Proceedings of the Eastern Asia Society for Transportation Studies, Vol.12,2019

Exploring Holiday Trip Patterns on Freeways Based on Electronic Toll Collection Data

Shiuan-Yu LOU [a], Wen-Yu HSU [b], Yu-Ting HSU [c]

[a,b] *Transportation Division, Department of Civil Engineering, National Taiwan University, Room 305, Civil Engineering Building, No. 1, Roosevelt Road, Sec. 4*
[a] *E-mail: d05521011@ntu.edu.tw*
[b] *E-mail: r07521517@ntu.edu.tw*
[c] *E-mail: yutinghsu@ntu.edu.tw*

Abstract Based on the trip data derived from the Electronic Toll Collection system for the freeway system in Taiwan, this study seeks to explore the flow pattern and related characteristics of trips made on national holidays, particularly long holidays. The daily travel data are converted into an origin-destination table, where the origins and destinations include the 15 major counties and cities in Taiwan. K-means clustering is used to group each daily O-D table into 5 groups to distinguish different types of daily trip patterns. This study further employs logistic regression to analyze the characteristics of each group in terms of how traffic management measures and holiday-related characteristics may affect trip distribution across the freeway. The results suggest that the trip pattern on normal workdays is significantly different from it on holidays; the latter one generally involves more long-distance trips.

Keywords: Electronic Toll Collection, Origin-Destination Table, National Holidays, K-Means, Logistic Regression

1. INTRODUCTION

The national holidays in Taiwan include Chinese New Year, Tomb Sweeping Festival (Quin Ming Festival), Dragon Boat Festival, Moon Festival, National Day, and Labor Day, etc., on which a large number of trips are generated, especially on long consecutive holidays (or long weekends). Trips on these holidays have apparent impact on island-wide transportation systems and particularly impose great challenge on freeway management. To alleviate the potential holiday congestion on freeways, the government has proposed several management strategies, such as ramp metering, high-occupancy vehicle lanes, discount of toll in off-peak periods, seeking to divert trips to alternative transportation modes (or different departure time slots) and to utilize freeway capacity more efficiently. For the described purposes, better understanding or even prediction of trip patterns across the freeway system may enable freeway operators to determine/select appropriate management strategies and thereby attain enhanced effectiveness.

In practice, the prediction of trip patterns during long weekends or holidays is made based on the data of the same genre of holidays in the past. For example, in 2018, there was a three-day holiday for Moon Festival, and the relevant freeway traffic conditions was conjectured from the data of a previous three-day holiday. However, such conjecture is made conceptually and roughly by experience without definite scientific and theoretical basis.

Hence, this study seeks to analyze trip pattern and relevant characteristics across consecutive holidays based on the data of the Electronic Toll Collection (ETC) system for the freeway system in Taiwan, which contain the Origin-Destination (O-D) record of each trip. Considering that people may try to extend holidays by planning trips that start earlier or finish later than the actual holiday span, the data of trips occurring before the start of the holiday and after the end of it are included in this study as well. These data are explored to cluster daily trip patterns during holidays by K-means, and thereby the daily trip patterns that are similar and share common features are identified. Based on the groups of daily trip patterns derived by K-means, a multinomial logit model is further developed to investigate the O-D distribution across the freeway in each group and how it may be associated with traffic management measures and holiday-related characteristics. It is expected that the developed model can help the freeway operators better infer how trips may be generated during long consecutive holidays, which can be the reference for relevant management decision-making.

2. DATA DESCRIPTION

2.1 ETC Traffic Data

The ETC system has been implemented for Taiwan freeway system since 2006 for the purpose of paperless payment and reducing interruption to freeway traffic flows. In 2014, the system was further upgraded by setting up the charging zone (ETC gantry) between every pair of adjacent ramps/interchanges, and thereby the trajectories of vehicles can be traced, which is the key to enable the “charge by mileage” pricing scheme. On the other, the data derived from the ETC system also allow the freeway operator to better monitor sectional traffic flow conditions across the freeway and to more flexibly deploy traffic control measures as well. In recent years, Freeway Bureau also develops an Open Data platform to release these to the society and academia for research use or other value-added service. “M08A” the data used in the study contain the record of each trip, with the details of origin, destination, type of vehicle, which can be also aggregate to O-D demand or sectional vehicular flow volumes. There are about 500 thousand piece of “M08A”data in one day. The samples in this study is a O-D table which is an aggregates of “M08A”data in each national holiday and the days before and after holidays.

