Study of rail public transportation network: demand analysis and forecasting

Submitted to
Graduate School of Information Science and Technology Osaka University

January 2024

Tissawat ASAVANANT

List of achievements

Authors	Title	Journal/Proceedings
Tissawat ASAVANANT and Hiroshi MORITA	Analysis of demand variability in public transportation planning	PROCEEDINGS OF THE SCHEDULING SYMPOSIUM, 2022
Tissawat ASAVANANT and Hiroshi MORITA	Short-term passenger demand forecasting from real-time data collection”	PROCEEDINGS OF THE SCHEDULING SYMPOSIUM, 2023
Tissawat ASAVANANT and Hiroshi MORITA	FORECASTABILITY ANALYSIS FOR PASSENGER RAIL NETWORK DEMAND DURING COVID-19	SCIENTIAE MATHEMATICAE JAPONICAE, 2023
Tissawat ASAVANANT and Hiroshi MORITA	Short-term origin-destination demand forecasting in rail transit systems: parallel model architecture and gravity approach	International Journal of Mathematics in Operational Research, 2024

Acknowledgments

I would like to acknowledge and give my warmest thanks to my supervisor Professor Hiroshi Morita who made this work possible. His guidance and advice carried me through stages of writing my project. I would also like to thank my committee members for letting my dissertation defense be an enjoyable moment, and for your kind and helpful comments and suggestions, thanks to you.

I would also like to give special thanks to my friends, colleagues, and my family for their continued support and understanding when undertaking my research and writing my project.

Abstract

This study focuses on the usage of IC card data of the metropolitan public transportation (MPT) system to forecast short-term ridership demand based on the ridership's behaviour. The MPT network is a complex structure, serving as the paths for city-wide travel and logistics of a city. MPT network stretches across the city like a net or web. Modifications to the MPT is regarded as very long-term projects due to limitations subjected upon the existing infrastructures. Such limitations include installations covering large area within highly populated areas, allocating traffics of people and vehicles without inflicting damages to transportation capability and nearby infrastructures, and lastly, the capital needed for running modification projects. For these reasons, maximization of utility of existing networks is the most important aspect of the MPT network study.

The utility of the transportation network, when generalized, is explain by three simple components: demand (ridership), supply (capacity), and network constraints (schedule, network connectivity, etc). The combination of these three components forms a spatial-temporal problem with transport restrictions and limitations. Commonly the demand component is derived from the IC card data, also known as Origin-destination matrix (ODM). The task of estimating the ODM is, however, difficult since it is a human induced event that is sequential in a network defined with both spatial and temporal variables. Transportation planners rely on expertise on their specific MPT network to make inference on the network's unique characteristics to design utility maximization methods. Understanding the available data is thus crucial for accurately estimating the ODM in different conditions. To achieve forecasting accuracy within the acceptable margin of error, forecasting model must be tailored to the data.

In this thesis, IC card data from subway network of Bangkok city is processed and analyzed. The result of the analysis indicates that the typical characteristics of ODM during the study period misaligned with the assumptions used in ODM estimation literature. Proportionality and normality are found after heavy processing of the data. Historical average (HA) is the baseline model which is obtained using statistical mean after partitioning data into multiple separated sets according to their temporal states. Additionally, multiple distributions were found within the assumed stochastic process. To obtain a satisfactory forecast, we redefined the gravity model into multivariate model such that the relationships between variables are traceable and well-defined. We found that partitioning data into multiple clusters and applying different models with Parallel model architecture (PMA) improves the forecast significantly. Lastly, we subject the forecasting problem with real-time data availability restrictions by defining multiple forecasting cases. The analysis and forecasting resulted in satisfactory error level of 26.14%, significantly lower than existing forecasting methods. Comparison to other methodologies is done qualitatively since the suitability of their models are inapplicable to the proposed methodology.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The MPT system refers to the public transportation system located inside an area of a significantly densely populated zone. MPT system operates following a decided path with or without schedule of the vehicles as to answer the transportation needs of the population travelling within their designated areas. City-wide infrastructure planning is conducted under constraints of existing infrastructures. The main objective of the MPT system is to provide a cost-effective travelling option to the population and alleviate the congestion within a populated area. City-wide transportation network consists of multiple transportation system including personal vehicles, taxis, buses, trains, and pedestrians. Compared to less populated areas, the MPT is more accessible within the urban area and the accessibility creates a complex relationship between the ridership and the travelling options. Socioeconomic level of a specific area also plays a significant role in travelling behaviour and thus, in a macroscopic scale, resulted in highly complex social problems.

On a microscopic scale, each system within a certain transportation network can be defined. Whether a specific MPT system can or cannot be defined separately from the network depends on the characteristics of that specific system. For example, personal vehicles, taxis and buses, all share street to traverse. Street transportation can thus be defined into segments only when the observed data and the objective of the study matches the segment definition. For example, when schedule delay is not considered, a bus system can be well defined, while if the schedule delay is included due to the objective of the study, inclusion of other vehicles sharing the paths is unavoidable. For rail transportation, due to the separation of the travelling paths from street transportation, can be considered as a separated system on its own since its relation to the other systems is negligible.

The transportation problem consists of a combination of three simple components: demand, supply, and constraints from the characteristics of a considered system. The demand of the transportation problem is defined as the ridership needs for using a set of paths to move between a pair of origin and destination. The supply is defined as the capacity of a specific system to accept the ridership. What differentiate a transportation system within a network is the last component, the constraints. The constraints are the system-specific characteristics/limitations imposed by the design. The constraints can be many things like cost of travelling, needs of vehicles and driving permits, travelling time, accessibility of the system, etc.

The utility maximization is defined as the adjustment of supply in accordance with the transportation constraints to fit to the ridership demand and optimize a specific objective function. The demand used in the transportation problem is mined from the historical data. The demand's volatility is usually undermined due to the difficulty in modelling a realistic human-induced needs. While a stable flow of demand is well-represented using historical data, the complex transportation network within a densely populated area is highly affected with a significant deviation of the demand since the balance of the demand-supply is easily broken. The form of imbalance can be the congestion increasing the standby ridership that needs recovery for the congestion to not cause operation delay. An overflow of demand within segments of a specific travelling path is a common occurrence in metropolitan areas.

The adjustment of supply is therefore, a common mean to solve the transportation problem. However, understanding the demand of the transportation network/system is necessarily to proactively deal with the possible imbalance of transportation demand-supply relationship. As the deviation of demand is abrupt and complex. There is a need of forecast of demand within a realistic scope.

1.1 Rail-transportation

Rail-transportation is a type of public transportation system (PTS) that is highly cost-effective due to its huge capacity of handling passengers and its relatively short travelling time from using rails as their specific travelling path. Even amongst the many PTSs, rail transportation proves its usefulness within the history of mankind in logistics. However, the cost of operation of rail transportation is high and constant.

In metropolitan areas, rails transportation systems are added to city's infrastructure. Due to limitations from existing infrastructures, it is common for the rail transportation systems to be set above or below ground level. In addition, stations, relay stations, and location for storing/repair trains are needed to provide stops for passengers, move the trains in case of emergencies, and when conducting maintenance on the vehicles. Thus, to provide wide coverage to the transportation without interfering with the existing infrastructures, rigorous planning is required and due to the large capital needed for the project, the lifetime of rail transportation system is long. Trains are the largest and the highest maintenance known land transportation vehicles.

The PTS system of interests in this thesis is the subway system, an underground rail transportation network within an urban area. While considering the cost-effective mode of transportation, the weakness of the rail-transportation is its reliance on rail. The entry-exit points of a rail transportation are, when compared to

other PTS, spatially wide. So, commonly it is a transportation mode choice for ridership that needs to travel a long distance and uses it as in-between transportation path, or when the destination is located near the station. So, considering the connectivity of a rail network with a limited area coverage, ridership may opt for alternative transportation mode instead.

This rigidity (reliance on rail) is why the rail transportation is distinctly different from other mode of PTS. Thus, it can be expected that the demand of the rail transportation network may have its own unique characteristics. Nevertheless, we are considering the context of passenger's transportation and not the freight transportation due to higher expected volatility and its limitation to the scheduling of the passenger rail transportation.

. Commonly, the rail transportation scheduling is decided with a parameter called "maximum headway". Maximum headway is the organizational parameter in spacing the vehicles along the rail. There are multiple objectives from the perspective of ridership to decide the maximum headway (i.e., comfort level, passenger density per area, or peak/non-peak period). A common strategy for optimizing the utility of the rail transportation is to locate an imbalance of ridership's origin-destination. The general imbalance on the context of rail transportation is the bi-directional imbalance. Focusing the utility of the vehicles on a specific segment along the full path lowers the number of required fleet size and consequentially minimize the operation cost and improve demand-supply balance. Thus, it is crucial to be able to sufficiently forecast the future demand of the rail transportation.

1.2 Origin-destination prediction

Historical data that represents rail transportation demand is the ODM. An ODM is defined to be the ridership matrix within a certain defined temporal time window. Under normal situation, a statistical average ODM is sufficient to represent the demand due to the scheduling of the rail system. The capacity of the normal operation of a rail transportation system is loosely defined within a similar time window. The deviation under normal operation, in many cases, can be handled with the next train arrival and thus, is a good enough estimate for planning a general operation of a rail transportation.

Asides from the general planning, we can categorize the rail transportation problems into simple short-term and long-term problems. Long-term problems, in similar nature to the general planning requires single demand representative of ODM. The prediction of the ODM for long-term problems involve statistical, distributive, or static model that give a rigid estimation. These estimations are deterministic in nature.

For the short-term context, available data is used to make a forecast within a harsh limitation of time and is usually not within the range of the historical data. These limitations make short-term predictions difficult because the understanding of ODM characteristics, relationships between temporal and spatial factors, and many unknown variables must be taken into account before the prediction problem can be defined. Additionally, a criterion of the acceptable margin of error under the context of prediction is generally unique to the system of interest.

1.3 Problem statement

In this thesis, the target of the prediction is the ODM in the earliest predictable interval using historical information and trip generation as the inputs. The time interval is defined as 60 min in this study, same time frame as the system of interest's time interval of estimated capacity. The ODM *M* and trip generation *N* are extracted from IC card data and are defined in Eqs. (1) to (3). The aggregation of the IC card data is entry-based. The framework of the forecasting problem is explained in detailed in Chapter 4.

$$M^{d,t} \in R^{n\times n} = \begin{bmatrix} m_{11}^{d,t} & m_{12}^{d,t} & m_{13}^{d,t} & \cdots & m_{1j}^{d,t} \\ m_{21}^{d,t} & m_{22}^{d,t} & m_{23}^{d,t} & \cdots & m_{2j}^{d,t} \\ m_{31}^{d,t} & m_{32}^{d,t} & m_{33}^{d,t} & \cdots & m_{3j}^{d,t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{i1}^{d,t} & m_{i2}^{d,t} & m_{i3}^{d,t} & \cdots & m_{ij}^{d,t} \end{bmatrix} \quad (1)$$

$$N^{d} \in R^{n\times t} = \begin{bmatrix} n_{1}^{d,1} & n_{1}^{d,2} & n_{1}^{d,3} & \cdots & n_{1}^{d,t} \\ n_{2}^{d,1} & n_{2}^{d,2} & n_{2}^{d,3} & \cdots & n_{2}^{d,t} \\ n_{3}^{d,1} & n_{3}^{d,2} & n_{3}^{d,3} & \cdots & n_{3}^{d,t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n_{i}^{d,1} & n_{i}^{d,2} & n_{i}^{d,3} & \cdots & n_{i}^{d,t} \end{bmatrix} \quad (2)$$

$$\begin{bmatrix} n_{1}^{d,t} \\ n_{2}^{d,t} \\ n_{3}^{d,t} \\ \vdots \\ n_{i}^{d,t} \end{bmatrix} = \begin{bmatrix} \sum_j m_{1j}^{d,t} \\ \sum_j m_{2j}^{d,t} \\ \sum_j m_{3j}^{d,t} \\ \vdots \\ \sum_j m_{ij}^{d,t} \end{bmatrix} \quad (3)$$

where $m_{ij}^{d,t}$ is the OD pair from station *i* to station *j* in time interval *t* on day *d*. $\sum_j m_{ij}^{d,t}$ is the sum of OD pairs that enter station *i* in time interval *t* on day *d*. $n_i^{d,t}$ is the trip

generation from entering station *i* in time interval *t* on day *d.* Note that in our problem definition, we do not assume that $n_i^{d,t} = \sum_j m_{ij}^{d,t} = \sum_i m_{ij}^{d,t}$ [1], or more specifically, we do not establish trip generation/trip attraction due to data availability delay from unobservable ODM due to the incompleteness of the trip of the entry-based aggregation. The station IDs are replaced with numbers according to their location on the track in order, from inbound to outbound.

The most recent available real-time OD matrix is decided by the greatest integer function in Eq. (4):

$$t' = \left\lfloor t - \frac{q}{v} \right\rfloor \qquad (4)$$

where time interval t' is of the most recent available real-time OD matrix when forecasting the OD matrix at time interval t, q is the maximum trip duration time, and v is the range of the time interval. Eq. (4) is defined to realistically design forecasting cases. In addition, we also denote the time interval gap $t - t'$ as follows:

$$\theta = t - t' \qquad (5)$$

θ dictates the earliest available forecast target from the start of the operation and is vital to the formulation of the chained forecasting case. Limitations are placed on the earliest forecastable time interval. However, forecasts which use previous days' ODMs of the same time interval $M^{d,t} = f\left(\{M^{d-x,t}\}_x, \{N^{d,t-y}\}_y\right)$ with $x = 1,2,3,\dots; y = 1,2,3,\dots$ [1], and N^d are derived directly from $M^{d,t}$ (Eqs. (1) to (3)), which can cause an overlap of inputs. Thus, with the interval gap from Eq. (5) in addition to the condition of $y < x$, the forecast can be expressed as

$$M^{d,t} = f\left(\{M^{d,t-x}\}_x, \{N^{d,t-y}\}_y\right), x = \theta, \theta+1, \theta+2, \dots; y = 1,2,3,\dots; y < x \qquad (6)$$

where $M^{d,t}$ is the ODM in time interval *t* on day *d.* While prior studies generally used the ODM in the last several time intervals as model inputs for short-term forecasting [2], we propose a trip generation-OD matrix input for realistic forecasting.

Chapter 2: Literature Review

2.1 Origin-destination matrix

Three generalized components of any transportation networks are demand, supply, and network's constraints. For subway networks, the demand of the subway is the ridership represents with ODM, commonly serves as the input for trip planning and other transportation problems [3], The supply is the capacity within a given time of the network interlinked with the network constraints, and the network constraints are the timetable, network connectivity given certain strategic scheduling.

In this thesis, "OD matrix" is used to refer to a demand matrix at a certain time interval. Short-term OD pair prediction has three characteristics: 1) data availability: real-time OD pairs are unavailable during the operation (delayed data availability); 2) data dimensionality: the dimension of the OD data is much higher than the cardinality of transportation networks; and 3) data sparsity: OD data are spatiotemporally sparse [1]. Although boarding and alighting demands at metro stations have received much attention in various studies, the short-term demand of OD pairs of a rail network is still relatively underrepresented in research [4].

Additionally, anomalous scenarios study uses representative static ODM with the addition of abnormal data to make inference to the deviation from the standard operation, the representative static ODM can be derived using HA or statistical distributive estimation [5], [6]. Nevertheless, given the short life of the trained model, such representative static ODM will outlive its usefulness due to the gradual/sudden shift of the demand. In a fortunate situation, such shift will be minimal relatively to the volatility of the demand and will still be a good enough representation until the static ODM is updated. However, at the start of 2020, a significant change in ridership was observed [7], [8], and has continued to persist for more than three years.

Completely observed ODM are readily obtainable in a rail network with two-way automated fare collection (AFC) or a trip-chaining system [9], [10]. The aggregation of trips made by users is either entry-time-based or arriving-time-based. For a one-way AFC, the recorded time is the time of either entry into or exit from the network. The starting-time-based method provides observations of destination choice demand. Demand is also commonly simulated with a multinomial distribution or multivariate normal distribution although the volatility is commonly underestimated [11]–[13].

2.2 Usage of origin-destination matrix

In this section, we reviewed the usage of ODM in the transportation research to identify common necessities that is desirable in ODM as per their usage. Congestion measurement and management are not included in the review due to the multiple types of congestion dependent on the definition given to it in accordance with a given network. Additionally, congestion management is beyond the scope of this thesis. The usage of ODM in transportation research can be classified into three general categories.

The first category is the ODM forecast. In this category, the common point is the usage of representative static ODM. The representative is used to model the relationship between certain variables to the deviation, thus resulting in an adjusted ODM. Changes in trip frequency, routes, location, and capacity were used with piecewise linear approximation to obtain adjusted ODM of freight train delivery [14]. Trip boarding and alighting were used with a recursive bi-proportional model to add constraints to the adjusted ODM of bus network [15]. ANN with randomized seed matrix are used with network capacity, physical location, and timetable headway to estimate the rail's ridership [16]. Licensed plate matching and sampling ratio were used with Bayesian estimation for making inference of the route usage [17] .

