Research on GSTAR-SVM Traffic Prediction Model Based on Wavelet Transform

Jinghao Hu¹, Shandong Wang** and Jian Mao²

¹ School of Earth Sciences and Engineering, Hohai University, Nanjing, Jiangsu, 210098, China
² School of Geographic and Oceanographic Sciences, Nanjing University, Nanjing, Jiangsu, 210023, China
*Corresponding author’s e-mail: wsd@hhu.edu.cn (Shandong Wang)

Abstract. With the continuous advancement of urbanization in China, the number of urban vehicles has increased rapidly, and traffic congestion has become an urgent problem for modern cities. Therefore, accurate real-time traffic flow prediction is of great significance for solving this problem. However, traffic flow, as a special type of prediction object, has the characteristics of complexity, uncertainty and nonlinearity, which bring great difficulties to the prediction, and the external spatial correlation also has a great influence on the prediction results. In this paper, based on the above problems, the GSTAR-SVM traffic prediction model based on wavelet transform is constructed to predict the short-term traffic flow, and the model is verified based on the data. The experimental results show that the proposed model has higher prediction accuracy.

1. Introduction
Traffic flow prediction is the use of the past traffic data to predict traffic flow for a certain period of time in the future by establishing relevant predictive models. Short-term traffic flow prediction generally refers to traffic flow prediction within a forecast period of less than 15 minutes. At present, many scholars have conducted a lot of related researches on this issue. There are two types of methods that are more commonly used: (1) Predictive models based on statistical theories such as Kalman filtering, time series models, Bayesian networks and Markov chains; (2) Data-driven machine learning prediction models such as decision trees, support vector machines, BP neural network and other methods.

The above prediction methods for short-term traffic flow are based on the relevant factors affecting traffic or in time series, but do not pay attention to the spatial correlation characteristics existing between traffic flows. In real life, the road network segment does not exist in isolation. The road segments are regularly connected together, and the traffic flow propagates through the road carrier in a certain way. In order to predict the traffic flow of the road section, in addition to the law existing in the time dimension, it is also necessary to take the spatial correlation between the road section and the road section into account. In geography, spatiotemporal data refers to a collection of sets of time series with spatial correlation. Spatiotemporal modeling refers to the process of using time-space data to find an analytical method to model and predict the value of an attribute without observing the spatiotemporal position. Martin extends the spatial dimension based on the ARMA model and proposes a space-time autocorrelation moving average model (STARMA) for space-time prediction [1]. Weiguo Han used this model to predict regional short-term traffic flow, and verified the feasibility of STARMA model in short-term traffic flow prediction [2]. Although the model has good applicability to stationary spatiotemporal
sequences, most of the research objects of traffic flow are non-stationary. Although the different method can be used to deal with non-stationary problems, it also causes the loss of data information. At the same time, the model requires the same parameter values for all locations, so the model performance is poor.

In view of the above problems, this paper proposes a combination of generalized space-time autoregressive model (GSTAR) and support vector machine model (SVM) for short-term traffic flow prediction. Wavelet transform has great advantages for nonlinear and non-stationary data processing. It is applied to spatiotemporal sequence analysis, which helps to separate trend items and stochastic fluctuation items, which provides the possibility to establish a combined model. Firstly, the traffic stream data is decomposed into new sequences of different frequency features by wavelet transform. After reconstruction, the two types of models are used to predict respectively. The low-frequency trend components are selected by SVM model prediction, the high-frequency wave components are predicted by GSTAR model. Finally, add up to get the final forecast. In this paper, the simulation experiment is carried out based on an actual case. The experimental results show that the model has higher prediction accuracy.

2. Wavelet transform
The traditional Fourier transform process uses a series of different frequencies to superimpose the sine or cosine signals of the initial phase to maximize the approximation of the original signal. This approach has a great drawback of using only regular smooth signals for approximation. In order to improve this defect, a new processing method based on Fourier transform-wavelet transform has been proposed in recent years [3]. This transformation method is gradually multi-scale refinement through expansion and translation changes, which makes it have outstanding local optimization properties in both time-frequency and spatial frequency domains. It can be applied to spatio-temporal sequence modeling analysis to separate low-frequency trend components and high-frequency fluctuation components, which can be used for combined model prediction to obtain better approximation effects.

