Predicting Next-Month Passenger Traffic at Taiwan Railway Stations to Optimize Restaurant Food Preparation

Team 5 | 111078501 Hsiang-Jung Cheng, 111078503 Yu-Kang Lai
111078506 Shu-Ting Chen, 112078510 Li-En Tsai

Executive Summary

Small and medium-sized restaurants near train stations face challenges in accurately predicting daily customer numbers, leading to food waste and operational inefficiencies. Unpredictable fluctuations in customer demand and the need to balance food freshness and waste prevention create a dilemma for these establishments. To address this, our goal is to enhance food preparation processes by providing strategic recommendations that minimize waste and optimize readiness for varying customer flows targeting small restaurants near the train stations.

We utilized Daily Passenger Traffic data from Taiwan Railway stations available at https://data.gov.tw/dataset/8792, focusing on four stations in the Hsinchu area, Hsinchu, Zhunan, Zhudong, and Neiwan. We reframed the data with a new metric, **totalPass,** representing the sum of entry and exit counts. Data from 2017 to most recent data Oct 2023, was prioritized to align with our business goal. Time plots illustrate daily passenger counts for selected stations. We decompose time series, analyze the impact for COVID-19, employ training models with data spans Nov 2017 to Oct 2022 and data spans Nov 2022 to Oct 2023 as validation. Our performance evaluation emphasizes minimizing over-forecasting in response to our stakeholders' needs. Our evaluation metrics and forecasting plot using data in Nov 2023 revealed that the forecasting can provide fine prediction for validation.

To aid restaurant decision-making, forecasts will be provided by 2 pm each Sunday. Moreover, visual representation of forecasted passenger numbers, similar to Google Maps' "Popular Times graph" will enhance accessibility. Besides, original value and percentage for better readability will be provided to enhance restaurant's intuitiveness to make decisions.
Future improvements include real-time updates, collaboration with the Taiwan Railway Administration for weekly data, and incorporating sales data to enhance forecast precision. The ultimate goal is to extend the forecasting system nationwide to benefit restaurants surrounding all train stations in Taiwan, improving operational efficiency on a broader scale.

1. Business Problem

The accurate forecasting of daily customer numbers is challenging for small and medium-sized restaurants located in proximity to train stations after interviewing four restaurant owners near Hsinchu train station. While experienced restaurant operators may cultivate an intuitive understanding of daily customer flows, unanticipated circumstances, such as sudden surges in demand unrelated to holidays or special events, disrupt their established operational routines. The preservation of dish freshness is another critical consideration, and excessive food preparation poses a noteworthy risk, especially for restaurants lacking adequate refrigeration capacity. Consequently, as articulated by the restaurant owners, many of them adopt a conservative approach to ingredient preparation to minimize food waste. However, this creates a problem. This surge not only strains their ability to fulfill the immediate needs of existing patrons but also hampers their capacity to attract and serve new customers.

In response to these challenges, the purpose of this report is to provide insightful recommendations aimed at enhancing food preparation processes for these restaurants near train stations. The focus will be on developing strategies that strike a balance between minimizing food waste and ensuring readiness for unpredictable fluctuations in customer demand, ultimately contributing to improved operational efficiency and sustained growth.

2. Forecasting Goal

Given that many small and medium-sized restaurants lack accurate daily customer records, we propose an alternative approach by utilizing the daily railway inflow and outflow data provided by the Taiwan Railway Administration (TRA). The primary focus of our forecasting initiative is directed towards restaurants situated near selected train stations in the Hsinchu area. Our forecasting horizon spans 7 days, taking into account that most food ingredients can last from several days to one week. This forecasting period can allow the restaurants to plan their inventory and restocking activities based on a reasonable timeframe that considers the freshness and quality of the ingredients. The results will consist of two parts: an accurate numerical forecast value and a percentage indicating the difference compared to the average passengers of the same weekday over the past 30 days.

The evaluation criteria for the forecasting model's performance will prioritize minimizing over-forecasting. The performance evaluation will be performance charts, affording a visual assessment of forecast accuracy. Addressing over-forecasting is particularly crucial, as it carries the risk of causing food waste—a situation restaurants are keen to avoid. Considering our plan for roll-forward forecasting, we will also take into account maintenance costs and model simplicity. By emphasizing accuracy and ease of model maintenance in our forecasting approach, our goal is to empower restaurants to make proactive decisions in managing their inventory and reducing the likelihood of unnecessary waste.

3. Data

3.1 Data Description

We obtained Daily Passenger Traffic data encompassing all 238 Taiwan Railway stations from the official government open data platform, available at https://data.gov.tw/dataset/8792. The dataset includes entry and exit counts and covers the period from 2005 to November 2023 (Figure 1 in Appendix).

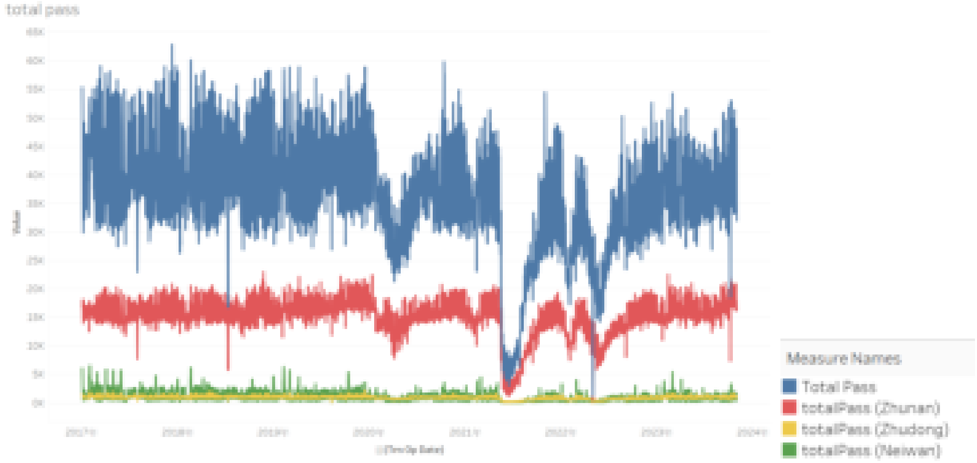
Our forecasting initiative focuses on four key stations in the Hsinchu area: Hsinchu, the primary hub; Zhunan, a station of moderate size, and Zhudong and Neiwan, smaller yet significant stops that hold importance for the local residents. Given the varying sizes and significance of the stations, we can observe distinct forecasting outcomes. The time plots depicting the entry and exit counts for the four stations can be found in Figure 2 in Appendix.

3.2 Data Preparation

We encountered challenges in the original dataset, including unfiltered stations, existing missing values, yearly information scattered across files, and inconsistencies in column names throughout these files. We resolved this by filtering the dataset by station code to focus on the 4 stations, consolidating data from 2005 to 2023, checking existing gaps, and standardizing column names. In addition to these adjustments, we introduced a new column named "**totalPass**," representing the sum of entry and exit counts and serving as a key metric for our daily passenger forecasts.

The modified dataset now consists of four distinct subsets for Hsinchu, Zhunan, Zhudong, and Neiwan stations. Each subset provides detailed information on daily passenger inflow, outflow, and total passenger counts. We believe that recent data holds greater significance and is more reflective of the current situation. As a result, we've opted to focus on data after 2017, roughly two years prior to the pandemic.
This process was implemented to ensure that our dataset aligns with our business goals and analysis objectives (Figure 3 in Appendix).

Figure 4. The time plots of the total daily passengers for the four stations from 2017



4. Methods

4.1 Data Decomposition and Adding External Variable

After the initial preparation of our dataset, we decomposed the 4 time series to determine the patterns respectively. (Figure 5) We found that there are monthly and daily seasonality in all 4 time series, and there was a drastic drop during the COVID-19 Period. Therefore, we decided to add dummy variables to highlight the impact of COVID-19 and determined the start and end time of that period. The start of COVID period was Jan 01, 2020, determined based on the activation of the Central Epidemic Command Center; while the end of this period was December 10, 2022, determined based on the removal of the entry cap and by our observation of the time series plot that shows the recovering trend during the beginning of December 2022.