The second category is the demand-supply matching. In this category, the cost of operation is directly minimized to the level of demand using geographical data. There are many types of objective functions for the demand-supply matching problem, but in general, the objective is considered from either the perspective of consumers or suppliers. Zone boundaries were assigned for the express route to minimize the service frequency during peak transit period [18]. GPS and call details record (telephone data) were monitored and directly used to infer the real-time demand of transit's ridership to estimate demand and sequentially assign appropriate amount of supply with the objective of minimizing carbon emissions and operation cost [19].

The last category is the scheduling problem [20]–[24]. The primal objective of the scheduling is to decide the optimal fleet size given an imbalance of the demand. The representative ODM (generally during the peak period) is used to infer the imbalance of the flows (inbound and outbound) of the ridership. The bidirectional travel routes with stations nodes are considered with multiple strategic objectives such as express, dead heading, and short turning. The feasible pairs of nodes are considered such to eliminate the imbalance of the bidirectional ridership flows that results in smaller fleet size. The scheduling problem is specifically for public transportation with timetable.

2.3 Origin-destination prediction

In this section, an introduction to the ODM prediction models is given. The models included in this section is the basis model used to formulate the forecasting model used in this thesis. The more advanced models are introduced in Chapter 4 for performance comparison. The basic model used for ODM prediction is called the gravity model. There are multiple variations with different usage for each of its variation, but in general, the gravity model is modelled after the attraction force between objects which takes the form like the physics attraction force due to gravitation between two objects [1]. The masses in the gravitational force equation are replaced with ridership's factors called trip generation and trip attraction. Trip generation and trip attraction can be defined as either the number of entry and exit [1] or represented by other appropriate factor such as population density [25]. The distance decay is used similarly as the gravitation force equation and although multiple nodes (objects) are considered, each pair is specifically considered as having exclusive relationship hence, impedance from other nodes are not considered. In addition, the constant of the equation is the only parameter that needs estimation but in general, every pair of nodes uses the same constant. A similar version to the traditional gravity model is the gravity trade model. The gravity trade model assumes proportionality property between representative factors of trip generation and trip attraction to model the movement of goods [25]. The gravity trade model also considered each pair as exclusive.

As for the non-exclusive gravity variations, proportionality is calculated as to define the relationship between all pairs simultaneously. The simplest form of the variations is the frictionless gravity model [26]. The frictionless gravity model uses all nodes and assume the distribution of the trip generation (as number) is proportional, using the weight calculate with the ratio between the representative factor of trip attraction of a specific node to the summation, without the distance decay. Similarly, multinomial logistic regression model and multinomial logit model work in the same manner but the weights are calculated using exponential forms of estimated utility scores representing the trip attraction of each nodes [10], [27]–[35].

The frictionless gravity model, multinomial logistic regression model and multinomial logit model, however, fall short on the balancing of the rows and columns of the ODM since the proportions are considered from the trip attraction one-sidedly. Thus, a doubly constrained gravity model were introduced with the biproportional constraints using cost decay function to balance the property of both trip generation and trip attraction [3], [36].

The gravity models are static and mainly used for well-defined movements such as goods and long-distance migration. As the need of a more volatile ODM estimation is needed, multivariate distributions or conditional probabilistic models were used to model the demand [37], [38]. However, such proportions assumption were reported as unsatisfied and suggested that the parameters for the gravity models may be less than satisfactory for modelling the behavior of certain types of transport since the changes in the behavior itself shorten the usefulness of the trained model [39]. Additionally, while simulation of the ODM with statistic distributions are acceptable in practice, with the gradual changes of transportation behavior over time, there are not enough data to reach conclusive model.

The additional problem of ODM predictions is the discrete form commonly defined into the matrix form $\{T_{ij}\}$. The size of matrix $\{T_{ij}\}$ is significantly large and the the distribution of the concentration of the demand add an additional step to evaluate the prediction. Yao et al. (2021) reported 6.78% of the forecasted pairs in their study resulted in more than 200% error for segments with relatively low average magnitude of demand for the tested rail network while the tested rail network has relatively dense usage.

As for the data size, a K-mean clustering for data classification was proposed to substitute the temporal classification which provide greater number of observations per class by Kirby et al. (1997). And the data filtration is then done with classification. The suitability of the generalization of this method, is however, unclear in literature as of now.

While there are many factors and design of the ODM prediction, the suitability of the model needs to be evaluated and the understanding of the data generated from the network of interest is needed. The usage of the prediction is also a factor to consider when designing the prediction model as the requirements changes depending on the objectives.

In Fig. 1 the level of consideration of demand variability is shown. The variability of the demand increases with the level of consideration of temporal-spatial factor. For example: considering the design of entry-exit a specific station, an estimation of boarding and alighting is needed; for basic timetable/headway decision, a static ODM is sufficient for decision making; while for congestion relief or rescheduling, a dynamic ODM prediction is needed to minimize the network operation recovery rate. In this thesis, we focus on the usefulness of the real-time forecasting of the ODM, so the spatial level is the OD pair level considered on a dynamic estimation.

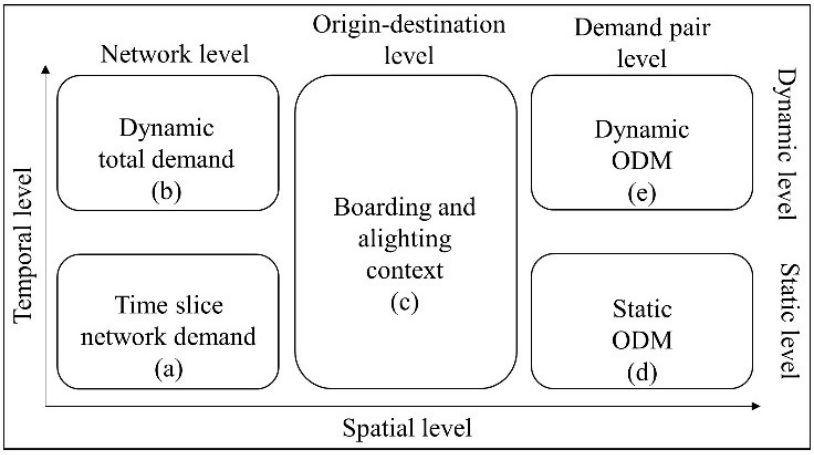

Figure 1: Level of consideration of demand variability.

Chapter 3: Data

The data used in this thesis is provided by Mass Rapid Transit Authority of Thailand (MRTA). MRTA is the largest subway network in the capital, Bangkok as shown in Fig 2. The format of the data is the aggregated ODM with entry-based record with the time frame of one hour between record year of 2019 to 2020. The network is divided into two main lines P002 and P502 with different inbound and outbound. The network has multiple transfer points which are connected within the subway network and including transfers outside to the other rail networks. Up to date, MRTA is the only organization in Thailand involved in developing underground rail network. The summary of the datasets is shown in Table 1. Special note on dataset P502: 1) The dataset P502 record is from the start of its operation unlike dataset P002 which has been established for much longer; 2) The number of records of P502 is smaller than P002 by 1.2 million; Lastly, 3) P502 is used in analysis in Chapter 4 and excluded from forecasting in Chapter 5 due to the fundamental instabilities in ridership behavior and operational instability.

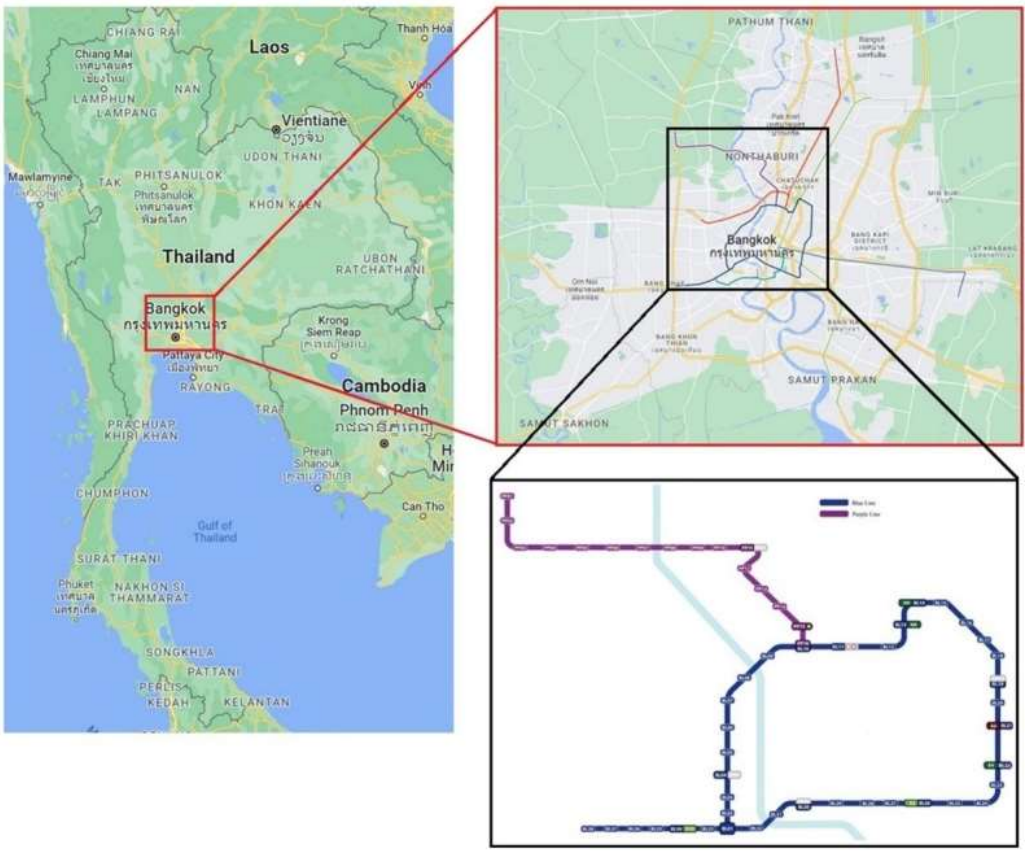

Figure 2: Subway map of Mass Rapid Transit Authority in Bangkok, Thailand.

Table 1: Datasets summary.

Description	P002	P502
Date	January 1, 2019 to December 29, 2020	December 3, 2019 to December 29,2020
Week number	105	57
Data record	2.9 million	1.7 million
Station number	34	35
Matrix dimension	34×34	35×35
Time interval	60 min	60 min
Matrix number in a	24	24
Type of records	Automated (entry-first)	
Same entry-exit	Allowed	
Transfer station	Yes	
Special note	None	Data record from the start of operation

The summary of the data statistics is shown in Table 2. P002 consists of blue line system and P502 consists mainly of purple line with additional overlaps with blue line stations. The ODM processing is, however, separated into 2 datasets based on entry-exit and is the official ODM released by MRTA. Since, the aggregated data consists of many records daily, the only statistics on significant time frames are shown, e.g., peak demand, and daily demand. Special note on the statistics: The statistics are lower than the actual due to the records within quarantine period. W3 and W4 specify the operational condition of non-peak and peak's capacity in relation to headway, a scheduling parameter deciding the hourly capacity of ridership. The range of the maximum capacity is decided with the comfort level, namely the density of ridership within the train. The correlations between significant time frames are shown in Table 3. The correlation between one/two-hour time frame to daily ridership of P502 are lower than expected indicating the inconsistency of ridership behavior and thus excluded from forecasting since it shows a lack of relation of peak and non-peak which is against the requirements of the design of forecasting study in this thesis.

Table 2: Data statistics.

2019-2020		*P002*			*P502*		
		1-h (M-Peak)	*2-h (M-Peak)*	*Daily*	*1-h (M-Peak)*	*2-h (M-Peak)*	*Daily*
Statistics	Mean	9073.04	16755.67	69121.38	3039.00	6073.72	28891.65
	Standard error	349.14	640.97	2101.41	172.27	354.71	1720.29
	Median	10369	19375	77059	2910	5801	25558
	Standard deviation	3577.57	6567.97	21533.01	1300.60	2677.99	12987.92
	Kurtosis	-0.14	-0.10	-0.27	-0.43	-0.55	-0.39
	Skewness	-1.07	-1.08	-0.99	0.22	0.17	0.79
	Range	13353	24052	82848	5241	10642	50796
	Minimum	707	1238	12146	439	779	6829
	Maximum	14060	25290	94994	5680	11421	57625
	Max Cap (W3)	9200-12880			9200-12880		
	Max Cap (W4)	11860-16604			11860-16604		
Basic Info	Number of stations	34			35		
	Special note (*)				Operation started December 2019		

Table 3: Correlation of demand between time slices.

Time slices		*Dataset*		
X	*Y*	*P002*	*P002**	*P502*
1-h	2 h	0.995	0.997	0.992
1-h	Daily	0.953	0.955	**0.768**
2-h	Daily	0.953	0.957	**0.760**

Chapter 4: Analysis of transportation data

The premise of this Chapter is, to check through data analysis on some general assumptions in forecasting problem on lower dimensions, and thus, decide on feasible approaches to the ODM level forecasting problem. The data used in this thesis is, in essence, known to be highly volatile. Thus, data analysis is essential in the design of the forecasting problem. Wide range of variables affects the predictability of ODM. In the real-time prediction, multiple steps are required to work with the understanding of data and thus designing the model and make logical comparison and analysis of the results to add to the literature. In this section, Historical average (HA) and statistical multinomial maximum likelihood are used to evaluate the volatility of the data. In addition to the two approaches, we include univariate and multivariate aspects of the analysis as the spatial level increases, the number of predictions and estimates increases. Proportionality is investigated as it is one of the most common properties in ODM design. The total demand and boarding analysis are used in designing the forecasting problem of ODM level in Chapter 5. Distributive forecasting model is tested on ODM level for clarification of the dataset.

4.1 Univariate analysis

For the univariate case, total demand's distribution is investigated. Normality is a common assumption in prediction problem. Chi-square frequency test is suitable for discrete data. To illustrate, we select two peak time frames of total demand of ODM from peak periods from dataset P002 and P502 including its daily total The data statistics for total demand is shown in Table 4. Due to the mixture of high-low density data, the discrete range of the total demand is significantly wide, and its standard deviation likewise vary.

To test the proportionality, we transform the data into fraction of demand between the time frame and daily, the result is shown in Table 5. The proportional data by time now show similar level of standard deviation. Overlapping qualities are desired between datasets. Cross-referencing amongst dataset adds validation to the assumptions that will be used in the forecasting process.

Table 4: Total demand data statistics.

Dataset	Mean	SD	Range
P002-1	7115.7	2827.8	9715
P002-2	8296.4	3411.3	12231
P002	60775.4	22109.1	78397
P502-1	3018.4	1335.1	5401
P502-2	3028.9	1270.41	5241
P502	27170.7	12124.3	49629

Table 5: Proportional data statistics.

Dataset	Mean	SD	Range
P002-1p	0.115	0.0214	0.123
P002-2p	0.134	0.0242	0.148
P502-1p	0.111	0.0242	0.145
P502-2p	0.112	0.0216	0.132

Table 6 shows the normality test of total demand. The statistics shows that the proportional data shows a significant level of normality. In addition, Fig. 2 shows the boxplot of total demand before and after transformed. The boxplots indicate two different data behavior of P002 and P502. While, for P002,the temporal proportion transform shows a normal distribution while for P502, the temporal proportion doesn't improve its normality. In this, case, we can safely disregard HA for a viable method of comparison for dataset P002 for total demand estimation.

Table 6: Normality test (total demand).

