COMP4801

Final Year Project

MTR TRAIN DATA ANALYSIS

Eashan Trehan | Aditya Mehta | Pang Ming Kin

Detailed Intermediate Report | 2nd February 2020

ABSTRACT

An increase in local connectivity within countries and cities in recent decades has sparked a new expansion in the railway industry, in an effort to ease transfer of human and commercial capital across regions. Developing nations and emerging markets around the globe, and in particular in Asia and Africa, with rapidly expanding economies, are actively attempting to connect citizens residing in rural areas to major urban centres and metropolis'. Well developed cities such as Hong Kong, which is a major global financial hub and boasts of quality infrastructure and transportation networks, are expanding investments in fields such as technology. Hong Kong's rail system which offers connectivity throughout the city has a good performance record in terms of reaching destinations on time. However, rail systems are not foolproof and may suffer from disruptions, such as technical failures and accidents, leading to delays. At present, there exists no conclusive technological response platform for mitigating the impact of such mishaps and offering relevant, real-time recommendations for optimising the handling of incidents. The project explores a solution in the form of a recommendation engine which can simulate various scenarios and thereby propose effective responses to potential mishaps. The report therefore, acts as a platform for discussing the potential application, benefits and challenges of utilising technology in transport systems and in particular, railway networks. The work discusses the available technological choices and various development approaches and compares them, also discussing relevant works of literature offering domain expertise, constraints faced during the project and remedial measures taken to tackle them. Future prospects of the project include automation of the entire rail network, from automated generation of timetables to driverless trains.

ACKNOWLEDGEMENTS I

We would like to express our deep and sincere gratitude to our project supervisor, Dr. Reynold Cheng, Associate Professor, Department of Computer Science, University of Hong Kong (HKU), for giving us the opportunity to work under his much valued guidance and supervision, throughout the course of this project. He has imparted his technical expertise and knowledge to us, in the field of software engineering and data analytics, which has enabled us to carry out research and meet the required deliverables. It has been a great honour and privilege for us to study and work under his guidance.

We are extremely grateful to the faculty members at the University of Hong Kong (HKU), including Dr. Micheal Ng, Director of Research Division for Mathematical and Statistical Science, Dr. Wai Keung Li, Chair of Statistics, Dr. Wai-Ki Ching, Professor of Mathematics, Dr. Benjamin Kao, Professor of Computer Science, Dr. Philip Yu, Associate Dean of the Depart of Statistics, and Mr. Kwan, Assistant IT Director. Without their expert inputs and invaluable experience, our project could not have been as successful.

We are extremely grateful to our fellow students belonging to the Department of Computer Science at University of Hong Kong (HKU), Ms. Wenya Sun, PhD Candidate, Mr. Tobias Grubenmann, Postdoctoral Researcher, and Ms. Kai-i Lam (Nancy), Penultimate Year Undergraduate Student, for their support and patience during the discussions we had on research work and preparation.

TABLE OF CONTENTS II

1. Introduction……..……..……..7
    1. Mass Transit Railway Corporation Limited (MTR)……7
    2. Problem Statement……..……8
    3. Objective and Deliverable……8
    4. Scope……9
    5. Potential Impact and Benefits……..9
    6. Personnel Involved…..……..……..……10
    7. Report Outline..……..……..……10
2. Methodology…..……..……..……..……11
    1. Bayesian Networks…….……..……..……11
    2. Agile Approach…….……..……11
    3. Hardware Requirements..……..……..……11
    4. Software Development……..……..……12
    5. Data Collection and Software Requirements……..……..……..……..13
3. Literature Review……15
    1. Introduction……..……..……15
    2. Analytical Models…….……..15
    3. Machine Learning Models..……..……..……16
    4. OpenTrack……..……..……..……..……..……16
    5. Miscellaneous……..……..……..……..……..16
4. Results and Discussion……..……..……..18
    1. Initial Findings……..……..……..……18
    2. Data Pipeline..……..……..……..……..……18
    3. Data Analysis…….……..……..……..……19
    4. Machine Learning Model…..……..……..……..……21
    5. Limitations and Difficulties Encountered……..……..……22
5. Future Work..……..……..……..24
6. Conclusion……..……..……..25
7. References……..……..……..……..26

LIST OF FIGURES III

Figure 1 Market Share of Public Transport in Hong Kong…..……7
Figure 2 Kwun Tong Line (KTL) Track Layout……..…….…….…….……9
Figure 3 Software Development Cycle…….…….…….…….……..12
Figure 4 Published Papers on Railway Network Simulation.…….…….……..15
Figure 5 Average Time Spent at a Station….……..…….……..19
Figure 6 Train Delay Distribution on Sunday Morning….……..…….…….……20
Figure 7 Kwun Tong Line Track Layout.…….…….…….……20
Figure 8 Timeline.…….…….24

LIST OF TABLES IV

Table 1 Average Dwell Time of Stations….……..……..……21
Table 2 Performance Analysis of Various Models.…….……22
Table 3 Risk Matrix……..…….…….…….……23

ABBREVIATIONS V

API Application Program Interface

CAES Centre for Applied English Studies

CMB China Merchants Bank

GPU Graphics Processing Unit

HKU The University of Hong Kong

HSBC Hong Kong and Shanghai Commercial Bank

KTL Kwun Tong Line

LSE The London School of Economics and Political Science

ML Machine Learning

MTR Mass Transit Railway Corporation Limited

UofT The University of Toronto

TERMINOLOGY VI

Accumulative Delay	The end-to-end delay from the scheduled journey time
Arrival Delay	Difference between the scheduled and actual arrival time
Auto-Sharding	Database partitioning into smaller parts
Black Swan	Extreme statistical anomaly
Caffe	A deep learning framework
Departure Delay	Difference between the scheduled and actual departure time
Dwell Time	Time taken between a train entering and leaving a station
GPU	Electronic device for processing graphics
Headway	Time elapsed from the track occupancy of the previous train
Journey Time	Time taken in an end-to-end journey between terminal stations
Keras	A neural network API
Keras-vis	Toolkit for visualising neural network models
MongoDB	Document database program
Neural Network	Algorithm to determine relationships in a data set
Python	A particular programming language
Pytorch	Python based scientific computing package
R	A particular programming language
Scheduled Time	Scheduled running or dwell time corresponding to the track segment
Scikit-Learn	Python based machine learning platform
Sci-Py	Open source Python library for scientific computing
SQL Server	Relational database management system by Microsoft
Station Type	Whether the station is an interchange station
TensorFlow	Open source library for machine learning
Theano	Python based library for mathematical operations

1 INTRODUCTION

The objective of this chapter is to introduce this progress report, beginning with a brief background on the Mass Transit Railway Corporation Limited (MTR), followed by a concise problem statement describing the current issues, the purpose and final deliverable of the project, as well as a well defined scope. The chapter then discusses the potential benefits of the project, the planned schedule to be followed, a description of the personnel involved and lastly, an overview of the structure and content covered in the report.

