This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2017.2716114, IEEE Internet of Things Journal

Data Driven Congestion Trends Prediction of Urban Transportation

Rui Jia, Pengcheng Jiang, Lei Liu*, Lizhen Cui, Yuliang Shi
School of Computer Science and Technology, Shandong University, Jinan, China
l.liu@sdu.edu.cn

***Abstract*—Smart traffic prediction system provides significant benefits in solving the city traffic congestion. However, existing smart transportation system needs a lot of real-time traffic data and accurate location information to display the traffic condition. We hope that we can use the data which is easy to be obtained, and then predict a reliable congestion time. To address this problem, this paper studied a smart traffic forecasting system based on SWARIMA model. The system includes three steps: (1) Use the Sliding Windows to calculate and process real-time data stream; (2) Establish the SWARIMA model and make regression analysis; (3) From a statistical point of view, calculate the elastic interval and predict the congestion trend. Our system is capable of accepting the real-time traffic data stream for the congestion prediction, in addition, we reduce the actual running parameters to three attributes: speed, time and location information. When faced with the challenges of real-time traffic congestion, the system can timely and effectively calculate the congestion trends and provide three reliable elastic intervals: warning, congestion and mitigation, which has significance to improve traffic condition and alleviate urban road congestion.**

***Index Terms*—Smart traffic, Congestion trends prediction, Data mining, ARIMA model, Sliding window**

I. Introduction

With the acceleration of urbanization process and the limitations of existing road networks in city, people's demand for an unobstructed road has been surpassed the expansion speed of urban road traffic networks. So the traffic congestion during the peak period has become a great challenge for every city [1], which causes a large amount of economic losses to the local city and will directly affect the city's image and development of the city [2]. A 2011 US urban transport report shows that the travel time delays and fuel consumption are estimated to be $ 121 billion in economic losses and it will increase to $ 190 billion in the future [3]. In China, the most influential of the top ten cities every day cause economic losses estimated at $ 1 billion because of traffic congestion. In addition to the United States and China, traffic congestion in the rest of the world is also a severe challenge [4]. There are many researches on alleviating and solving traffic congestion [5], [6], [7], but the smart prediction system has its unique contribution in alleviating traffic congestion. Assuming that people have arrived at the congestion section before updating the real-time congestion information, this congestion information has little significance for people who are in the congestion. If it is possible to predict a reliable congestion time or to provide a congestion elastic interval in advance, it will be of great benefit to ease congestion and reduce travel time. A simple and efficient smart real-time traffic prediction system will play an important role to the development of cities. The smart traffic system will be not only beneficial to the city's energy-saving and emission reduction, but also be able to make a more reasonable and efficient plan for people's peak travel, meanwhile it can provide constructive suggestions for the road networks of the city. This paper is working on studying a simple and effective traffic prediction system, making it more effectively to predict the whole process of congestion in advance and provide the elastic range, so that people can know the approximate formation time of congestion and then reduce the possibility of congestion.

There are many ways to deal with short term traffic data stream at present, for example, Jia-Dong Zhang et al focused on the aggregating and sampling methods for processing GPS data stream for traffic state estimation [8]; M. C. Tan et al utilized the Autoregressive Integrated Moving Average (ARIMA) to predict traffic stream [9]; while Lv Yisheng et al predicted the traffic stream with deep learning [10]. The traffic congestion is not a simple ON-OFF problem, because its formation has a certain trend and a number of points of mutation. But most of the predictions can't give an elastic interval for adjustment effectively. In addition, the traffic flow of urban roads has certain periodicity and weak stationarity [11]. For example, assuming a section in city often has traffic jams. If it congests during the morning peak on Monday, while at the same time of the next Monday, it will also congest, what's more its severity also has a certain similarity. In this paper, we found that most of the vehicles in the traffic congestion are low-speed, and the low-speed vehicles are also the main body of congestion. So most of the traffic congestions are formed by rapid increasement of low-speed vehicles at the beginning and while the real congestion will be formed, the number of low speed vehicles will tend to be stable. However, the congestion relief is just the opposite of this phenomenon, it is started by a relatively rapid reduction in the number of low-speed vehicles, and when the real congestion will be over, the low-speed vehicles will tend to be stable again in a period of time.

Therefore, in this paper, we make the following contributions based on this practical phenomenon. (1) Processing the original data stream with the preprocessor: Using sliding window method to analyze real-time traffic data stream and extract data features. (2) Using ARIMA model to deal with the above results, then performing mathematical modeling and solving regression equations. (3) Using a probabilistic model to find the feature points, resulting two elastic intervals which can be used as a prediction: congestion is about to occur in the first interval and ease in the second interval. (4) With a large number of actual traffic data and experiments, the results show that our method can well predict the traffic congestion. What's more, we compared our experimental results with the online analysis results of Baidu map, GaoDe map and other map softwares and the results show that the smart traffic prediction system does have a great advantage.

The rest of this paper is organized as follows. We introduce the preliminary knowledge of our work in Section II. Section III presents the details of SWARIMA model and algorithm. We introduce our experimental process, show a large number of processing results and evaluations in Section IV. Section V describes the application of the smart prediction system with comparing the results of our study and other map softwares in detail. We conclude the paper in Section VI.

II. Preliminary

A. Sliding Windows Model

Sliding windows model based on the data stream is one of the most common models of mining data. For example, SeungJong Noh et al studied the vehicle detection in highway environment with sliding window [12]; Leandro Pralon et al used the sliding window approach to study Convergence rate, Gaussian Sources, and Spatial Correlation [13]; G Sibley et al concerned with improving the range resolution of stereo vision for entry [14]; L Jin et al used a time-sensitive sliding window to mining recent frequent itemsets over data stream [15]. Gradually, with data stream constantly coming, the old data point which had been read will be thrown out or replaced by the new data point because of the less contribution [16]. In this model, both two ends of the window are identified as two sliding endpoints. When the new data stream arrives to the preprocessor, data which meet the requirement of the current window will be read. When the last data is filtered out, the window moves forward with time and the old data will be thrown out. This process is called as sliding.

Fig.1 gives an example of the sliding window, the horizontal axis indicates the time that the data stream arrives, and the vertical axis represents the speed, where S1 and S2 represent the speed range filtered by the preprocessor. For example, in this figure, P3 is an existed old data point, time window will slide to the solid box position because of the arrival of new data P6 and P7, meanwhile P3 is thrown. Besides, P4, P7 and other data without being read because they do not belong to the screening speed range.