Dataset	statistics	p-value
P002-1	37049.547	0.000
P002-2	47584.580	0.000
P002-3	319478.076	0.000
P502-1	17792.546	0.000
P502-2	6087.848	0.000
P502-3	45004.196	0.000
P002-1p	0.012	1.000
P002-2p	0.018	1.000
P502-1p	0.082	1.000
P502-2p	0.060	1.000

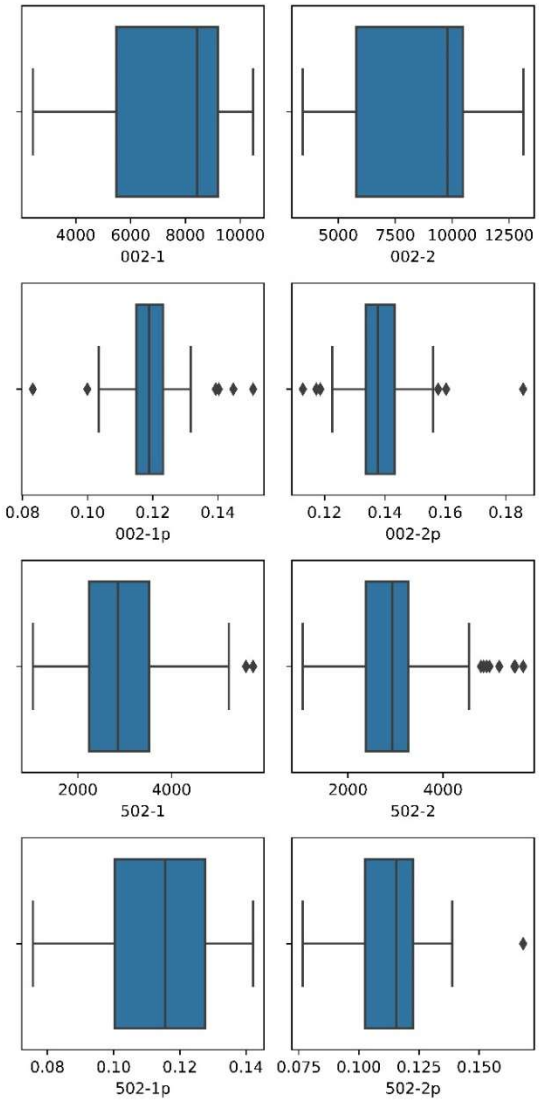

Figure 3: Total demand boxplots.

4.2 Multivariate analysis

For the multivariate case, boarding's multivariate normality is investigated. HZ statistics is applied for the trip generation. Poisson distribution is assumed for the multivariate analysis for the individual boarding points. Poisson distribution is identical to Multinomial of different sample size or total network demand n:

$$P(\bar{O}|n) \sim MultiNom(n, \pi)$$

where, O is the trip generation matrix, and π is the proportionality parameter matrix. If the proportionality condition is satisfied, the transformation of the data into sampling proportion distribution reduce the variance by a degree of ${O_i}^{-1}$ where O_i is the trip

generation factor of origin i. And the relation to the normality of the transformed data can be normally approximated as:

$$\frac{O_i}{n} \sim Normal\left(\pi_i, \theta\sigma_{O_i}\right); 0 < \theta \ll 1$$

where, $\theta\sigma_{O_i} \approx {O_i}^{-1/2}$ is the standard deviation. The Mahalanobis distance was selected to illustrate the degenerate case of the assumed multinomial distribution. Nearest pseudo-covariance matrix is calculated when the covariance matrix is singular. The normality of the sample size in the univariate level and the multivariate normality of the trip generation were analyzed with test statistics in addition to boxplots to visualize the changes on the normality level.

Table, 7 shows the HZ statistics and only proportional data of P002-1 resulted in a significant Gaussian behavior. An example of boxplot of randomly selected origin point is shown in Fig. 4.

Table 7: Multivariate normality test (trip generation).

Dataset	HZstatistic	p-value
P002-1	**47.990**	**0**
P002-2	**47.990**	**0**
P002-3	**47.990**	**0**
P502-1	**192**	**0**
P502-2	**180**	**0**
P502-3	**192**	**0**
002-1p	**1.000**	**0.110**
002-2p	**47.992**	**0**
002-3p	**47.992**	**0**
502-1p	**192**	**0**
502-2p	**192**	**0**
502-3p	**92**	**0**

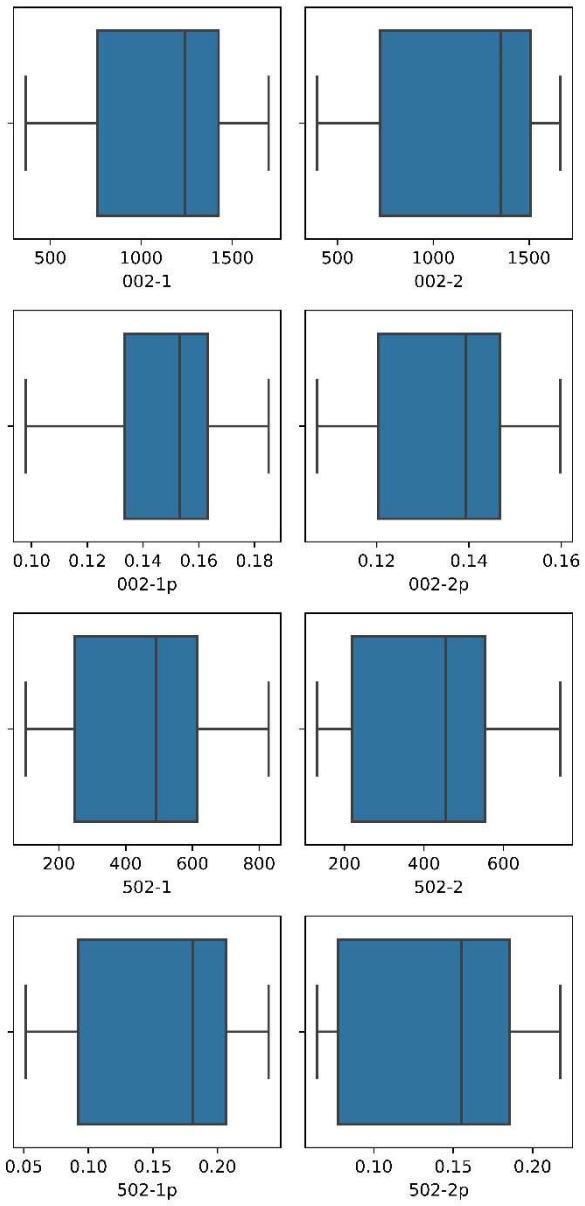

Figure 4: Trip generation boxplot (randomly selected origin point).

From the univariate and multivariate analysis, we can safely conclude that for MRTA datasets, normality should not be included into the assumptions for the ODM prediction since the increase in dimension from total demand, boarding, to ODM level will likely result in insignificant Gaussian behavior.

4.3 Distributive estimation

After, the analysis of the data, in this section, we will investigate the distributive estimation of the trip generation and ODM. Here, we will use traditional indicators for measuring transportation demand error.

4.3.1 Trip generation distributive prediction

In this section, we assume a multinomial distribution with the relation between total demand and trip generation. The varying size of total demand n is assumed to be random with constant parameter π estimated from Maximum likelihood estimator (MLE). And the estimate of π_i is

$$\hat{\pi}_i == \frac{O_{i+}}{n_+}$$

Where, (+) is the summation sign, hence, O_{i+} is the summation of O_i by its observations and likewise n. The estimate of the trip generation is

$$O_i = n\pi_i.$$

Total Demand Deviation (TDD) and Mean Squared Error (MSE) were selected as the error measure of variability.

$$TDD(\omega, \hat{\omega}) = \frac{\left|\sum_{i=1}^{N} \omega_i - \sum_{i=1}^{N} \hat{\omega}_i\right|}{\sum_{i=1}^{N} \omega_i}$$

$$MSE(\omega, \hat{\omega}) = \frac{1}{N}\sum(\omega_i - \omega_i)^2$$

Here, ω and $\hat{\omega}$ are respectively the real and forecasted values, and *N* is the length of data. The measure of TDD is used to measure a known network deviation in comparison to the past measure of TDD or to compare between TDD of different networks. TDD and MSE are shown in Fig.5 and Fig.6 accordingly.

The TDD of the dataset P002-1 and P502-1 are of a similar value with P002-1 having the minimum TDD out of all datasets. However, for another selected time frame of P002-2 and P502-2, the TDD increases substantially with P502-2 increases from P501-1 significantly.

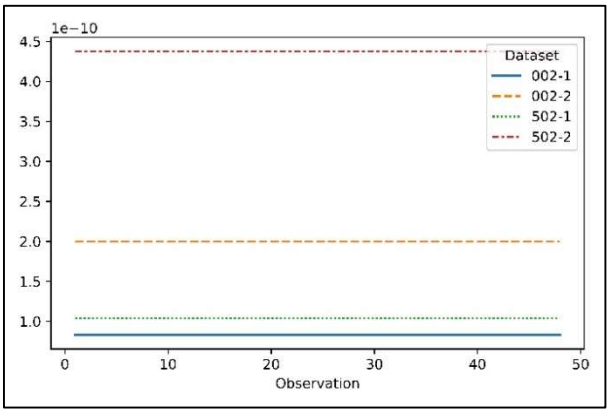

Figure 5: Total demand deviation.

MSE however, follow the trend of the selected observation, which indicates that although from different time frames, the daily effects suggests that since the two networks are under the same area, they are under the same influence.

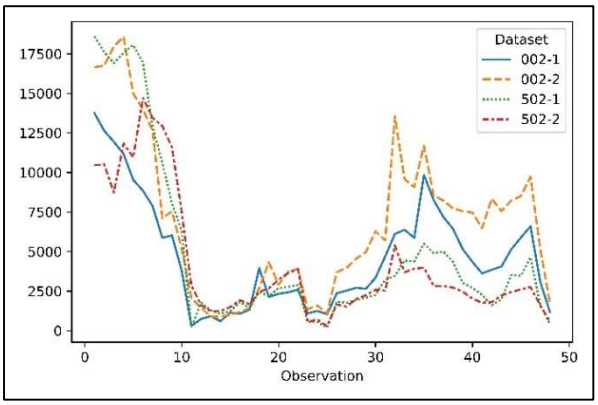

Figure 6: Mean squared error.

The total demand by tested observation is shown in Fig. 7. The trend of the of the measurement of error follows the magnitude of the total demand. When the comparison is made, it shows that P502 is more volatile than P002 since the magnitude of total demand of P502 is on average, lower than P002.

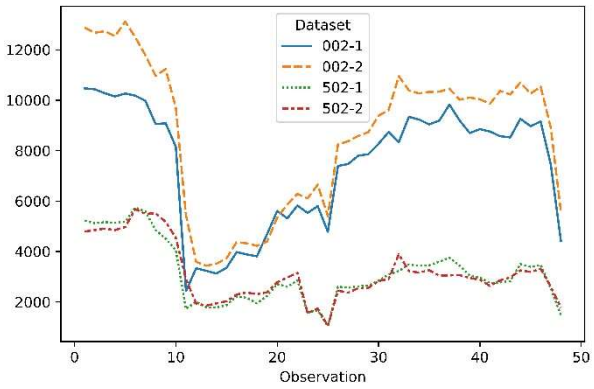

Figure 7: Total demand by observation.

4.3.2 Origin-destination distributive prediction

In this section, we used the multinomial model for the estimation of ODM and maximum likelihood estimator for the parameters. Revise to evaluations method for ODM estimation were designed for lower demand segments. Two networks from the same area and date of observations were analyzed as comparison of the effects of socioeconomic factor .

For ODM level, unlike, total demand and trip generation, multiple considerations for the design are needed. Due to the higher dimensions of the problem, the volatility of the target of prediction may not result in satisfactory level of accuracy. Low ridership density in each pair of ODM is apparent. Cut-off function or weight design is necessary for logical evaluation of the predictability of ODM. GEH statistics is used to evaluate the distributive model prediction and a statistical cut-off function is proposed to evaluate higher resolution ODM.

4.3.2.1 Methodology

The framework is summarized in Fig 8, consisting of Multinomial model for the estimation based on statistical distribution, the parameter estimation based on Maximum likelihood estimator, and the validation of the model. Given a network of size *m,* with Poisson-distributed demand, the assumptions of volume and variance are as follows.

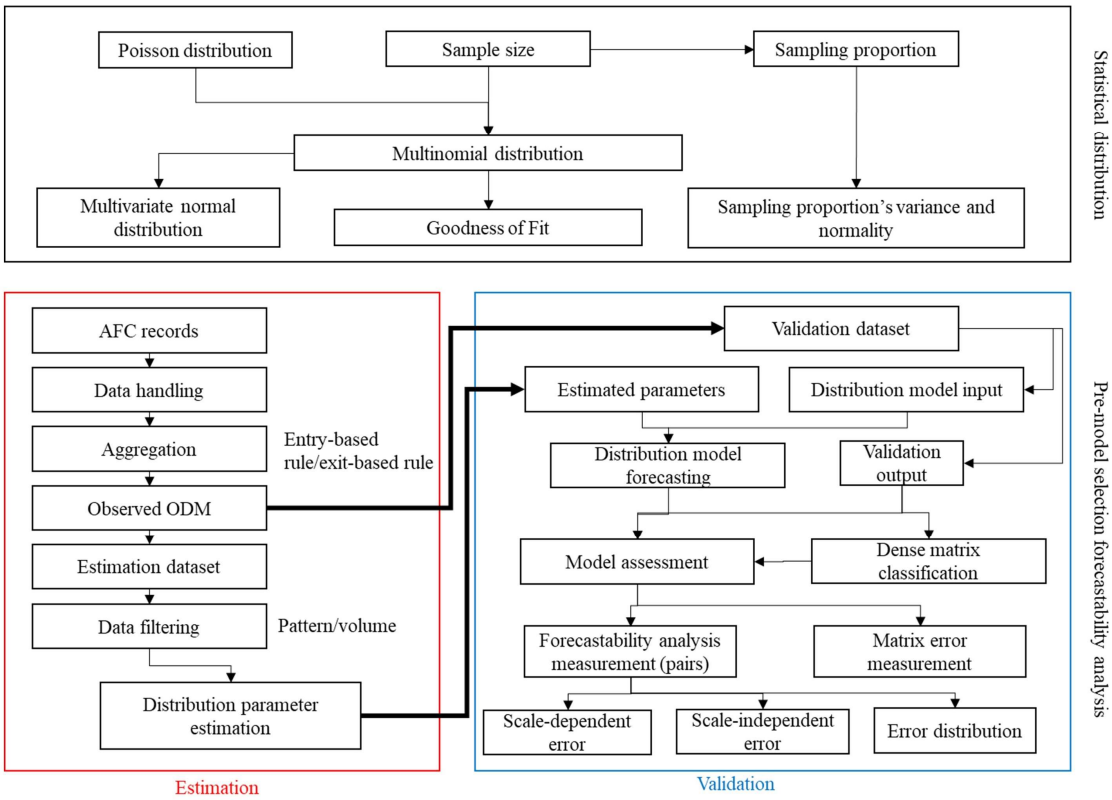

Figure 8: Framework for distributive model prediction.

Assumption

The travel demand flow from origin i to destination j at time t of day type r $d_{ij}^{t,r,,\gamma}$ of the OD pairs of follows an independent Poisson distribution with mean parameter a linear function of random variable socioeconomic level γ. The approximation to the Multinomial distribution of $d_{ij}^{t,r,,\gamma}$ is shown in Eq. (7).

$$d_{ij}^{t,\gamma} \sim Multinomial(\pi_{ij}^{t,r}, \lambda_i^{t,r,\gamma}) \quad (7)$$

where, $d_j^{t,\gamma} = \{d_{1j}^{t,\gamma}, d_{2j}^{t,\gamma}, \dots, d_{mj}^{t,\gamma}\}$ is the demand matrix, $\lambda_i^{t,r,\gamma}$ is the Poisson parameters representing observed passenger entries at origin i at time t, day type r, and socioeconomic level γ. γ is estimated from another network inside the assumed intra-cluster, the correlation between the networks is presented in Section 3.3.1. And $\pi_{ij}^{t,r}$ is the multinomial parameter estimated from MLE. γ is assumed homogenous across the networks of interests for this study. In the homogenous case, γ is equivalent to the

sampling proportion ρ of the true population originating from a traditional data collection strategy of sample surveying, where demand T_{ij} and sampling demand t_{ij} follow $T_{ij} = t_{ij}/\rho$. A similar model was proposed by Spiess (1987), offering a more robust solution with Lagrangian optimal solution ${\alpha_i}^*$ and ${\beta_j}^*$ for the dual maximum likelihood of doubly constrained in the form of $T_{ij} = t_{ij}/\left(\rho_{ij} + {\alpha_i}^* + {\beta_j}^*\right)$. Geva et al. (1983) proposed, for a sampling of full proportion, the solution $T_{ij} = t_{ij}/\left(1 - {\alpha_i}^*{\beta_j}^*\right)$ and $T_{ij} = t_{ij}/\left({\alpha_i}^* + {\beta_j}^*\right)$ for the binomial and multinomial models, respectively. The level of sampling of certain link flows is at the ρ-level of representation of the actual population value. This study adopts this principle to produce an estimate of the actual full sampling and the variability of the demand by substituting ρ with γ, hence, the solution to the demand estimation problem changed from 'assumed proportion' ρ to 'level of demand' γ similarly to gravity model.