In wavelet transform, the function is the mother wavelet function, which satisfies the condition:

\[ \int_{-\infty}^{\infty} \varphi(t) = 0 \]  \hspace{1cm} (1)

Or satisfying the equivalence condition after Fourier transform:

\[ \int_{0}^{\infty} \left| \Psi(\omega) \right| \frac{d\omega}{\omega} < \infty \]  \hspace{1cm} (2)

The function \( \varphi(t) \) can obtain the wavelet basis function after translation and scale transformation, which is expressed as:

\[ \varphi_{a,b}(t) = \left| a \right|^{\frac{1}{2}} \left( \frac{t-b}{a} \right) \]  \hspace{1cm} (3)

In the formula, \( a \) is the scale factor, \( b \) is the translation factor, and the change of the scale factor can compress the mother wavelet function. When \( a > 1 \), the stretch transformation is performed, and the larger the value is, the greater the degree of stretching will be. The translation factor controls the movement of the mother wavelet function over the time axis. During the transformation of the two factors, the mother wavelet function generates a series of irregular wavelet basis functions.

3. Method of prediction

3.1. GSTAR model
In 1975, Martin proposed a space-time autocorrelation function and a space-time partial correlation function to measure the correlation of space-time, and based on this, he proposed a space-time autocorrelation moving average model (STARMA). The model extends the spatial dimension on the basis of ARMA, and considers the mutual influence of spatial neighbors and realizes the real integration of time and space. Because the model fully considers the characteristics of geospatial data such as space-
time autocorrelation, it is more suitable for analyzing and processing geospatial data. However, STARMA requires that the values of the parameters be the same for all locations. Based on this, Borovkova introduced a more flexible model, namely the generalized space-time autoregressive model [4]. This model allows model parameters to vary from spatial location to more realistic location.

The formula for the $GSTAR(p_1)$ model is:

$$Z(t) = \sum_{k=1}^{p_1} (\Phi_{k_1} + \Phi_{k_2} W) Z(t-k) + e(t)$$

(4)

Where in, $\Phi_{k_1} = diag(\phi_{k_1}^1, ..., \phi_{k_1}^N)$, $\Phi_{k_2} = diag(\phi_{k_2}^1, ..., \phi_{k_2}^N)$, the spatial weight matrix needs to satisfy:

$$w_i = 0, \sum_{j=1}^{N} w_j = 1$$

(5)

As an example, a three-position based on spatiotemporal sequence $GSTAR(1)$ model can be written:

$$Z(t) = [\Phi_{k_0} + \Phi_{k_1} W] Z(t-1) + e(t)$$

(6)

Similar to ARMA modeling, GSTAR also uses a three-stage modeling process, namely model identification, parameter estimation and model checking. Model identification is to determine the specific form of GSTAR performance by calculating STACF (space-time autocorrelation function) and STPACF (time-space bias autocorrelation function) to determine the specific form of GSTAR performance. Parameter estimation generally uses least square estimation parameters, using estimated parameters to evaluate the correctness of the model. The model test refers to test the residual after fitting the model. STACF (space-time autocorrelation function) can be used to check whether the residual is a random error.

3.2. SVM model

In 1995, Vapnik proposed a machine learning method for solving multidimensional function prediction—support vector machine [5]. The learning strategy of this method is interval maximization, which is essentially an optimization algorithm for solving convex quadratic programming. The support vector machine constructs the optimal classification hyperplane in high-dimensional space by means of kernel function, which realizes linear separability. The method is suitable for small sample and nonlinear data, and has the advantages of global optimization and small structural risk. The kernel functions of support vector machines commonly used are linear kernel function, polynomial kernel function, exponential radial kernel function, Gaussian radial kernel function, etc [6]. The Gaussian radial kernel function is chosen in this paper.

The basic model of the support vector machine is $f(x) = \omega^T \phi(x) + b$, where $\omega$ is the weight vector and $b$ is the offset vector, both are parameters for determining the model. $\phi(x)$ is nonlinear mapping function for introducing slack variables in the base model, SVM can be formalized as:

$$\min \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*)$$

$$f(x_i) - y_i \leq \varepsilon + \xi_i$$

$$y_i - f(x_i) \leq \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0, i = 1, 2, ..., m$$

(7)

Where $C$ is the penalty parameter, $\xi_i, \xi_i^*$ are relaxation variables, $\varepsilon$ is the insensitivity factor.
3.3. Combined forecasting model

The modeling steps of the combined model in this paper are as follows: wavelet transform decomposition, which decomposes the traffic flow data into new sequences of different frequency features and reconstructs them. Perform a stationarity test on the new sequence to see if it can satisfy the construction conditions of the spatiotemporal sequence model. The stationary wavelet transform sequence is predicted using GSTAR, and the non-stationary sequence is predicted using SVM model.