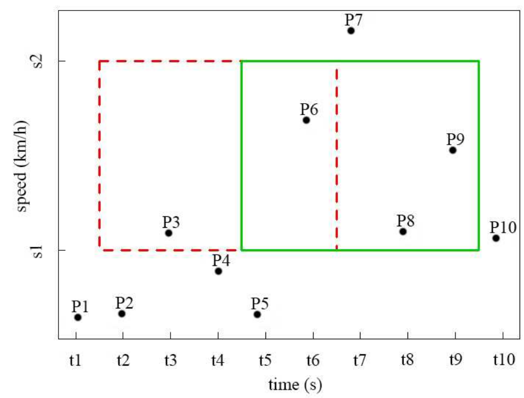


Fig. 1. An example of sliding windows.

B. ARIMA

Autoregressive integrated moving average (ARIMA) is one of the most widely used linear regression models [17], especially in the application of time series prediction in social economy [18] and networks [19]. For example, Kalid Yunus et al studied the modeling of wind speed time series based on ARIMA [20]; Carlos Alberto Villacorta Cardoso et al used ARIMA models and artificial neural network to forecast natural gas consumption [21]; W Zhang et al established a model of multiple seasonal ARIMA on time serial data and predicted for the seasonal time series [22]. In addition, many researchers have used this model to detect abnormal value or mutation value and to solve practical problems.

The AR model is also used to conduct traffic modeling, prediction and congestion control for high-speed networks [23]. It mainly displays the relationship between the time t arrived data X_t and the data X_{t-1}, X_{t-2} and so on before arriving. The definition is as follows:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon \quad (1)$$

Where c is a constant, p is the order of the model, $\phi_i (i = 1, 2, \ldots, p)$ are the parameters and ε is the error term that represents the difference between the calculating result and the true value.

The above AR model represents the relationship between the X_t and the correct terms. In order to show the relationship between it and the error terms, we need to use the MA model, which is defined as follows:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (2)$$

Where μ is the average value of series, $\varepsilon_i (i = t-1, t-2, \ldots, t)$ are the error terms, q is the order of the model and $\theta_i (i = 1, 2, \ldots, q)$ are the parameters. The above two models can be unified in a formula to describe and then the ARIMA model can be deduced. The definition is as follows:

$$X_t = \theta_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \quad (3)$$

Where $\theta_i (i = 1, 2, \ldots, q)$ and $\phi_i (i = 1, 2, \ldots, p)$ are the parameters. p and q are the orders, usually as an integer. X_t is the sample value of the data stream. $\varepsilon_i (i = t-1, t-2, \ldots, t)$

are the noise error terms. Obviously, if p is equal to 0, ARIMA model will be simplified as MA model; if q is equal to 0, ARIMA model will be simplified as AR model [24].

The ARIMA model shows well in prediction, so many researchers have used it for time series prediction. Because the ARIMA model can express the difference between the data,we use it to study the change of sensitive data in real-time data stream and then make accurate judgment and prediction in this paper.

III. SWARIMA Model and Algorithm Details

TABLE I
LIST OF NOTATIONS

Parameter	Description
X	Points come from data stream
t	Time of points from data stream
s	Speed of points form data stream
C	central speed of a sliding window
sr	Velocity radius of a sliding window
tr	Time radius of a sliding window
$tstep$	time step of windows
ε	error terms with a mean of zero
θ, ϕ	parameters of ARIMA mode
W	The point representing one window
o	The index of current window, which represents a short period of time
n	The current value of current window
u	The iteration variables in the process
z	The result of the smoothing operation for n
Z	The relatively smooth series include n of every window
mod	The regressive result of current window and windows before current window
Var_o	The excepted variance of W_o
$P_{i,j}$	The probability of W_j using $\hat{n}_{i,j}$ and Var_j
CS_o	The change point of current window W_o
FCS_o	The smoothed change point of current window W_o

A. SWARIMA Model Introduction

In this paper, we propose a new method named SWARIMA, which includes three stages that will be detailed in Section II, Section III and Section IV. And the Fig.2 shows the flow diagram model of SWARIMA in detail.

The first stage: pretreatment stage. SWARIMA process the raw traffic data stream through a preprocessor and extract features. First, we obtain a current point as the beginning of sliding window. While the sliding window detects the data stream, it filter out the points that not match the window and perform the operation of adding data. When the detection is completed, the sliding window with data stream will speed forward automatically. Repeat the above process to continuously detect the data stream until the end.

The second stage: modeling stage. When the detection of one window is completed, SWARIMA uses the information of this window to update the regression model, and tries to use the new model to predict the attribute information of the current point and the next point.

The third stage: calculating the score of this change process. First, we obtain the prediction scores of the current data point, which is calculated by the regression model of the previous window. Second, we calculate the score of current point in the current window. And finally compare the difference between the two scores. In this paper, the probability model is used to evaluate the difference. If the score exceeds a certain threshold, the corresponding time of the current point is considered to be a drastic change.

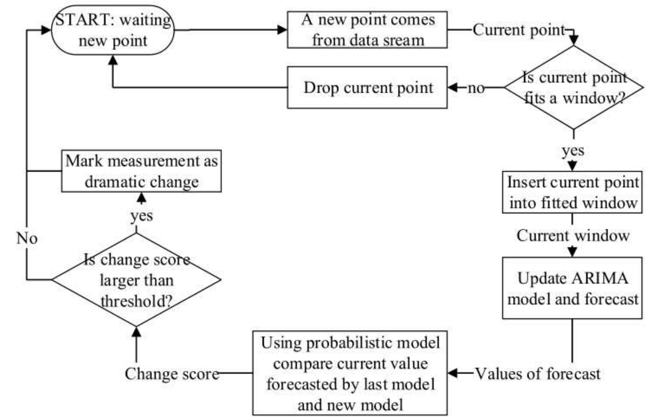


Fig. 2. A flow diagram of SWARIMA model.

B. Preprocessing Stage

In this paper, the raw data which is used to deal with is source from the measurements at the intersections in a designated city in China. The raw data contains five basic attributes: arrived time, the license plate number, vehicle speed, vehicle arrival direction and crossing number. In order to convert the raw data into preprocessing file which is available for the use of the second stage expediently, we design a preprocessor to complete it. In Fig.3, the operation of the preprocessor is divided into four steps: grouping, removal, selection and statistics, while the grouping and removal can be implemented at a same time. In order to save time cost, each data point is read only once, and the processor use online processing to deal with a large number of real-time data streams. Specific steps are as follows:

The first step: grouping. In order to simplify the subsequent statistical work, we first group the raw data. When the data stream arrives at a certain time, the preprocessor first groups the data according to the attributes of the data, such as vehicle arrival direction, crossing number and so on.