1.1 Mass Transit Railway Corporation Limited (MTR)

The MTR is a 40 year old organisation, acting as a major public transport network, offering railway connectivity in Hong Kong, covering all 18 districts and on average, enabling over 12 million passenger journeys, every weekday (MTR, 2019). The MTR likely has a positive reputation with regards to service quality and punctuality, given that it has achieved an on-time rate of 99.9% between February 2018 and January 2019, inferring that on average, 999 out of 1,000 passengers generally arrive at their desired destination within a 5-minute period of the scheduled time (MTR, 2019).

This is further reaffirmed by figure 1, as shown below, which highlights the market share of various means of public transport in Hong Kong, indicating the market leading popularity of the MTR, with a 41% market share (Legislative Council Secretariat, 2017).

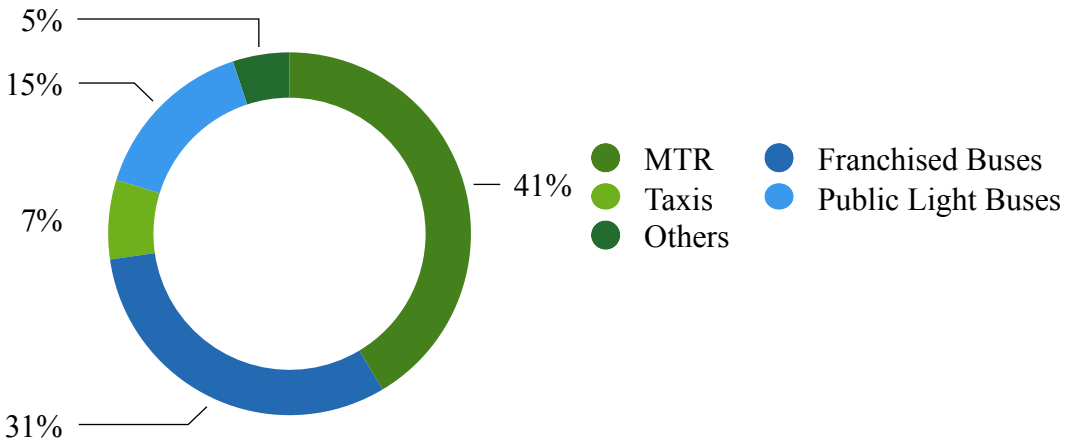


Figure 1: Market Share of Public Transport in Hong Kong

1.2 Problem Statement

As aforementioned, the MTR has delivered satisfactory performance for its patrons, however, there continue to be incidents leading to passenger delays. For instance, on the 25th of October 2018, train services at Po Lam station suffered from a significant delay, due to power supply issues, prompting the MTR to arrange shuttle buses for affected passengers (Wong, 2018). Similarly, on the 5th of August 2017, the Kwun Tong Line (KTL) suffered a delay of more than 10 hours, due to signalling issues (Chung, 2017). Such incidents may cause significant inconvenience to passengers as well as the MTR, since, in the event of a delay of over 5 minutes, the MTR is required to pay a penalty to the Hong Kong government, while commuters may also suffer from the loss of valuable time.

At present, when facing disruptive scenarios, such as the ones listed above, the MTR generally relies on possibly the past experience and technical expertise of their train operators for minimising the negative impact. However, this may potentially lead to an increase in the risk of a human error taking place, on part of the train operator, possibly further compounding the problem.

Hence, a problem statement may be summarised as follows - The MTR, is at present, likely reliant on train operators and officials to minimise the delay caused by a potential disruption, thereby potentially heightening the risk of a human error occurring. Furthermore, there is a possible need to reduce delays, in order to deliver potentially improved service to its commuters and reduce the probability of incurring a fine from the government.

1.3 Objective and Deliverable

The goal of this project is to work with the MTR, in order to develop a software platform that analyses and visualises train schedule and movement, for example, the amount of time the train spends at a platform, the departure time of the train, the amount of time for which the train stops in a tunnel etc. with the aim of reducing the delay suffered. The project is intended to reduce the number of trains that suffer from a short delay, i.e. greater than or equal to 5 minutes and long delay, i.e. 30 minutes or above.

The end deliverables of the project include a working simulation model, visualised, to take into account various scenarios that a train may face, such as a signal fault, overcrowding of passengers

etc. which could potentially cause a disruption to train service, and build a software/algorithm that can provide relevant recommendations to deal with such a scenario in an optimised manner.

1.4 Scope

The scope of the project is limited to a simple visualiser which can possibly simulate a finite set of scenarios in a 2-Dimensional space, i.e. the visualiser cannot account for all the infinite possibilities and the exact scenarios which are possible. The recommendations that are made by the software/ algorithm are expected to be in approximately real-time and are limited to certain options and constraints that have been communicated to the team by the MTR, for instance, the software may recommend the train operator to offboard passengers from the train, however it cannot recommend changing the direction of movement of the train in the opposite direction. Furthermore, the project is specifically focused on the Kwun Tong Line (KTL) in particular, as shown in figure 2 below, and does not take into account other lines of service, such as the Tsuen Wan Line etc.

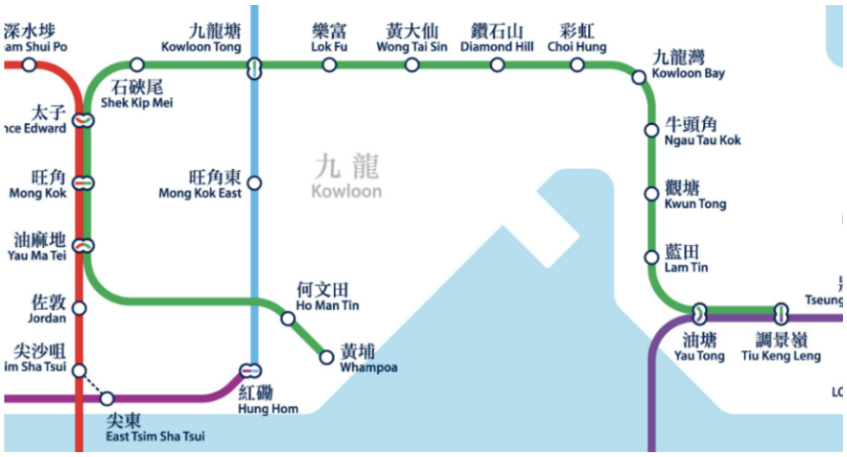


Figure 2: Kwun Tong Line (KTL) Track Layout

1.5 Potential Impact and Benefits

As the objective of the project is to minimise the delays suffered by the train, the potential benefits from the impact of this project include, but are not limited to, a reduction in delays, improved efficiency and time management, lower probability of the MTR being penalised by the government for extensive delays and time being saved for commuters. Should the project be successful in implementing a simulation model and recommendation system, it may possibly pave the way for driverless trains in the future, as the current MTR trains are operated by train operators.