Data filtering and parameter estimation

The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \dots, x_N)^T$ from a set of observations with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \dots, \mu_N)^T$ and covariance matrix Σ^{-1} is defined as

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})}$$

The Moore-Penrose pseudo-covariance matrix is calculated when the covariance matrix is singular. The filtering consists of raw data and transformed data using MLE for individual data matrix from the dataset using:

$$\hat{\pi}_{ij}^{t,r} = \frac{d_{ij}^{t,\gamma}}{d_{i+}^{t,\gamma}}, \qquad (8)$$

4.3.2.2 Derivation of the maximum likelihood estimator

Maximum likelihood estimator of multinomial distribution: single observation case (data transformation)

Considering the parameter estimation from a single observation dataset,

$$PMF: f(d_{i1}, d_{i2}, d_{i3}, \dots, d_{im}) = \frac{k!}{d_{i1}!\, d_{i2}! \dots d_{im}!} \pi_{i1}^{d_{i1}} \pi_{i2}^{d_{i2}} \dots \pi_{im}^{d_{im}}$$

where

$$d_{i+} = k \; and \; \pi_{i+} = 1$$

Log-likelihood is given by

$$L(\pi_{i1}, \pi_{i2}, \dots, \pi_{im}|\overrightarrow{d_\iota}) = log\left(\begin{matrix} k \\ d_{i1} d_{i2} \dots d_{im} \end{matrix}\right) + \sum_{j=1}^{m} d_{ij} log(\pi_{ij})$$

Maximization of the log-likelihood function with Lagrange function is

$$\mathcal{L}(\pi_{i1}, \pi_{i2}, \dots, \pi_{im}|\lambda) = L(\pi_{i1}, \pi_{i2}, \dots, \pi_{im}|\overrightarrow{d_\iota}) + \lambda(1 - \pi_{i+})$$

where λ is the Lagrange multiplier (not to be confused with the Poisson mean).

For λ,

$$\frac{\partial \mathcal{L}}{\partial \pi_{ij}} = \frac{d_{ij}}{\pi_{ij}} - \lambda \overset{\text{set}}{=} 0$$

$$\lambda \pi_{ij} = d_{ij}$$

$$\sum_{j=1}^{m} \lambda \pi_{ij} = d_{i+}$$

$$\lambda \pi_{i+} = k$$

From

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \pi_{i+} \overset{\text{set}}{=} 0$$

$$\pi_{i+} = 1$$

$$\lambda = k$$

The maximum likelihood estimator of multinomial parameters of a single observation is

$$\hat{\pi}_{ij} = \frac{d_{ij}}{k} = \frac{d_{ij}}{d_{i+}} = \frac{d_{ij}}{O_i}$$

Maximum likelihood estimator of multinomial distribution: multiple observations (destination choice model)

We repeat the experiments B times with different k_l values given the data matrix of each Origin i $\mathcal{D}_i = \{d_{ijl}\}$ with $l \in \{1,2, \dots, B\}$ of the destination choice model:

$$PMF: f(\mathcal{D}_i|\vec{\pi}_i, \vec{k}) = \prod_{l=1}^{B} f(\vec{d}_{il}|\vec{\pi}_i, k_l) = \prod_{l=1}^{B} \left(\begin{matrix} k_l \\ d_{i1l} \dots d_{iml} \end{matrix}\right) \pi_{i1}^{d_{i1l}} \dots \pi_{im}^{d_{iml}}$$

$$f(\mathcal{D}_i|\vec{\pi}_i, \vec{k}) = \left[\prod_{l=1}^{B} \left(\begin{matrix} k_l \\ d_{i1l} \dots d_{iml} \end{matrix}\right)\right] \pi_{i1}^{d_{i1+}} \dots \pi_{im}^{d_{im}}$$

where

$$\pi_{i+} = 1, k_+ = K, d_{i+l} = k_l, d_{i++} = K$$

Log-likelihood:

$$L\left(\pi_{i1}, \pi_{i2}, \dots, \pi_{im} | \overrightarrow{d_{il}}\right) = log\left(\prod_{l=1}^{B} \binom{k_l}{d_{i1l} \dots d_{iml}}\right) + \sum_{j=1}^{m}\left(\sum_{l=1}^{B} d_{ijl}\right) log\left(\pi_{ij}\right)$$

Maximization of the log-likelihood function with the Lagrange function:

$$\mathcal{L}(\pi_{i1}, \pi_{i2}, \dots, \pi_{im} | \lambda) = L\left(\pi_{i1}, \pi_{i2}, \dots, \pi_{im} | \overrightarrow{d_{il}}\right) + \lambda(1 - \pi_{i+})$$

where λ is the Lagrange multiplier.

For λ,

$$\frac{\partial \mathcal{L}}{\partial \pi_{ij}} = \frac{d_{ij+}}{\pi_{ij}} - \lambda \overset{\text{set}}{=} 0$$

$$d_{ij+} = \lambda \pi_{ij}$$

$$d_{i++} = \lambda \pi_{i+} = K$$

From

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \pi_{i+} \overset{\text{set}}{=} 0$$

$$\pi_{i+} = 1$$

$$\lambda = K$$

The maximum likelihood estimator of multinomial parameters of a repeated-observation dataset is

$$\hat{\pi}_{ij} = \frac{d_{ij}}{K} = \frac{d_{ij}}{d_{i++}}$$

4.3.2.3 Statistical cut-off

For large samples, the sample proportion is approximately normally distributed, with a mean of $\mu_{prop_{ij}} = \pi_{ij} = \lambda_{ij}/\lambda_{i+}$ and a variance of

$$\sigma_{prop_{ij}} = \sqrt{\pi_{ij}\left(1 - \pi_{ij}\right)/\lambda_{i+}} = \left(\lambda_{i+}\sqrt{\lambda_{i+}}\right)^{-1} \sqrt{\lambda_{i+} - \lambda_{ij}} \sqrt{\lambda_{ij}} = \theta \sigma_{\lambda_{ij}}$$

It can be inferred that $0 < \theta \ll 1$. Hence, $\sigma_{prop_{ij}} \ll \sigma_{\lambda_{ij}}$. Hence, for the transformed data, the normal approximation and lower bound are

$$\hat{\pi}_{ij}^{t,r} \sim Normal\left(\pi_{ij}^{t,r}, \theta\sigma_{\lambda_i^{t,r,\gamma}}\right); 0 < \theta \ll 1,$$

The observation of Multinomial parameter follows the sampling distribution. A sample is large if the interval

$$\left[\pi_{ij} - 3\sigma_{prop_{ij}}, \pi_{ij} + 3\sigma_{prop_{ij}}\right] \qquad (9)$$

lies within the interval $[0,1]$. For unknown π_{ij} and a repeated experiment, the given

$$\hat{\pi}_{ij} = \frac{d_{ij}}{d_{i++}}$$

For the demand of a transportation network, it can be inferred that the lower bound of Eq (9) can be binding. The estimated of the mean and variance of the sampling distribution are, $\pi_{ij} \cong \hat{\pi}_{ij}$ and $\sigma_{prop_{ij}} \cong \theta\sigma_{\lambda_{ij}}$, hence,

$$\left[\hat{\pi}_{ij} - 3\theta\sigma_{\lambda_{ij}}, \hat{\pi}_{ij} + 3\theta\sigma_{\lambda_{ij}}\right]$$

Based on the trip purpose and its stability, a variance smaller than $\sigma_{\lambda_{ij}}$ can be expected due to randomness being a smaller portion in the Poisson point process. The approximation is appropriate when the Poisson parameter λ_{ij} is sufficiently large. Solving the lower bound equation in relation to sample size:

$$\hat{\pi}_{ij} - 3\theta\sigma_{\lambda_{ij}} \geq 0$$

where

$$\theta\sigma_{\lambda_{ij}} = \frac{\sqrt{O_i - \lambda_{ij}}\sqrt{\lambda_{ij}}}{O_i\sqrt{O_i}}$$

and,

$$\hat{\pi}_{ij} = \frac{\lambda_{ij}}{O_i}$$

The non-linear lower bound is:

$$\lambda_{ij} \geq \frac{9O_i}{O_i + 9}$$

4.3.2.4 Evaluation

While the problem of validation based on scale-independent error needs some reference for comparison to comprehend the meaning of the value, multiple networks and error

were selected for demonstration. The notation k for matrix indexing is for evaluating individual OD pair with K being the total number of matrices, while notation i is for evaluating individual ODM with N being the total number of OD pairs.

Matrix evaluation

Relative root mean square error (RRMSE), root mean square error (RMSE), total demand deviation (TDD), mean absolute error (MAE), and root mean weighted fractional error (RMWFE) are selected for matrix level evaluation. The errors are calculated for each matrix of observation.

$$RRMSE(\omega, \hat{\omega}) = \frac{1}{\sum_{i=1}^{N} \omega_i} \sqrt{\frac{\sum_{i=1}^{N} (\omega_i - \hat{\omega}_i)^2}{N}}$$

$$RMSE(\omega, \hat{\omega}) = \sqrt{\frac{1}{N} \sum (\omega_i - \omega_i)^2}$$

$$TDD(\omega, \hat{\omega}) = \frac{|\sum_{i=1}^{N} \omega_i - \sum_{i=1}^{N} \hat{\omega}_i|}{\sum_{i=1}^{N} \omega_i}$$

$$MAE(\omega, \hat{\omega}) = \frac{\sum_{i=1}^{N} |\omega_i - y_i|}{N}$$

$$RMWFE(\omega, \hat{\omega}) = \sqrt{\frac{1}{\sum_{i=1}^{N} \omega_i} \sum_{i=1}^{N} \frac{(\omega_i - y_i)^2}{\omega_i}}$$

Origin-destination pair evaluation

The assessment of the OD pairs level contains a scale-independent measure, an error measurement distribution, and a scale-dependent measurement, as shown in Eqs. (11)-(13), MSE, mean percentage error (MPE), and weighted mean average percentage error (WMAPE), respectively.

$$MSE_{OD\ pair}(\omega, \hat{\omega}) = \frac{\sum_{k=1}^{K} (\omega_k - \hat{\omega}_k)^2}{K} \quad (11)$$

$$MPE_{OD\ pair}(\omega, \hat{\omega}) = \frac{1}{\sum_{k=1}^{K} \omega_k} \sum_{k=1}^{K} \omega_k - \hat{\omega}_k \quad (12)$$

$$WMAPE_{OD\ pair}(\omega, \hat{\omega}) = \frac{\sum_{k=1}^{K} |\omega_i - \hat{\omega}_i|}{\sum_{k=1}^{K} |\omega_k|} \quad (13)$$

Additionally, Geoffrey E. Havers (GEH Statistic) is empirical statistics error for ODM

prediction and is compared to Eq. (13).

$$GEH_{OD\ pair}(\omega, \hat{\omega}) = \sqrt{\frac{2(\omega-\hat{\omega})^2}{\omega+\hat{\omega}}} \quad (14)$$

4.3.2.5 Prediction result

The test dataset of dataset P502 is half of P002 due to smaller number of data points. The selected test dataset of P502 matches by date with the selected test dataset of P002. The multinomial is divided the same as the type of record, entry based. Higher numbers of sparse components are observed in P502. The plotting of the prediction error for both datasets are matched by date.

Evaluation of Matrix level predictions

The matrix assessment shown in Fig. 9 consists of the RMMSE, RMSE, TDD, RMSE, and RMWFE, in the form of individual matrix evaluation. As indicated, there exists high correlation between the error measure of both networks that can be easily seen in the plots of RRMSE, TDD, and MAE, all of which are scale-dependent error measures. According to the investigation of the abnormally high error, although network P002 and P502 has different spatial coverage, the shift in ridership density appeared at the same date. In addition to the comparison between datasets, the shift of ridership density is apparent across the peak-period and affects the daily total demand significantly.

The RRMSE and TDD show a relatively low error for the observations when excluding observations that are likely outliers (less than 0.3% and 5%, respectively). The error measures for network P502 are significantly higher than for network P002. The predictability of P502 needs further investigation in the recent data if the behavior of the predictability persists (if the predictability is unique for the network P502 or not). but that is only if we consider an additional baseline network. Overall, the matrix level evaluation of P002 appears to be satisfactory with average RRMSE of 0.0005, RMSE of 1, TDD which shows a negligent error except for when the observations with significant ridership density shift, MAE of 3-4 for shorter time frames but range between 5-24 for daily demand due to higher total ridership count. At the matrix level, it can be inferred that network P502 lacks the quality to represents the ridership demand just referring to IC card data as input.

Figure 9: Matrix forecast error.

Evaluation of origin-destination pair level predictions

The summary of the main performance evaluation at OD pair level is shown in Table. 8.

The main performance evaluation consists of two methods (raw data feed and normalized data feed methods), two classes of evaluation (no cut-off class and filtered class), and two relative indicators (WMAPE and GEH). Table 8 shows the performance of OD pair evaluations by calculating the percentage of the frequency assign when the error is higher than acceptable margin, (30% for WMAPE, and 5 for GEH).

The normalized data feed significantly increases the performance of the prediction for dataset P002. However, improvements are smaller as the time frame increases. For dataset P502, the opposite appears as improvements are shown when the time frame increases and for smaller time frame, discrete data feed seems to be better. Nevertheless, the percentage of OD pairs within acceptable margin is very low for dataset P502. In addition, GEH, while indicating large amounts of acceptable margin similarly for dataset P002, overestimate the predictability of dataset P502 and significantly overestimate the predictability as the magnitude of the demand within OD pair grows due to the increase of time frame. The expected variability of demand is high for network P502 but weren't finely evaluated by GEH statistics.

Table 8: Summary of main performance results.

Network	*Time*	*Method*	*Class*	*% frequency of WMAPE*		*% frequency of GEH*	
				0-30	*30+*	*0-5*	*5+*
P002	1-h	Discrete	Nonfilter	25.0	75.0	N/A	N/A
			Filtered	***80.5***	19.5		
		Transformed	Nonfilter	28.2	71.8	***80.2***	19.8
			Filtered	***85.0***	15.0		
	2-h	Discrete	Nonfilter	36.1	63.9	N/A	N/A
			Filtered	***81.2***	18.8		
		Transformed	Nonfilter	38.5	61.5	***87.1***	12.9
			Filtered	***84.3***	15.7		
	Daily	Discrete	Nonfilter	73.4	26.6	N/A	N/A
			Filtered	***81.5***	18.5		
		Transformed	Nonfilter	74.4	25.6	***95.5***	4.5
			Filtered	***82.4***	17.6		
P502	1-h	Discrete	Nonfilter	19.7	80.3	N/A	N/A
			Filtered	***39.8***	60.2		
		Transformed	Nonfilter	15.0	85.0	***57.4***	42.6
			Filtered	***36.9***	63.1		
	2-h	Discrete	Nonfilter	18.2	81.8	N/A	N/A
			Filtered	***41.6***	58.5		
		Transformed	Nonfilter	15.2	84.8	***67.1***	32.9
			Filtered	***40.4***	59.6		
	Daily	Discrete	Nonfilter	17.8	82.2	N/A	N/A
			Filtered	***34.8***	65.2		
		Transformed	Nonfilter	17.9	82.1	***75.1***	24.9
			Filtered	***35.1***	64.9		

Fig. 10 and Fig. 11 show the components of distribution of error by class of dataset P002 and P502 respectively. The dense class is the nonfilter case, the low base

class is the OD pairs that need to be excluded, and the normal class is the filtered OD. MSE, MPE, and WMAPE are shown as boxplot in the figures.

The low base class undermine the average MSE of the actual prediction error by a large margin, While for MPE and WMAPE of P002, low base class OD pairs significantly contribute to the overall error. However, for MPE and WMAPE of P502, the opposite is true, and the overestimation of predictability is apparent.

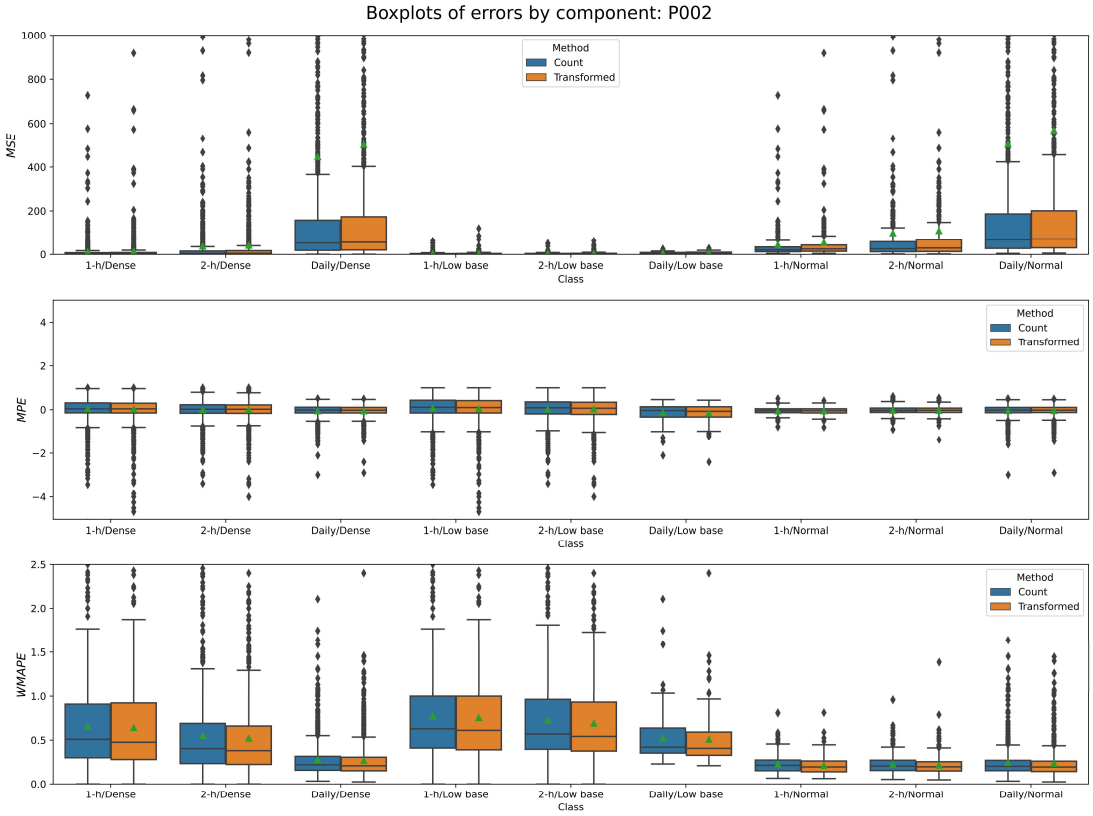

Figure 10: Boxplots of P002 error distribution.

