Data-Driven Prediction Methodology of Origin–Destination Demand in Large Network for Real-Time Service

Soomin Woo, Sehyun Tak, and Hwasoo Yeo

Prediction of origin–destination (O-D) demand is an important topic in transportation engineering because it is a crucial input for a dynamic traffic management and control system. Previous literature has focused primarily on estimation and prediction of O-D demand with Kalman filtering; however, these forecasts lack efficiency when unusually fluctuating O-D demand of a large O-D network is predicted in real time. With true, real-time O-D demand data from South Korean expressways, a data-driven prediction framework of O-D demand in a large network for real-time service is proposed by modifying the *k*–nearest neighbor (*k*-NN) algorithm. Three strategies that implement different feature vectors for *k*-NN prediction of single-level O-D demand, multilevel O-D demand, and single-level point demand are proposed. The strategies were tested on hourly O-D demand in South Korea. The average mean absolute percentage of error values of the three strategies in terms of total demand are 5.52%, 5.34%, and 3.36%, respectively; single-level point demand performs slightly better than do the other two strategies. Similarly, for the average mean absolute percentage and weighted average mean absolute percentage in terms of individual O-D demand, single-level point demand performs better than do the two other strategies, especially for O-D pairs with larger demand and for further prediction horizons. In addition, the single-level point demand shows the highest computation efficiency. Therefore, the single-level point demand strategy for *k*-NN prediction shows the best combination of accuracy and computation efficiency among the three strategies. Furthermore, a historical database size of at least 300 dates for this data-driven prediction algorithm seems required for accuracy.

An advanced traffic management system takes the important role of mitigating increasing congestion on the roadways. Functionalities of the system include prediction of the traffic conditions and simulation for control purposes. Also, providing information from origin to destination has an important role in the advanced traffic management system because it may influence people's movement, for instance, by prediction of route travel time and suggestion of optimal routes (*1, 2*). Recently, there have been studies to develop this management system even further to optimize control strategies on the road by using online traffic simulation for unusual situations, such as accidents or road work, and future origin–destination (O-D) demand is a crucial input to this simulation. The system will require a prediction method that uses real-time data for better accuracy and that captures sudden fluctuations of O-D demand. Also, the system will require future O-D demand from the whole O-D network, which is often large and requires high computation efficiency.

Most of the literature on O-D demand prediction methods has taken a time series approach, implementing various modifications of the Kalman filter (*3–14*). These studies estimate the O-D demand from point traffic measurements with an assignment matrix and predict the deviations of O-D demand from structural fluctuation. Many of these studies have attempted to modify the conventional time series prediction to real time by adopting real-time point measurements, such as density and flow, and updating their parameters in real time or by sampling travel time with wireless fidelity (Wi-Fi) or Bluetooth technology and updating their O-D seed matrix (*3, 4, 7, 9–12*). The core purpose of real-time O-D demand prediction is to use the most up-to-date information to accurately predict future demand such as sudden fluctuations that may even happen hourly; however, these autoregressive methods are limited in efficiency in capturing these sudden changes. For instance, a study by Zhou and Mahmassani updated the autoregressive terms to incorporate daily structural fluctuations, but this method may not cover the large hourly fluctuations of O-D demand (*3*). If the autoregressive terms with higher complexity are used to cover sudden fluctuations, the computation burden will also increase with less robust efficiency.

Some studies have attempted to apply time series prediction to a large-scale network, but this method has to overcome the curse of dimensionality from a large number of O-D pairs (*10, 11*). Principal component analysis extracts O-D pairs of reduced dimension and uses a similar Kalman filtering method. Still, these studies have the short prediction horizon of the Kalman filter, which may be shorter than the travel times between some origins and destinations in a large network. Also, the principal components could be found with a simple configuration of the network; however, for a large network with a highly complex configuration like that of the highway network of South Korea, finding the principal components may have to sacrifice accuracy, since each O-D pair may not have high relevance to the others.

O-D demand prediction so far has been developed mostly with parametric time series methods, which lack the capture of the high fluctuations of hourly O-D demand with real-time data and cannot be applied efficiently to a large network with a highly complex configuration. The limitations of these studies have largely contributed to the absence of observable information of true O-D demand or an assignment matrix. However, true O-D demand of a large network in the form of an historical data reserve as well as real-time data are

Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-Ro, Yuseong-Gu, Daejeon, South Korea. Corresponding author: H. Yeo, hwasoo@kaist.ac.kr.

Transportation Research Record: Journal of the Transportation Research Board, No. 2567, Transportation Research Board, Washington, D.C., 2016, pp. 47–56.
DOI: 10.3141/2567-06

available for almost the entire highway network of South Korea. The majority of the network is governed by the Korea Expressway Corporation and is priced by an automatic toll collection system that records the entrance and exit of all vehicles in the network. Therefore, there is the opportunity to explore the possibility of an O-D demand prediction methodology that is independent of the estimation process and makes full use of historical data. Then it is possible to use an alternative method to predict the future O-D demand in a data-driven approach, in order to capture large fluctuations of O-D demand from real-time data and to cover a large O-D network with high computation efficiency.

One candidate to meet the goals just mentioned is the simple and intuitive pattern-matching method using *k*–nearest neighbors (*k*-NN). Modified from the conventional version, the proposed method will search from the large historical data reserve for the *k*-number of dates with the most similar traffic pattern to the current traffic state, which should have a similar traffic pattern change as the future traffic state. Also, this method must achieve computation efficiency for real-time applications. To improve the computation speed, three strategies with different feature vectors for the proposed *k*-NN prediction are compared.

The two objectives of this study are to develop a framework to predict the future O-D demand of a large network for real-time service by using a data-driven approach and to compare three matching strategies for the proposed method to achieve the best combination of accuracy and computation efficiency. In the following, a real-time framework for the proposed O-D demand prediction is explained with a focus on three pattern-matching strategies. The comparison criteria and data used are described as well as the results of the comparison in terms of accuracy and efficiency.

REAL-TIME FRAMEWORK OF DATA-DRIVEN O-D DEMAND PREDICTION

The framework for O-D demand prediction in a data-driven environment for real-time service is shown in Figure 1. It has three major steps: data preprocessing, *k*-NN matching, and *k*-NN prediction. First, in the data preprocessing, the necessary data are accumulated, filtered, and sent to the *k*-NN matching module. Second, in the *k*-NN matching, the subject data are compared with historical data to calculate their similarity in traffic patterns and find the *k*-NNs. Finally, the historical *k*-NN data are integrated as prediction values.

