2015 IEEE International Conference on Data Science and Data Intensive Systems

A Prediction Model of Traffic Congestion Using Weather Data

Jiwan Lee*, Bonghee Hong*, Kyungmin Lee* and Yang-Ja Jang†
*Department of Electrical and Computer Engineering, Pusan National University, Busan, South Korea
Email: {wldhks85, bhhong, ykm4223}@pusan.ac.kr
†Big Data Processing Platform Research Center, Pusan National University, Busan, South Korea
Email: yjjang@pusan.ac.kr

***Abstract*—Weather factors such as temperature and rainfall in residential areas and tourist destinations affect traffic flow on the surrounding roads. In this study, we attempt to find new knowledge between traffic congestion and weather by using big data processing technology. Changes in traffic congestion due to the weather are evaluated by using multiple linear regression analysis to create a prediction model and forecast traffic congestion on a daily basis. For the regression analysis, we use 48 weather forecasting factors and six dummy variables to express the days of the week. The final multiple linear regression model is then proposed based on the three analytical steps of (i) the creation of the full regression model, (ii) the removal of the variables, and (iii) residual analysis. We find that the R-squared value of the proposed model has an explanatory power of 0.6555. To verify its predictability, the proposed model then evaluates traffic congestion in July and August 2014 by comparing predicted traffic congestion with actual traffic congestion. By using the mean absolute percentage error valuation method, we show that the final multiple linear regression model has a prediction accuracy of 84.8%.**

***Keywords*—Prediction Model, Multiple Linear Regression Analysis (MLRA) Model, Traffic Big data**

I. Introduction

Intelligent transport systems collect traffic data such as traffic volume and speed on every roads and provide statistical summary services, usually on traffic congestion. Traffic congestion is the state in which the increase in the number of vehicles delays the traffic flow. In this regard, knowledge of change in the traffic congestion due to weather is an essential element for decision making when individuals consider using public transportation or planning a long trip. The introduction and development of the intelligent transport system in Korea has resulted in more reliable traffic information gathering, analyzing, and processing, thereby providing more time-relevant and precise traffic analytics and prediction to users. However, while Korea's intelligent transport systems have provided accurate traffic information, they have not shown the services to predict traffic congestion dependent upon forecasted weather.

Studies have used various approaches to examine whether traffic indices depend on the temperature and rainfall in summer and would be able to forecast traffic speed. In [1], the extent to which speed delays were caused by rainfall is analyzed and an artificial neural network used to estimate traffic speed by level of rainfall. The [2] analyzes the influence of the prevailing weather conditions on daily traffic volume. However, such analyses of weather conditions and traffic changes are based on points of observation. The analysis of observed data is to evaluate the influence of a certain weather condition over traffic congestion, speed, and volume. Instead of a specific point, we'd like to analyze and predict traffic information with a set of observed traffic data over forecasted weather. In order to predict future traffic congestion on different roads or highways, we need to develop a new different prediction model by using a large number of collected traffic data with designated nodes and links.

Different roads require to have different prediction models because of different traffic characteristics. Of course, different regions require to have different prediction models. Development of different prediction model require huge traffic data processing.

In this study, we propose a big data processing system to create a prediction model of traffic congestion over specific roads and regions. The system is composed of the Hadoop and R open source codes. In the Hadoop part, big traffic data are processed to produce daily traffic congestion of tourist destination by Mapreduce programing model. In the R part, we performed to generate predictive model and prediction results using statistics library. In particular, we suggest a predictive model that uses multiple linear regression (MLR) analysis utilizing both weather forecast data in residential areas and tourist destinations and traffic congestion data in tourist destinations. Seoul is selected as the residential area, and Ocean Beach (Korean name: Gyeongpodae), which is close to Seoul about 200km, as the tourist destination. The time span of the data used in the analysis is July and August 2013-2014.

MLR analysis was performed in three steps, namely (i) creating a full regression model, (ii) removing the variables, and (iii) performing residual analysis. Firstly, a full MLR model that included 54 variables was constructed. Secondly, by using the variable removal method on this full MLR model, the variables that lacked influence, those with higher levels of correlations between the independent variables, and those for which the statistical significance was above a certain level were removed. Thirdly, residual analysis was performed to verify the normality and evenness of the distribution in order to develop the final model.

The final MLR model was created by using 48 weather forecast factors and the days of the week as the 6 dummy variables. By using this model, the traffic congestion on the roads near Ocean Beach in July and August 2014 were predicted. The evaluation of the mean absolute percentage error

978-1-5090-0214-6/15 $31.00 © 2015 IEEE
DOI 10.1109/DSDIS.2015.96

(MAPE) of the actual traffic congestion value and the estimated value showed that it was 0.06 in 2013 and 0.152 in 2014. By using the data from 2013, the 2014 forecast was then presented, showing an error rate of 15.2%. The estimation accuracy of the proposed model was thus 84.8%, which confirms its reliability.