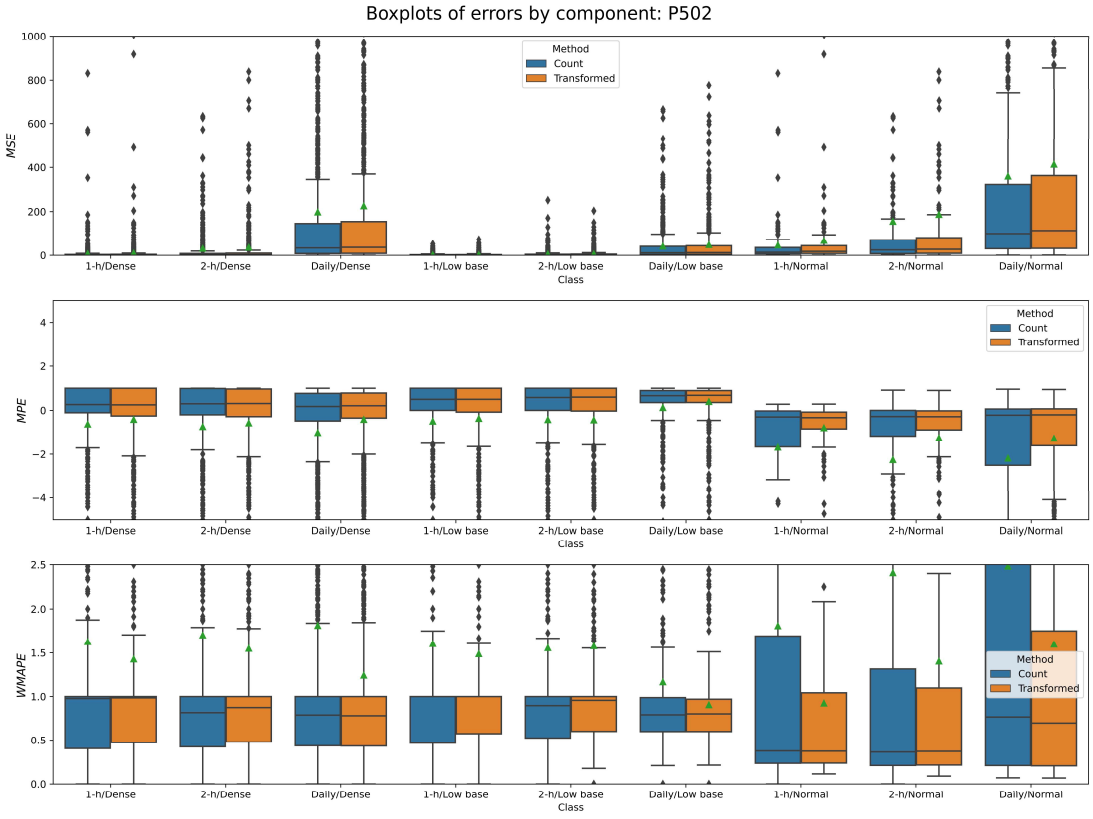

Figure 11: Boxplots of P502 error distribution.

4.4 Summary of analysis

The datasets tested in this section, P002 and P502 were analyzed on aspects of its predictability, the volatility of the record ODM exhibits low proportionality, high number of sparse, and deterioration under different spatial and temporal scales. Comparison between datasets locating near the capital exhibits similar trend of volatility when matching with dates where shift of ridership density was observed. In addition, the size of time frame under peak-period significantly affects the daily ridership number. Normality can be observed in lower resolution of spatial level from total demand (total ridership per time frame) to, with multiple modifications to the design of problem, the trip generation with weak multivariate normality. As the spatial scale reaches OD pair, the problem of low average density in certain pairs become prominent and the prediction lacks legitimacy when solved with HA and distributive multinomial.

Specifically on dataset P502, the predictability analyzed throughout this chapter is unsatisfactory and will be excluded from Chapter 5 henceforth. However, we will apply the trend of the predictability of P502 as cross referencing to potentially filter out

technical outliers from the observation as the design of the forecasting in Chapter 5 requires classification and thus, is expected to include outliers into the classification as shown in Fig 12.

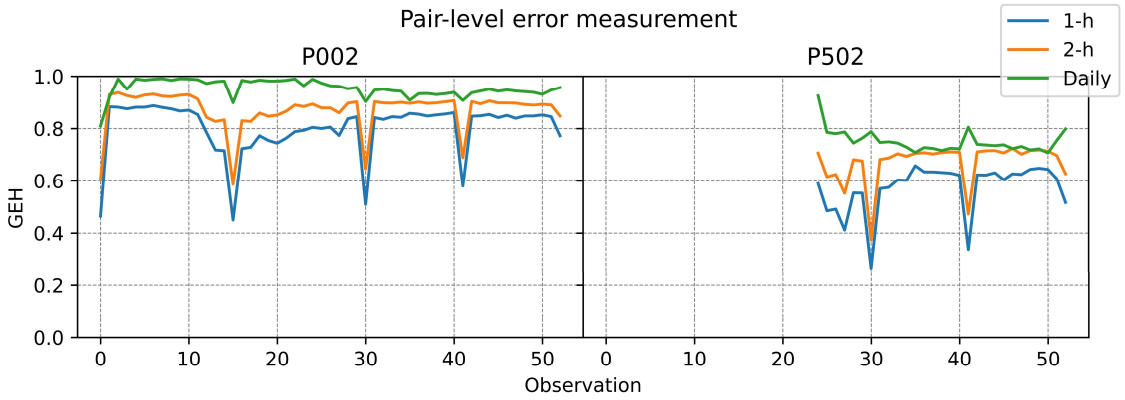

Figure 12: GEH statistics origin-destination pairs acceptance proportion by observations.

For dataset P002 which will be used in the next chapter, our hypothesis is that the assumption of homogenous socioeconomic factors, throughout the entire datasets, attributes to the increases in prediction error. Thus, the multivariate nature of the demand was inferred to be correlated to a single behavior throughout the entire observations. The design of the forecasting problem will need to tackle a multi-behavior of ODM.

GEH statistics will not be adopted into the forecasting evaluation since the empirical indicator was tested to be "naïve" when the magnitude of the demand increases but very sensitive to abnormal behavior.

This chapter does not include the design of the real-time data feed to the prediction and is focused on analysis. A detailed design of data feed will be considered and presented in Chapter 5. A utility function with cost decay may prove more effective if the predictability is unsatisfactory, however, in this thesis, only IC card data will be included since feeding utility value as variables are not possible with real-time case.

Chapter 5: Forecasting

5.1 Introduction

In this section, we proposed a short-term origin-destination demand forecasting in rail transit systems: parallel model architecture and gravity approach. As mentioned in Chapter 4, the predictability of the ODM is correlated to its volatility, a pre-analysis of predictability of lower resolution suggests that deterministic designs (statistical and distributive) are not suitable for the dataset used in this study. Additionally, technical outliers were detected, and the spread of ridership density suggest that multiple relationship designs may prove useful for approaching the forecasting case. Derivative of parallel model architecture is proposed and tested. Comparison is made with existing literature and summary of the findings is presented.

Extra considerations in this section include real-time forecasting design, short-term forecasting cases, magnitude of scale, noise, skewness of ODM, multi-pattern ODM, and unforeseen ODM. The problem definition, from gravity model, is redesigned with vectorized gravity model and applied on concatenation and chained forecasting cases for approximated forecasts.

5.1.1 Motivation

The motivation for the study of short-term ODM forecasting are as follows:

(1) Problems in the formulation of the forecasting due to the delayed availability of OD matrices in real-time and its high dimensionality.

(2) The unpredictable characteristics in OD demand and the challenges these pose to the forecasting problem, which have not been adequately addressed in existing studies (treated as outliers).

(3) Sparsity in OD pairs, which heavily reduces the accuracy and requires an appropriate response to improve forecasting.

In view of these issues, we introduce an adjusted parallel model architecture of an origin-based vector sum projection gravity model (APMA-OVG). In the proposed model, aggregated OD data and real-time inflow information are used as inputs for the delayed availability problem, and an origin-based formulation of the forecasting problem is implemented to reduce dimensionality. APMA solves the problem of multiple OD patterns by defining the model matrices uniquely for each input; thus, while behaving similarly to k-nearest neighbor (KNN) algorithms, it relies on forming relations between input variables and not distance. APMA also works with the clustering of the data by

separating the data by magnitude and patterns based on the formulation of OVG, which solves the dense and sparse features in the OD matrix. Experiments were conducted with the proposed model on two real-world datasets containing multiple patterns in OD demand in the Bangkok subway as presented in Chapter 4. Compared to benchmarks established in recent years, satisfactory results were obtained, although a direct comparison was difficult due to being unable to train the comparison models on datasets and in the datasets. The multi-distribution of the demand density in our test datasets are more numerous than other datasets and thus, were not treated as outliers in our study.

5.1.2 Contributions

The main contributions of this study are summarized as follows:

(1) Anomalous data are introduced to test the extrapolation ability of the model and the performance is compared to several benchmarks. To the best of our knowledge, this is the first time that such an experimental case has been introduced in short-term OD forecasting outside from anomalous studies.

(2) Relationships between trip generation and historical OD matrices are developed based on the simple-to-understand gravity model by considering their dependencies. We proposed a delayed OD matrix input constraint to formulate a data availability scenario fitting a real-world case. This constraint is also used as the basis for considering forecast scenarios due to its mechanism of defining the interval gap between real-time input and the earliest possible forecast target.

(3) APMA is introduced to separate inputs based on magnitude and pattern for the formulation of the OVG. In general, temporal indexing suits data with gradual changes over time, but in the case of anomalous changes, APMA is introduced to avoid “expired” models.

5.2 Methodology

This section presents the methodology, algorithms, and key concepts of the proposed method. First, the short-term OD forecast is defined and the relationship between trip generation and trip attraction is introduced. Next, the formulation of the OVG equations from the gravity model is presented following the adoption of APMA from parallel model architecture (PMA). The APMA-OVG comprises two forecasting cases of historical and real-time data, respectively.

5.2.1 Model development

The forecast framework is shown in Fig. 13. Network constraints are used to calculate the delay data availability constraint, in addition to historical data and real-time ODM from data collection, the OVG is defined and splitting of the data is processed. Next, the APMA process the large input into multiple clusters and the data clusters went into data refinement, model preparation, and model matrix defining. The parallel forecasting is then conducted, and performance evaluation is given for comparison with existing literature.

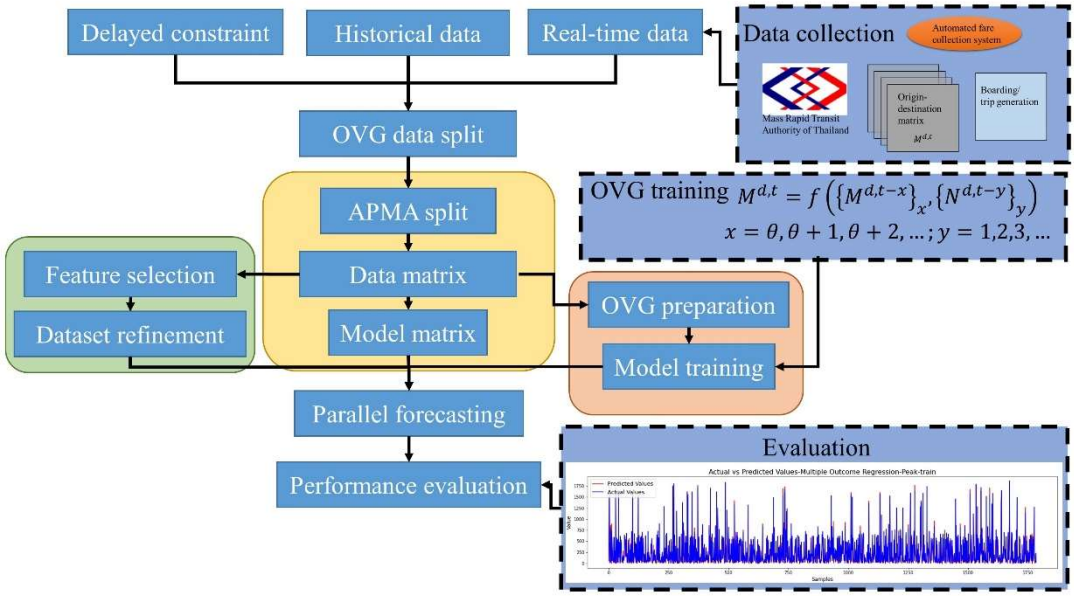

Figure 13: Framework for APMA-OVG.

5.2.2 Origin-based vector sum projection gravity model

The gravity model is a statistical model that uses OD data to predict OD pairs. Non-linear relationships between variables can be transformed using available feature space transition techniques. The gravity model with the distance decay assumption is presented in Eq. (15) [3]:

$$m_{pq}^{d,t} = \frac{\alpha}{dis_{pq}^2} n_p^{d,t} \sum_i m_{iq}^{d,t} \quad (15)$$

where α is the homogeneous parameter, $\sum_i m_{ij}^{d,t}$ is the trip attraction, and the column sum of matrix $M^{d,t}$, and dis_{pq} is the physical distance between origin p and destination q.

To formulate a real-time forecast based on the gravity model, we first define a pseudo-dimensional space while assuming that the actual locations of each node are unknown, as shown in Fig. 14.

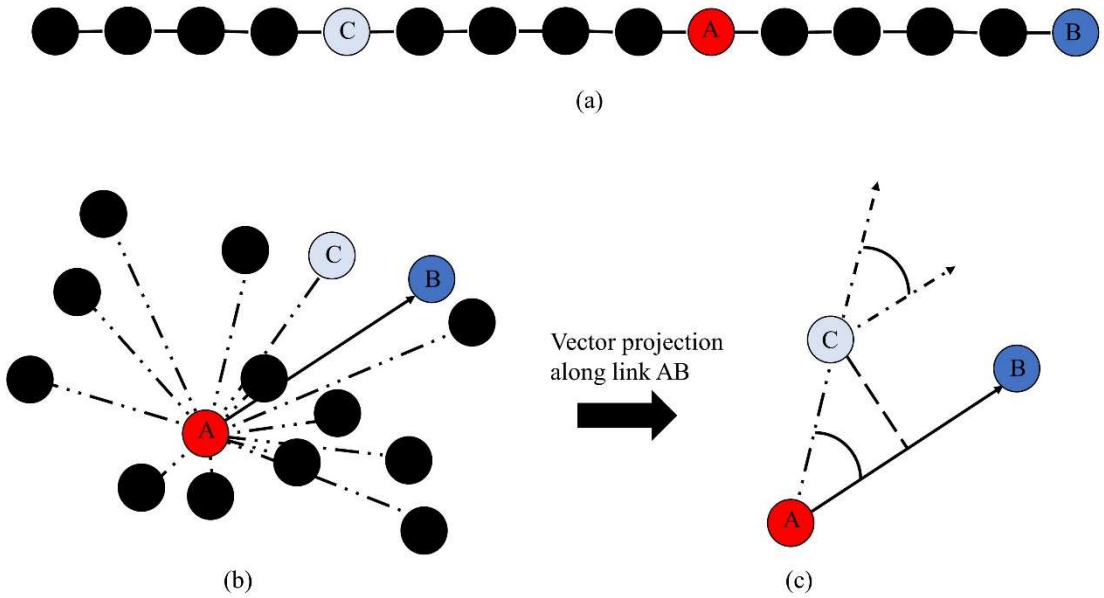

Figure 14: A visualization of nodes in physical distance space is shown in (a) pseudo-dimensional space is shown in (b). Node A is the origin node and node B is the destination node. The OD pair from node A to node B is marked with a solid line. An example of vector projection of node C is shown in (c).