In order to test the prediction performance of the model, we use the MAE, MSPE, RMSE and MAPE to evaluate the fitting and prediction accuracy of the model [7].

4. Experiment and Analysis

4.1. Experimental Scene and Data Processing

The experimental data is derived from the real-time data of the expressway collected by the traffic control department of Ma’anshan. A total of 26 sections in the road network of Yushan Road, Yinshan Road, Jiuhua Road, Hongqi Road, Huxi Road and Hudong Road in Ma’anshan are selected for experiment. The traffic flow in this area is stable and suitable for predictive analysis experiments. Select the traffic volume from December 7 to December 20, 2015 as the experimental data, predict the traffic flow on December 21, 2015, compare the actual data with the forecast results and use the performance indicators mentioned above, then determine the accuracy of the forecast results. Past results show that 15 minutes is a suitable time period for traffic flow prediction, so the time interval is 15 minutes and the statistical time is 7:30-18:15.

The topological relationship of the urban road network reflects the connection relationship between the various road sections. The relationship between any two road sections is two. One is directly adjacent to a road intersection, and the other goes through multiple other road sections and roads. The intersections are indirectly adjacent. Urban traffic road networks can be abstracted as undirected or directed graphs. If only the static network topology is considered, it is an undirected graph; if a road traffic flow is considered, it is a directed graph. As shown in the figure, there is a directed graph of the traffic flow direction of the upper and lower sections of the study area.

![Traffic flow directed graph](image.png)

Figure 1. Traffic flow directed graph

The spatial weight matrix is to express the adjacency relationship between spatial objects from the network topology. According to the influence characteristics of traffic flow, this experiment establishes the first-order and second-order adjacency matrix of the traffic network. In order to reflect the degree of influence between the road segment and each adjacent road segment, the row-standardization method is used to convert the spatial adjacency matrix into the spatial weight matrix.
4.2. Traffic Flow Decomposition
The most important thing in wavelet transform is to choose an appropriate base wavelet. The $\text{dnN}$ wavelet has good decomposition characteristics to deal with non-stationary spatiotemporal sequences. The wavelet has the characteristics of time-frequency tight support, orthogonality and high regularity, which is suitable for traffic flow. Decomposition of spatiotemporal sequences. At the same time, the selection of the number of decomposition layers should be appropriate. It is necessary to select the appropriate number of decomposition layers for many experiments. The random decomposition can not extract the characteristics of high frequency information and low frequency information, which has an impact on the prediction results. As for comprehensive experimental analysis, this paper chooses The $\text{dbN}$ wavelet with order $N=4$ performs the second-order wavelet coefficient decomposition. After selecting the base wavelet, wavelet decomposition of the traffic time-space sequence of each road segment is performed, and the decomposition result is reconstructed. The figure shows the results of the decomposition and reconstruction of the road segment 1. The decomposition sequence 2 is the low frequency part of the sequence, and the curve shape of the original sequence is maintained; the decomposition sequence 1 is the high frequency part of the sequence.

![Traffic flow decomposition chart]

Figure 2. Traffic flow decomposition chart

4.3. Establish GSTAR Model
The intuitive meaning of a stationary spatiotemporal sequence is that there is no trend (or pattern) in the sequence. The statistical significance is that the mean, variance, covariance, etc. do not change as time and spatial position changes. Direct modeling of non-stationary models can lead to pseudo-regression, so it is firstly necessary to determine whether the data is stable before modeling. For space-time variables, consider time-stable first. In general, stationarity is too strict for spatial variables. There is only theoretical feasibility and no application value. Therefore, only second-order stationary or intrinsic assumptions are required. There are many methods to determine whether the sequence is stable. It is common to judge the ACF and PACF and the ADF root test according to the sequence [8]. The ACF and PACF graphs require certain experience and have certain subjectivity. Therefore, this paper chooses to use the ADF root test. In the unit root ADF test, the ADF statistic = -6.3214, the corresponding p value <0.05, indicating that the sequence is stationary. The ADF root checklist is shown as below:

| Augmented Dickey-Fuller test statistic | t-Statistic | Prob* |
|---------------------------------------|------------|-------|
| Test critical values                  |            |       |
| 1% level                              | -3.4212    |       |
| 5% level                              | -2.8521    |       |
| 10% level                             | -2.7631    |       |

|                      | Augmented Dickey-Fuller test statistic | t-Statistic | Prob* |
|----------------------|---------------------------------------|------------|-------|
| Test critical values  |                                       |            |       |
| 1% level              |                                       | -6.3214    | 0.0004|

In order to determine the form of the GSTAR model, it is necessary to calculate a spatiotemporal autocorrelation function (ACF) and a spatiotemporal partial correlation function (PACF) to identify the autocorrelation parameter $p$. 