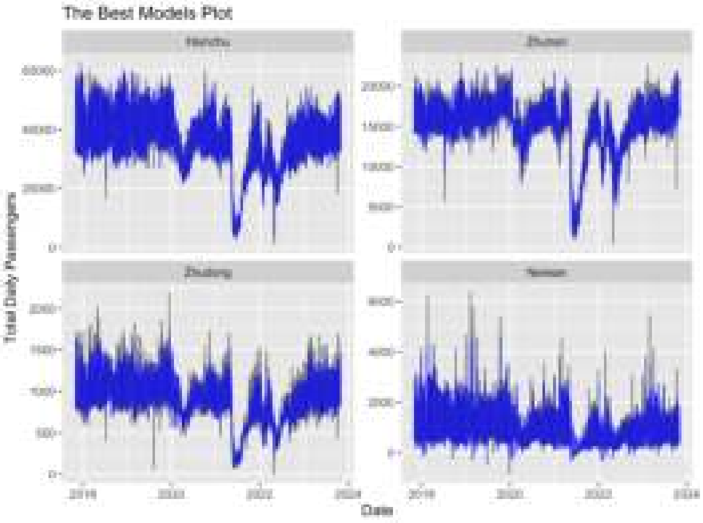
4.2 Data Partitioning and Modeling

As we considered it necessary to train and validate the forecasting model with enough time series, we partitioned our data into 2 periods: Nov 2017 to Oct 2022 for training and Nov 2022 to Oct 2023 for validation. Meanwhile, we would like to compare the final forecast result of Nov 2023 with actual values for evaluation.

Due to the clear trend and seasonality of our data and the manipulation of dummy variables, we applied NAIVE, SNAIVE, Exponential Smoothing (ETS), regression (TSLM) and ARIMA for model training. As for the performance measurement, aligning with stakeholder insights, we checked the boxplot of the forecast errors (Figure 6) and occurrence of over forecast using performance charts (Figure 7).

By comparing the performance charts and measurement, we found that ARIMA had the best performance with the minimum forecast errors and relatively low occurrence of over forecast in all 4 time series (Figure 8). The parameters chosen from automated ARIMA selection are shown in Figure 9 (Appendix). Furthermore, we set the horizon span as 7 days and applied a one-week-ahead roll-forward forecasting to the validation period in order to meet our forecast goal. (Figure 16)

Figure 8. ARIMA performance chart of the 4 time series

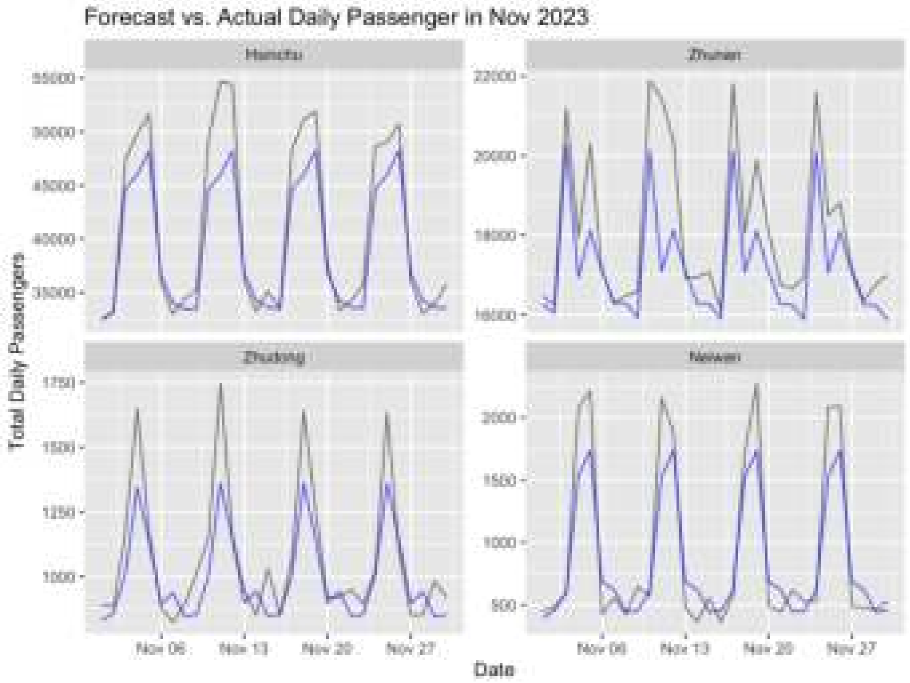


4.3 Performance Evaluation

We further combined the training and validation period data and set the dummy variables (covid) of Nov 2023 to 0 to forecast the total passengers of the 4 stations with ARIMA. In order to evaluate the performance of the forecasting model, we compared the actual numbers of total passengers of each 4 stations with the forecasted values using the performance chart (Figure 10) and calculated the forecast errors, as well as the relative errors (Figure 11 in Appendix). To more effectively assess the model performance, we calculated the average, minimum and maximum relative errors (Figure 12 in Appendix). The result shows that the average relative errors are all less than 6%, with the minimum relative errors under 1% and maximum relative errors approximately around 20%(except for Neiwan Station, where an outlier existed on Nov 14). Besides, our analysis brought to light that days affected by typhoons recorded markedly lower actual passenger numbers, resulting in instances of over-forecasting. It is necessary to remind the stakeholders that the forecast could be less accurate when it comes to typhoon days or perhaps holidays.

In alignment with our forecasting goals, we computed the average weekday passengers for the 4 series in Oct 2023 and the percentage differences between the values of Oct and Nov 2023 (Figure 13 in Appendix). In addition, we visualized the results to offer more clear and understandable information for the stakeholders.

Figure 10. The performance chart of ARIMA in Nov 2023

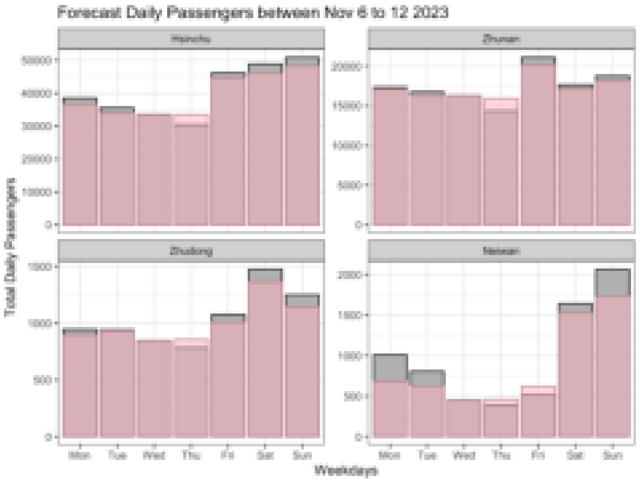


5. Conclusion

5.1 Advantages & Operational Recommendations

Considering the conventional practice of restaurants ordering food or ingredients either one day before or early on the day, we aim to provide forecast results by 2 pm each Sunday. This timing allows restaurants to make informed decisions regarding ingredient quantities for the next day, following the peak lunch hours. In addition to providing the precise numerical forecast, we'll be presenting the forecasted passenger numbers in a user-friendly visual format. Think of it like the "Popular Times graph" on Google Maps. In this graph, the gray bar represents the average number of passengers, while the red one represents the forecasted value. This way, our stakeholders can easily gauge whether a particular day is busier or quieter compared to the norm. It adds a visual element that makes the information more intuitive and accessible (See Figure 14).

Figure 14. The comparison chart of Oct and Nov 2023 by weekdays(Nov 6 ~ Nov 12, 2023)



Given some negative forecast results in our original performance charts, we plan to adjust the initial forecast values. Specifically, we will replace any negative values with zero to prevent confusion among the restaurants.

5.2 Limitations & Future Improvements

Given that the TRA updates its data on a monthly basis, our forecasting system will operate in real-time, incorporating the latest information every month.

Looking forward, we hope to establish a collaborative partnership with the TRA to receive data on a weekly basis. This would allow us to update our forecasts weekly and implement one-week-ahead roll-forward forecasting (Figure 15), and to present the bar chart data in a week-by-week format. Furthermore, we will engage with stakeholders and gather their daily sales data. This collaborative effort will enable us to analyze the correlation between daily passenger numbers and sales data. By incorporating sales data as an external variable, we aim to enhance the precision and customization of our forecasts.

In our broader vision, we aspire to extend the reach of our forecasts beyond the four stations in the Hsinchu area. We aim to make our forecast applicable to all train stations throughout Taiwan, aiding small and medium-sized restaurants in enhancing their food preparation and overall operations. These future implementations will play a crucial role in significantly improving operational efficiency for restaurants surrounding train stations nationwide.