1.6 Personnel Involved

The project is a large scale collaboration between the MTR and the University of Hong Kong (HKU), involving multiple departments, including the department of Computer Science, Statistics and Mathematics. The project is supervised by Dr. Reynold Cheng, Associate Professor of the Department of Computer Science, HKU, Dr. Cheng's main research area is large-scale data management and he has worked on the modelling, querying, cleaning, mining, and system development of databases.

The primary team of this project consists of three members, namely Eashan Trehan, Aditya Mehta and Pang Ming Kin. Eashan Trehan, the author of this report, is pursuing a Computer Science and Finance double major at HKU and has attended exchange programs at the University of Toronto (UofT) and the London School of Economics and Political Science (LSE), with past work experience at J.P. Morgan and Deutsche Bank in Hong Kong. Aditya Mehta is also a Computer Science and Finance double major, with experience in Machine Learning and has previously worked at the Hong Kong and Shanghai Banking Corporation (HSBC). Pang Ming Kin is also a Computer Science student, with experience in Data Analysis, and has previously worked as an intern at the Alibaba Group and China Merchants Bank (CMB). The team is supported by Ms. Wenya Sun, a PhD student at the Department of Computer Science and Kai-i Lam, a penultimate year Computer Science major at HKU.

1.7 Report Outline

The following chapters of the report are arranged as follows. Chapter 2 discuss the methodology and approach adopted, including various software and hardware choices, data collection requirements and rationale behind certain decisions made. Followed by Chapter 3 on literature review, discussing relevant research and reports pertaining to the project. Chapter 4 then entails the initial findings and limitations of the project thus far, and how certain risks can be mitigated, concluding with a discussion on future work to be completed in Chapter 5. A summary of the current progress is then discussed in Chapter 6.

2 METHODOLOGY

The following chapter discusses the methodology and certain technical concepts adopted in the project, such as Bayesian networks, agile approach, software development process and hardware requirements. The chapter concludes with a discussion on data requirements and the rationale behind various technical decisions and choices made.

2.1 Bayesian Networks

Bayesian network is a type of statistical model which represents the conditional dependencies of a group of variables in the form of a directed acyclic graph. There are several advantages of opting for Bayesian networks, such as avoiding over-fitting of data and better handling of missing data points.

A dynamic bayesian network in particular can be utilised for a time series prediction for the train, using the day and time as explanatory variables, since certain times during particular days may tend to be more crowded than others. Using the data collected by the MTR, a Bayesian inference network may be trained, as the delay time of the train is treated as a stochastic process.

2.2 Agile Approach

An agile software development approach has been adopted for the purpose of this project. Agile refers to an iterative method of working, involving regular feedback between the developers and various stakeholders, with the aim of keeping stakeholders updated on the progress and direction of the project, while gaining a better sense of their expectations and requirements. Hence, as part of the project, weekly internal meetings are organised, along with bi-weekly meetings with the supervising professor and monthly meetings with officials representing the MTR.

2.3 Hardware Requirements

As the project is data intensive and based on software development, the hardware requirements are likely low and include Graphical Processing Units (GPUs) with high processing capabilities that are likely required in order to successfully process vast amounts of data in a short time frame. Based on the data files provided, it is estimated that one standard hard drive (1 TeraByte) should suffice for processing data volumes up to one year. Further analysis may require processing additional data on a longer timeline. That would require extended secondary storage. Since the nature of the data is

extremely sensitive, it obviates the usage of commercial cloud storage providers such as Google Drive, GitLab, Microsoft OneDrive, etc. Any viable solution must be offline and proprietary. For the training of the Machine Learning model, significant computing capability will be required. The scope provided by the MTR include keeping the model training time within acceptable limits, which is defined as less than one hour of processing for a 24-hour data extract. To satisfy these constraints, specialised hardware called GPUs, mentioned earlier, will be utilised. These are specialised for parallel computations that are required for training Machine Learning models and the MTR has been requested for procurement of "NVIDIA RTX 2080", which is a state-of-the-art hardware required to fulfil this purpose.

2.4 Software Development

As illustrated by Figure 3 below, there are potentially four key steps to developing a working simulation model that can be visualised and provide relevant recommendations. Firstly, the algorithm accepts various forms of data, variables such as the train number, the track it is located at, the station it is leaving or arriving at, what kind of situation or incident is the train experiencing etc. Based on these data points, a recommendation may be produced, to possibly optimise the handling of any extraordinary situation. The recommendation is then to be simulated, to judge whether it could produce optimal results and the final selected recommendation is then simulated visually.

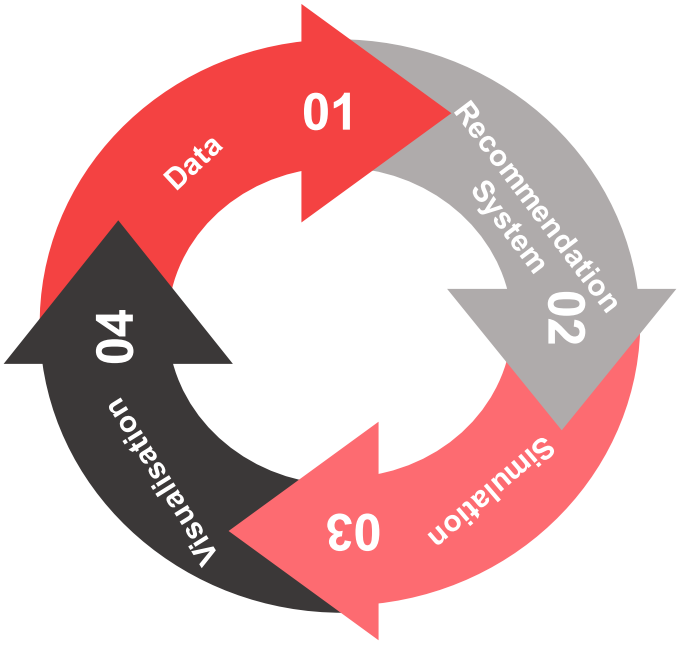


Figure 3: Software Development Cycle

To accomplish the above deliverable, first the track layout of the railway should be understood and relevant data should be collected. Certain data points such as the day, time of day, track segment, station name, direction etc. should be specified in the data collected. Next, a visualisation model is required to be built to showcase the train movement, the following chapter discusses the various options available for creating such a model. The data collected earlier should be used to simulate the train movement in the visualisation model. Once the data can be accurately simulated in the visualisation model, a recommendation system can be created using "Bayesian Networks" as aforementioned at the beginning of this chapter, taking into account various scenarios that may occur, including incidents such as technical failures.