Data Preprocessing

In the data preprocessing step, necessary data for prediction including real-time data and historical data are preprocessed to increase accuracy and efficiency. First, the real-time data are acquired every hour by using the Open API interface. These raw data are reformatted to standard feature vector frames in Module 1. The processed real-time data are sent to the *k*-NN matching step and saved in the historical database. To increase the accuracy of the prediction, only the historical data of the same day type as the current real time are filtered in Module 2 and used for the *k*-NN matching. The format of the historical data is identical to that of the preprocessed real-time data.

k-NN Matching

In the *k*-NN matching step, the subject real-time data are compared with historical data to find the *k*-NN dates that will provide possible

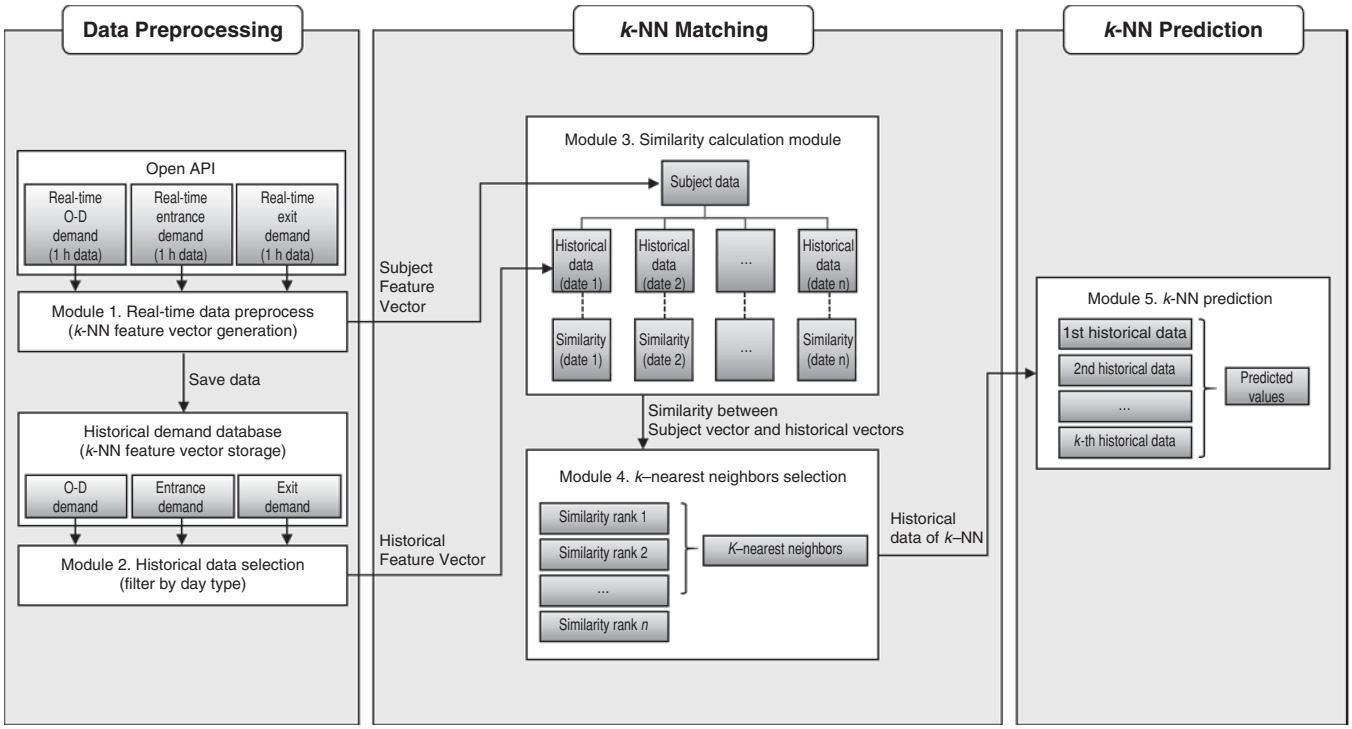

FIGURE 1 Framework of data-driven O-D demand prediction for real-time service.

future demand values. Module 3 receives both the real-time subject feature vector and the historical feature vector from the previous step and calculates the similarity between the subject data and each item of the historical data. Next, Module 4 ranks the similarity of each historical datum, extracts the *k*-NN data with the most similarity from the historical data, and passes them on to the *k*-NN prediction step.

Conventional k-*NN Matching Procedure*

Before three modifications of the *k*-NN matching procedure are introduced, the conventional approach is briefly described. An example of the conventional feature vector of dimension *d* is shown in Equation 1. The subject feature vector x and historical feature vector y^i have identical formats. Then the similarity between x and each y_i value for all i in the historical data size of N is calculated. The most conventional similarity metric, the Euclidean distance between x and y_i as S_i, is expressed in Equation 2, which corresponds to Module 3. Then the *k*–number of historical data with the best similarity values or least Euclidean distance are selected, with a parameter *k*. This step corresponds to Module 4.

$$x = (x_1, x_2, \ldots, x_d) \qquad y^i = \left(y_1^i, y_2^i, \ldots, y_d^i\right) \tag{1}$$

$$S\left(x, y^i\right) = \sqrt{\sum_{n=1}^{C} \left(x_n - y_n^i\right)^2} \qquad \forall i \in 1, \ldots, N \tag{2}$$

where n is a dimension of the feature vector.

This simple pattern-matching technique can be modified for the prediction of O-D demand data by changing the matching strategy and its feature vectors. Three modifications are shown in the following sections: single-level O-D demand, multilevel O-D demand, and single-level point demand.

Single-Level O-D Demand Matching

In this modification, the O-D demand from both real-time and historical data is directly used for similarity matching. From the conventional Kalman filter methods of O-D demand prediction, there is a temporal correlation of demand for each O-D pair (*3–11*). Therefore, the *k*-NN feature vector can be modified to include both temporal and spatial dimensions in a matrix for efficiency of computing. The temporal range of the feature vector can be defined as $[t - \tau, t - 1]$ for both subject and historical data. The value τ represents how much of the temporal correlation of the immediate history is wanted to find a similar pattern. Also, the target O-D pairs for prediction are all possible O-D pairs in the network, in which each toll gate station can serve as both entrance and exit. With R toll gate stations, the feature vector for O-D demand prediction can be expressed as follows:

$$x(t) = \begin{bmatrix} d_{1,1}(t-\tau) & d_{1,2}(t-\tau) & \cdots & d_{1,R}(t-\tau) & \cdots & d_{R,R}(t-\tau) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{1,1}(t-1) & d_{1,2}(t-1) & \cdots & d_{1,R}(t-1) & \cdots & d_{R,R}(t-1) \\ d_{1,1}(t) & d_{1,2}(t) & \cdots & d_{1,R}(t) & \cdots & d_{R,R}(t) \end{bmatrix} \tag{3}$$

where

$x(t)$ = subject O-D demand data for each O-D pair in network at time t,
$d_{s,e}(t)$ = O-D demand at time t from station s to station e,
s = origin toll gate,
$s \in 1, \ldots, R$, e = destination toll gate, and
$e \in 1, \ldots, R$ and τ = temporal range of feature vector.

$$y^i(t) = \begin{bmatrix} d_{1,1}^i(t-\tau) & d_{1,2}^i(t-\tau) & \cdots & d_{1,R}^i(t-\tau) & \cdots & d_{R,R}^i(t-\tau) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{1,1}^i(t-1) & d_{1,2}^i(t-1) & \cdots & d_{1,R}^i(t-1) & \cdots & d_{R,R}^i(t-1) \\ d_{1,1}^i(t) & d_{1,2}^i(t) & \cdots & d_{1,R}^i(t) & \cdots & d_{R,R}^i(t) \end{bmatrix} \tag{4}$$

where $y^i(t)$ is the historical O-D demand data for each O-D pair in the network at time t and $i \in 1, \ldots, N$.

