of this function is the knowledge derived from historical data. It works in integration with the monitoring and real-time data; thus, the result would be accordingly updated.
Optimization System	This function proposes a optimization approach to mitigate the anticipated risks.

3.1.3. Advanced Schedule Management

Amid dynamic and stochastic supply chain environment, one of the most prominent concern in regard to the transportation sector is that loading schedule cannot always be used as it was planned because of the unforeseen disturbance, therefore, to cope with this problem, it is necessary to have a clear guideline on the rescheduling scheme [57].

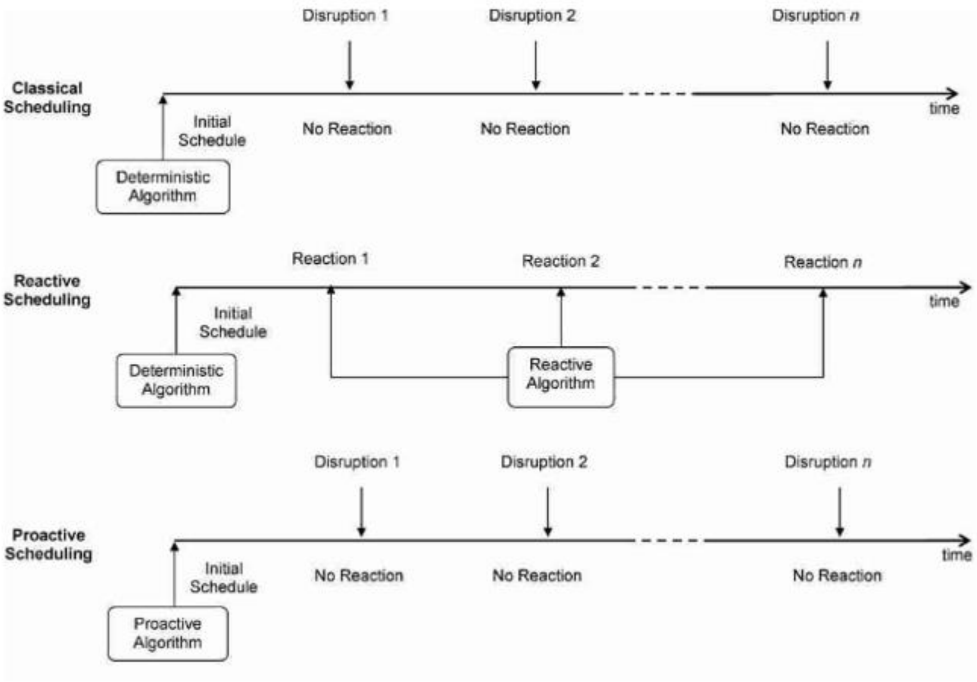


Figure 3.3 Scheme of scheduling strategy [58]

Theoretically, there are three general approaches in scheduling process, which are the classing scheduling, reactive scheduling and proactive scheduling as explained in [58] and shown in Figure 3.3. Classic scheduling is the type of scheduling based on deterministic algorithm that does not have set of mitigation plan when there are disruptions. On the contrary, when generating initial schedules, reactive scheduling does not explicitly incorporate uncertainty, but it does update the schedule when unexpected events or interruptions occur. To put it another way, reactive scheduling tries to figure out how to mitigate disturbances after they occur. The reactions are usually either altering the current original timetable or fully creating a new one from scratch.

This research considers a real-time information that will be updated periodically as one of the important input variables to develop intelligent rescheduling system, therefore it is more reasonable to opt for the reactive strategy as the main concept of the conceptual framework.

3.1.4.Factors Influencing Arrival

Since the loading facility can be considered as part of road transportation network, inbound arrival time of trucks are influenced by several peculiar factors includes either internal or external factors. Survey-based research conducted in [4] has successfully figured out the most probable and rational factor causing deviation in arrival time from the perspective of truck drivers. The response of the truck drivers is then analyzed and then transformed into a weighted average value that indicates the impact of a certain factor relative to others on the deviation of truck arrival time as it is shown in Table 3.2.

Table 3.2 Survey Result of Factor Influencing Arrival [4]

Factor	Weight
Congestion	25 %
Previous Job	21.4 %
Road Diversion	20 %
No Communication	13.3 %
Technical Breakdown	11 %
Force of Nature	9.3 %

As it can be observed, troubles related to the transportation network (“congestion” and “road diversion”) contribute to almost half (45%) of the total impact of early/late arrival time. Since this research only focus on the loading facility process in which the actors are the truck drivers and facility planner, thus the “previous job” factor that accounts for the second highest impact is out of the scope. Other important factor that must be considered is the “no communication” that constitutes of 13.3% impact of arrival time. Taking into account these 3 largest factors, the reasons of deviation of arrival time can be categorized into 2 solid classes, which are the road problem and the lag in communication system. Hence, this research will focus on the improvement of schedule management with respect to those two major factors of delay.

3.1.5.Presence Probability

With respect to uncertainty of arrival time, the theory of determining presence probability of certain truck at particular time slot tis explained. This presence probability could be derived from the predictive model as it is proposed in [46]. Additional random variable is defined to indicate the presence of truck j, as following:

X_{arr}: Random variable representing the time step in which truck j arrives.

Theoretically, statistic approach of distribution function could be used to capture the uncertain arrival of trucks. In this case, the concept is applied to illustrate whether a particular truck j arrives before the time step t, Cumulative Distribution Function (CDF) is formulated, as following:

$$F_{X_{arr}}(t) = P(X_{arr} \leq t)$$

This form of function could be translated into Probability Density Function (PDF) by applying derivative transformation, defined as:

$$f_{x_{arr}}(t) = \frac{d(F_{x_{arr}} \leq t)}{dt}$$

Since it is assumed that arrival time is the only stochastic indicator to determine whether trucks is present in facility, thus the requirement of being present could be simply defined as:

$$g_{pres} = F_{x_{arr}}$$

From historical data, delay arrival time with respect to the Scheduled of Arrival Time (STA) is plotted (Figure 3.4). Probability Density Function (PDF) of the arrival time of a particular inbound truck is constructed as it is illustrated in the upper graph. In the same manner, the Empirical Cumulative Distribution (ECDF) is also composed as it is illustrated in the lower graph. Both graphs are placed in a same timeline, whereby the zero-minute delay points are aligned at their respective STA.

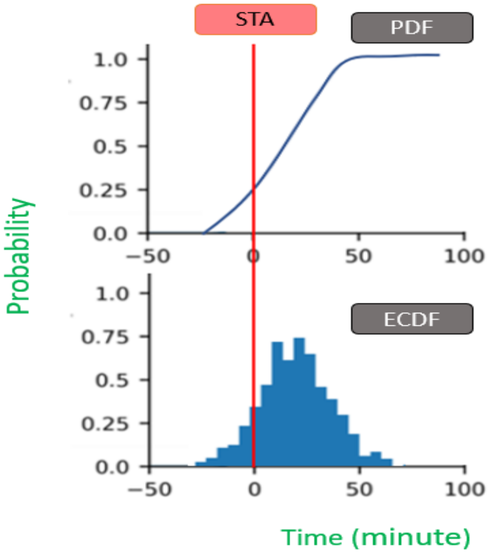


Figure 3.4 Comparison between PDF and ECDF

ECDF models the observed (empirical) data, instead of the hypothetical model as the result of the normal CDF. The justification of choosing the ECDF is due to its advantage, namely that ECDF would always result on discrete value even though the data is not drawn from the discrete distribution, while the normal CDF could not guarantee the discrete value.

Clear comparison between deterministic and stochastic approach is illustrated by plotting the binary scheduled time of a certain truck (red line) indicating the deterministic approach and the presence probability curve (green line) indicating the stochastic approach of the same truck as it shown at Figure 3.5.

In other words, the discrete presence probability value representing how likely a truck is present could be derived from the ECDF curve constructed for each scheduled truck in certain rolling time horizon. The integration of this concept with respect to the stochastic optimization model would be further explained in Section 3.2.2 .

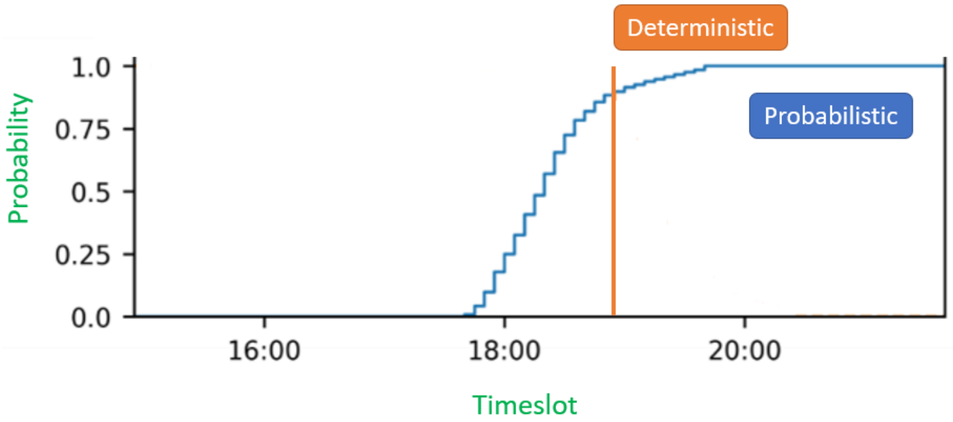


Figure 3.5 Deterministic vs Stochastic Approach of Arrival Probability, adapted from [21], [46]

Theoretically and ideally, in the case where the departure time is also considered as stochastic variable then the presence probabilistic is constructed by combining the ECDF of arrival time and ECDF of departure time. However, Since the presence probability is indicated by the time at which the truck has arrived at the facility (after factoring delay), thus the curve of presence probability could be directly adapted from the ECDF graph of truck arrival. For example, as it is shown in Figure 3.5, there is approximately 0.8 probability that truck j would be present at 19.00, compared to the deterministic approach that suggest the truck j would be present at 19.00 with an absolute certainty. In this research, probability of an individual truck being early, on-time, or late with respect to the real-time ETA information, the discrete probability value would be generated using predictive model.

3.1.6.Machine Learning for Predictive Model

According to [59] utilization of real-time flow of information would reduce the unknown factor in arrival time information. However, realistically, it still yields a certain degree of uncertainty[60]. To address this concern, a presence probabilistic value could be assigned to indicate how likely the truck would actually arrive according to its real time information.

Since it is understood that delay variability in transport network does not follow obvious pattern, for example, it could be in the form of within days variability, between days variability, based on weather, based on origin, etc. To that end, ML technique that learns the subtle pattern, trend, or structure embedded in the historical data is a suitable approach [61]–[63]. Specific to this case, the data from which the ML would learn is the historical records of truck arrival time including the real-time parameters, while the target is to predict the probability value of a certain truck being early, on-time, or late.

- Difference between AI, ML, and DL

Knowledge domain that try to make machine could simulate how human intelligence is consisted of 3 connected concepts as it is explained in [64] and shown in Figure 3.6. The widest term for technologies that aim to replicate human intellect is AI (Artificial Intelligence). It could anticipate, automate, and optimize processes that have traditionally been done by humans, such as voice and facial recognition, decision-making, and data interpretation. ML is a subset of AI in which computers learn from data. It refers to the intersection of computer science and statistics, in which algorithms are utilized to execute a given task without being explicitly coded; rather, they identify patterns in data and generate predictions. .DL (Deep Learning) is a subset of ML that automates much of the feature extraction piece of the process, eliminating some of the manual human intervention required. The backbone of DL is the ANN (Artificial Neural Network) which resembles the neural network inside human brain., resulting on more advanced learning process compared to standard ML model.

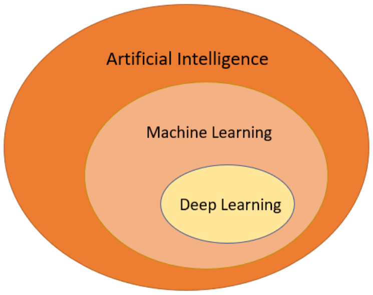


Figure 3.6 Relation between AI, ML, and DL

As it is shown in Figure 3.7, there are 3 important layers, namely input layer, middle layer, and output layer. Since their values aren't visible in the training set, the middle layers are referred to as hidden layers. Hidden layers, to put it simply, apply weights to the inputs and guide them through an activation function as the output. The hidden layers perform nonlinear transformations of the inputs entered into the network. Frankly, a DL network is defined as an ANN that has two or more hidden layers.

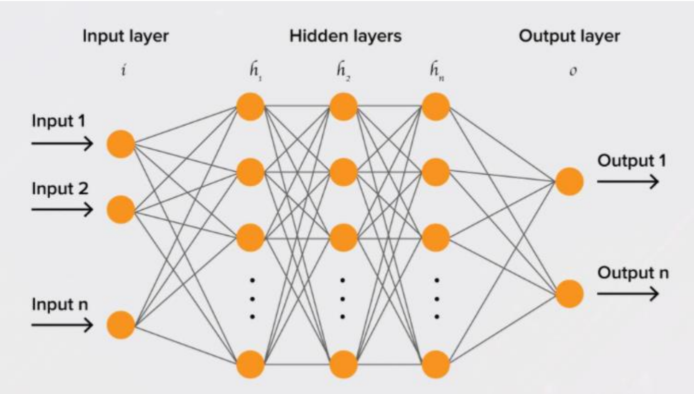


Figure 3.7 Basic structure of ANN, adapted from[64]

- Supervised and Unsupervised Learning

One distinguished characteristic between one model to another is the applied algorithm. ML algorithm is basically a method by which the system would conduct the prediction. There is no straightforward guideline on determining which ML algorithm should be opted or which one would result on best fit in understanding the historical data [64]. The process requires extensive trial-and-error experiments that involve several ML algorithms. However, since ML algorithm can be categorized based on its utilization, the pool could be narrowed down according to the context and goal. Since this research aims to predict the arrival class, namely being early, on-time, or late, it falls in the domain of multi-class classification. Accordingly, multiple ML algorithms that are deemed feasible to this circumstance are tested to find the best fit of the dataset. The categorization of machine learning type is as following:

Supervised learning is a ML approaches that are trained on labeled datasets. This technique aims to find a relation between variables; thus, it requires function to map the input variable (X) and output variable (Y) based on training set (X,Y). In other words, the foundation of supervised learning is similar to the concept of function approximation or curve fitting. The advantage is the high prediction accuracy compared to other learning method, but the downside is the requirement of human intervention to properly label the training data and to provide feedbacks on the output.

Moreover, this technique is commonly used to predict outcome of either classification or regression problem. Classification is technique used to assign an unknown test data into specific categories according to its attribute fit to training data, while the regression is technique used to comprehend the relation between dependent and independent that could be useful to predict the numerical values based on different data points. The supervised learning can be broken down into two categories, which are the generative model and discriminative model, the comparison between them is shown in Table 3.3.

Table 3.3 Key Difference between Generative and Discriminative Learning

	Generative	**Discriminative**
Accuracy	Highly dependent on training set	Dependent on training set and algorithm configuration
Requirement	Necessary to model both observed and hidden variable, resulting on high amount of training	Quality of training data does not have to be superbly good
Computation cost	Relatively low because it is based on graphical method following Bayesian rule	Relatively higher It relies on optimization of convex function with significant tolerable error
Constraint	Most algorithms assumes certain degree of independence features	Most algorithms accommodate dependency among features

Unsupervised learning is a machine learning approaches that are trained on a pattern inferred form the unlabeled dataset. This technique aims to discover hidden structure/pattern of regularities and irregularities from the input data without a need of human prediction. The most common utilization of this techniques is in clustering, and dimensionality reduction problem. Clustering is technique used to group unlabeled dataset based their characteristic, while the dimensionality reduction is technique used to eliminate number of features in a given datasets without decreasing its data integrity.

Direct comparison that shows key differences between supervised and unsupervised learning based on several important parameters is shown in Table 3.4. Accordingly, it can be inferred that the supervised learning is the most suitable for the predictive model in this research.

Table 3.4 Key Difference between Supervised and Unsupervised Model

	Supervised learning	**Unsupervised learning**
Data type	Labeled	Unlabeled
Goals	To predict outcome of new data based on the precedent information.	To discover hidden pattern or inherent structure of the large dataset
Application	Classification and regression	Clustering and dimensional reduction
Complexity	Less complex, it does not require super computational power.	More complex, it requires powerful tools and large amount of generic data.
Drawbacks	In some cases, it is time-consuming to label the data and to train the model	High sensitivity that could lead to a wildly inaccurate or invalid results

- Determining ML Algorithms

Previously, it is decided to apply supervised learning algorithm, specifically the probabilistic classifier algorithm. The detail of supervised learning algorithm is shown in Figure 3.8. More specific study is required to determine which exact algorithm should be applied. To fully understand which specific supervised learning algorithm is the most suitable for this case, more elaboration is provided.

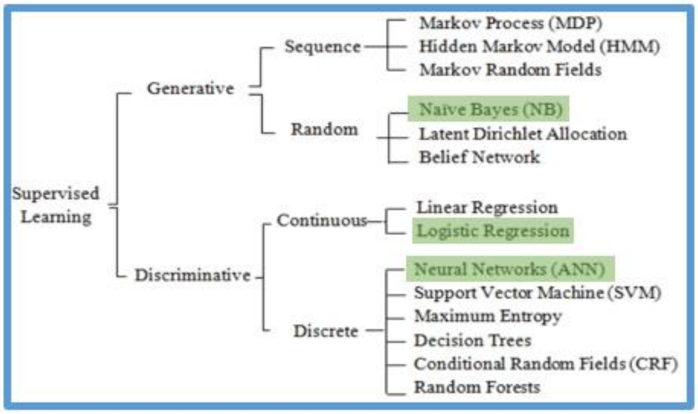


Figure 3.8 Taxonomy of Supervised Learning Algorithm, adapted from [65]

Similar studies to the context of this research suggested that Gaussian Naïve Bayesian (GBN), Logistic Regression (LR), and Artificial Neural Network (ANN) are among the best of classifier algorithm. However, the best performing ones varies according to the input data. Therefore, it is interesting to test those 3 algorithms would perform in the context of this research.

On one hand, the GBN algorithm is a fast-working algorithm. It may be used to solve challenges involving multi-class prediction. Furthermore, if the assumption of feature independence remains true, it can outperform other models while requiring far less training data.[66]. However, The pitfall is, for instance, when two features are extremely correlated (or just imagine adding the same feature ‘a’ twice), the GBN could not differentiate them, therefore it would result on overestimation [67], [68].

On the other hand, Logistic Regression offers numerous advantages over GBN , as it is explained in [69], LR does not require statistically independent variables, but it assumes collinearity relatively low. By contrast, LR offers robustness in dealing with multicollinearity.

Ultimately, ANN has a number of advantages over other algorithms. One of the clear advantage its ability to comprehend non-linear and complex connections since many interactions are not straightforward and one dimensional. Furthermore, unlike other algorithms ANN could generalize unknown relationships from unfamiliar data. The only downside of this technique is its requirement of heavy computational power.

3.2.Conceptual Framework

In the previous sections, relevant theories to address the issues in this research have been established. The gained knowledge will be used as the basis to design a conceptual framework of advanced schedule management.

This section provides an elaboration on the proposed solution that includes general flow and the detailed conceptual framework of rescheduling system. The main goal of the framework is to integrate predictive model and optimization model.   Furthermore, an example of use-case implementation is also presented to clearly show how the proposed solution works.

3.2.1.Intelligent Rescheduling System

In regard to the general loading process explained in Section 1.2 and to acknowledge the rescheduling problem due to the uncertain truck arrival time, an intelligent rescheduling system that could provide more operational efficiency and robustness against real-time disturbance is introduced as main solution.

The general flow of the system is illustrated in Figure 3.9. As it is mentioned, it considers stochastic nature in transportation represented by uncertain arrival time. This stochastic variable is captured as probability value that will be generated by the predictive model that learns from historical data. Continuously, the real-time ETA data will be inputted to the predictive model that will assign a probability of whether a truck would be early, on-time, or delay. Based don that information, the optimization model could generate an adjusted schedule that could facilitate a more efficient and robust operation.

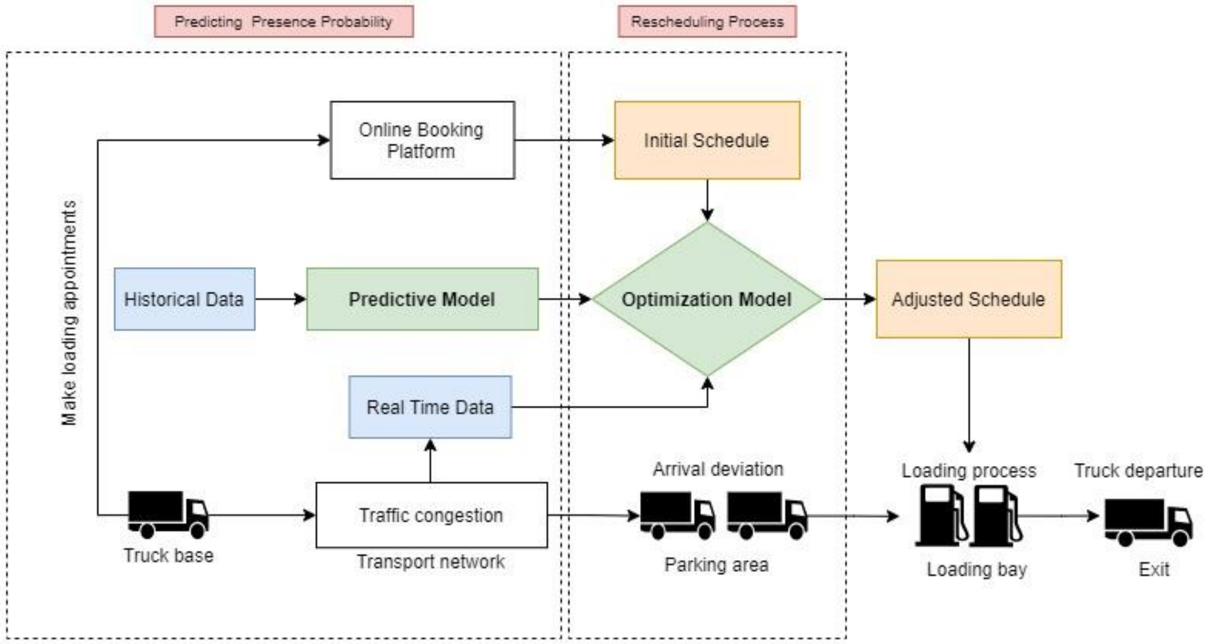


Figure 3.9 General Flow of Proposed Intelligent Rescheduling System

3.2.2.Integration of Predictive and Optimization Model

This section will specifically explain the proposed method in integrating the predictive model (probabilistic classifier powered by ML algorithm) and the optimization model (exact method). The core idea of this framework is to input the real-time ETA information into the predictive model that would generate the probability of a certain truck being early, on-time, or late considering the trend/pattern of historical data learnt by the ML algorithm. Subsequently, the result of the predictive model would be inputted into the optimization model that would generate the adjusted schedule. The flow process is detailed in Figure 3.10.

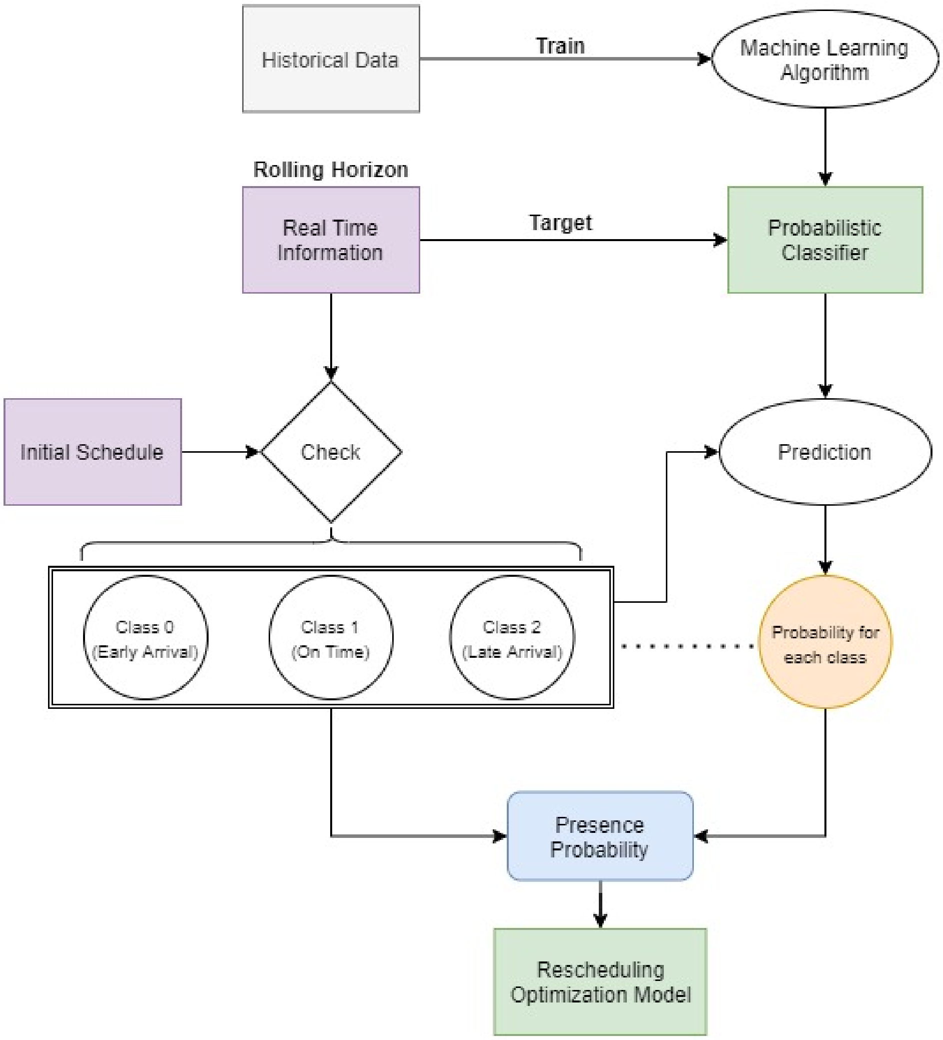


Figure 3.10 Conceptual Framework of Integration between Predictive and Optimization Model

In this case arrival class relative to the initial schedule is used as reference point (checking step). This is done to determine whether rescheduling is triggered. The arrival class is varied from 0, 1, and 2 in which the former figure means that the truck would arrive early (at minimum 10 minutes early), the middle figure means that the truck would arrive on time (between 10 minutes early and 5 minutes late), and the latter figure means that the truck would arrive late (at minimum 5 minutes late). According to that classification, it defined that the class 0 and class 2 would require rescheduling, whereas the class would stick to the initial schedule.

With respect to the defined classes, the predictive model could assign probability value for each truck that indicates how likely a truck belongs to a certain class. This value is subsequently perceived as the presence probability of truck at a certain timeslot. By doing this, the uncertainty embedded in the real-time ETA information is acknowledged and captured. As the final result, a pair of ETA information and its corresponding presence probability value is prepared to be inputted into the optimization model (furtherly explained in Section 5.8)

Furthermore, in regard to the time horizon, the set of real-time ETA information will be updated based on the defined rolling horizon. In this research, it is assumed that the rolling horizon is 30 minutes, therefore there will be new set of real-time ETA information for every 30 minutes. For each period, the probabilistic classification model that has been trained using the historical data would use the newly updated ETA information including all the features as the input data to predict the probability value of a particular truck to be present according to its ETA information.

3.2.3.Real Time Information & Presence Probability

For a certain period in the rolling horizon, the planner in the loading facility receive a real-time information on the arrival of trucks scheduled in that day, shown in Table 3.5. The real-time data includes several features data points that could be inputted to the predictive model; thus, it could indicate the arrival class of the corresponding truck. The real-time ETA information consists of product type and its corresponding origin which is the base of truck company. Moreover, a time indicator that would possibly influence the delay behavior is also presented, namely day of the week, initial scheduled departure from the origin, actual time of departure from origin, estimation of traveling time, total amount of time that has passed since the departure, total distance to the loading facility, and the estimated time of arrival to the loading facility.

Table 3.5 Example of Real-Time Information Data

Product type	Origin	Day of Week	Distance Group	Scheduled Departure	Actual Departure	Departure Delay (minutes)	Estimated Elapsed Time (EET)	Scheduled Arrival	Estimated Time of Arrival (ETA)
P0	A	Monday	3	8:00 A.M.	8:05 A.M.	5	1h:22m	10:20 AM	10:27 A.M.
P3	C	Monday	..	..	..		..	..	..
P2	F	Monday	..	..	..	..	..	..	..
P1	K	Monday	..	..	..	..	..	..	..
P5	E	Monday	..	..	..	..	..	..	..

The real-time information must be checked with respect to the initial schedule of trucks to determine whether the truck would arrive accordingly. The information can be interpreted as Truck P0 is experiencing congestion on the road that would likely lead to late arrival according to the ETA value, then the predictive model would assign a presence probability value in regard to the corresponding class as it is shown in Table 3.6. Moreover the difference approach between the probabilistic and deterministic approach is also clearly shown.

Table 3.6 Example of Difference between Deterministic and Probabilistic of Truck Presence

	Presence Probability					
	Deterministic			Probabilistic		
	Early	On-Time	Late	Early	On-Time	Late
P0	0	0	1	0.1	0.2	0.7

In situation where the ETA information is considered deterministic, this arrival is simply classified as Class 2 (probability value = 1), consequently, it would immediately trigger rescheduling. However, in the case where the ETA information is assumed stochastic, the predictive model would provide the probability value of a class that a certain truck should belong to. In this example, it would indicate that there is 0.7 probability value of Truck 9 being classified as class 2 (arrive at minimum 5 minutes late). As it is explained earlier, this value will be considered as the presence probability of truck according to its ETA.

3.3.Chapter Overview

Relevant theories as the foundation to build the conceptual framework have been established. The operational magnitude of petrochemical supply chain requires a decision support system (DSS) to ensure more efficiency and profitability. Given the scope of study, DSS that integrates historical data and real-time information is required to enable advanced schedule management that could reduce problem stemming from uncertainty of truck arrival time.

Specific to the case of petrochemical loading facility, disruption related to transport network (congestion, road disruption, etc.) contributes the most to deviation of trucks from initial schedule. Real time information does not entirely eliminate the uncertainty of truck arrival time. To acknowledge this issue, a presence probability value that denotes the likelihood of truck being early, on-time , and late is assigned to corresponding ETA.

Given the randomness of the data, predictive model powered by ML algorithm is a good fit in making prediction. ML algorithm could identify subtle pattern/trend in historical data and is also able to understand the non-linear and complex relationship among parameters. According to the goal of the predictive model, classification algorithm will be used. Since there is no absolute guideline in determining best ML algorithm, three different classifier algorithms, namely GBN, LR, and ANN will be tested and evaluated.

Ultimately, based on the established theories and knowledge gap, a conceptual framework as the main tool to answer the Research Question 1 is built. The main elements of this conceptual framework is the predictive model (probabilistic classifier powered by ML algorithm) and optimization model (exact approach for rescheduling). The main inputs are the historical data and the real-time information. This conceptual framework allows checking whether a certain truck would be able to comply with its initial schedule based on the real-time ETA, therefore a certain truck could be classified as early, on-time, or late. According to the pattern in the historical data, the predictive model would assign a presence probability value for that certain arrival class, which then would be the input of the optimization model.

4. Data Analysis

In the previous sections, a conceptual framework of rescheduling system that aims to improve the operational efficiency of loading facility has been proposed. The main elements are the integration of the real-time ETA information and the historical data of truck arrival time. Since the real-time ETA is straightforward and provided by external provider, the only concern is the historical data. Historical data analysis must be done because it would serve as the basis to predict future outcomes.

This chapter provides sequential elaboration on input dataset used in this research. It consists of data generation phase where a historical dataset is synthetically generated, and data exploration that includes initial data analysis (IDA) and exploratory data analysis (EDA) where the dataset is pre-processed and analyzed using varied method. Hence, the purpose of this section is not only to deliver a hypothetical valuable insight derived from historical data, also to ensure that the final data has decent quality to be inputted into both predictive model and optimization model.

4.1.Data Generation

To verify the proposed method, a synthetic data is generated from the publicly available database because the real data cannot be obtained. Since data that specifically fit the context of this research, namely the truck scheduling in loading operation could not be found, it is decided to adapt and manipulate the closest data resembling the problem of this research, which is the historical data of airline on-time performance in US. The data is publicly accessible and can be downloaded in [70]. This record consists of almost 2.46 million rows of individual flight data that correspond to 31 variables of air transportation indicators.

The justification of using this sort of dataset is the availability of historical pattern of key parameters that have direct relation with delay prediction such as the complete list of ETD, ETA, actual departure, and actual arrival. Moreover, this dataset also provides other relevant data that could influence the delay to a certain extent such as the deviation between estimated travel time and actual travel time, time indicator (year, month, day, and day name), origin, destination, and the corresponding distance between origin and destination. Hence, the combination of this correlated data could possibly explain and improve the generalization of delay level.

4.2.Data Description

The synthetic dataset serves as the basis to train, validate, and testing the different ML algorithm. The original synthetic datasets contains arrival information of trucks in the period from 1 January 2016 until 31 March 2021. Since only 1 loading facility is being considered, there is only 1 single location for all the trips that represents the loading facility itself, therefore the full dataset is reduce to 141517 rows datapoints representing trip of trucks. Furthermore, each row contains information or features of the operational attributes, for instance, origin location, scheduled arrival, scheduled departure, elapsed time, distance, etc. The full list of all features are presented in Table 4.1.

In this research, the main subject is the truck which has been assigned a single particular product type to load. Specifically, in the dataset, the type of trucks are defined and according to their pre product type, instead of the name or the company or the index number of the truck. This is done because the focus of this research is not to emphasize on the performance of the truck provider, but more on the distribution of particular product types that possibly be loaded at a particular petrochemical loading facility. Hence, since there are only 8 product types are assumed available in the loading facility, the column ‘Truck’ would consist of 8 variation of values that directly corresponds to the predetermined product type. For example, ‘P0’ means truck that is assigned to load and carry a product type P0.

Table 4.1 Features in Historical Data

Features	Description	Specification
Product Type	A predetermined product type which each truck supposedly loads at the loading facility	The truck directly corresponds to its predetermined product type. Complete product type is provided in Appendix B: Synthetic Data.
Origin	The location that trucks depart from.	The set of origin is composed of 20 different locations. Complete list of origin name is provided in Appendix B: Synthetic Data
Destination	The location of loading facility.	In this case, loading facility is only one and indicated by 'FACILITY'
Distance group	The distance category between various origin to the loading facility.	The categories are based on a specific range of distance between origin and destination. The detail is shown in Appendix B: Synthetic Data
Year, Month, Day Name, Date	Time indicator of the corresponding schedule.	• Year ranges from 2016-2020 • Month ranges from 1-12 • Day of Week ranges from Monday to Sunday • Date ranges from 1 till 30/31 • Hour ranges 00:00 till 23:59
Scheduled Departure Time	Initial schedule of departure from origin	This feature is formatted in hour and minutes
Scheduled Arrival Time	Actual realization of departure from origin and arrival to destination.	This feature is formatted in hour and minutes
Arrival Delay	Deviation between scheduled and actual for both arrival and departure.	This feature is in minutes basis • Early arrival denotes by negative value • On-time arrival denotes by 0 • Late arrival denotes by positive value
Elapsed Time	Realization of travel time from a certain origin location at a certain time of the day/week	This feature is formatted in minute

4.3.Data Exploration

This section provides elaboration on the historical data that includes both initial data analysis and exploratory data analysis. The purpose of the IDA (Initial Data Analysis) is to assess the data quality in regard to errors, missing data, statistical outliers. The final output is the final form of historical data that is ready to be used in this research.

The main purpose of EDA (Explanatory Data Analysis) is to understand the relation between variables and underlying pattern of the historical data, therefore valuable insights related to the transportation and schedule management could be derived. Besides that, the result of exploration could also indicate whether the synthetic data is logical and valid to be used for model verification.

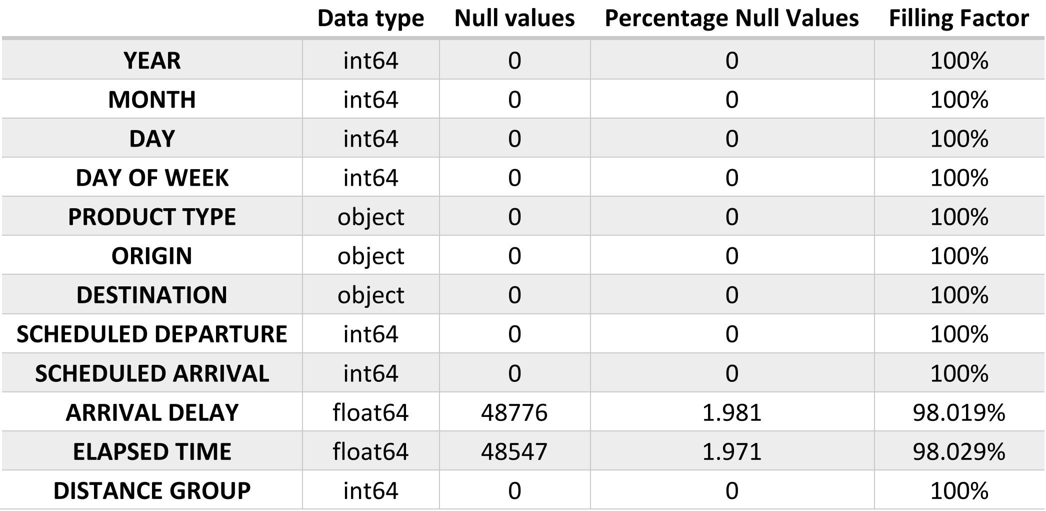
4.3.1.IDA (Initial Data Analysis)

Since there are irrelevant, empty, and outlier data points in the raw dataset, some pre-processing steps are required to avoid excessive noises that leads to bias result. The identification of these problems are important because the low data quality could heavily influence the prediction as this set of historical data is the main source of information for the ML algorithm. In the case where the missing values are not significant relative to the whole set of data, if necessary then the empty data cells and its corresponding data points would be dropped. Despite of the effort of ensuring this data is as ideal as possible, there will be inevitable bias since this dataset is retrieved from different domain. The pre-processing that will be done in this stage includes cleaning, formatting and filtering.