2.5 Data Collection and Software Requirements

The project potentially involves extensive data requirements, including but not limited to - the distance between various stations, the planned timetable for different times and days of the week, the dwell time at each station etc. These datasets are preferred to be available in .csv or Excel format in particular, for ease of use, and may potentially be provided by the MTR. The quality, quantity, and availability of data are the primary concerns for any data driven project. Machine learning models show increasingly better performance when trained using more volume of data (Cong et al., 2007). For statistical models, estimated parameters exhibit lower variance and greater predictive ability with increasing sample size (Rosenfeld, 2018), therefore, getting access to relevant data is essential. The MTR has provided three data extracts till date, covering a period of 9 days. The first data file covers the 3 hour period from 2000h – 2300h on August 11, 2019. The second data file cover the whole of August 4, 2019. The third includes the week of 6th December to 12th December 2019. A request for access to additional data has been filed, which is currently subject to internal approvals within the MTR Corporation.

"Python" and "MongoDB" have been chosen to analyse the data. Python was chosen in particular, over the likely alternative "R", as the latter is mostly used for statistical analysis, while the former can be used more widely for data science purposes in general and provides ease of replicability and accessibility, in comparison. MongoDB was preferred over "SQL server", as the former offers better availability and scalability, due to auto-sharding and makes it easier to represent complex relationships between different variables in the data. The project involves the use of Machine Learning, which will be executed by utilising libraries such as SciPy and Scikit-Learn, while Deep Learning can be implemented via Caffe, TensorFlow, Theano or Keras and be visualised through

Pytorch or Keras-vis. Each of these options present their own respective advantages and disadvantages and a final decision will be taken after all the relevant data from MTR and studied it in detail. Given the sensitive nature of the project and data privacy requirements, the data and code assembled for the project is being stored in an online platform called “GitLab”, rather than the more popular alternative called “GitHub”, as the former provides a private account free of charge, unlike the latter. The visualisation created to showcase the simulation of the project is built through “Unity”, a platform and programming language which is often used for creating games, films etc. and is apt for creating a 2-Dimensional simulation model for the purpose of this project. Unity was chosen in particular, over other options, such as “Arena”, "Unreal Engine” and “Photon”, as it is widely compatible with a diverse range of platforms, the visuals produced can adapt to different devices and screens without compromising quality and there is extensive documentation and tutorials that are available for new developers to learn quickly.

3 LITERATURE REVIEW

The chapter in discussion highlights key technical aspects and areas of domain expertise which are potentially vital to obtaining a fair understanding of the concepts and knowledge required by the project, as learned from reviewing certain relevant works of literature.

3.1 Introduction

The existing body of literature focusing on simulating the operations of light railway networks was examined during the first phase of our project.

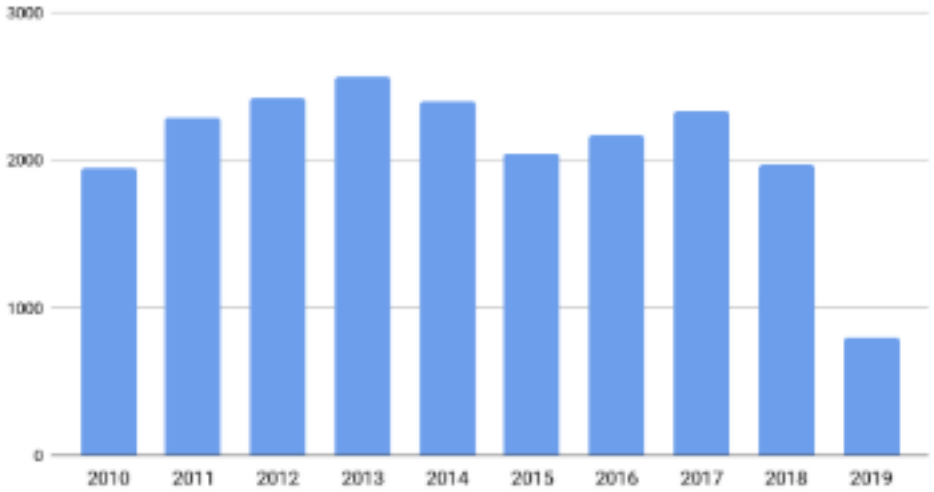


Figure 4: Published Papers on Railway Network Simulation

As evidenced by figure 4 above, a considerable volume of research papers has been published on the subject of the topic of 'Railway Network Simulation' worldwide. Based on the frequency of citations, a curated selection of papers pertinent to our use cases was chosen and reviewed in detail. It was discovered that three broad approaches to railway network simulation are well documented, these include - utilisation of statistical techniques and models, applying machine learning models, and using the open-source software 'OpenTrack'.

3.2 Analytical Models

Before considering more technologically driven approaches, the presently prevalent analytical methods that are being utilised to study rail networks were researched first. Sahin (2017) evaluates the efficacy of using Markov Chains to model "disruptions and disturbances". Viewing the departure and arrival times as probabilistic processes allows the prediction of steady-state delays. A physics-based approach to predicting rail timings using mechanical properties such as weight of train, acceleration, drag, et cetera was also examined (Goodman et al., 1998). However, this

research did not yield promising results. The feasibility of modelling the light-rail network as Space-Time network (Dessouky & Leachman, 1999) was also evaluated and discounted.

3.3 Machine Learning Models

The past three years have witnessed a dramatic rise in the application of Machine Learning (ML) technology to a myriad of use-cases. Railway simulation has also witnessed a concomitant increase in the application of ML techniques to achieve higher prediction accuracy than traditional approaches. Ostensibly uncorrelated factors such as weather and temperature can offer insight into predicting train network delays (Wang & Zhang, 2019). This approach may offer positive collateral benefits, such as a lower operating cost. A closely related approach is to train multiple mini-models for every station, instead of one global model (Zhang & Nguyen, 2013). However, mini-models can only be trained if access to highly granular (station-level) data is available.