The remainder of this paper is organized as follows. In Section 2, the method of converting the traffic speed data into traffic congestion data is introduced. In Section 3, the approach to creating the MLR model using the weather forecast data and traffic congestion data is explained. In Section 4, we describe how the prediction model was generated by using Hadoop and R open source. Section 5 presents the prediction results using the regression model created in Section 3. In Section 6, we introduce the existing literature on the analysis of weather and traffic information. Lastly, in Section 7, the conclusion of the paper is presented and future work is discussed.

II. Definition of Traffic Congestion

Traffic congestion[3] is a condition on road networks that is characterized by slower speeds, longer trip times. Information on traffic congestion is required by vehicle drivers, traffic police, and road construction planners. Drivers want to know the status of the road to their destination to avoid traffic congestion. Traffic police demand information on intersections to reduce traffic jams through improvement of signal system. The road construction planners use traffic information as one of the factors for choosing locations where road expansion is required.

Speed limit: 60 km/h

	$link_1$: 2km	$link_2$: 4km	$link_3$: 8km
t_1	15km/h	30km/h	45km/h
t_2	30km/h	45km/h	30km/h
t_3	15km/h	30km/h	45km/h
t_4			

Fig. 1: Example of traffic data

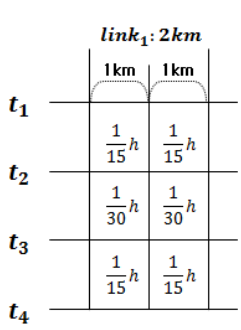
We describe two issues with the aggregation of traffic congestion and introduce a formula for computing the congestion score for spatial and temporal aggregation [4]. To compute traffic congestion, the key factors are speed, travel time, and traffic volume [5]. This study considers only travel speed on a link because of the difficulty collecting traffic volume owing to its high cost and because vehicle drivers recognize traffic congestion by the decreasing travel speed. Moreover, traffic congestion can be measured qualitatively and quantitatively. We use quantitative information to predict traffic congestion in this study. The formula for calculating traffic congestion is as follows:

$$S_{tc}(LV_i, RS_i) = \begin{cases} 100\%, & \text{if } LV_i \leq 0 \\ 0\%, & \text{if } LV_i \geq RS_i \\ (1 - \frac{LV_i}{RS_i}) \times 100\%, & \text{otherwise} \end{cases} \quad (1)$$

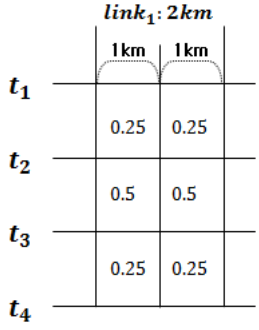
where i is a specific link identifier, LV_i is the velocity of $link_i$, and RS_i is the speed limit of $link_i$.

The traffic congestion score (TCS) is calculated as a percentage by using function $S_{tc}()$, which ranges from 0% to 100% and consists of three parts: the velocity of the link is less than 0 km/h, the velocity of the link is over a reference speed, and the velocity of the link is over 0 km/h but less than the reference speed.

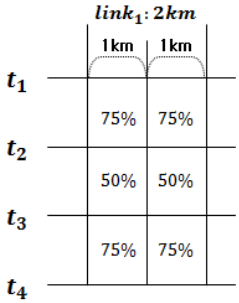
We compute the TCS by using the temporal aggregation method to obtain daily traffic congestion. Temporal aggregation is used to calculate the TCS of a specific link for the given time intervals. To calculate the TCS, we perform three steps: calculation of the time required to travel 1 km (the travel time unit, or TTU hereafter), calculation of the approximation ratio of TTU (ATTU), and aggregation. We explain each step in Fig. 2. Given the sample data in Fig. 1, we calculate TTU by using the velocity of link 1 as follows (see Fig. 2(a)):



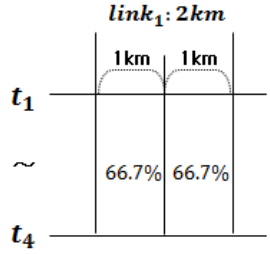
(a) Results of TTU calculation



(b) Results of ATTU calculation



(c) Results of the TSC



(d) Results of temporal aggregation

Fig. 2: Example of calculating temporal aggregation

$$TTU = \frac{d}{v} \quad (2)$$

where v is velocity, d is distance.

$ATTU$ is then calculated as follows (see Fig. 2(b)):

$$ATTU(t_1, t_2) = \frac{TTU_{rs}}{TTU_v} \quad (3)$$

where TTU_v is TTU based on the current velocity on the link and TTU_{rs} is TTU based on the speed limit.

According to Eq. (3), $ATTU$ is 1/60h 1/15/h = 0.25 given the time interval from T_1 to T_2.

The temporal aggregation of the TCS is calculated by the summation of each time of the TCS as follows (see Fig. 2(c)):

$$(1 - ATTU_{t_1,t_2}) \times 100\% \tag{4}$$

The temporal aggregation equation is as follows (see Fig. 2(d)):

$$\{1 - \sum_{t_n}^{t_{m-1}} \frac{ATTU_{t_n,t_n+1}}{d(t_n,t_m)}\} \times 100\% \tag{5}$$

where n is the start time of aggregation, m is the end time of aggregation, and T_i is the interval of collected time for the data.