- **Data Cleaning**

The main process of this stage includes the elimination of noisy data that incorporate irrelevant and meaningless data that could not be interpreted by the machine learning algorithm. Furthermore, the data type, missing values, and filling factor of each deemed relevant attribute is checked. Table 4.2 shows that most of the attributes have complete data points, whereas the ones that have missing values still have roughly 98% completeness, therefore it is not necessary to fill the missing values and the data rows can be safely removed.

Table 4.2 Data Type and Missing Values of Features



	Data type	Null values	Percentage Null Values	Filling Factor
YEAR	int64	0	0	100%
MONTH	int64	0	0	100%
DAY	int64	0	0	100%
DAY OF WEEK	int64	0	0	100%
PRODUCT TYPE	object	0	0	100%
ORIGIN	object	0	0	100%
DESTINATION	object	0	0	100%
SCHEDULED DEPARTURE	int64	0	0	100%
SCHEDULED ARRIVAL	int64	0	0	100%
ARRIVAL DELAY	float64	48776	1.981	98.019%
ELAPSED TIME	float64	48547	1.971	98.029%
DISTANCE GROUP	int64	0	0	100%

- **Data Transformation**

The purpose of this process is to transform and manipulate the data in such a way that enables appropriate format suitable for subsequent processes. In this case, this first step is the time conversion to combine year, month, day, and day of the week data into the date time format which is convenient to work with. Another tine conversion issue to deal with is the raw ETA and ETD data is inputted as a float whereby the two first digits indicate the hour and the last two indicate the minutes.

- **Data Reduction**

In this stage, data reduction technique is applied to eliminate the inappropriate data and statistical outliers. Besides that, the purpose is also to decrease the data volume without reducing the data quality. As the result, computational efficiency is increased, therefore data exploration could be comfortably done.

Since the records consist of full year of arrival data that ranges from 2016-2020 , it results on extremely large dataset. It is decided to cut the data down to only contain the flight records of 1 month period for each year. In practice, the month of February is chosen, then arrival records of February for each year (2016, 2017, 2018, 2019, 2020) is extracted. The combination of this subset of data results on 90198 rows of data points. Other justification in limiting to a certain month period is to reflect a recurring trend in monthly basis which is deemed sufficient for a short-term planning.

To mimic the situation of having a single loading facility in a certain location, the destination data is also set to only include a certain destination. This action significantly reduces the size of the data because this initial data include all trips between 20 different origin and destination airports. Lastly, the data is normalized by filtering out the statistic outlier such as the extreme late arrival time that is defined as actual arrival time having deviation value higher than 120 minutes and the extreme early arrival time that is defined as actual arrival time having deviation value lower than -15 minutes. To check if the data is correctly transformed, a basic statistical analysis is conducted in which the result is shown in Table 4.3. It shows that the latest arrival of trucks is roughly around 2 hours, whereas the earliest arrival of trucks is less than 15 minutes. Furthermore, it also indicates that the most data points belong to the P0. In terms of mean arrival delay, Truck P0 has the lowest mean of delay, whereas Truck P4 has the highest mean of delay. The final set of historical datasets that will be furtherly used in this research is provided in Appendix B: Synthetic Data

Table 4.3 Basic Statistical Information of Dataset

Product Type	Arrival Delay			
	Min (min)	**Max (min)**	**Count**	**Mean (min)**
P7	-15	118	950	7.6
P6	-15	120	2295	7.3
P5	-15	119	3336	5.8
P3	-15	120	4760	7.2
P4	-15	120	4777	8.0
P2	-15	120	5854	6.0
P1	-15	120	11831	3.3
P0	-15	120	57116	1.7

4.3.2.EDA (Explanatory Data Analysis)

In this section , EDA is conducted to profoundly analyze and investigate the underlying characteristics of the large dataset beyond the obvious appearance and formal statistic method. The objective is to discover hidden patterns and anomalies. Visualization are used to provide better interpretation and mapping between different variables. With respect to the scope of this research, EDA could assure that the dataset is valid and to a sufficient extent, relative to the existing theories in literatures. Therefore, it is safe to utilize this synthetic dataset in verifying the proposed conceptual framework.

Moreover, it is interesting to take into account individual perspective in analyzing the data. According to the dataset, the highest number of data point correspond to the truck type of ‘P0’, thus this truck is being considered as the individual sample to be analyzed. As it is explained, the label of the truck means that the particular truck is assigned to load a certain product type whereby the origin location of the truck could be varied. To realize this concept, the historical data of a specific truck type with a specific origin is considered. In this sample, the data correspond to the truck type ‘P0’ and origin location ‘A’ are analyzed.

Some analysis technique is applied to provide information on this individual delay issue, as following:

- **According to Day of the Week**

In the perspective of day-to-day within a week, the historical data indicates non-significant variations in delay of truck arrival time. The box plot shown in Figure 4.1, visualizes the range of arrival delay is between approximately 4-3- minutes. The median of the arrival delay is around 15 minutes for all days in the week. Furthermore, the most severe condition relative the other days is Tuesday, in which the highest arrival delay is roughly 80 minutes. The results are perfectly in line with [71] which inferred that, unless there is a major accident in the traffic, the variability of delay would be similar in day-to-day comparison.

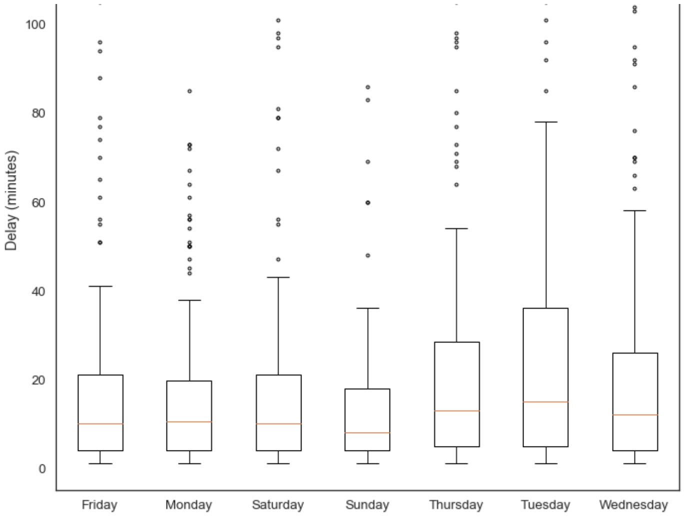


Figure 4.1 Delay Variability Between Days of the Week

- **According to Time of the Day**

In road transport network that includes the highway, random behavior of arrival time differd considerably depending on parts of day [71]. Furthermore, to check whether the synthetic historical data complies with the literatures, analysis that focuses on pattern of the delay that occurs in a day is conducted. To do so, the mean delay is aggregated according to the time of the day. The result is plotted as it is shown in Figure 4.2. This illustration would provide valuable information to comprehend the delay behavior in regard to the time of the day. As it can be seen, the peak delay tends to occur at later time of the day.

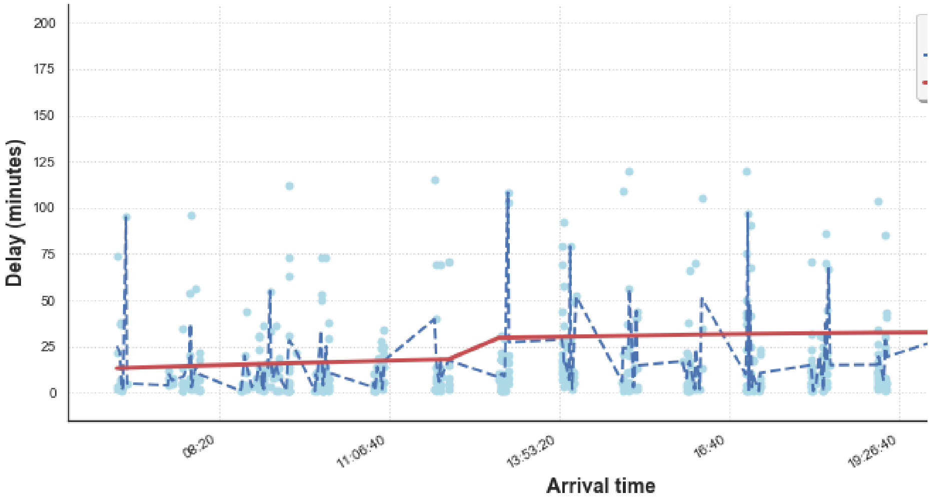


Figure 4.2 Delay Variability Within Day

4.4.Chapter Overview

Data exploration as the initial step in building predictive model, whereby the large unstructured historical dataset is deeply examined to discover hidden patterns, unique characteristics and important point of interests. This process successfully serves their intended goals which are to provide valuable insights in hypothetical sense derived from large historical dataset and to check the validity of the synthetic data with respect to logical sense and the relevant literatures. Hence, it can be concluded that the synthetic data is sufficient to be used as foundation in building predictive model.

The IDA process consisting of cleaning, transforming and filtering indicates that the historical data is statistically sufficient. Since the total missing values is not significant, all the data rows that contain can be safely removed. Moreover, data outliers such as extreme early arrival and extreme late arrival have been filtered out because the values are non-sense in this context of research and could probably distort the analysis. Finally, the data is fixed in which there are 90198 rows of data points with 10 parameters.

The EDA is designed to emphasize on the arrival time variability based on different parameters. Concerning time of the day, the results show that late arrival is likely to occur in the afternoon, while the early arrival is evenly distributed across the operational hours. Regarding time of the week, the median delay is almost the same for all days. Furthermore, comparing delay in monthly basis illustrates a similar pattern for each month in a year. In terms of product type and truck origin location, it is impossible to generalize the pattern since they are highly random. Ultimately, the results proves that the synthetic historical data is valid with respect to alignment to existing theories and logical sense of real word.

5. Predictive Model

In the previous sections, the historical data has been thoroughly explored. Some valuable insights with respect to the pattern of truck arrival are successfully derived. Moreover, The analysis prove that the synthetic data is sufficient, in the sense of representativeness of reality and alignment with precedent researches, hence will be applied as the input for predictive model.

In this chapter, a predictive model based on ML algorithm will be built. The first goal of this chapter is to correctly predict the probability of a certain truck being in a particular class of arrival based on set of parameters. Several learning algorithms are chosen and assessed on defined KPIs to the best predictive model. Moreover, the second goal of this chapter is to investigate the added value of real-time information with respect to performance of predictive model. To do so, different scenarios related to availability of real-time information will developed. Accordingly, a sensitivity analysis will be conducted.

5.1.Real Time Data Description

The basis of the predictive model is the historical data of truck arrival time. The utilization of real-time information enables obtaining extra parameters (labeled red) that could potentially provide more information for the predictive model. Therefore, in building the predictive model, two different historical datasets could be prepared, as it is shown in Table 5.1.

Table 5.1 Comparison of Features between Without and With Real Time Information

No.	Without real-time information	With real-time information
1	Product Type	Product Type
2	Origin	Origin
3	Destination	Destination
4	Distance **G**roup	Distance Group
5	Day of Week	Day of Week
6	Scheduled Departure Time	Scheduled Departure Time
7	Scheduled Arrival Time	Scheduled Arrival Time
8	Elapsed time	Actual Departure Time
9	Arrival Delay	Departure Delay
10		Estimated Elapsed Time (EET)
11		Estimated Time of Arrival (ETA)
12		Arrival Delay

Both datasets contains historical record of truck arrival time. Specifically, the first dataset contains only the common information of truck arrival, whereas the second dataset includes the real-time elements. The main difference between them is the real-time information provides more in-depth knowledge representing the uncertainty in logistic and transportation. The additional parameters enabled by the technology of real-time information is marked by red label. As it is clearly shown by the comparison of historical and real-time dataset, the distinction is the set of parameters from which the ML algorithm would try to generalize the relation, yet the prediction target is still the arrival delay.

For majority part in this chapter, datasets with real-time data will be utilized to produce a predictive model, while the other datasets will be used in sensitivity analysis.

5.2.Classification Techniques

As Section 3.1.6 justifies the chosen ML algorithms, the fundamental of each one of them is elaborated in this section.

5.2.1.Naïve Bayesian Classifier

In a simple term, Naïve Bayesian Classifier is a classification algorithm based on the Bayes Theorem. There are varied branches of model in this family in which all of them rely on fundamental that every feature is independent of each other and has equal contribution to the outcome . The drawback of this approach is the naïve design and apparently oversimplification, namely the assumption that all features are independent and have equal weight that are not truly precise in real-word situation. However, some researches proved that this flaw is quite negligible, thus this method still produces decent result in practice [72].

Bayes Theorem, also known as conditional probability, is defined as probability of an event occurring based on probability of another event that has already occurred [72]. Let a problem instance of classes, denoted by a vector $x = (x_1, \ldots , x_n)$ that represents n features (independent variables). The instance probability for each of K possible outcome (class Ck) can be formulated as:

$$p(C_k \mid x_1, \ldots, x_n)$$

Bayes' theorem is stated mathematically as the following equation:

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

In which, $p(C_k|x)$ is the conditional probability of event C_k occurring given the event x is true, also known as posterior. $p(x|C_k)$ is the conditional probability of event x occurring given that event C_k is true, also known as likelihood of event x given a fixed event C_k. Both $p(C_k)$ and $p(x)$ are the probability of observing $C_k$and x or known as prior and evidence, respectively without any given conditions, known also as prior probability. In mathematical formulation, this term can be written as:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Naïve Bayesian concept assumes that each class is distributed according to a Gaussian distribution. In generating prediction, this method follows Maximum Posteriori Assignment (MAP) rule in which it would attempt find to the most probable calculation of a certain class given the observed data the fit of distribution. The assignment of class label y is determined of predictive model is determined by the highest posterior probability value. This formula can be written as:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

5.2.2.Logistic Regression

According to [64], LR (Logistic Regression) is a classification algorithm that is underlined by the concept of logistic/logit function which naturally emerges from the ratio of normal probability density functions. The LR suits the case because of the assumption that the value of classes are normally distributed

Let t can be defined as the linear function of a single explanatory variable x. The t variable can be expressed, as following:

$$t = \beta_0 + \beta_1 x$$

As stated in [76], probabilistic classification model, p(x) is interpreted as the probability of the dependent variable given some linear combination of the predictors, β_0 is the constant value from the linear regression function, and $\beta_k \; x_k$ is the regression coefficient multiplied by some value of predictor. Regardless the value of his linear regression combination (t) that may vary from negative to positive infinity, the transformation process would result on the $p(x)$ ranges between 0 and 1. A extended general logistic function can be written as:

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots \beta_k \, x_k)}}$$

Normally, Logistic Regression algorithm only supports binary class classification problem in which it is able predict the True or False figure according to the sigmoid loss function. Although, this algorithm could be possibly extended to support the multi-class classification problems by transforming the loss function into the cross-entropy loss function and also by converting the classification problem into multiple binary classification problems.

Theoretically, LR can be modified to solve the multi-class problems by altering the type of probability distribution [64]. In this case, the multinomial probability distribution, instead of the binomial probability distribution that is applied in the binary case. In other words, logistic regression model that is adapted to learn and predict a multinomial probability distribution is referred to as Multinomial LR, and this is the one that is applied in this research.

To simply define it, Multinomial LR, also known as softmax regression is classification algorithm that considers target variable y that ranges over more than 2 classes, and the objective is to predict the probability of y being in each potential class $c \in C$, thus the probability that soft max function computes can be stated as $p(y = c|x)$. Similar to the normal Logistic Regression, this softmax function take a vector $z = [z_1, z_2, \dots, z_k]$ of k arbitrary values and plots them to a probability distribution, whereby each value ranges from 0 to 1 and the total of all values is 1. In other words, the denominator would normalize the values into a probability [77]. The softmax can be defined as following:

$$softmax(z) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \; 1 \leq i \leq k$$

Hence, by applying the equation to the vector $z = [z_1, z_2, \dots, z_k]$, the probability value can be calculated and can be written as following:

$$probability\,(z) = softmax(z) = \frac{e^{z_1}}{\sum_{i=1}^{k} e^{z_i}}, \frac{e^{z_2}}{\sum_{i=1}^{k} e^{z_i}}, \dots, \frac{e^{z_k}}{\sum_{i=1}^{k} e^{z_i}}$$

5.2.3.Artificial Neural Network (ANN)

- **Structure**

Neural Network – also known as ANN (Artificial Neural Network) is a network of connected perceptrons (artificial neurons) consisting of a function. The architecture of the network resembles a neural network in human's brain that could process information. A basic structure of neural network is shown in Figure 5.1. There are two important elements of neural network structure are the perceptron and the connector. The main structure of neural networks consists of three different layers, as following:

- Input layer: layer that receives the input data.
- Hidden layer: layers of mathematical functions that are designed to produce specific form of output to be passed through layers. This is also where the optimizing of weight and bias being done. A neural network's structure with more than 1 hidden layer of perceptron is categorized as DNN (Deep Neural Network).
- Output layer: layer that is responsible to concretely provide a final prediction result.

-

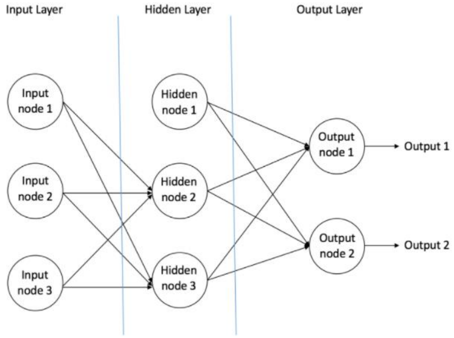


Figure 5.1 Basic Structure of Artificial Neural Network

As if a human's brain, this system could perceive input data, then process them through layers to generate appropriate result based on the existing knowledge. The knowledge is obtained through a learning process in which perceptrons are trained using variant of algorithm that rakes into account the error made by the networks. This is done iteratively, resulting on reinforcement process that would decrease the error. Specifically, perceptron is fed a set of training data whereby it would make its prediction. Subsequently, for every wrong prediction, it reiterates the connection weights that would supposedly result on correct prediction.

Basically, there are two main parts of the learning process, which are the front propagation and the back propagation. The flow of the learning process of the ANN is shown in Figure 5.2. Forward propagation refers to the process whereby the information is received by the input layer, passed through hidden layers and ended up at the output layer. Back propagation refers to the process whereby the error is calculated and passed backward, so the network can adjust the weight or bias to get a reduced error, in other words, better prediction result.

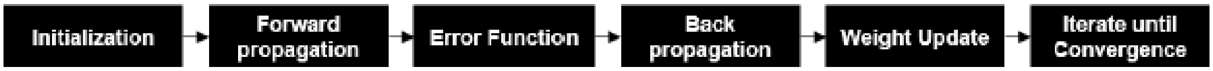


Figure 5.2 Flow of ANN Algorithm (adapted from: [64]

- **Activation function**

Activation function is a critical part of the Neural Network. Its main role is to define how the weighted sum of the inputs would be transformed into an output from nodes in a layer [78]. The choice of activation function would influence the type of prediction the model is capable of making, therefore different activation functions may be used at different parts of the layers the model. However, typically all hidden layers use a same activation function, whereas the activation function of output layer differ according to the type of prediction required by the model. Activation function allows restricting values from the nodes within a certain range and accommodating non-linearity element in ANN.

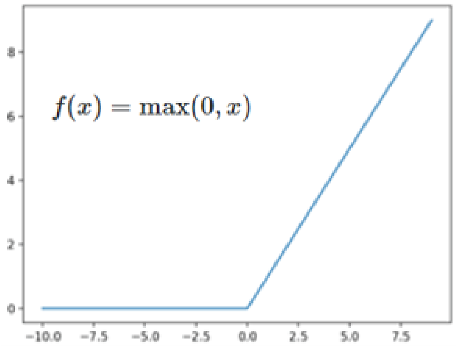


Figure 5.3 Plot of ReLU Function (adapted from: [78])

ReLU is the most popular because it is simple to implement and less computationally expensive. The formula of ReLU can be written as max(0,x), meaning that if the input value (x) is negative, then a value of 0 is returned, otherwise value of x is returned as it is shown in Figure 5.3. Therefore, this function guarantees a node can only be activated when the value is positive. Given the advantages and good fit to the problem, ReLU function will be applied in this research.

- **Hyperparameter**

Aside from choosing the suitable activation function, some variables explicitly determines the network structure known as hyperparameters. The hyperparameters include characteristic such as number of hidden layers, batch size, , number of epochs and learning rate. Intuitively, total hidden layers describes the exact quantity of layers between input and output layer. The number of layers would impact the fitness of the model. Learning rate defines how quickly a network updates its parameter. Low learning rate results on slow process but better, while higher learning rate results on fast process but worse convergence. Number of epochs refers to how many cycles in which training data is run through the network while training. Batch size defines the number of sub sample of training data used in one iteration.

All those stated parameters can be tuned to obtain a better prediction. However, due to some limitation in computational capability in this research, there will be no hyperparameter tuning that will be done. In other words, the model would accept the default setting provided by the module. The constructed predictive model is considered as the baseline model that could potentially be improved in the future.

- **Loss function**

Loss function is defined as evaluation algorithm that would measure how well the predictive model performs by calculating the error during the optimization process. The loss function is chosen based on the requirement of the model that is categorized as regression model or classification model. Since this predictive model aims to be a classification, then the explanation would focus on the common loss function used in classification model. To that end, the most popular loss function is the cross entropy/log loss Figure 5.4.

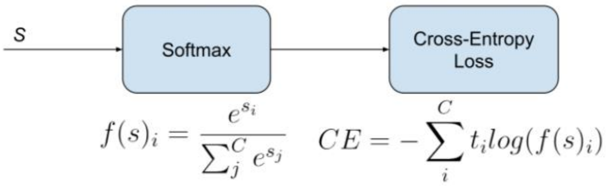

s
Softmax
Cross-Entropy Loss
$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}}$
$CE = -\sum_i^C t_i log(f(s)_i)$

Figure 5.4 Loss Function for ANN, adapted from [64]

Although the underlying mathematic is the same, many literatures usually use the term of log loss for binary class problems, while the cross-entropy for the multi-class problems. Since the predictive model developed in this research is a multi-class problem, then the Cross-Entropy loss function will be implemented.

- **Solver**

The optimizer is an optimization algorithm used to adjust the attribute of the neural network – weight, bias, learning rate, etc. –to minimize the losses with regard to the training data set, therefore the model would be able to achieve its optimal performance in terms of speed and accuracy of results. One of the most popular optimizers is the stochastic gradient descent.

Calculating the gradient of the target function with regard to the precise values of the input values is the foundation of this approach. Because the gradient is uphill, the negative of each input variable's gradient is followed downhill, resulting in new values for each variable and a lower assessment of the target function. The gradient is scaled using a step size. This method is repeated until the goal function's minimum is found, a maximum number of potential solutions is assessed, or another stop condition is reached. Since the gradients of the target function are normally noisy, the algorithm is called "stochastic" (e.g., a probabilistic approximation)

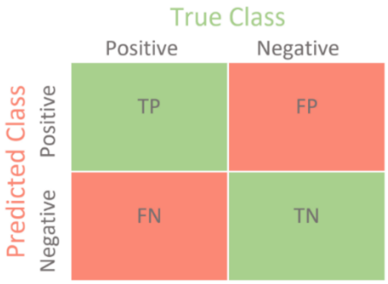
Another more advanced algorithm is Adaptive Moment Estimation (ADAM). Adam optimization is an improved Stochastic Gradient Descent (SGD) in which the method is based on adaptive estimation for each parameter that provides more advantages compared to the normal SGD. This research will use ADAM method.

5.3.Performance Metric

The following metrics are used to evaluate the performance of the classifier model [64]:

5.3.1.Confusion Matrix

As it is illustrated in Figure 5.5, Confusion matrix is a technique to summarize the performance of a classification model that is based on the counts of the records correctly predicted and incorrectly predicted by the algorithm. The parameters used in this evaluation method is the True Positive, False Positive, True Negative, and False Negative. Given the simple interpretation of it, high-performing classification model would have a dominant result located in the diagonal of the matrix.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Figure 5.5 Illustration of Confusion Matrix

The description of the parameter is provided in Table 5.2, as following:

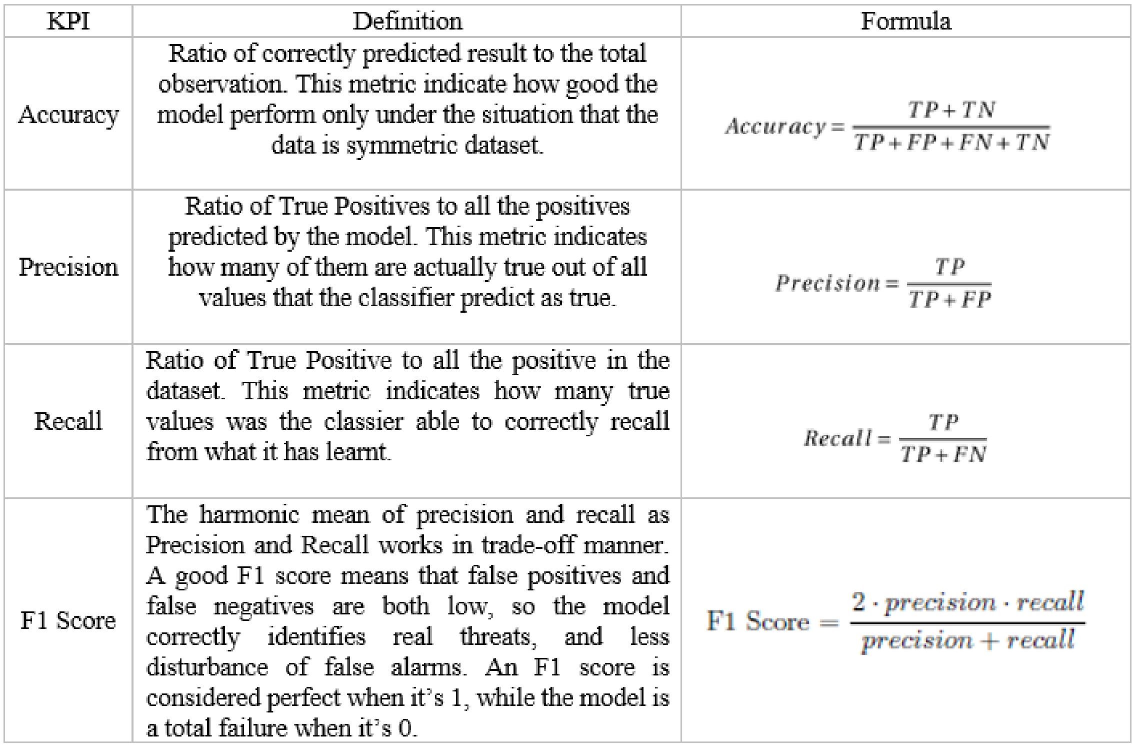
Table 5.2 Description of Parameter in Confusion Matrix

Parameter	Description
TP	Observation is positive, and prediction positive
FN	Observation is positive, yet prediction is negative
TN	Observation is negative, and prediction is negative
FP	Observation is negative, and prediction is positive

5.3.2.Accuracy, Precision, Recall & F1 Score

A set of commonly used KPIs to evaluate the ML performance is shown in Table 5.3

Table 5.3 Set of Standardized KPIs



KPI	Definition	Formula
Accuracy	Ratio of correctly predicted result to the total observation. This metric indicate how good the model perform only under the situation that the data is symmetric dataset.	$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$
Precision	Ratio of True Positives to all the positives predicted by the model. This metric indicates how many of them are actually true out of all values that the classifier predict as true.	$Precision = \frac{TP}{TP+FP}$
Recall	Ratio of True Positive to all the positive in the dataset. This metric indicates how many true values was the classier able to correctly recall from what it has learnt.	$Recall = \frac{TP}{TP+FN}$
F1 Score	The harmonic mean of precision and recall as Precision and Recall works in trade-off manner. A good F1 score means that false positives and false negatives are both low, so the model correctly identifies real threats, and less disturbance of false alarms. An F1 score is considered perfect when it’s 1, while the model is a total failure when it’s 0.	$\text{F1 Score} = \frac{2 \cdot precision \cdot recall}{precision + recall}$

5.4.Pre-processing

Referring to the current form of the dataset shown in Section 4.3.1 , pre-processing steps are required to assure that the data is adequate and suitable to be inputted into ML algorithm. The pre-processing are applied to historical datasets with and without the real-time ETA information.

5.4.1.Confusion Matrix between Features

To examine a relation between features, a Pearson correlation matrix is constructed as it is shown in Figure 5.6. Further explanation about the Pearson correlation matrix can be found Appendix C: Predictive Model. The value of the correlation matrix ranges between 0 and 1. Value 0 indicates no correlation between two variables, while value 1 indicates perfect correlation between two variables. Alternatively, it could also be utilized as a proxy to diagnose whether the data is logical or not.

For example, the relation between ‘Elapsed Time’ and the ‘Distance Group’ is at the high of 0.96 which indicates that both variables are strongly correlated, meaning that a more distance would result on longer time to travel between the distance. Hence, the relation between variables could point out whether the dataset yields valid and logical behavior concerning the real-world events.

With respect to the input for ML algorithm, correlation between features should be thoroughly examined. Removing features that do not provide significant knowledge in the learning because could make training process more efficient. This phenomenon is known as multicollinearity problem which could be identified by looking at the features having high correlation value. The justification is that multicollinearity causes a very similar pattern between those variables and one feature can be linearly predicted from the others with a high degree of accuracy; therefore it is unnecessary to include both features in the training process.

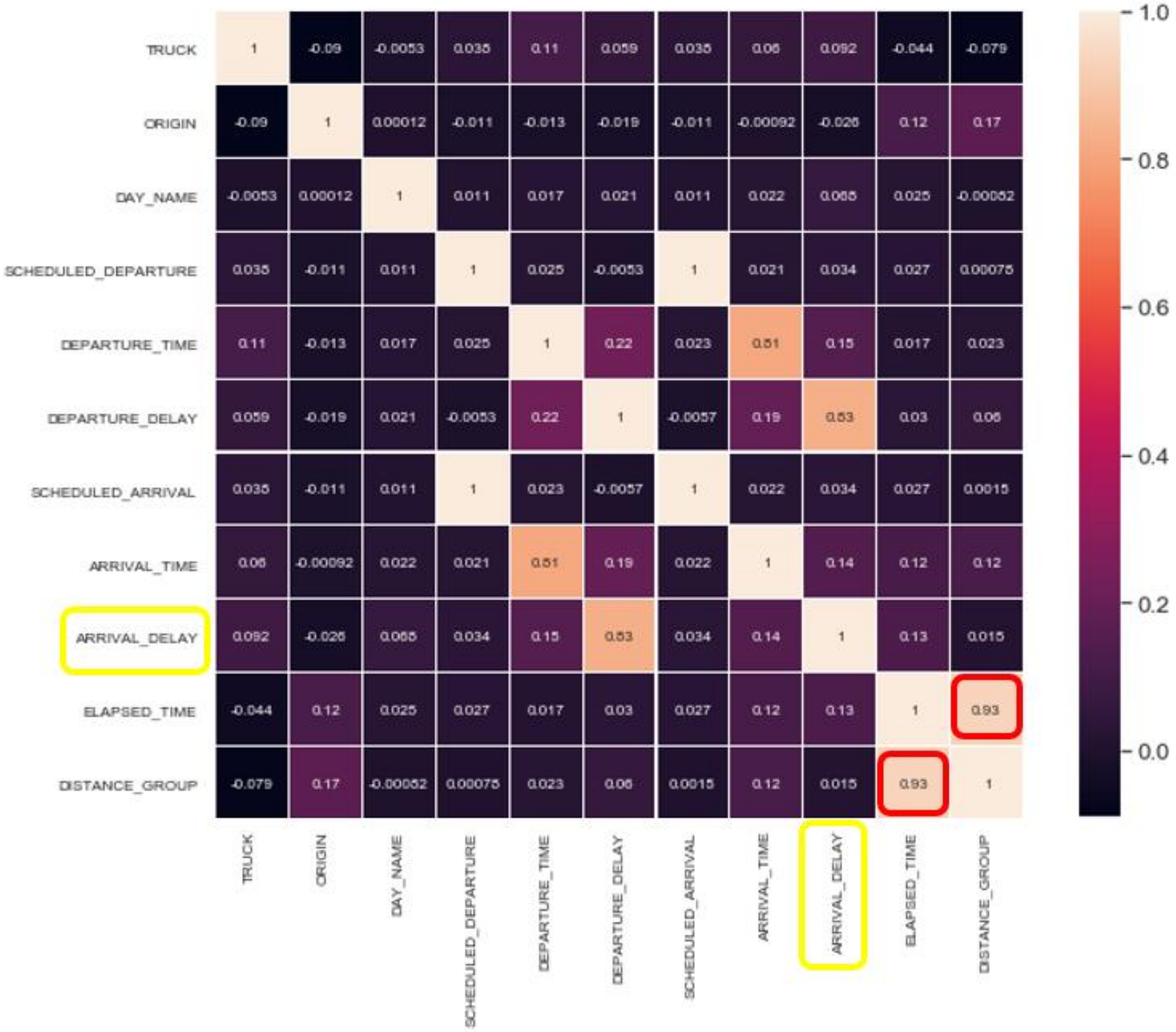


Figure 5.6 Pearson Correlation Matrix between Features

5.4.2.Feature Selection

In this section, some unnecessary features are removed according to the result of Pearson Correlation Matrix. The purpose of this step is to ensure that the dataset only consists of the data that contribute the most, thus would result on more efficient process and more accurate prediction. As it is explained, high collinearity provides more intuitiveness for the ML algorithm in predicting output because the value of a feature would be dependent of another. When a pair of features yields a correlation of 0.8 or higher, then one of the related features can be safely removed because the presence of both would not improve the model's ability to learn and make prediction. Consequently, in this research, "Elapsed Time", and "Distance Group" have a high positive correlation, therefore it is reasonable to eliminate two of them. In this case, "Distance Group" are removed. The reason in choosing the "Elapsed Time" that refers to the actual travel time because the value is volatile, representing the disruption in transport network, . Moreover, apart from it, there is no extremely high correlated features in the dataset, therefore it is well-founded to assume this dataset would provide optimal learning for the ML algorithm

5.4.3.Feature Encoding

Since ML algorithm is unable to process string or date-time data type, therefore all values in dataset must be converted into numerical value. In other words, feature encoding process is required . This is done using a built-in Python library called Label Encoder, whereby features that are in form of string are consistently converted into number values.

5.4.4.Data Scaling

due to the lack of ability to deal with features that vary in magnitudes, the dataset is scaled. When feature scaling is not done, there is a tendency of ML algorithm to perceive greater values higher than smaller values, regardless of the unit of values. To solve this problem, standardization that transforms the data to ensure mean value of 0 and the standard deviation of 1 is applied.

5.5.Model Training

This section provides the result of the probabilistic classifier model and analysis on the performance of the chosen ML algorithm based on the defined evaluation metrics.

5.5.1.Feature Creation of Prediction Target

In line with the explanation in Section 3.2.2, the proposed system is designed to trigger rescheduling based on the arrival class which are early, on-time, and late arrival. In order to realize this concept, new feature called "Delay Level" is created as the target/output of the predictive model. In this case, the terms of being present could possibly be denoted by the arrival class of either "0" and "1", and "2" in which the Class 0 means that the truck would arrive earlier, the Class 1 means that the truck would arrive on time, and the Class 2 means that the truck would arrive late. The definition and requirement to be classified as certain class is defined in Table 5.4.

Table 5.4 Requirement for Arrival Classification

Class	Requirement
0 = Early Arrival	ETA indicates that the truck would arrive more than 30 minutes early than initial schedule. $(\boldsymbol{Actual - Scheduled < -30})$
1 = On-Time	ETA indicates that the truck would arrive less than 30 minutes early up to the start initial schedule. $(\boldsymbol{-30 \leq Actual - Scheduled \leq 0})$
2 = Late Arrival	ETA indicates that the truck would arrive late than the start of initial schedule. $(\boldsymbol{Actual - Scheduled > 0})$

The proportion of the arrival class for eacth product type is checked to ensure that the data is sufficient. As it is shown in Figure 5.7, the distribution pattern of arrival class is quite similar across different product types in which the on-time arrival is more dominant compared to late arrival and early arrival. It also points out that the largest volume of data belongs to the product type P0.