2.2 Descriptive Statistics

This study retrieves the trip data of national holidays during the period from 2014 to October 2018 and the trip data one day before and after these holidays. Some holidays are coincident with the impact of typhoons; the associated freeway traffic characteristics are significantly different from those on the other national holidays. Hence, the data of the holidays coincident with typhoons are removed from this study. Consequently, there are total 185 samples of daily trip data (one sample per day), which are summarized in Table 2.1.

Table 2.1 Description of Samples

Characteristics of Holidays		Sample amount
Type of Holiday	Moon Festival	11
	New Year’s Day	21
	Peace Memorial Day	22
	Chinese New Year	43
	National Day	22
	Tomb Sweeping Festival	30
	Labor Day	14
	Dragon Boat Festival	22
Which day in a holiday does the sample belong to?	The Day before the Holiday	30
	The First Day of the Holiday	36
	The Middle Day of the Holiday	58
	The Last Day of the Holiday	31
	The Day after the Holiday	30
Which day in a week does the sample belong to?	Monday	31
	Tuesday	22
	Wednesday	18
	Thursday	19
	Friday	31
	Saturday	32
	Sunday	32
Length of holiday	1 day	5
	2 days	2
	3 days	80
	4 days	48
	5 days	7
	6 days	32

	9 days	11
Measure of High Occupancy Vehicle (HOV)	Southbound	5
	Northbound	12

Among the national holidays in Taiwan, Chines New Year introduces most of the samples compared with others. The distribution of the samples shows the centralization on Friday, Saturday, Sunday, and Monday. In terms of the length of holidays, three-day holiday accounts for the majority of the national holidays in the data, while there are still few one-day holidays (if the holiday happens to fall on Wednesday). In addition, during the considered data period (from 2014 to October 2018), High Occupancy Vehicle (HOV) lanes have been implemented for seventeen times on National Freeway No.1 and No.3, out of which there are five times for southbound and twelve times for northbound, particularly for handling holiday traffic.

3. METHODOLOGY

To classify travelers, Chiou et al. (2014) divide travelers into three groups on the basis of travelers' residence types and develop a mixed logit model to analyze travelers' mode choice behavior. Tong and Feng (2018) employ K-means, random forest, and latent class modeling for clustering travelers. According to the clustering results, they further develop a logit model for each identified traveler group and investigate the behavior of switching across different groups. We can find out that these researches cluster travelers and then develop a logit model to explore different transportation mode choices, no one used it to analyze the trip patterns. In addition, due to obtaining O-D data is not easy, lots of researchers establish algorithm to estimate. Lin et al.(2018) make use of the one-section smart card fare tallying record, combined with the smart card records and regional zoning information, in order to establish a three-stage trip algorithm to estimate the destinations. David Martin (2018) uses the analysis of travel to work flows by combining separate geodemographic classifications of origins and destinations. However, this study uses actual O-D data to analyze. We can find out that these researchs cluster travelers and then develop a logit model to explore different transportation mode choices; no one used it to analyze the trip patterns. I-Shin Yeh (2017) collect and refine the records which include go and return transaction in the EasyCard of Taipei MRT to organize passenger's behavior, and give the following four practical cases to verify Taipei MRT policy:(i)the MRT crowded analysis.(ii)Wenhu line shift scheduling analysis.(iii)New Year's Eve distribution analysis. (iv)The crowd analysis of MRT transfer stations. Although the study uses actual O-D data, it only analyzes the how many passengers entering and leaving the stations on weekdays, weekends and New Year's Eve. It does not elaborate on the comparison of travel characteristics between weekdays, weekends and New Year's Eve, nor does it explore the reasons for the travel characteristics. In this study, clustering of trip patterns across the freeway is directly implemented to the entire O-D table. K-means is adopted to for clustering the holidays with similar entire O-D tables (trip patterns) into the same group. The number of groups are determined based on the statistical result of clustering. Then, logistic regression is further constructed to analyze the factors associated with each (trip pattern) group and compare the behavioral difference across different groups.

3.1 K-Means

K-means is the most widely used algorithm in non-hierarchical methods for data clustering. Before implementing the K-means algorithm, the number of clusters needs to be specified in advance. If there is the lack of *a priori* knowledge of data pattern and relevant structure, an appropriate number of clusters can be determined by testing different pre-specified numbers of clusters with the algorithm. The steps of the K-means algorithm are briefly summarized as follows:

1) Specifying the number of clusters and defining the centroid for each cluster.
2) Calculating the distances between each data point and the defined cluster centroids, and then associating it to the nearest centroid.
3) Re-calculating new centroids of the clusters resulting from the newly added data point (in the previous step).
4) Regrouping every data point according to new centroids.
5) Repeating Steps 2-4 until all data points are assigned and all the clusters do not change anymore.