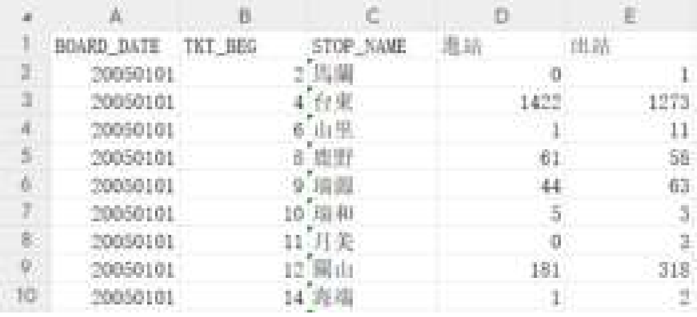
Appendix

1. R code and Data Files

https://github.com/shu9911/2023-BAFT.git

2. Figures

Figure 1. The first 10 rows of the raw data



◢	A	B	C	D	E
1	BOARD_DATE	TKT_BEG	STOP_NAME	進站	出站
2	20050101	2	馬蘭	0	1
3	20050101	4	台東	1422	1273
4	20050101	6	山里	1	11
5	20050101	8	鹿野	61	58
6	20050101	9	瑞源	44	63
7	20050101	10	瑞和	5	3
8	20050101	11	月美	0	2
9	20050101	12	關山	181	318
10	20050101	14	海端	1	2

Figure 2. The time plots of the entry and exit counts for the four stations

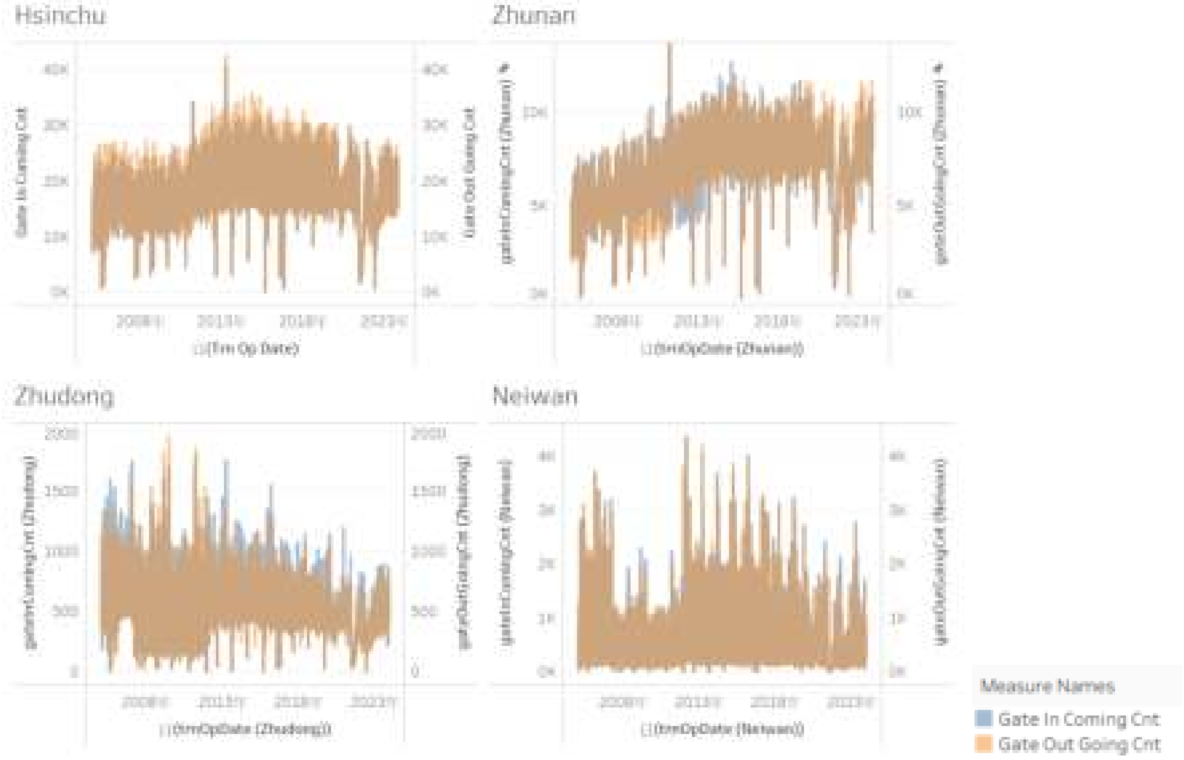
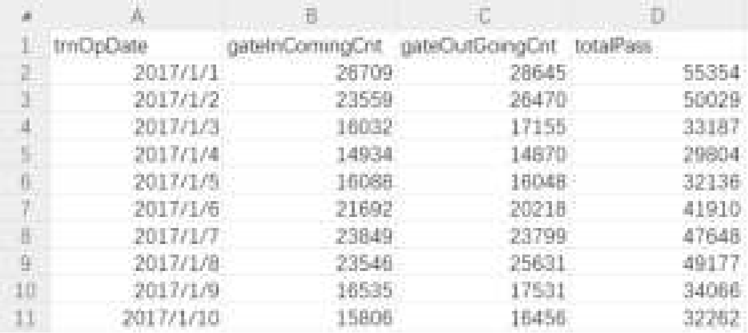


Figure 3. The first 10 rows of the data of Hsinchu station



◢	A	B	C	D
1	trnOpDate	gateInComingCnt	gateOutGoingCnt	totalPass
2	2017/1/1	26709	28645	55354
3	2017/1/2	23559	26470	50029
4	2017/1/3	16032	17155	33187
5	2017/1/4	14934	14870	29804
6	2017/1/5	16088	16048	32136
7	2017/1/6	21692	20218	41910
8	2017/1/7	23849	23799	47648
9	2017/1/8	23546	25631	49177
10	2017/1/9	16535	17531	34066
11	2017/1/10	15806	16456	32262