The similarity calculation can be modified further to consider weighting on each demand by its contribution to the total demand, as shown in Equation 5. The historical data N correspond to those of the same day type as the subject data. Next, the *k*-NNs with the least distance are selected as Module 4.

$$\text{dist}_i^{\text{single}}(t) = S^{\text{weight}}\left(x(t), y^i(t)\right) \qquad \forall i \in 1, \ldots, N \tag{5}$$

where

$$S^{\text{weight}}\left(x, y^i\right) = \sqrt{\sum_{n=1}^{C}\sum_{m=1}^{L} w_{m,n}\left(x_{m,n} - y_{m,n}^i\right)^2}$$

$$w_{m,n} = \frac{x_{m,n}}{\sum_{n=1}^{C}\sum_{m=1}^{L} x_{m,n}}$$

and

$\text{dist}_i^{\text{single}}(t)$ = dissimilarity measure at time t between subject data and historical data i,
S^{weight} = weighted Euclidean distance,
$w_{m,n}$ = weighting factor from demand size,
$x_{m,n}$ = value at mth row and nth column of x,
$y_{m,n}^i$ = value at mth row and nth column of y^i,
L = number of rows of matrix, and
C = number of columns of matrix.

When the number of toll gate stations R increases, the possible combinations of O-D pairs increase rapidly and the computation burden becomes much heavier. Therefore, two alternative strategies for *k*-NN matching are proposed next in an attempt to increase the computation efficiency.

Multilevel O-D Demand Matching

The first alternative strategy for increasing the computation efficiency of *k*-NN matching adopts the idea of "divide and conquer," also proposed by Cantelmo et al. (*13*) and Frederix et al. (*15*). The authors

suggested dividing the network into subnetworks and predicting the O-D demand in a hierarchical approach: first the O-D demand between subnetworks is considered and second O-D demand in the original dimension.

Under this multilevel strategy, *k*-NN matching occurs twice: primary and secondary matching. In the primary matching, a feature vector with reduced dimension is achieved from aggregating O-D demand between toll gates into O-D demand between subnetworks. The feature vector of aggregated demand is shown in Equations 6 and 7. The spatial dimension is reduced from R^2-value to $(R^{agg})^2$-value, where the size of subnetworks is R^{agg}. For instance, the demand data between 345^2 O-D pairs from the Korean Expressways Corporation can be aggregated to become demand data between only 8^2 O-D subnetwork pairs. The similarity is calculated as Equation 8 for each historical datum with size *N*. Then the historical data in order of least distance are selected for the primary matching result with parameter $k_{primary}$.

$$x^{agg}(t) = \begin{bmatrix} d^{agg}_{1,1}(t-\tau) & d^{agg}_{1,2}(t-\tau) & \cdots & d^{agg}_{1,R^{agg}}(t-\tau) & \cdots & d^{agg}_{R^{agg},R^{agg}}(t-\tau) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d^{agg}_{1,1}(t-1) & d^{agg}_{1,2}(t-1) & \cdots & d^{agg}_{1,R^{agg}}(t-1) & \cdots & d^{agg}_{R^{agg},eR^{agg}}(t-1) \\ d^{agg}_{1,1}(t) & d^{agg}_{1,2}(t) & \cdots & d^{agg}_{1,R^{agg}}(t) & \cdots & d^{agg}_{R^{agg},R^{agg}}(t) \end{bmatrix} \quad (6)$$

$$y^{agg,i}(t) = \begin{bmatrix} d^{agg,i}_{1,1}(t-\tau) & d^{agg,i}_{1,2}(t-\tau) & \cdots & d^{agg,i}_{1,R^{agg}}(t-\tau) & \cdots & d^{agg,i}_{R^{agg},R^{agg}}(t-\tau) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d^{agg,i}_{1,1}(t-1) & d^{agg,i}_{1,2}(t-1) & \cdots & d^{agg,i}_{1,R^{agg}}(t-1) & \cdots & d^{agg,i}_{R^{agg},R^{agg}}(t-1) \\ d^{agg,i}_{1,1}(t) & d^{agg,i}_{1,2}(t) & \cdots & d^{agg,i}_{1,R^{agg}}(t) & \cdots & d^{agg,i}_{R^{agg},R^{agg}}(t) \end{bmatrix} \quad (7)$$

where

$x^{agg}(t)$ = aggregated subject O-D demand for each O-D subnetwork pair at time *t*,
$y^{agg,i}(t)$ = aggregated historical O-D demand for each O-D subnetwork pair at time *t* and $i = 1, \ldots, N$, and
$d^{agg}_{s^{agg},e^{agg}}(t)$ = aggregated O-D demand at time *t* from subnetwork s^{agg} to subnetwork e^{agg} and $s^{agg}, e^{agg} \in 1, \ldots, R^{agg}$.

$$\text{dist}^{agg}_i(t) = S^{weight}\left(x^{agg}(t), y^{agg,i}(t)\right) \qquad \forall i \in 1, \ldots, N \quad (8)$$

where $\text{dist}^{agg}_i(t)$ is the dissimilarity measure at time *t* between aggregate subject data and aggregate historical data *i*.

In the secondary matching, a feature vector with original dimension is used. This time, the subject data are compared with only the selected historical data of $k_{primary}$ from the primary matching instead of all historical data with size *N*. The feature vector for the secondary matching looks identical to those of Equations 3 and 4. The similarity calculation is shown in Equation 9 now with $k_{primary}$ number of iterations. Last, among the candidates from $k_{primary}$, the $k_{secondary}$ number of historical data with the least distance is selected and used for prediction.

$$\text{dist}^{secondary}_j(t) = S^{weight}\left(x(t), y^j(t)\right) \qquad \forall i \in 1, \ldots, k_{primary} \quad (9)$$

where $\text{dist}^{secondary}_j(t)$ is the dissimilarity measure at time *t* between subject data and historical data *j* from the first matching.

k-NN Matching: Single-Level Point Demand

Another alternative is to use demand data separated from the O-D pair information: the entrance and exit demand. A similar approach was taken by Ashok and Ben-Akiva, in which departure rate deviations at origins were considered for O-D demand prediction (*7, 16*). This method is an efficient way to reduce the size of the feature vector, whose dimension becomes $2R$ from R^2 because the feature vector has entrance counts and exit counts from each individual toll gate station.

The matching procedure is single level with slightly different feature vectors, since the entrance and exit demand are separated as shown in Equations 10 and 11. The similarity calculation is as shown in Equation 12. The *k*-number of historical date indexes that are found to have the most similar traffic patterns to the subject data are extracted, and the corresponding O-D data are used for the *k*-NN prediction step. The final *k*-NNs contain O-D data of the historical dates that have the most similar entrance and exit demand pattern to the subject. The corresponding O-D demand is fed into the prediction step.

$$x^{in}(t) = \begin{bmatrix} d^{in}_1(t-\tau) & d^{in}_2(t-\tau) & \cdots & d^{in}_R(t-\tau) \\ \vdots & \vdots & \vdots & \vdots \\ d^{in}_1(t-1) & d^{in}_2(t-1) & \cdots & d^{in}_R(t-1) \\ d^{in}_1(t) & d^{in}_2(t) & \cdots & d^{in}_R(t) \end{bmatrix} \quad (10a)$$

$$x^{out}(t) = \begin{bmatrix} d^{out}_1(t-\tau) & d^{out}_2(t-\tau) & \cdots & d^{out}_R(t-\tau) \\ \vdots & \vdots & \vdots & \vdots \\ d^{out}_1(t-1) & d^{out}_2(t-1) & \cdots & d^{out}_R(t-1) \\ d^{out}_1(t) & d^{out}_2(t) & \cdots & d^{out}_R(t) \end{bmatrix} \quad (10b)$$

where

$x^{in}(t)$ = subject entrance demand at time *t*;
$d^{in}_s(t)$ = entrance demand at station *s* at time *t*;
$s = 1, \ldots, R$;
$x^{out}(t)$ = subject exit demand at time *t*;
$d^{out}_e(t)$ = exit demand at station *e* at time *t*; and
$e = 1, \ldots, R$.