Although the centroids of clusters will be updated through the described iterative process, selecting different initial centroids (in Step 1) can more or less lead to distinct clustering results. Hence, several runs of the algorithm with randomly selected initial centroids may need to be conducted to avoid biased or sub-optimal clustering results.

3.2 Logistic Regression

Categorical variables and the effects of associated factors are often analyzed by logistic regression in terms of conditional probability $P(y_i|x_i)$ denoting the probability of the occurrence of Event *i*. based on the relevant independent variable x_i. $y_i = 1$ indicates the occurrence of Event *i* (and it is observed); otherwise, $y_i = 0$ means that no such an event is observed. Assuming there is a linear relationship between y_i with x_i, which can be represented as:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{3.1}$$

then, $P(y_i|x_i)$ can be further written as Equation (3.2):

$$P(y_i = 1|x_i) = P[(\alpha + \beta x_i + \varepsilon_i > 0)] = P[\varepsilon_i > (-\alpha - \beta x_i)] \tag{3.2}$$

where ε_i is assumed to follow a logistic distribution, and the logistic distribution is symmetric. Hence, we can derive Equation (3.3) from Equation (3.2):

$$P(y_i = 1|x_i) = P[\varepsilon_i \le (\alpha + \beta x_i)] = F(\alpha + \beta x_i) = \frac{1}{1+e^{-\varepsilon_i}} \tag{3.3}$$

x_i is the independent variable (regressor); α and β are the intercept term and the coefficient associated with x_i. By expressing the formula in one-way regression and letting the probability that Event i occurs $p_i = P(y_i = 1|x_i)$, the logistic regression can then be derived as a nonlinear probability function:

$$p_i = \frac{1}{1+e^{-(\alpha+\beta x_i)}} = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \tag{3.4}$$

Equation (3.4) can be also converted into a linear function. By contrast, the probability that Event i does not occur can be calculated as Equation (3.5):

$$1 - p_i = 1 - \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right) = \frac{1}{1+e^{\alpha+\beta x_i}} \tag{3.5}$$

We then further define *odds* as the ratio of (the probability that the event will occur) / (the probability that the event will not occur):

$$odds = \frac{p_i}{1-p_i} = e^{\alpha+\beta x_i} \tag{3.6}$$

Consequently, the natural logarithm of *odds* results in a linear function of x_i, which is also known as the logit of y_i

$$logit = ln(odds) = \alpha + \beta x_i \tag{3.7}$$

By developing logistic regression as Equation (3.7), we obtain the insight of how variable x_i affects the occurrence of Event i by β. If there are more than two (types of) events, one may employ multi-category logistic regression for similar analysis. Assuming that there are three events *a*, *b*, and *c*, the associated probabilities of occurrence are p_a, p_b and p_c, respectively, we can select Event *c* as the base case. Then, based on Equations (3.8) and (3.9), we can evaluate how the relevant factors (*x*) influence the occurrence probabilities of Event *a* and Event *b* against Event *c*.

$$ln\left(\frac{p_a}{p_c}\right) = \alpha_a + \beta_a x \tag{3.8}$$

$$ln\left(\frac{p_b}{p_c}\right) = \alpha_b + \beta_b x \tag{3.9}$$

4. RESULT ANALYSIS

4.1 Clustering of Daily Trip Patterns

This study test Different number of cluster by using the K-means algorithm, if number of clusters less than 5, it can't divide Chinese New Year into one cluster. Whereas number of clusters more than 5, the day before the holiday and the day after the holiday will distributed in many clusters, the characteristics are not significant. Finally, the 185 samples in this study are divided into 5 clusters. The basic holiday characteristics of each cluster are presented in Table 4.1. The features of the trip patterns in Cluster 2 and Cluster 3 are comparatively discernible. The major feature of Cluster 2 is that the associated trip patterns result from the traffic during Chinese New Year (which is the longest holiday in Taiwan), except the last day of it. Additionally, the policy of HOV lane is generally implemented on the first day of Chinese New Year. Cluster 3 basically contains the samples of the days before and after holidays.