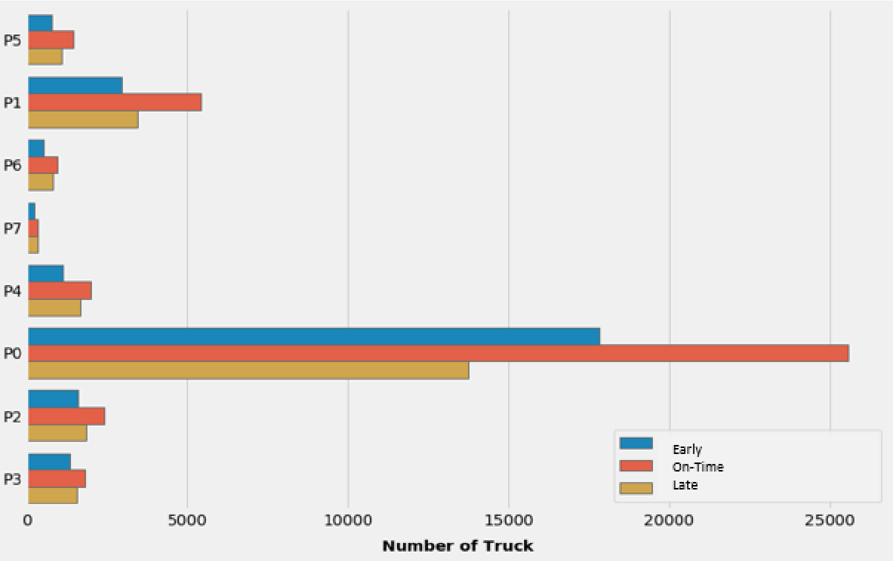


Figure 5.7 Proportion of Arrival Class per Truck

5.5.2.Splitting and Balancing Data

Since the current dataset is already simplified and encoded to assure that the dataset is perfectly compatible with the machine learning algorithm, the next step is to determine the input features and the target for prediction. As the objective is to predict the delay class, thus the Delay Level is separated from the main dataset. This subset would be converted into new variable that represents the targeted class (y) which the model would try to predict.

Subsequently, the historical dataset would serve multiple purpose, as the ML model is trained, and validated/tested on it. To do so, the full dataset must be split into multiple subsets with varied ratio. The training set is described as the source of knowledge data that would be used to fit the model, while the validation/testing set is the holdout/unseen data set that would be used to provide unbiased evaluation of the model performance based on the defined metrics. Although being commonly used interchangeably, the slight difference between validation and testing is the context of doing it, the former is used for tuning model hyperparameter, whereas the latter is used to assess the final model fit. In this research, tuning model hyperparameter is disregarded as the default setting is preferred (Appendix C: Predictive Model), so there is no urgency to differ the testing and validation stage and they will be considered as the same process. It is decided to split the full historical data into training set and testing set t with ratio of 80% and 20%, respectively. The clear description of data category for training, validating, and testing is shown in in Table 5.5.

Table 5.5 Description of Proportion of Historical Data in Training and Validating Model

Timeline	Ratio	Variable	Description
2016-2020	80%	X_train	Input features data for training the model
		y_train	The actual output for training the model
2016-2020	20% (holdout)	X_validate	Input features data for testing the trained model
		y_validate	The actual output for testing the trained model (this would be compared with the y_pred)

The balance of the training data is checked. The class proportion contained in variable X_train is not balance. This concern should be addressed because otherwise the algorithm could predict 'class 1' all the times and still get high performance score. A common technique to solve this imbalance issue is called Over-Sampling technique. Specifically in this case, a built-in library called Synthetic Minority Over-sampling Technique (SMOTE) is used. It allows increasing the minority class sample by generating 'fake' samples to resemble the minority samples, thus it would balance the class distribution. The final result after applying this technique is shown in Table 5.6.

Table 5.6 Initial Imbalance Data and After Implementation of SMOTE Method.

Class	Original Count	SMOTE Count
0 (Early)	31955	31955
1 (On-Time)	21186	31955
2 (Late)	19594	31955

5.5.3.Model Fitting

Model fitting is a metric for how effectively a machine learning model generalizes input that is comparable to that with which it was trained. When given unknown inputs, a good model fit refers to a model that properly approximates the output. To build a machine learning model, an algorithm is performed on data for which the target variable is known. The model's outputs are then compared to the actual data to investigate the accuracy.

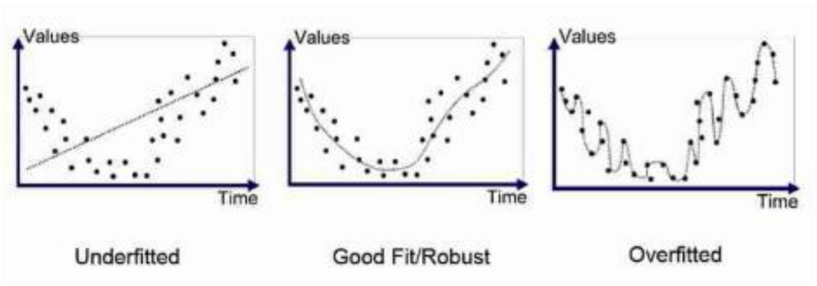


Figure 5.8 Pitfalls in Model Fitting [64]

As it is shown in Figure 5.8, there are two common causes of poor performance of ML model, which are underfitting and overfitting. The ideal goal in building the machine learning is to have good fit/robust model. In statistical term, underfitting is defined when the model could not capture the complex behavior or underlying trend of the data, whilst overfitting is defined when the model inappropriately captures too much data that includes the noise in the dataset, resulting on overgeneralization.

There are several possible solutions to tackle the overfitting problem, as following:

- Simplify the model by selecting less parameters or attributes.
- Obtaining more data.
- Eliminating the noise in training data.

There are several possible solutions to tackle the underfitting problem, as following:

- Opting for more powerful model.
- Inputting better features to learning algorithm. This could be done by better feature engineering.
- Reducing the constraints on the model.

This process utilizes the Python library called “Scikit learn” to fit the training data. Two different learning algorithms which are the Naïve Bayesian Classifier and Multinomial Logistic Regression are applied to the fit the X_train with the y_train data. After the fitness process, the model can be considered as trained and is ready to produce prediction. As this experiment utilizes a rather smaller database, it is likely to make ML-based predictions more inherently challenging and prone to overfitting.

5.6.Model Validation

Last step in building the ML-based classification model is the testing process. The testing utilizes the holdout/unseen 20% of the data (Table 5.5) as it is previously explained. In this example, the output variable (y_test) is used as the threshold to evaluate the predicted output (y_pred). As one of the goals of this chapter is to find the best way to build a predictive model, therefore 3 different ML algorithms are chosen and will be tested on the dataset.

5.6.1.Result

- **Naïve Bayesian Classifier**

According to the KPIs (Figure 5.9 and Table 5.7), the Naïve Bayesian Classifier generated unreliable predictions that contain large amount of error in the form of both False Positive and False Negative, as it is shown by the Accuracy score of only 0.50, indicating that the model mistakenly predicted almost half of total prediction.

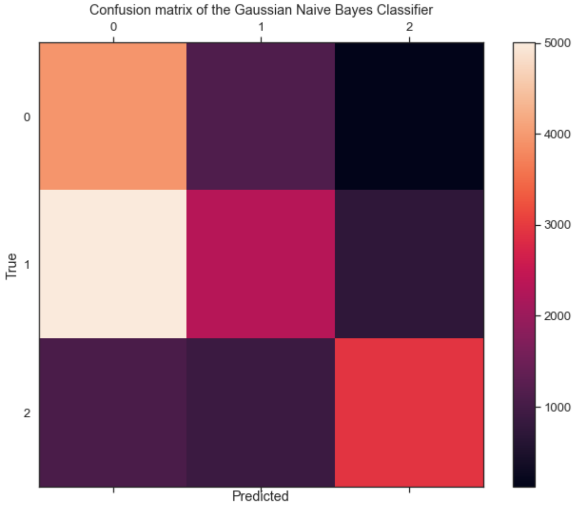
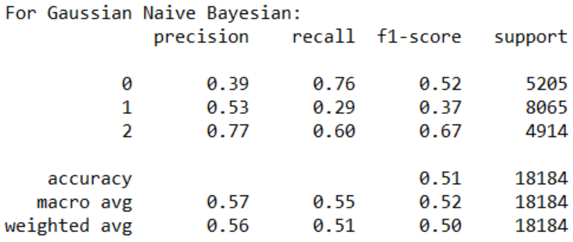


Figure 5.9 Confusion Matrix for GBN

Table 5.7 Evaluation Result of GBN based on KPIs



```
For Gaussian Naive Bayesian:
              precision    recall  f1-score   support

           0       0.39      0.76      0.52      5205
           1       0.53      0.29      0.37      8065
           2       0.77      0.60      0.67      4914

    accuracy                           0.51     18184
   macro avg       0.57      0.55      0.52     18184
weighted avg       0.56      0.51      0.50     18184
```

- **Logistic Regression Classifier**

According to the KPIs (Figure 5.10 and Table 5.8), the Multinomial Logistic Regression algorithm resulted on better performance in which more than half of prediction are correct as it is indicated by the Accuracy score of 64%. The major errors come from the misclassification of early arrival (Class 0) and late arrival (Class 2) as the on-time arrival (Class 1)

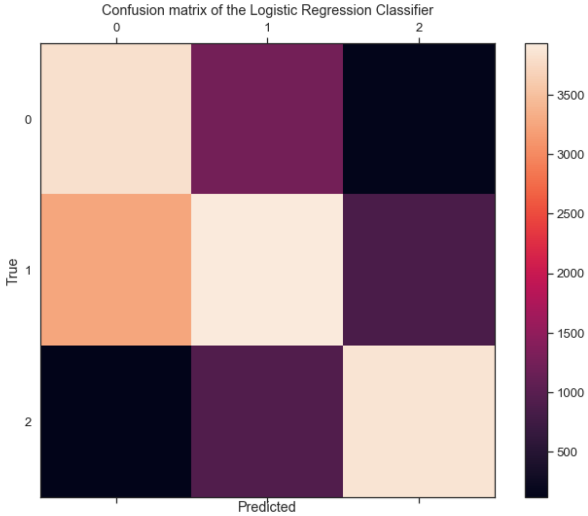
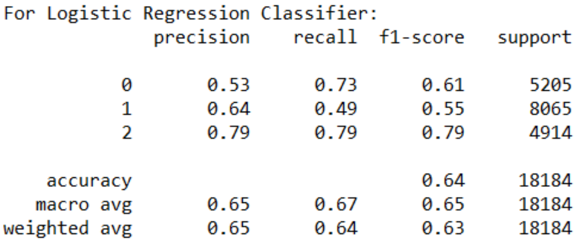


Figure 5.10 Confusion Matrix for LR

Table 5.8 Evaluation Result of LR based on KPIs



```
For Logistic Regression Classifier:
              precision    recall  f1-score   support

           0       0.53      0.73      0.61      5205
           1       0.64      0.49      0.55      8065
           2       0.79      0.79      0.79      4914

    accuracy                           0.64     18184
   macro avg       0.65      0.67      0.65     18184
weighted avg       0.65      0.64      0.63     18184
```

- **Artificial Neural Network**

Lastly, the ANN model outperforms all other models in every KPI (Figure 5.11 and Table 5.9), thus it can be concluded that this is the best model out of all tested algorithms. It correctly predicted approximately 70% of the total data according to its Accuracy score. It shows equal ability in predicting both early and later arrival with roughly 72% correctness in Precision and Recall.

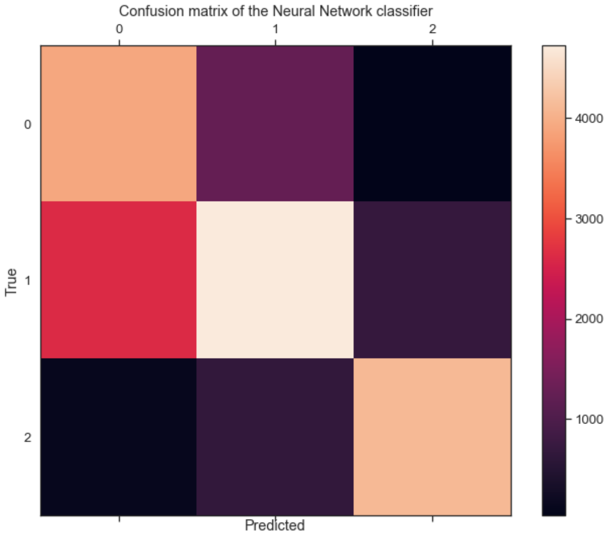
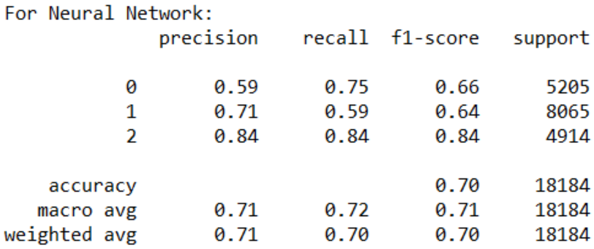


Figure 5.11 Confusion Matrix for ANN

Table 5.9 Evaluation Result of ANN based on KPIs



```
For Neural Network:
              precision    recall  f1-score   support

           0       0.59      0.75      0.66      5205
           1       0.71      0.59      0.64      8065
           2       0.84      0.84      0.84      4914

    accuracy                           0.70     18184
   macro avg       0.71      0.72      0.71     18184
weighted avg       0.71      0.70      0.70     18184
```

5.6.2.Analysis

According to the results of the validation process, the ANN outperforms the other two algorithms (GBN and RF) in all defined KPI metrics. In terms of Accuracy, the ANN algorithms score 72% accuracy which is the highest among the other ML algorithms. Moreover, as stated earlier, this research considers equal importance for precision and recall metric, thus ideally, all KPI scores should be higher than 0.5, meaning that there are more True Positives than False Negatives (Recall) and more True Positives than False Positives (Precision). According to the Table 5.7, Table 5.8, and Table 5.9 all ML algorithms fulfill this ideal criterion.

To benchmark the result of the predictive models, a comparison with other predictive models found in similar studies is conducted. The ANN's accuracy of 70% is sufficient and competitive compared to predictive model proposed in [17], [20], [40], [43], [45]–[47], [61], [79], [80] whose accuracy is approximately ranges from 40% - 80% Considering the complexity level, the proposed predictive model even outperforms the ones in which the target classes are binary. However, this could not be stated with absolute certainty because the results could be affected by many factors, such as data quality, context, computational power, etc. Despite of the potential flaws , it still can be concluded that predictive model proposed in this research is acceptable.

All being said, It is justified to state that ANN model is best possible model to be applied to the conceptual framework. Subsequently, a more detailed explanation on how to implement and incorporate this predictive model in integration to the optimization model will be provided in the last section.

5.7.Sensitivity Analysis

In this section, a sensitivity analysis is conducted to measure the added value of real-time information with respect to the result of predictive model.

5.7.1.Scenario

Since feature combination inputted in the training phase would heavily affect the performance of predictive model, it is necessary to test the model under different setting. Moreover, in line with one of the goals of this chapter which is to observe the added value of real-time ETA information, some scenarios consisted of different feature combination is developed and its corresponding intention is explained in Table 5.10.

Table 5.10 Description of Scenarios to Investigate the Added Value of Real Time Information

	WITHOUT ETA INFORMATION	WITH ETA INFORMATION
FEATURES	• Historical: Truck, Origin, Day Name, Scheduled Departure Time, Scheduled Arrival, Scheduled Travel Time, Distance.	• Historical: Truck, Origin, Day Name, Scheduled Departure Time, Scheduled Arrival, Scheduled Travel Time, Distance. • Real-Time: Actual Departure, Estimated Arrival Time, Estimated Elapsed Time.
CONTEXT	This scenario simulates situation where only historical data consisting relevant parameter except the ones that might directly indicate the arrival time are available.	This scenario simulates the situation where real-time ETA information on top of the standard features are available.
INTENTION	This scenario aims to investigate the added value of applying historical data for predictive analysis.	This scenario aims to investigate the added value of real-time ETA and historical data for predictive analysis

5.7.2.Comparison

Given the set of scenarios a comparative study is conducted to investigate the added value of real-time ETA information in improving the predictive model. For each defined KPI, a macro average is calculated. The evaluation result of predictive models under different scenarios are presented in Table 5.11.

Table 5.11 Result of Comparison Between Scenario of Without and With Real Time Information

		Without real-time Information	With real-time information
GNB	Accuracy	0.36	0.51
	Recall	0.41	0.55
	Precision	0.38	0.57
	F1	**0.35**	**0.52**
LR	Accuracy	0.45	0.64
	Recall	0.48	0.67
	Precision	0.46	0.65
	F1	**0.44**	**0.65**
ANN	Accuracy	0.5	0.70
	Recall	0.51	0.71
	Precision	0.5	0.72
	F1	**0.51**	**0.70**

5.7.3.Evaluation

Based on the result of the sensitivity analysis, the ANN algorithm trained using the historical dataset that contains real-time information is best performing predictive model. Importantly, the comparison indicates a significant improvement across all ML algorithms compared if the real-time information is incorporated, to the ones trained based on dataset without ones. The real-time features enhances the prediction results by approximately 20% increase in all metric of defined KPI. Therefore, the real-time information certainly provides an essential added value to the predictive model.

Another key finding derived from the sensitivity analysis is that the inclusion of feature having a larger than 0.5 correlation value (‘Departure Delay’) with the prediction target (‘Arrival Delay’) in the learning process is important to get the potentially best result. Hence, it is important to note that in building a predictive model, all features with significant correlation must be selected.

5.8.Model Implementation

As it is explained, the benefit of using multiclass probabilistic classifier is the ability predict a class where given inputs likely belongs to, based on the prediction class having highest probability value. In addition, it also provides to probability of predicted classes which is deemed essential in the proposed conceptual framework of rescheduling system (Section 3.2.2). According to the evaluation on the previous section, the best ML algorithm to fit the given historical dataset of truck arrival is the Artificial Neural Network (ANN), therefore it would be applied on the chosen sample dataset of truck arrival at the loading facility which aims to verify the conceptual framework.

The sample dataset is the real-time ETA information of set of trucks scheduled to arrive in that certain day, including all the features inputted in the training dataset. The snippet of the sample data is shown in Table 5.12, while he full sample data is shown in Appendix C: Predictive Model. The full sample dataset correspond to the set of trucks initially scheduled to load products at that day. There are a total of 42 trucks that are already assigned to certain timeslot and loading bay to load its predetermined product. For example, in the, the row illustrate the identification of a truck and its relevant attributes of information, in which the

Truck with index 0 is tasked to load Product Type P5. The truck has departed from origin ‘A’ on Monday 8 February 2021 at 5:21. Based on the real-time information, the truck is estimated to arrive at 7:03. Moreover, the real-time flow of information also includes the actual travel time since the truck has departed, and the category of the origin location with respect to distance to the loading facility.

Table 5.12 Description of Features Included in Real Time Information

INDEX	PRODUCT TYPE	ORIGIN	DATE	DAY	DEPARTURE TIME	ESTIMATED ARRIVAL TIME	ESTIMATED ELAPSED TIME	DISTANCE GROUP
0	P5	A	2/8/2021	Monday	5:21:00	7:03:00	103	3
1	P1	C	2/8/2021	Monday	5:05:00	6:57:00	110	2
2	P6	F	2/8/2021	Monday	5:18:00	7:01:00	89	3
3	P0	G	2/8/2021	Monday	5:03:00	7:10:00	135	3
4	P4	D	2/8/2021	Monday	5:41:00	7:27:00	103	3
5	P0	I	2/8/2021	Monday	5:41:00	7:29:00	108	3
6	P5	O	2/8/2021	Monday	6:59:00	8:29:00	106	2
7	P3	F	2/8/2021	Monday	6:37:00	8:38:00	109	3
8	P7	G	2/8/2021	Monday	6:48:00	8:55:00	132	4
9	P3	G	2/8/2021	Monday	6:38:00	8:41:00	126	4
10	P4	A	2/8/2021	Monday	7:21:00	9:21:00	122	4
11	P0	C	2/8/2021	Monday	7:32:00	9:05:00	106	2
12	P6	K	2/8/2021	Monday	7:51:00	9:39:00	88	3

The sample dataset is inputted to the predictive model as the target data. In other words, the predictive model would try to predict the probability of a certain data row being a certain class which class (early, on-time, or late) based on the historical pattern learnt by the ML algorithm. The classification of arrival class follows the requirement explained in Table 5.4. The snippet of prediction result is shown in Table 5.13.

Table 5.13 Final Output of Predictive Model

Index Truck	**Early**	**On Time**	**Late**
0	0.675	0.286	0.038
1	0	0	1
2	0.582	0.11	0.308
3	0	0.001	0.999
4	0.451	0.47	0.08
5	0.56	0.345	0.095
6	0	0	1
7	0.261	0.392	0.346
8	0.356	0.291	0.353
9	0.481	0.326	0.193
10	0.655	0.273	0.072
11	0	0	1
12	0.827	0.073	0.1

As it is described on Section 3.2.3, the probability value will be incorporated into the optimization model as a way to capture the uncertainty of truck arrival time.

5.9.Chapter Overview

This chapter aims to present a predictive mode as an approach to answer the Research Question 2. Accordingly, probabilistic classifier powered by ML algorithm has been built and evaluated. Since there is no clear guideline in determining which ML algorithms would provide the best result. It is decided to explore 3 different algorithms, namely GBN, LR, and ANN. To increase the efficiency of the predictive model, the features are selected according to their correlation value to minimizes the negative impacts caused by multicollinearity problem. The evaluation results according to the standardized KPI shows the ANN is the best algorithm to fit the input historical data with overall F1 score of 71%. Relative to other tested models, the ANN model outperforms the LR model and GBN model by averagely 5% and 20%, respectively, in all the KPIs metrics. A benchmark done by comparing the result with other similar studies highlights that the ANN model has a better performance although it could be affected by many factors. Importantly, to serve the first goal of this chapter, the best performing and most suitable predictive model to be applied in this research is the ANN model.

As the second goal which is to see the extent to which real-time information would improve the predictive model, 2 different scenarios of features combination is applied. The first one represents a situation where historical data only consists of conventional information, while the second one represents a situation where parameter of real-time information is incorporated. The results clearly indicates that addition of feature enabled by incorporation of real-time information enhance the predictive model by roughly 20%. Hence, it amplifies the importance of implementing advanced collaboration of logistic planning and real-time traffic data that allows transferring real-time information.

Lastly, an example on how the probabilistic data could be incorporated as the supporting tool is explained. The result clearly highlights the relevance and compatibility of using this predictive model as part of the rescheduling system.

-

6. Optimization Model

In the previous sections, the predictive model has been successfully built. The main result is the presence probability of truck based on the historical data. According to conceptual framework, the value will be the input for the optimization model.

This chapter aims to develop optimization model in which the objective is to increase the operational efficiency of the loading facility that considers arrival time as stochastic variable. Technically, the model will adapt an exact approach and will be based on the concept of MIP (Mixed Integer Programming). To incorporate the stochastic element, which is the truck arrival time, a concept of expected value will be utilized.

6.1. Expected Value

Several parameters in transportation system are not always crisp/precise but are rather fuzzy/random. This has been an issue in optimizing the problem. Intuitively, to solve the problem, the random variable could be transformed into its crisp equivalence. To do so, there are some appropriate methods, such as expected value technique 80. In the context of optimization under uncertainty, chance-constrained programming is a common method in dealing with random parameter in which it ensures that the probability of complying with a certain constraint is above a certain level, as demonstrated in [80]–[82]. Normally, it requires some sort of nonlinear optimization or special analysis technique that is extremely dependent on corresponding probability distribution, thus it is often difficult to solve. Another alternative in solving the random parameter is called expected value constraint in which it describes a bound on the expected value of a random parameter. The main features of the expected value constraint are that it combines second-stage decision variables from several scenarios into a single constraint.

As it is stated in [82], expected value of a random variable is defined as a generalization of the weighted average of a large number of realizations of that certain random variable. The weight could be interpreted as the probability of random variable taking a specific value. The expected value could be derived by multiplying the alternative outcomes with the probability that each of them will occur, and then adding all of those values together., as following:

$$EV = \sum P(X_i)\, Xi$$

In a more general term, expected value could be defined as a way to assign a measurement value based on its probability of occurrence to a random variable/parameter. In the context of this research, the uncertainty parameter is the truck arrival that is classified into 3 different classes, namely early, on-time, and late. As it is described earlier, probability value generated by the predictive model is assigned to each class. To fit the concept of expected value into this circumstance, the indicator function that will equal 1 if the event occurs and zero otherwise can be used. Therefore, the expected value of a certain random event equals to the probability. This relationship can be used to translate properties of expected values into properties of probabilities.

Assuming indicator function of truck being present (1) and not being present (0) at a certain timeslot, a corresponding probability value is assigned to each situation. As it is shown in Table 6.1, assuming a certain truck scheduled to load in a certain timeslot, the conversion of presence probability value which is the result of predictive model into the expected value of being present which is the generalization of a random variable can be done by simply multiplying both values.

Table 6.1 Example of Expected Value

Time slot	Probability	Presence Indicator	Expected Value
9.00 AM	30%	1	0.3
9.30 AM	60%	1	0.6
10.00 AM	80%	1	0.8
10.30 AM	100%	1	1

For this instance, the truck is initially scheduled at 9.00 AM slot to load its predetermined product. However, according to the predictive model, there expected value of the truck to be able to follow its initial schedule is only 0.3, thus it is likely that the truck would require a reassignment to a further timeslot. In similar fashion, the predictive model could generate a presence probability of that certain truck in other further timeslots. Based on the expected values of being present in other set of timeslots, the adjusted schedule of that certain truck could be optimized.

6.2.Rolling Horizon

As loading process is a consistent event that is highly dependent on the actual arrival of truck, then application of rolling window method is suitable [83]–[86], thus will be applied in the proposed solution. The schedule will be optimized for a fix period of time. As it is illustrated by Figure 6.1, for every iteration, the time horizon moves forward by one scheduling period, therefore the start period i and end period W_i will be updated accordingly with respect to the continuous update of the real-time ETA and all other variables with respect to the truck and loading facility. In this case, the iteration will refer to rescheduling process that constantly occurs along the horizon of operational day.

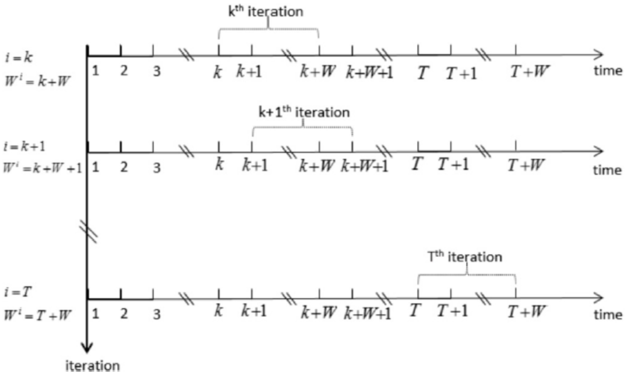


Figure 6.1 Rolling Horizon Approach [83]

6.3.Problem Description

In regard to the optimization model, several problems are likely to emerge during the described process in Section 1.2. Firstly, the scheduling concept is based on the notion that the trucks would have arrived in the facility according to the initial schedule based on their preferred/requested time. However, in practice, the trucks arrive stochastically due to inevitable causes associated with transport network disruption and would enforce rescheduling process. Secondly, the parking area is supposed to accommodate all incoming trucks in the planning horizon, but in reality, the parking space is limited, and the excess trucks would load the side of access road, thus they would cause detrimental impact to the transport network in a form of congestion and the livelihood around the facility as the queued truck would contribute to high level of emission gas. Thirdly, this assignment of loading bay is not interchangeable because the trucks are ordered to load predetermined product that is available at a certain loading bay. Consequently, problem also potentially occurs when the assigned loading bay is still being occupied by other trucks, then the truck must wait until the that certain loading bay comes available.

6.4. Assumption

According to [11] , several variables are relevant in modeling the rescheduling system as it is shown in Table 6.2. However, only selected variables labeled in green color are being considered in this research , as following:

Table 6.2 Description of Variables in Loading Facility

Location	Variables
General	Time horizon
	Slot
	Site opening hour
	Site over time
	CO2 emission factor of idle trucks
	CO2 emission cost
	Rescheduling cost
	Space rent cost
	Site operation cost
Parking Area	Waiting time in parking area for a certain truck
	Average number of queued trucks in parking area at certain time period
	Discrete arrival time of scheduled trucks
Facility Area	Waiting time in facility area for a certain truck
	Average number of queued trucks in a certain loading bay
	Total loading bays available
	Total identical loading bays with respect to the product type
	Type of product
	Service time
	Discrete departure time of scheduled trucks
	Average service rate of certain loading bay

Some assumptions are made to acknowledge the data limitation and to ensure this research be positioned within the stated scope, as following:

- The rescheduling cost is constituted of the difference between the adjusted schedule and initial schedule multiplied by generalized cost which is assumed 1 monetary value/slot. All the other costs, namely the CO2 emission cost, space rent cost, site operation cost, etc. are omitted.
- The space in facility area is adequate to accommodate all scheduled trucks within the planning horizon.
- The serviceability of loading bays are constantly available over the planning horizon. In other words, disruption such as machine breakdown or maintenance are not considered.
- The loading/service time is constant for all the timeslots in the planning horizon which is assumed to be sufficient within the timeslot
- Consequently, the departure time of truck could also be certainly calculated, thus the departure time from loading facility is deterministic.
- Rescheduling could be triggered only for truck that is classified as Class 0 (early arrival and Class 2 (late arrival) in which some rules in line with the Feature Creation (Section 5.4.2) are defined.

6.5.Mathematical Model

This research aims to develop model to optimize the rescheduling process that incorporates iterative rescheduling process and also takes into account the uncertain arrival time of trucks at the loading facility. The objective is to minimize the expected cost of rescheduling that would lead to more efficient and robust operation. The process is converted into a mathematical model in which the schedule assignment of the truck is subject to expected value of the truck's presence in the loading facility. Following the concept explained in Section 6.1, non-deterministic setting can be transformed into its equivalent deterministic mode, thus, solved straightforward MIP method. Common approach of SAP (Slot Assignment Problem) is adapted to model the rescheduling problem. This model is then referred as SRP (Slot Rescheduling Problem). This chosen model is inspired from literatures in similar problem [87] with appropriate modifications to fit this research. The following section provides the definition of sets, parameters, decision variables, objective and constraints that are required to solve the SRP problem.

6.5.1.Set

It is important to emphasize that the context being taken in building this optimization model is the perspective of schedule manager of chemical loading facility, therefore the model focuses on the interest and relevance of that particular actor. Set of variables are introduced in Table 6.3.

Table 6.3 Set of Variables.

Notation	Definition
T	Set of time slots within time horizon
J	Set of scheduled trucks within time horizon
L	Set of loading bays at the chemical facility
P	Set of product type
$\boldsymbol{J}$	Set of incoming trucks in a operation day

6.5.2.Parameter

The parameters represent the characteristic and configuration of loading facility that are being considered in the optimization model. The parameters are in Table 6.4.

Table 6.4 Set of Parameters.

Notation	Definition
$\boldsymbol{Y_j}$	Minimum time difference between new and initial schedule of truck j
$\boldsymbol{g_{jl}}$	Generalized cost in assigning truck j to loading bay l
$\boldsymbol{s_{jt}}$	Presence indicator of trucks, the value is 1 if the truck j has non-zero presence chance in time slot t, 0 otherwise
$\boldsymbol{r_{lj}}$	Assignment of certain product type, the value is 1 if certain product type meant for truck j can be loaded at loading bay l, 0 otherwise
$\boldsymbol{c_{jl}^t}$	Current schedule of trucks, the value is 1 if truck j is scheduled to loading bay l at time slot t, 0 otherwise

6.5.3.Decision Variables

The decision variable of this mathematical model is defined as following:

$$\boldsymbol{x_{jl}^t} = \begin{cases} 1, & if\ truck\ j\ is\ reassigned\ to\ loading\ by\ l\ in\ time\ slot\ t \\ 0, & otherwise \end{cases}$$

6.5.4.Objective Function

The objective of this model is to minimize the total expected rescheduling cost of the operation at loading facility. Technically, the expected cost is the accumulation of rescheduling cost of all trucks in all loading slots within the planning horizon. As it is explained, the expected cost equals to the difference between adjusted and initial scheduled multiplied by generalized cost which is 1 monetary value/slot. This objective function can be formulated as:

$$\textbf{Objective Function}: \text{Min} \sum_{t} \sum_{l} \sum_{j} \left(x_{jl}^{t}\, t \; - c_{jl}^{t}\, t\right)^{2} \;\; g_{jl} \tag{1}$$

There are several scheduling policies in which the goal might be differed, the form of objective function depends on the requirement and preference of the output. As deviation between initial and adjusted slot which could be either in negative (in case if truck being moved to earlier slots) or positive value (in case if truck being moved to later slots), therefore a quadratic cost function is chosen because it perfectly fits the context, instead of linear form that would prefer rescheduling to earlier slots than adhering initial schedule. This model is now categorized as MIQP (Mixed Integer Quadratic Programming) problem.

Moreover, it would ensure eliminating extreme value while maintaining overall deviation for all rescheduled trucks. To put it into perspective, given a set of 10 trucks getting delayed, would prefer to have 10 delayed trucks (Truck 1st – Truck 10th) to be reassigned between 2-4 later slots, rather than assigning 9 delayed trucks (Truck 1st – Truck 9th) to immediate next slots (1 slot) and 1 last delayed truck (Truck 10th) to be moved by 10 slots.

6.5.5.Integrating Expected Value Constraint

In the deterministic sense, the presence of particular truck in a certain time slot is assumed to be fully known and the delay probability is not taken into account. To capture the stochastic nature of arrival time, the model needs to be adapted. As it is proposed by [22], the idea of “presence probability” is incorporated into the deterministic SRP model. This modified model is further referred as P-SRP. Instead of using the binary constraint to define the truck presence, the probability of truck being present at a certain timeslot are used as replacement. This notion could be applied by incorporating probabilistic value to a constraint (6).

The changes represent the idea of allowing multiple trucks to be scheduled in a certain time slot as long as the overlapping probability does not violate the permissible threshold value. The r is a defined parameter that denotes the maximum permissible overlap probability, while the p_{jt} is the presence probability of truck j at time slot t. The p_{max} is defined as the maximum allowable presence probability of another truck to be simultaneously assigned at the same loading bay l and same time t with respect to the defined value of r. To satisfy the relation between both parameters, the equality formula can be written as:

$$r = p_{jt}\, p_{max} \tag{2}$$

Overlap probability at loading bay l at the time slot t is defined as the probability of two or more trucks j to be simultaneously assigned to a loading bay l at time slot r. In order to ensure the compatibility with the basic model a scaling function is needed. The scaling function is formulated in such way that fulfill the condition of the total presence probabilities p_{jt} is equal to 1 if the result of total presence probabilities of the corresponding trucks equal to the defined maximum permissible overlap probability r. Hence, this could be formulated as:

$$f\left(p_{jt}, r\right) p_{jt} + \; f\left(p_{jt}, r\right) p_{max} = 1 \tag{3}$$

By substitution, the equation (3) can be rewritten, resulting on:

$$p_{jt\;scaled} = \; f\left(p_{jt}, r\right) p_{jt} = \frac{p_{jt}^{2}}{r + p_{jt}^{2}} \tag{4}$$

Given the description on how to derive the expected value of the probabilistic variable and the stated objective function (1), set of constraints are defined for the MIQP, as following:

Constraint 1: Ensure that all trucks are assigned exactly once at a certain time horizon.

$$\sum_{l \in L} x_{jl}^{t} \, s_{jt} = s_{jt} \qquad ,\forall\, j \in J\,, t \in T \tag{5}$$

Constraint 2: Allow simultaneous assignment of trucks in a certain slot based on the presence probability .

$$\sum_{j \in J} x_{jl}^{t} \, f(p_{jt}, r) \, p_{jt} \; \le 1 \qquad ,\forall\, l \in L\,, t \in T \tag{6}$$

The presence probability ranges from 0 to 1 in which the presence value of 1 means that it is absolutely certain that truck would be present. Hence, the summation is limited to be smaller or equal to 1 because it would serve as threshold indicating if there is a truck having presence probability of 1, then it does not make sense to simultaneously assign another truck to that slot regardless its corresponding probability.

Constraint 3: Ensure maximum number of trucks that can be simultaneously assigned to measure the risk

$$\sum_{j \in J} x_{jl}^{t} \; \le 3 \qquad ,\forall\, t \in T, l \in L \tag{7}$$

Constraint 4: Enforce different new slot for delayed trucks while allowing same slot for on-time truck

$$x_{jl}^{t} \; + c_{jl}^{t} \; \le 2 \qquad ,\forall\, j \in J, t \in T,, l \in L \tag{8}$$

Constraint 6: Ensure that trucks are loaded with its corresponding product type.

$$x_{jl}^{t} \;\; \le \; r_{lj} \qquad ,\forall\;\, j \in \; J\,, t \in T, l \in L \tag{9}$$

Constraint 7: Satisfy the minimum time difference between new and initial slot.

$$x_{jl}^{t} \, t - c_{jl}^{t} \, t \ge Y_j \qquad ,\forall\;\, j \in \; J\,, t \in T, l \in L \tag{10}$$

Constraint 8: Domain of decision variable.

$$x_{jl}^{t} \;\; \in \{0,1\} \qquad ,\forall\;\, j \in \; J\,, t \in T, l \in L \tag{11}$$

6.5.6.Example of Implementation

To provide the rationale behind this approach, an example on how the stochastic constraint would allow incorporating delay probability into the schedule optimization model is exhibited. As it is described, the basic deterministic is simply transformed into the stochastic model by incorporating flight presence probability as replacement constraint to the binary constraint of truck arrival time. The basic concept is to ensure that the overlap probability that indicates conflicts between trucks scheduled to arrive at the timeslot does not exceed the specified value of maximum permissible threshold (r).