Figure 5. Decomposition of the 4 time series

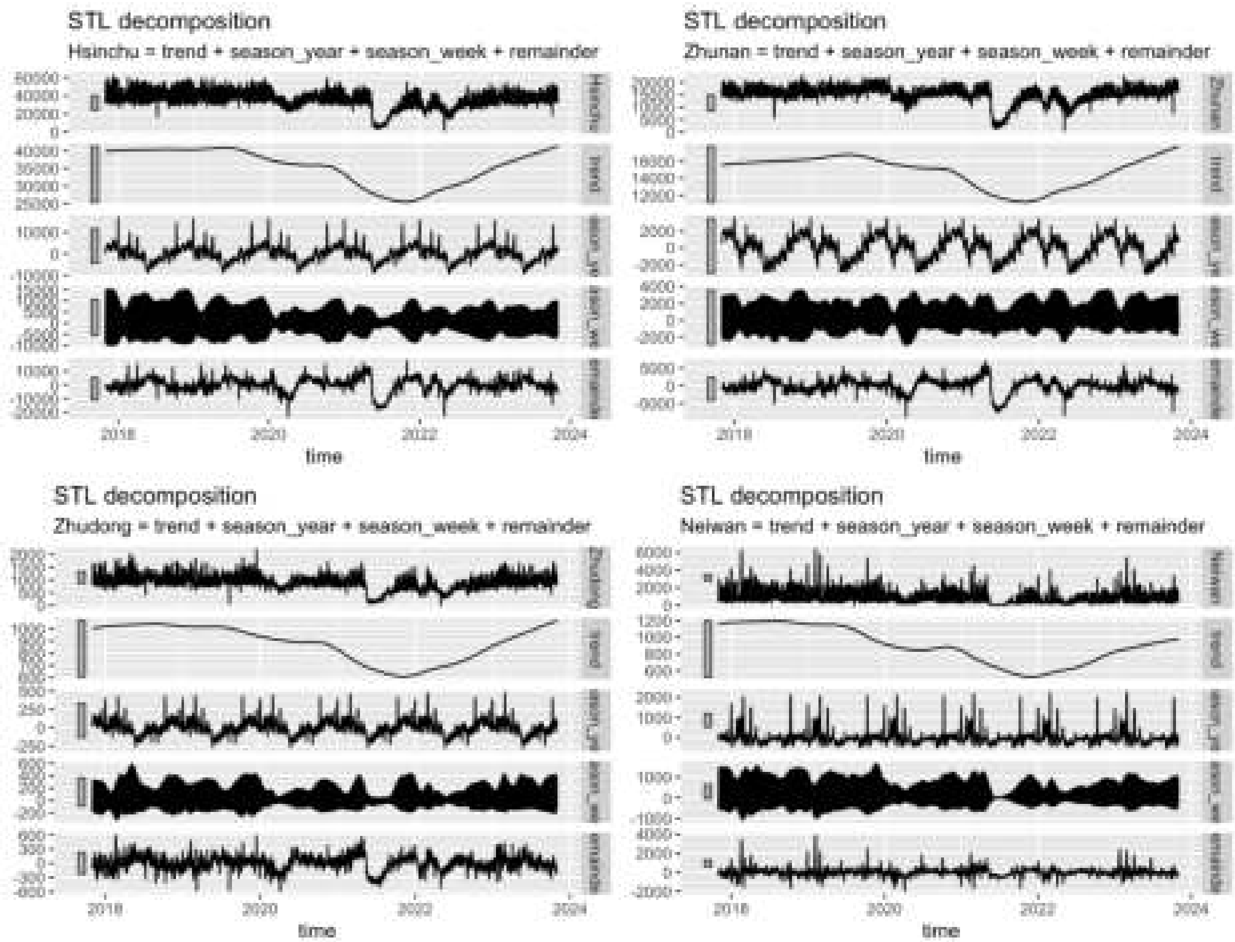


Figure 6. Boxplot of forecast error of the 4 time series

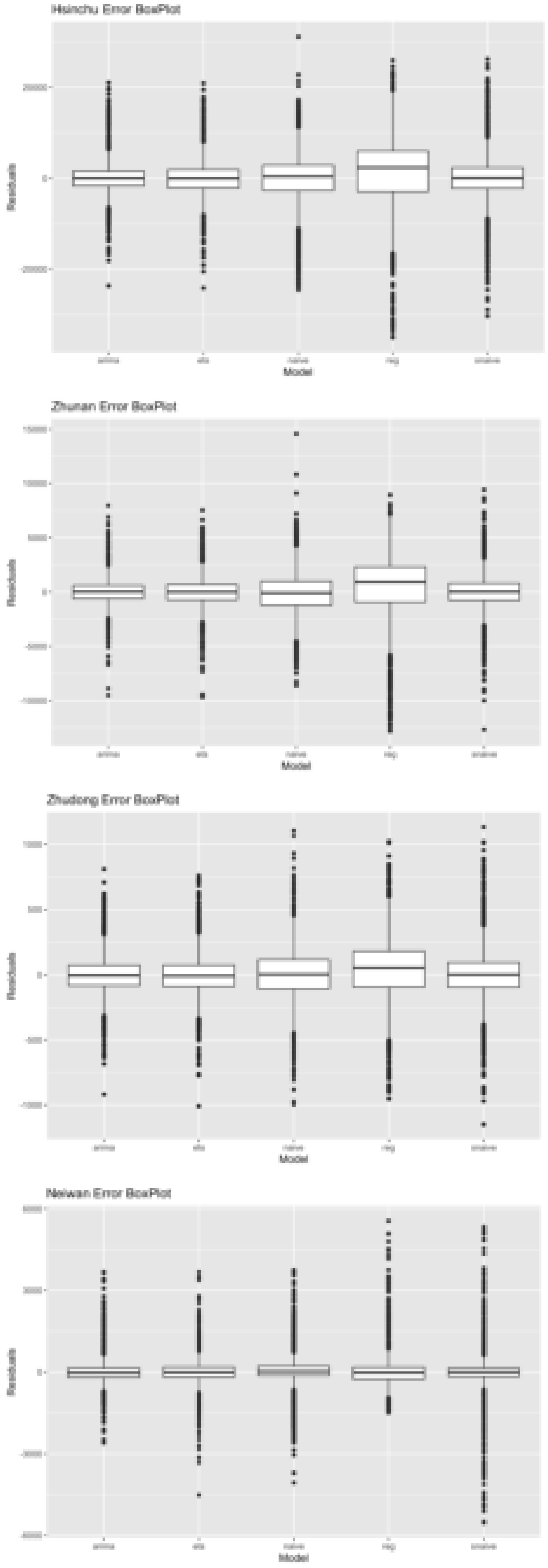
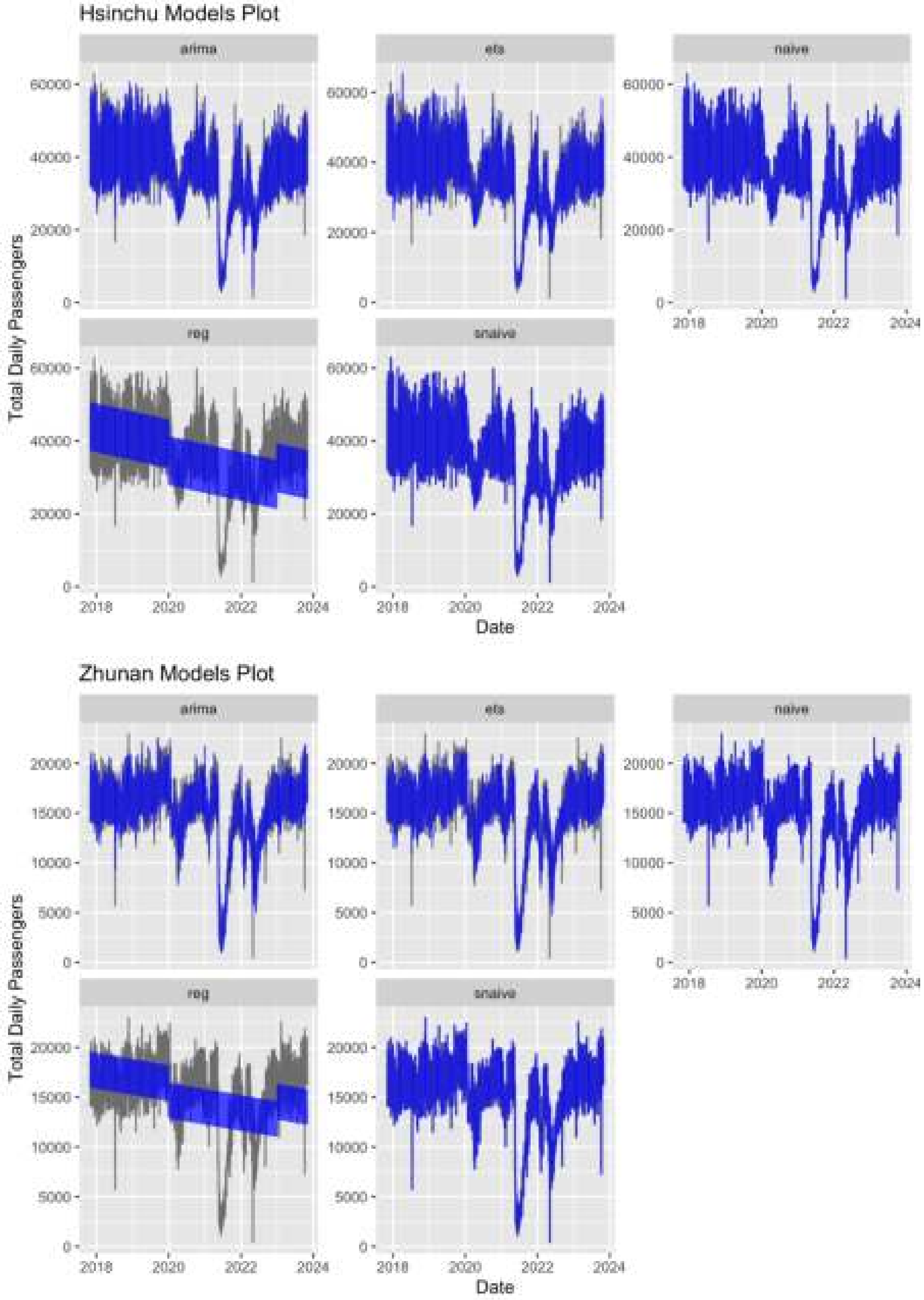
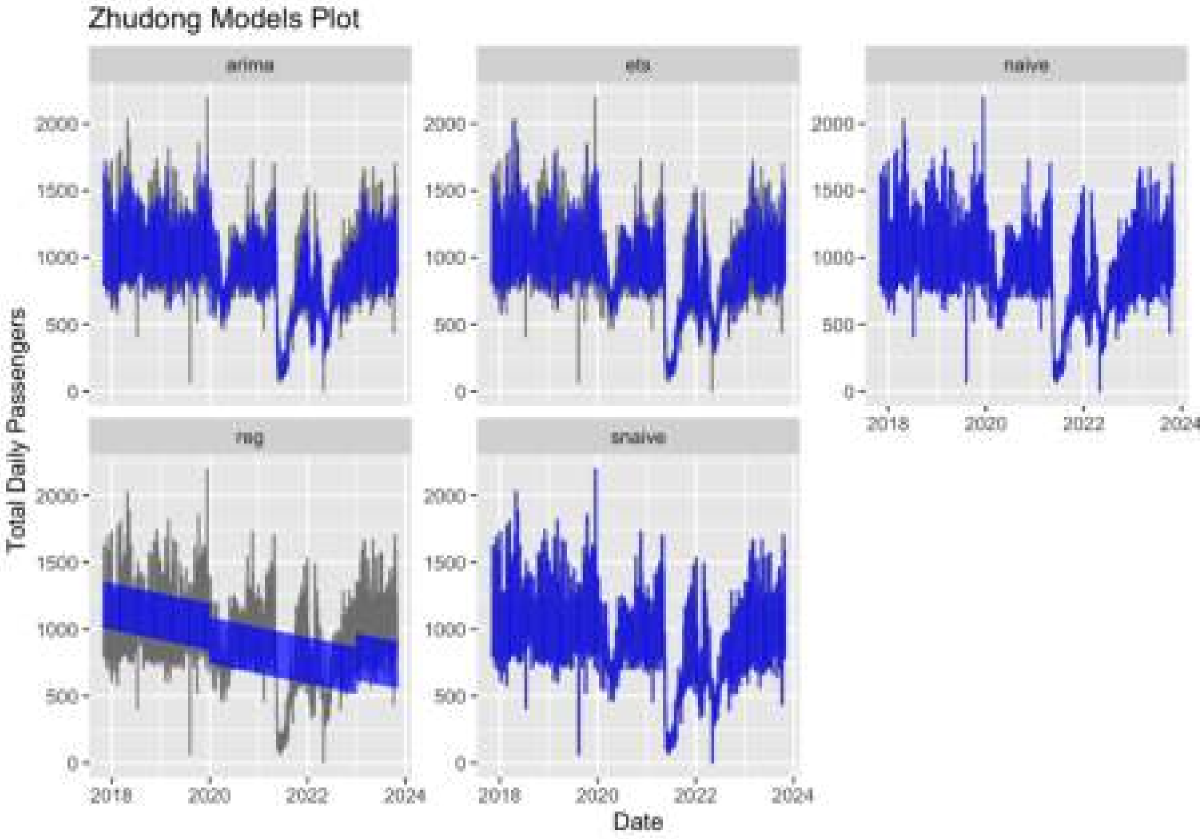