The second step: removal. After grouping, there are still a lot of attributes redundant in the data set. The removal step not only saves the time cost of the follow-up steps but also reduces the storage cost of the system. In fact, we only need to read the useful data properties into the fitting array, while ignoring other irrelevant attributes or the attributes of small impact.

The third step: selection. This step, actually, can be carried out with the removal. In this paper, we mainly study the number of low speed vehicles that can respond to the change of congestion. Therefore, it is necessary for us to filter the high speed vehicles data and some strange data points. Thus, only

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2017.2716114, IEEE Internet of Things Journal

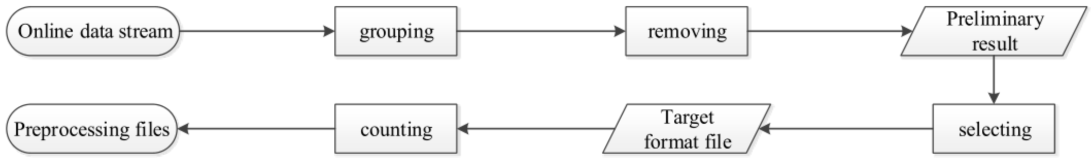


Fig. 3. A flow diagram of Preprocessing Phase.

the data points that meet the requirements of the window speed are obtained, so as to generate the standard format files. While this step is also the most important step in the preprocessor.

The fourth step is statistics. The statistics plays an important role in the sliding window model, aiming at processing data points after above the three steps according to the sliding window method into the preprocessing file for data regression. First define the time center and the time radius of the sliding window tr and then begin to statistics on each standard format files from the first time window. If a data point is in a time range of a sliding window, the sliding window corresponding to the array of values plus one. If a data not only belong to the previous window but also belongs to the next window, then this one will be recorded by the two windows at the same time. With the sliding window sliding, we can traverse the standard format file to obtain the standardized files which contain the number of each sliding window. The achievement of pseudo code in algorithm 1.

Algorithm 1 The standardization of raw data

Require: X,C,sr,tr,tstep
Ensure: win
//when an object X comes, get and process it.
GET X
LET $t = GetSeconds(X)$
LET $sp = GetSpeed(X)$
if $C - sr < sp \&\& sp <= C + sr$ **then**
  //compute the numbers of leftmost and rightmost fitted windows
  LET LW = (t - 2 * tr) / tstep + 2
  LET RW = t / tstep + 1
  for $i = LW$ to RW **do**
    $win[i] = win[i] + 1$
  end for
end if

C. Modeling Stage

The output of the previous stage will be the input for this stage. In order to facilitate the description of the subsequent steps, we define a current window, which indicates that the previous window of statistics has been completed and output to the window in this stage and the internal data is processed by the two-dimensional data. SWARIMA is a method that can be used to predict multi-dimensional data, while in this paper we only use two-dimensional. Before the data regression, we observed that the overall sample is large and the fluctuation of sliding window data. In order to reduce the impact of these data on the experimental results, we will make an average of each value and the first two values as the current value of the subsequent calculations. Subsequently, SWARIMA predicted the current and the next value and used a probabilistic model to compare the predictive value and the actual value, so as to reflect the road conditions. As mentioned in Section II of this paper, the ARIMA model has great advantages and convenience in solving this kind of problem and the SWARIMA model also inherits the advantage of the ARIMA. The algorithm 2 is the prediction process. The variables in the specific operation process are defined as follows:

Algorithm 2 The congestion trends forecast

Require: value n_o of current window W_o ; values of windows before W_o
Ensure: the predicted value of W_o and W_{o+1}
$n_o \leftarrow (n_{o-2} + n_{o-1} + n_o) \div 3$
$u_o = (1-r)n_{o-1} + rn_o$
$z_o = n_o - u_o$
Z.add(z_o)
/* using ARIMA(p,d,q) regress $z_i (i = 1, 2, \ldots, o)$*/
$mod \leftarrow ARIMA(Z, p, d, q)$
$\hat{z}_{o,o} \leftarrow prdict(mod, Z_{1:o})$
$\hat{n}_{o,o} = \hat{z}_{o,o} + u_o$

The current point is represented by W_o or (o, n), where o represents the current window of time, and n_o represents the current value. The new sequence N to be used for analysis consists of the current n and the values before it, while o is the index of the current window and the result model is defined as M_o, which is calculated by $n_i \, (i = 1, 2 \ldots o)$. In the formula, p, d and q are predefined parameters of the ARIMA model.

Next, we define a new symbol $\hat{n}_{i,j}$,which means that the value of the n_j is predicted by model M_i. The SWARIMA model, for example, uses the current results to predict the value of the current point $\hat{n}_{o,o}$, and the value of the next point $\hat{n}_{o,o+1}$. And such a n_o will reflect the congestion trend in the subsequent steps. Finally, $\hat{n}_{o,o}$ and $\hat{n}_{o,o+1}$ will be processed by the next step in the system.

D. Calculating Score of The Change Process

Firstly, we assume that n is affected by the Gauss distribution, and the expectation and variance are necessary in the Gauss distribution model. By the introduction of step C, we can assume that expectations are $\hat{n}_{o,o}$ and $\hat{n}_{o-1,o}$, and the variance can be calculated by the equation (4).

$$Var_o = r \cdot Var_{o-1} + (1-r) \cdot (n_o - \hat{n}_{o,o})^2 \qquad (4)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2017.2716114, IEEE Internet of Things Journal

Equation (5) is the Gauss distribution formula with the mean μ and variance σ^2. Then we can use $\hat{n}_{o,o}$ and $P_{o-1,o}$ respectively calculate the possible value of n_o, in order to facilitate the description, we use $P_{o,o}$ and $P_{o-1,o}$ to represent these two values.

$$P(x\left|\mu,\sigma^2\right.) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{5}$$

Then we use $\Delta P_{o-1,o}$ to express the difference between $P_{o,o}$ and $P_{o-1,o}$. At this point we can use the score calculation formula to calculate the value of CS_o, the formula (6).

$$CS_o = -\ln(|P_{o,o} - P_{o-1,o}|) \tag{6}$$

Taking into account the volatility of the data, this paper uses the average method to calculate the final score of FCS_o, and its value is equal to the average value of CS_{o-2}, CS_{o-1}, and CS_o. When the FCS_o of the window W_o exceeds the threshold, then it is called the time point of the window has changed dramatically. In this paper, we mainly study the traffic congestion of urban traffic peak, so SWARIMA model will use the average change points of continuous several peak windows as threshold. And in the process of a large number of experiments, we have used data from consecutive days and the same day of consecutive weeks to perform the experiment and verification. The pseudo code calculation process is shown by the algorithm 3, and it consists of three steps: calculating the variance, using the Gauss distribution to calculate the probability and finally using the SWARIMA model to calculate the FCS. If the score is greater than the threshold, then check whether the change point is the beginning or the end of the congestion, otherwise the state of the road has not changed.