III. Multiple Linear Regression Analysis Using Weather Data

In this section, the MLR analysis for predicting the changes in traffic congestion based on weather data is presented. The data used in the analysis include weather data in the residential area and the destination as well as traffic congestion data in the destination. The data and process of developing the MLR model are explained in the following subsections.

A. The Real Data

1) Weather forecast data: The Korean Meteorological Administration provides 3-hour, 12-hour, and 24-hour forecasts on 12 weather factors such as temperature, humidity, cloud, rainfall, and wind velocity. Since this study aims to develop a model that forecasts daily traffic congestion, these weather forecast data were converted into daily frequencies and provided in 5 km $\times$ 5 km cells. The residential areas used in the analysis were Seoul, Destination- Ocean Beach (Gyeongpodae). Fig. 3 shows these locations in Seoul and the corresponding weather forecast data used.

2) Traffic data: Data from SK Planet, which are collected in every hour and cover the traffic speed on each link, were used for the analysis. The link is a segment of road. As shown Fig 3(c), target road has included a lot of links. The traffic data consist of specific link id, time, and speed. As mentioned in Section 2, speed data were converted into the TCS (range 0% to 100%). As in the case with weather data, the traffic congestion point was also converted into daily averages. The roads used in the analysis were selected from areas near the Ocean Beach.

B. MLR Analysis Model

Altogether, 48 independent variables were used in the regression analysis, including temperature, humidity, cloud, rainfall, and wind velocity. The days of the week were additionally used as 6 dummy variables. Regression models were not created for each day of the week. Instead, a single model for the entire week was created. By using this model, we found that the R-squared level of the independent weather variables to explain traffic congestion on nearby roads was 0.941. If we use only the explanatory power for the estimation, the weather forecast in both the residential and the destination areas seems to affect traffic congestion in the destination area significantly. However, in a regression analysis, a larger number of independent variables may contribute to increased explanatory power. Therefore, it is necessary to use the variable removal method in order to leave the optimal variables only.

C. Variable Removal Method

In a regression analysis, after first including all the variables, the variables that are not thought to be important are removed as the regression model is developed. This practice is necessary for preventing an increase in the R-squared value because of the large number of independent variables and for removing the independent variables that are highly correlated. This method comprises three steps.

1) Step1.Backward: This method simplifies the model by removing the unnecessary variables one by one. After making the regression model that includes all the independent variables fit using the least squares method, the regression coefficient of the independent variable that has the extra sum of square is verified. The extra sum of square means the effect of the independent variable when it is included in the model for the last time. If the result of the verification is not significant, the corresponding independent variable is removed from the model. Then, a new regression model is made to fit. This process is repeated until the regression coefficient becomes statistically significant, meaning that only the independent variables that are relevant to the model remain.

As a result, 18 variables were removed from the regression model that originally included all weather forecast information for Seoul and Ocean Beach. Table I shows the results. The three key variables were temperature at Ocean Beach, humidity in Seoul's forecast, and rainfall. As the number of independent variables was still too large, however, it was necessary to remove more variables by diagnosing their multicollinearity.

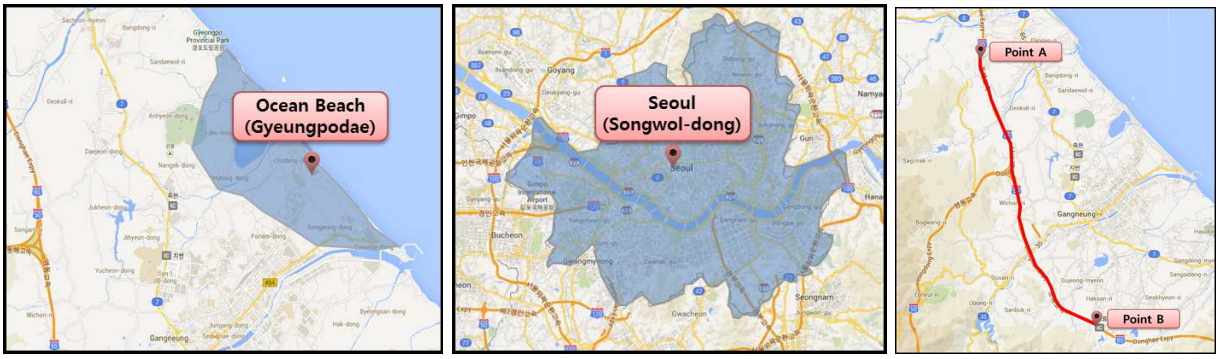
2) Step2.Multicollinearity: In regression analysis, multicollinearity exists as a level of the linear relationship between the independent variables and therefore must be addressed. If the R-squared value is too high, the explanatory power of the regression function is high and the p-values of the independent variables are high, meaning that the individual factors are no longer significant. In such a case, a high level of correlation between the independent variables can be suspected.

