Research Article

The short-term prediction of the mobile communication traffic based on the product seasonal model

Li-Na Wang[1,2] · Chen- Rui Zang[3] · Yuan-Yuan Cheng[1]

Abstract

The short-term forecast can be used to respond to the unexpected business shocks in advance, thus guaranteeing user experience. We present a practical communication traffic forecasting technology based on autoregressive moving average model. The proposed short-term prediction method is mainly based on the product seasonal model. The orders of product seasonal ARMA equation are recognized by the Akaike information criterion, and the parameters of equation are estimated by the maximum likelihood method. The unit root test is used to judge the stability and reversibility of the model. The performance of the proposed method is evaluated. The experimental result shows that the method has high prediction accuracy. The short-term forecast can provide a basis for mobile operators to accurately expand capacity.

Keywords Mobile communication · Prediction · Time series · Product seasonal model

1 Introduction

Nowadays, mobile communication has become an inseparable part of human society. It serves to the public, makes people's life convenient and promotes the development of society. With the widespread use of mobile devices, a large number of high-resolution spatial and temporal data are collected. Scientists use different methods to analyze mobile communication data to study human behavior [1, 2], predict trends in mobile communication data [3], establish the traffic signal advisory system [4] and analyze the future development of mobile services [5, 6]. With the rapid development of mobile network and the universality of network services, the storage of traffic increases. In order to ensure the normal operation of mobile system and efficient use of business resources, traffic forecasting has become a hot research topic. We studied the mobile communication traffic data from four aspects: model determination, model recognition, testing and predicting.

Analyzing the traffic volume data, the optimal model is determined. Then, it is used to predict the situation of the future day. The prediction accurately depicts the actual double-peak trends, and the prediction accuracy is high. The mobile communication network contains a large amount of data. Mining mobile communication data can provide suggestions for operation optimization, protocol improvement and architecture design.

In 2013, a new wavelet multi-resolution analysis and prediction algorithm combined with Fourier spectral prior knowledge was proposed [7]. This method focuses on the long-term trend prediction of multi-period and non-fixed mobile communication series. In 2014, a time series prediction method based on the echo state network and multiplication seasonal autoregressive integrated moving average (ARIMA) model was proposed [3]. The empirical results show that the method has a good prediction effect. In 2015, Mao et al. [2] used the Ivorian mobile phone data to analyze the relationship between

the national communications network and the socioeconomic dynamics. In 2017, Ren et al. [8] constructed a social network based on the call records of China mobile service providers and proposed a new model to simulate the process of information dissemination. In the same year, the interaction between user demography and social behavior was simulated by modeling more than 7 million users and 1 billion mobile communication records [9]. Furthermore, the extent to which users' demographic data can be inferred from their mobile communication patterns is studied. In 2018, according to the characteristics of communication network performance indicators, a time series prediction algorithm based on big data was proposed [10]. By analyzing, fitting, modeling and forecasting massive data, the algorithm decomposes granularity value of the time series. In 2019, the fault data of mobile communication was analyzed to distinguish valid faults (faults caused by infrastructure problems) and invalid faults (faults caused by equipment defects or other problems), in order to achieve the purpose of filtering invalid faults [11].

The prediction of mobile communication data can provide a basis for future business planning. Time series prediction methods can be roughly summarized into three categories: (1) classical prediction methods based on statistics. Classical time series prediction models and their extensions have been applied in many fields, and the effectiveness has been proved by many research studies; (2) time series forecasting method based on artificial intelligence. The method incorporates the latest machine learning technology. The prediction accuracy is improved, but the operation is complicated; (3) the combined prediction method, combining the classical time series forecasting and machine learning technology. When analyzing and predicting the mobile communication traffic data, many methods can be used. However, the Elman neural network method is suitable for data with nonlinear and abrupt changes [12], the computation of traditional back-propagation (BP) neural network is complex [13], and ARIMA model is suitable for non-stationary sequence [14]. Because of the stationarity and periodicity of our data, we use the product seasonal ARMA method to model and predict in this paper.