To showcase the nature of difference between the deterministic and stochastic with respect to the arrival time, a simple example is provided. Let assume that there are 5 trucks (j = 1, … ,5) that are inbound to the facility in the planning horizon. Due to the transport network problem, the real-time information on ETA indicates that the trucks are getting delayed in which the degree of deviation of predicted arrival time with respect to the initial schedule is represented by the corresponding variable of presence probability, p_{jt}. The data of discretized presence indicator provided by the ETA information of all trucks assigned to certain loading bay l in a certain time slot t. is shown in Table 6.5.

Table 6.5 Example of Binary Presence

	Truck 1	Truck 2	Truck 3	Truck 4	Truck 5
Presence indicator $\boldsymbol{s_{jt}}$	1	0	0	1	1

In the deterministic model, the Constraint 2 only allows assignment of single truck (either Truck 1,4, or 5) to the loading bay l in time slot t. In mathematical formula, the deterministic constraint would result on:

$$1\, x_{1l}^{t} + 0\, x_{2l}^{t} + 0\, x_{3l}^{t} + 1\, x_{4l}^{t} + 1\, x_{5l}^{t} \leq 1$$

On the other hand, in case where the real time ETA information is considered stochastic, then on top of presence indicator, the presence probability as the output of predictive model could also be assigned. The example of data is shown in Table 6.6.

Table 6.6 Example of Probabilistic Presence

	Truck 1	Truck 2	Truck 3	Truck 4	Truck 5
Expected value $(\boldsymbol{p_{jt}}, \boldsymbol{r})$	0.85	0.45	0.2	0.7	0.55
Presence indicator $(\boldsymbol{i})$	1	1	1	1	1
Presence probability $(\boldsymbol{p})$	0.85	0.45	0.2	0.7	0.55

The presence probability is the result of simple multiplication between the presence indicator and its corresponding confident level, which can be mathematically written as:

$$Pjt, r = \ i\ x\ p$$

The stochastic model would allow integrating the presence probability according to the latest ETA information that is generated by the predictive model, instead of using the binary presence. The constraint is transformed into, as following:

$$0.85\, x_{1l}^{t} + 0.45\, x_{2l}^{t} + 0.2\, x_{3l}^{t} + 0.7\, x_{4l}^{t} + 0.55\, x_{5l}^{t} \leq 1$$

According to this constraint, it allows simultaneously assigning Truck 2 and Truck 5 to same loading bay j at time slot t, in which the estimated overlap probability of those incoming trucks is 0.2475 (0.45 x 0.55). However, by directly applying the given presence probability as its given Table 6.6 might lead to undesirable or bias result. This is shown by the fact that the despite having overlap probability of merely 0.17 (0.85 x 0.2) which is lower than the former combination, Truck 1 and Truck 3 cannot be assigned simultaneously according to the constraint. Hence, to solve this issue, scaling function is necessary to connect the relation and to maintain the consistency between trucks that simultaneously assigned to the same loading bay l in same time slot t and the maximum permissible overlap probability, r.

Scaling function $f(p_{jt}, r)$ is implemented to the presence probability given in Table 6.6. By assuming that the value of r is equal to 0.1, then the constraints would be redefined as following:

$$0.88\, x_{1l}^{t} + 0.67\, x_{2l}^{t} + 0.29\, x_{3l}^{t} + 0.83\, x_{4l}^{t} + 0.75\, x_{5l}^{t} \leq 1$$

With respect to this constraint, the only feasible combination of trucks to be simultaneously assigned to same loading bay l in same time slot t is Truck 2 and Truck 3. To check whether the overlap probability of this combination comply with the maximum permissible overlap probability (r=0.1), the presence probabilities of Truck 2 and 3 are multiplied (0.45 x 0.2 = 0.09). The result shows that this combination suffice both the presence constraint and the threshold value of overlap probability.

6.6.Chapter Overview

In this chapter, a probabilistic optimization model aims to increase the operational efficiency of loading facility has been developed.

The P-SRP model incorporates the expected value of presence probability constraint instead of the binary presence in assigning truck to a certain loading slot has been developed. This model acknowledges the uncertainty of arrival time by considering a presence probability value to each arrival information. Consequently, the advantage of this model is that it allows simultaneous assignment based on the accumulated presence probability of trucks in a certain timeslot.

This concept could be applied to mitigate the further disruption caused by trucks that cannot comply with the initial schedule due to early/late arrival. The implementation of this concept is done in the objective function which minimize the expected cost consisting of deviation between adjusted and initial schedule multiplied by generalized cost of rescheduling.

7. Verification

In the previous sections, both predictive and optimization model have been established as part of conceptual framework.

This chapter aims to verify whether the framework will perform in the expected behavior. In doing so, both predictive and optimization model will be tested using a dataset that is designed to simulate a situation where congestion is significant, and rescheduling is necessary. The predictive model will be used to generate presence probability, while the P-SRP is then applied to generate an adjusted schedule. The results will be evaluated to show how the P-SRP would provide an improvement in dealing with uncertain truck arrival, compared to the current system applied in the loading facility.

7.1.Configuration of Loading Facility

There are set of elements that represent the vital function/characteristic of loading facility. Specifically in this research, the standard configuration of loading facility is constituted of loading bays, and product types. The loading bay is designed to load a certain defined product, in which the arrangement is shown in Table 7.1. The product type P0, P1, P2, P3, can only be loaded in Loading Bay 0, while the product type P4, P5, P6, P7 can only be loaded at Loading Bay 1.

Table 7.1 Configuration of Loading Bay and Available Product Type

Loading bay	Available Product type
0	P0, P1, P2, P3
1	P4, P5, P6, P7

In reality, the number of loading slots in each loading bay and its duration are varied, nevertheless, this research assumed them as constant figures in which there are 24 loading slots with 30 minutes duration in 12 hours of operation in a day, so the result would be in more general manner. Normally, the earliest loading slot is at 6:00 and the latest loading slot is at 17.30. Currently, the loading facility applies appointment system that utilizes an internet-based service booking provider. The complete list of available timeslots for loading process is provided in Table 7.2. The index of the loading slot is described in Column 1, while the start time of each loading slot is described in Column 2 .

Table 7.2 List of Available Time Slot during Operation Hours

Timeslot	Start Time	Timeslot	Start Time	Timeslot	Start Time
0	6:00	8	10:00	16	14:00
1	6:30	9	10:30	17	14:30
2	7:00	10	11:00	18	15:00
3	7:30	11	11:30	19	15:30
4	8:00	12	12:00	20	16:00
5	8:30	13	12:30	21	16:30
6	9:00	14	13:00	22	17:00
7	9:30	15	13:30	23	17:30

7.2.Initial Schedule

For the evaluation purpose, a sample set consisting of 42 trucks labeled from 0 to 41 is generated (Table 7.3). For each operational day, a set of initial schedules for loading process is released in which a set of trucks are assigned to a certain timeslot and loading bay according to their predetermined product type. The set of time slots have an interval of 30 minutes. Moreover, it is assumed that service time is deterministic, and would always be sufficient to finish all the required loading process. The start time indicates how late the truck should arrive in order to be eligible into the assigned initial loading slot. The administration process that includes document checking and technical preparation for the loading process is included within the timeslot.

To provide more comprehensive view, the initial schedule is plotted and shown in Figure 7.1. It can be derived that the initial schedule ensure that there will be more than 1 truck being assigned to a particular loading bay at a particular timeslot, so there will be no overlapped loading slot. Almost all the timeslots of both loading bays are fully occupied that shows that the loading operation is supposedly non-stop, except at timeslot 12 and timeslot 16 in which only loading bay 0 and loading bay 1 is occupied, respectively. The reason is because those slots are reserved as the buffers to accommodate the delayed trucks from previous sessions which is the only and actual mitigation plan applied in current practice.

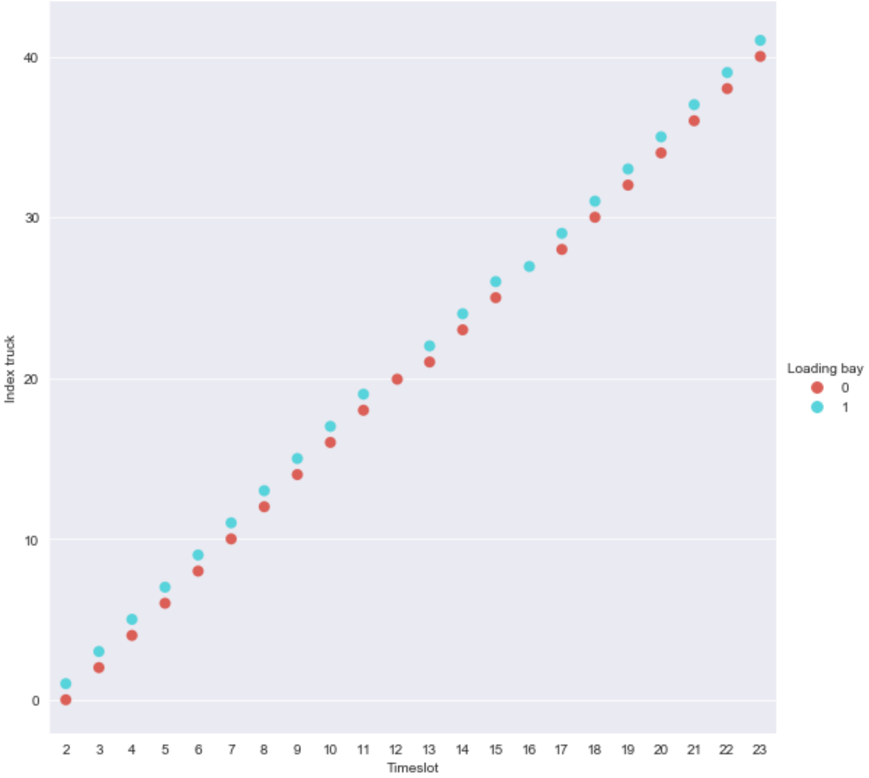


Figure 7.1 Assignment Matrix for Initial Schedule

Table 7.3 Data of Initial Schedule

Index Truck	Product Type	Start Time	Timeslot	Loading Bay
0	P5	7:00:00 AM	2	0
1	P1	7:00:00 AM	2	1
2	P6	7:30:00 AM	3	0
3	P0	7:30:00 AM	3	1
4	P4	8:00:00 AM	4	0
5	P0	8:00:00 AM	4	1
6	P5	8:30:00 AM	5	0
7	P3	8:30:00 AM	5	1
8	P7	9:00:00 AM	6	0
9	P3	9:00:00 AM	6	1
10	P4	9:30:00 AM	7	0
11	P0	9:30:00 AM	7	1
12	P6	10:00:00 AM	8	0
13	P3	10:00:00 AM	8	1
14	P5	10:30:00 AM	9	0
15	P3	10:30:00 AM	9	1
16	P5	11:00:00 AM	10	0
17	P0	11:00:00 AM	10	1
18	P4	11:30:00 AM	11	0
19	P2	11:30:00 AM	11	1
20	P6	12:00:00 PM	12	0
21	P5	12:30:00 PM	13	0
22	P2	12:30:00 PM	13	1
23	P5	1:00:00 PM	14	0
24	P1	1:00:00 PM	14	1
25	P4	1:30:00 PM	15	0
26	P3	1:30:00 PM	15	1
27	P3	2:00:00 PM	16	1
28	P6	2:30:00 PM	17	0
29	P0	2:30:00 PM	17	1
30	P5	3:00:00 PM	18	0
31	P2	3:00:00 PM	18	1
32	P4	3:30:00 PM	19	0
33	P3	3:30:00 PM	19	1
34	P5	4:00:00 PM	20	0
35	P3	4:00:00 PM	20	1
36	P7	4:30:00 PM	21	0
37	P3	4:30:00 PM	21	1
38	P7	5:00:00 PM	22	0
39	P2	5:00:00 PM	22	1
40	P6	5:30:00 PM	23	0
41	P0	5:30:00 PM	23	1

7.3.Real Time ETA

Inevitably, most of the trucks would be experiencing disruption in the transport network that will likely result on inability to comply with the initial schedule. The real-time information (Table 7.4) indicates an ETA and its deviation from the start time of the initial schedule where the positive value refers to later than initial point, whereas the negative value refers earlier than initial point. Consequently, reactive action, namely rescheduling is required to mitigate the risk. Otherwise, most of the trucks would potentially experience conflicted timeslot allocation upon their arrival that would inflict larger cost. In other words, the main benefit of this real-time information is that it allows rescheduling to be processed and communicated hours in advanced.

Table 7.4 Real Time ETA Data

Index Truck	ETA	Deviation	Index Truck	ETA	Deviation
0	7:12:00 AM	12.20	21	1:38:00 PM	68.40
1	7:27:00 AM	27.60	22	12:50:00 PM	20.30
2	8:06:00 AM	36.44	23	1:12:00 PM	12.40
3	7:24:30 AM	-5.50	24	1:08:00 PM	8.40
4	8:13:00 AM	13.60	25	2:02:00 PM	32.10
5	8:01:00 AM	1.20	26	1:44:00 PM	14.40
6	8:39:00 AM	9.10	27	1:58:00 PM	-2.00
7	8:38:00 AM	8.20	28	3:10:00 PM	40.70
8	9:02:00 AM	2.30	29	2:26:48 PM	-3.20
9	9:02:00 AM	2.10	30	3:12:00 PM	12.50
10	9:46:00 AM	16.30	31	3:22:00 PM	22.40
11	9:47:00 AM	17.40	32	3:35:00 PM	5.00
12	10:41:00 AM	41.20	33	4:04:00 PM	34.20
13	9:56:12 AM	-3.80	34	3:50:18 PM	-9.70
14	10:19:24 AM	-10.60	35	4:33:00 PM	33.20
15	10:17:36 AM	-12.40	36	5:00:00 PM	30.40
16	10:31:36 AM	-28.40	37	5:26:00 PM	56.50
17	11:02:00 AM	2.30	38	5:14:00 PM	14.20
18	11:24:24 AM	-5.60	39	5:06:00 PM	6.20
19	11:34:00 AM	4.20	40	5:24:30 PM	-5.50
20	11:57:42 AM	-2.30	41	4:57:48 PM	-32.20

7.4.Presence Probability

According to the proposed conceptual framework (Section 3.2.2), the next step is that the real-time ETA will subsequently be inputted to the predictive model in which its ML algorithm has been trained using the historical data. The predictive model is run to predict the probability value of whether a certain truck would be early, on-time, or late as it is explained in Section 5.8. The prediction result given the sample data in this use-case implementation is shown in Table 7.5. From the real-time ETA information, the class of each truck could be checked with respect to its initial schedule. The presence probability is then can be derived.

Table 7.5 Presence Probability of Truck

Index Truck	Status	Presence Probability
0	Late	0.14
1	Late	0.15
2	Late	0.214
3	On-Time	0.358
4	Late	0.127
5	Late	0.162
6	Late	0.366
7	Late	0.215
8	Late	0.227
9	Late	0.176
10	Late	0.181
11	Late	0.156
12	Late	0.247
13	On-Time	0.362
14	On-Time	0.36
15	On-Time	0.342
16	On-Time	0.352
17	Late	0.3
18	On-Time	0.351
19	Late	0.228
20	On-Time	0.335
21	Late	0.288
22	Late	0.384
23	Late	0.366
24	Late	0.349
25	Late	0.535
26	Late	0.243
27	On-Time	0.314
28	Late	0.431
29	On-Time	0.284
30	Late	0.401
31	Late	0.388
32	Late	0.493
33	Late	0.404
34	On-Time	0.273
35	Late	0.434
36	Late	0.627
37	Late	0.741
38	Late	0.741
39	Late	0.532
40	Late	0.229
41	Early	0.109

7.5.Rescheduling Result

By comparing the ETA real-time ETA information, the decision-makers could understand which trucks would be able to comply with the initial schedule. As the consequence, a rescheduling process is required as the mitigation plan. The result of the process is the adjusted schedule. Since the real-time ETA information is continuous, meaning that there will be constant updates, thus it is important to note that the rescheduling process would be iterative along the rolling horizon.

Two different approaches in dealing with the rescheduling problem will be explored given the previously stated sample dataset. The probabilistic optimization model proposed in this research will be verified and evaluated. The result would show the added value of using a real-time ETA information in integration with a predictive model. To provide a perspective on how well it would enhance the operation, a current state analysis will also be implemented as the benchmark that represent current situation.

As it is explained, this research only assumes the arrival time as the stochastic element, while the rest of the variables are deterministic. Both optimization models is written using Python programming language and the MILP optimization problem is solved using Gurobi module.

7.5.1.Current Practice

In this research, the current practice in loading facility is characterized by the inexistence of ETA information or any other significant communication channel between the incoming truck and decision-maker in loading facility. The ramification is the inability to do a rescheduling process ahead in time.

The current rules that regulates the loading process are, as following:

a. A set of slots during operation hours are available in which trucks could choose their most preferred time. The starts of booking window starts from 4 working days and ends at 1 working day before the planned loading date.
b. Every day, the decision-makers at the loading facility will possess the full day schedule that details the initial schedule containing exact time slot and assigned loading bay.
c. There are two conditions where trucks cannot comply with the appointment:
   - Early arrival in which the options are either waiting until the booked slot or getting rescheduled to earlier slots if there are available slots that suit the requirement.
   - Late arrival in which the options are either getting rescheduled to later slots if there are available slots compatible to the trucks' requirement.

In hypothetical sense, assuming an actual observation has been done, a plot that maps where and when a certain truck being assigned according to their arrival time can be illustrated in Figure 7.2. Accordingly, several overlapped assignments are occurred, indicated by orange circles. This could provide an insight on the magnitude of disruption that possibly occur in the operation. Consequently, according to current practice, the only solution is to manually move the trucks to next possible slots upon its arrival. However, in order to do that, the truck must wait indefinitely, and if there is no slots available at that day, then the truck will be moved to the next day schedule. Therefore, this current practice is not sufficient.

7.5.2.Baseline Model

Unlike the current situation where the rescheduling is not an option, one alternative is to implement a more advanced communication technology, such as the real-time ETA information. This addition enables rescheduling process according to the ETA value by applying the optimization model.

A baseline model that will be used as the benchmark to analyze the added values of the P-SRP is introduced. The baseline model refers to the rescheduling system of loading facility that considers the real-time ETA information as deterministic variable. In other words, the initial schedule will be according to fully known and certain arrival information. Consequently, to do so, the P-SRP model described in Section 6.5.5 is slightly modified by removing the expected value of probabilistic variable in Constraint (6). The rest of it is still exactly the same, including the objective function and other set of constraints. Further on, this model will be referred as Baseline model.

Assuming the same sample as previous ones, the result of rescheduling is illustrated in Figure 7.3. Based on the result, this method could be sufficient in situation where the nearest possible slots are still empty and no conflict possibility with other truck that is initially assigned to those certain slots. Nonetheless, in reality where almost all the remaining slots are already being reserved for other trucks, this approach is not that efficient. In addition the drawback of manual scheduling would be heavily experienced in late operational, assuming the operational of loading facility ends at 18.00, then the last possible loading slot is the Timeslot 23. Trucks that are scheduled to afternoon/evening loading slot (shown in red circle) would suffer from the delay propagation effect and the fact that there is no other later slots available in that day, thus the only solution for them is to return to home and come again another day, which would require additional transportation cost. Hence, a more advanced rescheduling approach is appropriate.

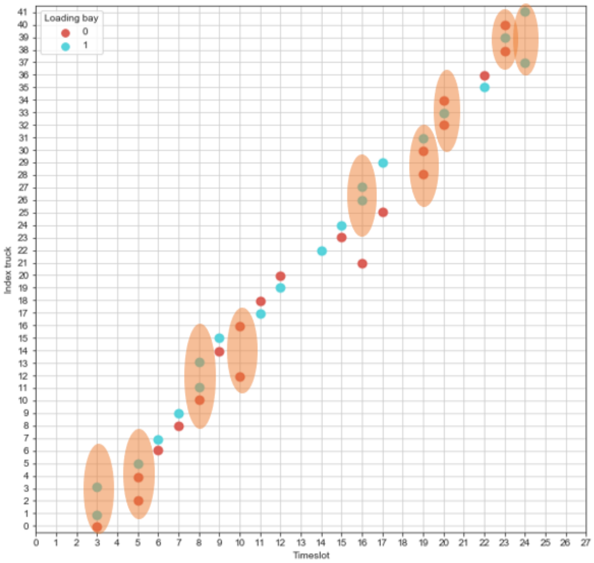


Figure 7.2 Hypothetical Conflicts in Current Situation

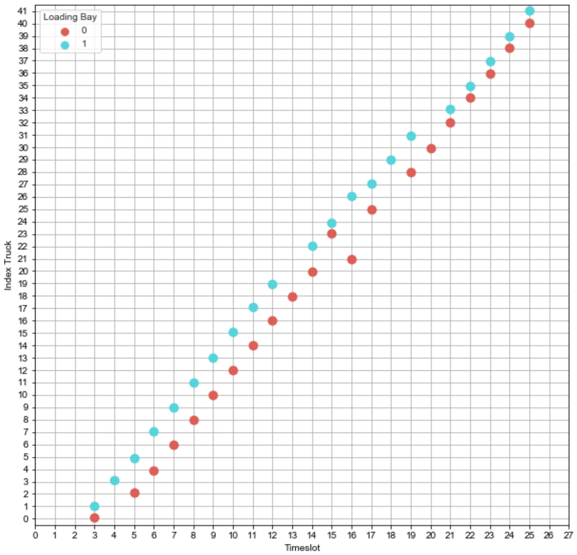


Figure 7.3 Rescheduling Result of Baseline Model

7.5.3. P-SRP Model

This model is part of the conceptual framework proposed in this research . Fundamentally, the probabilistic optimization model incorporates presence probability of a certain truck in a certain timeslot as the main indicator to determine whether it is possible to simultaneously assign 2 or more trucks in a same time slot.

Set of information relevant to the rescheduling process could be obtained as it is provided in Table 7.6. Given the comparison between predicted arrival and its initial schedule, arrival class, namely early (Class 0), on-time (Class 1), and late (Class 2) of which a certain truck likely belongs to could be derived. Based on that classification, a status whether a particular truck should be rescheduled could be generated, whereby the ‘Reschedule’ label means that truck would expectedly require rescheduling while the ‘Initial label means that truck is still expected to attend its initial slot schedule

Moreover, from the ETA information, expected time gap between initial and ETA time could be derived. The values imply that given the current ETA, the trucks must be rescheduled in such a way that the minimum gap between initial and adjusted schedule satisfy the time difference. The values could be either negative or positive, which means that if negative the truck could be reassigned to earlier slots, while if positive the truck could only be reassigned to later slots.

Set of presence probabilities for each truck with respect to the corresponding ETA is appropriate to represent that truck arrival is still a stochastic element. The values are derived from the predictive model. It is a quantitative measurement that represents the probability of a certain truck being truly in its respective class based on trend/pattern of its historical data. In regard to the optimization, the value also indicates how likely the overlapping issue with the currently scheduled truck at the same loading slot will occur. As it is explained in Section 6.5.5, to maintain the correct behavior between presence probability and overlapping situation between 2 or more trucks, a scaled function is applied, thus the presence probability will be scaled with respect to the permissible overlap parameter, r, with value of 0.15. Given the real-time ETA information and tis corresponding presence probability generated by the predictive model, an adjusted schedule as the result of the probabilistic optimization model is shown in Figure 7.4.

Table 7.6 Input Data to the P-SRP

Index truck	Arrival Class	Status	Time difference	$p_{jt\ scaled}$
0	2	Reschedule	0:30	0.089
1	2	Reschedule	0:30	0.101
2	2	Reschedule	1:00	0.187
3	1	Initial Schedule	0:00	0.591
4	2	Reschedule	0:30	0.393
5	2	Reschedule	0:30	0.439
6	2	Reschedule	0:30	0.402
7	2	Reschedule	0:30	0.411
8	2	Reschedule	0:30	0.423
9	2	Reschedule	0:30	0.134
10	2	Reschedule	0:30	0.381
11	2	Reschedule	0:30	0.387
12	2	Reschedule	1:00	0.409
13	1	Initial Schedule	0:00	0.395
14	1	Initial Schedule	0:00	0.260
15	1	Initial Schedule	0:00	0.133
16	1	Initial Schedule	0:00	0.370
17	2	Reschedule	0:30	0.310
18	1	Initial Schedule	0:00	0.382
19	2	Reschedule	0:30	0.475
20	1	Initial Schedule	0:00	0.360
21	2	Reschedule	1:30	0.293
22	2	Reschedule	0:30	0.425
23	2	Reschedule	0:30	0.401
24	2	Reschedule	0:30	0.378
25	2	Reschedule	1:00	0.588
26	2	Reschedule	0:30	0.344
27	1	Initial Schedule	0:00	0.330
28	2	Reschedule	1:00	0.481
29	1	Initial Schedule	0:00	0.287
30	2	Reschedule	0:30	0.297
31	2	Reschedule	0:30	0.430
32	2	Reschedule	0:30	0.548
33	2	Reschedule	0:30	0.286
34	1	Initial Schedule	0:00	0.271
35	2	Reschedule	1:00	0.485
36	2	Reschedule	0:30	0.663
37	2	Reschedule	1:30	0.733
38	2	Reschedule	0:30	0.132
39	2	Reschedule	0:30	0.586
40	1	Initial Schedule	0:00	0.208
41	2	Reschedule	-0:30	0.133

Based on the rescheduling result, several main findings can be derived:

- All trucks are assigned to loading bays that can serve the predetermined product it supposedly carries out, therefore there is no violation with respect to predetermined product type.
- All the delayed trucks are appropriately reassigned to a new loading slot before the normal operational hours, except Truck 37 that must be reassigned to an extra loading slot
- Compared to the manual rescheduling in which some trucks cannot be served even though arriving before the end of operational hours, the probabilistic method appropriately solves this problem. It means that, in general, the probabilistic could provide more efficient operation, indicated by total expected cost of only 17.25, while hypothetically, the current rescheduling method results on 29.75.
- Lastly, allowing simultaneous assignment to solve overlapping problem is one major advantage of P-SRP model, which is clearly illustrated by the loading slot 3,5,8,10, etc., In this situation, possible conflicted trucks keep their initial schedule because their overlap probability is still lower than the maximum permissible value.

In regard to managerial insights, there are several main benefits of applying the P-SRP model. The additional incorporation of real-time ETA information and predictive model provides more valuable information to make a better planning of the operation. By grounding the rescheduling decision on the presence probability value, instead of binary value a more measured decision could be taken. For example, on one hand, if the value is low then simultaneously assigning another truck could mitigate the risk of having idled loading slots. On the other hand, if the value is high, then the slots could be confidently allocated to the initial truck. Hence, the implication of this strategy is not only the potential of reducing gaps between initial and adjusted schedule, also preventing underutilization of loading slots early notification on the possible number of trucks that would likely miss their initial schedule could be used as the base to determine manpower allocation and additional loading slots outside the normal operational hours.

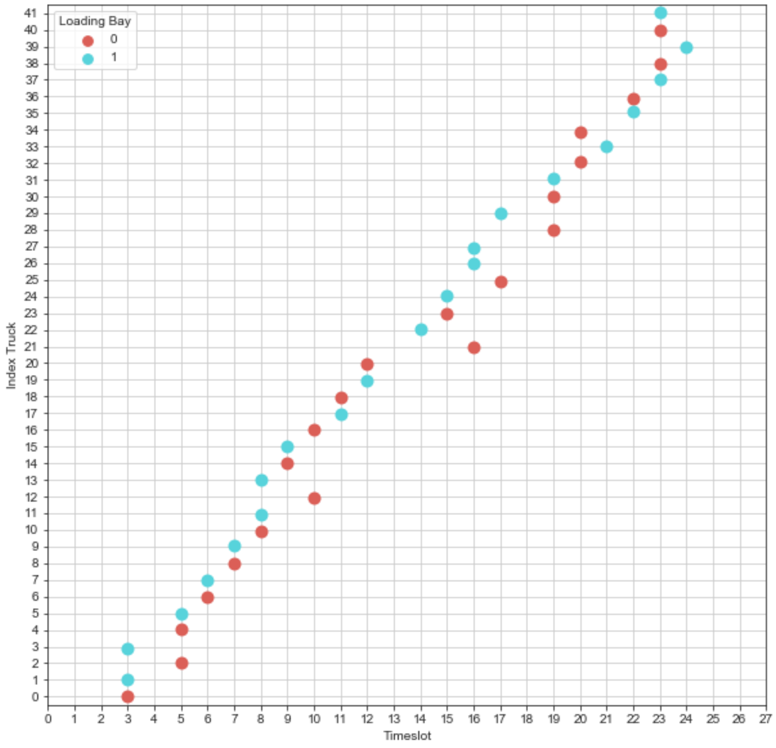


Figure 7.4 Rescheduling Result of P-SRP

7.6.Dynamic Simulation

In this section, practical implementation of the proposed rescheduling system will be simulated in a dynamic setting. The main goal is to exhibit how this system would realistically improve the operational efficiency of loading facility. The advantage of the P-SRP model which utilizes additional feature of predictive model embedded in that could provide the presence probability in comparison with the deterministic approach will also be explored . Due to limitations to conduct a real empirical study, some assumptions and synthetic data will be used.

7.6.1. Temporal Setting

As it is previously explained, the system relies on the continuous stream of real-time ETA that allows rescheduling to be done ahead in time. The system is designed to work on a rolling horizon period, in which this system would be iteratively executed in every certain period of time along the daily operational hours. Therefore, the loading schedule will be accordingly adjusted as a mitigation effort in order to accommodate truck that will not be able to comply with its initial schedule. Due to the time limitation and computational power, some simplifications are required to achieve the goal of this study, as following:

- Assuming that there are 3 working shifts in a day, this experimental study focuses only on the morning shift that spans from 7.00 AM – 11.30 AM.
- Extreme perspective that the operation has 100% slot utilization rate is assumed. It means that if there is a truck that miss its reserve slot, then the truck has to wait indefinitely or will be moved to next day schedule.
- Experiments is conducted discretely. The rescheduling system is set to be executed for every 2 hours. Thus there will be two iterations of adjusted schedules according to the updates of real-time ETA that will be analyzed.
- A synthetic data of actual arrival time is generated that will be served as a benchmark to show the potential of the proposed rescheduling system.

To test the proposed system, a normal operational condition of loading facility and the corresponding initial schedule is assumed and shown in Table 7.7.

Table 7.7 Initial Schedule for Dynamic Simulation

Initial Schedule					
First period (7.00 AM – 9.00 AM)			**Second period (9.30 AM – 11.30 AM)**		
Index Truck	Timeslot	Loading bay	Index Truck	Timeslot	Loading bay
0	2	0	**10**	7	0
1	2	1	**11**	7	1
2	3	0	**12**	8	0
3	3	1	**13**	8	1
4	4	0	**14**	9	0
5	4	1	**15**	9	1
6	5	0	**16**	10	0
7	5	1	**17**	10	1
8	6	0	**18**	11	0
9	6	1	**19**	11	1

7.6.2.First iteration

According to their real-time ETA, inevitably, most of the trucks are not able to adhere the initial schedule. To mitigate this problem, rescheduling is done to adjust the schedule (Table 7.8). The baseline and P-SRP are compared. The baseline model considers ETA as perfect information, whereas the P-SRP acknowledges the uncertainty and to compensate it, a presence probability as the result of predictive model is assigned to the ETA.

By comparing the performance of these models, the main practical advantage of the implementing probabilistic approach in real situation (rather than in theoretical sense shown in Section 6.5.6) could be showcased. To achieve this goal, an Actual Time of Arrival (ATA) is synthetically and carefully generated. This data will be used to see the extent to which the proposed rescheduling system could cope with uncertainty and maintain the operational efficiency of loading facility.

Table 7.8 First Iteration of Rescheduling and Actual Arrival

ETA: First iteration						Actual
Baseline			P-SRP			
Index Truck	Timeslot	Loading Bay	Timeslot	Loading Bay	Presence probability	Timeslot
0	2	0	2	0	0.34	2
1	3	1	3	1	0.15	3
2	3	0	3	0	0.94	3
3	5	1	6	1	0.30	6
4	4	0	4	0	0.32	4
5	6	1	7	1	0.69	7
6	6	0	6	0	0.62	6
7	7	1	5	1	0.80	5
8	7	0	6	0	0.12	Unknown
9	8	1	6	1	0.25	Unknown
10	8	0	8	0	0.13	Unknown
11	9	1	8	1	0.73	Unknown
12	9	0	8	0	0.59	Unknown
13	10	1	9	1	0.64	Unknown
14	10	0	9	0	0.12	Unknown
15	11	1	10	1	0.21	Unknown
16	11	0	11	0	0.94	Unknown
17	12	1	10	1	0.44	Unknown
18	12	0	12	0	0.79	Unknown
19	13	1	13	1	0.87	Unknown

In the first iteration of the rescheduling that is conducted at the beginning of the operational hours (7.00 AM), 70% of the trucks will not be able to attend their initial scheduled according to the ETA information. Consequently, the rescheduling will be triggered to provide an adjusted schedule as a solution. An analysis on a simple case is done to highlight the benefit of the proposed system. For instance, based on initial schedule, Truck 3 will arrive late and miss its initial schedule of Timeslot 3. In order to mitigate the possible detrimental effect, rescheduling is triggered.

In the baseline, the Truck 3 is then reassigned to the Timeslot 5 which is initially reserved for Truck 7. This rescheduling comes at the expense of Truck 7 being moved further to Timeslot 7. At first glance, this adjustment makes sense and would possibly avoid congested operation, however, this might actually cause inefficiency in the form of potential idled slot because it assumes perfect information of ETA. It means that this approach disregards a factor of which truck is more likely to be present when adjusting the schedule.

The possible practical weakness of this method is exposed when the adjusted schedule is evaluated with respect to the actual arrival time of trucks (Table 7.8). Turned out, Truck 3 still could not attend the Timeslot 5 because it actually arrived at the beginning of Timeslot 6, while Truck 7 that has been reassigned to a further schedule arrive on time for its initial schedule of Timeslot 5. The ramification is that Timeslot 5 being idled with no truly beneficial impacts in any possible senses for the loading operation. So, with benefit of hindsight, it can be concluded that the adjusted schedule does not provide as much robustness and efficiency as it is projected.

The P-SRP acknowledges this issue by considering presence probability value on the ETA. According to the predictive model, Truck 7 will present with probability of 0.8 while truck 3 only present at 0.3. Therefore, instead of altering initial schedule of truck 7, it is more logical to reassign truck 3 to later slots, since the probability of being present of truck 3 is higher as the time progresses. This is exactly what happened when the P-SRP is utilized. Evaluation with respect to the actual arrival data proves that it is a better and smarter solution because it could reasonably reduce a certain loading slot being idled (Timeslot 5) without unnecessarily making too many changes in the schedule, Hence, considering presence probability provides adjusted schedule that could properly cope with the uncertainty of truck arrival time.

Moreover, based on actual arrival data, Truck 8 and Truck 9 have not arrived yet at the end of the first part of morning shift, this is the exact reason why the arrival status of them are stated as unknown in Table 7.8. As the ETA is dynamic and live, it can be interpreted that there might additional congestion or problem experienced by Truck 8 and 9 that deviates their ETA This is not a problem since the rescheduling process is conducted in iterative manner, thus the current schedule is deemed void and instead will be adjusted according to the ETA in the next iteration.

7.6.3.Second iteration

Table 7.9 Second Iteration of Rescheduling and Actual Arrival

ETA: Second iteration						Actual
Baseline			P-SRP			
Index Truck	Timeslot	Loading Bay	Timeslot	Loading Bay	Presence probability	Timeslot
0				Arrived		
1				Arrived		
2				Arrived		
3				Arrived		
4				Arrived		
5				Arrived		
6				Arrived		
7				Arrived		
8	7	0	7	0	0.65	7
9	7	1	7	1	0.11	9
10	8	0	8	0	0.12	10
11	8	1	7	1	0.75	7
12	9	0	8	0	0.79	8
13	9	1	8	1	0.69	8
14	10	0	9	0	0.67	9
15	11	1	11	1	0.94	11
16	11	0	11	0	0.64	11
17	10	1	10	1	0.34	10
18	13	0	13	0	0.27	Unknown
19	12	1	12	1	0.68	Unknown

At this time, it is understood that the Truck 0 – Truck 7 have all arrived and loaded their products. Truck 8 and Truck 9 are also supposed to arrive but have not. Thus, both trucks required a second rescheduling that will be included in the second iteration. As the system is working live and dynamic, Table 7.9 clearly indicates not only the changes in the ETA, also the presence probability as it is affected by the fluctuation of transport network's condition. To mitigate the negative impact of prospective disruptions, the second iterations of rescheduling, both in deterministic and probabilistic approach are executed.