Algorithm 3 Computing the change scores

Require: predicted value $\hat{n}_{o,o}$ and $\hat{n}_{o-1,o}$, real value n_o, previous variance Var_{o-1}, threshold value T_o

Ensure: the predicted road state, which may be smooth, warning, congestion, or mitigation

```
/*Compute the variance and mean of */
Var_o ← r · Var_{o-1} + (1 − r) · (n_o − n̂_{o,o})^2
/*compute the change score and make sure the state*/
P_{o,o} ← Gaussian(μ = n_{o,o}, σ^2 = Var_o)
P_{o-1,o} ← Gaussian(μ = n_{o-1,o-1}, σ^2 = Var_o)
CS_o ← −ln(|P_{o,o} − P_{o-1,o}|)
FCS_o ← (CS_{o-2} + CS_{o-1} + CS_o) ÷ 3
if CS_o ≥ T_o then
    state_o ← check(W_o)
else
    state_o ← normal
end if
```

IV. Experimental Results and Evaluation

The experimental data used in this paper is the intersection traffic data which is obtained from the designated city in China for a month, and its contents include time, license plate number, vehicle speed, vehicle direction and intersection number. And there are more than five million records every day. After the grouping step, we mainly use two important attributes: arrival time and the vehicle speed. In order to intuitively understand the characteristics of the data, we directly use the original data to draw a diagram, as shown in Fig.4, the horizontal axis represents the window time of one day, and the vertical axis represents the number of vehicles. As can be seen from the Fig.4, the number of vehicles in the designated city in the morning and evening peak will produce significant mutations in the vicinity of some points in time, and these mutations are located just in the time range of the morning and evening peak. What is more, we can notice that in many time windows, there are some obvious data fluctuations which will influence our observation and the processing of data stream, so it is necessary for this paper to use the mean value to preprocess more smoothly. In addition, we can observe that the formation and remission of the morning peak is much more intense than the evening peak, which is also in line with our understanding of the actual situation. Therefore, the actual experiment in this paper is to calculate the morning and evening peak data respectively to ensure the scientific results.

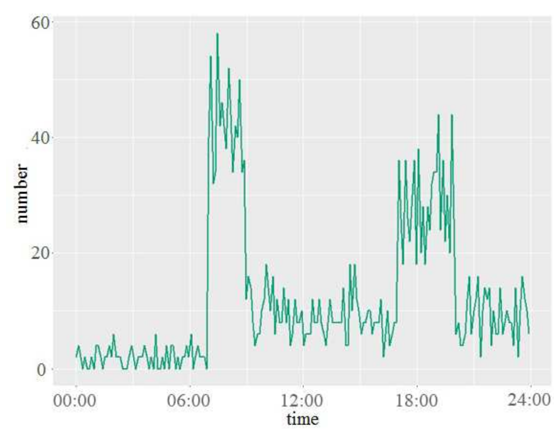
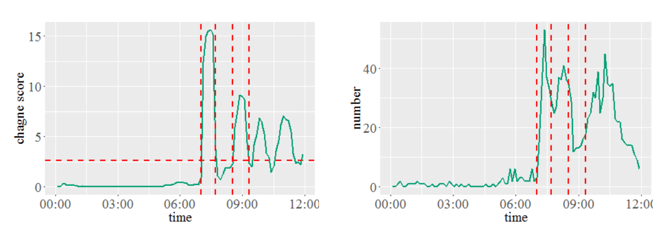


Fig. 4. Real data in the whole day

Fig.4 is directly drawn from the original data. From this figure, we can roughly see the morning and evening peak congestion time and congestion trends.



(a) The process results of SWARIMA (b) Original real data.

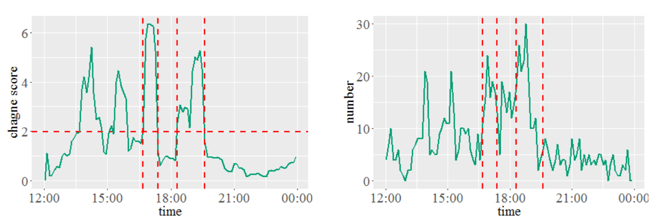
Fig. 5. The results and the original data for morning peak.

Fig.5(a) and Fig.5(b), respectively, are the morning results of the score calculation and original data map. The red line which is parallel to the horizontal axis is the threshold calculated, and the calculation method is to take the average of the change of 7:00 to 9:30 in the morning before several days.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2017.2716114, IEEE Internet of Things Journal

Based on the intersections of the threshold line and the original image, we can draw some vertical lines that are perpendicular to the horizontal axis, and these intersections are the critical time we require.

As shown in Fig.5(a), we can observe that these several time points are 7:00 a.m, 7:35 a.m, 8:30 a.m and 9:15 a.m, while the four time points show that the number of low-speed vehicles vary most dramatically at these time points. Before the full formation of congestion, the number of low-speed vehicles has gone through a change from less to more, and then from more to less, which is like to the quadratic curve. While before the complete remission, the number of low-speed vehicles change from less to more and then from more to less again. Obviously, this result is consistent with the view in the Introduction Section. The former change process is called as the early warning period, the latter is known as the mitigation period, and the real congestion is just located between these two elastic intervals. In short, every formation of morning peak will go through three periods: early warning period, congestion period and remission period. While a large number of experimental results also verify the existence of such a phenomenon. From the point of life experience, the first interval can be considered as the beginning of congestion and the starting point is considered to be the starting point of all stages; the second interval can be considered to be the end of the congestion and the end point is considered to be the end point of all stages. So far, the prediction process of morning peak can be fully displayed.



(a) The process results of SWARIMA (b) Original real data.

Fig. 6. The results and the original data for evening peak.

Fig.6(a) and Fig.6(b) are the evening time results of the score calculation and original data map respectively. The same to the description of the last paragraph, we can observe the evening peak congestion also contains three periods from the Fig.6(a), while the four time nodes are 16:41 p.m, 17:23 p.m, 18:19 p.m and 19:36 p.m. Between 16:41 p.m and 17:23 p.m are early warning period; Between 17:23 p.m and 18:19 p.m are congestion period; Between 18:19 p.m and 19:36 p.m are mitigation period. Although the mitigation period is later than the experience conclusion, the prediction results still provide a relatively reliable reference value in a certain range of allowable error.