$$y^{in,i}(t) = \begin{bmatrix} d^{in,i}_1(t-\tau) & d^{in,i}_2(t-\tau) & \cdots & d^{in,i}_R(t-\tau) \\ \vdots & \vdots & \vdots & \vdots \\ d^{in,i}_1(t-1) & d^{in,i}_2(t-1) & \cdots & d^{in,i}_R(t-1) \\ d^{in,i}_1(t) & d^{in,i}_2(t) & \cdots & d^{in,i}_R(t) \end{bmatrix} \quad (11a)$$

$$y^{out,i}(t) = \begin{bmatrix} d^{out,i}_1(t-\tau) & d^{out,i}_2(t-\tau) & \cdots & d^{out,i}_R(t-\tau) \\ \vdots & \vdots & \vdots & \vdots \\ d^{out,i}_1(t-1) & d^{out,i}_2(t-1) & \cdots & d^{out,i}_R(t-1) \\ d^{out,i}_1(t) & d^{out,i}_2(t) & \cdots & d^{out,i}_R(t) \end{bmatrix} \quad (11b)$$

where

$y^{\text{in},i}(t)$ = historical entrance demand at time t; $i = 1, \ldots, N$;
$d_s^{\text{in},i}(t)$ = entrance demand at station s at time t; $s = 1, \ldots, R$;
$y^{\text{out},i}(t)$ = historical exit demand at time t; $i = 1, \ldots, N$; and
$d_s^{\text{out},i}(t)$ = exit demand at station s at time t.

$$\text{dist}_i^{\text{singlepoint}}(t) = \frac{1}{2} \cdot S\left(x^{\text{in}}(t), y^{\text{in},i}(t)\right) + \frac{1}{2} \cdot S\left(x^{\text{out}}(t), y^{\text{out},i}(t)\right) \quad (12)$$

where $\text{dist}_i^{\text{singlepoint}}(t)$ is the dissimilarity measure at time t between subject data and historical data i.

k-NN Prediction

Finally, the future O-D demand is predicted by integrating the O-D demand of the final k-NNs from the k-NN matching step over a time span of the prediction horizon. An example of the predicted value of an O-D pair is shown in Equation 13. The calculation method for integrating these candidate values to make prediction values can be various, such as average or weighted average. Here, the simple average method is chosen for the prediction values.

$$p_{s,e}(t, c) = \frac{\sum_{q=1}^{k} d_{s,e}^{q}(c)}{k} \qquad c \in [t+1, t+\sigma] \quad (13)$$

where

$p_{s,e}(t+\sigma)$ = predicted O-D demand from station s to station e at time c,
t = time at prediction execution,
σ = prediction horizon,
$d_{s,e}^{k}(c)$ = O-D demand value of historical date q from station s to station e at time c, and
k = number of k-NNs selected for prediction.

COMPARISON METHOD

The performance of the three proposed prediction strategies is compared in terms of mean absolute percentage of error (MAPE) and weighted MAPE as shown in Equations 14*a* and 14*b*, respectively. In addition, computation time is evaluated for an efficiency comparison:

$$\text{MAPE} = \frac{1}{D} \sum_{c=t}^{t+\sigma} \sum_{s=1}^{R} \sum_{e=1}^{R} \left| \frac{A_{s,e}(c) - p_{s,e}(c)}{A_{s,e}(c)} \right| \times 100\% \quad (14a)$$

$$\text{weighted MAPE} = \sum_{c=t}^{t+\sigma} \sum_{s=1}^{R} \sum_{e=1}^{R} \left(\left| \frac{A_{s,e}(c) - p_{s,e}(c)}{A_{s,e}(c)} \right| \times \frac{A_{s,e}(c)}{\sum_{\forall s,e} \sum_{c=t}^{t+\sigma} A_{s,e}(c)} \right) \times 100\% \quad (14b)$$

where

$A_{s,e}(c)$ = actual demand value at time c from station s to station e,
$p_{s,e}(c)$ = predicted demand value at time c from station s to station e, and
D = number of elements in predicted O-D demand equal to $R^2(\sigma + 1)$.

The parameters for the k-NN search are as follows: the temporal range of searching k-NN τ is 4 h, and the temporal prediction horizon σ is 6 h. The historical search space ranged from January 7, 2013, to April 15, 2014, and the number of historical dates was about 70 for each day type. The final k-number of nearest neighbors for all three methods chosen is 3. For the multilevel O-D demand strategy, k_{primary} was 10.

All three strategies were tested for 31 days from July 1 to July 31, 2013, at three different times—09:00, 13:00, and 17:00—for a total of 93 samples. At each iteration, the subject date was removed from the historical dates; this step would otherwise make the distance in the similarity calculation zero and would make the traffic pattern a perfect match. The computing capacity used for this study was as follows: Intel Core i5-4670 processor with 3.40 GHz and 16.0 GB of RAM. Each prediction method was allocated half a core to process its entire procedure with R software.

DATA

The O-D demand data (Table 1) and entrance and exit demand data (Table 2) were provided by the Korea Expressway Corporation data portal (http://data.ex.co.kr). The highway network contains 345 toll gates for recording of O-D demand and has on average more than 2 million vehicle trips a day. The network spans the country in the mainland with a total expressway length of approximately 4,000 km (*17*).

RESULTS

This section is in the following order. First, accuracy in terms of total demand is compared between the three prediction strategies. Second, accuracy in terms of demand for individual O-D pairs is compared. Third, computation time is compared. The section concludes with a recommendation of the best prediction strategy.

TABLE 1 Samples of Data Used: O-D Demand on South Korean Highways

System Date	System Time	Toll Gate Code		Vehicle Count by Type						Total Count
		Origin	Destination	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	
20101001	11	101	101	7	1	1	0	0	0	9
20101001	11	101	102	5,413	109	469	21	16	299	6,327
20101001	11	101	103	16,595	666	1,587	93	58	868	19,867

TABLE 2 Samples of Data Used: Entrance and Exit Demand on South Korean Highways

System Date	System Time	Toll Gate Code	In or Out	Vehicle Count by Type						Total Count
				Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	
20101001	11	101	1	470	10	11	13	4	42	550
20101001	11	101	0	447	4	7	0	0	24	482
20101001	11	102	1	614	7	7	19	10	36	693
20101001	11	102	0	603	5	20	0	0	43	671

Accuracy of Prediction: Total Demand

The performance of the three strategies is compared in terms of total demand prediction (Figure 2). In Figure 2*a*, the MAPE values for three prediction strategies are shown for different prediction horizons. The means for the first prediction horizon are 5.52%, 5.34%, and 3.36% for single-level O-D demand, multilevel O-D demand, and single-level point demand, respectively. The single-level point shows better performance with slightly smaller mean and smaller variance compared with the two other methods. In Figure 2*b*, an example case of total demand prediction is shown. The single-level point demand outperforms the two other strategies that use O-D demand.