Since the locations in dimension space are unknown, we substitute Eq. (15) with the vector form broken down for each node in relation to Eq. (6). Thus, from Eq. (15), the new gravity equation is given by

$$m_{pq}^{d,t}=\sum_x \sum_i \frac{\alpha_{iq}^{d,t-x} n_{pq}^{d,t-x} m_{iq}^{d,t-x}}{\left(dis_{iq}^{d,t-x}\right)^2}, x=\theta,\theta+1,\theta+2,\dots;i=1,2,\dots,n \quad (16)$$

Here, the homogeneous parameter α is transformed into non-homogenous parameter specific to each OD pair. From Eq. (16), it can be inferred that, by extension, the equation can be expressed as a linear equation, as shown in Eq. (17).

$$m_{pq}^{d,t}=\sum_x \sum_i \beta_{iq}^{d,t-x} m_{iq}^{d,t-x}, x=\theta,\theta+1,\theta+2,\dots;i=1,2,\dots,n \quad (17)$$

with the coefficients

$$\beta_{iq}^{d,t-x}=\frac{\alpha_{iq}^{d,t-x} n_{pq}^{d,t-x}}{\left(dis_{iq}^{d,t-x}\right)^2}.$$

Finally, we introduce $n_{pq}^{d,t-y}, y < x$ into Eq. (17) and the completed OVG that satisfies Eq. (6) is given by Eq. (18).

$$m_{pq}^{d,t} = \sum_x \sum_i \beta_{iq}^{d,t-x} m_{iq}^{d,t-x} + \sum_y \beta_p^{d,t-y} n_p^{d,t-y}, x = \theta, \theta+1, \theta+2, \dots; i = 1,2, \dots, n; y = 1,2,3, \dots; y < x \quad (18)$$

5.2.3 Adjusted parallel model architecture

PMA is a strategic organization of the models designed to maximize the performance within constraints at any instance of time. This strategy was first proposed to improve the accuracy of electricity load forecasting [42]. Later, an addition of an annual trend weighted forecasting step were applied the PMA strategy to forecast gas consumption [43]. PMA is unsuitable for application to relatively complex models due to its restriction on data volume. The algorithm can be generally described by the following steps:

(1) Split the data into sub-datasets according to temporal index.

(2) Predict each sub-dataset with a prepared model.

(3) Summarize and rearrange the final inputs.

The difference between PMA and APMA lies in the data splitting criteria, as shown in Fig. 15, PMA uses temporal indexing for the time series (months are most used for temporal indexing), which has a major drawback of significantly relying on large-scale temporal variables and potentially splitting the data into too many sets. In contrast, APMA relies on a combination of information on historical data, real-time data, and delayed constraints to perform k-clustering on the assumption that the output is unknown for the real-time data since it is a forecast of a future OD matrix. The Monte Carlo method is then applied with minimization of the weighted absolute percentage error (WAPE) as the objective to maximize predictability. Note that root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R^2) are bias errors, and thus, not suitable for maximizing predictability. RMSE and MAE heavily depend on magnitude of the OD pairs and R^2 is an indicator for the proportion of explained variance to total variance which does not reflect predictability.

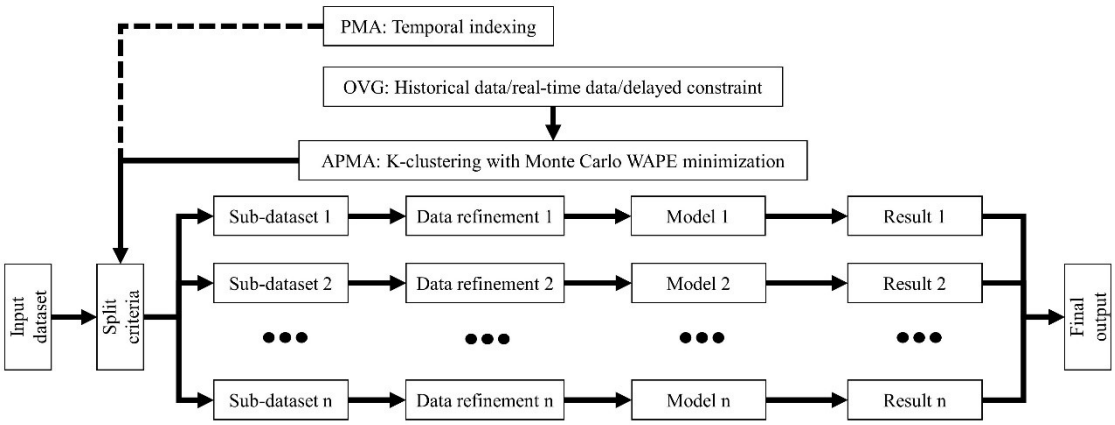

Figure 15: Overview of PMA and APMA.

5.3 Experiment

For this section, we tested the proposed method with two real-world datasets and compared it with benchmark methods. An analysis is also given in this section.

5.3.1 Two forecasting cases

The delayed availability OD data play a significant role in our problem definition. We design two forecasting cases to clarify the appropriateness of the OVG formulation. The first case, the concatenation case, involves using all OD and boarding data available from the start of the operation to make a direct forecast of the ODM period of interest. The second case, the chained case, involves modeling the time-step forecasting to estimate missing OD matrices within the interval gap θ ($\theta = 2$ in this study). Eq. (18) transformed into $m_{pq}^{d,t} = \sum_x \sum_i \beta_{iq}^{d,t-x} m_{iq}^{d,t-x}, x = 1,2,3,\dots; i = 1,2,\dots,n$ in the final forecast of the chained case, as shown in Fig 16.

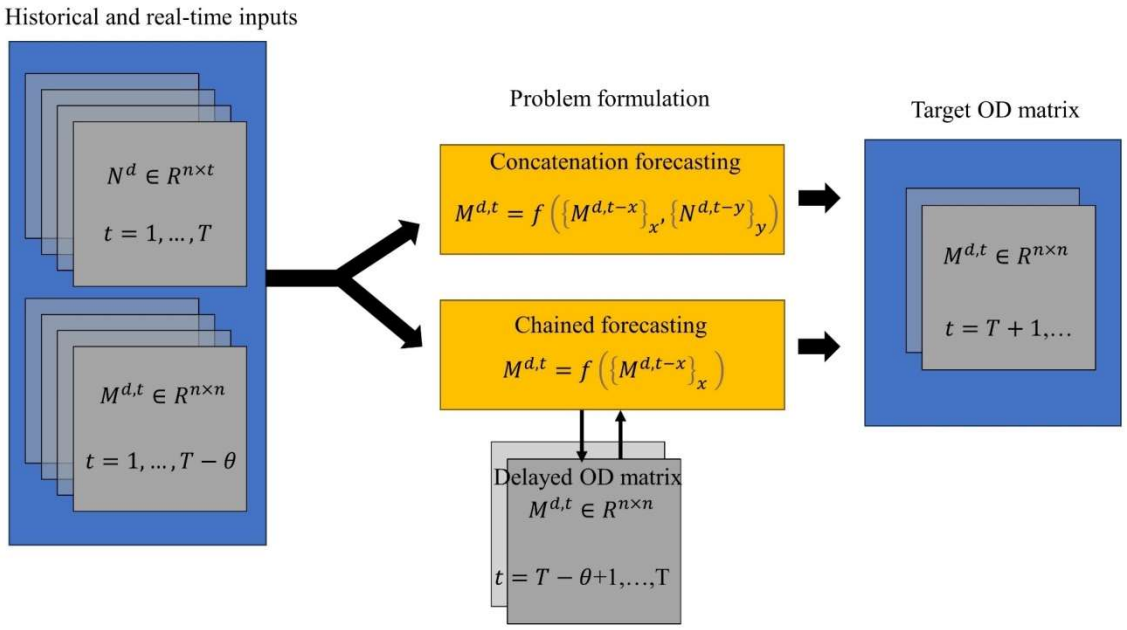

Figure 16: Overview of the two forecasting cases.

The implementation of the models was coded in Spyder 5.4.2 (Python 3.8.5) and the amount of training data was half of each entire dataset and based on each forecasting case. The hyperparameters and parameters were not updated in the testing period. P502 dataset is provided in the discussion of the method's limitations.

5.3.2 Evaluation methods

Indicators MAE, RMSE, WAPE, and R^2 for matrix evaluation are respectively given in Eqs. (19) to (22).

$$\mathrm{MAE}(\omega, \hat{\omega}) = \frac{1}{N}\sum_{i=1}^{N}|\omega_i - \hat{\omega}_i| \quad (19)$$

$$\mathrm{RMSE}(\omega, \hat{\omega}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\omega_i - \hat{\omega}_i)^2} \quad (20)$$

$$\mathrm{WAPE}(\omega, \hat{\omega}) = \frac{\sum_{i=1}^{N}|\omega_i - \hat{\omega}_i|}{\sum_{i=1}^{N}|\omega_i|} \times 100\% \quad (21)$$

$$\mathrm{R}^2(\omega, \hat{\omega}) = 1 - \frac{\sum_{i=1}^{N}(\omega_i - \hat{\omega}_i)^2}{\sum_{i=1}^{N}(\omega_i - \bar{\omega}_i)^2} \quad (22)$$

Since MAE, RMSE, and R^2 are bias indicators, comparison between datasets use mainly WAPE.

5.3.3 Result and discussion

The conceptual design of the forecasting study is based on data's characteristics. Comparison in this section is purely on performance of the models. The data characteristics in referred literature are unspecified but conjectures can be made with the usage of the models. Most of the methods reported in the literature like the temporal regularized matrix factorization (TRMF) [44] and ConvLSTM [45] cannot use trip generation as extra inputs in their formulation and must assume all historical OD matrices are known which is not the case for our design. Regarding the requirements of the models, the choice of models for benchmarking is lacking. ConvLSTM has a strict chronological order and requires denser data than OD matrices. Since the test dataset is expected to be highly sparse, the chronological order is not obvious enough and will poorly represent ConvLSTM's performance. Lastly, the simplicity and complexity of the models are problematic. Relatively more complex models use higher data volumes, so when the data volume is limited, a simpler model might perform better. However, the simpler models do not account for special characteristics, for example, HA is limited by reliance on centric behavior. Additionally, the gravity model variations that OVG is based on are relatively static models, which are more suitable for exclusive locational motions, like immigration/movement of goods between long distances.

In view of these factors, the comparison will focus on predictive ability. Multiple linear regression (MLR) is used to solve the APMA-OVG problem with comparison to categories of designed regressors that solve OVG: k-nearest neighbors (KNN), support vector regressor (SVR), and light gradient boosting machine (LGBM). KNN solves the regression problem by finding the minimum distance between the input and the training dataset to infer the regression result, while SVR finds a hyperplane to fit the data points instead of a line. As for LGBM, it performs efficiently for small datasets with a tree ensemble algorithm. KNN, SVR, and LGBM can all model non-linear relationships.

5.3.3.1 Application of feature selection algorithms

Table 9 summarizes the breakdown of the data refinement process of APMA-OVG. Various feature selection methods are chosen, namely, PCA for covariance analysis, kernel PCA (kPCA) for non-linear feature space transformation, incremental PCA (iPCA) for high volume PCA, sparse PCA (sPCA) for locating sparse components and reconstructing data, non-negative matrix factorization (NMF) for non-negative data, and truncated singular value decomposition (TSVD) for data matrix factorization. Each algorithm tests properties of the OD matrices data and evaluates their suitability. The results for OVG on the P002 dataset are summarized in Table 9 and visualized in Fig. 17.

Table 9: Performance comparison of different feature selection algorithms.

Feature selection	Concatenation forecasting				Chained forecasting			
	MAE	RMSE	WAPE	R2	MAE	RMSE	WAPE	R2
None	*22.44*	*44.66*	*38.59%*	*0.7732*	*25.7*	*51.99*	*44.19%*	*0.7034*
PCA	*22.47*	*44.56*	*38.65%*	*0.7743*	*25.56*	*51.61*	*43.95%*	*0.7056*
kPCA	***18.84***	***36.18***	***32.40%***	***0.8327***	***19.07***	***36.85***	***32.79%***	***0.8238***
iPCA	*22.88*	*45.51*	*39.35%*	*0.7624*	*25.26*	*51.46*	*43.44%*	*0.7063*
sPCA	*22.94*	*45.74*	*39.46%*	*0.7628*	*25.48*	*51.87*	*43.81%*	*0.7041*
NMF	*22.78*	*45.58*	*39.18%*	*0.7658*	*25.55*	*51.94*	*43.94%*	*0.7037*
TSVD	***19.41***	***37.26***	***33.38%***	***0.8256***	***20.84***	***39.95***	***35.83%***	***0.8014***

Our preliminary assumption is that either a feature selection algorithm for sparseness or a non-linear feature space will result in higher performance. Since the Monte Carlo method is applied to the model setup, the feature selection algorithm is not homogeneous but is set to minimize WAPE.

(1) For the concatenation case, the results show that kPCA and TSVD respectively reduce the WAPE error by 16.04% and 13.50%, reduce RMSE by 18.99% and 16.57%, reduce MAE by 16.04% and 13.50%, and increase R^2 by 7.70% and 6.78%.

(2) For the chained case, the results show that kPCA and TSVD respectively reduce the WAPE error by 25.80% and 18.92%, reduce RMSE by 29.12% and 23.16%, reduce MAE by 25.80% and 18.91%, and increase R^2 by 17.12% and 13.93%.

We can conclude that the low covariance in relation to temporal features between matrices is the main factor reducing the predictability of P002 in this case. Results for P502 were inconclusive due to its predictability reported in Chapter 4, so we excluded them from the report.

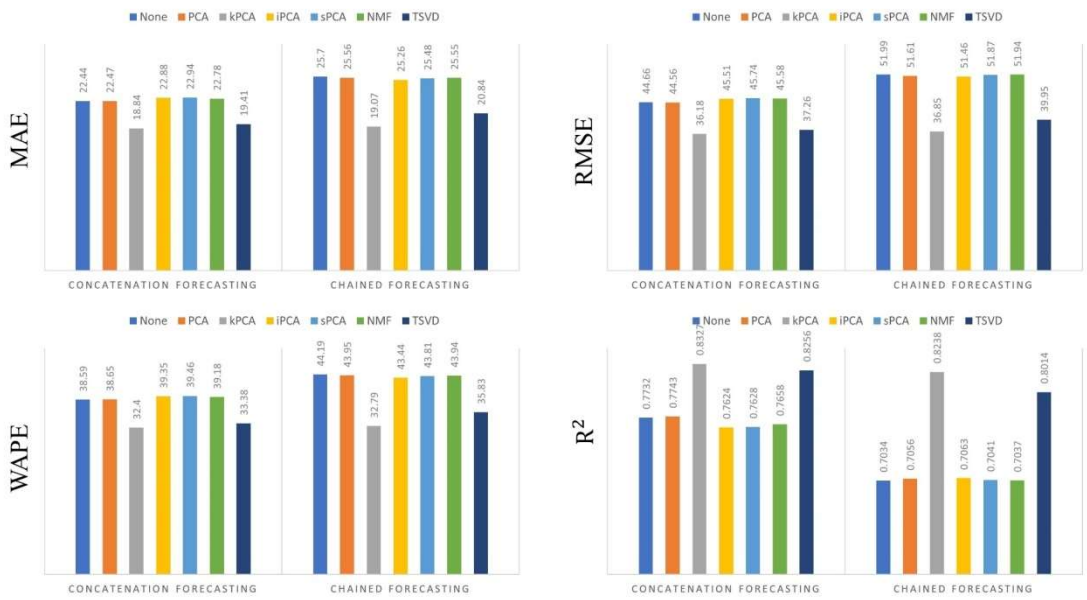

Figure 17: Performance comparison of different feature selection algorithms.

5.3.3.2 APMA-OVG

We show four randomly selected OD pairs to illustrate the categorical data ordering of APMA in Fig. 18. Each OD pair is separated by the OD flow following the pattern of its corresponding cluster. One disadvantage of APMA is the emphasis on categorical order, which results in ignoring chronological order. However, for the datasets tested in this thesis, APMA can establish an appropriate order to the data by abandoning the chronological order. Five clusters are shown, and four randomly selected OD pairs are split after the data were processed.