In addition, we also try to carry out simulation experiments to verify the reliability of the SWARIMA method. On the basis of experimental data, the data of simulation experiment is created by the Poisson distribution, which rely on the average number of vehicles to distinguish between peak period and other periods, and the mean of the peak is significantly greater than the mean of other periods. Fig.7 is drawn directly from the simulated data. Fig.8 is the original figure and the processed figure for morning, and Fig.9 is the original figure and the processed figure for evening.

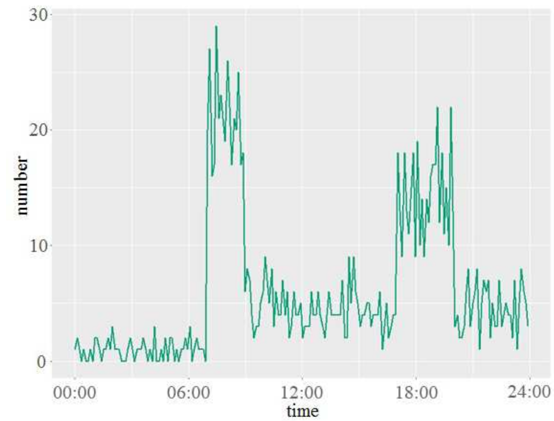
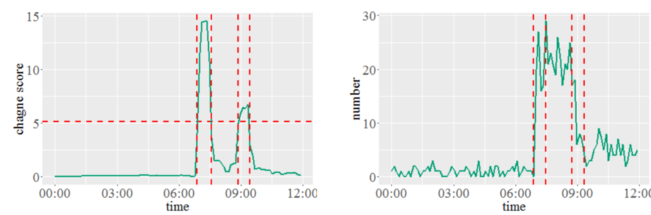
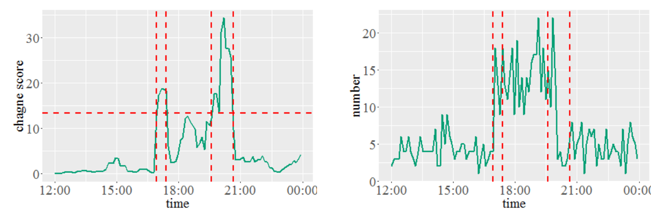


Fig. 7. Simulation data of a whole day.



(a) The process results of SWARIMA (b) Original simulation data

Fig. 8. The results and the Simulation for morning peak.



(a) The process results of SWARIMA (b) Original simulation data.

Fig. 9. The results and the Simulation for evening peak.

From Fig.8, we observe four time nodes are 7:00 a.m, 7:35 a.m, 8:52 a.m and 9:27 a.m. The four time nodes in Fig.9 are 17:00 p.m, 17:30 p.m, 19:42 p.m and 21:27 p.m. The mitigation period is significantly later than the actual data. The reason is that we deliberately delayed the evening peak in order to test the sensitivity of the SWAR model in the production of simulation data. Obviously, the improved model also has excellent performance in the processing of simulation data and can clearly determine in which time window the original data exists dramatic changes and find out the corresponding time points. This also confirms the reliability of our experiment from another angle.

What is more, we conducted a large number of data experiments. Fig.10 to Fig.14 are the processing figure and original data figure of the morning and evening peaks in five different sections. Next we will explain in detail how to clearly understand and distinguish the range and time interval of congestion. The first step, by smoothing process we process the original graph into a graph which can be analyzed. And this process is clearly described in the preprocessing and modeling stages; The second step, we need a threshold to divide the time point whether there is congestion. In this paper, we will use the change score FCS as our threshold, and the paper will calculate threshold of each half day to ensure the accuracy of the threshold in different images. Once the curve in the image is below this threshold, we consider this region not congested. on the contrary, we believe that this region exists congestion and conduct the third step analysis; The third step, a large number of data results show that there is a formation period in the occurrence of each congestion and there is a mitigation period in the disappearance of each congestion. Both these two periods have the start time point and end time point. Therefore, after the complete processing, we can clearly find that each threshold has four intersections with the image. And the meaning of these four intersection coordinates is the beginning of congestion formation, the end of congestion formation, the beginning of congestion mitigation and the end of congestion mitigation. Two images as a set, from the left image we can see that the red horizontal line is the threshold which we calculated and it has four intersections with the image. While we identify these four time points on the right original data image, it can be seen that the entire peak time is completely included by the beginning time point and the end time point. In addition, each region of the predicted results not only has the elastic range of same characteristics, and the elastic range is also in line with the actual situation, which shows that our prediction scheme is accurate and effective.

V. Application

In this paper, based on the theory of intelligent transportation, we process, analyze and regress the real-time data stream obtained from the electronic sensing at the intersection, and obtain the SWARIMA model curve of the peak time, and predict the whole congestion process simply and effectively. In a short time, we can smartly transform traffic information and instant messaging data into predictive services so as to provide it to users through a variety of ways. This smart service is of great significance for the majority of city traffic, especially for the roads in large cities and special sections. By effectively forecasting peak time of the congestion, administrators can timely relieve traffic pressure, improve the traffic conditions of the congested road sections, and reduce the occurrence of traffic accident [25].

A. System Characteristics

(a) Universal: Firstly, since the system gets data from the city's basis traffic facilities, such as the intersection of detection equipment, so we can easily get a lot of experimental raw data. Secondly, due to the standardization of urban traffic monitoring equipment, we can obtain unified data format. Thirdly, the large number of verification tests in this paper show that SWARIMA model will not fluctuate obviously due to different intersection status. Fourthly, the processing results of the system is convenient for secondary analysis and post-processing.

(b) Accuracy: The data used in this paper is source from the real data sets. From a certain angle, the precision of the equipment also determines our acquisition accuracy. But the results show that our original data set has fully met the requirements of the SWARIMA model. In addition, the system's data analysis algorithm and each step of the data processing methods are effective, and the final results also reflect the accuracy of the process.

(c) Prediction and Elastic Intervals: Compared with other systems, this is the most important difference of the system in this paper. Real-time traffic system uses the vehicles positioning equipment to carry out the statistics, and visually display the real-time congestion curve. In this paper, mathematical algorithm is used to analyze the change of the number of vehicles, so as to predict the approximate time of the congestion, and give the warning time and mitigation time of two elastic intervals for people to refer to. Early prediction allows the users to adjust the travel route as early as possible, avoid traffic congestion, reduce traffic congestion and improve road conditions.