The structure of this paper is as follows: The mobile communication traffic volume data is introduced in Sect. 2.1. The periodic and the doublet characteristics of time series are observed. Then, in the remainder of Sect. 2, the main methods are introduced, including the product seasonal model, unit root test and Ljung–Box test. We use these methods to judge the stability of the mobile communication time series, to test the significance of the parameters and to test the stationarity and the reversibility. In Sect. 3, the product seasonal ARMA model is used to analyze the mobile communication traffic data, according to three steps: model determination, model recognition and model test. Furthermore, a short-term prediction is given, and the performance of the method is evaluated. The experimental result shows that the method has high prediction accuracy. Finally, the conclusion and discussion are mentioned in Sect. 4.

2 Materials and methods

2.1 Data

The data comes from the voice traffic collected automatically by the base station of an operator in a city. Nearly 90,000 pieces of data were collected in six days, from 1:00 on February 22, 2017, to 24:00 on February 27, 2017. In terms of hours, a time-averaged series of voice traffic is obtained, as shown in Fig. 1a. The time series exhibits a bimodal feature, and the traffic volume varies with time. As shown in Fig. 1b, the first peak of the day appears at 10:00 and 12:00; the second peak of the day appears at 18:00; the trough between the peaks appears at 14:00; before 8 o'clock, the traffic is small; and the traffic begins to decrease sharply at 20 o'clock. In addition, the traffic volume on February 25 and February 26 is significantly lower, which coincides with Saturday and Sunday, and the traffic on February 27 is the highest, which is Monday. Overall, the traffic data exhibits the following: traffic on Monday > traffic on Friday > traffic on others (Tuesday, Wednesday or Thursday) > traffic on weekends.

Fig. 1 Time series diagram of mobile communication traffic volume
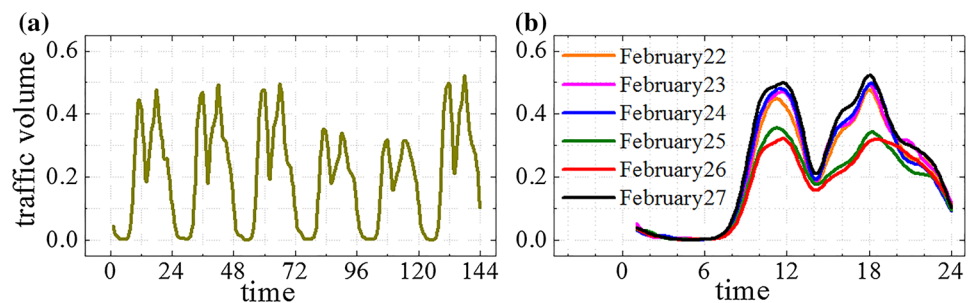
The time series data is divided into two parts: The first part is used as a training set with a length of 120, from 1:00 on February 22 to 24:00 on February 26; the second part is used as a test set with a length of 24, from 1:00 on February 27 to 24:00 on February 27. In this paper, we will analyze the mobile communication traffic volume of the first 120 h, construct the optimal model, test the applicability of the model and then predict the change trend of the future 24 h.

2.2 Product seasonal model

Autoregression and moving average model (ARMA model) is one of the most important methods for analyzing the time series. The model can be applied in wide field and has the characteristic of the small prediction error. For the stationary time series, the implementation process for autoregressive model (AR model) and moving average model (MA model) is relatively easy. However, in many cases, in order to fit the dynamic process of the data adequately, the problems of too many parameters or too complex models will arise, if you just use one model (AR model or MA model). In order to avoid these problems, ARMA model is presented. It is a special combination of the AR model and the MA model. The ARMA model not only fits the stationary time series accurately, but also reduces the number of parameters. ARMA model and its combination models have been widely used in many different fields, such as prediction [15, 16], electricity consumption [17–19], hydrological research [20] and so on.