One method of diagnosing multicollinearity is by calculating the variance inflation factor (VIF) [6] using Eq. (6). Normally, multicollinearity is thought to be problematic when the VIF value exceeds 10:

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}, \tag{6}$$

where R_i^2 is the coefficient of determination of the regression equation.

In our study, the VIF was calculated to diagnose multicollinearity between the independent variables in the regression model. Although the VIF value is normally set to 10, to achieve a stronger multicollinearity diagnostic performance, the number was set to 4 or more. The outcome of the diagnosis on multicollinearity was to remove 18 more variables from the results of the backward variable removal method. The number



(a) Ocean Beach (Gyeongpodae) (b) Seoul (c) Target road

Fig. 3: Study areas

TABLE I: Results of the MLR model using the backward method

Variables	coef.	p-value	Variables	coef.	p-value	Variables	coef.	p-value
Mon	-0.165	0.105	G_Max_T	0.147	0.045	S_Avg_H_PD	0.011	0.469
Tue	-0.221	0.110	G_Avg_T	-0.149	0.176	S_Avg_WV_PD	0.052	0.509
Wed	-0.277	0.038	G_Max_H	-0.003	0.624	S_L_Cloudy_PD	0.089	0.637
Thu	-0.142	0.132	G_Min_H	0.001	0.887	S_Cloudy_PD	0.079	0.484
Fri	-0.096	0.353	G_Clear	0.376	0.075	S_RF_P_PD	0.005	0.207
G_Max_T_PD	-0.135	0.142	G_L_Cloudy	0.334	0.043	S_Daily_RF_PD	-0.001	0.156
G_Avg_T_PD	0.156	0.191	G_Cloudy	0.103	0.253	S_Max_T	-0.046	0.559
G_Max_H_PD	0.012	0.234	G_RF_P	0.010	0.043	S_Min_T	0.095	0.297
G_Min_H_PD	0.016	0.070	S_Max_T_PD	0.118	0.173	S_Avg_H	0.020	0.381
G_Avg_H_PD	-0.018	0.304	S_Avg_T_PD	-0.120	0.490	S_L_Cloudy	0.158	0.332
G_RF_P_PD	-0.006	0.192	S_Max_H_PD	-0.013	0.222	S_Daily_RF	-0.001	0.235
G_Daily_RF_PD	0.005	0.340	S_Min_H_PD	-0.005	0.619			
R-Squared	0.9293							

TABLE II: Results of the MLR model using multicollinearity

Variables	coef.	p-value	Variables	coef.	p-value	Variables	coef.	p-value
Mon	-0.165	0.105	G_Min_H	0.001	0.887	S_RF_P_PD	0.005	0.207
Tue	-0.221	0.110	G_Clear	0.376	0.075	S_Daily_RF_PD	-0.001	0.156
Wed	-0.277	0.038	G_Cloudy	0.103	0.253	S_Min_T	0.095	0.297
Thu	-0.142	0.132	S_Max_H_PD	-0.013	0.222	S_L_Cloudy	0.158	0.332
Fri	-0.096	0.353	S_Avg_WV_PD	0.052	0.509	S_Daily_RF	-0.001	0.235
G_Daily_RF_PD	0.005	0.340	S_Cloudy_PD	0.079	0.484			
R-Squared	0.6979							

of variables decreased significantly in the diagnosis because the independent variables of the weather forecast overlapped. In the case of a temperature forecast, the temperature variable is composed of three variables, namely the lowest, highest, and average temperature forecast. This shows that the correlation between these three variables may be higher. During the multicollinearity diagnosis, two variables were removed. The results are shown in Table II.

3) Step3.Significant Probability(p-value): After removing the unnecessary variables (Steps 1 and 2), we must determine whether the analysis model is statistically significant by focusing on the p-values of the remaining independent variables. If the p-value is 0.05 or higher, the weather forecast assumed in this study does not affect traffic congestion. The p-value was used to leave only the relevant variables as the independent variables. As a result, eight independent variables were removed. Table III shows the remaining variables.

TABLE III: Final variables selected after the three-step variable removal method

Variables	coef.	p-value
Wed	-0.277	0.033
G_Daily_RF_PD	0.005	0.342
G_Min_H	0.004	0.888
S_Daily_RF_PD	-0.007	0.153
S_Min_T	0.097	0.291
R-Squared	0.6431	

D. Final MLR Analysis Model

This model has an explanatory power of 0.655. The final list of the weather forecast variables included the rainfall forecast of Ocean Beach, forecast of the lowest humidity for Ocean Beach on the previous day, daily rainfall forecast in Seoul, and forecast of the lowest temperature in Seoul on the previous day. Of the selected variables, the forecast of the

lowest temperature in Seoul on the previous day has the highest influence. The traffic congestion around the destination has a positive influence on the lowest temperature in Seoul and a negative influence on rainfall in Seoul. A low temperature in Seoul means that tropical nights have begun in the area. In turn, this means that Seoul residents escape the city for a vacation in Ocean Beach, resulting in heavier traffic. The weather forecast for Ocean Beach (i.e., rainfall and lowest humidity) also affects traffic congestion in a positive manner.