(d) Low-Complexity: This system is divided into two parts: preprocessing and data analysis. The preprocessing section is optimized so that the complexity is reduced to O(n), that is, the sliding window traverses the original data once to obtain the preprocessed file. Subsequent processing using the R language standard library function to ensure the operating efficiency and low time complexity. The whole system does not have a complex processing mechanism, and its processing results can be presented in a short time.

(e) Visualization: We believe that the visualization of the final processing results is necessary. Although we can clearly show the elasticity interval of congestion occurrence and mitigation, it is easier for people to understand and use our results [26] in the form of web pages. As shown in Fig.15, we combine the results of the system with clocks and watches, and use different colors to distinguish different road conditions. When the users input the date, time, the road number and the direction, the system will automatically generate a circular chart on the right side of the page. The four mark points in the graph are the change points of the road condition. The four colors of green, yellow, red and blue in the ring indicate four traffic states: smooth, warning, congestion and mitigation.

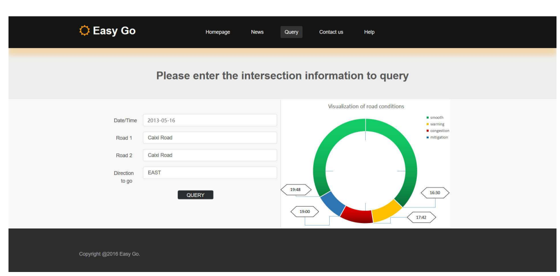
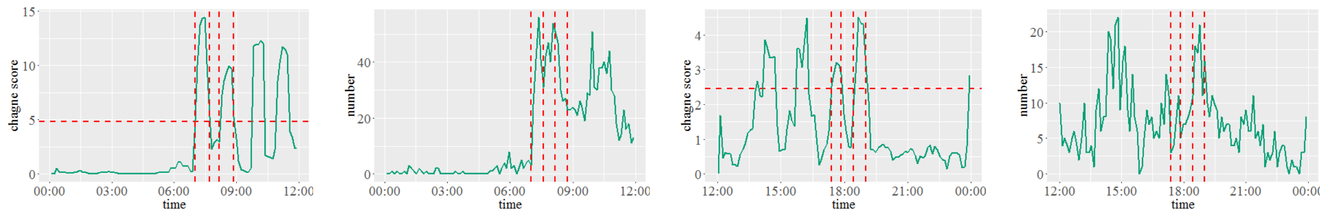


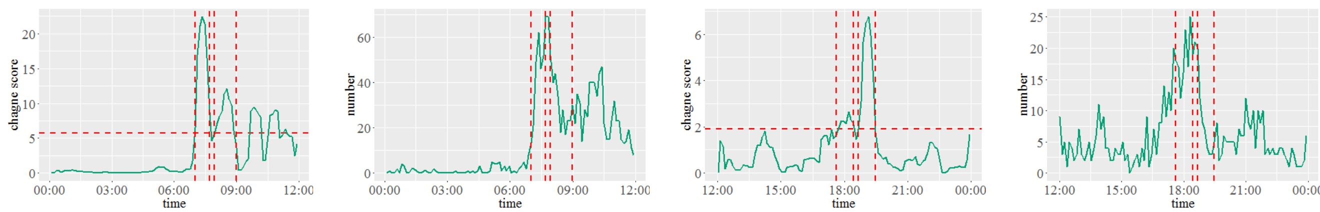
Fig. 15. The query page of smart traffic prediction.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2017.2716114, IEEE Internet of Things Journal



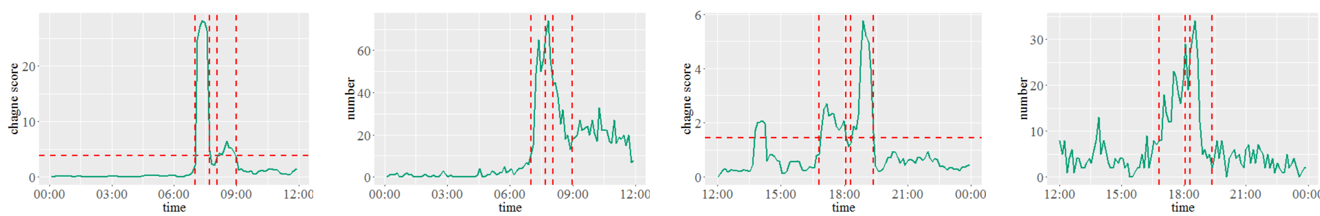
(a) morning-processed (b) morning-original (c) evening-processed (d) evening-original

Fig. 10. The results and the original graphs of the first section.



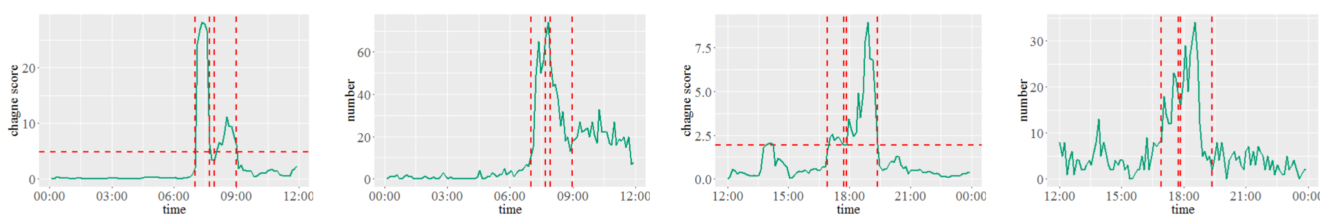
(a) morning-processed (b) morning-original (c) evening-processed (d) evening-original

Fig. 11. The results and the original graphs of the second section.



(a) morning-processed (b) morning-original (c) evening-processed (d) evening-original

Fig. 12. The results and the original graphs of the third section.



(a) morning-processed (b) morning-original (c) evening-processed (d) evening-original

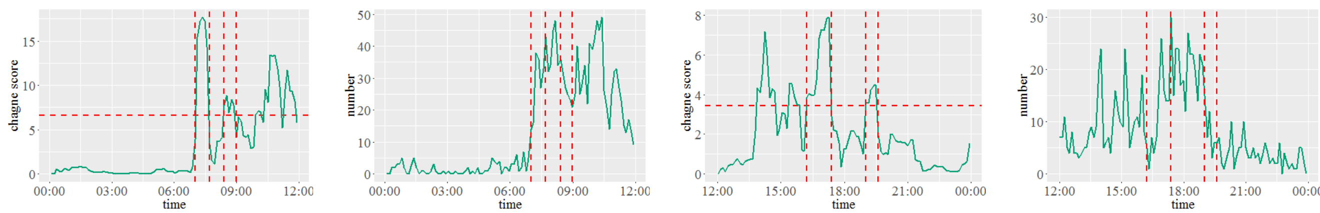
Fig. 13. The results and the original graphs of the fourth section.