3.4 OpenTrack

Started in 1990 as a research project at the Federal Institute of Technology, Switzerland, OpenTrack is now the gold-standard software package utilised by many global rail networks to plan routes, model scenarios and simulate crowd movement (Nash & Huerlimann, 2004). However, this software has primarily been used to evaluate infrastructure components such as crossing throughput and side channel efficiency. The MTR Corporation makes use of this software and the project team is in ongoing negotiations regarding access to this platform for gaining better understanding of the KTL structure.

3.5 Miscellaneous

Dessouky and Leachman (1995) in their paper discussing simulation modelling in complex rail networks highlight the development of time and event based train models for simulation of rail traffic in a likely realistic manner and how such models may be utilised to analyse delays and domino effects from mishaps in a railway network. One may infer from this research paper that a simulation model should take into account factors such as the time of day and day of the week to predict train delay. For instance, there may be more passenger traffic during the morning hours of weekdays as people rush to work, while on weekends there may be increased passenger traffic during the evening as friends and family go for an outing. This theory may further be advanced by taking into consideration factors such as consumption of energy, traction and speed of the train (Goodman, Siu and Ho, 1998).

The Nanjing Metro Line in China serves as a fair case study example, taking into account details pertaining to the train, track, controller and power supply (Wang and Cheng, 2012) which may be modelled in a queue based context through machine learning algorithms (Zhang, Nguyen and Zhang, 2013). Yaghini, Khoshraftar and Seyedabadi (2013) also discuss the application of a neural network based structure with a relatively high accuracy compared with other options such as decision trees and logistic regression models. These works of literature further encourage the use of "Bayesian Networks" as mentioned earlier in this paper for building a recommendation system, since Bayesian Networks are a neural network based structure and a type of machine learning algorithm.

3.6 Summary

In sum, it may potentially be stated that building an accurate simulation model requires use of several complex factors including but not limited to the train's features, track features, power supply etc. in order to create a realistic time series model, through means of a neural network based algorithm.

Based on the literature review, a multi-pronged approach has been adopted. Since conventional analytical models are not applicable to this use case, the project team has maintained a focus on Machine Learning Models and OpenTrack. Work is underway to recreate a global-local model approach (Zhang & Nguyen, 2013). Additional data is required to implement this model and has been requested from MTR. An application has also been made to the Computer Science department for the procurement of OpenTrack software for the team.

4 RESULTS AND DISCUSSION

The following chapter discusses certain initial findings at the current stage of the project, detailing the observations and inference derived from these findings. Followed by a discussion of various risks and challenges that have been encountered.

4.1 Initial Findings

Due to reasons pertaining to confidentiality, there are certain restrictions on information that can be shared with third parties, with a view to protect the MTR's proprietary data and functioning methodologies. Thus, certain findings and factors may not be disclosed.

In order to simulate train operations, running times and dwell times derived from the datasets are used to calibrate and validate the model. The ultimate goal of the development of model is to reproduce the behaviour pattern of the railway system in the past, and predict the effect of different types of events on the operations. While some related works computes running times and dwell times using train motion equations and traffic conditions, a data-driven approach is proposed in this project, since a large amount of track occupancy data is provided by MTR.

The statistical approach captures the distribution of running times and dwell times due to external conditions without manual tuning of parameters, and does not require the knowledge of train motion characteristics. It is also more adaptive for future changes in timetables and extensions to other lines. The terminology outlined earlier in the report before the introductory chapter lists the input variables of simulation models selected according to findings in exploratory data analysis

4.2 Data Pipeline

A robust data processing pipeline is the backbone of effective data analysis, which streamlines the iterations of algorithms and update of datasets. In our case, train operation logs received from MTR are encoded in a special format produced by the log management system. The pipeline currently includes transformation of the special format into CSV format, and combination of logs partitioned by hours. An automation toolset called doit is used to describe the tasks involved with their dependencies, similar to a Makefile. The generated dependency graph ensures only the necessary tasks are executed when datasets are updated. Parallel execution of tasks also increases the efficiency of the pipeline, resulting in a shorter feedback loop. In the current stage, the pipeline is

used for data cleaning and preprocessing. Training of simulation and recommendation models can be integrated when the design is finalised.

4.3 Data Analysis

As part of the project's objective, a data analysis has been conducted to find any relevant patterns and it has been discovered that on average there is a significant standard deviation in the average travel time of trains on KTL, of about 26.10, thereby suggesting that the simulation model may be required to take into account any deviation from the timetable which takes place and also factor in any cumulative delay which may be caused by a train suffering from disruption, leading to other trains also being delayed. Furthermore, the travel time of a train through a station on KTL, as illustrated in figure 5 below, including the time taken to decelerate, dwell time and accelerate is also varying significantly, with the lowest average time for a station being 91.35 seconds and the highest being 210.91 seconds. This likely indicates that the model may also need to factor in the station at which the train is arriving.

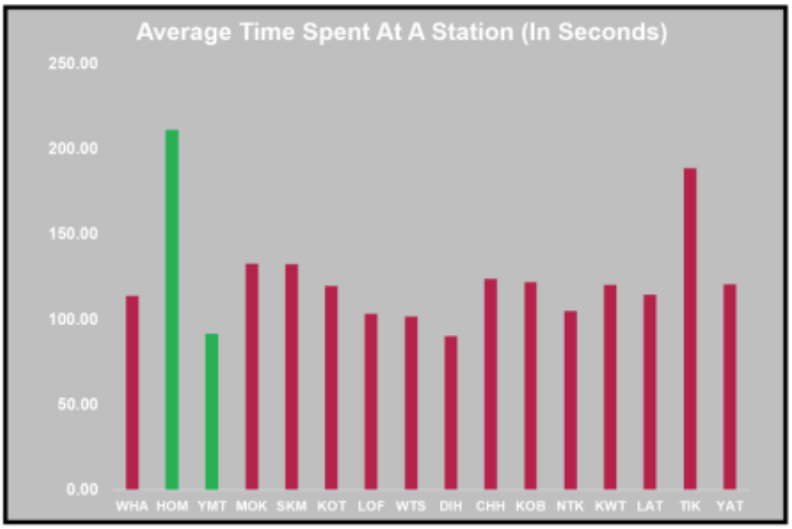


Figure 5: Average Time Spent at a Station

Post the aforementioned data analysis, a log file containing the travel time of various trains and stations over a 2-day period on Sundays in August was simulated using Unity, showcasing the movement and arrival of multiple trains at their respective destinations, while also providing a dashboard which displayed the logic and reasoning behind various actions, such as removal of a train from a track due to a disruption or introduction of a new train into KTL to meet passenger requirements, to likely provide greater transparency into the decision making rationale of the simulator, to the user.