The comparison result is shown in Table 7.9. There is not extremely different, for instance, Truck 14-Truck19 are rescheduled in exactly same manner. Some distinctive impact of taking into account presence probability is rescheduling is highlighted by the case of simultaneous assignments (highlighted yellow) that occurs in Timeslot 7 at Loading Bay 1 and Timeslot 8 at Loading Bay 0. This sort of features proved to be a competitive edge for the P-SRP over the baseline because this measurement would likely reduce the risk of having unutilized slot. Considering Timeslot 7 and Loading Bay 1 which is initially reserved by Truck 11 as example, probabilistic approach simultaneously reschedule Truck 9 whose presence probability is merely 0.11 and keep Truck 11 whose presence probability is 0.75. The implication is an existence of contingency plan to ensure that this loading slot will be utilized, highly likely by Truck 11 or less likely by Truck 9. On the contrary, the baseline approach does not employ this ability because it assumes perfect information of ETA and will treat all arrival with 100% uncertainty. This is why the Truck 9 is rescheduled to the Timeslot 7 Loading Bay 1, while Truck 11 is rescheduled to Timeslot 8 Loading Bay 1. As the consequence, there will be ripple-effect for the rest of the trucks that will likely to be shifted further.

In order to check the extent to which this rescheduling affect the fluidity of operation, a comparison with respect to the actual arrival data is done. Referring to Table 7.9, it is evident that P-SRP provides more operational efficiency by reducing probability of underutilization of loading slot. This can be observed at the case of Truck 11 and 12 whereby the P-SRP efficiently rescheduled both of them without any cost of other trucks being unnecessarily moved, whereas the baseline model will cause idled slots of Timeslot 7 at Loading Bay 1 and Timeslot 8 at loading Bay 0 because Truck 9 and 10 that are rescheduled to that slot experienced further delay and in reality, have arrived late than its current ETA. This suitability between the planned schedule and the actual arrival indicate that adjusted schedule generated by probabilistic method yields a higher degree of robustness against the real-time disturbance and uncertainty.

Furthermore, similar to the previous stage, the two last trucks, Truck 18 and 19 cannot be assessed in the second iteration because they actually have not arrived by the end of second phase, therefore they will be included in the next iteration. This exact iterative process will be executed along the rolling horizon of daily operation.

The result of analysis emphasizes the advantage of probabilistic approach and the importance of factoring presence probability in rescheduling process. The implementation of this proposed system could provide a more intelligent rescheduling process, in the sense that operational efficiency could be increased without inappropriately reducing the schedule robustness.

7.7.Chapter Overview

In this chapter, the conceptual framework has been completely applied and verified on a synthetic use-case sample. Expectedly the integration of predictive and optimization model provide better performance of loading facility in terms of rescheduling problem. In addition, this chapter also emphasizes on the importance of historical and real-time data to generate a presence probability for each arrival information.

The first step of this verification process is to synthetically generate the setting of the loading operation and its relevant data. The current practice of operation is established, in which there is a full initial schedule released in the beginning of operational hours that corresponds to the initial schedule of incoming trucks from all over Europe. Since deviation of arrival time is inevitable, most of the trucks are will not be able to adhere the initial schedule. As the result, a congested operation will likely occur and disrupt the entire supply chain process. Currently, the only mitigation plan is to manually find a free feasible slot in the later hours upon its arrival. It sounds sufficient, yet if the slots are fully booked then the problem becomes unsolvable, enforcing trucks to be fit into the next-day schedule which would inflict additional cost for both parties.

In contrast to the current practice that heavily lacks communication between incoming trucks and decision-makers in loading facility, a solution that incorporates ETA information is introduced. The main advantage is it allows execution of rescheduling process ahead of time before truck actually arrive at the loading facility. It is evident that this approach could mitigate the risk of uncertain truck arrival time to a certain extent. However, the baseline model assumes a perfect information of ETA, which is not correct because ETA yields a certain degree of uncertainty, therefore disregarding this aspect could lead to unreliable result.

As a way to acknowledge this concern, a probabilistic approach known as P-SRP model is introduced. The incorporation of predictive model allows simultaneous assignment based on the value of presence probability that directly improve the operational efficiency by reducing the occurrence of idled slots and unnecessary schedule adjustment. Based on the comparison result, the P-SRP model results on expected cost that is 42% lower than the P-SRP model.

Lastly, a dynamic experiment is conducted to clearly illustrate the real-world implementation and to observe the practical benefit of proposed solution. As it is explained, the rescheduling system is designed to work in rolling horizon. It means that the system will be iteratively following the continuous stream of real-time ETA data. Due to the limitation in this research, only 2 iterations with period of 2 hours are simulated. The results showcased the superiority of P-SRP compared to the baseline. On top of the higher operational efficiency, the analysis indicates that P-SRP provides robustness against uncertain arrival time, whereby the adjusted schedule could withstand and anticipate the fluctuation of actual arrival of incoming truck. Hence, it can be concluded that the main advantage of utilization of predictive model and P-SRP proposed in this research is a more intelligent rescheduling process, in which operational efficiency is improved without radically compromising schedule robustness. However, the results might be case-specific considering the limited data and experiments.

8. Numerical Experiment

In the previous sections, the predictive model and optimization model have been elaborated. The implementation of both models with respect to the conceptual framework have also been implemented on a use-case sample. The assessment of the results suggests that the proposed solution yields a promising direction. It is understood that the proposed system could produce adjusted schedule that can increase the operational efficiency while still being robust against the uncertain arrival time.

In this section, a numerical experiment that consists of sensitivity analysis and comparative study will be conducted to investigate how much improvement the proposed solution would provide in operation of loading facility given a different context of situation. The goal is to determine in which sort of situation this proposed system could be optimally applied considering the additional cost in requires in regard to the integration infrastructure and technology readiness.

In order to achieve the goal, the analysis will focus on the added value of combining a predictive model that assigns probability value to the real-time ETA information with a P-SRP model. A set of KPIs and scenarios are developed to properly evaluate the result. Moreover, to provide the scale of perspective, a baseline model (Section 7.5.2) that refers to a rescheduling system that assumes the real-time ETA information as deterministic values will be used as reference point to benchmark the comparison result.

Furthermore, the comparison also aims to highlight the advantage and disadvantage of the proposed probabilistic especially in the context of trade-off between operational efficiency and schedule robustness.

8.1.KPI

Loading facility is a part of complex transport network in supply chain in which disruption is constantly occurred. To fairly measure the performance of a loading facility, it is logical to analyze how well this rescheduling system would sustain the inevitable disruption. Two of most important attributes in the operational of loading facility is operational efficiency and the schedule robustness. An ideal loading facility should be able to maximize the loading slot utilization even under the uncertain situation of truck arrival. However, the schedule should also not adjust the schedule of incoming trucks in every minute. Assigning 10 trucks in 1 loading slot would ensure that the loading slot would be occupied, but it comes that the expense of doing intensive rescheduling for the other 9 trucks. In order to fairly measure the results of the proposed P-SRP in terms of operational efficiency and schedule robustness, a set of KPIs (Key Performance Indicators) is defined, as following:

a. Total expected cost

This metric is defined as the rescheduling cost. It can be calculated by the multiplying the deviation between initial and adjusted schedule with the generalized cost. In this case, it is assumed that the generalized cost is 1, therefore it can be inferred that the total expected cost is the quadratic value of initial and adjusted schedule. The value of this metric could indicate the efficiency of a loading operation.

b. Total rescheduled trucks

This metric is defined as the total number of trucks being rescheduled from their initial slot, either because of late/early arrival or delay propagation effect. As it is understood that rescheduling is a mitigation action to alleviate inefficiency problem, then it is interesting to compare how many trucks are required to be rescheduled in order for the optimization model to reach convergence. .

c. Total possible idled slots

This metric is defined as the total number of loading slot that is highly likely to be idled in reality. The purpose of implementing this metric is to investigate how many loading slots are possibly being empty because the trucks are unlikely to be present at the scheduled time, hence would badly affect the slot utilization rate. In this case, it is assumed that trucks having presence probability lower than 0.2 is considered as a no-show truck. The estimation method is detailed in Appendix E: Result

8.2.Scenario Development

Establishing set of scenarios is a popular method to capture uncertain situations in transportation study [88]. This approach allows investigating the expected impact of the system given a varied plausible future context. By observing the array of potential shift and its expected outcome, decision-maker could have better preparation/strategies to face the uncertainty.

Therefore, at this section, set of scenarios that will be used to analyze the performance of the P-SRP relative to the baseline model are introduced. Each scenario correspond to different contexts relevant to the operation of loading facility, namely congestion level in transport network, specification of loading infrastructure, scheduling policy, and the maximum permissible probability. The justification of having varied contexts of scenarios is because it would provide a more profound analysis to generalize on which exact conditions the P-SRP would generate maximum benefits. The detailed format data of scenarios is provided in Appendix D: Scenario Data.

8.2.1.Scenario 1: Congestion Level

As it is elaborated in Section 3.1.4, the uncertainty variable is the arrival time of the truck at the loading facility, whereby the largest impact stems from the troubles related to transport network, namely congestion.Varied delay ratios are assigned to each scenario of congestion level, namely mild, medium, and severe, respectively, as seen in Table 8.1. A simulation that is based on triangular probability is run to generate the set of arrival time according to a scenario. This probability distribution enables the randomization to have a particular value of lower bound, mode, and upper bound. The resulting set of arrival time for each corresponding scenario is shown in Appendix C.

Table 8.1 Scenario in regard to Congestion Level

	Mild	Medium	Severe
Congestion Level	25 %	40 %	70 %

8.2.2.Scenario 2: Loading Infrastructure

This scenario focuses on the variation in regard to configuration/specification of a loading infrastructure. Different number of loading bay is applied to denote the scale of operation. On top of that, other parameters that could be associated with this context is utilization rate that could indirectly indicate the amount of buffer time. As the scale of operation in a loading facility is proportionate to the complexity of schedule management, thus it is relevant to analyze how well the P-SRP would solve the increase of magnitude in operation of loading facility. The full variation of scenarios related to the loading infrastructure is shown in Table 8.2.

Table 8.2 Scenario in regard to Loading Infrastructure

	Small scale			Medium scale			Large scale		
Loading Bay	2			3			4		
Utilization Rate	50%	75%	100%	50%	75%	100%	50%	75%	100%

8.2.3.Scenario 3: Rescheduling Strategy

As the purpose of rescheduling is to mitigate the detrimental effect of uncertain truck arrival time, there are several other strategies beside the one explained in Section 6.5.4 in which the priority is to maintain the initial schedule. The rationale behind exploring other rescheduling policy is because of the possibility that a different approach could lead to a different sense of benefits in the operational of loading facility.

One most popular priority in rescheduling is to minimize the waiting time between actual arrival time and adjusted schedule, meaning that if trucks arrive early and there an earlier slot available, then the truck would be moved instead of waiting for its initial schedule. To simulate this situation, the objective function will be modified into a linear form. In addition, other possible strategy to mitigate conflict probability is by providing a stand-in loading bay which main purpose is to be used as an alternative option if the other loading bays are fully occupied. The full variation of scenario related to the rescheduling strategy is shown in Table 8.3.

Table 8.3 Scenario in regard to Rescheduling Policy

Priority	Minimizing waiting time $\textbf{Min} \sum_{t} \sum_{l} \sum_{j} (x_{jl}^{t}\, t - c_{jl}^{t}\, t)\; g_{jl}$		Maintaining initial schedule $\textbf{Min} \sum_{t} \sum_{l} \sum_{j} (x_{jl}^{t}\, t - c_{jl}^{t}\, t)^{2}\; g_{jl}$	
Stand-in Loading Bay	Without (3 Specific)	With (3 Specific + 1 General)	Without (3 Specific)	With (3 Specific + 1 General)

8.3. Analysis of Result

In this section, all defined scenarios will be applied to both baseline and P-SRP. The purpose of is to investigate in which certain context, implementing P-SRP yields more advantages. Therefore, KPI analysis will be conducted to fairly measure the result. The analysis will be done in the basis of each scenario.

The baseline model does not incorporate the predictive model; hence it assumes all ETA information as deterministic variable. On the other hand, P-SRP utilizes predictive model, allowing simultaneous assignments of trucks at the same loading bay based on the presence probability value.

8.3.1. Impact of Congestion Level in Transport Network

Scenarios of varied congestion level is applied to both baseline and P-SRP. The result is evaluated according to the set of KPIs stated in Section 8.1 and shown in Figure 8.1.

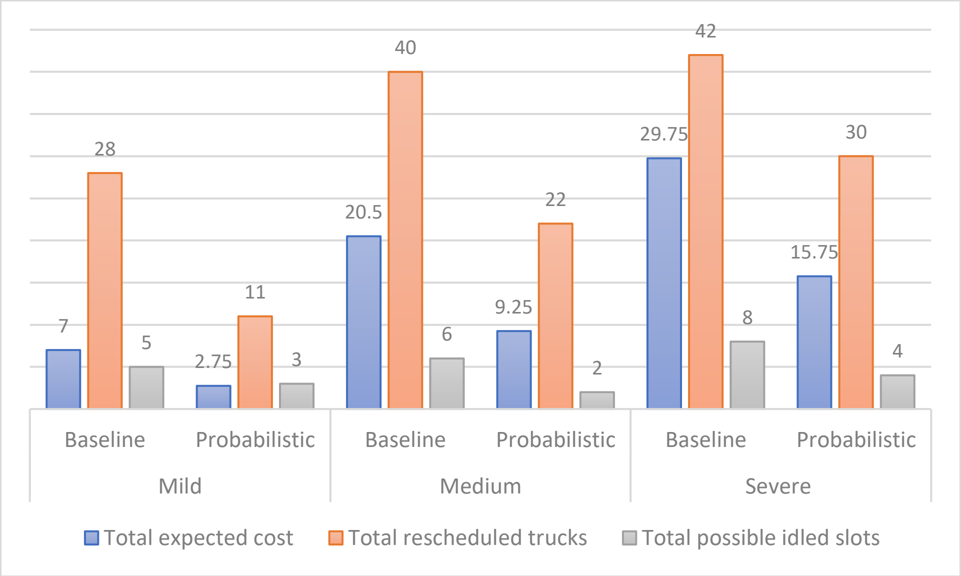


Figure 8.1 Result of Varied Congestion Level

In all variation of congestion level, P-SRP heavily outperforms the baseline model in terms of total expected generalized cost. The main reason behind it is because the advantage of P-SRP that allows assigning several trucks simultaneously based on the presence probability values and ETA information that would provide another alternative way in rescheduling the early/late trucks, thus there is a possibility of not changing the schedule of a truck even though. In other words, rescheduling could be done without additional cost to a certain extent. On the other hand, the baseline model could either move the truck to earlier or later slot in situation where rescheduling is necessary, hence the rescheduling process would always require additional cost. Consequently, given that logic of explanation, P-SRP would intuitively provide more added value in situation where congestion level is more severe. However, as it is proven by the comparison shown in Figure 8.1, this is not precisely what happened. In the mild and medium congestion level, the expected cost of the P-SRP is roughly 55% lower than the baseline model, whereas in the scenario of severe congestion level, the expected cost of the P-SRP is only 47% lower than the baseline model.

Furthermore, in terms of the total rescheduled trucks, in general, it is obvious that the baseline model reschedules more truck than the P-SRP. To be more specific, as it is indicated by Figure 8.1, in the scenario of mild congestion level and medium congestion level, the baseline model reschedules roughly twice as much as the P-SRP. However, the value significantly decreases in the scenario of severe level congestion level, where the gap between the baseline and probabilistic is merely 28%. The slump implies that in severely congested situation, the P-SRP model is not impactful. Accordingly, it can be inferred that the P-SRP slightly more suitable to be applied in condition where the necessity of rescheduling is higher due to variability of arrival, either early or late, instead of using severity of congestion as the main indicator. The reason is because normally, congestion would affect only the intensity of late arrival that implies higher number of late arrivals, with almost no occurrence of early or on-time arrival. Practically, the primary benefit of P-SRP could not be fully maximized since there is no other option except to move further the trucks to the later loading slots. Hence, the P-SRP offers less added values in the case of severe congestion compared to the medium and mild one because.

To further assess the performance gap between the probabilistic and baseline model, a KPI of possible idled slots are used. Hypothetical approach is taken to expose how inefficient the baseline model compared to the P-SRP dealing with risk of no-show truck. It means that baseline result would be evaluated based on the presence probability to check how many loading slots are assigned to trucks that most likely no going to be present at their schedule. It is assumed that truck with presence probability less than 0.2 is considered a no-show.

According to Figure 8.1, in all congestion level, proportionately, the P-SRP reduces the risk of idled slots. To get a more specific calculation of added values, Table 8.4 shows improvement provided by the P-SRP, whereby it reduced possibility of idled slots by 40% the mild congestion level, 66.67% in medium congestion level, and 50% in in severe congestion level. The main concern with the baseline model is that it assumes the ETA as perfect information and do rescheduling based on it. Consequently, ignoring the uncertainty element in truck arrival time would result on a high risk of wasting useable slots for trucks having very low or almost zero presence probability at the expense of trucks with higher presence probability.

From the combination of evaluation based on varied KPI values, the P-SRP optimally performs in the situation of medium congestion level, in the sense of lowering total expected cost, decreasing total trucks to be rescheduled and preventing idled slots that would lead to higher operational efficiency.

Table 8.4 Performance Comparison between P-SRP and Baseline in Scenario 1

	Reduced expected cost	Reduced rescheduled trucks	Reduced possible idled slots
Mild	54.27%	46.43%	40.00%
Medium	54.88%	45.00%	66.67%
Severe	47.06%	28.57%	50.00%

8.3.2. Varied Configuration of Loading Infrastructure

Scenarios of varied configuration of loading infrastructure is applied to both baseline and P-SRP. The result is evaluated according to the set of KPIs stated in Section 8.1 and shown in Figure 8.2.

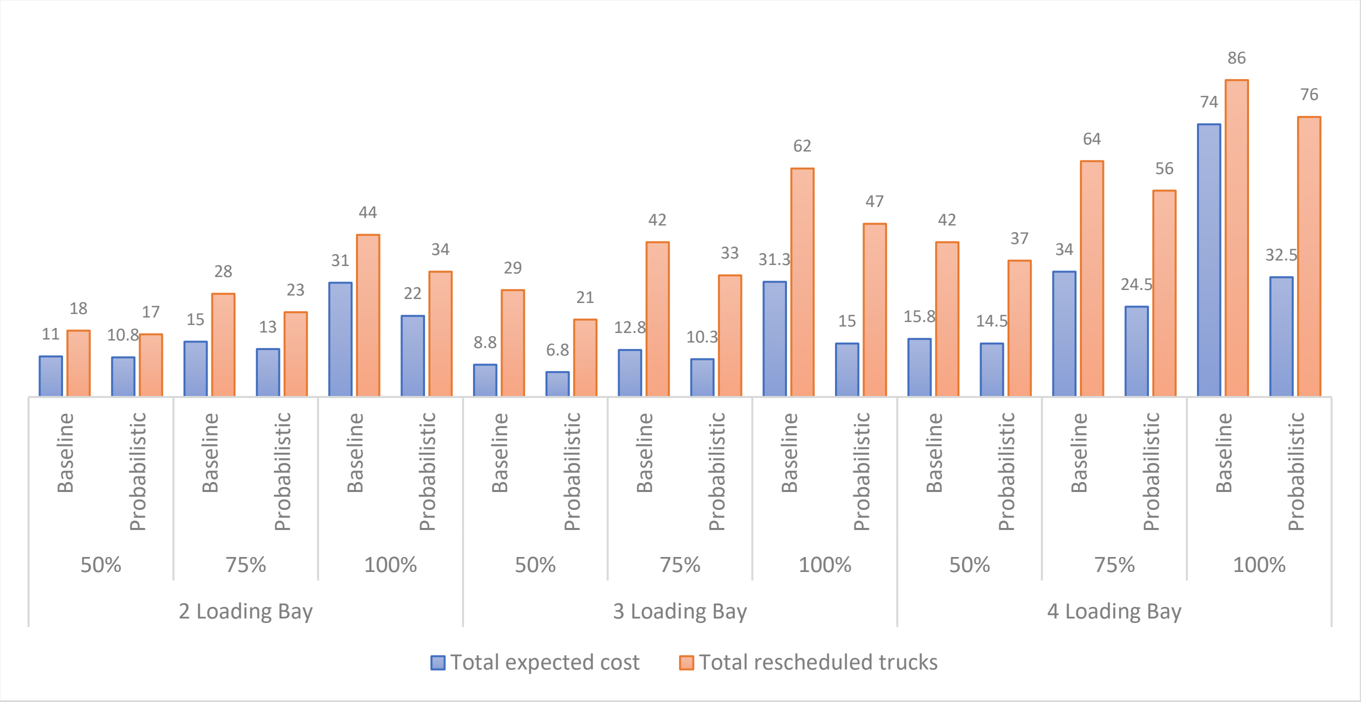


Figure 8.2 Result of Varied Specification of Loading Infrastructure

In general, the P-SRP outperforms the baseline model in terms of expected cost regardless the number of loading bays and the utilization rate. Although larger scale of operation would indeed proportionately increase the expected cost of both models, it can be distinguished that the impact of P-SRP increases as the magnitude of operation getting bigger compared to the baseline model.

To focus on scale of operation, the first analysis will be based on the variation of loading bay. As it is indicated in Table 8.2, in the operation with 4 loading bays with 100% utilization rate and 3 loading bays with 100% utilization rate, the P-SRP generates expected cost that is averagely 54.84% lower than the baseline model, whereas in the operation with 2 loading bays and 100% utilization rate, the P-SRP generates expected cost that is merely 29.03% lower than the baseline model. Therefore, it proves that the P-SRP is highly beneficial to be applied in a large-scale operation, rather than in the small-scale operation where the baseline could also provide similar result without additional cost of incorporating predictive model.

Interestingly, a more detailed analysis that focuses on variability of utilization rate points out that a decrease in utilization rate also reduce the performance gap between the P-SRP and baseline model. In the case of 100% utilization rate, the P-SRP could provide averagely 46.24% lower expected cost, indicating a significant improvement, however, the figure significantly diminishes to only averagely 11.02% and 20.29% in the situation of 50% and 75% utilization rate, respectively, indicating an almost equal performance. The implication is that implementing the P-SRP is low utilization rate case is not truly necessary considering the higher deployment cost to enable this technology since the availability of more empty slots in combination with baseline model level out the advantage of P-SRP. As less utilization rate of loading bay could also be interpreted as intense use of buffer slot strategy, it can be inferred that the P-SRP provide most added value when it is applied in situation where there is no buffer time in the operational schedule.

In general, increase in number of loading bays and utilization rate would consistently result on higher number of trucks being rescheduled in both baseline and P-SRP as it is indicated in Figure 8.2. This is a logical result because larger scale of operation implies higher number of incoming trucks, therefore the probability of having more trucks deviating from the initial schedule would proportionately increase as it is indicated by consistent upward trend in all different conditions. However, more in-depth analysis is required to comprehend the added values of P-SRP in terms of specification of loading infrastructure.

P-SRP constantly outperforms the baseline model in every case, as it is indicated by averagely 17.27% reduction in total number of rescheduled trucks between models regardless the number of loading bay and utilization rate. Based on the variation of total loading bays, the results shows that in situation of 3 loading bays, the P-SRP reduced the total rescheduled trucks by averagely 24.4%, while in situation of 2 and 4 loading bays, it reduced the total rescheduled trucks by only averagely 15.3% and 12.01%, respectively. Furthermore, based on variation of utilization rate, the comparison results are quite vague. As the consequence, it is not possible to directly derive the relation between infrastructure specification and the total rescheduled trucks. Therefore, there is no such distinctive generalization that can be derived to know at which specific configuration that the P-SRP would be beneficial in terms of total number of rescheduled trucks. Lastly, in this scenario, the KPI of possible idled slots are not applicable because the data of loading schedule, ETA information, and its corresponding presence probability are all constant, and would result on roughly similar impact in number of possible idled slot. Therefore, there is no point in utilizing this KPI in evaluating the added values of P-SRP in this scenario.

According to the result in this scenario, it can be inferred that the P-SRP offers most added values in situation of large-scale operation and high utilization rate, especially in reducing the total expected cost. In addition, in terms of total rescheduled trucks, the P-SRP provides most optimal improvement in the situation of 3 loading bays.

Table 8.5 Performance Comparison between P-SRP and Baseline in Scenario 2

	Utilization Rate	Reduced Expected Cost	Reduced Rescheduled Trucks
2 Loading Bay	50%	2.27%	5.56%
	75%	13.33%	17.86%
	100%	29.03%	22.73%
3 Loading Bay	50%	22.86%	27.59%
	75%	19.61%	21.43%
	100%	53.60%	24.19%
4 Loading Bay	50%	7.94%	11.90%
	75%	27.94%	12.50%
	100%	56.08%	11.63%

8.3.3. Impact of Varied Rescheduling Strategy

Scenarios of varied rescheduling strategy is applied to both baseline and P-SRP. The result is evaluated according to the set of KPIs stated in Section 8.1 and shown in Figure 8.3.

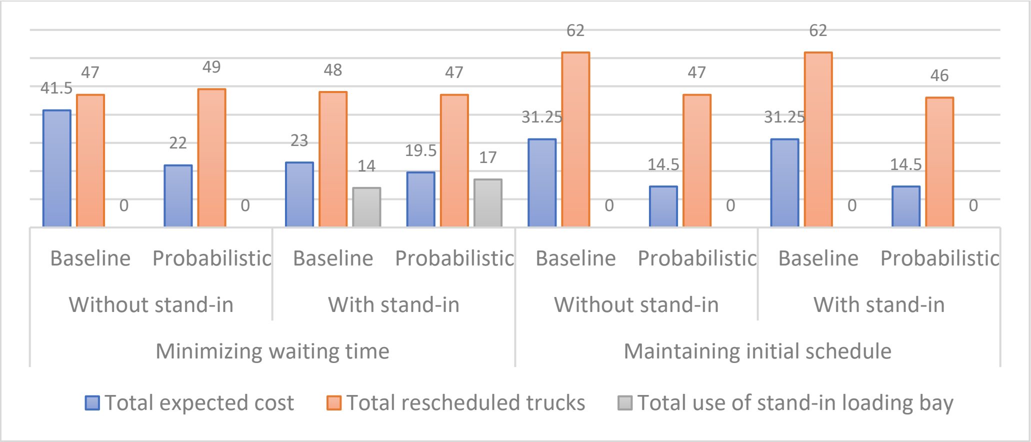


Figure 8.3 Result of Varied Rescheduling Strategy

Similar to the previous comparison result, the P-SRP results on lower expected cost compared to the baseline model. In combined situation in which the priority in rescheduling is to minimize the waiting time and there is stand-in loading bay that could serve all product types, the impact of P-SRP is insignificant, indicated of only

15.22% lower expected cost. On the other hand, in other conditions, the P-SRP reduces the expected cost by averagely 51.3%. The justification is because the existence of extra loading bay facilitates other alternative method in mitigating the early or late arrival, therefore instead of waiting for the later slot, the stand-in could be utilized to load the early or late trucks. However, this is not happening in the case where there is a stand-in loading bay, yet the rescheduling strategy is to maintain the initial schedule. When a truck arrive early than its initial schedule, this strategy is reluctant to find an early feasible loading slot if not necessary, thus having a stand-in loading bay would be meaningless.

In regard to the total number that has to be rescheduled, regardless the existence of stand-in loading bay, the result shows clear pattern which is the P-SRP requires less rescheduling process by almost 25% in the case where the strategy is to maintain the initial schedule. On the contrary, in the case where the strategy is to minimize the waiting time and there is a stand-in loading bay, the P-SRP merely offers a negligible amount of impact (2.08%), whereas in the case where there is no stand-in loading bay, the baseline model provides optimal schedule with less truck being rescheduled, although the margin is also negligible (-4.26%). To this end, since the latter comparison results illustrates a rather less obvious pattern and might be counter intuitive, the added value of P-SRP terms of total rescheduled truck with respect to rescheduling strategy could not be fully comprehended. A more experiments using a more diverse sample is necessary to investigate whether the current result is just statistical outlier. In similar fashion to the Scenario 2, the KPI of possible idled slot is not relevant here because the input data that influenced the possibility of idled slot is constant.

Based on the analysis with respect to the rescheduling strategy, it can be deduced that the P-SRP provides most added value in situation where the priority is to maintain the initial schedule. Moreover, the benefit of having a stand-in loading bay only exists in situation where the priority is to minimize waiting time, while there is no essence of having one if the preference is to maintain the initial schedule.

Table 8.6 Performance Comparison between P-SRP and Baseline in Scenario 3

		Reduced Expected Cost	Reduced Rescheduled Trucks
Minimizing waiting time	Without stand-in	46.99%	-4.26%
	With stand-in	15.22%	2.08%
Maintaining initial schedule	Without stand-in	53.60%	24.19%
	With stand-in	53.60%	25.81%

8.4. Trade-off between Robustness and Efficiency

8.4.1. Additional KPI: Possible Conflicted Slot

Ideally, robustness can be measured by observing the conflicted loading slots that occur in practice. More conflicted slots would require more rescheduling process, which results on lack of schedule robustness. However, since real observation is not possible in this research, then alternative approach based on probability value is utilized. So, this KPI is defined as the total number of loading slot in which several trucks are likely to be present at the same time and loading bay. In this case, it is assumed that trucks having presence probability higher than 0.75 is considered as absolute present, Hence, getting another/more trucks simultaneously assigned to the same loading slot would cause a conflicted slot.

8.4.2. Additional Scenario

Another parameter that can be explored to find the most ideal value is the maximum permissible overlap probability. This variable is basically the main factor in determining whether a truck could be simultaneously assigned in same timeslot and same loading bay. Consequently, the degree of this variable would directly affect the balance between robustness and efficiency of the proposed rescheduling system.

Furthermore, a set of scenarios containing varied value of the maximum permissible overlap probability is constructed. The purpose of this comparison is to see how much an increment in maximum value would impact the result of the optimization model. Hence, decision-makers could have insight whether it is worth implementing higher bound in overlap probability. To realize the idea, different amount of maximum overlap probability, namely 5%, 20%, 35%, and 50% are applied. To get a more variability on sample set, the diverse congestion levels , namely mild, medium, and severe (Scenario 1) are also added

8.4.3.Trade-off Analysis

Ideally, robustness could be easily achieved by providing large gap of buffer time, but this approach comes at the expense of slot efficiency in loading facility. For example, assigning only 1 certain truck in the slot of 2 hours would result on a less probability of rescheduling which means a very robust, but this is not operationally efficiency since the probability of slot sitting idled is increasingly high. In other words, the trade-off between robustness and efficiency should be analyzed

The core fundamental of the P-SRP is the capability to balance schedule robustness and efficiency of slot utilization which work in a trade-off manner. The key variable that could directly alter both values is the permissible overlap, denoted by r. To exhibit its influence on the schedule generated by the P-SRP, a varied value of maximum permissible overlap is applied. Furthermore, it is also intriguing to investigate the effect of congestion level towards the balance between robustness and efficiency. Hence, the final comparison would be 2 dimensional that includes a varied permissible overlap probability, and also different congestion level following the scenarios used in previous section.

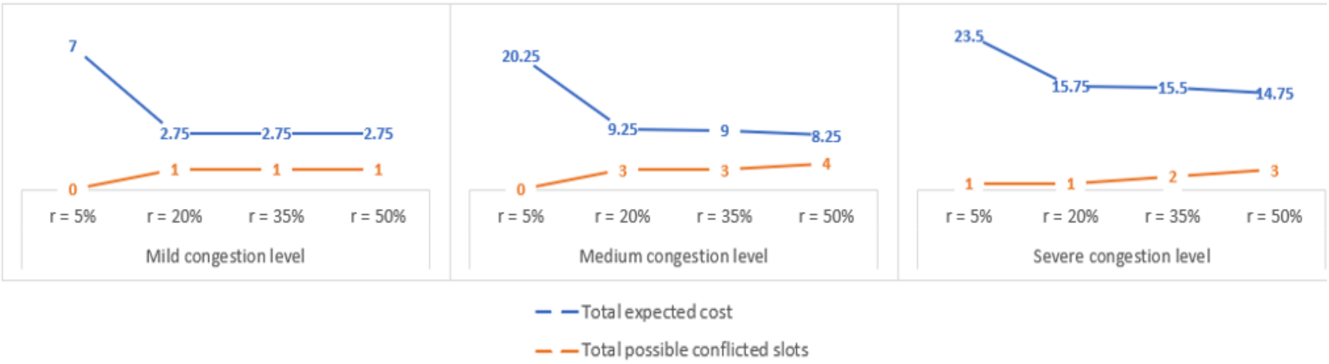


Figure 8.4 Trade-off between Schedule Robustness and Operational Efficiency

Total expected generalized cost could be used as the proxy to measure the efficiency of a loading facility. A deeper observation is done by varying the degree of permissible overlap probability, and the result is illustrated in Figure 8.4. In all variation of congestion level, applying higher value of permissible overlap probability (r) generated a less expected generalized cost, which means that the operation is more efficient.

The most important insight that can be derived is that in all scenarios of varied congestion level, the most significant increment in the efficiency occurs when the r value is shifted from 5% to 20%, and after that the improvement is quite negligible. To be more specific, in the scenario of mild congestion level, the shift from 5% to 20% in r value provides a 60% improvement in the efficiency, yet after that, the improvement is constant. Moreover, in the scenario of medium congestion level, raising 5% of r value to 20% of r value yields 52% higher efficiency, while after that increasing the r value only bears averagely 2.5% improvement in efficiency metric. Lastly, in the case of severe congestion level, given the same fashion, the first improvement is in the high of 32%, whereas after that, set of higher r values only yield averagely 3% increase in efficiency metric.

The total number of possible conflicted slots could be used as indicator to measure how often a rescheduling is required; hence the schedule is not robust. Given this logic, a higher number of possible conflicts would mean a lack of robustness. With respect to the variation of permissible overlap probability, the results indicates that higher r values generated a higher number of conflicted slots, even though the rate of change is insignificant. It bears a logical sense, because if the model allows a higher overlap probability, then more conflicts should be expected.

Hence, it can be inferred that improvement in efficiency at the loading operation does come at the expense of schedule robustness.

A closer look in the comparison indicates an interesting point that a higher level of congestion doesn't guarantee a higher number of conflicted slots, as it is indicated by scenario of severe congestion level. The experiment shows that the conflicted loading slots tend to happen more where the total simultaneous assignments is higher, as it is indicated by the scenario of medium congestion level compared to the rest of the scenarios.

Moreover, shifting the value of permissible overlap probability from 5% to 20% provides most impact on both KPIs compared to the rest of increments. One way to interpret is that 20% permissible overlap probability could potentially offer the optimal balance between operation efficiency and schedule robustness. Nonetheless, there is no clear guideline or precedent researches that indicate how to correctly determine the best value of permissible overlap probability. Ultimately, it is the privilege of decision-makers to set their own acceptable level of overlap probability given the knowledge of trade-off between the efficiency and robustness. One of important factors in determining the value is to consider the risk profile of the operation. It means that if the decision-makers prefer to have the least expected generalized cost regardless the number of conflicted slots, then efficiency is the priority, and vice versa. To this end, it is important to note that this research does not aim to find the optimal balance between efficiency and robustness, the goal of this research is rather to exhibit the trade-off between robustness and efficiency as the result of the proposed model.

8.5.Chapter Overview

A numerical experiment consisting of comparative study and sensitivity analysis has been done in this chapter. Three set of scenarios are developed which have different goals to achieve. This comparison will be multidimensional in regard to scenarios that includes varied parameters. It means that the performance will be analyzed in such way to clearly highlight the advantage and disadvantage of an approach given a certain context. The main focus to explore how well the P-SRP would perform under varied scenarios compared to the baseline model that assumes ETA as deterministic variable.

Consequently, the first analysis aims to exhibit the added value of incorporating predictive model to capture the uncertainty of ETA. Each scenario refers to a different context relevant to the operation of loading facility, namely the congestion level of transport network, the configuration/specification of loading infrastructure, and rescheduling strategy. The second one is to measure the trade-off between operational efficiency and schedule robustness by applying varied maximum value of permissible overlap.

In the scenario of varied congestion level, the comparison reveals that the P-SRP outperforms the baseline model in all KPIs evaluation. The P-SRP offers a rather equal improvements in terms of expected cost regardless the congestion level, whereby it reduces the expected cost by averagely 52.07% compared to the baseline model. With respect to the total rescheduled trucks, the P-SRP requires rescheduling roughly 45.7% less than baseline model, yet in the severe congestion level, the value decreases to only 28%. In preventing idled slots, the P-SRP is able to diminish the expected value by 66.67% lower than the baseline model compared to only 40% and 50% in mild and severe congestion level, respectively. Hence, it can be concluded that the probabilistic has most added values when being implemented in the situation of medium congestion level.