B. Application

There are many differences between the smart forecasting system in this paper and the real-time traffic system which is widely used in the market. Details are as follows:

(a) Essential Difference: The real-time traffic system is based on the wide range of GPS positioning and large data statistics to obtain the results. This approach requires a large amount of terminal equipment support [27]. The current mainstream map applications, such as GaoDe Map, Baidu Map [28], are through GPS or other mobile devices to collect real-time location and speed of the vehicle information and distinguish the number of vehicles on a road with different colors. The color is red when there is a large number of vehicles or even congestion, while it is green when there is a small number of vehicles. The smart prediction system uses the intersection monitoring equipment to collect and process the real-time data stream, and calculates the results by SWARIMA. Whenever our system receives the data stream, we calculate the mutation value of the number of low-speed vehicles according to the steps of preprocessing, regression analysis, calculating the change scores and then use the peak prediction method to predict the elastic interval of congestion occurrence and mitigation. And then we update the website with the latest results for users to query.

(b) The data used for processing is different: Compared with real-time traffic system, smart prediction system only uses the



(a) morning-processed (b) morning-original (c) evening-processed (d) evening-original

Fig. 14. The results and the original graphs of the fifth section.

low-speed vehicles obtained by the preprocessing to analyze. Because the decrease of the mean value of vehicle speed and the increase of the number of low-speed vehicles are the main characteristic of traffic congestion. This not only makes the sample data in the SWARIMA system more lightweight, saves a lot of time cost of system operation, but also can accurately grasp the information and alleviation of road congestion.

(c) System results: Real-time traffic system determines whether there is traffic congestion from the results. If the number of vehicles stranded on the road for more than a certain threshold, it will be judged to be congested. Therefore, a sudden change in color may occur at a certain time, and this also means it is difficult for people to know which stage of the elastic interval does the current road belongs to. The biggest advantage of smart prediction system is to predict the formation of congestion in advance and to provide a reference elastic interval for the three stages of congestion process. The four points provided by the system are all abrupt points in the change process of the low-speed vehicles'number. According to experience, we have the following division: the two endpoints of the warning phase indicate that the road has a congestion trend and the congestion will form immediately. The two endpoints of the mitigation phase indicate that the road has a mitigation trend and the road will be unobstructed immediately. And the time between these two phases are the real congestion phase. This results is also better in line with people's understanding of congestion formation and mitigation.

(d) Contradistinction: We took a screenshots of Baidu map and Gaode map displayed results of the same time, and compared it with the results of this paper. The specific contrast is as follows:

TABLE II
THE COMPARISON BETWEEN PREDICTIVE RESULTS AND BAIDU MAP.

Representative time	the morning peak	
	Prediction time	Real time in Baidu maps
warning time	7:00 to 7:35	7:25
congestion	7:35 to 8:30	8:03
mitigation	8:30 to 9:15	8:46

Three real time points in TABLE II are the whole process of congestion formation and mitigation in a section. We divide it into three levels. Level 1: the time point is 7:25 a.m, the road was yellow in this figure and the yellow color appears continuously from this moment, indicating that there were more low-speed vehicles than before, and the congestion may occur. Level 2: the time point is 8:03 a.m. The road in this figure was red and was the first time from yellow to red, which indicated that it had entered the congestion state. Level 3: the time point is 8:46 a.m, and the road in this figure turned yellow-green again while no red color appeared after that time, indicating that there was no congestion situation from this moment. We compared the predicted results (as shown in Fig.5) with the actual congestion in the section, as shown in Table II. The three elastic intervals given by our system are: warning time (from 7:00 a.m to 7:35 a.m), congestion (from 7:35 a.m to 8:30 a.m) and mitigation (from 8:30 a.m to 9:15 a.m). What is more, the above three points of time also fit in the elastic interval shown in Fig5, respectively. This shows that, within allowable error range, our smart prediction system successfully predicts the congestion situation of morning peak before it happened, which means we can predict a warning interval before the dramatic number increasement of low-speed vehicles, a congestion interval before the real congestion happens, and a mitigation interval before the complete remission.

Fig.16, 17 and 18 are the whole process of congestion formation and mitigation in a section. We also divide it into three levels. Level 1: the time point in Fig.16 is 5:08 p.m, and at this time the intersection appeared yellow first time, which said the number of low-speed vehicles were more and the congestion may occur. Level 2: Fig.17 corresponds to the time point of 5:24 p.m, this is the first time that the intersection appeared red and then continued to be red, said the intersection had been in a state of congestion already. Level 3: the time point in Fig.18 is 6:29 p.m, then no longer appeared red in this intersection, indicating that there was no congestion situation from this moment. We compared the predicted results (as shown in Fig.6) with the actual congestion in the section, as shown in Table III. The three elastic intervals given by our system are: warning time (from 16:41 p.m to 17:23 p.m), congestion (from 17:23 p.m to 18:19 p.m) and mitigation (from 18:19 p.m to 19:36 p.m). Meanwhile, these time points also respectively fit in the elastic interval shown in Fig6. This shows that, within allowable error range, our smart prediction system successfully predicts the congestion situation of evening peak before it happens. By far, the smart prediction system is finished.

VI. CONCLUSION

In this paper, we studied the smart traffic prediction system based on SWARIMA method. Inspired by an observation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2017.2716114, IEEE Internet of Things Journal

Fig. 16. 5:08 p.m. Level 1 in GaoDe Maps.

Fig. 17. 5:24 p.m. Level 2 in GaoDe Maps.

on the causes of traffic congestion, we found that most sections of the urban road networks had similar congestion characteristics. Therefore we propose a three-step model to solve the congestion prediction problem: (1) we first use the preprocessor to preprocess the original data directly, and then (2) the SWARIMA model is used to analyze the mathematical model and the regression analysis, and finally (3) the Gauss distribution model is used to calculate the change scores so as to further calculate three elastic intervals: warning, congestion the mitigation. In addition, this paper also carries out the simulation data test. Experimental results show that our method are superior to real-time traffic system in terms of the congestion trends prediction and providing the elastic range of congestion. Through the long-term accumulation of data, the public can be more reasonable, according to the congestion prediction, to plan their own itinerary and reduce energy consumption; managers are able to manage frequently-congested sections

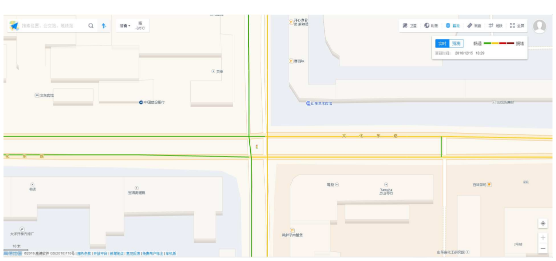


Fig. 18. 6:29 p.m. Level 3 in GaoDe Maps.