Figure 8. The performance charts of the 4 time series




Zhudong Models Plot
arima
ets
naive
reg
snaive
Total Daily Passengers
Date
0
500
1000
1500
2000
2018
2020
2022
2024

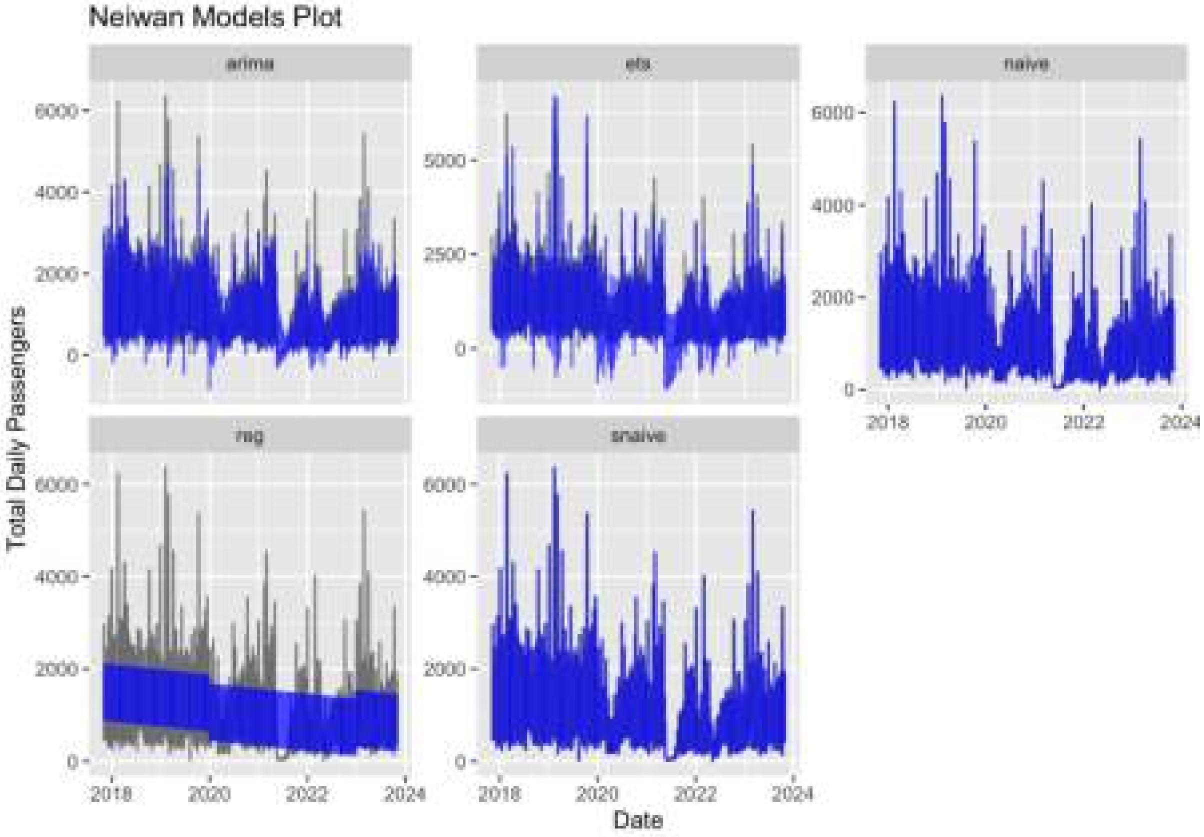
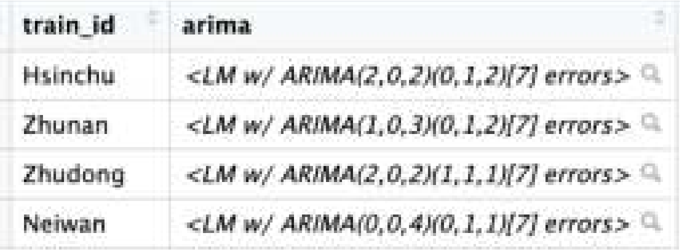
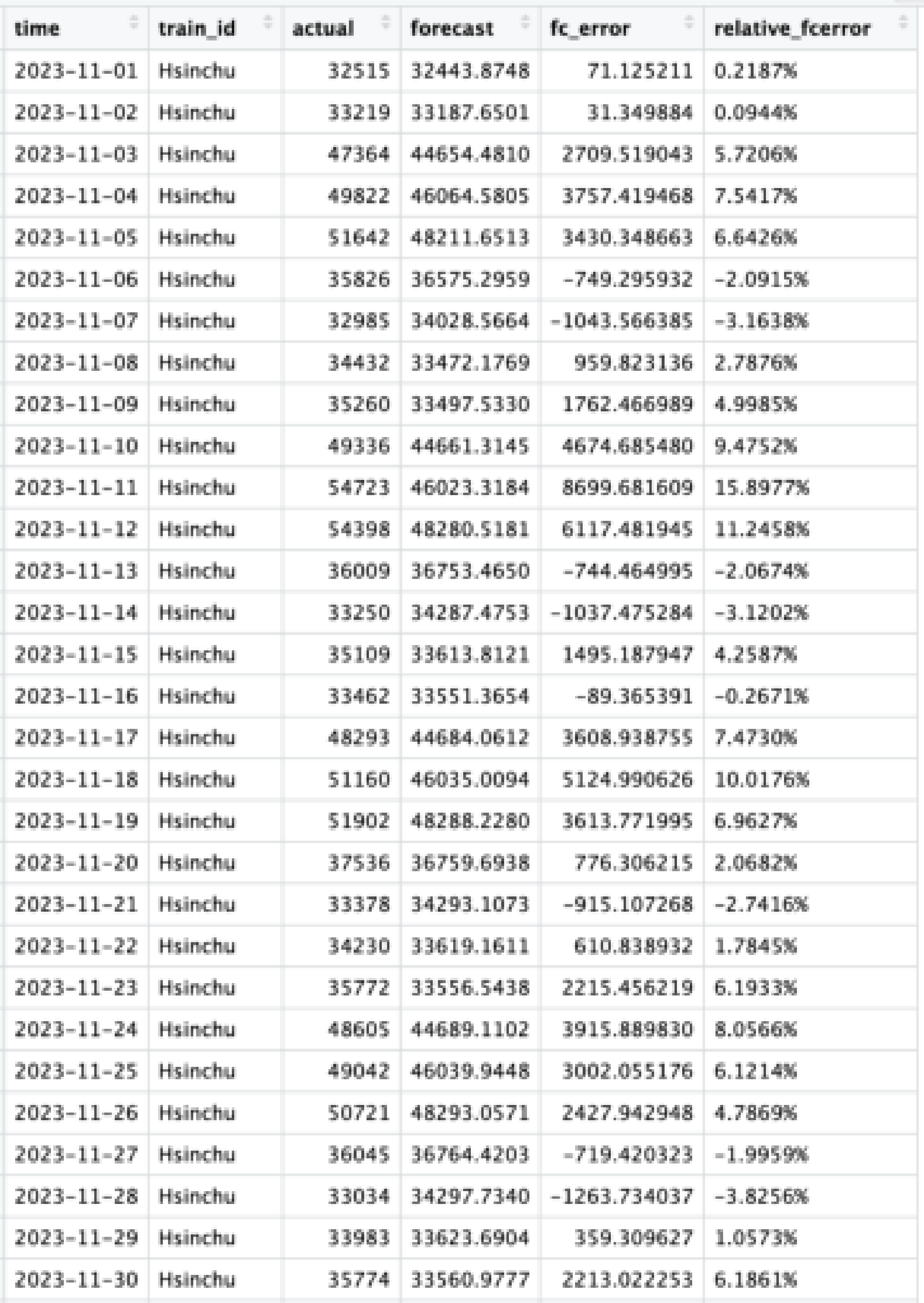