Eq. (7) shows the final MLR model, which selected the weather variables through the removal of the variables from the MLR model:

$$\begin{aligned} Y = {} & 1.0715432 + 0.0327241X_1 + 0.0210769X_2 \\ & -0.0677942X_3 + 0.0182411X_4 + 0.0419038X_5 \\ & +0.0501668X_6 + 0.0063967X_7 + 0.0018295X_8 \\ & -0.0019324X_9 + 0.0581072X_{10} \end{aligned} \tag{7}$$

where the variables are denoted by Table IV

TABLE IV: Description of the variables in predicted model

X_1	Monday
X_2	Tuesday
X_3	Wednesday
X_4	Thursday
X_5	Friday
X_6	Saturday
X_7	Daily rainfall in Ocean Beach
X_8	Lowest hum for Ocean Beach on the previous day
X_9	Daily rainfall in Seoul
X_{10}	Lowest temp in Seoul on the previous day

To determine whether the MLR analysis method is suitable, normality and homoscedasticity were verified. The data used for this analysis were daily data from July and August 2013. Since the number of samples was small, the Shapiro-Wilk method was used to assess normality. As shown in Table V, the p-value is 0.05 or higher, suggesting that the distribution was normal. Homoscedasticity was verified by using the Breusch-Pagan method. As the p-value was 0.05 or higher, the homoscedasticity requirement was also met. The verification of normality and homoscedasticity therefore showed that the MLR analysis is appropriate.

TABLE V: Verification of homoscedasticity and normality

Test / Verification	Shapiro-Wilk	Breusch-Pagan
p-value	0.9548	0.8293

IV. Hybrid Structure of the Prediction Model Generating System

A. Processing the Big Traffic Data based on Hadoop

Because traffic data are very large-scale and accumulated continuously by vehicle drivers, it is necessary to handle a large amount of data by using a big data processing platform such as Hadoop in order to provide meaningful information. The Map and Reduce phases of MapReduce are described in detail next.

1) Map Phase: The role of the Map phase is to group the raw traffic data and transfer them to the Reduce phase. The input data of the Map function stored in the HDFS are split into several files by using a splitter. Each split file is then transferred to the Map function. In the Map function, the raw traffic data are composed of a link id, time, and speed. We generate an intermediate result as a pair (key, values). The key is a unique link id that identifies the real road. The values included are time and speed data. Finally, the intermediate result sets transfer to the Reduce phase by combining and shuffling steps.

2) Reduce Phase: The role of the Reduce phase is to convert the raw traffic data into the TCS and store the final result in the HDFS. The input data of the Reduce function are sorted by the link id defined by the key in the Map phase. Here, we have to calculate the TCS by using speed data as mentioned in Section 2. Then, we generate a result as a pair (key, values). The key is same as that in the Map phase. The values included are time and the TCS. After finishing the Map and Reduce phases, the results are input into the R program.

B. Generating the Prediction Model based on R

R is a language and environment for statistical computing and graphics that provides a wide variety of statistical linear and nonlinear modeling. In this study, we focused on the MLR model because of its widespread use. A detailed description of the module used analysis in R is introduced below.

1) Statistical Model Generating Module: In the statistical model generating module, we create an MLR model including all the independent variables (48 variables). The library used in the statistical analysis step is the $lm()$ function. Because the MLR model is unsuitable as it has too many independent variables, we need to remove the unnecessary variables.

2) Variable Removal Module: This module is composed of three methods: backward, multicollinearity, and significance. The backward method is performed by using $step()$ in the R library. The multicollinearity diagnosis method is performed by $vif()$ and the significance test is performed by $summary()$. All the inputs of the variable removal method go into the regression model. We hence remove those variables that have the highest correlation by using these three methods.

3) Residual Analysis Module: It is important to analyze the residuals of the statistical model because if the result of the residual analysis is not satisfied, the statistical model has no value. Tests for homoscedasticity and normality are typical methods of residual analysis. The libraries typically used in residual analysis are the $shapirotest()$ and $bptest()$. After this step, the final MLR model is completed.

C. MLRA Algorithm

The MLR analysis algorithm is composed of two parts: Hadoop and R. First, we describe line by line the algorithm 1 being run in Hadoop. Line 1 initializes the variables for storing a pair (key, values) of the output of the Map function. Then, we have to perform a splitting string to filter the traffic data given the conditions shown from line 3 to line 5. In line 7, the filtered traffic data are stored in the intermediate results to be transferred to the Reduce function. Line 10 initializes the