TABLE III
THE COMPARISON BETWEEN PREDICTIVE RESULTS AND GAODE MAP.

Representative time	the evening peak	
	Prediction time	Real time in Gao De maps
warning time	16:41 to 17:23	17:08
congestion	17:23 to 18:19	17:24
mitigation	18:19 to 19:36	18:29

and adjust the traffic conditions comprehensively, thus helping to alleviate the traffic congestion in the city.

ACKNOWLEDGEMENT

This paper is partially supported by National Science Foudation China (61402262, 61572295). Lei LIU is the corresponding author of this paper.

REFERENCES

[1] S. Wongcharoen and T. Senivongse, "Twitter analysis of road traffic congestion severity estimation," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, July 2016, pp. 1–6.

[2] S. Liu, K. P. Triantis, and S. Sarangi, "A framework for evaluating the dynamic impacts of a congestion pricing policy for a transportation socioeconomic system," *Transportation Research Part A Policy & Practice*, vol. 44, no. 8, pp. 596–608, 2010.

[3] S. Wang, S. Djahel, Z. Zhang, and J. Mcmanis, "Next road rerouting: A multiagent system for mitigating unexpected urban traffic congestion," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2016.

[4] J. I. Levy, J. J. Buonocore, and V. S. Katherine, "Evaluation of the public health impacts of traffic congestion: a health risk assessment." *Environmental Health*, vol. 9, no. 1, pp. 1–12, 2010.

[5] S. Wang, S. Djahel, Z. Zhang, and J. McManis, "Next road rerouting: A multiagent system for mitigating unexpected urban traffic congestion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2888–2899, Oct 2016.

[6] G. Kim, Y. S. Ong, T. Cheong, and P. S. Tan, "Solving the dynamic vehicle routing problem under traffic congestion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8, pp. 2367–2380, Aug 2016.

[7] A. Ahmad, R. Arshad, S. A. Mahmud, G. M. Khan, and H. S. Al-Raweshidy, "Earliest-deadline-based scheduling to reduce urban traffic congestion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 4, pp. 1510–1526, Aug 2014.

[8] J. D. Zhang, J. Xu, and S. S. Liao, "Aggregating and sampling methods for processing gps data streams for traffic state estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1629–1641, Dec 2013.

[9] M. C. Tan, L. B. Feng, and J. M. Xu, "Traffic flow prediction based on hybrid arima and ann model," *China Journal of Highway & Transport*, vol. 4,20, 2007.

[10] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 1–9, 2014.

[11] C. Dong, C. Shao, D. Zhao, and Y. Liu, *Short-Term Traffic Flow Forecasting Based on Periodicity Similarity Characteristics*. Springer Berlin Heidelberg, 2013.

[12] S. Noh, D. Shim, and M. Jeon, "Adaptive sliding-window strategy for vehicle detection in highway environments," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 323–335, Feb 2016.

[13] L. Pralon, G. Vasile, M. D. Mura, J. Chanussot, and N. Besic, "Evaluation of ica-based ictd for polsar data analysis using a sliding window approach: Convergence rate, gaussian sources, and spatial correlation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 4262–4271, July 2016.

[14] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *Journal of Field Robotics*, vol. 27, no. 5, p. 587C608, 2010.

[15] L. Jin, D. J. Chai, Y. K. Lee, and K. H. Ryu, "Mining frequent itemsets over data streams with multiple time-sensitive sliding windows," in *International Conference on Advanced Language Processing and Web Information Technology*, 2007, pp. 486–491.

[16] F. Angiulli and F. Fassetti, "Detecting distance-based outliers in streams of data," in *Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November*, 2007, pp. 811–820.

[17] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and arima models for time series forecasting," *Applied Soft Computing*, vol. 11, no. 2, pp. 2664–2675, 2011.

[18] R. M. K. T. Rathnayaka, D. M. K. N. Seneviratna, W. Jianguo, and H. I. Arumawadu, "A hybrid statistical approach for stock market forecasting based on artificial neural network and arima time series models," in *2015 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, Oct 2015, pp. 54–60.

[19] S. M. T. Nezhad, M. Nazari, and E. A. Gharavol, "A novel dos and ddos attacks detection algorithm using arima time series model and chaotic system in computer networks," *IEEE Communications Letters*, vol. 20, no. 4, pp. 700–703, April 2016.

[20] K. Yunus, T. Thiringer, and P. Chen, "Arima-based frequency-decomposed modeling of wind speed time series," *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 2546–2556, July 2016.

[21] C. A. V. Cardoso and G. L. Cruz, "Forecasting natural gas consumption using arima models and artificial neural networks," *IEEE Latin America Transactions*, vol. 14, no. 5, pp. 2233–2238, May 2016.

[22] W. Zhang, Y. Q. Zhang, and X. Yang, "Model of multiple seasonal arima and its application to data in time series," *Acta Academiae Medicine Militaris Tertiae*, vol. 24, no. 8, pp. 955–957, 2002.

[23] B.-S. Chen, S.-C. Peng, and K.-C. Wang, "Traffic modeling, prediction, and congestion control for high-speed networks: a fuzzy ar approach," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 5, pp. 491–508, Oct 2000.

[24] G. P. Zhang, "Time series forecasting using a hybrid arima and neural network model," *Neurocomputing*, vol. 50, no. 1, pp. 159–175, 2003.

[25] X. Chen, Z. Yang, and J. Feng, "A study on the prediction technology of urban road intersection congestion," in *International Conference on Intelligent Computation Technology and Automation*, 2009, pp. 456 – 459.

[26] S. Koshman, "Information visualization: Human-centered issues and perspectives," *Journal of the Association for Information Science and Technology*, vol. 60, no. 12, p. 2588C2590, 2009.

[27] W. H. Lee, S. S. Tseng, and W. Y. Shieh, "Collaborative real-time traffic information generation and sharing framework for the intelligent transportation system," *Information Sciences An International Journal*, vol. 180, no. 1, pp. 62–70, 2010.

[28] W. Pan, Z. Chen, L. I. Xingmin, and S. O. Computer, "Decryption method of baidu map's coordinates based on artificial neural network," *Computer Engineering & Applications*, pp. 110–113, 2014.