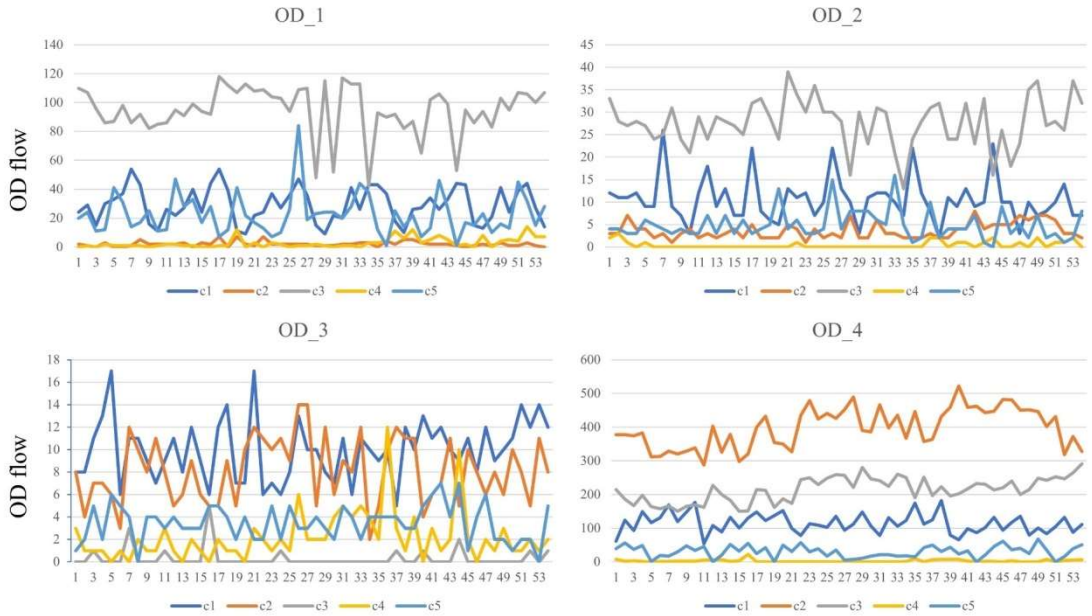

Figure 18: OD flow of randomly selected OD pairs from different clusters.

Next, the PMA, SVR, KNN, KGBM, and APMA algorithms are compared in terms of the overall performance, and a silhouette δ lower bound is made for the APMA algorithm. The comparison results are summarized in Table 10 and visualized in Fig. 19.

Table 10: Comparison of APMA-OVG with different clustering tolerances, PMA, and other models.

	Concatenation forecasting				Chained forecasting			
Model	MAE	RMSE	WAPE	R^2	MAE	RMSE	WAPE	R^2
MLR	*22.44*	*44.66*	*38.59%*	*0.7732*	*25.7*	*51.99*	*44.19%*	*0.7034*
PMA	*17.74*	*33.83*	*30.15%*	*0.8486*	*21.35*	*42.22*	*36.29%*	*0.7567*
APMA*(0.4< δ < 0.6)*	*16.55*	*32.41*	*26.9%*	*0.8676*	*21.08*	*42.8*	*34.26%*	*0.7718*
APMA ($\delta \geq 0.6$)	*16.25*	***31.69***	*26.93%*	*0.8693*	*20.46*	*41.37*	*33.92%*	*0.7789*
APMA-OVG	***14.57***	*32.37*	***24.16%***	***0.8729***	*18.49*	*37.37*	*30.64%*	*0.8073*
SVR	*21.66*	*50.82*	*41.41%*	*0.6563*	*23.06*	*64.55*	*39.65%*	*0.5993*
LGBM	*16.1*	*32.18*	*30.77%*	*0.7848*	***12.68***	***28.46***	***21.8%***	***0.8837***
KNN	*18.36*	*38.71*	*35.09%*	*0.7481*	*14.62*	*31.37*	*25.14%*	*0.8564*

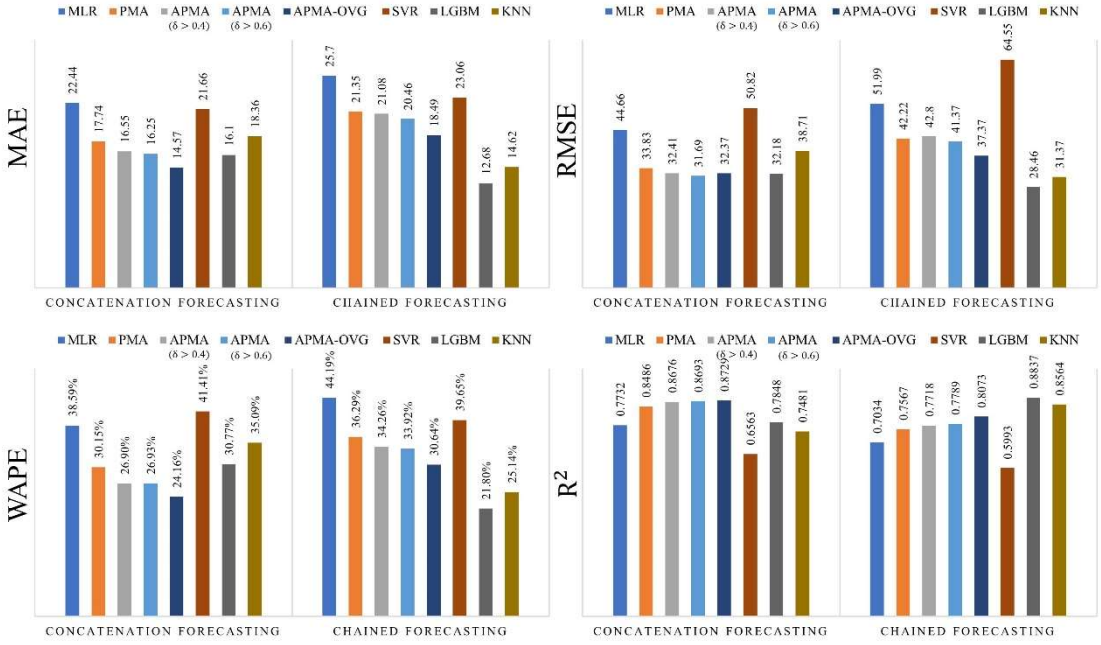

Figure 19: Overall performance comparison .of tested models.

The Table 10 and Fig. 19 can be summarized as follows:

(1) The minimization of WAPE resulted in a loose silhouette score that sets the lower limit for APMA. The APMA-OVG model selected kPCA for minimizing WAPE for both forecasting cases, similarly to APMA. APMA-OVG in the concatenation case has the lowest WAPE of 24.16%, including the lowest MAE and highest R^2. When compared to PMA, APMA with a minimum $\delta > 0.4$, and APMA with a minimum $\delta \geq 0.4$, WAPE is respectively lowered by 19.87%, 10.19%, and 10.29%, MAE is respectively lowered by 17.87%, 11.96%, and 10.34, and R^2 is respectively increased by 2.86%, 0.61%, and 4.14%.

(2) However, for the concatenation case of SVR, LGBM, and KNN, although their performance is better than that of the base MLR model, compared to the PMA variation models, lower performance is observed.

(3) For the chained forecasting case, based on the formulation of OVG, while APMA-OVG has the best performance among similar model types and higher performance than SVR, the best model is LGBM, followed by KNN.

(4) The chained forecasting case is expected to be more difficult to solve due to involving predicting delayed OD matrices (by replacing $\sum_y \beta_p^{d,t-y} n_p^{d,t-y}$

with an estimated $m_{iq}^{d,t}$). Table 11 summarizes the chained forecasting performance within the interval gap. The estimation of OD matrices within the interval gap of APMA-OVG, when compared to LGBM and KNN, has significantly worse performance, with higher WAPE for first- and second-time steps, respectively, by 18.31% and 25.76% compared to LGBM, and 22.35% and 56.11% when compared to KNN.

(5) KNN and LGBMR share similar traits to APMA-OVG. While KNN search for some historical data closest to the input, LGBMR is an ensemble algorithm. APMA-OVG categorize data and define strict relationships between variables. The difference in performance in the chained case is due to the "rigidity" of APMA-OVG since both KNN and LGBMR are black-box regressors.

Table 11: Performance comparison of delayed OD matrices estimation of APMA-OVG, LGBM, and KNN.

	first-time step				second-time steps			
Model	MAE	RMSE	WAPE	R^2	MAE	RMSE	WAPE	R^2
APMA-OVG	2.29	27.76	29.34%	0.6562	18.54	37.27	30.74%	0.8197
LGBM	1.9	16.93	24.8%	0.7637	13.27	28.94	22.82%	0.8761
KNN	1.84	16.42	23.98%	0.757	11.45	27.44	19.69%	0.8773

5.3.3.3 Analysis of APMA-OVG

Four randomly selected OD pairs' performance is discussed in this section. The comparison between actual and forecasting values is shown in Fig. 20. OD_1 has the lowest average actual value, and OD_4 has the largest range of actual value. APMA-OVG gives importance to OD pairs with higher actual value. The variance of the actual value is high for OD_2 to OD_4 when compared to OD_1. For both concatenation and chained cases, the forecasting values follow a similar trend due to the same design of the OVG model. This trend includes when the prediction is significantly deviating from the actual value. However, chained case exhibits higher deviations in most OD pairs due to the estimations of missing OD matrices. In addition, for relatively low actual value (OD_1), deviation is apparent when compared to concatenation case. The issue of numerous small OD flows was discussed in both [1] and [4]. As can be seen in OD_2 through OD_4, the fluctuation of OD flow is significant to generalize the model to assign ranking of OD pairs to address this issue.

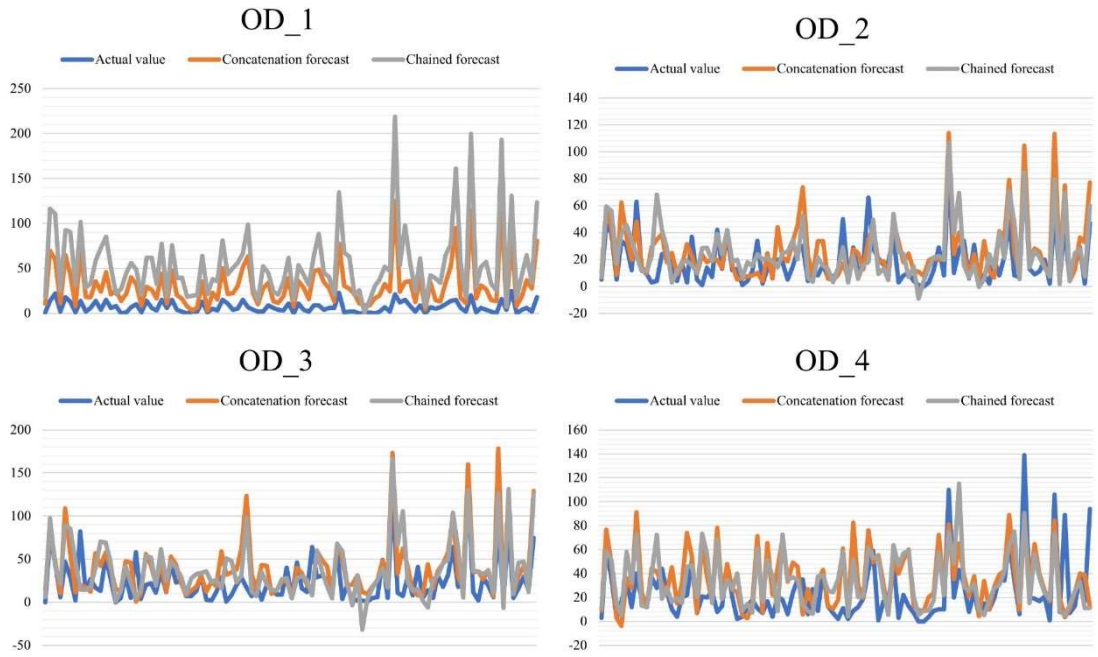

Figure 20: Comparison of actual and forecasting flows of randomly selected OD pairs.

In Fig. 21, the boxplot of forecast deviations shows a normal distributed behavior with chained case exhibit larger ranges of deviation on each OD pair. The outliers of the boxplots, however, are more balanced with chained case. For, OD_1, as explained above, because APMA-OVG gives less importance to OD pairs with small OD flow, distribution of deviations deviates from zero and the forecast value is higher than actual value. In comparison to OD_1, the normal distributed deviations of OD_2 to OD_4 are more zero centric, especially OD_2. The reason being that OD_2 has the lowest fluctuation of OD flow while not having small OD flow. The outliers in the boxplot results from the extremity within the fluctuation of significantly higher or lower actual OD flow. It is worth exploring the addition of external data such as weather data, or local events for improving the model to solve the issue of predictability of extreme fluctuation within OD pair. Whether the issue of small OD flow will also be addressed is difficult to say but within the scope of anomalous data, such fluctuations should be given higher priority.

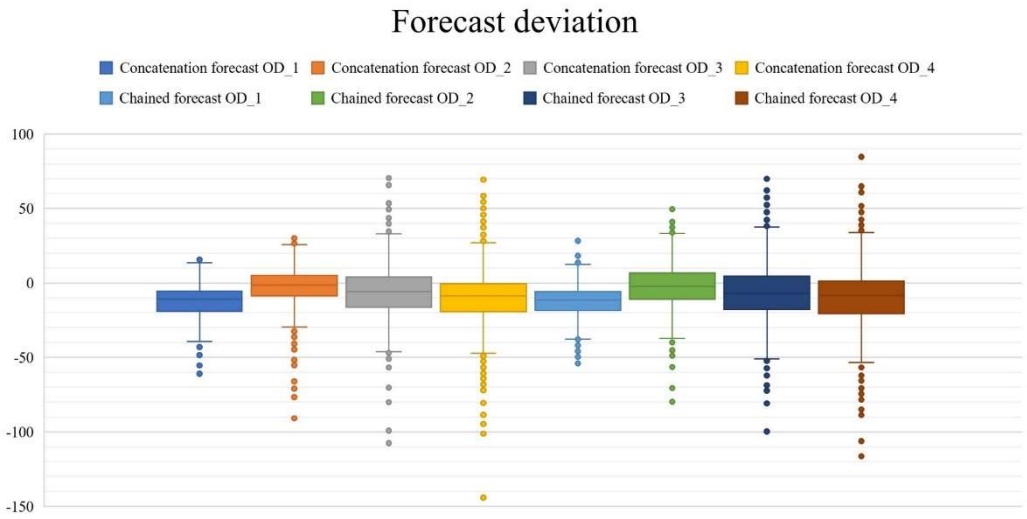

Figure 21: Boxplot of forecast deviation of randomly selected OD pairs (same as Fig. 20).

5.3.3.4 Comparison with other literature

The comparison of model performance is shown in Table 12. We compare APMA-OVG with the benchmarks established by [1] and [4]. WAPE, dataset, models, and their assumptions are summarized and presented. The analyses of the datasets (Guangzhou, Hangzhou, and Beijing) used in comparison are not presented in the literature. While the APMA-OVG has the lowest error among all the model, the suitability between model's usage and dataset's characteristics is missing so we can only conclude that the performance is satisfactory.

Guangzhou, Hangzhou, and Beijing dataset are dataset of stable operation while Bangkok dataset exhibits anomalous demand behavior. Additionally, the size of OD matrices and relative density of usage varies. Finally, the models of the benchmarks are not applicable to the Bangkok dataset due to 1) The requirements of benchmark models are not designed for anomalous data; 2) The complexity of the models are not suitable for the size of Bangkok dataset. While there are no range of acceptable performance indicator established for the short-term forecast, based on the benchmarks' performance, the WAPE of APMA-OVG suggests that it falls within acceptable range of error. In addition, when consider basic model like HA to the other models in literature, the improvement of the ODM estimation problem does not significantly differ from static estimation method. Henceforth, APMA-OVG can be expected to give good enough accuracy using datasets of networks with stable operation. Under stable operation, APMA-OVG may be reduced to a derivation of gravity model due to the centric of the data.

Table 12: Performance comparison between benchmarks and APMA-OVG.

Model	Dataset	WAPE	Usage
HW-DMD	Guangzhou	29.65%	small and zero flows
	Hangzhou	31.76%	
TRMF	Guangzhou	30.61%	Strict chronological order
	Hangzhou	34.02%	
FNN	Guangzhou	30.23%	N/A
	Hangzhou	33.58%	
HA	Guangzhou	31.21%	Strict demand behavior order
	Hangzhou	34.28%	
Conv-LSTM	Guangzhou	30.11%	Dense OD flow Strict chronological order
	Hangzhou	32.92%	
	Beijing	26.98%	
2D CNN	Beijing	26.94%	N/A
3D CNN	Beijing	27.00%	N/A
ConvGRU	Beijing	27.10%	Dense OD flow Strict chronological order
TrajGRU	Beijing	29.46%	Location variant pattern
CAS-CNN	Beijing	26.10%	small and zero flows
APMA-OVG	***Bangkok***	***24.16%***	***Anomalous demand behavior***

5.3.3.5 Forecasting cases

The tested model with highest performance for the concatenation case and the chained case are respectively, APMA-OVG and LGBMR. While the performance of the chained case is slightly better than the concatenation case, it is inappropriate to establish which forecasting case is superior. The interval gap θ in problem formulation (in this case is 2) can be considered relatively small. For larger θ, the increase in delayed data availability to each forecasting case results in the following:

(1) For the concatenation case, the number of input variable is decreased by $n-1$ for every increase of θ by 1 ($M^{d,t}$ is replaced with $N^{d,t}$).