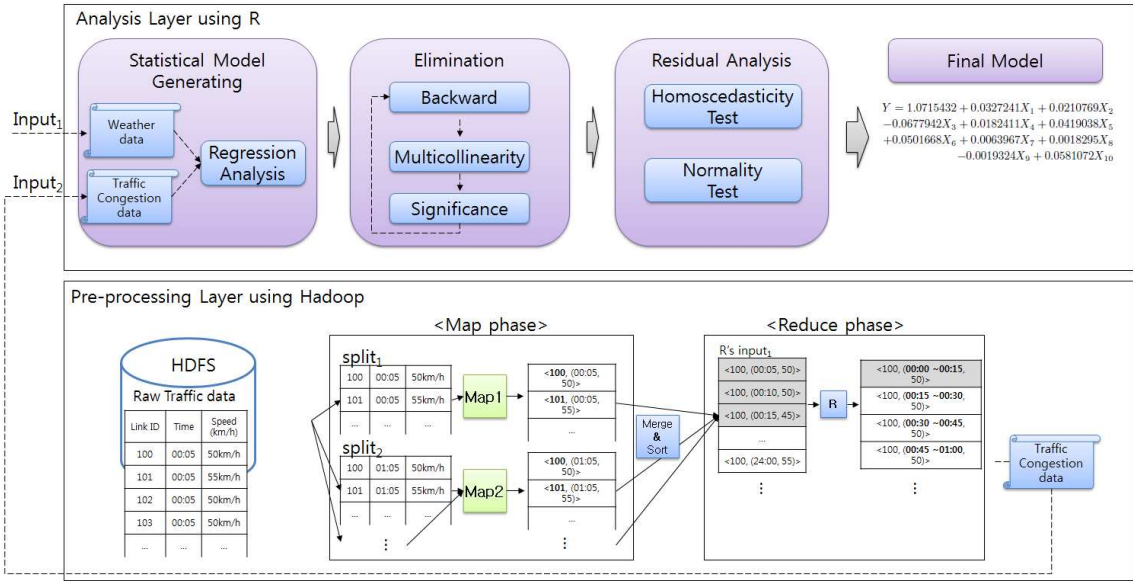


Fig. 4: Structure of the predicted traffic congestion system using weather data

variables for storing the TCS set and lines 12 and 13 are the variables for calculating the TCS of the specific link id. Lines 14, 15, and 16 then calculate the average TCS given the r.key. The final TCS is stored in the HDFS for transferring to the R program in line 18.

Algorithm 1: Big traffic data processing algorithm for calculating TCS

Input: Traffic data set TD
Output: Traffic congestion score TCS

1 Intermediate results $IR \leftarrow \phi$
2 **for** ***each*** *traffic data* $d \in D$ **do**
3 Link id $lid \leftarrow$ string split of d for the key
4 Time $t \leftarrow$ string split of d for the time
5 velocity $v \leftarrow$ string split of d for the speed
6 **if** *lid is included in interest link* **then**
7 $IR \leftarrow$ Emit(lid, $< t, v >$)
8 **end**
9 **end**
10 Traffic congestion score TCS $\leftarrow \phi$
11 **for** ***each*** *Intermediate result ir* $\in$ *IR* **do**
12 Count $c \leftarrow \phi$
13 Sum $s \leftarrow \phi$
14 **for** ***each*** *result* $r \in ir$ **do**
15 $c \leftarrow$ one increments
16 $s \leftarrow s$ + $S_t c$(r)
17 **end**
18 TCS $\leftarrow$ Emit($r.key$, $< r.time, s \div c >$)
19 **end**

In algorithm 2, we introduce how to create an MLR model using the R program. Line 1 creates the full model in step one of the regression analysis by using the $lm()$ function. The $lm()$ function has the TCS and weather data set as the inputs. Lines 3, 4, and 5 remove the unnecessary variables to omit the independent variables that are highly correlated. After that, we perform the residual analysis in line 6. There are two kinds of residual analyses: the shapirotest and the bptest. In the regression analysis, the final MLR model must satisfy the criteria of both tests. Otherwise, the selected regression model is unsuitable as a statistical model. Finally, the final model is selected in line 7.

Algorithm 2: Generating Multiple linear regression model

Input: Traffic congestion score set TCS, Weather data set WD
Output: multiple linear regression model $MLRM$

1 Multiple linear regression model $MLRM_0$
2 $MLRM_0 \leftarrow$ creation of initial model using $lm(TCS, WD)$
3 $MLRM_1 \leftarrow step(MLRM_0, backward)$
4 $MLRM_2 \leftarrow vif(MLRM_1)$
5 $MLRM_3 \leftarrow summary(MLRM_2)$
6 **if** $MLRM_3$ *is satisfied criterion of* $shapiro.test()$ *and* $bptest(MLRM_3)$ **then**
7 $MLRM_3$ is a final multiple linear regression model
Return $MLRM_3$
8 **end**

V. Prediction Result

A. Experimental Data

As discussed in the previous section, the input values used as the weather forecast variables include the rainfall forecast of the Ocean Beach, forecast of the lowest humidity for the Ocean Beach on the previous day, daily rainfall forecast in Seoul, and forecast of the lowest temperature in Seoul on the previous day. To compare the model's outcome with the actual values, actual traffic congestion data from July and August 2014 were used. These traffic congestion data were calculated by using the method introduced in Section 2.