Table 4.1 Cluster description

Characteristics \ Cluster		1	2	3	4	5
Number of sample		50	10	37	36	52
Proportion		27%	5.4%	20%	19.5%	28.1%
Which day is the Holiday on?	The Day before the Holiday	8	0	22	1	0
	The First Day of the Holiday	21	2	0	7	6
	The Middle Day of the Holiday	4	8	0	12	34
	The Last Day of the Holiday	3	0	0	16	12
	The Day after the Holiday	14	0	15	0	0
Name of Holiday	Moon Festival	3	0	3	2	3
	New Year's Day	4	0	7	2	8
	Peace Memorial Day	10	0	4	5	3
	Chinese New Year	6	10	4	7	16
	National Day	6	0	5	5	6
	Tomb Sweeping Festival	5	0	7	7	11
	Labor Day	10	0	2	2	0
	Dragon Boat Festival	6	0	5	6	5
What Day is the Holiday?	Monday	10	1	4	4	12
	Tuesday	6	0	7	2	7
	Wednesday	3	1	8	1	5
	Thursday	6	2	5	2	4
	Friday	5	2	13	5	6
	Saturday	17	2	0	8	5
	Sunday	3	2	0	14	13
Length of Holiday	1	2	0	0	1	2
	2	2	0	0	0	0
	3	33	0	17	20	10
	4	7	0	14	8	19
	5	0	0	2	0	5
	6	5	10	2	5	10
	9	1	0	2	2	6
High Occupancy Vehicle (HOV)	Southbound	2	2	0	0	1
	Northbound	0	3	0	1	8

4.2 Trip Pattern Investigation

To highlight the comparison of the clustering result, we further merge the daily O-D table from the county level (15×15 O-D pairs) to a regional level, which aggregates the fifteen counties into three regions along the freeway service corridor: North, Middle and South. Accordingly, a daily trip pattern is represented in a 3×3 O-D Table as illustrated in Figure 4.1(the relationship and position of O-D table is in Figure 4.2), where:

K: index for cluster, K=1,2,3,4,5

O,D：origin and destination regions, 1 for North, 2 for Central, and 3 for South

T_{KOD}: number of trips for a specific O-D pair in Cluster *k*

Table 4.2 summarizes the 3×3 O-D table of each cluster, where the cells in the table take the average number of daily trips across the O-D tables in the corresponding cluster.

		1					2					3				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	212911.9	4332.288	8574.673	1026.019	38122.17	2633.923	15924.27	11442.38	6445.135	1348.981	13172.21	16259.15	12899.5	8537.115	38636.25
	2	6197.75	137149.5	911.8846	6652.731	458.25	209.8269	3176.692	42408.44	726.6923	37352.94	829.7885	98122.17	8744.788	735.0385	962.5577
	3	8365.481	789.1154	150477.7	168.6346	3550.077	10762.6	1193.827	2102.981	61464.04	211.4231	6197.481	3434.385	2004.481	13872.17	2664.442
	4	877.8846	8298.423	128.1923	25142.29	72.01923	30.84615	307.8077	2084.519	114.0962	1094.788	90.5	10699.12	927.4615	75.65385	125.2692
	5	32272.94	428.1731	3649.846	93.57692	47728.4	1280.942	1109.115	1400.904	2594.731	128.1731	7764.808	2358.173	1258.173	4339.404	1858.731
2	6	2860.269	171.2115	8710.942	27.78846	1153.462	52401	341.6923	728.75	11416.6	48.26923	1235.404	918.1154	518.5769	1415.654	388.2115
	7	30215.04	1760.058	1069.115	289.2885	855.6154	286.8654	57154.33	6650.038	910.3269	489.3654	1059.019	8096.596	10628.17	875.8846	2165.462
	8	9491.308	12998.62	1458.673	1133.769	881.9808	531.5385	7702.192	158487.2	1443.058	3594.635	1490.385	33217.92	29226.52	1354.385	1475.865
	9	6654.154	698.6923	43606.79	150.7885	2836.538	27760.69	1080.077	2077.865	241409.3	239.5192	4307	2909.423	1789.269	7553.308	2348.231
	10	1641.942	39160.58	232.6923	1707.442	137.2115	61.5	721.4038	7414.596	295.9615	42409.19	193.4038	37941.88	1827.212	166.3654	235.9038
3	11	9346.288	632.9423	6061.865	95.36538	3399.942	1181.615	1009.423	2285.019	3591.519	155.0577	34906.81	4298.673	1336.865	17858.56	4577.788
	12	15466.38	72877.13	2977.904	22782.04	2020.096	916.7692	9924.615	107306.4	2245.038	24569.52	3840.846	265285.5	20364.69	2953.385	3046.712
	13	12139.4	5706.077	1456.442	937.3846	970	431.7115	25472.96	23090.17	1094.577	1506.135	1102.25	20563.85	90771.35	1095.904	1595.942
	14	8306.212	594.6731	26362.85	86.53846	3593.577	2241.077	941.7115	1945.942	9158.25	137.7308	9693.135	3310.019	1405.442	52275.23	2640.519
	15	39894.04	883.6346	3665.731	218.8846	9025.885	978.7115	2332.115	2577.25	2650.365	251.25	12595.33	4270.788	2303.192	4200.827	40906.08