The time series is a sequence of data collected at a certain time interval, that is, $\{x_t, t = 1, 2, \ldots, n\}$, which is considered to be a stationary time series when the mean and the variance do not change with time. Otherwise, it is a non-stationary time series. According to different series structures, the model AR(p), MA(q) or ARMA(p, q) can be used to predict the stationary time series. For non-stationary time series, the difference method is usually used to make it a stationary one. And then, the stationary series analysis can be used to predict.

The model ARMA(p, q)

$$
\begin{aligned}
x_t = {} & \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} \\
& + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}
\end{aligned}
\tag{1}
$$

is a special combination of two models, model AR(p)

$$
x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t
\tag{2}
$$

and model MA(q)

$$
x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}.
\tag{3}
$$

where ε_t is defined as white noise series with the mean 0 and the variance σ^2; p and q are, respectively, taken as the order of AR model and MA model; $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the autoregressive function, and $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the moving average function.

Considering the lag operator B, the AR(p) model can be rewritten as $\phi(B) \cdot x_t = \varepsilon_t$, where

$$
\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p
\tag{4}
$$

is p-order autoregressive coefficient polynomial. And $\phi(B) = 0$ is the characteristic equation of AR(p) model. Making use of the lag operator B, the MA(q) model can be rewritten as $x_t = \theta(B) \cdot \varepsilon_t$, where

$$
\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q
\tag{5}
$$

is q-order moving average coefficient polynomial. And $\theta(B) = 0$ is the characteristic equation of MA(q) model. Thus, after introducing the lag operator B, the ARMA(p, q) model can be denoted as

$$
\phi(B) \cdot x_t = \theta(B) \cdot \varepsilon_t.
\tag{6}
$$

where $\phi(B)$ and $\theta(B)$ are as shown in Eqs. (4) and (5), respectively.

If a time series shows similarity after s time intervals, this series has the periodicity characteristic. Here s is the periodic length. A series with the periodicity characteristic is also called as a seasonal time series. A simple seasonal model is to add periodic effects to other effects. That is $x_t = S_t + T_t + I_t$, where S_t is the periodic change, T_t is the trend item and I_t is the random fluctuation item. However, in most cases, seasonal effects and short-term correlations cannot be simply extracted by additive. It is necessary to consider the short-term correlations and the replication effect of seasonal effects. It is considered that the seasonal effect and the short-term correlations are product relations. The product seasonal model is actually the product of ARMA (p, q) and ARMA (P, Q)$_s$, denoted as the product seasonal ARMA process ARMA(p, q) $\times$ (P, Q)$_s$, where the right subscript s indicates the period. Similar to Eq. (6), the product seasonal model ARMA(p, q) $\times$ (P, Q)$_s$ is written as

$$
\Phi\big(B^s\big)\phi(B)x_t = \Theta\big(B^s\big)\theta(B)\varepsilon_t.
\tag{7}
$$

where there are

$$
\begin{cases}
\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \\
\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \\
\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \\
\Theta(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}
\end{cases}
\tag{8}
$$

and s is the periodic length. The first equation in Eq. (8) is the same as Eq. (4), and the third equation in Eq. (8) is the same as Eq. (5).

2.3 Unit root test

The unit root test can be used to judge the stability and reversibility of the model. For the ARMA model, the unit root means roots of the characteristic equation of the AR model and roots of the characteristic equation of the MA model. For the AR model, when $|\phi| < 1$, $D(x_t) = 0$, $E(x_t) = C < \infty$, $\text{Cov}(x_t, x_{t-k}) = C < \infty$ can be obtained. The condition of stationarity is satisfied. That means the time series is stationary. In this case, the absolute value of the root of the characteristic equation is greater than 1. Thus, when $|\phi| < 1$ or the absolute value of the root of the characteristic equation is greater than 1, the time series is considered as stationary. On the contrary, the time series is considered to be unstable.