In the scenario of varied specification/configuration of loading infrastructures, in general, the P-SRP has significant edges compared to the baseline model. With focus on the scale of operation, the P-SRP provides averagely 54.84% lower expected cost compared to baseline model in condition where there are 4 loading bays, whereas the improvement reduces to only 29.03% in condition where there are only 2 loading bays. Therefore, the result proves that P-SRP is promising to be implemented in large-scale operation. With respect to the utilization rate, lower utilization rate reduces the performance gap between probabilistic and baseline model, whereby in situation of 100% utilization rate, it provides averagely 46.24% lower expected cost, whereas the improvement only accounts for 11.02% and 20.29% in situation of 50% and 75% utilization rate, respectively. Hence, it can be derived that the P-SRP provides most added value in large scale of operation with more intense utilization rate of loading slots.

In the scenario of varied rescheduling strategy, the focus is to explore the impact of other rescheduling priority and to assess the merit of having a stand-in loading bay as an alternative loading. One interesting point is that in situation where the priority is minimizing waiting time and there is a stand-in loading bay, the P-SRP offers an insignificant improvement, only 15.22% lower than the baseline model, whereas in other combined situations the value reaches 51.39%. Moreover, it is observed that the stand-in loading bay is only utilized if the priority is to minimize the waiting time, otherwise, there is no substantial impact in having stand-in loading bay. Thus, it can be inferred that probabilistic would be optimal when the priority is to maintain the initial schedule and there is no stand-in loading bay available.

The last analysis is grounded upon the notion that efficiency and schedule robustness works in trade-off manner. The result shows P-SRP model offer advantage to balance the schedule robustness and operational efficiency by altering the maximum value of permissible overlap. It is found that shifting the value from 5% to 20% yields most significant improvement on the operational efficiency with negligible conflicted slot. However, the most optimal trade-off cannot be determined because it depends on the preference of decision-makers.

9. Conclusion & Recommendation

The previous chapters has presented key findings and discussions in regard to the purpose of this research. A literature review has been done to find the knowledge gap and to provide relevant theories in building conceptual framework to integrate real-time ETA information and historical data of truck arrival time. To understand the data quality and to derive valuable insights embedded in the large database, data exploration technique is executed. Subsequently, predictive model and optimization model have been successfully built and evaluated to highlight how promising the methods in solving the uncertainty issue in operation of loading facility. Moreover, comparative study and sensitivity analysis have also been conducted to point-out the added value and the pros-cons of this approach.

This chapter aims to conclude and provide recommendations both for the practical matter and for this research domain. In addition, this chapter also explains the limitations, as well as the contribution of this research.

9.1. Conclusion

In this section, the answers to all research questions are presented. The research is driven by the recommendation of previous study and the knowledge gap in the literatures of this study domain. The combination leads to the main research question, which is:

“To what extent an integrated system of logistics and traffic would improve the schedule management of petrochemical loading facility considering the stochastic nature in transportation?”

9.1.1. Answers to Sub Research Questions

A set of sub-research questions is composed to address the main research question sequentially and coherently, as following:

SQ 1: *Is it conceptually possible to design schedule management that is more efficient and robust against real-time disturbance?*

This first research question is mainly answered through literature reviews. Firstly, as is done Chapter 2, investigation that includes previous and state-of-the-art studies in the relevant domain is conducted. The result reveals that the research on schedule management of petrochemical loading facility lags behind its equivalent counterparts, such as airport operation and railway management. Since prediction of arrival time holds an extremely vital role in dealing with uncertainty in transport network, most of the recent researches utilizes Machine Learning (ML) techniques to generate better prediction results. On top of that, optimization based on MILP concept are proven to be sufficient in solving the rescheduling issues. Therefore, it is decided that those two models would be the core of the proposed conceptual framework.

Secondly, additional literature reviews done in Chapter 3 established theoretical foundation to adequately design conceptual framework for advanced schedule management at a petrochemical loading facility. Accordingly, it is derived that the framework would serve as a decision support-system (DSS) consisting of predictive and optimization system that would allow reactive scheduling based on historical and real-time ETA information. Moreover, a precedent survey indicates that factors having most influence in hindering the efficiency of a loading facility are related to transport network problem, therefore, given the scope of this study, it is logical to focus on dealing with uncertain truck arrival time. To capture the stochasticity, presence probability derived from historical arrival distribution is suggested as one solution.

By combining all the gathered knowledge, a rescheduling system is established. The main feature of this conceptual framework is the integration between real-time ETA information and historical data. For the implementation, rolling horizon is applied to accommodate a continuously updated real-time ETA information. So, to conclude, it is possible and potentially beneficial to use a rescheduling system that integrates the real-time ETA information and the historical data of truck arrival time.

SQ 2: *How to improve prediction of truck arrival time based on real-time information?*

The second research question is answered through data exploration (Chapter 4) and by developing predictive model (Chapter 5) based on ML algorithm. Data of historical arrival time for the past 5 years are synthetically generated due to the time limitation of this project. To derive valuable insights and to ensure the validity of the synthetic data with respect to the existing theories, data exploration is conducted. In addition, data exploration also exhibits the relation between parameters, which would be useful in pre-processing in developing predictive model. It is found that the most important factor in affecting deviation in arrival time is the time of the day when the truck departed from the origin location. One conclusion can be derived that the result of data exploration indicates that is the synthetic data represents the expectation and does not contradict existing knowledges, especially in regard to the pattern of the truck arrival time.

In developing predictive model based on ML techniques, several classification algorithms are applied, namely Gaussian Naïve Bayesian (GNB), Logistic Regression (LR) and Artificial Neural Network (ANN). The classification method is chosen because it is perfectly aligned with the purpose of the model which is to find a probability of a particular truck belongs to a certain arrival class (early, on-time, or late). Set of scenarios consisting different feature combination is developed, and based on the KPI scores, the features combination that include the features enabled by the incorporation of real-time information shows the best overall result. The evaluation results according to the set of KPIs shows that ANN model overall with F1 score of 0.7 outperforms the LR and GBN which only score 0.65 and 0.52 respectively.

SQ 3: *What are the benefits of optimization model that considers stochastic arrival time in improving the loading operation?*

The third research question is answered by developing an optimization model based on MIQP concept (Chapter 6). To capture the stochastic element, which is only the truck ETA, a concept of expected value is utilized. This concept provides way to facilitate a generalization of stochastic variable as its deterministic equivalent, resulting on more straightforward computation. P-SRP (Probabilistic Slot Rescheduling Problem) model that considers ETA information to be stochastic variable with certain probabilistic value generated by the predictive model is proposed.

As the utilization of real-time ETA enables rescheduling to be done in advance, there is additional advantage if that value is considered as probabilistic instead of as a perfect information. It facilitates a more intelligent rescheduling process, in the sense that presence probability of a certain truck could be taken into account. Moreover, acknowledging the presence probability also allows simultaneous assignment of several trucks to a loading slot which directly affects the expected rescheduling cost that consists of deviation between adjusted and initial schedule multiplied by generalized cost. According to the verification and evaluation (Chapter 7), the P-SRP results on 42% lower cost compared to baseline model that assumes ETA as deterministic variable.

Lastly, the proposed system is tested on a dynamic setting that represents a continuous operation of loading facility. As this system is designed to work on rolling horizon, some iterations are done to show how this proposed system would exactly work in practice. In order to exhibit the potential benefit of it, an observation data that represents actual arrival of truck is synthetically generated as the benchmark. The analysis shows that this proposed system provides a robust solution against uncertain arrival time as the adjusted schedule proves to be feasible compared to the actual arrival. Moreover, other important remarks, realistically, the proposed system could prevent underutilization of loading slot by reducing possible idled slot and unnecessary schedule shift.

To conclude, P-SRP model yields promising solution in mitigating the detrimental effect of uncertain arrival time. It could improve the operational efficiency of a loading facility, while still maintain a certain level of schedule robustness. However, the results might be case-specific considering the limited sample used in the verification and evaluation.

SQ 4: *How much added value does the proposed solution offer in dealing with uncertain arrival time?*

Lastly, the fourth research question is answered through conducting numerical experiments consisting of sensitivity analysis and comparative study (Chapter 8). The first part aims to highlight the added value of the predictive model and probabilistic optimization; thus the comparison is conducted between the baseline model which assumes the real-time ETA information as deterministic and the probabilistic which assigns a probability value to the real-time ETA information. Three set of scenarios are developed, the first one consists of varied congestion level, the second one consists of varied specification/configuration of loading infrastructure, and the last one consists of varied rescheduling strategies. Moreover, relevant KPIs to measure the operational efficiency are also defined.

Scenarios of varied congestion level indicate that P-SRP has the most significant impact medium congestion level, shown by reduced expected cost by averagely 52.07% and reduced risk of idled by averagely 52.23%. Interestingly, it can be deduced applying probabilistic in medium congestion level yields the best result in mitigating the potential of idled slot, because it demands more intensity of rescheduling (due to early and late arrival) than severe congestion level, not precisely because of the parameter congestion level itself. Scenario of varied configuration and specification points out that P-SRP has most significant edge over the baseline model in largest scale of operation (100% utilization rate and 4 loading bays), whereby it reduces the expected cost by averagely 54.84%. Lastly, in the scenario of varied rescheduling strategy, the P-SRP model offers best performance in situation without stand-in loading bays and the priority is to maintain the initial schedule, in which it decreases the expected cost by 51.39% compared to the baseline. However, in the strategy of minimizing waiting time and there is stand-in loading bay, the P-SRP only reduces the expected cost 15.22% compared to the baseline model.

As it is understood that the probabilistic approach provides a certain level of robustness, and it works in trade-off manner with operational efficiency, thus the second part aims to explore the extent of it. The trade-off could be balanced by altering the value of permissible overlap probability (r), thus scenarios of varied r value is added to the first comparison. The results prove that increased operational efficiency comes at the expense of schedule robustness. Importantly, the rate of change is not linear. The most significant improvement of efficiency occurs while shifting the r value from 5% to 20% that accounts for averagely 48% regardless the congestion level. Afterwards, constantly increasing the value only result on averagely 2.75% improvement in operational efficiency. The increased of efficiency does come at the expense of the reduced robustness, however, in this case the rate of change and the proportion of conflict is negligible. To put into perspective, when applying r value of 20%, the conflicted loading slots is roughly 5% of total slots which can be considered as relatively negligible.

To conclude, the proposed solution yields optimal added values in situation characterized by high number of reschedule being required, large scale of operation in terms of quantity of loading bay and utilization rate, the priority is to maintain initial schedule, and there is no stand-in loading bay available.

9.1.2. Answers to Main Research Question

This study shows that it is conceptually possible to improve the operational efficiency of a petrochemical loading facility by having an integrated system between logistic operation and traffic system. It proposes an advanced schedule management that allows intelligent rescheduling. The main elements of the rescheduling system are the predictive and optimization model that integrates the real-time information and historical data of truck arrival. The implementation potentially provides valuable managerial insights for operational planning in terms of mitigating propagation effect of arrival uncertainty and manpower/facility resources allocation.

9.1.3.Contribution

This research contributes to fill the stated knowledge gaps. In regard to the first knowledge gap which is the lack of exact method for capturing uncertainty in operational of loading facility , this study shows that considering a stochastic arrival time in the form of discrete probability could improve the operational efficiency of loading facility. In regard to the second gap, which is lack of predictive model based on ML algorithm, this study shows that predicting arrival class (early, on-time, late) is possible and beneficial in maintaining a high-performance operation.

The study is conducted in more general sense; therefore the proposed conceptual framework is not supposed to tackle a tailor-fit problem. Given the benefit of the generality of this study, the models and solutions presented in this paper are adaptable for other similar operational problem. Specifically, in situation where the assignment are object specific, unlike the ones in existing literatures where the assignment process is in universal sense.

9.2.Limitation

This research is established on set of assumptions and simplification which might lead to imperfect outcome. Nevertheless, this does not disprove the fact several principal insights might still be valid and relevant. The limitation in this research is detailed in the following section.

The time constraint in finishing the research has been a deterrent from pursuing more profound research. It leads to inability to obtain a real data and to conduct real observation at the loading facility, enforcing the proposed conceptual framework to be tested on synthetic data. Since the synthetic data is adapted from other relevant domain due to inexistent of publicly accessible data on petrochemical loading facility, there are several specific data characteristics of operation in loading facility that are missing, such as frequency of disrupted operation due to facility maintenance, specific routes taken by trucks, weather condition, live traffic data, etc.

In regard to development of predictive model, a computational power is a crucial element. This factor directly corresponds to capability of having ANN with more layers that could potentially generate more valid and reliable prediction. Moreover, it is understood that ML algorithm and data exploration technique applied in this research are highly dependent on both quality and quantity of input data. Although the result is sufficient in this case, there is a chance of overestimating the correctness due to the fact of using synthetic data which may contain statistical outliers and less contextual value, thus this matter should be heavily considered in the real implementation.

In regard to development of optimization model, a lack in specific facility data and access to the managerial personnel are the main concern. It results on inability to incorporate full variables relevant to the operation of loading facility, for instance, the generalized cost is limited to only the deviation between initial and adjusted schedule because there is no additional information to determine the weight/penalty of other factors that are essential in efficiency measure. The main reason is because the writer is not directly and legally involved in a certain project of petrochemical loading facility.

Lastly, there is no empirical proof on how significant the proposed conceptual method would improve the current situation. All studies done in this research is limited to the use of hypothetical analysis, therefore there is no absolute way in stating the increase of efficiency in practical implementation

9.3.Recommendation

According to the main findings and acknowledged limitations of this research, a set of recommendations could be inferred. In the following section, recommendations for future research direction will be presented.

- Implement real historical data

Synthetic data provides sufficient validity only to a limited extent. In implementing this technology to solve real problems, real data must be incorporated. The predictive model must learn from the real data because even the perfectly refined synthetic data still might lack the contextual value of the problem.

- Apply other assessment methods

The assessment done in this research could only be able to theoretically showcase the added value of the proposed conceptual framework and to exhibit the trade-off between operational efficiency and schedule robustness. However, in practice, there will be more consideration in implementing a new approach. To get a more comprehensive outlook, other assessment methods such as CBA or MCDA could be applied. Moreover, currently, this research does not aim to calculate the most acceptable level of trade-off between efficiency and robustness. Hence, by being able to weigh the tangible/intangible benefit against the cost of the trade-off, a more well-founded acceptable risk level could be determined.

- Consider more stochastic variables

This research only assumes truck arrival time as the stochastic element, while loading and finish time are still deterministic. It is indeed a promising direction to also consider the two latter variables as stochastic, thus it would enable a more advanced way to estimate presence probability by combining the ECDF of arrival and ECDF of departure time. The implication is the potential of applying dynamic or adaptive length of loading slots to increase the operational efficiency, instead of moving truck assignment based on its arrival class probability done in this research.

- Add more features

As it is understood that the predictive model based on ML algorithm could only generate results as good as the input of historical data. Given this key principle, there is merit in utilizing data containing more relevant attributes such as weather condition, type of road disruption (congestion, road diversion, road maintenance), the occurrence of accidents, routes taken by trucks, time record at series of GPS coordinate, etc.

- Improve predictive model

As technology is growing rapidly, the predictive model could be more sophisticated. The result would be more promising, yet it might come at the cost of a higher requirement of computational power. Since the predictive model built for this research is in default setting in terms of hyperparameter, another relatively simple approach that might yield positive results is a process called “hyperparameter tuning”.

- Improve stochastic model

It is interesting to apply the ideal stochastic programming that includes several stages of recourse stages. It could not be realized in this research due to data limitation that hinders the development of scenarios to capture uncertainty. Moreover, other optimization approaches such as stochastic job-shop scheduling that optimizes the order of service based highest presence probability value, could also be an interesting alternative.

- Improve automation of models

Currently, the execution of the integrated model is done manually to a certain extent. In order to achieve the ideal envisioned implementation, the algorithm should be improved. The iteration of the integrated system should be seamlessly automated following the flow of real-time information. This means that the prediction and optimization would be continuously done following the fluctuation of real-time ETA information over the rolling horizon. Hence, it is unnecessary for the planner to manually iterate the process for every certain period of time.

Bibliography

[1] A. Dolgui, D. Ivanov, and B. Sokolov, “Ripple effect in the supply chain: an analysis and recent literature,” *International Journal of Production Research*, vol. 56, no. 1–2, pp. 414–430, 2018.

[2] R. Chira, “The role of transport activities in logistics chain,” *Knowledge Horizons. Economics*, vol. 6, no. 3, p. 17, 2014.

[3] H. M. S. Lababidi, M. A. Ahmed, I. M. Alatiqi, and A. F. Al-Enzi, “Optimizing the supply chain of a petrochemical company under uncertain operating and economic conditions,” *Industrial & Engineering Chemistry Research*, vol. 43, no. 1, pp. 63–73, 2004.

[4] R. Ritu Raj, “Resilient Logistics & Distribution System: A Conceptual framework for ABC,” 2019.

[5] Port of Antwerp, “Antwerp remains the ideal location for logistics,” *https://www.portofantwerp.com/en/news/antwerp-remains-ideal-location-logistics*, Jul. 17, 2019.

[6] TomTom, “Antwerp Historical Traffic Data,” *https://www.tomtom.com/en_gb/traffic-index/antwerp-traffic/*, 2021.

[7] V. Sanchez-Rodrigues, A. Potter, and M. M. Naim, “Evaluating the causes of uncertainty in logistics operations,” *The International Journal of Logistics Management*, 2010.

[8] E. Zehendner and D. Feillet, “Benefits of a truck appointment system on the service quality of inland transport modes at a multimodal container terminal,” *European Journal of Operational Research*, vol. 235, no. 2, pp. 461–469, 2014.

[9] W. Zhao and A. v Goodchild, “Using the truck appointment system to improve yard efficiency in container terminals,” *Maritime Economics & Logistics*, vol. 15, no. 1, pp. 101–119, 2013.

[10] A. Ramírez-Nafarrate, R. G. González-Ramírez, N. R. Smith, R. Guerra-Olivares, and S. Voß, “Impact on yard efficiency of a truck appointment system for a port terminal,” *Annals of Operations Research*, vol. 258, no. 2, pp. 195–216, 2017.

[11] B. Wibowo and J. Fransoo, “Joint-optimization of a truck appointment system to alleviate queuing problems in chemical plants,” *International Journal of Production Research*, vol. 59, no. 13, pp. 3935–3950, 2021.

[12] Y. Huiyun, L. Xin, X. Lixuan, L. Xiangjun, J. Zhihong, and B. Zhan, “Truck appointment at container terminals: Status and perspectives,” in *2018 Chinese Control And Decision Conference (CCDC)*, 2018, pp. 1954–1960.

[13] R. Larbi, G. Alpan, P. Baptiste, and B. Penz, “Scheduling cross docking operations under full, partial and no information on inbound arrivals,” *Computers & Operations Research*, vol. 38, no. 6, pp. 889–900, 2011.

[14] R. Musaddi, A. Jaiswal, J. Pooja, M. Girdonia, and M. S. Minu, “Flight delay prediction using binary classification,” *Int. J. Emerg. Technol. Eng. Res.(IJETER)*, vol. 6, pp. 34–38, 2018.

[15] R. Nigam and K. Govinda, “Cloud based flight delay prediction using logistic regression,” in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 662–667.

[16] Y. Ding, “Predicting flight delay based on multiple linear regression,” in *IOP conference series: Earth and environmental science*, 2017, vol. 81, no. 1, p. 012198.

[17] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris, “Prediction of weather-induced airline delays based on machine learning algorithms,” in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, 2016, pp. 1–6.

[18] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, "A deep learning approach to flight delay prediction," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, 2016, pp. 1–6.

[19] M. Yaghini, M. M. Khoshraftar, and M. Seyedabadi, "Railway passenger train delay prediction via neural network model," *Journal of advanced transportation*, vol. 47, no. 3, pp. 355–368, 2013.

[20] V. Kumar, B. A. Kumar, L. Vanajakshi, and S. C. Subramanian, "Comparison of model based and machine learning approaches for bus arrival time prediction," in *Proceedings of the 93rd Annual Meeting*, 2014, pp. 14–2518.

[21] O. R. P. van Schaijk and H. G. Visser, "Robust flight-to-gate assignment using flight presence probabilities," *Transportation Planning and Technology*, vol. 40, no. 8, pp. 928–945, 2017.

[22] S. B. Boswell and J. E. Evans, *Analysis of downstream impacts of air traffic delay*. Lincoln Laboratory, Massachusetts Institute of Technology Lexington, 1997.

[23] E. Mueller and G. Chatterji, "Analysis of aircraft arrival and departure delay characteristics," in *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, 2002, p. 5866.

[24] Y. Zhou, L. Yao, Y. Chen, Y. Gong, and J. Lai, "Bus arrival time calculation model based on smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 81–96, 2017.

[25] H. Xu and J. Ying, "Bus arrival time prediction with real-time and historic data," *Cluster Computing*, vol. 20, no. 4, pp. 3099–3106, 2017.

[26] Z. Qingcheng, Z. Xiaoju, C. Wenhao, and Z. Xiaocong, "Optimization model for truck appointment based on BCMP queuing network," *Journal of Systems Engineering*, vol. 28, no. 5, pp. 592–599, 2013.

[27] J. Yang, Y. Ding, and F. Gen, "Research on Optimal Equipment Allocation in Container Terminals by Using Closed Queueing Networks," *Chinese Journal of Management Science*, vol. 14, no. 6, pp. 56–60, 2006.

[28] Z. Wu and J. Zhang, "Optimal Design of Container Terminal's Gate System Based on M/G/K Queuing Model," in *ICTE 2011*, 2011, pp. 2671–2676.

[29] N. Li, G. Chen, K. Govindan, and Z. Jin, "Disruption management for truck appointment system at a container terminal: A green initiative," *Transportation Research Part D: Transport and Environment*, vol. 61, pp. 261–273, 2018.

[30] M.-H. Phan and K. H. Kim, "Negotiating truck arrival times among trucking companies and a container terminal," *Transportation Research Part E: Logistics and Transportation Review*, vol. 75, pp. 132–144, 2015.

[31] Q. L. Xu, L. J. Sun, X. P. Hu, and L. R. Wu, "Optimization model for appointment of container trucks with nonstationary arrivals," *Journal of Dalian University of Technology*, vol. 54, no. 5, pp. 589–596, 2014.

[32] C. Caballini, J. Mar-Ortiz, M. D. Gracia, and S. Sacone, "Optimal truck scheduling in a container terminal by using a Truck Appointment System," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2525–2530.

[33] C. Yang and Z.-Q. Lu, "Optimizing time windows for delivering export container using genetic algorithm," *Application Research of Computers*, vol. 30, no. 6, pp. 1643–1646, 2013.

[34] F. Schulte, E. Lalla-Ruiz, R. G. González-Ramírez, and S. Voß, "Reducing port-related empty truck emissions: a mathematical approach for truck appointments with collaboration," *Transportation Research Part E: Logistics and Transportation Review*, vol. 105, pp. 195–212, 2017.

[35] N. Huynh and C. M. Walton, "Robust scheduling of truck arrivals at marine container terminals," *Journal of transportation engineering*, vol. 134, no. 8, pp. 347–353, 2008.

[36] N. Huynh, D. Smith, and F. Harder, “Truck appointment systems: where we are and where to go from here,” *Transportation Research Record*, vol. 2548, no. 1, pp. 1–9, 2016.

[37] A. Azab, A. Karam, and A. Eltawil, “A simulation-based optimization approach for external trucks appointment scheduling in container terminals,” *International Journal of Modelling and Simulation*, vol. 40, no. 5, pp. 321–338, 2020.

[38] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, “A review on flight delay prediction,” *arXiv preprint arXiv:1703.06118*, 2017.

[39] S. van der Spoel, C. Amrit, and J. van Hillegersberg, “Predictive analytics for truck arrival time estimation: a field study at a European distribution centre,” *International journal of production research*, vol. 55, no. 17, pp. 5062–5078, 2017.

[40] H. Lee, W. Malik, and Y. C. Jung, “Taxi-out time prediction for departures at Charlotte airport using machine learning techniques,” in *16th AIAA Aviation Technology, Integration, and Operations Conference*, 2016, p. 3910.

[41] J. J. Rebollo and H. Balakrishnan, “Characterization and prediction of air traffic delays,” *Transportation research part C: Emerging technologies*, vol. 44, pp. 231–241, 2014.

[42] S. Pongnumkul, T. Pechprasarn, N. Kunaseth, and K. Chaipah, “Improving arrival time prediction of Thailand’s passenger trains using historical travel times,” in *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2014, pp. 307–312.

[43] N. Marković, S. Milinković, K. S. Tikhonov, and P. Schonfeld, “Analyzing passenger train arrival delays with support vector regression,” *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 251–262, 2015.

[44] W. Barbour, J. C. M. Mori, S. Kuppa, and D. B. Work, “Prediction of arrival times of freight traffic on US railroads using support vector regression,” *Transportation Research Part C: Emerging Technologies*, vol. 93, pp. 211–227, 2018.

[45] L. Vorage, “Predicting Probabilistic Flight Delay for Individual Flights using Machine Learning Models,” 2021.

[46] S. Dutrieux, “Predicting Flight Delay Distributions: A Machine Learning-Based Approach at a Regional Airport,” 2021.

[47] A. Shapiro and A. Philpott, “A tutorial on stochastic programming,” *Manuscript. Available at www2. isye. gatech. edu/ashapiro/publications. html*, vol. 17, 2007.

[48] S. Sen and J. L. Higle, “An introductory tutorial on stochastic linear programming models,” *Interfaces*, vol. 29, no. 2, pp. 33–61, 1999.

[49] M. Şeker and N. Noyan, “Stochastic optimization models for the airport gate assignment problem,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 48, no. 2, pp. 438–459, 2012.

[50] L. Meng and X. Zhou, “Robust single-track train dispatching model under a dynamic and stochastic environment: A scenario-based rolling horizon solution approach,” *Transportation Research Part B: Methodological*, vol. 45, no. 7, pp. 1080–1102, 2011.

[51] H. Sahebi, S. Nickel, and J. Ashayeri, “Strategic and tactical mathematical programming models within the crude oil supply chain context—A review,” *Computers & chemical engineering*, vol. 68, pp. 56–77, 2014.

[52] C. Lima, S. Relvas, and A. P. F. D. Barbosa-Póvoa, “Downstream oil supply chain management: A critical review and future directions,” *Computers & Chemical Engineering*, vol. 92, pp. 78–92, 2016.

[53] F. F. A. Caniato and J. Rice, “Building a secure and resilient supply chain,” 2003.

[54] L. Chi, E. Hartono, C. W. Holsapple, and X. Li, "Organizational decision support systems: Parameters and benefits," in *Handbook on Decision Support Systems 1*, Springer, 2008, pp. 433–468.

[55] J. Zak, "Decision support systems in transportation," in *Handbook on Decision Making*, Springer, 2010, pp. 249–294.

[56] J. Casas, A. Torday, J. Perarnau, M. Breen, and A. R. de Villa, "Decision Support Systems (DSS) for traffic management assessment: Notes on current methodology and future requirements for the implementation of a DSS," *Trans Res Arena*, pp. 1–10, 2014.

[57] D. Ivanov, A. Dolgui, and B. Sokolov, "Robust dynamic schedule coordination control in the supply chain," *Computers & Industrial Engineering*, vol. 94, pp. 18–31, 2016.

[58] I. Sabuncuoglu and S. Goren, "Hedging production schedules against uncertainty in manufacturing environment with a review of robustness and stability research," *International Journal of Computer Integrated Manufacturing*, vol. 22, no. 2, pp. 138–157, 2009.

[59] I. Parolas, "ETA prediction for containerships at the Port of Rotterdam using Machine Learning Techniques," 2016.

[60] G. Fancello, C. Pani, M. Pisano, P. Serra, P. Zuddas, and P. Fadda, "Prediction of arrival times and human resources allocation for container terminal," *Maritime Economics & Logistics*, vol. 13, no. 2, pp. 142–173, 2011.

[61] S. Maiti, A. Pal, A. Pal, T. Chattopadhyay, and A. Mukherjee, "Historical data based real time prediction of vehicle arrival time," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2014, pp. 1837–1842.

[62] A. M. Hernández, D. Scarlatti, and P. Costas, "Real-time estimated time of arrival prediction system using historical surveillance data," in *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 2019, pp. 174–177.

[63] A. Balster, O. Hansen, F. Hanno, and A. Ludwig, "An eta prediction model for intermodal transport networks based on machine learning," *Business & Information Systems Engineering*, vol. 62, no. 5, pp. 403–416, 2020.

[64] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[65] H. Xue, S. Sun, G. Venkataramani, and T. Lan, "Machine learning-based analysis of program binaries: A comprehensive study," *IEEE Access*, vol. 7, pp. 65889–65912, 2019.

[66] A. Perez, P. Larranaga, and I. Inza, "Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes," *International Journal of Approximate Reasoning*, vol. 43, no. 1, pp. 1–25, 2006.

[67] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in neural information processing systems*, 2002, pp. 841–848.

[68] S. I. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 90–94.

[69] D. A. Belsley, "A guide to using the collinearity diagnostics," *Computer Science in Economics and Management*, vol. 4, no. 1, pp. 33–50, 1991.

[70] Bureau of Transportation Statistics, "Reporting Carrier On-Time Performance (1987-present)," *https://transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FGJ*.

[71] H. A. Rakha, I. El-Shawarby, M. Arafeh, and F. Dion, “Estimating path travel-time reliability,” in *2006 IEEE Intelligent Transportation Systems Conference*, 2006, pp. 236–241.

[72] D. J. Hand and K. Yu, “Idiot’s Bayes—not so stupid after all?,” *International statistical review*, vol. 69, no. 3, pp. 385–398, 2001.

[73] D. W. . Hosmer, S. Lemeshow, and R. X. . Sturdivant, *Applied logistic regression*. Wiley New York, 2000.

[74] D. Jurafsky and J. H. Martin, “Sequence labeling for parts of speech and named entities,” *Speech and Language Processing*, 2020.

[75] Sharma, “Activation Functions in Neural Networks,” *https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6*, 2017.

[76] L. Zonglei, W. Jiandong, and Z. Guansheng, “A new method to alarm large scale of flights delay based on machine learning,” in *2008 International Symposium on Knowledge Acquisition and Modeling*, 2008, pp. 589–592.

[77] H. Khaksar and A. Sheikholeslami, “Airline delay prediction by machine learning algorithms,” *Scientia Iranica*, vol. 26, no. 5, pp. 2689–2702, 2019.

[78] Y. H. Liu and M. Ha, “Expected value of function of uncertain variables,” *Journal of uncertain Systems*, vol. 4, no. 3, pp. 181–186, 2010.

[79] S. Ding, “Uncertain multi-product newsboy problem with chance constraint,” *Applied mathematics and computation*, vol. 223, pp. 139–146, 2013.

[80] Y. Gao, “Uncertain models for single facility location problems on networks,” *Applied Mathematical Modelling*, vol. 36, no. 6, pp. 2592–2599, 2012.

[81] X. Zhang and X. Chen, “A new uncertain programming model for project scheduling problem,” *Information: An International Interdisciplinary Journal*, vol. 15, no. 10, pp. 3901–3910, 2012.

[82] M. O’Brien, *Techniques for incorporating expected value constraints into stochastic programs*. stanford university, 2004.

[83] Z. Wang, P. Jochem, and W. Fichtner, “A scenario-based stochastic optimization model for charging scheduling of electric vehicles under uncertainties of vehicle availability and charging demand,” *Journal of Cleaner Production*, vol. 254, p. 119886, 2020.

[84] M. Samà, A. D’Ariano, and D. Pacciarelli, “Rolling horizon approach for aircraft scheduling in the terminal control area of busy airports,” *Procedia-Social and Behavioral Sciences*, vol. 80, pp. 531–552, 2013.

[85] B. Addis, G. Carello, A. Grosso, and E. Tànfani, “Operating room scheduling and rescheduling: a rolling horizon approach,” *Flexible Services and Manufacturing Journal*, vol. 28, no. 1–2, pp. 206–232, 2016.

[86] L. K. Nielsen, L. Kroon, and G. Maróti, “A rolling horizon approach for disruption management of railway rolling stock,” *European Journal of Operational Research*, vol. 220, no. 2, pp. 496–509, 2012.

[87] M. Y. Maknoon, F. Soumis, and P. Baptiste, “An integer programming approach to scheduling the transshipment of products at cross-docks in less-than-truckload industries,” *Computers & Operations Research*, vol. 82, pp. 167–179, 2017.

[88] P. Bishop, A. Hines, and T. Collins, “The current state of scenario development: an overview of techniques,” *foresight*, 2007.

Appendix A: Scientific Paper

Integrating Predictive Model and Optimization Model for Intelligent Schedule Management based on Real Time ETA Information

A Concept of Machine Learning and Exact Solution in Petrochemical Loading Facility

Emanuel Febrianto Prakoso

Delft University of Technology

Abstract

This study addresses the truck rescheduling problem as the consequence of uncertain arrival time. It proposes an integrated system of predictive model powered by machine learning algorithm and exact optimization model such that it is distinct from most existing literatures in this domain. The uncertainty of truck arrival time is captured as presence probability by the developed predictive model. Subsequently, a Mixed-Integer Quadratic Programming (MIQP) is built to solve the Probabilistic Slot Rescheduling Problem (P-SRP) in which the rescheduling is subject to expected value constraint that incorporates the presence probability of incoming trucks. The objective is to minimize the expected cost of rescheduling that would lead to more efficient and robust operation. In regard to the predictive model, evaluation according to the standardized KPI shows the ANN is the best algorithm to fit the input historical data with overall F1 score of 73%. Moreover, adding real-time elements enhances the prediction result by 20%. In regard to optimization model, the P-SRP model results on expected cost that is 42% lower than the P-SRP model. Numerical experiment based on multiple scenarios indicates that the proposed solution yields optimal added values in situation characterized by high number of reschedule being required, large scale of operation in terms of quantity of loading bay and utilization rate, the priority is to maintain initial schedule, and there is no stand-in loading bay available. Lastly, it is found that increasing operational efficiency comes at the expense of the schedule robustness, although the value is negligible. This study is limited to the conceptual setting and the use of synthetic data; therefore it should be extended to include empirical result.

Keywords: intelligent rescheduling, exact solution, probabilistic optimization, machine learning, uncertain arrival.

1. Introduction

1.1.Background

Transportation are one of the most common factors of disruption in the supply chain that contribute to almost one-third of total operational cost [1]. In petrochemical business, efficient and robust downstream supply chain system are important factors to achieve competitive edge [2]. One of the most important actors is the loading facility because of its role as a linkage between production and distribution [3].

In operational level of loading facility, the performance of loading facility is dependent on truck arrival time. The main concern is that, under stochastic environment, the trucks arrive at the loading facility with time uncertainty as the result of complex transport network problem, therefore deviation in forms of early or late arrival is inevitable [4]. Consequently, due to this disruption, operational efficiency of loading facility will be reduced, leading to various severe problems in different perspective. In terms of business, it increases total operation cost. Moreover, it also limits the space, resulting on higher risk of accidents and lower productivity of the site. In environmental perspective, excessive queued trucks emit pollutant gas that could endanger the surrounding livelihood [5].

According to [5], [6], and [7] the standard approach to mitigate the negative impacts of uncertainty in arrival time is implementing Truck Appointment System (TAS). Basically, TAS is defined as a platform for truck companies to request appointment before the actual arrival in order to streamline the arrival flow of incoming trucks. However, basic TAS system is not entirely sufficient at eliminating inefficiency in petrochemical loading facility, especially when rescheduling as reaction of unexpected events is required [8]. However, as observed in [9], literatures in truck rescheduling is limited in which most of the studies disregard the stochastic nature, thus only solve the problem to a limited extent in practice Moreover, rescheduling heavily relies on the information of truck arrival time. To a certain extent, gradual growth in ICT (Information and Communication Technology) that incorporates continuous update and real-time elements could enable more accurate ETA (Estimated Time of Arrival) data [10]. Utilization of this technology results on ability to adjust the schedule ahead of time to anticipate the potential disordering. Nonetheless, it still could not provide a perfect information of the truck arrival time because to many variables that could affect the estimation.

Hence, these shortfalls emphasize not only the urgency of having a better-optimized schedule management in loading facility that considers stochastic variables, also a more advanced system to predict the truck arrival time as the supporting tool in rescheduling decision

1.2.Literature Review

This section provides review of literatures related to the schedule management. There will be two main focuses which are to investigate the precedent researches in regard to prediction of arrival time, and to explore the concepts that are possibly applied to optimize the schedule system. The purpose is to identify the knowledge gaps, therefore the justification of conducting this research and its potential contribution to the existing body of knowledge will be clearly defined.

1.2.1.Predictive Model

To address the issue of inherent uncertainty in transport domain, a predictive model is commonly proposed as the solution. In general, predictive model in the context of predicting arrival time is based on historical data and the real-time data. Boswell & Evans [11] stated that delay can be described as discrete probability function in which the delay propagation. Mueller et al. [12] extends the study to explore type of distribution to best model the dela using probability density function. Exponential growth on transmission of information technology prompted researchers to utilize the real-time ETA as the basis of to derive delay arrival prediction in which all researches proved that real-time ETA provide lesser degree of stochasticity, thus could result on more accurate prediction. Larbi et al. [10] showcased that value of information in comparative study including scenario of no information, partial information, and full information, leading to conclusion that distant information yields small contribution in improving the schedule optimization model.