(2) For the chained case, the number of estimations needed to transform Eq. (6)

to $M^{d,t} = f(\{M^{d,t-x}\}_x)$ increases along with θ.

The earliest known ODM is the most important input for the forecasting problem. The correlations between ODMs are an important part and the usage of incomplete ODM should be considered in future research. Incomplete ODM is included in Eq. 18 of concatenation case. Usefulness in prediction of incomplete ODM is reported but the extend of its usefulness was not verified with combinations of earlier ODM inputs yet [46].

5.3.3.6 Limitations

The APMA-OVG was designed to fit the characteristics of dataset P002. The analysis in Chapter 3 shows that the data, when divided annually, has unstable total demand (summation of ODM). As shown in Table 12, Historical Average model gives higher WAPE compared to other models on dataset Guangzhou and Hangzhou. However, the difference in WAPE is less than 5% and 7.5% respectively for Guangzhou and Hangzhou datasets. The dataset used in this thesis is severely different from the dataset used in other literature and cannot be applied to other models without assumptions. In addition, the temporal prediction constraint used in this study is, in other literature, ignored or assumed negligible, making the inputs less temporally completed when compared to other studies.

The problem of predictability is a major problem since suitability between model and data is more important than the complexity of the model. To explain this aspect, the P502 dataset is explained. P502 was recorded from the start of operation in December 2019 until the end of 2020. The decision to include this dataset, which exhibits unstable ridership characteristics in addition to effects from the 2020 global travel restriction, was to show how predictable the proposed model is given different anomalous data.

Compared to P002, P502 was introduced very late, and has sparse OD pairs that comprise more than 50% of the OD matrices on average. P502 also contains significant ridership demand's shifts throughout the dataset with no clear tendency. Table 13 lists the best performing models, with the very best model being SVR with 40.36% and 47.27% WAPE error for the concatenation and chained forecasting cases, respectively. MAE and RMSE for this model are lower, but as stated before, both MAE and RMSE are bias indicators that depend on the magnitude of OD flows, with averagely low ridership of P502 compared to P002, it is understandable that those bias indicators have lower values. R^2 values for both forecasting cases are not satisfactory except for the PMA model on the chained forecasting case, though not when considering other indicators. APMA-OVG is excluded due to unsatisfactory performance. Relatively complex deep learning models are expected not to perform well on P502 due to the limited data volume.

Although we proposed APMA-OVG, which is appropriate for the type of dataset P002 belongs to, namely one that needs categorical more than chronological ordering, tests on P502 reveal the gap in predictability of extreme fluctuations. Thus, for short-term forecasting, suitability should be determined based on the datasets, especially those in which non-standard operation is observed.

Table 13: Performance comparison of delayed OD matrices estimation on dataset P502.

	Concatenation forecasting				Chained forecasting			
Model	MAE	RMSE	WAPE	R^2	MAE	RMSE	WAPE	R^2
SVR	7.53	16.97	40.36%	0.6211	8.82	20.06	47.27%	0.56
KNR	8.39	19.76	44.96%	0.5324	9.44	20.85	50.62%	0.2188
LGBMR	8.21	18.41	44.03%	0.2947	8.85	20.27	47.46%	0.1871
PMA	10.8	19.04	56.11%	0.1477	12.39	33.79	52.47%	0.727

Chapter 6: Conclusion

This study proposed forecasting model from the analysis of dataset's characteristics. Statistical and distributive's predictabilities were considered from multiple level of origin-destination perspectives (total demand, boarding, and origin-destination pair). The logical design of the proposed model considered the real-time short-term forecasting including, dense and sparse segments, constraints for magnitude of the demand and delay data availability, estimation of the delayed data availability, and classification of the data points.

The IC card datasets in this study is relatively unstable and problematic limiting options for usable models. After many considerations, APMA-OVG model was designed and tested. The proposed model is expected to be helpful when forecasting is needed during disruptions from stable operation is expected. The trained model is simple and easy to understand. Advanced models may struggle with extrapolating beyond range of inputs. In addition, short-term ODM forecasting under anomalous demand behavior may prove useful for the significant infrastructural changes to predict the demand until the changing of ridership reach saturation. The APMA-OVG model consisting of: APMA algorithm, delayed data availability constraint, data refinement, MLR, and OVG formulation. OVG is derived from gravity model and is a simple linear model. In our experiments, we collected historical datasets that reflect a large difference in relative network demand concentration, serving as testing datasets where the distribution under temporal variable deviated from standard operation. The conclusions of the present study can be summarized as follows:

(1) Adjusted parallel model architecture is a useful tool for improving accuracy when the data's characteristics call for categorical order instead of temporal order. OVG formulation is a simple formulation derived from gravity model that includes data availability delays, avoid input overlaps, and consider the effects of all destination nodes without homogeneity assumption.

(2) The generated hybrid model, APMA-OVG, provides satisfying performance comparable to some advanced models. While the usage of APMA-OVG is different from other advanced models, the performance suggests that suitability should be prioritized over complexity of the model.

This study concludes that effective forecaster for short-term ODM forecasting should take suitability to the instances of the forecasting targets as the priority when multiple patterns appear in OD demand. This study excludes external variables like

weather, seasonal, events, etc., also, questions remain for how to handle extremity of data, and how to detect changes that indicates need for switching between chronological and categorical order. Mathematical method or empirical method for evaluating these issues are needed to improve the model for higher accuracy and more interpretable forecast. The generality of the model is expected to give a good enough accuracy even for the case of stable range of demand's magnitude. However, the accuracy of the forecast, in comparison to other models, would not necessary be better due to the specific usage of the APMA-OVG for the anomalous events.

Reference

[1] J. Zhang, H. Che, F. Chen, W. Ma, and Z. He, "Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method," *Transp. Res. Part C Emerg. Technol.*, vol. 124, pp. 1–30, 2021, doi: 10.1016/j.trc.2020.102928.

[2] L. Liu, Z. Qiu, G. Li, Q. Wang, W. Ouyang, and L. Lin, "Contextualized Spatialoral Network for Taxi Origin-Destination Demand Prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3875–3887, 2019, doi: 10.1109/TITS.2019.2915525.

[3] I. Ekowicaksono, F. Bukhari, and A. Aman, "Estimating Origin-Destination Matrix of Bogor City Using Gravity Model," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 31, no. 1, 2016, doi: 10.1088/1755-1315/31/1/012021.

[4] Z. Cheng, M. Trépanier, and L. Sun, "Real-Time Forecasting of Metro Origin-Destination Matrices with High-Order Weighted Dynamic Mode Decomposition," *Transp. Sci.*, vol. 56, no. 4, pp. 904–918, 2022, doi: 10.1287/trsc.2022.1128.

[5] X. Ling, Z. Huang, C. Wang, F. Zhang, and P. Wang, "Predicting subway passenger flows under different traffic conditions," *PLoS One*, vol. 13, no. 8, p. e0202707, Aug. 2018, doi: 10.1371/JOURNAL.PONE.0202707.

[6] Z. Zeng and T. Li, "Analyzing Congestion Propagation on Urban Rail Transit Oversaturated Conditions: A Framework Based on SIR Epidemic Model," *Urban Rail Transit*, vol. 4, no. 3, pp. 130–140, 2018, doi: 10.1007/s40864-018-0084-6.

[7] I. G. Macola, "Analysing Covid-19's impact on the UK rail supply sector." https://www.railway-technology.com/analysis/analysing-covid-19-impact-on-the-uk-rail-supply-sector/ (accessed May 28, 2022).

[8] A. Tardivo, A. C. Zanuy, and C. S. Martín, "Covid-19 impact on transport: A paper from the railways' systems research perspective," *Transp. Res. Rec.*, vol. 2675, no. 5, pp. 367–378, 2021, doi: 10.1177/0361198121990674.

[9] J. Hora, T. G. Dias, A. Camanho, and T. Sobral, "Estimation of Origin-Destination matrices under Automatic Fare Collection: The case study of Porto transportation system," *Transp. Res. Procedia*, vol. 27, pp. 664–671, 2017, doi: 10.1016/j.trpro.2017.12.103.

[10] P. Kumar, A. Khani, and Q. He, "A robust method for estimating transit passenger trajectories using automated data," *Transp. Res. Part C Emerg. Technol.*, vol. 95, no. February, pp. 731–747, 2018, doi: 10.1016/j.trc.2018.08.006.

[11] A. R. Pitombeira Neto, F. M. Oliveira Neto, and C. F. G. Loureiro, "Statistical

models for the estimation of the origin-destination matrix from traffic counts," *Transportes*, vol. 25, no. 4, p. 1, 2017, doi: 10.14295/transportes.v25i4.1344.

[12] Y. Chen, F. Ord, and K. Palmer, "Confidence Intervals for OD Demand Estimation," no. December, pp. 1–30, 2006.

[13] W. Ma and Z. (Sean) Qian, "Statistical inference of probabilistic origin-destination demand using day-to-day traffic data," *Transp. Res. Part C Emerg. Technol.*, vol. 88, no. August 2017, pp. 227–256, 2018, doi: 10.1016/j.trc.2017.12.015.

[14] M. Fernández, E. L. Huamaní, A. Fernández, and A. Roman-Gonzalez, "Application of piecewise linear approximation method for the estimation of origin-destination matrix," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 674–680, 2020, doi: 10.14569/IJACSA.2020.0110487.

[15] P. G. Furth and D. S. Navick, "Bus route OD matrix generation: Relationship between biproportional and recursive methods," *Transp. Res. Rec.*, no. 1338, pp. 14–21, 1992.

[16] M. Gallo, G. De Luca, L. D'Acierno, and M. Botte, "Artificial neural networks for forecasting passenger flows on metro lines," *Sensors (Switzerland)*, vol. 19, no. 15, pp. 1–14, 2019, doi: 10.3390/s19153424.

[17] I. Geva, E. Hauer, and U. Landau, "Maximum-Likelihood and Bayesian Methods for the Estimation of Origin-Destination Flows.," *Transp. Res. Rec.*, vol. c, pp. 101–105, 1983.

[18] P. G. Furth, "Zonal Route Design for Transit Corridors.," *Transp. Sci.*, vol. 20, no. 1, pp. 1–12, 1986, doi: 10.1287/trsc.20.1.1.

[19] N. S. Hadjidimitriou, M. Lippi, and M. Mamei, "A Data Driven Approach to Match Demand and Supply for Public Transport Planning," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6384–6394, 2021, doi: 10.1109/TITS.2020.2991834.

[20] P. G. Furth, "Alternating Deadheading in Bus Route Operations.," *Transp. Sci.*, vol. 19, no. 1, pp. 13–28, 1985, doi: 10.1287/trsc.19.1.13.

[21] P. G. Furth, "Short Turning on Transit Routes.," *Transp. Res. Rec.*, no. 3, pp. 42–52, 1987.

[22] P. Delle Site and F. Filippi, "Service optimization for bus corridors with short-turn strategies and variable vehicle size," *Transp. Res. Part A Policy Pract.*, vol. 32, no. 1, pp. 19–38, 1998, doi: 10.1016/S0965-8564(97)00016-5.

[23] A. Ceder and H. I. Stern, "Deficit Function Bus Scheduling With Deadheading Trip Insertions for Fleet Size Reduction.," *Transp. Sci.*, vol. 15, no. 4, pp. 338–363, 1981, doi: 10.1287/trsc.15.4.338.

[24] A. A. Ceder, "Optimal design of transit short-turn trips," *Transp. Res. Rec.*, vol.

1221, no. 557, pp. 8–22, 1989.

[25] A. Almog, R. Bird, and D. Garlaschelli, “Enhanced gravity model of trade: Reconciling macroeconomic and network models,” *Front. Phys.*, vol. 7, no. MAR, pp. 1–27, 2019, doi: 10.3389/fphy.2019.00055.

[26] J. E. Anderson, “The gravity model,” *Annu. Rev. Econom.*, vol. 3, no. December 2010, pp. 133–160, 2011, doi: 10.1146/annurev-economics-111809-125114.

[27] A. A. M. Alsger, “Estimation of transit origin destination matrices using smart card fare data,” *Ph. D. Univ. Queensl.*, pp. 1–206, 2017.

[28] L. E. Carvalho and C. F. G. Loureiro, “A bayesian multinomial-poisson simplified model for network traffic inference based on link count data,” *World Confence Transp. Res. WCTR*, pp. 1–17, 2010.

[29] L. E. Carvalho, C. Felipe, and G. Loureiro, “A Dynamic Hierarchical Bayesian Model for the Estimation of day-to-day Origin-destination Flows in Transportation Networks,” *Networks Spat. Econ.*, vol. 20, no. 422464, pp. 499–527, 2020.

[30] S. Washington, M. Karlaftis, F. Mannering, and P. Anastasopoulos, *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC, 2020.

[31] D. Raghavarao, J. B. Wiley, and P. Chitturi, *Choice-based conjoint analysis : models and designs*. Chapman & Hall/CRC, 2011.

[32] E. Yao, J. Hong, L. Pan, B. Li, Y. Yang, and D. Guo, “Forecasting Passenger Flow Distribution on Holidays for Urban Rail Transit Based on Destination Choice Behavior Analysis,” *J. Adv. Transp.*, vol. 2021, 2021, doi: 10.1155/2021/9922660.

[33] J. Chen and P. Chitturi, “Choice experiments for estimating main effects and interactions,” *J. Stat. Plan. Inference*, vol. 142, no. 2, pp. 390–396, 2012, doi: 10.1016/j.jspi.2011.06.028.

[34] C. Rudloff and M. Ray, “Detecting travel modes and profiling commuter habits solely based on GPS data,” *89th Annu. Meet. Transp. Res. Board, Washington, DC*, no. September, pp. 1–15, 2010.

[35] W. E. Marshall and N. W. Garrick, “Effect of Street Network Design on Walking and Biking:,” *https://doi.org/10.3141/2198-12*, no. 2198, pp. 103–115, Jan. 2010, doi: 10.3141/2198-12.

[36] S. Ryu, “A bicycle origin-destination matrix estimation based on a two-stage procedure,” *Sustain.*, vol. 12, no. 7, pp. 1–14, 2020, doi: 10.3390/su12072951.

[37] P. Robillard, “Estimating the O-D matrix from observed link volumes,” *Transp. Res.*, vol. 9, no. 2–3, pp. 123–128, 1975, doi: 10.1016/0041-1647(75)90049-0.

[38] P. Krishnakumari, H. Van Lint, T. Djukic, and O. Cats, “A data driven method for

OD matrix estimation," *Transp. Res. Procedia*, vol. 38, pp. 139–159, 2018, doi: 10.1016/j.trpro.2019.05.009.

[39] K. Mikkonen and M. Luoma, "The parameters of the gravity model are changing - how and why?," *J. Transp. Geogr.*, vol. 7, no. 4, pp. 277–283, 1999, doi: 10.1016/S0966-6923(99)00024-1.

[40] H. R. Kirby, S. M. Watson, and M. S. Dougherty, "Should we use neural networks or statistical models for short-term motorway traffic forecasting?," *Int. J. Forecast.*, vol. 13, no. 1, pp. 43–50, 1997, doi: 10.1016/S0169-2070(96)00699-1.

[41] H. Spiess, "A maximum likelihood model for estimating origin-destination matrices," *Transp. Res. Part B*, vol. 21, no. 5, pp. 395–412, 1987, doi: 10.1016/0191-2615(87)90037-3.

[42] E. Ceperic, V. Ceperic, and A. Baric, "A strategy for short-term load forecasting by support vector regression machines," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4356–4364, 2013, doi: 10.1109/TPWRS.2013.2269803.

[43] N. Wei, L. Yin, C. Li, C. Li, C. Chan, and F. Zeng, "Forecasting the daily natural gas consumption with an accurate white- box model," *Energy*, vol. 232, p. 121036, 2021, doi: 10.1016/j.energy.2021.121036.

[44] H. Yu, "Temporal Regularized Matrix Factorization for High-dimensional Time Series Prediction," no. Nips, 2016.

[45] Z. Chao, F. Pu, Y. Yin, B. Han, and X. Chen, "Research on real-time local rainfall prediction based on MEMS sensors," *J. Sensors*, vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/6184713.

[46] R. R. Cura, R. Stickar, C. Delrieux, F. Tohmé, L. Ordinez, and D. Barry, "Modeling the origin-destination matrix with incomplete information," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10586 LNCS, pp. 121–127, 2017, doi: 10.1007/978-3-319-67585-5_13.