	1.North	2.Central	3.South
1.North	T_{K11}	T_{K21}	T_{K31}
2.Central	T_{K12}	T_{K22}	T_{K32}
3.South	T_{K13}	T_{K23}	T_{K33}

Figure 4.1 Merging O-D Table

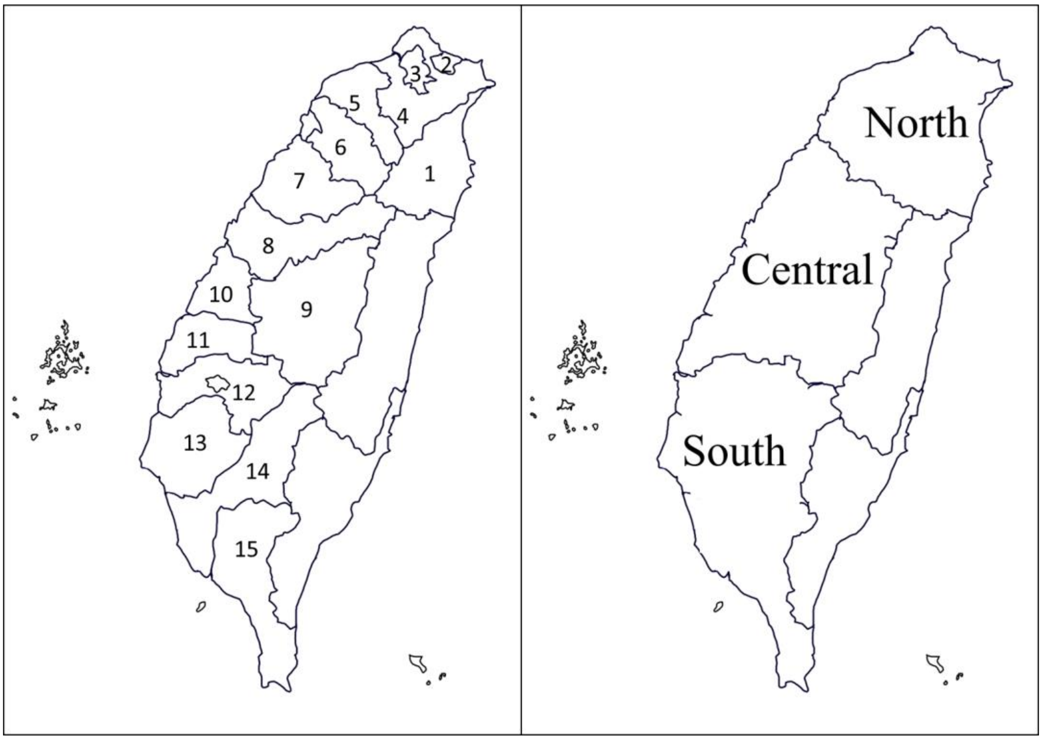

Figure 4.2 Position of O-D

Table 4.2 Average number of trips for each cluster

Cluster	O\D	North	Central	South
1	North	1,585,588	107,179	20,652
	Central	66,426	640,236	63,999
	South	13,256	47,281	693,477
2	North	1,129,233	127,454	33,608
	Central	105,015	519,057	87,687
	South	28,573	75,557	640,404
3	North	1,822,450	100,055	20,536
	Central	58,406	710,461	56,901
	South	12,555	40,311	749,725
4	North	1,327,833	99,604	18,007
	Central	87,940	558,580	65,108
	South	21,464	56,999	636,643
5	North	1,403,556	112,649	21,648
	Central	100,197	658,943	81,387
	South	26,469	70,351	720,888

Based on the distance of trips, we categorize trips into short- (T_{K_Short}), medium-(T_{K_Mid}), and long- (T_{K_Long}) trips; the short-distance trips are also considered regional trips. For those trips across regions, we further categorize them into south-bound (T_{K_South}) and north-bound (T_{K_North}) trips. The number of trips for each type in each cluster can be calculated according to the formula defined in Table 4.3, and Table 4.4 summarizes the calculation.

Table 4.3 Definition of trip type

Type of trip	Sum of cell
T_{K_Short}	$T_{K11} + T_{K22} + T_{K33}$
T_{K_Mid}	$T_{K21} + T_{K32} + T_{K12} + T_{K23}$
T_{K_Long}	$T_{K31} + T_{K13}$
T_{K_South}	$T_{K12} + T_{K13} + T_{K23}$
T_{K_North}	$T_{K21} + T_{K31} + T_{K32}$

From Table 4.4, we can find that Cluster 3 has the most of trips, and most of them are short-distance or regional trips. Conversely, Cluster 2 is the smallest cluster in terms of the number of trips. However, it contains the highest numbers of long-distance and cross-region trips (in both south-bound and north-bound directions).