A possible reason for better accuracy of the single-level point demand on total demand is its feature vector with a smaller dimension; this feature is advantageous in finding a similar historical total demand trend. If the feature vector of O-D demand in Equation 3 is compared with that of the point demand in Equation 6, the large dimension of O-D demand may create a larger bias than the point demand as shown in Figure 2*b*, since it divides the total demand by the large number of O-D pairs. In contrast, in Figure 2*b* the multilevel O-D demand follows a similar trend to the actual demand but with some error. One possible reason is that this strategy uses a low-dimensional feature vector in the primary matching step and finds similar historical dates as well; however, the strategy uses a feature vector of the same dimension as the single-level O-D demand in the secondary matching step and creates some bias.

Accuracy of Prediction: Demand for Individual O-D Pairs

Accuracy is compared in terms of individual O-D pair demand. First, the conventional MAPE value and weighted MAPE value are compared to evaluate the accuracy performance of the three strategies. Second, samples of prediction values and actual values are compared for an additional comparison of accuracy.

Accuracy Comparison with MAPE and Weighted MAPE

Figure 3 shows the MAPE values and weighted MAPE of individual O-D demand prediction with the *x*-axis showing different percentages of O-D demand explained by O-D pairs. For instance, the top 100% of O-D demand represents the demand of large-scale O-D pairs that contribute 100% of the total O-D demand—in other words, O-D demand with counts greater than zero. Likewise, the top

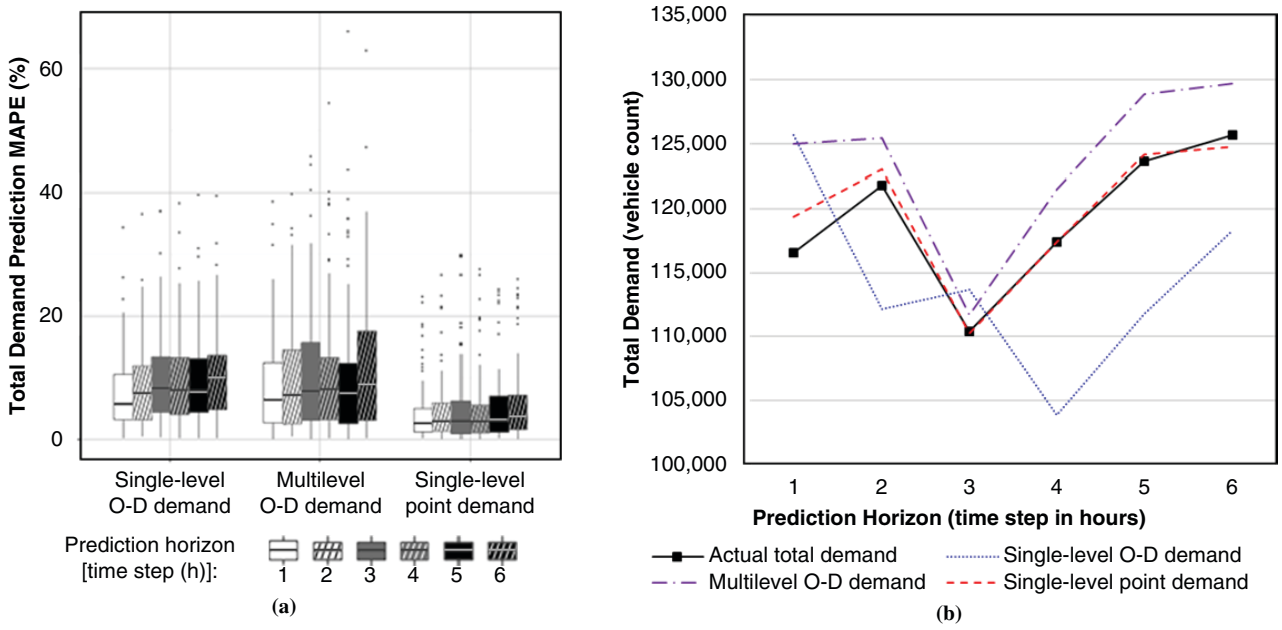

FIGURE 2 Comparison of accuracy with total demand prediction for three prediction strategies: (*a*) MAPE values of total demand prediction by prediction step and (*b*) time series plot of predicted total demand (predicted at 9:00 a.m. on July 1, 2013).

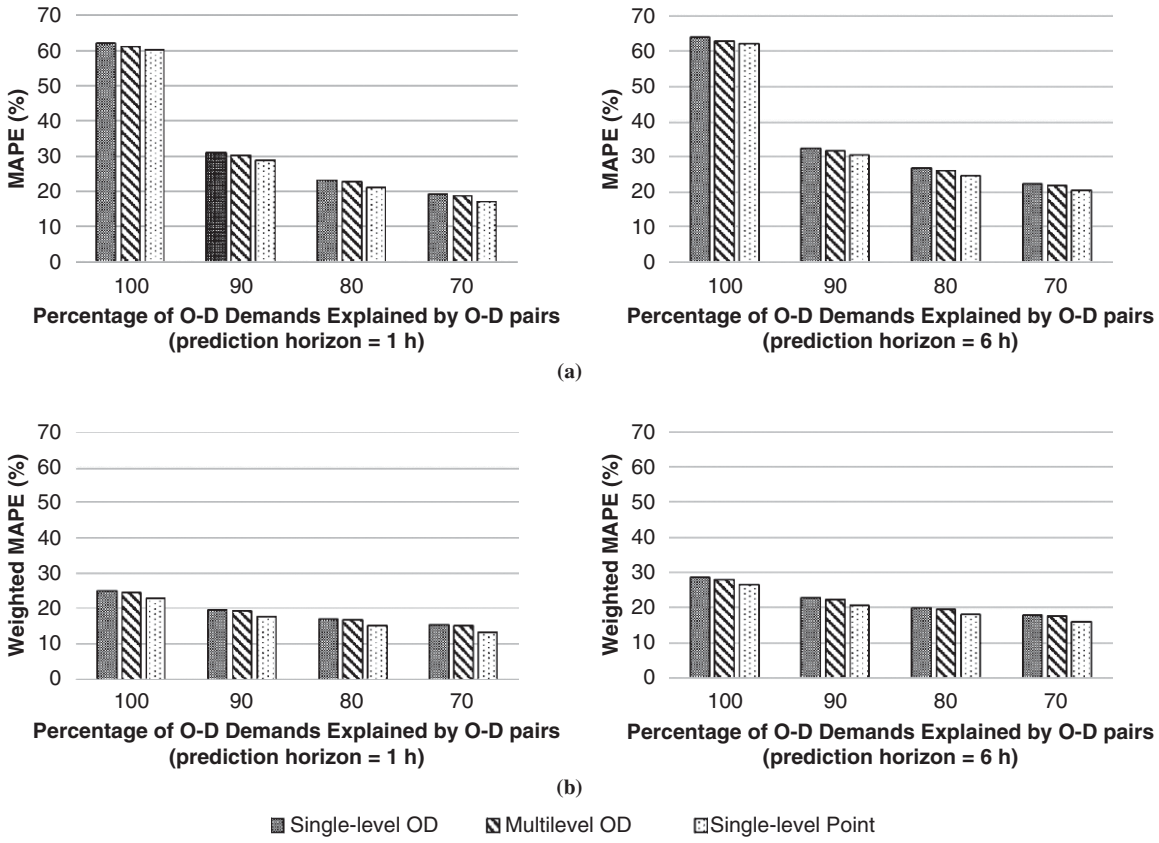

FIGURE 3 Comparison of MAPE and weighted MAPE for each prediction horizon: (*a*) mean MAPE values and (*b*) mean weighted MAPE values.