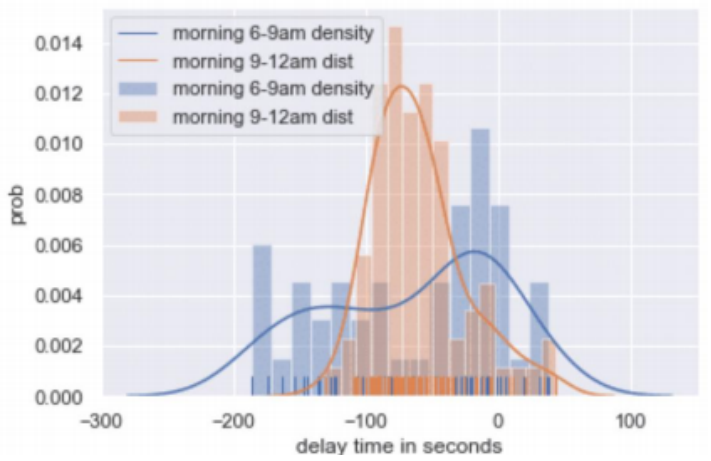


Figure 6: Train Delay Distribution on Sunday Morning

Figure 6 above shows that train movements are time-dependent, as train delays in different periods of a day vary significantly, which is likely to be affected by different train schedules and fluctuations in passenger demands. This finding supports the conclusion that time and location are two key factors in designing the simulation model.

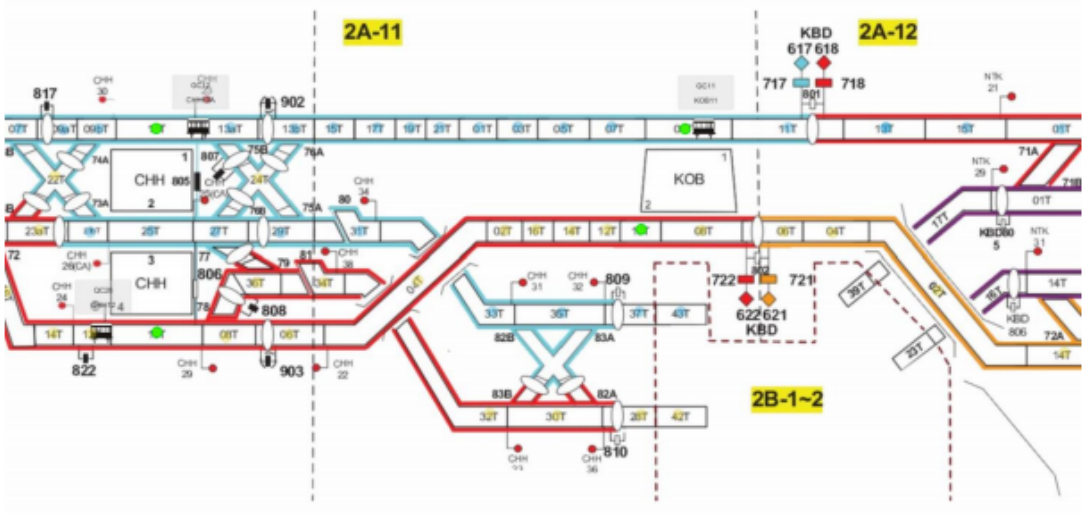


Figure 7: Kwun Tong Line Track Layout

The implementation of the visualisation module has been completed, as shown in figure 7 above, is now able to reproduce the train movements of KTL from the log files received from the MTR. It serves as a platform to present the project's findings in an intuitive manner, and uncover operational patterns of trains when they are located in terminal stations. For instance, apart from regular train service, the timetable also specifies some special trains during high traffic periods such as public holidays, that are aligned with our observations of operational logs. The operational patterns reveal the need to integrate the timetable into the simulation model to reflect the differences in train schedules.

4.4 Machine Learning Model

Due to a lack of sufficient data, a statistical-inference driven approach has been adopted to derive key insights. Based on our findings thus far, it is estimated that at least 4 – 8 weeks of daily ridership data will be required to accurately simulate the KTL train movement. This is further corroborated by the variance of the data fields, even within the small sample provided, that were highlighted in the previous section focusing on data analysis. One parameter, Average Dwell Time is depicted in Table 1 below. This is an important metric that needs to be accurately predicted.

Station Code	Dwell Time (s)
WHA	110
HOM	209
YMT	88
MOK	125
SKM	129
KOT	122
LQF	85
WTS	92
DIH	70
CHH	73
KOB	82
NTK	112
KWT	200
LAT	54
TIK	193
YAT	101

Table 1: Average Dwell Time of Stations

In spite of these issues, the development of the models has been initiated. Until actual data becomes available, user-generated dummy data has been used as model input, to enable the structural development of the models. The downside of this approach is the limited ability to evaluate the performance of the model. Based on the team's consideration and negotiations with the MTR, this is an acceptable trade-off.

Type of Regression

Station	Mean Squared Error (in seconds)	Linear Regression	Random Forest	Regression Tree + Bagging	Regression Tree + Boosting	Average
	WHA	35.363	30.1603	35.630	28.649	32.451
	HOM	8.261	8.33682	9.0934	8.361	8.513
	YMT	11.641	11.7592	11.6534	11.671	11.681
	MOK	6.411	6.178	6.307	6.017	6.228
	SKM	3.580	3.277	3.282	3.239	3.344
	KOT	5.716	5.284	5.166	5.215	5.345
	LOF	4.359	4.343	4.439	4.240	4.345
	WTS	4.489	4.365	4.334	4.339	4.382
	DIH	5.296	5.386	5.301	5.382	5.341
	CHH	4.367	4.285	4.586	4.329	4.392
	KOB	6.499	7.240	6.396	7.043	6.794
	NTK	4.325	4.224	4.093	4.179	4.205
	KWT	14.733	6.324	6.422	6.337	8.454
	LAT	5.279	5.364	5.270	5.349	5.316
	YAT	10.392	10.540	10.737	10.587	10.563
	TIK	34.311	23.206	33.661	21.423	28.150
	Global	23.383	18.309	23.029	17.636	20.589
	Average	11.083	9.328	10.553	9.059	10.006

Worst Best

Table 2: Performance Analysis of Various Models

Table 2 summarises the results of predictions in terms of mean squared errors in seconds, using local models trained for each station, and a global model for the whole KTL. Comparing mean squared errors from prediction using different models, the performance are similar in general, with linear regression being the least accurate, and boosting with regression tree being the most accurate. Experiments with larger dataset and other ensemble modeling approaches will be carried out to select the optimal model. It is also observed that errors for terminal stations (i.e. WHA and TIK) are larger than other stations, which highlights room for improvement using more specific models.