B. Result of Mean Absolute Percentage Error

In this section, we use the MAPE [7] to evaluate the proposed MLR analysis model. The MAPE measures the accuracy of a method of constructing fitted time series values in statistics. It is defined by Eq. (8):

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|, \tag{8}$$

where A_t is the actual value and F_t is the forecast value.

In Fig. 5, the predicted values using the MLR model developed in Section 3 and the actual traffic congestion data from July and August 2013 are compared graphically by using the MAPE value. The MAPE was around 0.059, which converted into a percentage represents an error value of 5.9%. Hence, the final regression model presented here is shown to have an accuracy of 94.1%.

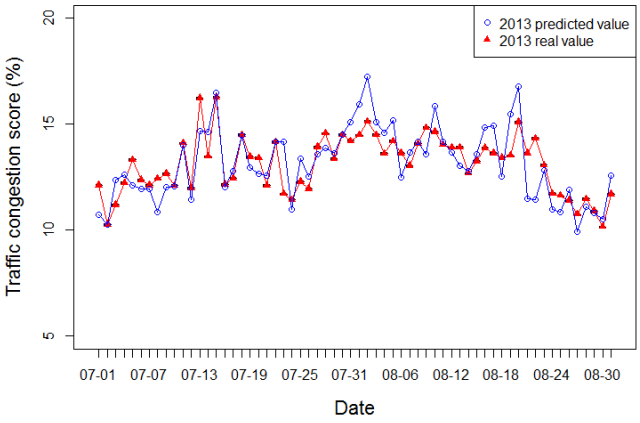


Fig. 5: Comparison of the predicted and actual values of 2013

In Fig. 6, we have applied the MLR model developed in Section3 to predict traffic congestion data from July and August 2014. The predicted values are shown highest error value in the second and fourth weeks of July. In fact, these differences values are a small error with value less than 5%. The MAPE of 2014 was around 0.152, which converted into a percentage represents an error value of 15.2%. Hence, the final regression model presented here is shown to have an accuracy of 84.8%. Because the weather and traffic congestion data from July and August 2013 are used to create predictive model, the accuracy of 2013 is higher than 2014.

In Fig. 7, we have evaluated days of the week of MAPE. The highest MAPE means that high prediction error. The MAPE of Monday is the highest error during the days as around 0.245. The MAPE converted into a percentage represents an error value of 24.5%. The final regression model presented here is shown to have accuracy more than 75.5%. In this experiement, we show that the final regression model proposed in Section 3 is suitable for predicting traffic congestion of weekday and weekend.

VI. Related Work

Falk [1] analyzes static and dynamic tourist demand models to examine whether tourist demand during the summer peak

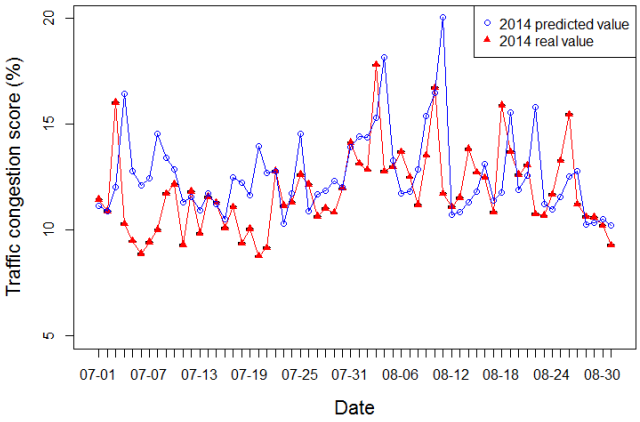


Fig. 6: Comparison of the predicted and actual values of 2014

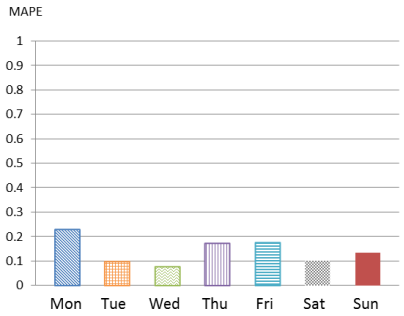


Fig. 7: Day of the week MAPE of July and August 2014

season influences average temperature, the duration of sunshine, and rainfall. In particular, the author evaluates influences on summer nights, focusing on domestic and international tourists, and finds a nonlinear U-shaped relationship. In [2], to develop a traffic accident risk index that depended on weather data, the statistics on traffic accidents are surveyed and areas with higher probabilities of traffic accidents indicated.

By using detector data and the RWIS, Eunbi Jeong [8] classify rainfall grade based on the subsequent speed delay due to this rainfall and analyze speed reduction patterns. By using artificial neural networks, the traffic speed for each level of rainfall is then forecast. By comparing the results with the AI delay speed estimation characteristics, the estimated speed characteristics by rainfall level are analyzed. The rainfall level classification criteria are shown to be 0.4 mm/5 min and 0.8 mm/5 min. The results of the distribution analysis on the speed per rainfall level and traffic volume thus suggest that differences exist based on rainfall level.