Neiwan Models Plot
arima
ets
naive
reg
snaive
Total Daily Passengers
Date
0
2000
4000
6000
2500
5000
2018
2020
2022
2024

Figure 9. The parameters of automated ARIMA selection



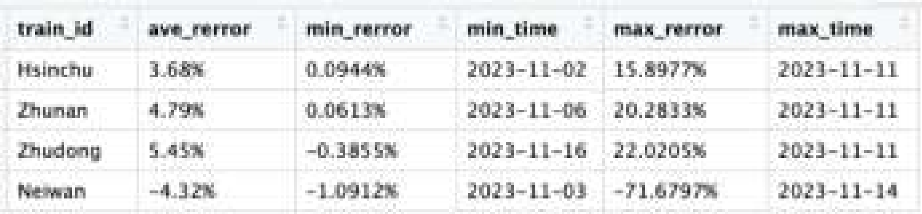
train_id	arima
Hsinchu	<LM w/ ARIMA(2,0,2)(0,1,2)[7] errors>
Zhunan	<LM w/ ARIMA(1,0,3)(0,1,2)[7] errors>
Zhudong	<LM w/ ARIMA(2,0,2)(1,1,1)[7] errors>
Neiwan	<LM w/ ARIMA(0,0,4)(0,1,1)[7] errors>

Figure 11. The forecast errors and relative errors of Hsinchu Station



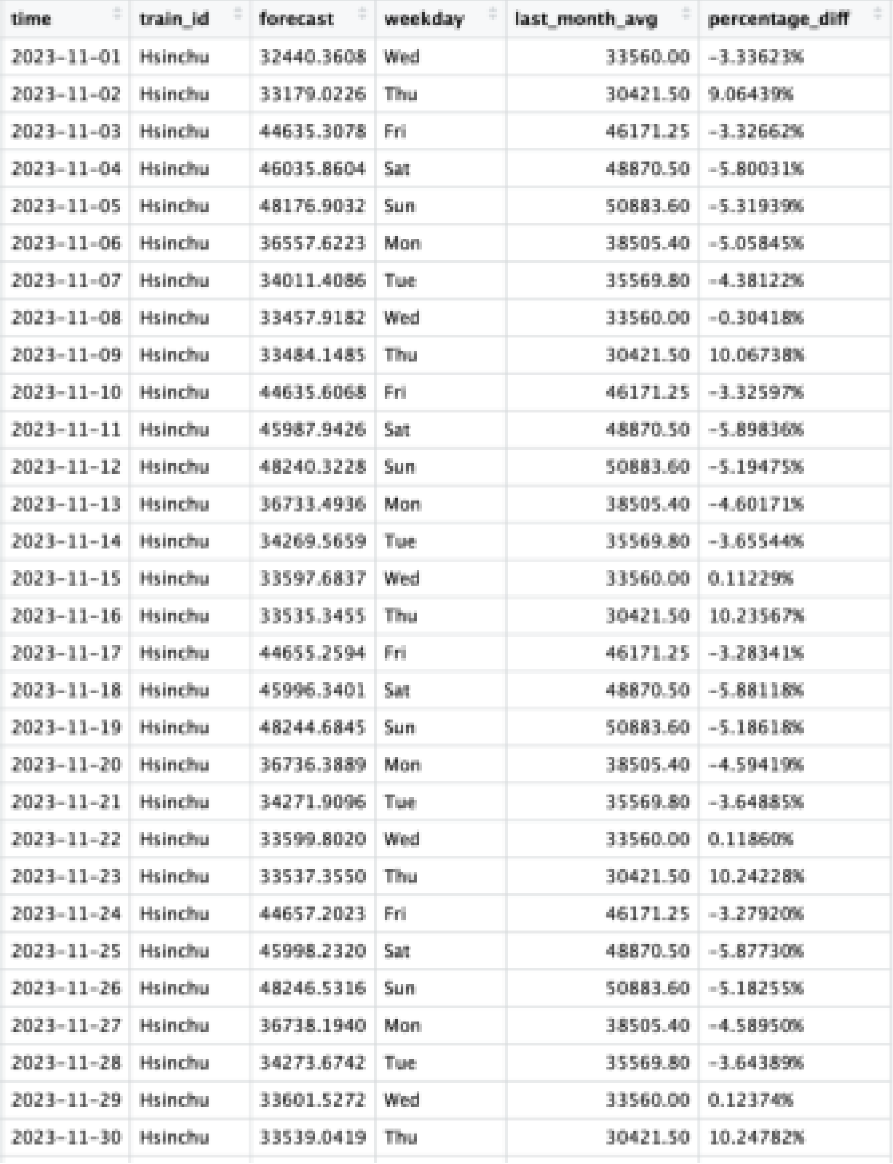
time	train_id	actual	forecast	fc_error	relative_fcerror
2023-11-01	Hsinchu	32515	32443.8748	71.125211	0.2187%
2023-11-02	Hsinchu	33219	33187.6501	31.349884	0.0944%
2023-11-03	Hsinchu	47364	44654.4810	2709.519043	5.7206%
2023-11-04	Hsinchu	49822	46064.5805	3757.419468	7.5417%
2023-11-05	Hsinchu	51642	48211.6513	3430.348663	6.6426%
2023-11-06	Hsinchu	35826	36575.2959	-749.295932	-2.0915%
2023-11-07	Hsinchu	32985	34028.5664	-1043.566385	-3.1638%
2023-11-08	Hsinchu	34432	33472.1769	959.823136	2.7876%
2023-11-09	Hsinchu	35260	33497.5330	1762.466989	4.9985%
2023-11-10	Hsinchu	49336	44661.3145	4674.685480	9.4752%
2023-11-11	Hsinchu	54723	46023.3184	8699.681609	15.8977%
2023-11-12	Hsinchu	54398	48280.5181	6117.481945	11.2458%
2023-11-13	Hsinchu	36009	36753.4650	-744.464995	-2.0674%
2023-11-14	Hsinchu	33250	34287.4753	-1037.475284	-3.1202%
2023-11-15	Hsinchu	35109	33613.8121	1495.187947	4.2587%
2023-11-16	Hsinchu	33462	33551.3654	-89.365391	-0.2671%
2023-11-17	Hsinchu	48293	44684.0612	3608.938755	7.4730%
2023-11-18	Hsinchu	51160	46035.0094	5124.990626	10.0176%
2023-11-19	Hsinchu	51902	48288.2280	3613.771995	6.9627%
2023-11-20	Hsinchu	37536	36759.6938	776.306215	2.0682%
2023-11-21	Hsinchu	33378	34293.1073	-915.107268	-2.7416%
2023-11-22	Hsinchu	34230	33619.1611	610.838932	1.7845%
2023-11-23	Hsinchu	35772	33556.5438	2215.456219	6.1933%
2023-11-24	Hsinchu	48605	44689.1102	3915.889830	8.0566%
2023-11-25	Hsinchu	49042	46039.9448	3002.055176	6.1214%
2023-11-26	Hsinchu	50721	48293.0571	2427.942948	4.7869%
2023-11-27	Hsinchu	36045	36764.4203	-719.420323	-1.9959%
2023-11-28	Hsinchu	33034	34297.7340	-1263.734037	-3.8256%
2023-11-29	Hsinchu	33983	33623.6904	359.309627	1.0573%
2023-11-30	Hsinchu	35774	33560.9777	2213.022253	6.1861%