4.5 Limitations and Difficulties Encountered

The project involves certain ambiguities and new skills that need to be learned, for instance, in order to gain a good understanding of the datasets, one needs to have obtain the necessary domain knowledge and a sound understanding of the concepts. As highlighted in table 3 below, there is a high probability of there being technical terms and concepts that are new to the members of the team, for instance, “dwell time” refers to the time that a train waits at a particular station, to allow passengers to onboard and offboard. Such key pieces of information may be learned by interviewing train operators, reviewing existing literature and reaching out to experts who are familiar with such concepts.

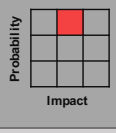
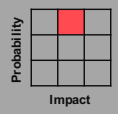
	Risks & Challenges	Impact-Probability	Mitigations
1.	**Lack of Domain Knowledge**		• Review existing literature on relevant topics • Reach out to experts to gain knowledge • Interview train operators
2.	**Limited Access to Internal Data of MTR**		• Negotiate for access to necessary data • Understand the various data requirements
3.	**Exploratory Nature of Project**		• Strong emphasis on teamwork • Maintain flexibility • Regular communication and feedback • Agile approach to software development

Table 3: Risk Matrix

Another key challenge is gaining access to data, since a lot of the data is confidential and proprietary, hence the team needs to be aware of what forms of data are necessary and negotiate with the MTR officials for the required access. For example, data on the distance between various stations is required in order to calculate the required speed and time for travelling. Given that the students involved are new to the field of railways, the project is exploratory in nature, this poses a key risk which can be overcome through an emphasis on teamwork, maintaining flexibility, regular communication and adopting an agile approach to the development process. These are some of the main challenges that may be faced by the team, during the course of this project, certain other challenges, such as multitasking, ability to meet deadlines etc. may also pose a risk.

5 FUTURE WORK

The figure attached below, showcases a “Gantt Chart” outlining the proposed software development cycle. The project is broadly partitioned into three parts. In Phase 1, from August to September, exploratory data analysis was carried out on a dataset provided by the MTR to understand the possible extent of the problem being faced. Following which, during the next phase from October to January, major components of the simulator may potentially be implemented, with a simulation model validated by actual data. In the final phase planned, from February to April, the implementation of the simulator will likely be finalised, with comprehensive testing and validation. A prototype of the recommendation engine is also expected to be developed.

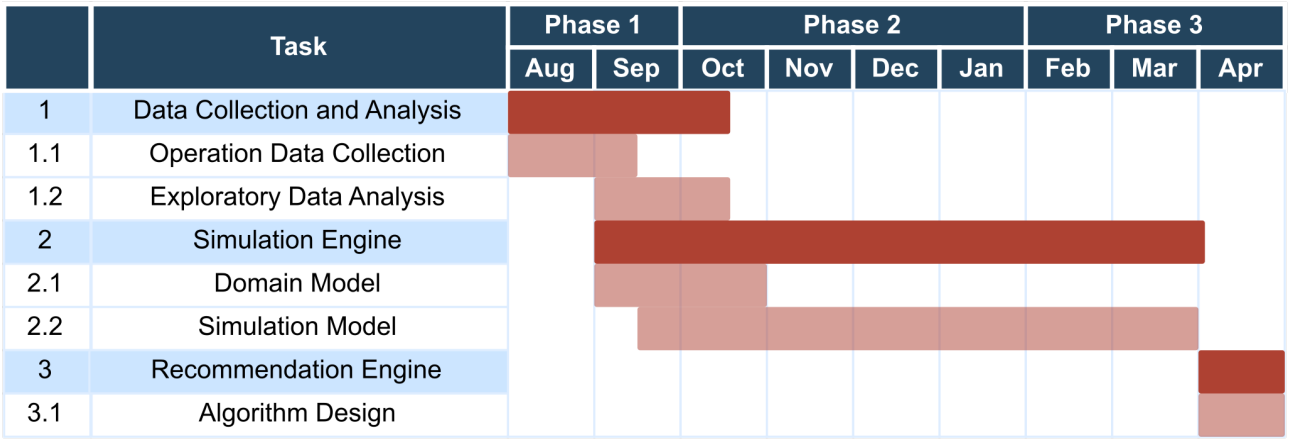


Figure 8: Timeline

The key milestones of the project include creation of a visualised simulation model, a software/ algorithm for providing recommendations and linking the simulation model with the recommendation system to provide real-time recommendations. In future, the project may also lead to opportunities in automation of the railway network, such as driverless trains and automatic timetable generation.

6 CONCLUSION

The project intends to attempt the creation of a visualised simulation model, showcasing complex pieces of information, and also a recommendation system which could provide relevant advice on handling a wide variety of disruptions which may cause passengers loss of valuable time. At present, challenges such as a lack of technical knowledge in the rail transportation industry and certain limitations on access to internal data of the MTR have surfaced. To tackle these constraints, steps such as interviewing experts, reviewing relevant pieces of literature and actively negotiating with the MTR have been undertaken. The currently undergoing second phase has involved the creation of a simulation model, visualised using Unity software and the data provided by the MTR. With rigorous analysis of data obtained thus far and in the future, the third phase will involve generating useful recommendations which can lead to likely more informed decisions.