Table 4.4 Calculation results of trip types in each cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
T_{K_Short}	2,919,302	2,288,695	3,282,637	2,523,058	2,783,389
T_{K_Mid}	284,887	395,714	255,675	309,653	364,585
T_{K_Long}	33,909	62,182	33,092	39,471	48,117
T_{K_North}	191,832	248,749	177,493	182,720	215,684
T_{K_South}	126,964	209,146	111,274	166,404	197,019
$\sum_{O,D} T_{KOD}$	3,238,099	2,746,592	3,571,405	2,872,183	3,196,093

In Table 4.5, the trip patterns in the identified clusters are directly compared in terms of the values of the cells in the O-D tables, and Clusters 2 and 3 are selected comparison references for their definite features aforementioned. Table 4.5 shows the relative percentage differences between the numbers of trips for specific O-D pairs in two clusters, which are calculated by using Equation (4.1):

$$\delta(O,D)_{K=a/K=b} = \frac{T_{bOD} - T_{aOD}}{T_{bOD}} \times 100 \tag{4.1}$$

Table 4.5 shows the comparison of clusters in terms of the number of trips, where Clusters 2 and 3 are taken as the base cases. It can be found from Table 4.5 that Cluster 2 has larger trip volumes for those cells related to medium- and long-distance trips, while in Cluster 3 the cells of short-distance trips take larger numbers.

Table 4.5 Comparison with respect to Clusters 2 and 3

	2			3		
1	-28.78%	18.92%	62.73%	14.94%	-6.65%	-0.57%
	58.09%	-18.93%	37.01%	-12.07%	10.97%	-11.09%
	115.55%	59.80%	-7.65%	-5.28%	-14.74%	8.11%
2				61.39%	-21.50%	-38.90%
				-44.38%	36.88%	-35.11%
				-56.06%	-46.65%	17.07%
3	-38.04%	27.38%	63.65%			
	79.80%	-26.94%	54.10%			
	127.57%	87.43%	-14.58%			
4	-14.96%	27.96%	86.64%	37.25%	0.45%	14.04%
	19.42%	-7.08%	34.68%	-33.58%	27.19%	-12.60%
	33.12%	32.56%	0.59%	-41.50%	-29.28%	17.76%
5	-19.54%	13.14%	55.25%	29.85%	-11.18%	-5.14%
	4.81%	-21.23%	7.74%	-41.71%	7.82%	-30.09%
	7.95%	7.40%	-11.16%	-52.56%	-42.70%	4.00%

Table 4.6 compares the difference between Clusters 1, 4, and 5. It shows that all cells in Cluster 5 have larger numbers compared with those in Clusters 1 and 4, except T_{111} of Cluster 1 is larger than T_{411} and T_{511}. In contrast to Cluster 1, Cluster 4 has a large number of trips in the northbound direction.

Table 4.6 Comparison between Clusters 1, 4 and 5

	1			5		
4	19.41%	7.60%	14.69%	5.70%	13.10%	20.22%
	-24.46%	14.62%	-1.70%	13.94%	17.97%	25.00%
	-38.24%	-17.05%	8.93%	23.32%	23.42%	13.23%
5	12.97%	-4.86%	-4.60%			
	-33.71%	-2.84%	-21.36%			
	-49.92%	-32.79%	-3.80%			

The characteristics and distribution patterns of trips in each cluster are summarized as follows:

Cluster 1

Cluster 1 has the second largest number of trips among all clusters and has higher numbers of short-distance trips to the north region.

Cluster 2

Cluster 2 primary contains the trip patterns for the Chinese New Year, which has higher numbers of long- and medium-distance trips.

Cluster 3

Cluster 3 is characterized by the trip patterns for the workdays (right before and after holidays), having higher numbers of short and regional trips.

Cluster 4

Averagely, Cluster 4 has the fourth largest number of the total trip volume among all clusters and has a little bit higher number of northbound trips.

Cluster 5

Cluster 5 contains the third largest number of the total trip volume among all clusters; however, compared with other clusters, it seems to have no specific characteristics in its trip distribution.

4.3 Cluster Comparison

4.3.1 Comparison between clusters 1 and 3

Samples from the day before and after the holiday are only assigned to Clusters 1 and 3 (except one day assigned to Cluster 4). The samples in Cluster 1 are primarily from the days before holidays, the days after holidays and some other days during holidays. Cluster 3 contains only the days before and after holidays.