Andrey et al. [9] is a comparative study of the accident probabilities in Canada. Six cities (Halifax-Dartmouth, Ottawa, Quebec, Hamilton, Waterloo Region, and Regina) are analyzed by using standardized methods to understand the relationship between the weather in Canadian cities and travel risks. On average, rainfall is shown to have a 75% influence in traffic collisions among all meteorological data; however, snowfall is an even more significant factor.

Sohn Chul [10] analyze how the weather conditions in-

fluence entering Gangreung, Korea. As for the meteorological data, ASOS observation data managed by the Meteorological Administration were used. Daily traffic data were based on tollgate data provided by Korea EX. The result of the regression model estimation shows that during spring, summer, and fall, an increase of 1 mm in rainfall results in a 0.4% to 0.7% reduction in traffic influx. In the case of spring, the influence of temperature on traffic influx into Gangreung is not significant, however. During the summer, the temperature in the capital area rather than in Gangreung turns out to affect traffic influx.

VII. Conclusion

In this study, weather forecast data for Seoul and Ocean Beach and traffic congestion data on the roads surrounding Ocean Beach were used to perform an MLR analysis in order to forecast traffic congestion on a daily basis. The time span of the data used in analysis was July and August 2014 and the spatial range covered Seoul and the Ocean Beach area. Altogether, 48 weather forecast variables were used including the rainfall forecast of Ocean Beach, forecast of the lowest humidity for Ocean Beach on the previous day, daily rainfall forecast in Seoul, and forecast of the lowest temperature in Seoul on the previous day. According to the model results, traffic congestion around the destination has a positive influence on the lowest temperature in Seoul and a negative influence on rainfall in Seoul. The lowest temperature in Seoul means that tropical nights have begun in the area, suggesting that Seoul residents escape the city for a vacation in Ocean Beach, resulting in heavier traffic.

In this study, the final MLR model was then used to measure the MAPE in July and August 2013 and 2014. We found that the estimated MAPE in 2013 was 0.059, while it rose to 0.152 in 2014, suggesting an error rate of around 15.2%. Therefore, the estimation accuracy of the proposed model was 84.8%, which confirms its reliability.

The findings presented in this paper can be used to forecast traffic jams around the Ocean Beach area in Seoul. Further, they may be used as reference data for predicting traffic congestion on the roads into Ocean Beach. To increase the accuracy of the regression model, data from the past five years could have been collected and used in the regression model analysis to reflect the fundamental load of traffic congestion. In future work, we will thus employ mining techniques as well as statistical methods and compare their prediction accuracy with the findings presented herein.

Acknowledgment

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1011) supervised by the IITP(Institute for Information & communications Technology Promotion)

References

[1] M. Falk, "Impact of weather conditions on tourism demand in the peak summer season over the last 50years," *Tourism Management Perspectives*, vol. 9, pp. 24–35, 2014.

[2] H. Uk, K.-H. Kim, S.-O. Han *et al.*, "High-impact weather research center, forecast research division," *Korean Meteorological Society*, pp. 638–639, 2013.

[3] "Traffic congestion," http://en.wikipedia.org/wiki/Traffic_congestion.

[4] J. Lee and B. Hong, "Congestion score computation of big traffic data," in *Big Data and Cloud Computing (BDCloud), 2014 IEEE Fourth International Conference on*. IEEE, 2014, pp. 189–196.

[5] H. K. Seungjun Lee, Taeyoung Kim and K. Bok, "The evaluation of existing congestion indices' applicability for development of traffic condition index," *Journal of Korean Society of Road Engineers*, vol. 10, no. 3, pp. 119–128, 2008.

[6] "Variance inflation factor," https://en.wikipedia.org/wiki/Variance_inflation_factor.

[7] "Mean absolute percentage error," https://en.wikipedia.org/wiki/Mean_absolute_percentage_error.

[8] S. H. Eunbi Jeong, Cheol Oh, "Prediction of speed by rain intensity using road weather information system and vehicle detection system data," *Korea Intelligent Transport Systems*, vol. 12, no. 4, pp. 44–55, 2013.

[9] J. Andrey, B. Mills, M. Leahy, and J. Suggett, "Weather as a chronic hazard for road transportation in canadian cities," *Natural Hazards*, vol. 28, no. 2-3, pp. 319–343, 2003.

[10] K. G. Sohn Chul, "Influences of weather on the inbound traffic volume of a tourist destination," *The Korea Spatial Planning Review*, pp. 99–111, 2014.

Appendix

Table VI show a description of the variables abbreviation used in our MLR analysis. For example, G_RF_P_PD means rainfall probability in Ocean Beach (Gyeongpodae)'s forecast on the previous day.

TABLE VI: Abbreviation for variables

Abbreviation	Explanatione
G	Ocean Beach (Gyungpodae)
S	Seoul
T	Temperature
H	Humidity
WV	Wind velocity
Clear	Clear sky
L_Cloudy	Little Cloudy
RF	Rainfall
RF_P	Rainfall probability
PD	Previous day