70% of O-D demand represents large-scale O-D pairs that make up 70% of the total demand. On average, O-D pairs had demand counts greater than 0, 3.14, 8.65, and 18.09 for the top 100%, 90%, 80%, and 70% of O-D demand, respectively.

As shown with mean MAPE values in Figure 3*a*, single-level point demand performs generally better than do the other two prediction strategies. This strategy especially performs better than the other two for O-D pairs with larger demand—that is, moving from the top 100% of O-D demand to 70%—and for a longer prediction horizon—that is, a 1-h prediction horizon rather than a 6-h horizon. This is shown in Table 3, where the mean MAPE difference of single-point demand (upper part) compared with the other two methods is increasing for both size of demand and prediction horizon. When calculated with a weighted MAPE value, the results are similar; however, the weighted MAPE value seems to be a more reliable accuracy measure than the MAPE value, as observed in Table 3 (lower part), where the different percentages of O-D demand show little difference in the weighted MAPE values. This finding is because the weighted MAPE value considers the size of the demand and shows less fluctuation in accuracy values as demand size gets larger.

Time Series of Individual O-D Pair

The slight difference in accuracy between the foregoing strategies makes it difficult to state which strategy is the best, especially when other studies do not have true O-D demand data and have evaluated accuracy in terms of estimated or indirect values. Therefore, the accuracy is evaluated further with time series plots of actual and predicted demand of an individual O-D pair. Figure 4 shows actual values and prediction values of O-D demand from Seoul, South Korea, to two different destinations, Kiheung and Icheon, South Korea. The O-D demand to Kiheung is relatively larger compared with that to Icheon. Two examples are shown, one at a normal commuting time of Monday at 17:00 (Figure 4, *a* and *c*), and another on a noncommuting time of Sunday at 9:00 (Figure 4, *b* and *d*).

First, for the normal commuting time in Figure 4, *a* and *c*, all three prediction strategies perform very well and follow the time series trend of the two O-D pairs. This finding is because the normal commuting pattern is repeated daily with small variance and is easily found in the historical database. Many studies of O-D prediction targeted their experiments at commuting hours; however, the clear repetitive pattern of commuting actually makes the prediction performance much better than irregular times, such as a weekend, with a data-driven method. For the prediction of the irregular traffic pattern in Figure 4, *b* and *d*, the single-level O-D demand and point demand perform well and the multilevel O-D demand has slightly more error than the other two methods. Both the single-level O-D demand and the point demand capture the unusual increase of demand for both O-D pair examples, which are difficult to predict with conventional time series models.

Computation Efficiency of Prediction

As discussed in the framework, the multilevel O-D demand and single-level point demand were developed to reduce the computation

TABLE 3 Differences in Accuracy Values

Prediction Horizon	Value Description	Value Difference by Percentage of O-D Demand Explained by O-D Pairs			
		100	90	80	70
MAPE Value Difference					
1 h	$\overline{MAPE^{Single\ Point}} - \overline{MAPE^{Single\ O\text{-}D}}$	–2.72	–4.21	–4.66	–4.83
	$\overline{MAPE^{Single\ Point}} - \overline{MAPE^{Multi\ O\text{-}D}}$	–1.11	–2.28	–2.63	–2.86
6 h	$\overline{MAPE^{Single\ Point}} - \overline{MAPE^{Single\ O\text{-}D}}$	–3.02	–4.76	–5.54	–5.66
	$\overline{MAPE^{Single\ Point}} - MAPE^{Multi\ O\text{-}D}$	–1.35	–2.33	–2.84	–2.91
Weighted MAPE Value Difference					
1 h	$\overline{wMAPE^{Single\ Point}} - \overline{wMAPE^{Single\ O\text{-}D}}$	–4.46	–4.71	–4.76	–4.73
	$\overline{wMAPE^{Single\ Point}} - \overline{wMAPE^{Multi\ O\text{-}D}}$	–2.54	–2.74	–2.80	–2.83
6 h	$\overline{wMAPE^{Single\ Point}} - \overline{wMAPE^{Single\ O\text{-}D}}$	–5.32	–5.72	–5.78	–5.76
	$\overline{wMAPE^{Single\ Point}} - MAPE^{Multi\ O\text{-}D}$	–2.85	–3.11	–3.16	–3.19

NOTE: *w*MAPE = weighted MAPE.