As the technology rapidly advances, number of studies involving machine learning as a supporting tool to predict delay has exponentially increased, yet most of the studies are done in the domain of airport operation and railway management. Machine learning is deemed powerful to provide accurate prediction in which the concept requires learning process based on historical data [13]. Various popular algorithms, namely RF (Random Forest) [14], [15], k-NN (k-Nearest Neighbor) [16], SVR (Support Vector Regression) [17], [18], and ANN (Artificial Neural Network) [19], [20] could be applied to predict arrival time. However, there is no such generalization to determine the best algorithm because it depends on characteristic of datasets or conditions, thus comparative approach should be taken.

1.2.2.Optimization Model

In general, optimization model has been extensively used to address the problem in scheduling system. In earlier period, most of the studies [21], [22], and [23] rely on queuing theory as the tool to model the truck scheduling. However, Huiyun et al. [9] claimed that queuing model is not sufficient in providing improvement in rescheduling system because the concept is only able to analyze the appointment data yet could not offer proper decision-making assistance in the scheduling system. Furthermore, most of recent researches on truck schedule management utilized MIP (Mixed Integer Programming) model, whereby this approach allows better quantitative evaluation.

In the domain of truck scheduling in loading facility, various studies of deterministic MILP model have been conducted with different objectives. There are researches that constructed model with single objective, namely, to minimize emission cost prompted from idle trucks [24], to minimize penalty cost due to delay in arrival time [27], and to minimize turn over time of trucks [28], and to produce truck schedule that maximize slot efficiency considering sufficient service for all trucks [27][5], [28]. In all those studies, regardless the variation of measurement indicator being applied, the optimization model is potential approach that could enable better operation. Since in reality most of the problems are not constituted by a single factor, more recent researches incorporated multi-objective [8] or generalized cost [29] to provide more valid result. The majority of studies in the specific domain of truck scheduling system assumed that arrival time is deterministic variable, therefore it only solves the real problems to a limited extent. It is indeed necessary to explore the study that incorporates the arrival time as stochastic variable.

Most of studies build stochastic optimization model based on the concept of mathematical programming, but the method used in integrating the stochastic elements are varied. There are two main SP (Stochastic Programming) approaches that are suitable and have been normally used in the study of scheduling system, which are two-stage stochastic and probability constraints [30]. Therefore the literature review on the application of stochastic model focus on those specific approaches. Since most of the schedule management models are based on the deterministic approach, parts of model have to be adapted when encounter stochastic variables. Several studies assumed stochastic elements as random variables. Seker & Noyan [31] incorporated disruptions such as delay in arrival time, idle time, and buffer time as random variables which is integrated in the constraints of MILP optimization model. The objectives quantified total conflicting slots and idle slots as robustness measures.

Stochastic model that directly considers the uncertainty variables as discrete probabilistic distribution is also an alternative approach in dealing with stochastic variables. In railway scheduling system research conducted by Meng & Zhou [32], many attributes associated with high degree of uncertainty, such as travel time, arrival time, and departure time are considered as stochastic variable, thus individual probability distribution is assigned. However, since in the distributions used in these experiments are not the empirical distributions from real statistical data over certain periods, the benefit of this model is limited.

Furthermore, Visser & van Schaijk [33] transformed deterministic slot assignment method to be able to consider stochastic elements by replacing the binary scheduled arrival time constraint with presence probability to capture the inherent stochastic delay. This approach would assign arrival to a certain location according to the maximum permissible overlap probability. Higher permissible value would result in a more compact schedule, therefore lower assignment cost. However, this research does not incorporate specific assignment, in which arrival can be universally assign to any available slots. This approach could not be directly applied here because, in this research, since specific product type could only be loaded in a certain slot, hence it rises the complexity of the problem.

1.3.Objective and Contribution

The literature reviews have pointed out that most optimization models to solve truck rescheduling problem are designed in a deterministic fashion. Few papers have addressed the uncertain arrival time by applying exact method, yet to author's best knowledge, none has been done in the context of truck rescheduling system. Moreover, almost of all the researches in the domain of truck rescheduling have not incorporated predictive model powered by ML algorithm to capture the uncertainty of arrival time. Since this approach proved beneficial in similar cases in other domain, applying it in this research might be a promising direction.

As a first research gap, it is interesting to see the extent to which an exact optimization model that considers stochastic variable could increase the performance of loading facility with respect to operational efficiency Furthermore, as studies in the truck schedule management domain are limited to the use of simple heuristic to simulate the deviation in arrival time, as the second gap, it is compelling to explore the feasibility of utilizing ML algorithm as supporting tool to predict truck arrival time.

The first objective of this research is to develop more intelligent rescheduling system for petrochemical loading facility that could mitigate uncertain arrival time, and therefore would ensure higher operational efficiency. Moreover, since this research focuses on utilizing real-time ETA information to enable rescheduling ahead of time, the second objective is to observe the added value of the technology in operation of loading facility. The result will be evaluated relative to the current situation where the real-time information is not available. Lastly, the third objective is to observe the added value of integrating a predictive model and optimization model that considers stochastic variable. The result will be assessed relative to the baseline approach where the predictive model is not utilized, and the optimization model assumes all values as deterministic.

A loading facility owned and operated by [redacted] located in Antwerp is chosen as the use-case sample of this study, although the outcome is not intended as a tailor-fit solution for their particular problems.

2. Intelligent Rescheduling System

In a typical petrochemical loading facility owned and operated by [redacted] in Antwerp, the main flow of loading process can be described according to [8]. There are two main components that can be considered as a series of queue system which are the parking area and plant area. The purpose of parking area is where the incoming trucks will be waiting until their scheduled loading slot is available. In between the parking area and facility area, there is a gate that acts as regulator in determining which trucks to enter the facility area. A decision-maker controls the inflow of trucks into the facility area based on the initial schedule or the adjusted schedule as the result of stochastic arrival. Upon checking process, the trucks proceeds to the loading bays inside the facility area the where the trucks will queue according to its assigned loading bay.

In order to acknowledge the rescheduling problem due to the uncertain truck arrival time, an intelligent rescheduling system that could provide more operational efficiency and robustness against real-time disturbance is introduced as main solution. The general flow of the system is illustrated in Figure 2.1. As it is mentioned, it considers stochastic nature in transportation represented by uncertain arrival time. This stochastic variable is captured as probability value that will be generated by the predictive model that learns from historical data. Continuously, the real-time ETA data will be inputted to the predictive model that will assign a probability of whether a truck would be early, on-time, or delay. Based don that information, the adjusted schedule could be optimized.

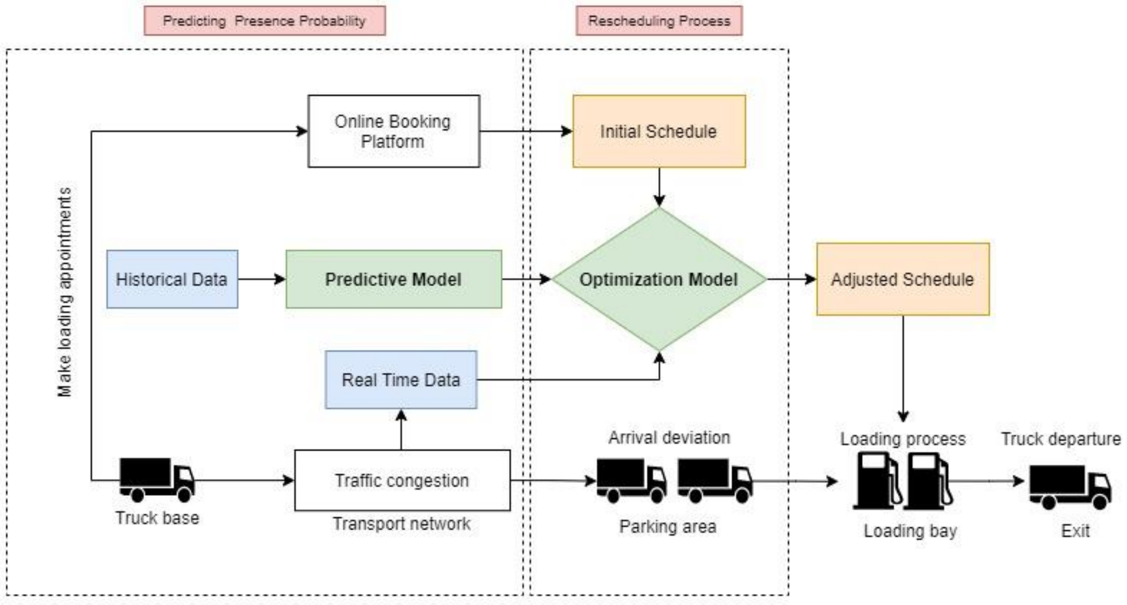


Figure 2.1 General Flow of Intelligent Rescheduling System

The core idea of this framework is to input the real-time ETA information into the predictive model that would generate the probability of a certain truck being early, on-time, or late considering the trend/pattern of historical data learnt by the ML algorithm. Subsequently, the result of the predictive model would be inputted into the optimization model that would generate the adjusted schedule. The general flowchart of this process is detailed in Figure 2.2.

In this case arrival class relative to the initial schedule is used as reference point (checking step). This is done to determine whether rescheduling is triggered. The arrival class is varied from 0, 1, and 2 in which the former figure means that the truck would arrive early (at minimum 10 minutes early), the middle figure means that the truck would arrive on time (between 10 minutes early and 5 minutes late), and the latter figure means that the truck would arrive late (at minimum 5 minutes late). According to that classification, it defined that the class 0 and class 2 would require rescheduling, whereas the class would stick to the initial schedule.

With respect to the defined classes, the predictive model could assign probability value for each individual truck that indicates how likely a truck belongs to a certain class. This value is subsequently perceived as the presence probability of truck at a certain timeslot. By doing this, the uncertainty embedded in the real-time ETA information is acknowledged and captured. As the final result, a pair of ETA information and its corresponding presence probability value is prepared to be inputted into the optimization model (furtherly explained in Section **Error! Reference source not found.**)

Furthermore, in regard to the implementation period, the set of real-time ETA information will be updated based on the defined rolling horizon. In this research, it is assumed that the rolling horizon is 30 minutes, therefore there will be new set of real-time ETA information for every 30 minutes. For each period, the predictive model that has been trained using the historical data would use the newly updated ETA information including all the features as the input data to predict the probability value of a particular truck to be present according to its ETA information.

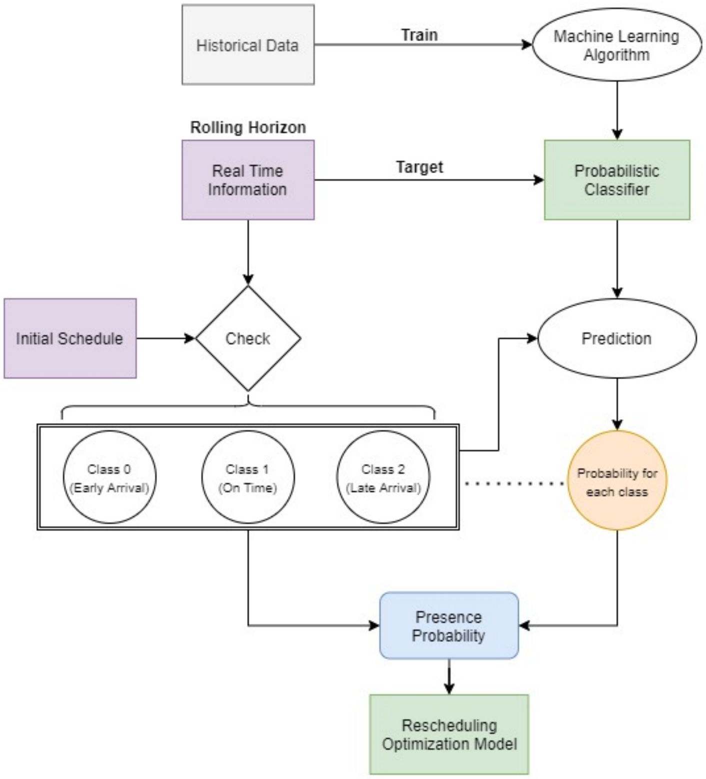


Figure 2.2 Conceptual Framework of integration between predictive model and optimization model

3. Data

Previously, a conceptual framework of scheduling system that aims to improve the operational efficiency of loading facility has been proposed This chapter provides sequential elaboration on input dataset used in this research. It consists of data generation phase where a historical dataset is synthetically generated, and data exploration that includes IDA (Initial Data Analysis) and EDA (Exploratory Data Analysis) where the dataset is pre-processed and analyzed using varied method. Hence, the purpose of this section is not only to deliver a hypothetical valuable insight derived from historical data, also to ensure that the final data has decent quality to be inputted into both predictive model and optimization model.

For the purpose of verifying the proposed method, a synthetic data is generated from the publicly available database because the real data cannot be obtained. Since data that specifically fit the context of this research, namely the truck scheduling in loading operation could not be found, it is decided to adapt and manipulate the closest data resembling the problem of this research, which is the historical data of airline on-time performance in US. The data is publicly accessible and can be downloaded in [36].

3.1.Data Description

The original synthetic datasets contains arrival information of trucks in the period from 1 January 2016 until 31 March 2021. Since only 1 loading facility is being considered, there is only 1 single location for all the trips that represents the loading facility itself, therefore the full dataset is reduce to 141517 rows datapoints representing trip of trucks. Furthermore, each row contains information or features of the operational attributes. The full list of all features are presented in Table 3.1.

Table 3.1 Features in Historical Data

Features	Description	Specification
Product Type	A pre-defined product type which each truck supposedly loads at the loading facility	The truck directly corresponds to its pre-defined product type. Complete product type is provided in Appendix B.
Origin	The location that trucks depart from.	The set of origin is composed of 20 different locations. Complete list of origin name is provided in Appendix B
Destination	The location of loading facility.	In this case, loading facility is only one and indicated by 'FACILITY'
Distance group	The distance category between various origin to the loading facility.	The categories are based on a specific range of distance between origin and destination. The detail is shown in Appendix B
Year, Month, Day of Week, Date	Time indicator of the corresponding schedule.	• Year ranges from 2016-2020 • Month ranges from 1-12 • Day of Week ranges from Monday to Sunday • Date ranges from 1 till 30/31 • Hour ranges 00:00 till 23:59
Scheduled Departure Time	Initial schedule of departure from origin	This feature is formatted in hour and minutes
Scheduled Arrival Time	Actual realization of departure from origin and arrival to destination.	This feature is formatted in hour and minutes
Arrival Delay	Deviation between scheduled and actual for both arrival and departure.	This feature is in minutes basis • Early arrival denotes by negative value • On-time arrival denotes by 0 • Late arrival denotes by positive value
Elapsed Time	Realization of travel time from a certain origin location at a certain time of the day/week	This feature is formatted in minute

3.2.Data Exploration

The IDA (Initial Data Analysis) consisting of cleaning, transforming and filtering is conducted to ensure that historical data is statistically sufficient. Since the records consist of full year of arrival data that ranges from 2016-2020 , it results on extremely large dataset. It is decided to cut the data down to only contain the flight records of 1 month period for each year. In practice, the month of February is chosen, then arrival records of February for each year (2016, 2017, 2018, 2019, 2020) is extracted. The combination of this subset of data results on 90198 rows of data points. Other justification in limiting to a certain month period is to reflect a recurring trend in monthly basis which is deemed sufficient for a short-term planning. To mimic the situation of having a single loading facility in a certain location, the destination data is also set to only include a certain destination. This action significantly reduces the size of the data because this initial data include all trips between 20 different origin and destination airports. Lastly, the data is normalized by filtering out the statistic outlier such as the extreme late arrival time that is defined as actual arrival time having deviation value higher than 120 minutes and the extreme early arrival time that is defined as actual arrival time having deviation value lower than -15 minutes.

Furthermore, EDA (Exploratory Data Analysis) is conducted to profoundly analyze and investigate the underlying characteristics, such as pattern, trend, and anomalies of the large dataset beyond the obvious appearance and formal statistic method. It is designed to emphasize on the arrival time variability based on different parameters. With respect to time of the day, the results show that late arrival is likely to occur in the afternoon, while the early arrival is evenly distributed across the operational hours. In regard to time of the week, the median delay is almost the same for all days. Furthermore, comparing delay in monthly basis illustrates a similar pattern for each month in a year. In terms of product type and truck origin location, it is impossible to generalize the pattern since they are highly random. Ultimately, the results proves that the synthetic historical data is valid with respect to alignment to existing theories and logical sense of real word.

4. Predictive Model

4.1.Pre-Processing

Pre-processing steps are required to assure that the data is adequate and suitable to be inputted into ML algorithm. The pre-processing are applied to historical datasets with and without the real-time ETA information. Firstly, feature selection is conducted to remove multicollinearity variables that could renders ML model inefficient. In order to do so, a Pearson correlation method is utilized. Secondly, since ML algorithm is unable to process string or date-time data type, therefore all values in dataset must be converted into numerical value. In other words, feature encoding process is required . This is done using a built-in Python library called Label Encoder, whereby features that are in form of string are consistently converted into number values. Lastly, due to the lack of ability to deal with features that vary in magnitudes, the dataset is scaled. When feature scaling is not done, there is a tendency of ML algorithm to perceive greater values higher than smaller values, regardless of the unit of values. To solve this problem, standardization that transforms the data to ensure mean value of 0 and the standard deviation of 1 is applied.

4.2.Model Training and Validation

The proposes solution is designed to trigger rescheduling based on the arrival class which are early, on-time, and late arrival. In order to realize this concept, new feature called “Delay Level” is created as the target/output of the predictive model. In this case, the terms of being present could possibly be denoted by the arrival class of either “0” and “1”, and “2” in which the Class 0 means that the truck would arrive earlier, the Class 1 means that the truck would arrive on time, and the Class 2 means that the truck would arrive late. The definition and requirement to be classified as certain class is defined in Table 4.1.

Table 4.1 Requirement for Arrival Classification

Class	Requirement
0 = Early Arrival	ETA indicates that the truck would arrive more than 30 minutes early than initial schedule. $(\boldsymbol{Actual - \ Scheduled < \ -30})$
1 = On-Time	ETA indicates that the truck would arrive less than 30 minutes early up to the start initial schedule. $(\boldsymbol{-30 \leq \ Actual - \ Scheduled \leq 0})$
2 = Late Arrival	ETA indicates that the truck would arrive late than the start of initial schedule. $(\boldsymbol{Actual - \ Scheduled > \ \ 0})$

Since there is no clear guideline in determining which ML algorithms would provide the best result. It is decided to explore 3 different algorithms, namely GBN, LR, and ANN. To increase the efficiency of the predictive model, the features are selected according to their correlation value to minimizes the negative impacts caused by multicollinearity problem. The evaluation results according to the standardized KPI shows the ANN is the best algorithm to fit the input historical data with Overall F1 score of 73%. Relative to other tested models, the ANN model outperforms the LR model and GBN model by averagely 5% and 20%, respectively, in all the KPIs metrics.

To benchmark the result of the predictive models, a comparison with other predictive models found in similar studies is conducted. The ANN’s accuracy score of 0.70 is sufficient compared to predictive model proposed in [16], [18]–[20], [35]–[40] whose accuracy is approximately ranges from 0.4 - 0.8. Considering the complexity level, the proposed predictive model even outperforms the ones in which the target classes are binary. However, this could not be stated with absolute certainty because the results could be affected by many factors, such as data quality, context, computational power, etc. Despite of the potential flaws , it still can be concluded that predictive model proposed in this research is acceptable.

4.3.Sensitivity Analysis

Two different scenarios to that differs in features combination is developed. The first only includes the feature from historical data, while the second includes the real-time information enabled by ICT implementation. The feature combination is detailed in Table 4.2.

Table 4.2 Description of Scenarios to Investigate the Added Value of Real Time Information

	WITHOUT ETA INFORMATION	WITH ETA INFORMATION
FEATURES	• Historical: Truck, Origin, Day Name, Scheduled Departure Time, Scheduled Arrival, Scheduled Travel Time, Distance.	• Historical: Truck, Origin, Day Name, Scheduled Departure Time, Scheduled Arrival, Scheduled Travel Time, Distance. • Real-Time: Actual Departure, Estimated Arrival Time, Estimated Elapsed Time.

Given the set of scenarios a comparative study is conducted to investigate the added value of real-time ETA information in improving the predictive model. For each defined KPI, a macro average is calculated. The evaluation result of predictive models under different scenarios are presented in **Error! Reference source not found.**.

Table 4.3 Result of Comparison Between Scenario of Without and With Real Time Information

		WITHOUT REAL-TIME INFORMATION	WITH REAL-TIME INFORMATION
GNB	Accuracy	0.36	0.51
	Recall	0.41	0.55
	Precision	0.38	0.57
	F1	**0.35**	**0.52**
LR	Accuracy	0.45	0.64
	Recall	0.48	0.67
	Precision	0.46	0.65
	F1	**0.44**	**0.65**
ANN	Accuracy	0.5	0.70
	Recall	0.51	0.71
	Precision	0.5	0.72
	F1	**0.51**	**0.70**

Based on the result of the sensitivity analysis, the ANN algorithm trained using the historical dataset that contains real-time information is best performing predictive model. Importantly, the comparison indicates a significant improvement across all ML algorithms compared if the real-time information is incorporated, to the ones trained based on dataset without real-time information. The real-time features enhances the prediction results by approximately 20% increase in all metric of defined KPI. Therefore, the real-time information certainly provides an essential added value to the predictive model.

Another key finding derived from the sensitivity analysis is that the inclusion of feature having a larger than 0.5 correlation value (‘Departure Delay’) with the prediction target (‘Arrival Delay’) in the learning process is important to get the potentially best result. Hence, it is important to note that in building a predictive model, all features with significant correlation must be selected.

5. Optimization Model

5.1.Problem Description

In regard to the optimization model, there are several problems that are likely to emerge during the described process in Section **Error! Reference source not found.**. Firstly, the scheduling concept is based on the notion that the trucks would have arrived in the facility according to the initial schedule based on their preferred/requested time. However, in practice, the trucks arrive stochastically due to inevitable cause associated with arrival time. Secondly, the parking area is supposed to accommodate all incoming trucks in the planning horizon, but in reality, the parking space is limited, and the excess trucks would load the side of access road, thus they would cause detrimental impact to the transport network in a form of congestion and to the livelihood around the facility as the queued truck would contribute to high level of emission gas. Thirdly, this assignment of loading bay is not interchangeable because the trucks are ordered to load a certain product that correspond to a certain loading bay. Consequently, problem also potentially occurs in the plan area in case where the assigned loading bay is being occupied by other trucks, then the truck must wait until the loading bay comes available.

Hypothetically, a plot that maps where and when a certain truck being assigned according to their arrival time is illustrated in Figure 5.1. Accordingly, several overlapped assignments are occurred, indicated by orange circles. The result of this conflicts is that the conflicted trucks are either to wait indefinitely until there is available slots that fulfill its requirement or returning to the base and then getting the slot appointment days later.

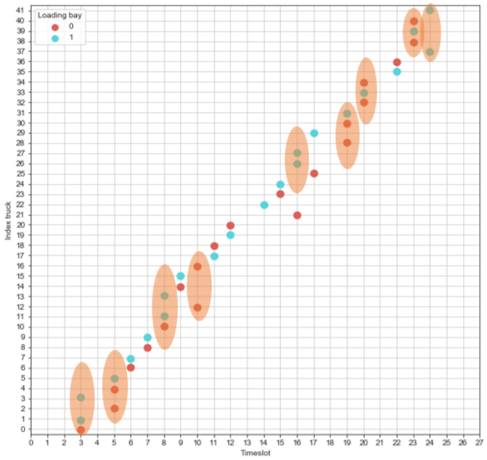


Figure 5.1 Hypothetical Conflicts in Current Situation

5.2.Mathematical Model

Some assumptions are made to acknowledge the data limitation. The rescheduling cost is constituted of the difference between the adjusted schedule and initial schedule multiplied by generalized cost which is assumed 1 monetary value/slot. All the other costs, namely the CO2 emission cost, space rent cost, site operation cost, etc. are omitted. The space in facility is adequate to accommodate all scheduled trucks within the planning horizon. The serviceability of loading bays are constantly available over the planning horizon. In other words, disruption such as machine breakdown or maintenance are not considered. The loading/service time is constant for all the timeslots in the planning horizon which is set to be 15 minutes per truck. Consequently, the departure time of truck could also be certainly calculated, thus the departure time from loading facility is deterministic. Rescheduling could possibly be triggered only for truck that is classified as Class 0 (early arrival and Class 2 (late arrival).

Table 5.1 Set of Variables.

Notation	Definition
T	Set of time slots within time horizon
J	Set of scheduled trucks within time horizon
L	Set of loading bays at the chemical facility
P	Set of product type
$\hat{J} \subseteq J$	Set of delayed trucks
$\bar{J} \subseteq J$	Set of on-time trucks

Table 5.2 Set of Parameters.

Notation	Definition
Y_j	Minimum time difference between new and initial schedule of truck j
g_{jl}	Generalized cost in assigning truck j to loading bay l
s_{jt}	Presence indicator of trucks, the value is 1 if the truck j has non-zero presence chance in time slot t, 0 otherwise
r_{lj}	Assignment of certain product type, the value is 1 if certain product type meant for truck j can be loaded at loading bay l, 0 otherwise
c_{jl}^{t}	Current schedule of trucks, the value is 1 if truck j is scheduled to loading bay l at time slot t, 0 otherwise
x_{jl}^{t}	$\begin{cases} 1, & if\ truck\ j\ is\ reassigned\ to\ loading\ by\ l\ in\ time\ slot\ t \\ 0, & otherwise \end{cases}$

The formulation of the mathematical model is as follows:

$$\text{Min} \sum_{t} \sum_{l} \sum_{j} \left(x_{jl}^{t}\, t\ - c_{jl}^{t}\, t\right)^{2}\ \ g_{jl} \tag{1}$$

Subject to:

$$\sum_{l\, \in\, L} x_{jl}^{t}\, s_{jt} = s_{jt} \qquad ,\forall\, j \in J\, , t \in T \tag{2}$$

$$\sum_{j\, \in J} x_{jl}^{t}\, f\left(p_{jt}, r\right)\, p_{jt}\ \le 1 \qquad ,\forall\, l \in L\, , t \in T \tag{3}$$

$$\sum_{j\, \in J} x_{jl}^{t}\ \ \le 3 \qquad ,\forall\, t \in T, l \in L \tag{4}$$

$$x_{jl}^{t}\ + c_{jl}^{t}\ \le 2 \qquad ,\forall\, j \in\ \hat{J}\, , t \in T, , l \in L \tag{5}$$

$$x_{jl}^{t}\ \ \le\ r_{lj} \qquad ,\forall\ j \in\ J\, , t \in T, l \in L \tag{6}$$

$$x_{jl}^{t}\, t - c_{jl}^{t}\, t \ge Y_j \qquad ,\forall\ j \in\ J\, , t \in T, l \in L \tag{7}$$

$$x_{jl}^{t}\ \ \in \{0,1\} \qquad ,\forall\ j \in\ J\, , t \in T, l \in L \tag{8}$$

Common approach of SAP (Slot Assignment Problem) is adapted to model the rescheduling problem. This model is then referred as SRP (Slot Rescheduling Problem). This chosen model is inspired from literatures in similar problem [41] with appropriate modifications to fit the context of this study case. In the deterministic sense, the presence of particular truck in a certain time slot is assumed to be fully known and the delay probability is not taken into account. To capture the stochastic nature of arrival time, the model needs to be adapted. As it is proposed by [33], the idea of 'presence probability' is incorporated into the deterministic SRP model. This modified model is further referred as P-SRP (Probabilistic Slot Rescheduling Problem). Instead of using the binary constraint to define the truck presence, the probability of truck being present at a certain timeslot are used as replacement. The changes represent the idea of allowing multiple trucks to be scheduled in a certain time slot as long as the overlapping probability does not violate the permissible threshold value.

The objective function (1) is to minimize the total expected rescheduling cost of the operation at loading facility. Technically, the expected cost is the accumulation of rescheduling cost of all trucks in all loading slots within the planning horizon. As it is explained, the expected cost equals to the difference between adjusted and initial scheduled multiplied by generalized cost which is 1 monetary value/slot. As deviation between initial and adjusted slot which could be either in negative (in case if truck being moved to earlier slots) or positive value (in case if truck being moved to later slots), therefore a quadratic cost function is chosen because it perfectly fits the context, instead of linear form that would prefer rescheduling to earlier slots than adhering initial schedule. This model is now categorized as MIQP (Mixed Integer Quadratic Programming) problem

Constraint (2) ensures that all trucks are assigned exactly once at a certain time horizon. Constraint (3) allows simultaneous assignment of trucks in a certain slot based on the presence probability and is limited to smaller or equal to 1 because it would serve as threshold indicating if there is a truck having presence probability of 1, then it does not make sense to simultaneously assign another truck to that slot regardless its corresponding probability. Constraint (4) ensures maximum number of trucks that can be simultaneously assigned to measure the risk. Constraint (5) enforces different new slot for delayed trucks while allowing same slot for on-time truck. Constraint (6) ensures that trucks are loaded with its predetermined product type. Constraint (7) ensures satisfaction of minimum time difference between new and initial slot. Constraint (8) determines the domain of decision variable.

5.3.Verification

The proposed rescheduling system is completely applied and verified on a synthetic use-case sample. It is clear that the utilization of ETA yields a significant advantage in terms of rescheduling process, whereby it allows rescheduling to be done ahead of time. In order to measure the benefit of the P-SRP model, a baseline model that considers ETA as a deterministic variable is introduced for comparison (Figure 5.2 and Figure 5.3). The main feature of the P-SRP is a more intelligent rescheduling, in which it allows simultaneous assignment based on the value of presence probability. The results clearly highlight the shortcoming of the baseline model in terms of preventing occurrence of idle slots and unnecessary schedule shift. Consequently, the P-SRP results on expected cost that is 42% lower than the baseline model.

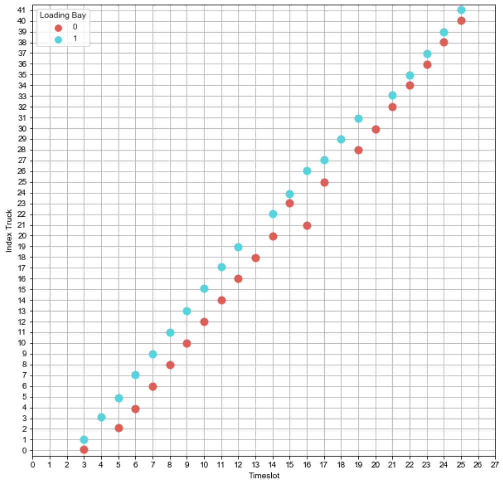


Figure 5.2 Rescheduling Result of Baseline Model

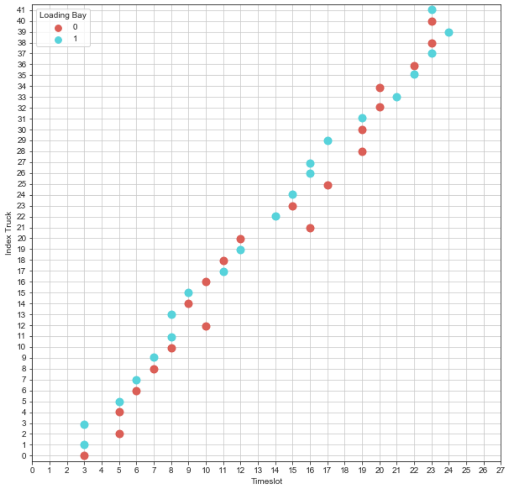


Figure 5.3 Rescheduling Result of P-SRP

Lastly, the proposed rescheduling system is tested on dynamic experiment based on rolling horizon to simulate the actual implementation and continuous update of real-time ETA data. Due to the limitation in this research, only 2 iterations with period of 2 hours are simulated. The results showcased the superiority of P-SRP model compared to the baseline. On top of the higher operational efficiency, the analysis indicates that P-SRP provides robustness against uncertain arrival time, whereby the adjusted schedule could withstand and anticipate the fluctuation of actual arrival of incoming truck. Hence, it can be concluded that the main advantage of utilization of predictive model and P-SRP proposed in this research is a more intelligent rescheduling process, in which operational efficiency is improved without radically compromising schedule robustness. However, the results might be case-specific considering the limited data and experiments.

6. Numerical Experiment

6.1.Scenario Development

Set of scenarios are developed to analyze the performance of the P-SRP relative to the baseline model. Each scenario correspond to different contexts relevant to the operation of loading facility, namely congestion level in transport network, specification of loading infrastructure, scheduling policy, and the maximum permissible probability. The justification of having varied contexts of scenarios is because it would provide a more profound analysis to generalize on which conditions the P-SRP would generate maximum benefits.

6.1.1.Scenario 1: Congestion Level

Varied delay ratios are assigned to each scenario of congestion level, namely mild, medium, and severe, respectively, as seen in Table 6.1. A simulation that is based on triangular probability is run to generate the set of arrival time according to a scenario

Table 6.1 Scenario in regard to congestion level

	Mild	**Medium**	**Severe**
Congestion Level	25 %	40 %	70 %

6.1.2. Scenario 2: Loading Infrastructure

This scenario focuses on the variation in regard to configuration/specification of a loading infrastructure. Different number of loading bay is applied to denote the scale of operation. On top of that, other parameters that could be associated with this context is utilization rate that could indirectly indicate the amount of buffer time. As the scale of operation in a loading facility is proportionate to the complexity of schedule management, thus it is relevant to analyze how well the P-SRP would solve the increase of magnitude in operation of loading facility. The full variation of scenarios related to the loading infrastructure is shown in Table 6.2.

Table 6.2 Scenario in regard to loading infrastructure

	Small scale			**Medium scale**			**Large scale**		
Loading Bay	2			3			4		
Utilization Rate	50%	75%	100%	50%	75%	100%	50%	75%	100%

6.1.3.Scenario 3: Rescheduling Strategy

One most popular priority in rescheduling is to minimize the waiting time between actual arrival time and adjusted schedule. In addition, other possible strategy to mitigate conflict probability is by providing a stand-in loading bay which main purpose is to be used as an alternative option if the other loading bays are fully occupied. The full variation of scenario related to the rescheduling strategy is shown in Table 6.3.

Table 6.3 Scenario in regard to rescheduling policy

Priority	**Minimizing waiting time**		**Maintaining initial schedule**	
	$\textbf{Min} \sum_t \sum_l \sum_j (x_{jl}^t\, t - c_{jl}^t\, t)\; g_{jl}$		$\textbf{Min} \sum_t \sum_l \sum_j (x_{jl}^t\, t - c_{jl}^t\, t)^2\; g_{jl}$	
Stand-in Loading Bay	Without (3 Specific)	With (3 Specific + 1 General)	Without (3 Specific)	With (3 Specific + 1 General)

6.2.Added Values of P-SRP

A baseline model that will be used as the benchmark to analyze the added values of the P-SRP is introduced. The baseline model refers to the scheduling system of loading facility that considers the real-time ETA information as deterministic variable. In other words, the initial schedule will be according to fully known and certain arrival information. Consequently, to do so, the P-SRP model slightly modified by removing the expected value of probabilistic variable in Constraint (3). Further on, this model will be referred as Baseline model.

In the scenario of varied congestion level, the comparison reveals that the P-SRP outperforms the baseline model in all KPIs evaluation. The P-SRP offers a rather equal improvements in terms of expected cost regardless the congestion level, whereby it reduces the expected cost by averagely 52.07% compared to the baseline model. With respect to the total rescheduled trucks, the P-SRP requires rescheduling roughly 45.7% less than baseline model, yet in the severe congestion level, the value decreases to only 28%. In preventing idled slots, the P-SRP offers decreasing the amount by 66.67% lower than the baseline model compared to only 40% and 50% in mild and severe congestion level, respectively. Hence, it can be concluded that the probabilistic has most added values when being implemented in the situation of medium congestion level

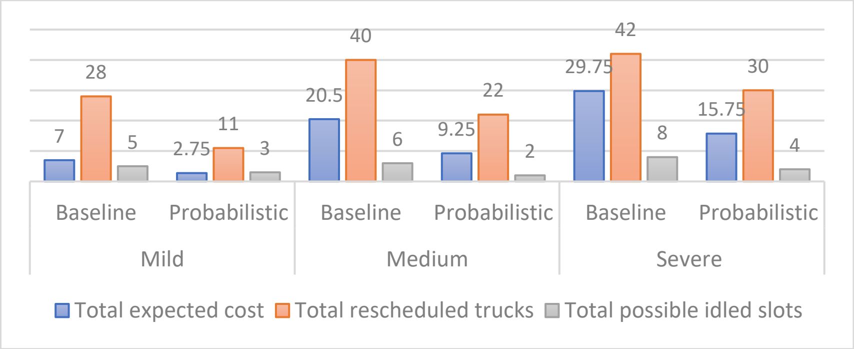


Figure 6.1 Result of Varied Congestion Level

In the scenario of varied specification/configuration of loading infrastructures, in general, the P-SRP has significant edges compared to the baseline model. With focus on the scale of operation, the P-SRP provides averagely 54.84% lower expected cost compared to baseline model in condition where there are 4 loading bays, whereas the improvement reduces to only 29.03% in condition where there are only 2 loading bays. Therefore, the result proves that P-SRP is promising to be implemented in large-scale operation. With respect to the utilization rate, lower utilization rate reduces the performance gap between probabilistic and baseline model, whereby in situation of 100% utilization rate, it provides averagely 46.24% lower expected cost, whereas the improvement only accounts for 11.02% and 20.29% in situation of 50% and 75% utilization rate, respectively. Hence, it can be derived that the P-SRP provides most added value in large scale of operation with more intense utilization rate of loading slots.