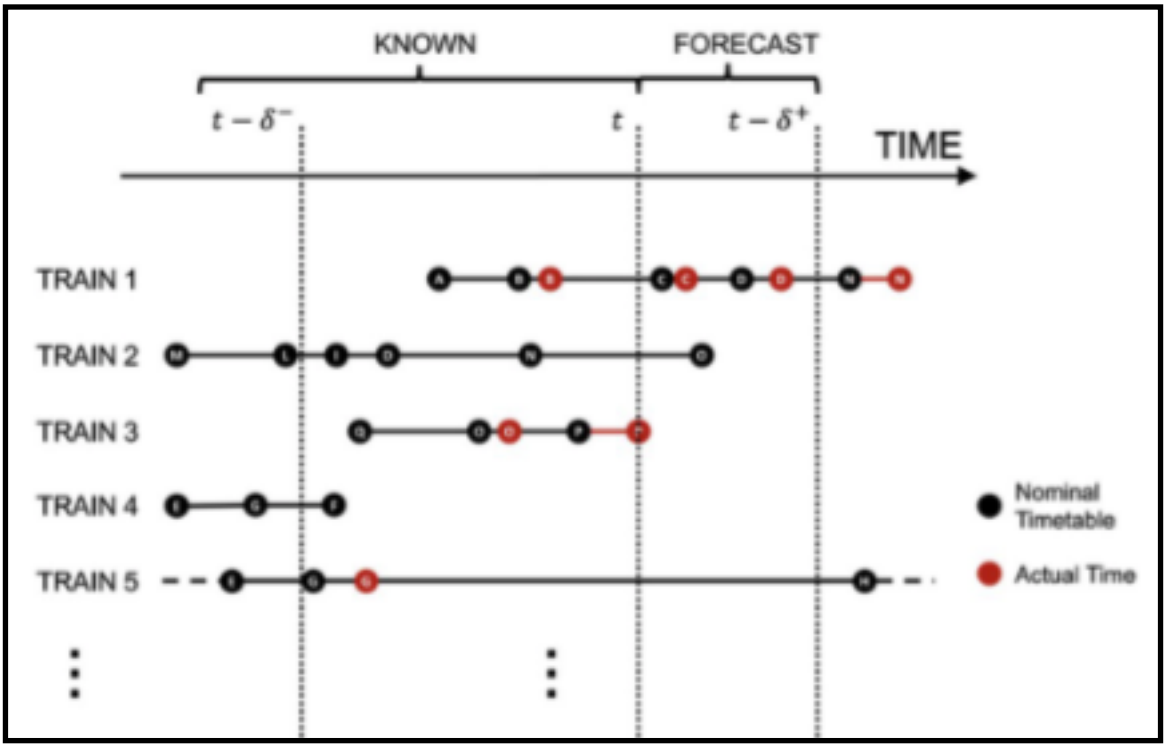
7 REFERENCES

1. Chung, N.K. (2017). *MTR apologises for recent Hong Kong service delays and promises to invest in system maintenance*. South China Morning Post. Retrieved from https://www.scmp.com/news/hong-kong/economy/article/2106320/mtr-apologises-recent-hong-kong-service-delays-while
2. Cong, G., Fan, W., Geerts, F., Jia, X., & Ma, S. (2007). *Improving Data Quality: Consistency and Accuracy... Proc*. Int'l Conf. Very Large Data Bases (VLDB). 315-326.
3. Corman, F., & Kecman, P. (2018). *Stochastic prediction of train delays in real-time using Bayesian networks*. Transportation Research Part C: Emerging Technologies, 95, 599-615.
4. Dessouky, M.M. and Leachman, R.C. (1995). *A Simulation Modelling Methodology for Analysing Large Complex Rail Networks*. Simulation: Transactions of the Society for Modelling and Simulation International.
5. Elgabli, A., Khan, H., Krouka, M., & Bennis, M. (2018). *Reinforcement learning based scheduling algorithm for optimizing age of information in ultra reliable low latency networks*. arXiv preprint arXiv:1811.06776.
6. Goodman, C.J., Siu, L.K. and Ho, T.K. (1998). *A review of simulation models for railway systems*. International Conference on Developments in Mass Transit Systems.
7. Hao, W., Meng, L., Corman, F., Long, S., & Jiang, X. (2019). *A train timetabling and stop planning optimization model with passenger demand*. In Rail Norrköping 2019. 8th International Conference on Railway Operations Modelling and Analysis (ICROMA), Norrköping, Sweden, June 17th–20th, 2019 (Vol. 69, pp. 390-406). Linköping University Electronic Press.
8. Ismail, S. (2017). *Journal of Rail Transport Planning and Management*. 101-113. Elsevier Publishing.
9. Jespersen-Groth J. et al. (2009). *Disruption Management in Passenger Railway Transportation*. In: Ahuja R.K., Möhring R.H., Zaroliagis C.D. (eds) Robust and Online Large-Scale Optimization. Lecture Notes in Computer Science, vol 5868. Springer, Berlin, Heidelberg.
10. Kecman, P., & Goverde, R. M. (2015). *Predictive modelling of running and dwell times in railway traffic*. Public Transport, 7(3), 295-319.
11. Legislative Council Secretariat. (2017). *MTR Train Service Performance*. Retrieved from https://www.legco.gov.hk/research-publications/english/1718issh07-mtr-train-service-performance-20171220-e.pdf

12. Mannino, C., Lamorgese, L. C., & Piacentini, M. (2017). *Integer Programming Techniques for Train Dispatching in Mass Transit and Main Line. Advances and trends in optimization with engineering applications*.
13. Mass Transit Railway Corporation. (2019). *MTR Train Service Performance*. Retrieved from http://www.mtr.com.hk/en/customer/main/MTR-train-service-performance-jan-2019.html
14. Mass Transit Railway Corporation. (2019). *MTR 40th Anniversary*. Retrieved from http://www.mtr.com.hk/en/customer/main/brand-story.html
15. Nash A. & Huerlimann D. (2004). *Railroad simulation using OpenTrack*. Swiss Federal Institute of Technology.
16. Rosenfeld, M. (2018). *Practical Applied Statistics for Sociologists*. Stanford Publishing.
17. Wang P. & Zhang Q. (2019). *Train delay analysis and prediction based on big data fusion*. In: Transportation Safety and Environment, vol 1. Oxford.
18. Wang, W. and Cheng, M. (2012). Simulation of Nanjing Metro Line 1 Using Metro Simulator. School of Electrical Engineering, Southeast University, Nanjing 210096, China.
19. Wong, O. (2018). *MTR delays hit Hong Kong passengers for second time in October*. South China Morning Post. Retrieved from https://www.scmp.com/news/hong-kong/article/2170088/mtr-delays-hit-hong-kong-passengers-second-time-october
20. Yaghini, M., Khoshraftar, M.M. and Seyedabadi, M. (2012). *Railway passenger train delay prediction via neural network model*. Journal of Advanced Transportation.
21. Zhang, Y., Nguyen, L.T. and Zhang, J. (2013). *Wait Time Prediction: How to Avoid Waiting in Lines?* ACM Conference 2013 on Pervasive and ubiquitous computing adjunct publication.

APPENDICES VII

Appendix A Delay Time Prediction Chart



Appendix B Basic Track Layout Model