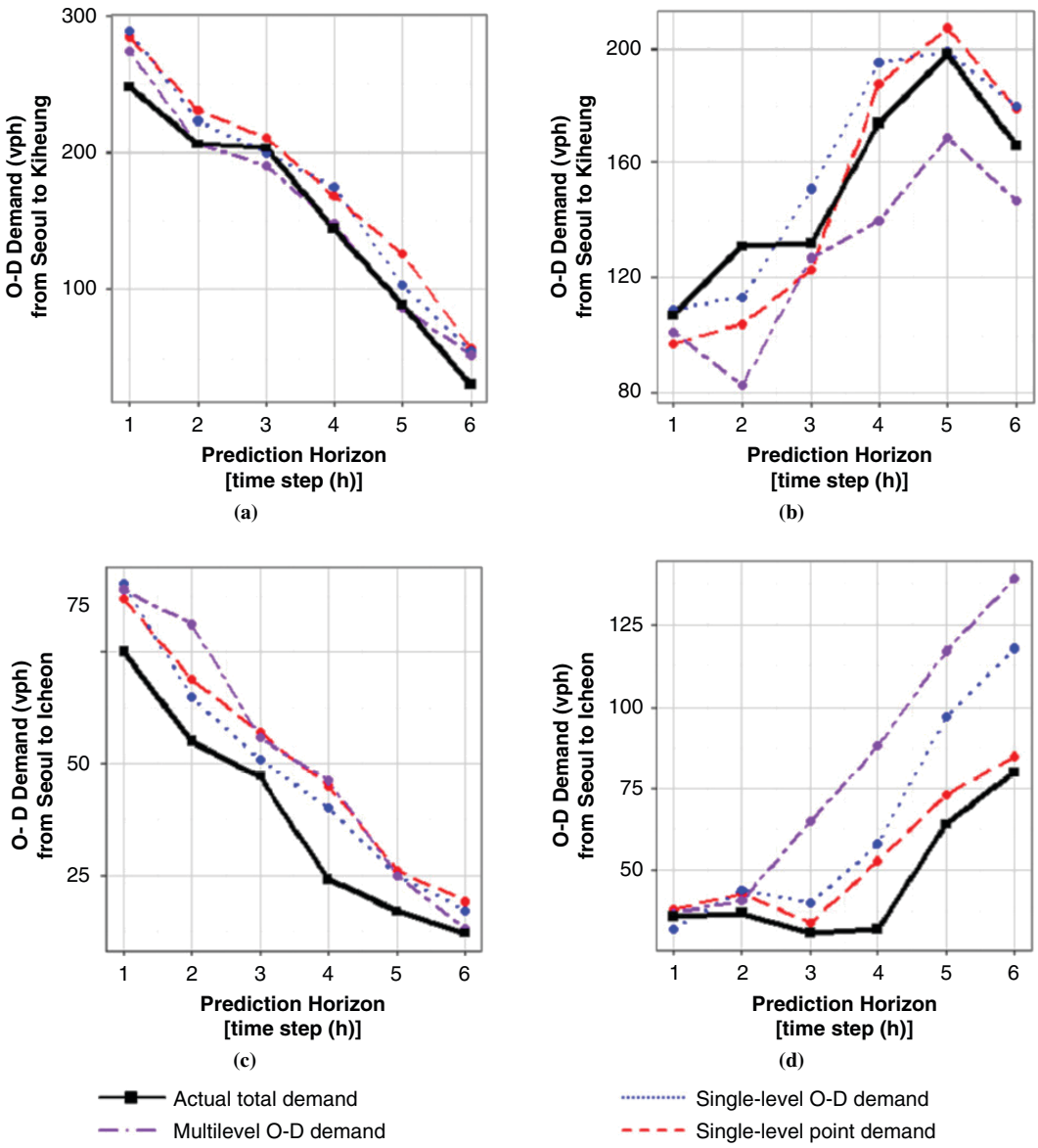

FIGURE 4 Comparison of accuracy on time series for individual O-D pair: (*a*) commuting hours, Seoul to Kiheung; (*b*) noncommuting hours, Seoul to Kiheung; (*c*) commuting hours, Seoul to Icheon; and (*d*) noncommuting hours, Seoul to Icheon (vph = vehicles per hour).

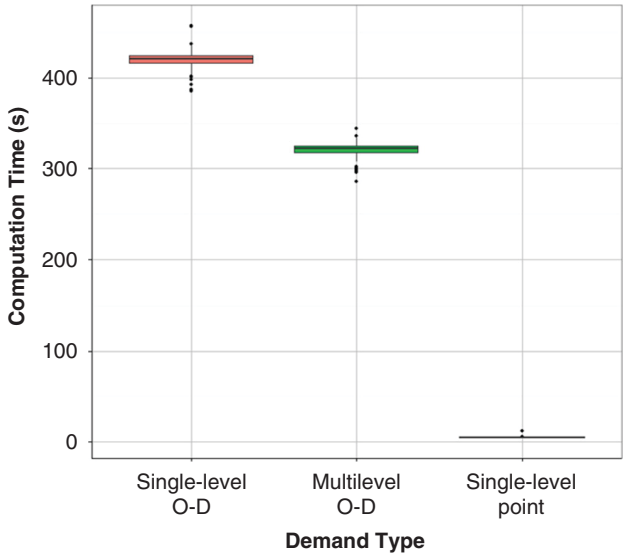

FIGURE 5 Comparison of computation time for three strategies.

burden of single-level O-D demand, which has a large dimensional feature vector to search for all the historical dates. The computation time for the entire prediction process is shown in Figure 5. The mean computation times for single-level O-D demand, multilevel O-D demand, and single-level point demand were 419.49, 320.90, and 5.16 s, respectively.

The single-level point demand performed in a matter of seconds, much faster than the two other strategies. This superior efficiency is crucial in determining the prediction strategy for real-time service, especially a meaningful result for the targeted driving population, whose free flow is more than 80 km/h on the highway. Moreover, the entire traffic management system will include other time-consuming modules, such as data transmission and the core simulation itself, which requires faster O-D demand prediction.

From Equation 3, it is understandable why single-level O-D demand takes a very long time. One feature vector has entries of $(\tau + 1)R^2$, which have to be compared for every historical data set. Especially when there is an increased number of digits for each demand entry, such as weekends or the peak holiday season, the prediction time increases greatly. For instance, the outliers of the largest computation time belonged mostly to the end of July, which in South Korea is the start of the summer holiday.

One possible reason why multilevel O-D demand has little improvement from the single-level O-D demand is the aggregating process of individual O-D demand to subnetwork O-D demand in the first matching step. For instance, on July 7, 2014, at 9:00, the computation time for the first matching step was 290.07 s with a total of $(\tau + 1)(R^{agg})^2$, whereas the second matching step with individual O-D demand took 44.34 s with entries of $(\tau + 1)R^2$. Even with historical data already converted as subnetwork O-D demand in storage, converting the real-time individual O-D demand into network O-D demand involves heavy computation. As well, the outliers of computing the multilevel prediction were also similar and mostly in the peak holiday season.

Compared with the prediction strategies using O-D demand data, the single-level point demand data performed much faster. This prediction has a feature vector with many dimensions of $2R$ and does not require an aggregation process since it is directly acquired. Therefore, considering the accuracy results with the other two strategies, the strong computation efficiency of single-level point demand makes it more advantageous for predicting O-D demand of a large O-D network for real-time service.

Sensitivity to Historical Database Size

In addition, a sensitivity study on the historical database size is briefly done, as shown in Figure 6. The effect of historical database size is

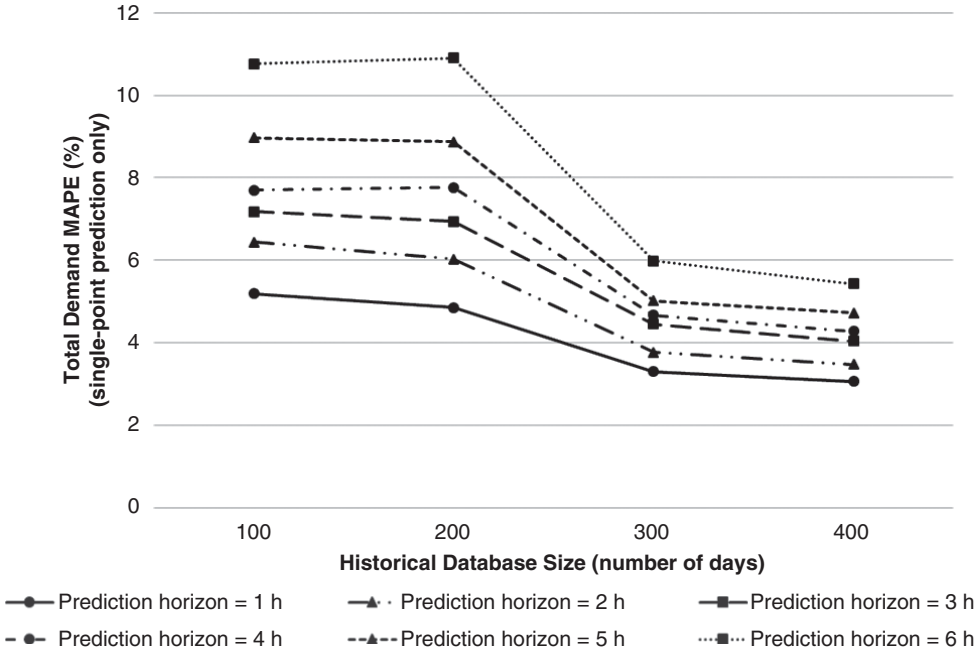

FIGURE 6 Sensitivity of accuracy with historical database size.

similar for all prediction methods; therefore, only the case of single-level point demand prediction is shown with total demand MAPE for differing prediction horizons. The historical date periods were January 6 to April 15, 2014; September 28, 2013, to April 15, 2014; March 11, 2013, to April 15, 2014; and January 7, 2013, to April 15, 2014, for 100, 200, 300, and 400 dates, respectively. As shown, the accuracy increases as historical size increases, regardless of prediction horizon, and there was more improvement in the interval from 200 to 300 days than in other intervals. Especially for prediction further in the future, the improvement was even larger as seen by a more significant drop of the MAPE. The graph shows that a historical database size of at least 300 is required for a data-driven prediction of O-D demand.