Figure 12. The average, minimum and maximum relative errors of each station



train_id	ave_rerror	min_rerror	min_time	max_rerror	max_time
Hsinchu	3.68%	0.0944%	2023-11-02	15.8977%	2023-11-11
Zhunan	4.79%	0.0613%	2023-11-06	20.2833%	2023-11-11
Zhudong	5.45%	-0.3855%	2023-11-16	22.0205%	2023-11-11
Neiwan	-4.32%	-1.0912%	2023-11-03	-71.6797%	2023-11-14

Figure 13. The percentage differences of weekdays in Nov 2023(Hsinchu Station)



time	train_id	forecast	weekday	last_month_avg	percentage_diff
2023-11-01	Hsinchu	32440.3608	Wed	33560.00	-3.33623%
2023-11-02	Hsinchu	33179.0226	Thu	30421.50	9.06439%
2023-11-03	Hsinchu	44635.3078	Fri	46171.25	-3.32662%
2023-11-04	Hsinchu	46035.8604	Sat	48870.50	-5.80031%
2023-11-05	Hsinchu	48176.9032	Sun	50883.60	-5.31939%
2023-11-06	Hsinchu	36557.6223	Mon	38505.40	-5.05845%
2023-11-07	Hsinchu	34011.4086	Tue	35569.80	-4.38122%
2023-11-08	Hsinchu	33457.9182	Wed	33560.00	-0.30418%
2023-11-09	Hsinchu	33484.1485	Thu	30421.50	10.06738%
2023-11-10	Hsinchu	44635.6068	Fri	46171.25	-3.32597%
2023-11-11	Hsinchu	45987.9426	Sat	48870.50	-5.89836%
2023-11-12	Hsinchu	48240.3228	Sun	50883.60	-5.19475%
2023-11-13	Hsinchu	36733.4936	Mon	38505.40	-4.60171%
2023-11-14	Hsinchu	34269.5659	Tue	35569.80	-3.65544%
2023-11-15	Hsinchu	33597.6837	Wed	33560.00	0.11229%
2023-11-16	Hsinchu	33535.3455	Thu	30421.50	10.23567%
2023-11-17	Hsinchu	44655.2594	Fri	46171.25	-3.28341%
2023-11-18	Hsinchu	45996.3401	Sat	48870.50	-5.88118%
2023-11-19	Hsinchu	48244.6845	Sun	50883.60	-5.18618%
2023-11-20	Hsinchu	36736.3889	Mon	38505.40	-4.59419%
2023-11-21	Hsinchu	34271.9096	Tue	35569.80	-3.64885%
2023-11-22	Hsinchu	33599.8020	Wed	33560.00	0.11860%
2023-11-23	Hsinchu	33537.3550	Thu	30421.50	10.24228%
2023-11-24	Hsinchu	44657.2023	Fri	46171.25	-3.27920%
2023-11-25	Hsinchu	45998.2320	Sat	48870.50	-5.87730%
2023-11-26	Hsinchu	48246.5316	Sun	50883.60	-5.18255%
2023-11-27	Hsinchu	36738.1940	Mon	38505.40	-4.58950%
2023-11-28	Hsinchu	34273.6742	Tue	35569.80	-3.64389%
2023-11-29	Hsinchu	33601.5272	Wed	33560.00	0.12374%
2023-11-30	Hsinchu	33539.0419	Thu	30421.50	10.24782%

Figure 15. Roll-forward one-week-ahead forecasting of Nov 2023, Hsinchu Station

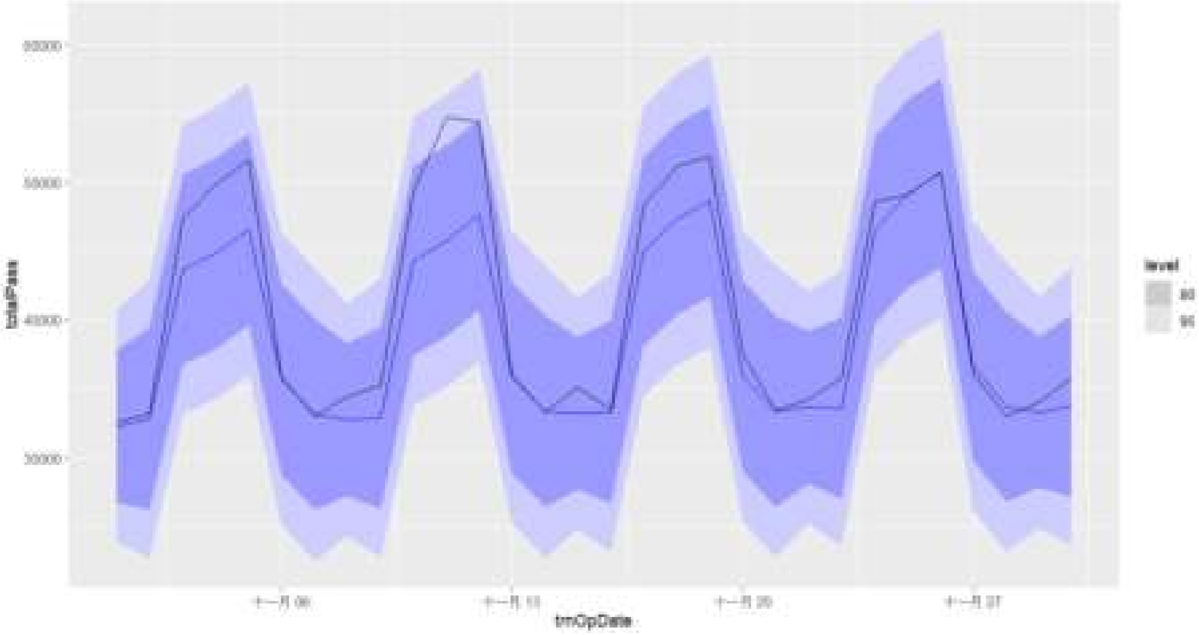
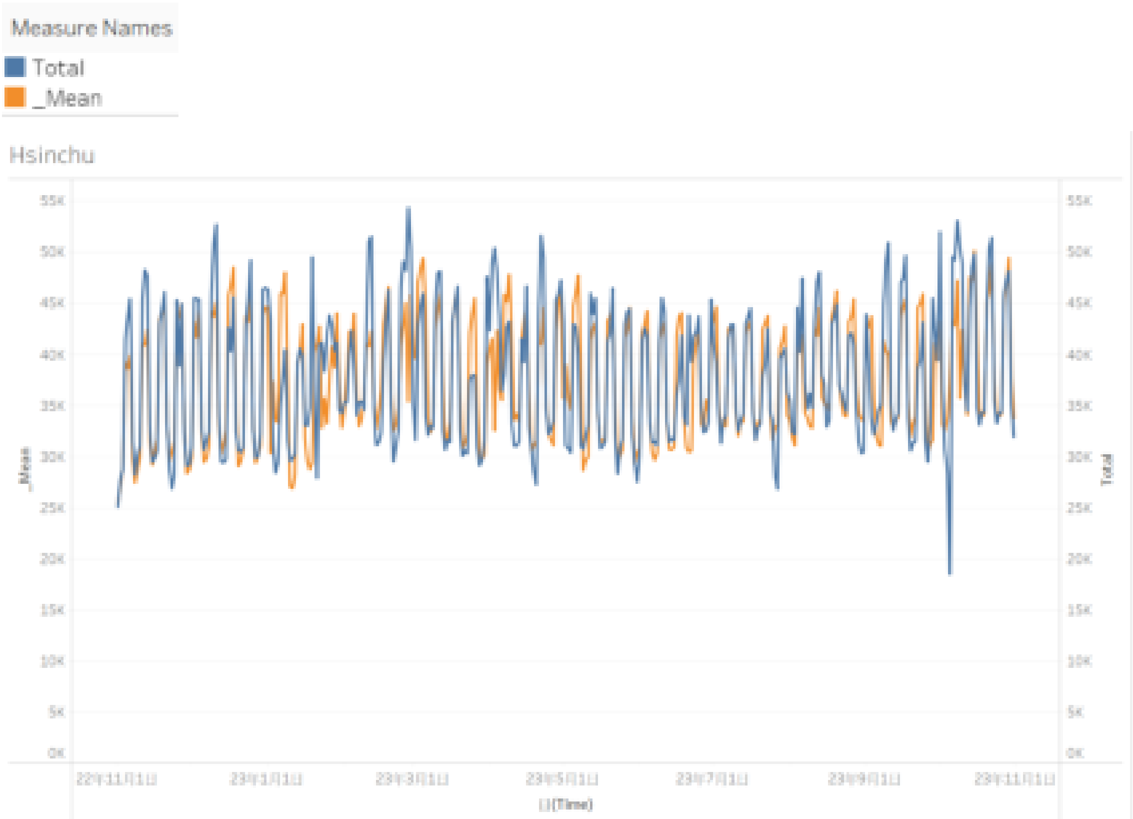
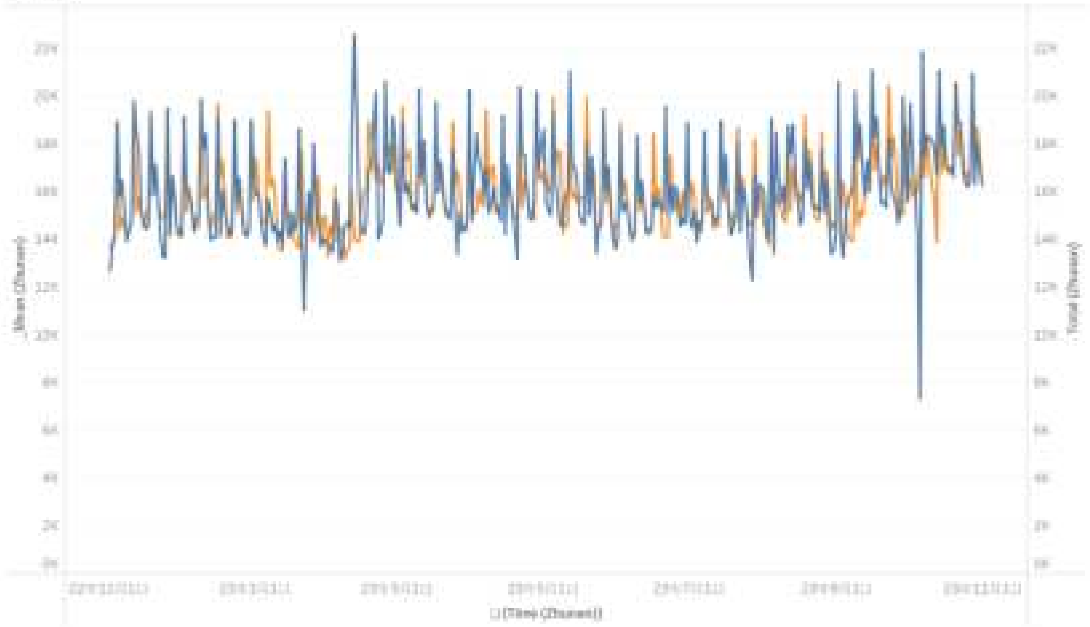