The days before and after holidays in Cluster 1 also reveal the trip pattern of the holiday; however, the trip distributions of the days before and after holidays in Cluster 3 may be more similar to the trip pattern of a general workday. In order to more explicitly distinguish the trip characteristics of the days before and after holidays, we take Cluster 3 as a reference, seeking to screen out the difference between Cluster 1 and Cluster 3 and use logistic regression for relevant analysis. The analysis results are shown in Table 4.7. If the sign of a variable is positive, it implicates that the variable will increase the probability of the day (before or after a holiday) assigned to Cluster 1; that is, it may be more of the characteristics of the holiday trip pattern. If the sign is negative, it implies that the associated variable will increase the probability of the day (before or after a holiday) belonging to Cluster 3, which is more likely to reveal the trip pattern of general workdays. The cells of “Sign” column colored red mean that the variable is significant at the level of 0.05. The analysis results show that if the day (before or after a holiday) is on Monday, its probability of belonging to Cluster 1 increases; if it is on Friday, then its probability of belonging to Cluster 1 decreases. If the oil price rises, the probability of a sample assigned to Cluster 1 also increases. Additionally, the longer the length of the holiday, the higher the probability that the associated samples belong to Cluster 1, but this effect is not significant.

Table 4.7 Analysis results of the days before and after holidays in Clusters 2 and 3 by logistic regression

Variable	**Sign**	**Description**
Intercept	-	
MON	+	If the day is on Monday, the value is 1, otherwise is 0.
FRI	-	If the day is on Friday, the value is 1, otherwise is 0.
Total_day	+	The length of the holiday
Oil price	+	The 95 unleaded gasoline price, TWD/liter.

4.3.2 Comparison between clusters 1, 4 and 5

By observing Clusters 1, 4 and 5, it is difficult to determine the influencing factors that distinguish the samples in these three clusters. Hence, a multi-category logistic regression model is developed for the comparison and the analysis results are summarized in Table 4.8. If Sign(1) of a variable is positive, it indicates that the variable can increase the probability of a sample belonging to Cluster 1. If Sign(5) of a variable is positive, the variable can increase the probability of a sample belonging to Cluster 5. If Sign(1) and Sign(5) are both positive, it means that the associated variable can reduce the probability of a sample belonging to Cluster 4.

In Table 4.8, the cells of "Sign(1)" and "Sign(5)" columns colored red also mean that the variable is significant at the level of 0.05. The analysis results show that the first day of a holiday may increase the probability of a sample belonging to Cluster 1 (but not significant) and reduce the probability of a sample belonging to Cluster 5. The last day of a holiday can reduce the probability of a sample belonging to Clusters 1 and 5, thereby increasing the probability of it belonging to Cluster 4. Since both the first day and last day of a can reduce the probability of a sample belonging to Cluster 5, it is inferred that the middle day(s) of a holiday may lead to a higher probability for a sample to belong to Cluster 5. In addition, if the oil price rises, it may also reduce the probability of a sample belonging to Clusters 1 and 5 and by contrast increase the probability of it belonging to Cluster 4.

Table 4.8 Analysis results of Clusters 1, 4 and 5 by logistic regression

Variable	**Sign(1)**	**Sign(5)**	**description**
Intercept	+	+	
Total_day	-	-	The length of Holiday
First_day	+	-	If it is the day before the holiday, the value is 1, otherwise it is 0.
Final_day	-	-	If it is the first day of the holiday, the value is 1, otherwise it is 0.
Rainfall_TPE	+	+	The rainfall of the Taipei Weather Station on that day, in millimeters.
Rainfall_TAI	-	-	The rainfall of the Taichung Weather Station on that day, in millimeters.
Rainfall_KAO	+	+	The rainfall at the Kaohsiung Weather Station on that day, in millimeters.
Oil price	-	-	The 95 unleaded gasoline price, TWD/liter.
New years	-	+	If it belongs to New Year's Day, the value is 1, otherwise it is 0.
Chinese New Year	+	+	If it belongs to Chinese New Year, the value is 1, otherwise it is 0.
Peace	+	+	If it belongs to Peace Memorial Day, the value is 1, otherwise it is 0.
Tomb Sweeping	-	+	If it belongs to Tomb Sweeping Festival, the value is 1,

			otherwise it is 0.
Dragon Boat	-	+	If it belongs to Dragon Boat Festival, the value is 1, otherwise it is 0.
Moon	+	+	If it belongs to Moon Festival, the value is 1, otherwise it is 0.

4.4 Cluster Description and Discussion

Considering the clustering and model analysis results, the following descriptions with respect to each cluster can be drawn as a reference for the traffic management measures on holiday:

Cluster 1

The length of the holiday can reduce the probability of a sample (O-D table) belonging to Cluster 1 (but not significant), and we can infer that if the length of the holiday is shorter, the trip distribution pattern on the first day of holiday may have shorter-distance trips in Northern Taiwan. In addition, if the day after the holiday is on Monday, there are usually more commuting trips, so there are also more short-distance trips within regions, which may be the major feature of the trip distribution pattern in this cluster.