CONCLUSION

A prediction method is proposed for O-D demand of a large network for real-time service in a data-driven approach. With the large data reserve of direct O-D demand measurements in the South Korean highway network, a *k*-NN pattern-matching prediction was developed that searched for a *k*-set of historical dates that potentially had a similar traffic pattern as the subject time and provided possible values for the future O-D demand for prediction. This method overcomes several weaknesses of the current literature with time series prediction techniques, such as lack of efficiency in prediction of unusual O-D demand and real-time O-D demand prediction of a large network.

Along with the framework for real-time prediction of O-D demand, three strategies of *k*-NN prediction were developed to experiment with different feature vectors. The single-level O-D demand strategy uses a feature vector with O-D demand of all possible O-D pairs in the network for one-step pattern matching. The multilevel O-D demand strategy adopts a divide-and-conquer concept and takes two matching steps for prediction: primary searching with aggregated O-D demand in a low dimension and secondary searching with original O-D demand in a high dimension. Last, the single-level point demand uses a feature vector of only the entrance and exit demand, which reduces the dimension and computation load.

The three strategies were compared on hourly prediction of O-D demand on South Korean expressways, with 119,025 O-D pairs. Single-level point demand prediction performed slightly better in terms of total demand as well as for individual O-D demand with increasing demand size and prediction horizon. Moreover, the computation time for the single-level point demand strategy was very efficient with an average of 5.16 s, whereas single-level O-D demand and multilevel O-D demand took 419.49 s and 320.90 s, respectively. Therefore, the single-level point demand strategy of the *k*-NN pattern-matching method is recommended for the prediction of the O-D demand of a large network for real-time service. In addition, a sensitivity study of the historical database size on accuracy shows that a historical database of at least 300 dates is required.

For future studies, it is necessary to evaluate the performance of this prediction framework with O-D data from South Korean toll gates. There are many potential technologies that may be manipulated to sample or measure O-D demand indirectly, for instance, mobile GPS data of those who use navigation services and data obtained through dedicated short-range communication. The appropriate data processing and matching strategies can then be evaluated for specific data sets.

ACKNOWLEDGMENTS

This research was supported by a Microsoft Research Award in Urban Informatics and by the U-City Master and Doctor Course Grant Program of the South Korean Ministry of Land, Infrastructure, and Transport.

REFERENCES

1. Oh, S., Y.-J. Byon, K. Jang, and H. Yeo. Short-term Travel-time Prediction on Highway: A Review of the Data-driven Approach. *Transport Reviews,* Vol. 35, No. 1, 2015, pp. 4–32.
2. Tak, S., S. Kim, K. Jang, and H. Yeo. Real-Time Travel Time Prediction Using Multi-Level *k*–Nearest Neighbor Algorithm and Data Fusion Method. *Computing in Civil and Building Engineering,* 2014, pp. 1861–1868.
3. Zhou, X., and H. S. Mahmassani. A Structural State Space Model for Real-Time Traffic Origin-Destination Demand Estimation and Prediction in a Day-to-Day Learning Framework. *Transportation Research Part B: Methodological,* Vol. 41, 2007, pp. 823–840.
4. Camus, R., G. E. Cantarella, and D. Inaudi. Real-Time Estimation and Prediction of Origin-Destination Matrices per Time Slice. *International Journal of Forecasting,* Vol. 13, 1997, pp. 13–19.
5. Antoniou, C., M. E. Ben-Akiva, and H. N. Koutsopoulos. Dynamic Traffic Demand Prediction Using Conventional and Emerging Data Sources. *IEEE Proceedings on Intelligent Transport Systems,* Vol. 153, No. 3, 2006, pp. 199–212.
6. Ashok, K., and M. E. Ben-Akiva. Estimation and Prediction of Time-Dependent Origin-Destination Flows with a Stochastic Mapping to Path Flows and Link Flows. *Transportation Science,* Vol. 36, No. 2, 2002, pp. 184–198.
7. Ashok, K., and M. E. Ben-Akiva. Alternative Approaches for Real-Time Estimation and Prediction of Time-Dependent Origin-Destination Flows. *Transportation Science,* Vol. 34, No. 1, 2000, pp. 21–36.
8. Ashok, K. *Estimation and Prediction of Time-Dependent Origin-Destination Flows.* PhD dissertation. Massachusetts Institute of Technology, Cambridge, 1996.
9. Bierlaire, M., and F. Crittin. An Efficient Algorithm for Real-Time Estimation and Prediction of Dynamic O-D Tables. *Operations Research,* Vol. 52, No. 1, 2004, pp. 116–127.
10. Djukic, T., J. W. C. van Lint, and S. P. Hoogendoorn. Application of Principal Component Analysis to Predict Dynamic Origin–Destination Matrices. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2283,* Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 81–89.
11. Djukic, T. *Dynamic O-D Demand Estimation and Prediction for Dynamic Traffic Management.* PhD dissertation. Delft University of Technology, Netherlands, 2014.
12. Barceló, J., and L. Montero. A Robust Framework for the Estimation of Dynamic O-D Trip Matrices for Reliable Traffic Management. *Transportation Research Procedia,* Vol. 10, 2015, pp. 134–144.
13. Cantelmo, G., F. Viti, C. Tampère, E. Cipriani, and M. Nigro. Two-Step Approach for Correction of Seed Matrix in Dynamic Demand Estimation. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2466,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 125–133.
14. Antoniou, C., J. Barceló, M. Breen, M. Bullejos, J. Casas, E. Cipriani, B. Ciuffo, T. Djukic, S. Hoogendoorn, V. Marzano, L. Montero, M. Nigro, J. Perarnau, V. Punzo, T. Toledo, and H. van Lint. Towards a Generic Benchmarking Platform for Origin–Destination Flows Estimation/Updating Algorithms: Design, Demonstration and Validation. *Transportation Research Part C: Emerging Technologies,* Vol. 66, 2016, pp. 79–98.
15. Frederix, R., F. Viti, and C. Tampère. A Hierarchical Approach for Dynamic Origin-Destination Matrix Estimation on Large-Scale Congested Networks. In *Proceedings of 14th International IEEE Conference on Intelligent Transportation Systems (ITSC),* IEEE, New York, 2011, pp. 1543–1548.
16. Ashok, K., and M. E. Ben-Akiva. Alternative Approaches for Real-Time Estimation and Prediction of Time-Dependent Origin–Destination Flows. *Transportation Science,* Vol. 34, No. 1, 2000, pp. 21–36.
17. *Traffic Statistics.* Korea Expressway Corporation. http://www.ex.co.kr. Accessed June 14, 2015.

The Standing Committee on Transportation Network Modeling peer-reviewed this paper.