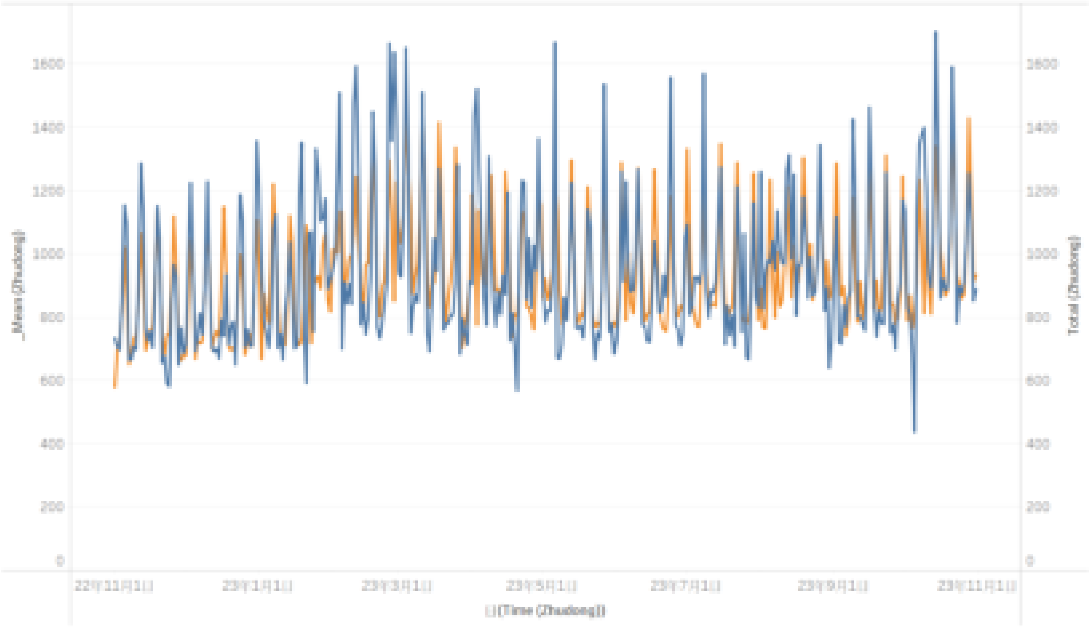
Figure 16. Roll-forward one-week-ahead forecasting of the validation period

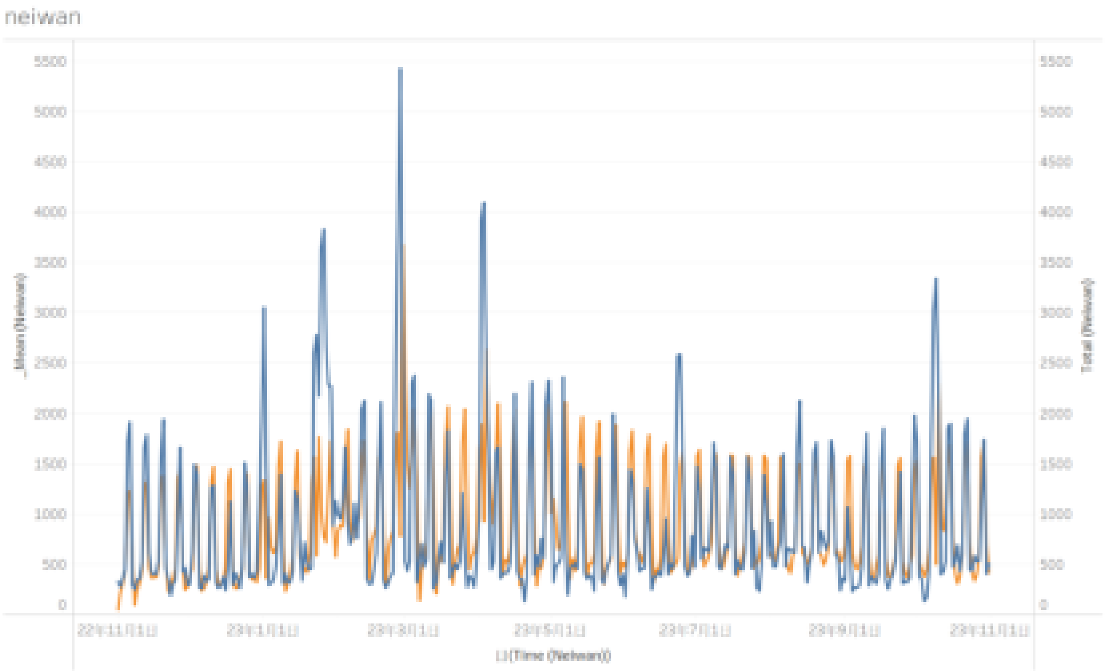


zhunan



zhudong





3. Stakeholder Interviews

Interview objectives:

(1) Explore the potential use of passenger forecasting for our stakeholders
(2) Understand if the volume of passenger forecasting will help stakeholders' business
(3) Reach out to our potential stakeholders

Key stakeholders:

Vendors in markets, new stores, and stores without refrigeration systems nearby the train station

Takeaways:

(1) Store owners rely on experience to decide on the preparation of ingredients.
This is an opportunity to provide passenger forecasting since they lack sales records.
(2) Vendors prefer to prepare less food than prepare too much without being able to sell out and causing food waste.
Our predictions give vendors near train stations a reference for traffic, allowing them to decide how much food to prepare so as not to cause food waste.
(3) Sales flow has risen to the same level compared to before the Covid.
Our forecasting can capture the volume of passengers before the epidemic.