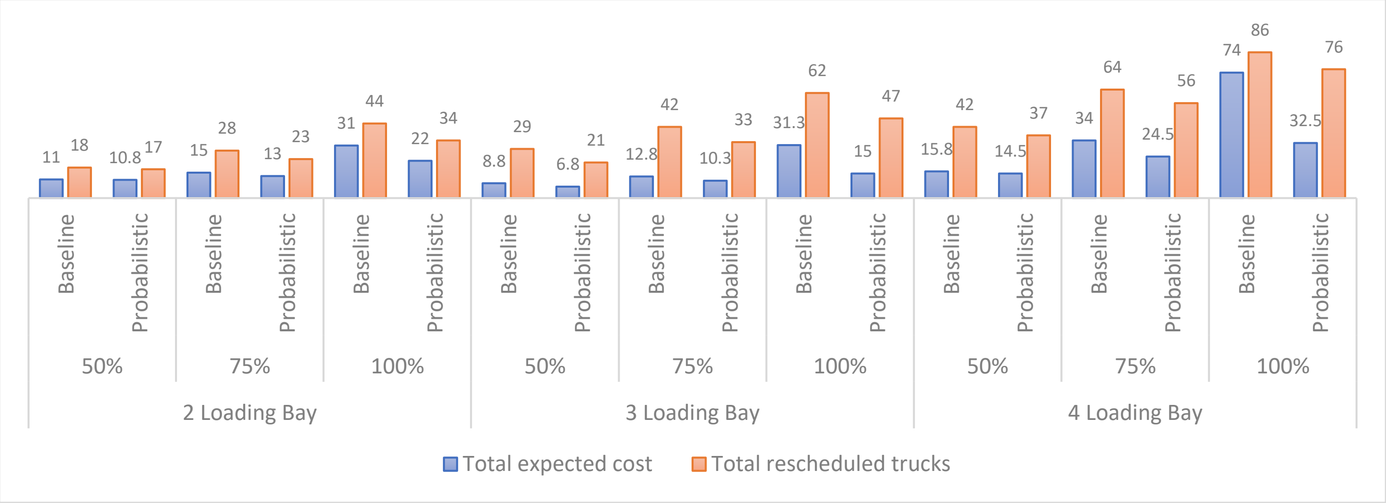


Figure 6.2 Result of Varied Specification of Loading Infrastructure

In the scenario of varied rescheduling strategy, the focus is to explore the impact of other rescheduling priority and to assess the merit of having a stand-in loading bay as an alternative loading. One interesting point is that in situation where the priority is minimizing waiting time and there is a stand-in loading bay, the P-SRP offers an insignificant improvement, only 15.22% lower than the baseline model, whereas in other combined situations the value reaches 51.39%. Moreover, it is observed that the stand-in loading bay is only utilized if the priority is to minimize the waiting time, otherwise, there is no substantial impact in having stand-in loading bay. Thus, it can be inferred that probabilistic would be optimal when the priority is to maintain the initial schedule and there is no stand-in loading bay available.

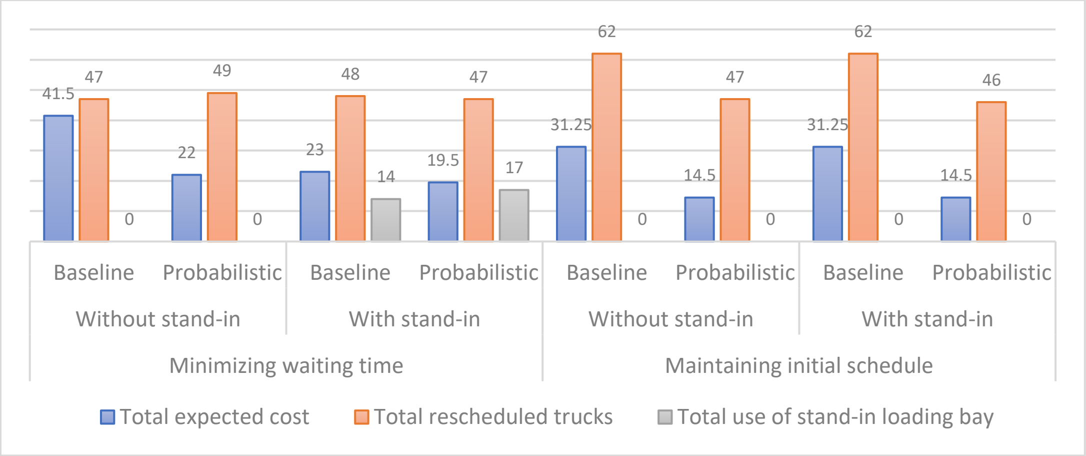


Figure 6.3 Result of Varied Rescheduling Strategy

6.3.Trade-off between Efficiency and Robustness

Ideally, robustness can be measured by observing the conflicted loading slots that occur in practice. More conflicted slots would require more rescheduling process, which results on lack of schedule robustness. However, since real observation is not possible in this research, then alternative approach based on probability value is utilized. So, this KPI is defined as the total number of loading slot in which several trucks are likely to be present at the exact same time and loading bay. In this case, it is assumed that trucks having presence probability higher than 0.75 is considered as absolute present, Hence, getting another/more trucks simultaneously assigned to the exact same loading slot would cause a conflicted slot.

Another parameter that can be explored to find the most ideal value is the maximum permissible overlap probability. This variable is basically the main factor in determining whether a truck could be simultaneously assigned in same timeslot and same loading bay. Furthermore, a set of scenarios containing varied value of the maximum permissible overlap probability is constructed. The purpose of this comparison is to see how much an increment in maximum value would impact the result of the optimization model. Hence, decision-makers could have insight whether it is worth implementing higher bound in overlap probability. To realize the idea, different amount of maximum overlap probability, namely 5%, 20%, 35%, and 50% are applied. To get a more variability on sample set, the diverse congestion levels , namely mild, medium, and severe (Scenario 1) are also added.

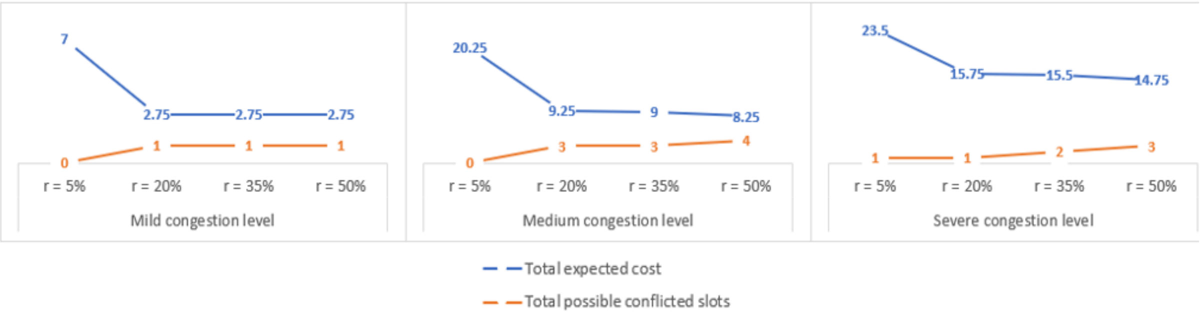


Figure 6.4 Trade-off between robustness and efficiency with respect to varied congestion level

The most important insight that can be derived is that in all scenarios of varied congestion level, the most significant increment in the efficiency occurs when the r value is shifted from 5% to 20%, and after that the improvement is quite negligible. To be more specific, in the scenario of mild congestion level, the shift from 5% to 20% in r value provides a 60% improvement in the efficiency, yet after that, the improvement is constant. Moreover, in the scenario of medium congestion level, raising 5% of r value to 20% of r value yields 52% higher efficiency, while after that increasing the r value only bears averagely 2.5% improvement in efficiency metric. Lastly, in the case of severe congestion level, given the same fashion, the first improvement is in the high of 32%, whereas after that, set of higher r values only yield averagely 3% increase in efficiency metric.

The total number of possible conflicted slots could be used as indicator to measure how often a rescheduling is required; hence the schedule is not robust. With respect to the variation of permissible overlap probability, the results indicates that higher r values generated a higher number of conflicted slots, even though the rate of change is insignificant. It bears a logical sense, because if the model allows a higher overlap probability, then more conflicts should be expected. Hence it can be inferred that improvement in efficiency at the loading operation does come at the expense of schedule robustness., however, in this case the rate of change and the proportion of conflict is negligible. To put into perspective, when applying r value of 20%, the conflicted loading slots is roughly 5 of total slots which can be considered as relatively negligible.

7. Conclusion

This study shows that it is conceptually possible to improve the operational efficiency of a petrochemical loading facility by having an integrated system between logistic operation and traffic system. It proposes an advanced schedule management that allows intelligent rescheduling. The main elements of the rescheduling system are the predictive and optimization model that integrates the real-time information and historical data of truck arrival. Uncertainty is captured by the presence probability of a truck at a certain timeslot generated by a ML model. Rescheduling strategy is adapted as reactive action to mitigate detrimental impact of arrival deviation by developing MIQP model (P-SRP). The proposed system is designed to work under a rolling horizon in which it significantly outperforms the baseline in terms of lower expected cost of rescheduling, less possibility of idled slots, less unnecessary schedule shift and higher robustness when being compared to actual arrival. Furthermore, based on numerical experiments, the proposed system yields optimal added values in situation characterized by high number of reschedule being required, large scale of operation in terms of loading bay and utilization rate, the priority is to maintain initial schedule, and there is no stand-in loading bay available. Lastly, as operational efficiency and schedule robustness works in trade-off manner, P-SRP model offers advantage to balance them by altering the maximum value of permissible overlap. It is found that shifting the value from 5% to 20% yields most significant improvement on the operational efficiency with negligible conflicted slot. However, the most optimal trade-off cannot be determined because it depends on the preference of decision-makers.

This research contributes to fill the stated knowledge gaps. In regard to the first knowledge gap which is the lack of exact method for capturing uncertainty in operational of loading facility , this study shows that considering a stochastic arrival time in the form of discrete probability could improve the operational efficiency of loading facility. In regard to the second gap, which is lack of predictive model based on ML algorithm, this study shows that predicting arrival class (early, on-time, late) is possible and beneficial in maintaining a high-performance operation. The study is conducted in more general sense; therefore the proposed conceptual framework is not supposed to tackle a tailor-fit problem. Given the benefit of the generality of this study, the models and solutions presented in this paper are adaptable for other similar operational problem. Specifically, in situation where the assignment are object specific, unlike the ones in existing literatures where the assignment process is in universal sense.

According to the main findings and acknowledged limitation of this research, set of recommendation could be inferred. For instance, applying other assessment method (CBA or MCDA) to find the tangible trade-off, considering operational and departure as stochastic variable to allow dynamic or adaptive length of loading slots, adding more features of parameter in predictive model (weather condition, type of road disruption, occurrence of accidents, routes taken by trucks, time record at series of GPS coordinate, etc.), conducting hyperparameter-tuning to get a more precise prediction, and exploring the potential of stochastic programming that includes multiple recourse actions and set of scenarios to capture uncertainty.

References

[1] R. Chira, “The role of transport activities in logistics chain,” *Knowledge Horizons. Economics*, vol. 6, no. 3, p. 17, 2014.

[2] H. M. S. Lababidi, M. A. Ahmed, I. M. Alatiqi, and A. F. Al-Enzi, “Optimizing the supply chain of a petrochemical company under uncertain operating and economic conditions,” *Industrial & Engineering Chemistry Research*, vol. 43, no. 1, pp. 63–73, 2004.

[3] R. Ritu Raj, “Resilient Logistics & Distribution System: A Conceptual framework for ABC,” 2019.

[4] V. Sanchez-Rodrigues, A. Potter, and M. M. Naim, “Evaluating the causes of uncertainty in logistics operations,” *The International Journal of Logistics Management*, 2010.

[5] E. Zehendner and D. Feillet, “Benefits of a truck appointment system on the service quality of inland transport modes at a multimodal container terminal,” *European Journal of Operational Research*, vol. 235, no. 2, pp. 461–469, 2014.

[6] W. Zhao and A. v Goodchild, “Using the truck appointment system to improve yard efficiency in container terminals,” *Maritime Economics & Logistics*, vol. 15, no. 1, pp. 101–119, 2013.

[7] A. Ramírez-Nafarrate, R. G. González-Ramírez, N. R. Smith, R. Guerra-Olivares, and S. Voß, “Impact on yard efficiency of a truck appointment system for a port terminal,” *Annals of Operations Research*, vol. 258, no. 2, pp. 195–216, 2017.

[8] B. Wibowo and J. Fransoo, “Joint-optimization of a truck appointment system to alleviate queuing problems in chemical plants,” *International Journal of Production Research*, vol. 59, no. 13, pp. 3935–3950, 2021.

[9] Y. Huiyun, L. Xin, X. Lixuan, L. Xiangjun, J. Zhihong, and B. Zhan, “Truck appointment at container terminals: Status and perspectives,” in *2018 Chinese Control And Decision Conference (CCDC)*, 2018, pp. 1954–1960.

[10] R. Larbi, G. Alpan, P. Baptiste, and B. Penz, “Scheduling cross docking operations under full, partial and no information on inbound arrivals,” *Computers & Operations Research*, vol. 38, no. 6, pp. 889–900, 2011.

[11] S. B. Boswell and J. E. Evans, *Analysis of downstream impacts of air traffic delay*. Lincoln Laboratory, Massachusetts Institute of Technology Lexington, 1997.

[12] E. Mueller and G. Chatterji, "Analysis of aircraft arrival and departure delay characteristics," in *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, 2002, p. 5866.

[13] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, "A review on flight delay prediction," *arXiv preprint arXiv:1703.06118*, 2017.

[14] H. Lee, W. Malik, and Y. C. Jung, "Taxi-out time prediction for departures at Charlotte airport using machine learning techniques," in *16th AIAA Aviation Technology, Integration, and Operations Conference*, 2016, p. 3910.

[15] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transportation research part C: Emerging technologies*, vol. 44, pp. 231–241, 2014.

[16] S. Pongnumkul, T. Pechprasarn, N. Kunaseth, and K. Chaipah, "Improving arrival time prediction of Thailand's passenger trains using historical travel times," in *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2014, pp. 307–312.

[17] N. Marković, S. Milinković, K. S. Tikhonov, and P. Schonfeld, "Analyzing passenger train arrival delays with support vector regression," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 251–262, 2015.

[18] W. Barbour, J. C. M. Mori, S. Kuppa, and D. B. Work, "Prediction of arrival times of freight traffic on US railroads using support vector regression," *Transportation Research Part C: Emerging Technologies*, vol. 93, pp. 211–227, 2018.

[19] L. Vorage, "Predicting Probabilistic Flight Delay for Individual Flights using Machine Learning Models," 2021.

[20] S. Dutrieux, "Predicting Flight Delay Distributions: A Machine Learning-Based Approach at a Regional Airport," 2021.

[21] Z. Qingcheng, Z. Xiaoju, C. Wenhao, and Z. Xiaocong, "Optimization model for truck appointment based on BCMP queuing network," *Journal of Systems Engineering*, vol. 28, no. 5, pp. 592–599, 2013.

[22] J. Yang, Y. Ding, and F. Gen, "Research on Optimal Equipment Allocation in Container Terminals by Using Closed Queueing Networks," *Chinese Journal of Management Science*, vol. 14, no. 6, pp. 56–60, 2006.

[23] Z. Wu and J. Zhang, "Optimal Design of Container Terminal's Gate System Based on M/G/K Queuing Model," in *ICTE 2011*, 2011, pp. 2671–2676.

[24] N. Li, G. Chen, K. Govindan, and Z. Jin, "Disruption management for truck appointment system at a container terminal: A green initiative," *Transportation Research Part D: Transport and Environment*, vol. 61, pp. 261–273, 2018.

[25] M.-H. Phan and K. H. Kim, "Negotiating truck arrival times among trucking companies and a container terminal," *Transportation Research Part E: Logistics and Transportation Review*, vol. 75, pp. 132–144, 2015.

[26] Q. L. Xu, L. J. Sun, X. P. Hu, and L. R. Wu, "Optimization model for appointment of container trucks with nonstationary arrivals," *Journal of Dalian University of Technology*, vol. 54, no. 5, pp. 589–596, 2014.

[27] C. Caballini, J. Mar-Ortiz, M. D. Gracia, and S. Sacone, "Optimal truck scheduling in a container terminal by using a Truck Appointment System," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2525–2530.

[28] C. Yang and Z.-Q. Lu, “Optimizing time windows for delivering export container using genetic algorithm,” *Application Research of Computers*, vol. 30, no. 6, pp. 1643–1646, 2013.

[29] F. Schulte, E. Lalla-Ruiz, R. G. González-Ramírez, and S. Voß, “Reducing port-related empty truck emissions: a mathematical approach for truck appointments with collaboration,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 105, pp. 195–212, 2017.

[30] S. Sen and J. L. Higle, “An introductory tutorial on stochastic linear programming models,” *Interfaces*, vol. 29, no. 2, pp. 33–61, 1999.

[31] M. Şeker and N. Noyan, “Stochastic optimization models for the airport gate assignment problem,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 48, no. 2, pp. 438–459, 2012.

[32] L. Meng and X. Zhou, “Robust single-track train dispatching model under a dynamic and stochastic environment: A scenario-based rolling horizon solution approach,” *Transportation Research Part B: Methodological*, vol. 45, no. 7, pp. 1080–1102, 2011.

[33] O. R. P. van Schaijk and H. G. Visser, “Robust flight-to-gate assignment using flight presence probabilities,” *Transportation Planning and Technology*, vol. 40, no. 8, pp. 928–945, 2017.

[34] Bureau of Transportation Statistics, “Reporting Carrier On-Time Performance (1987-present),” *https://transtats.bts.gov/DL_SelectFields.asp?gnoyr_VQ=FGJ*.

[35] L. Zonglei, W. Jiandong, and Z. Guansheng, “A new method to alarm large scale of flights delay based on machine learning,” in *2008 International Symposium on Knowledge Acquisition and Modeling*, 2008, pp. 589–592.

[36] H. Khaksar and A. Sheikholeslami, “Airline delay prediction by machine learning algorithms,” *Scientia Iranica*, vol. 26, no. 5, pp. 2689–2702, 2019.

[37] S. van der Spoel, C. Amrit, and J. van Hillegersberg, “Predictive analytics for truck arrival time estimation: a field study at a European distribution centre,” *International journal of production research*, vol. 55, no. 17, pp. 5062–5078, 2017.

[38] V. Kumar, B. A. Kumar, L. Vanajakshi, and S. C. Subramanian, “Comparison of model based and machine learning approaches for bus arrival time prediction,” in *Proceedings of the 93rd Annual Meeting*, 2014, pp. 14–2518.

[39] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris, “Prediction of weather-induced airline delays based on machine learning algorithms,” in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, 2016, pp. 1–6.

[40] I. Parolas, “ETA prediction for containerships at the Port of Rotterdam using Machine Learning Techniques,” 2016.

[41] M. Y. Maknoon, F. Soumis, and P. Baptiste, “An integer programming approach to scheduling the transshipment of products at cross-docks in less-than-truckload industries,” *Computers & Operations Research*, vol. 82, pp. 167–179, 2017.

Appendix B: Synthetic Data

Specification of Synthetic Data

As it is explained, the synthetic data is adapted from publicly accessible database of airport operation. To fit the context of this study adjustment in terminology of origin location is necessary. The initial origin data consists of IATA code of airport which is being transformed into a more general term based on alphabet. There are 20 origin location assumed in this research as it is shown in Table B.1.

Table B.1 Look-up Table for Adaptation Done in Origin Location

ORIGIN			
Raw	Adapted	Raw	Adapted
MCO	A	LAX	K
FLL	B	DCA	L
TPA	C	CLT	M
DFW	D	JAX	N
MSY	E	PHL	O
MIA	F	PBI	P
ORD	G	DEN	Q
LGA	H	HOU	R
RDU	I	DTW	S
BWI	J	JFK	T

In terms of distance group and product type, the specification is shown in the following Table B.2.

Table B.2 Look-up Table for Distance Group and Product Type

DISTANCE GROUP		Product Type	
Code	Range (Kilometers)	Code	Product
1	0-50 KM	P0	Product 0
2	50-100 KM	P1	Product 1
3	100-150 KM	P2	Product 2
4	150-200 KM	P3	Product 3
5	200-250 KM	P4	Product 4
		P5	Product 5
		P6	Product 6
		P7	Product 7

Final Format of Historical Data

The final form of the historical data after going through all the preparation process is shown in Table B.3 and Table B.4. The former shows the ones without the real time information, whereas the latter shows the ones with the real time information.

Table B.3 Snippet of Final Version of the Synthetic Dataset Without Real Time Information

INDEX	PRODUCT TYPE	ORIGIN	DESTINATION	DAY NAME	SCHEDULED DEPARTURE	SCHEDULED ARRIVAL	ARRIVAL DELAY	DISTANCE GROUP
0	P0	B	Facility	Monday	2/1/2016 0:55	2/1/2016 6:24	13	5
1	P3	G	Facility	Monday	2/1/2016 5:20	2/1/2016 6:49	-14	2
2	P0	J	Facility	Monday	2/1/2016 5:20	2/1/2016 7:28	-11	2
3	P1	R	Facility	Monday	2/1/2016 5:30	2/1/2016 8:30	-2	3
4	P2	A	Facility	Monday	2/1/2016 5:30	2/1/2016 7:59	-11	2
5	P2	S	Facility	Monday	2/1/2016 5:30	2/1/2016 8:51	-10	3
6	P2	D	Facility	Monday	2/1/2016 5:35	2/1/2016 7:32	1	1
7	P4	A	Facility	Monday	2/1/2016 5:35	2/1/2016 8:59	-7	4
8	P5	I	Facility	Monday	2/1/2016 5:40	2/1/2016 9:05	6	4
9	P6	E	Facility	Monday	2/1/2016 5:40	2/1/2016 7:30	-8	3
....	....	....	....	....	....	....	....	....
90198	P5	H	Facility	Saturday	2/29/2020 20:49	2/29/2020 23:57	26	3

Table B.4 Snippet of Final Version of the Synthetic Dataset With Real Time Information

Index	TRUCK	ORIGIN	DESTINATION	DAY NAME	SCHEDULED DEPARTURE	DEPARTURE TIME	DEPARTURE DELAY	SCHEDULED ARRIVAL	ACTUAL/ESTIMATED ARRIVAL TIME	ARRIVAL DELAY	ESTIMATED ELAPSED TIME	DISTANCE GROUP	DELAY LEVEL
0	P5	B	Facility	Monday	2020-02-17 10:29:00	10:31:00	2	2020-02-17 12:13:00	12:07:00	-6	96	2	1
1	P6	G	Facility	Monday	2020-02-17 10:47:00	11:15:00	28	2020-02-17 12:07:00	12:37:00	30	82	2	2
2	P2	J	Facility	Monday	2020-02-17 18:59:00	18:48:00	-11	2020-02-17 20:30:00	20:42:00	12	114	2	2
3	P3	R	Facility	Monday	2020-02-17 06:00:00	05:59:00	-1	2020-02-17 08:17:00	09:01:00	44	182	3	2
4	P4	A	Facility	Monday	2020-02-17 19:57:00	19:43:00	-14	2020-02-17 21:18:00	21:11:00	-7	88	2	1
5	P1	S	Facility	Monday	2020-02-17 10:29:00	10:31:00	2	2020-02-17 12:13:00	12:07:00	-6	96	2	1
6	P2	D	Facility	Monday	2020-02-17 10:47:00	11:15:00	28	2020-02-17 12:07:00	12:37:00	30	82	2	2
7	P0	A	Facility	Monday	2020-02-17 19:29:00	19:44:00	15	2020-02-17 21:31:00	21:51:00	20	127	3	2
8	P0	I	Facility	Monday	2020-02-17 11:49:00	11:49:00	0	2020-02-17 13:00:00	12:58:00	-2	69	1	1
9	P3	E	Facility	Monday	2020-02-17 06:00:00	06:05:00	5	2020-02-17 11:24:00	11:14:00	-10	189	7	0
…	....		....	....	....	....	....	....	....	....	....	....	....
90198	P6	B	Facility	Monday	2020-02-17 19:29:00	19:44:00	15	2020-02-17 21:31:00	21:51:00	20	127	3	2

Appendix C: Predictive Model

Pearson Correlation

Pearson correlation is a method that evaluates the linear correlation between the input features and the target variable and the linear correlations between features themselves. Features which have a high linear correlation with each other can lead to multicollinearity for some machine learning models. Pearson correlation is calculated between two features of n represent the sample size, where $\bar{x}$ and $\bar{y}$ represent the means respectively of the features x and y. The formula is shown as following:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

The Pearson correlation for every feature combination is calculated, which results in a so-called ’Pearson correlation matrix’. Table 2.1 shows a selection of the Pearson correlation matrix, as the whole matrix is too big to fit on paper. In this selection, all features are found with are dropped for having an absolute correlation value of 0.8 or higher.

Validation Result

The metric score of KPI presented in Section 5.6.1 are based upon the result of confusion matrix shown in this following Table C.1, Table C.2, and Table C.3. The detail of default hyperparameter used in the predictive model is shown in Figure C.4

Table C.1 Result of GNB

		Predicted		
		Early	On Time	Late
Actual	Early	3947	1142	116
	On Time	5038	2282	745
	Late	1086	891	2937

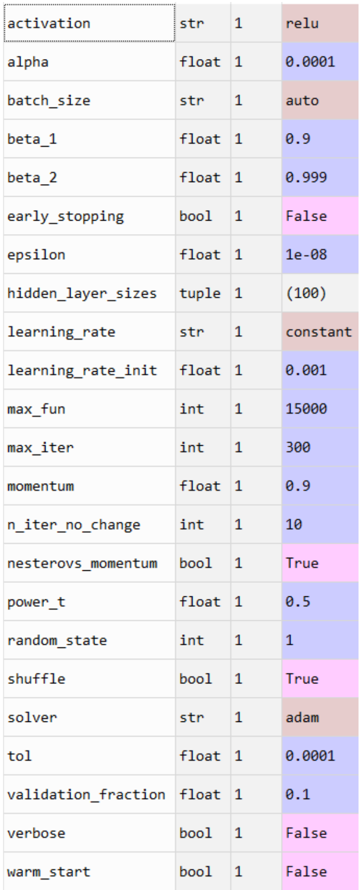
Table C.2 Result of Logistic Regression

		Predicted		
		Early	On Time	Late
Actual	Early	3851	1209	145
	On Time	3302	3895	868
	Late	119	945	3850

Table C.3 Result of Artificial Neural Network

		Predicted		
		Early	On Time	Late
Actual	Early	3966	1205	34
	On Time	2822	4543	700
	Late	145	646	4123

Table C.2 Hyperparameter setting of ANN structure



activation	str	1	relu
alpha	float	1	0.0001
batch_size	str	1	auto
beta_1	float	1	0.9
beta_2	float	1	0.999
early_stopping	bool	1	False
epsilon	float	1	1e-08
hidden_layer_sizes	tuple	1	(100)
learning_rate	str	1	constant
learning_rate_init	float	1	0.001
max_fun	int	1	15000
max_iter	int	1	300
momentum	float	1	0.9
n_iter_no_change	int	1	10
nesterovs_momentum	bool	1	True
power_t	float	1	0.5
random_state	int	1	1
shuffle	bool	1	True
solver	str	1	adam
tol	float	1	0.0001
validation_fraction	float	1	0.1
verbose	bool	1	False
warm_start	bool	1	False

Input for Predictive Model

As it is explained, there are three required steps to ensure the suitability of data, which are selection, encoding and scaling . The example format of each step in following Table C.5, Table C.6 and Table C.7, respectively.

- **Table C.5 Snippet of Selected Features as the Input of Predictive Model**

PRODUCT	ORIGIN	DAY	SCHEDULED	DEPARTURE	DEPARTURE	SCHEDULED	ARRIVAL	DISTANCE
		NAME	DEPARTURE	TIME	DELAY	ARRIVAL	TIME	GROUP
P1	31	Monday	09:50	10:06	16	12:25	12:55	2
P0	159	Tuesday	09:45	09:42	-3	12:05	12:07	3
P5	41	Thursday	10:14	10:10	-4	13:30	13:40	3
P2	11	Saturday	10:20	10:20	0	14:44	14:32	4
P6	90	Friday	13:34	14:16	42	15:27	16:07	2
P0	131	Wednesday	15:39	15:35	-4	16:33	16:55	3

Table C.6 Snippet of Encoded Features as the Input of Predictive Model

TRUCK	ORIGIN	DAY	SCHEDULED	DEPARTURE	DEPARTURE	SCHEDULED	ARRIVAL	DISTANCE
		NAME	DEPARTURE	TIME	DELAY	ARRIVAL	TIME	GROUP
1	31	1	29728	293	16	38565	418	2
0	159	2	26977	134	-3	35219	336	3
5	41	4	33915	533	-4	43790	754	3
2	11	6	19681	492	0	26014	695	4
6	90	5	25303	439	42	33171	573	2
0	131	3	33781	133	-4	43582	275	3

Table C.7 Snippet of Scaled Features as the Input of Predictive Model

TRUCK	ORIGIN	DAY	SCHEDULED	DEPARTURE	DEPARTURE	SCHEDULED	ARRIVAL	DISTANCE
		NAME	DEPARTURE	TIME	DELAY	ARRIVAL	TIME	GROUP
-0.59439	1.84158	1.4162	0.38759	-1.38157	-0.467826	0.401389	-1.2474	0.0816757
2.46211	-0.92874	0.93298	0.928682	0.0514862	-0.514236	0.920721	0.230595	0.0816757
-0.59439	-1.63306	-0.516679	-0.181422	-0.0957699	-0.328595	-0.156358	0.021979	0.668796
-0.59439	0.221646	1.4162	0.257036	-0.286125	1.62063	0.277297	-0.409397	-0.505445
-0.59439	1.18422	0.93298	0.918231	-1.38516	-0.514236	0.908118	-1.46309	0.0816757
1.23951	-1.79741	1.4162	1.49239	1.49172	0.0890961	1.47817	1.31611	-1.09257

Appendix D: Scenario Data

A snippet of data format used in the numerical experiments is shown in Table D.1. It corresponds to the Scenario 1 of varied congestion level. For the other scenarios, the format is still the same, but adaptation is required to fit the context of scenario, especially in number of trucks and utilization rate of loading slot. The data is not shown because it is too large.

Table D.1 Snippet Data for Scenario 1

Index Truck	Product type	Initial Schedule	Mild		Medium		Severe	
			ETA	Delay	ETA	Delay	ETA	Delay
0	0	7:00:00 AM	6:43:31 AM	-16.49	7:08:00 AM	8.19	7:12:00 AM	12.20
1	4	7:00:00 AM	6:31:21 AM	-28.6562	7:25:00 AM	25.52721	7:27:00 AM	27.6
2	3	7:30:00 AM	7:05:28 AM	-24.5363	7:51:00 AM	21.3843	8:06:00 AM	36.44
3	6	7:30:00 AM	6:40:43 AM	-49.2766	6:53:20 AM	-36.6672	7:24:30 AM	-5.5
4	1	8:00:00 AM	7:52:10 AM	-7.83978	7:37:54 AM	-22.1056	8:13:00 AM	13.6
5	5	8:00:00 AM	8:15:00 AM	15.67567	7:31:54 AM	-28.1045	8:01:00 AM	1.2
6	3	8:30:00 AM	7:33:35 AM	-56.424	8:33:00 AM	3.509122	8:39:00 AM	9.1
7	7	8:30:00 AM	8:04:37 AM	-25.3828	8:23:08 AM	-6.86204	8:38:00 AM	8.2
8	0	9:00:00 AM	9:24:00 AM	24.82127	8:22:12 AM	-37.7981	9:02:00 AM	2.3
9	6	9:00:00 AM	8:50:09 AM	-9.85763	9:08:00 AM	8.989969	9:02:00 AM	2.1
10	0	9:30:00 AM	9:35:00 AM	5.385718	9:20:01 AM	-9.98646	9:46:00 AM	16.3
11	6	9:30:00 AM	9:23:45 AM	-6.24448	9:14:43 AM	-15.2801	9:47:00 AM	17.4
12	3	10:00:00 AM	10:21:00 AM	21.48931	9:05:10 AM	-54.8298	10:41:00 AM	41.2
13	7	10:00:00 AM	9:43:40 AM	-16.3388	9:55:58 AM	-4.0383	9:56:12 AM	-3.8
14	2	10:30:00 AM	10:00:50 AM	-29.1695	10:32:00 AM	2.890046	10:19:24 AM	-10.6
15	4	10:30:00 AM	10:12:49 AM	-17.1759	10:37:00 AM	7.320136	10:17:36 AM	-12.4
16	2	11:00:00 AM	10:40:04 AM	-19.9286	11:11:00 AM	11.50301	10:31:36 AM	-28.4
17	6	11:00:00 AM	11:26:00 AM	26.75084	11:19:00 AM	19.25437	11:02:00 AM	2.3
18	1	11:30:00 AM	11:21:17 AM	-8.72214	11:20:14 AM	-9.7734	11:24:24 AM	-5.6
19	4	11:30:00 AM	11:19:21 AM	-10.6452	10:57:17 AM	-32.7238	11:34:00 AM	4.2
...	..	...	...	...	...	...	...	...
30	..	3:00:00 PM	2:44:08 PM	-15.8593	2:41:36 PM	-18.3918	3:12:00 PM	12.5
31	0	3:00:00 PM	3:21:00 PM	21.24461	3:28:00 PM	28.6515	3:22:00 PM	22.4
32	4	3:30:00 PM	3:21:00 PM	-9.0012	3:43:00 PM	13.4	3:35:00 PM	5
33	1	3:30:00 PM	3:19:59 PM	-10.0167	3:22:06 PM	-7.90658	3:44:00 PM	14.2
34	5	4:00:00 PM	3:52:17 PM	-7.71064	3:31:36 PM	-28.3928	3:50:18 PM	-9.7
35	0	4:00:00 PM	3:44:10 PM	-15.8342	4:12:00 PM	12.24141	4:33:00 PM	33.20
36	7	4:30:00 PM	4:02:29 PM	-27.5228	4:43:00 PM	13.2	5:00:00 PM	30.40
37	0	4:30:00 PM	4:44:00 PM	14.8	5:05:00 PM	35.4	5:36:00 PM	66.50
38	4	5:00:00 PM	5:26:00 PM	26.77818	5:08:00 PM	8.2	5:14:00 PM	14.20
39	3	5:00:00 PM	4:44:23 PM	-15.6173	5:36:00 PM	36.92256	5:12:00 PM	12.40
40	6	5:30:00 PM	4:59:00 PM	-31	5:15:00 PM	-15	5:24:30 PM	-5.50
41	1	5:30:00 PM	5:36:00 PM	6.235576	5:40:00 PM	10.4	5:52:00 PM	22.50

Appendix E: Result

Corresponding to Section 8.3, the detailed calculation of comparison between baseline and probabilistic is provided in below. The value can be interpreted as how well the P-SRP outperforms the baseline. Table E.1 refers to Scenario 1, Table E.2 refers to Scenario 2, and Table E.3 refers to Scenario 3.

Table E.1 Performance improvement offered by P-SRP compared to baseline in Scenario 1

	Mild	**Medium**	**Severe**
Reduced expected cost	54.17%	54.88%	47.06%
Reduced rescheduled trucks	46.43%	45.00%	28.57%
Reduced possible idled slots	40.00%	66.67%	50.00%

Table E.2 Performance improvement offered by P-SRP compared to baseline in Scenario 2

	2 Loading Bay			**3 Loading Bay**			**4 Loading Bay**		
	50%	75%	100%	50%	75%	100%	50%	75%	100%
Reduced expected cost	2.27%	13.33%	29.03%	22.86%	19.61%	53.60%	7.94%	27.94%	56.08%
Reduced rescheduled trucks	5.56%	17.86%	22.73%	27.59%	21.43%	24.19%	11.90%	12.50%	11.63%

Table E.3 Performance improvement offered by P-SRP compared to baseline in Scenario 3

	Minimizing waiting time		**Maintaining initial schedule**	
	Without stand-in	With stand-in	Without stand-in	With stand-in
Reduced Expected Cost	47.0%	15.2%	53.6%	53.6%
Reduced Rescheduled Trucks	-4.3%	2.1%	24.2%	25.8%

The method used to estimate possible idled slot will be explained. For example, according to Table E.4, there are 3 trucks categorized as possible idled slot because their presence probability is lower than 0.2. In the baseline model, each truck is assigned to a loading bay without considering their presence probability, thus there are 3 slots that are probably being idled that could result on less efficient operation. On the other hand, the P-SRP takes into account the presence probability that allows it to simultaneously assign other truck to that same slot in situation where there is possibility of idled slot (highlighted yellow).

Table E.4 Investigation on Possible Idled Slots

Index Truck	**Loading Bay**	**Initial**	**Presence Probability**	**Baseline**	**Probabilistic**
1	L1	2	0.150	3	3
3	L1	3	0.538	2	3
…	…	…	…	…	…
9	L1	6	0.176	7	7
11	L1	7	0.355	8	7
…	…	…	…	…	…
36	L0	21	0.627	22	22
38	L0	22	0.174	23	22