Table 1. ADF root checklist
From the table analysis, the space-time autocorrelation coefficient is decremented and smeared, and the partial correlation coefficient tends to zero after the 1st order spatial delays and the 1st time delays, then it can be preliminarily judged that it is the 1st time for the sample to delay and for the 1st order space to delay. The delayed autocorrelation process, the selected model is the space-time autocorrelation model GSTAR (1), and its specific expression is:

$$Z(t) = [\phi_0 + \phi_1 W] Z(t-1) + e(t)$$

4.4. Establish SVM model
The SVM prediction model is used to model the non-stationary data after wavelet decomposition. Before modeling with SVM, first you need to determine the type of kernel function and the size of the optimal parameters. This paper chooses the Gaussian Radial Basis (RBF) function, and the regression machine type is E-SVR. There are many methods for parameter optimization, such as K-CV, genetic algorithm, empirical selection method, gradient descent method and cross-validation method. K-CV is the best way to combine speed and accuracy. The minimum value of K is chosen to be 2, but it should not be too large. Excessive K value will lead to over-learning and over-fitting. Generally, in case of big data, in order to ensure the generalization of learning, K is usually set to 10. After many experiments, we find that in the case of K=3, the operation speed is acceptable and the accuracy is also high. The SVM is trained on the training set with the obtained optimal parameters to obtain the prediction result.

4.5. Combined model accuracy
Combine the above single model results, and use each single model to predict 26 traffic flows on December 21, 2015, compare the prediction accuracy of the combined model with a single model. Due to the space limitations, this paper only gives a comparison of the actual and predicted values of the three sections on December 21, 2015.

Figure 3. Comparison of actual and predicted values

The prediction accuracy of the three models is evaluated by using the above-mentioned accuracy evaluation indexes. It can be seen from the table that the four precision indexes of MAE, MSPE, MAPE and RMSE of the combined model are significantly improved.

| Model       | MAE   | MSPE  | MAPE  | RMSE  |
|-------------|-------|-------|-------|-------|
| GSTAR       | 1.6353| 0.0030| 0.0064| 1.7943|
| SVM         | 1.8104| 0.0032| 0.0054| 1.9136|
| Combined model | 1.1678| 0.0019| 0.0032| 1.1529|

5. Conclusion
Based on the above analysis of the results, we can see that:
Traffic flow data is a non-linear and non-stationary spatiotemporal sequence. Through wavelet transform, SVM is used to predict its trend term, and GSTAR model is used to predict its random term. The superposition of the predicted values is used as the final prediction result, which can truly reflect the actual situation of traffic flow. This method combines the advantages of SVM and GSTAR to make up for the limitations of a single prediction model with better prediction and accuracy.

In the GSTAR model we can take both time and space characteristics into account when dealing with traffic data. This model allows model parameters to be more flexible due to spatial location. However, the accuracy of GSTAR model in processing nonstationary spatiotemporal data prediction is not high, so wavelet decomposition processing is applied, which is processed into a stable spatiotemporal sequence, and verified by relevant data.

As a method of data mining, GSTAR-SVM hybrid model is not only suitable for traffic flow prediction, but also suitable for many nonlinear fields of spatiotemporal prediction. It is worthy of popularization and application.

References
[1] Martin R L, Oeppen J E. (1975) The identification of regional forecasting models using space-time correlation functions. Transactions of the Institute of British Geographers, 66:95-118.
[2] Han, W.G. Forecasting and Analysis of Regional Traffic Flow in Space and Time. Journal of Highway and Transportation Research and Development, 24:93-96
[3] Gilles J. Empirical wavelet transform. IEEE Transactions on Signal Processing, 2013,61(16):3999-4010.
[4] S. Borovkova, H. P. Lopuhaa and B. N. Ruchjana, Consistency and Asymptotic Normality of Least Squares Estimator in Generalized STAR Models. Statistica Neerlandica, nr.4, pp. 482-508 (2008).
[5] Vapnik V N. Statistical Learning Theory. NewYork: John Wi-ley and Sons,1998.
[6] O Chapelle, V Vapnik. Choosing multiple parameters for SupportVector Machines. [J]. MechineLearning, 2002, 46:131-150
[7] Taylor, Francis. Spatial Data Analysis: Theory and Practice. Journal of the American Statistical Association, 2005, Vol.100 (469), pp.353-353
[8] Econometric analysis. Economic science, Zhang X.T,2000