For the MA model, when $|\theta| < 1$, the equation is reversible. In this case, the absolute value of the root of the characteristic equation is greater than 1. Thus, when $|\theta| < 1$ or the absolute value of the root of the characteristic equation is greater than 1, the time series is considered to be as reversible. On the contrary, the time series is considered to be irreversible. The existence of unit roots in the autoregressive part indicates that the model is unstable. That is, the model does not have the trend of returning to the specified values over time. The existence of unit roots in the moving average part indicates that the sequence is irreversible. That is, the sequence cannot be represented as an autoregressive equation of a series of deviated observations.

2.4 Ljung–Box test

The Ljung–Box test is mainly used to test the residual sequence of the fitting model. According to the test results, the sufficiency of extracting information from the model is analyzed. Null hypothesis and alternative hypothesis are set as follows: H_0: $\rho_1 = \rho_2 = \cdots = \rho_m = 0$ versus H_1: at least one $\rho_i \neq 0$ ($i = 1, 2, \ldots, m$). The test statistic is given as

$$Q(m) = N(N + 2) \sum_{i=1}^{m} \frac{\hat{\rho}_i^2}{N - i} \tag{9}$$

where N denotes the amount of data, $\hat{\rho}_i$ represents the autocorrelation value at lag i and m is the value of the detected lag period. $Q(m)$ follows a Chi-square distribution. The null hypothesis will be accepted, when the p value is larger than the significant level α (the default $\alpha = 0.05$). At this time, the residual sequence is considered as white noise sequence.

The Ljung–Box test has been widely used in time series analysis and other fields. Xing et al. [21] used a flat gene filter based on the Ljung–Box test to screen and identify differentially expressed genes in biological experiments. This method helps to understand gene functions and to identify key genes at specific stages of plant development. Chen et al. [22] proposed a blind robust detection method based on Ljung–Box test. The simulation results show that the method achieves significant improvement in detection performance when the correlation between receiving antennas is low. Bogusz et al. [23] used the least squares method, the median absolute deviation criterion and the t test, respectively, to remove the deterministic part of the time series. Then, the Ljung–Box test is used to test the residual sequence and prove the self-similarity of the stochastic part of the GPS time series.

3 Results

The statistical characteristics of mobile communication traffic data can reflect the economic situation and human dynamics in a certain area. In this section, we analyze the traffic volume data of 120 h and then construct the optimal model and test it. Finally, the change trend of the future twenty-four hours is predicted. The product seasonal ARMA model is used to analyze the mobile communication traffic data and make a short-term prediction, according to the following four steps:

Step 1 model determination. The stability and the seasonality of time series are judged according to the ACF (autocorrelation function) graph, the PACF (partial autocorrelation function) graph and the ADF (augmented Dickey–Fuller) test. The product seasonal model is determined.

Step 2 model recognition. The order and the parameters of the seasonal ARMA model are recognized by the EACF (extended autocorrelation function) and the AIC (Akaike information criterion). The product seasonal ARMA equation is obtained.

Step 3 model test. The applicability of the model is detected from three aspects: significance test of the parameters, stationarity and reversibility test and the residual sequence test. That means the equation needs to satisfy the following: the parameter is significant, the inverse for the root of the characteristic equation is within the unit circle, and the residual sequence is a white noise one.

Step 4 prediction. The mobile communication traffic volumes of the future twenty-four hours are predicted, and the prediction error is measured with the mean-square error (MSE) and the mean absolute error (MAE). This section may be divided by subheadings. It should provide

a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

3.1 Model determination

In general, the data correlation and ADF test are used to test the stationarity of time series. First, we apply the graph of ACF and PACF to initially determine the stationarity of data. Then, the ADF test method accurately tells us that the time series is stationarity. The autocorrelation function slowly attenuates in a regular sine form when the lag phase gradually increases, and the partial autocorrelation function slowly attenuates in an irregular sine form when the lag phase gradually increases. That is, the autocorrelation plot and the partial autocorrelation plot exhibit a trailing decay. So it is preliminarily judged that the time series is stationary.