Cluster 2

The 6-day long Chinese New Year may generally involve a high volume of home-returning trips, so during the first five days of Chinese New Year, trips are usually of long or medium distances.

Cluster 3

This cluster contains mostly the trips during weekdays. All the samples in the cluster are the days before and after the holidays, which are significantly characterized by short-distance or regional trips.

Cluster 4

This cluster has the largest number of the samples from the last day of a holiday. The analysis results show that the last day of a holiday can significantly increase the probability of a sample belonging to Cluster 4, and it reveals a strong tendency that people travel back to the place of their residence for the start of the workways on the next day.

Cluster 5

This cluster has the largest number of the middle days during holidays. The analysis results also suggest that the middle days of a holiday significantly increase the probability of a sample belonging to Cluster 5. Therefore, the numbers of trips in this cluster are relatively large, but more evenly distributed without explicit characteristics related to trip directions, and the number of regional trips is larger. It is inferred that the trips made during the middle days of a holiday (such as the 2nd day in a 3-day holiday or the 2nd and 3rd days in a 4-day holiday) tend to go back and forth to a certain location within a single day, thereby increasing the number of regional trips.

5. CONCLUSION

Based on the investigation of daily O-D tables across the freeway system in Taiwan, this study identifies that the difference in trip distribution patterns between workdays and holidays is significant. Most of the trips made on weekdays are short-distance, within a region, and there are more long- or medium-distance trips during holidays. Because there are more short-distance trips on weekdays, the number of trips is also the largest.

Regarding the days before and after holidays, the analysis results suggest that if the day after a holiday is on Monday, the trip distribution pattern be similar to it of a holiday trip, while if the day before a holiday is on Friday, the trip distribution pattern may reveal to the pattern of weekday trips. A likely reason is that the freeway users are not inclined to take a leave on Friday (the day before the holiday), but they tend to ask for a leave on Monday (the day after the holiday) so that they can extend the holiday.

In terms of the characteristics of holiday trips, it is generally considered that if the national holiday is a folk holiday, the purpose of trip is to complete custom activities, such as tidying up the gravesite, returning home, and so on, probably revealing some specific trip characteristics. However, according to analysis of this study, the trip distribution pattern on freeway is less affected by the type of holiday, and the variables related to holiday types are not significant. The exception is Chinese New Year; its trip distribution pattern characteristics are relatively distinguishable. According to experience, the 6-day long Chinese New Year can result in more long- or medium-distance trips.

In addition, according to analysis of this study, the oil price affects the trip distribution pattern on freeway, and the variables are significant. The correlation between oil price and short-distance trips is negative. When oil price raised, the number of short-distance trips would reduce, and increase the probability of a sample at the days before and after holidays belonging to Cluster 1 and on national holidays belonging to Cluster 4.

Therefore, planning traffic management on national holidays should consider the oil price, when oil price is lower, traffic management should be stricter because of short-distance trips increasing.

REFERENCES

Wang, J.-C., Guo, J.-G. (2004) *Logistic Regression Models*, Taiwan: WU-NAN.

Freeway Bureau, Ministry of Transportation and Communications, Taiwan, Republic of China (2015) *Traffic Data Collection System Manual*.

Chiou, Y.-C., Wen, C.-H., Chen, W.-Y., Wang, M.-H, Fu, C., Tseng, H.-M., Lin, Y.-S. (2014) *Modeling Mode Choice Behaviors by Using Mixed Logit Models*, Institute of Transportation (IOT), Ministry of Transportation and Communications, Taiwan, Republic of China.

Tong C.-C., Feng C.-H. (2018) T*aste Variation in Disaggregate Choice Model：A Comparative Study of User Segmentation Techniques*. Tamkang University.

Buttrey, S.E., Karo, C. (2002) Using K-nearest-neighbor classification in the leaves of a tree. *Computational Statistics and Data Analysis*, 40, 27-37.

Lin, C.-K., Chang, C.-H., Su, J.-M., Chang, C.-N., Shen, M.-H., Tsai, C.-T. (2018) Applying smart card data to estimate and validate the destination of individual trips in transit systems. *Transportation Planning Journal*, 47(1), 1-28.

Martin, M., Gale, C., Cockings, S., Harfoot, A. (2018) Origin-destination geodemographics for analysis of travel to work flows. *Computers, Environment and Urban Systems*, 67, 68-79.

Yeh, I.-S. (2017) The analysis of the crowd moving in taipei mrt system, Journal of the Chinese Statistical Association Vol. 55, pp. 69–95.