Further, the exact results are calculated by ADF test, which is used to judge stationarity of the time series through the unit root detection. The original hypothesis will be rejected when statistics t is less than the significance level threshold, indicating that there is no unit root in the time series. It indicates that the time series is stationary. Otherwise, it is non-stationary. As shown in Table 1, the t value is obviously less than the critical value of the significance level of 1%, that is, $-5.2397 < -4.0405$. It holds that the time series is stationary at 99% level of confidence. To sum up, data correlation detection and ADF test indicate that the time series is stationary. In addition, the correlation of ACF and PACF diagrams is very strong at the time points when the lag interval is 12 times. That is the typical performance of the periodicity of time series, which indicates that the time series has seasonal characteristics. In summary, the time series satisfies stationarity and seasonality.

3.2 Model recognition

Because of the stationarity and seasonality of the time series, the product seasonal ARMA model is chosen. The product seasonal ARMA model is a hybrid model of non-seasonal model ARMA(p, q) and seasonal model ARMA(P, Q)$_s$. The order p and q of the non-seasonal model are determined by the EACF. The order P and Q of the seasonal model are more complex, but not larger than 2 usually. Periodic s is determined by the characteristics and actual conditions of the data. The EACF contains triangles made up of "0," and the position of the upper left corner of the triangle is defined as (p, q). For the non-seasonal model ARMA(p, q), $p=0$, $q=4$ is obtained. For the seasonal ARMA(P, Q)$_s$ model, P and Q take various combinations of values not exceeding 2. Moreover, the period $s=24$ is determined by the characteristics of the data. According to the AIC shown in Table 2, the best one is established as product seasonal ARMA(0, 4)$\times$(1, 0)$_{24}$ in a number of possible fitted models.

The parameters and related statistical information of the optimal model ARMA(0, 4)$\times$(1, 0)$_{24}$ are given in Table 3, according to maximum likelihood estimation. After taking two decimals, the model equation is as follows:

$$\left(1 - 0.91B^{24}\right)x_t = \left(1 + 1.01B + 0.92B^2 + 0.82B^3 + 0.25B^4\right)\varepsilon_t \tag{10}$$

3.3 Model test

Model diagnosis is used to verify the fitting sufficiency of the model ARMA(0, 4)$\times$(1, 0)$_{24}$. The following will be conducted from three aspects: coefficient significance test,

Table 3 Test table of the model

	Estimation	SE	t value
ma1	1.0120	0.0959	10.5527
ma2	0.9158	0.1056	8.6723
ma3	0.8152	0.1061	7.6833
ma4	0.2490	0.1034	2.4081
sar1	0.9075	0.0266	34.1165

Estimation represents the estimated value, SE represents the estimation error, and t value represents the t statistic value

Table 1 ADF test for the time series

Augmented Dickey–Fuller test	t statistic	Prob
	−5.2397	0.0002
Test critical values		
1% level	−4.0405	
5% level	−3.4497	
10% level	−3.1501	

Table 2 AIC value table for several fitting models

p	q	P	Q	AIC	p	q	P	Q	AIC	p	q	P	Q	AIC
0	4	0	2	−426.87	0	4	0	1	−392.84					
0	4	1	2	−490.76	0	4	1	1	−488.58	0	4	1	0	−506.94
0	4	2	2	−489.65	0	4	2	1	−488.60	0	4	2	0	−491.41

stationarity and reversibility test and residual sequence test.

3.3.1 Significance test

Coefficient test refers to detecting the significance of each coefficient by t test. The main function is to test whether coefficient of the variable (or control variable) is significant at a specific level of significance. Significant results can indicate that the variables have obvious explanatory power. Otherwise, it is not. Null hypothesis and alternative hypothesis are set as follows: H_0: $\mu = 0$ versus H_1: $\mu \neq 0$, and the test statistic is given as

$$t = \frac{\hat{x}}{s / \sqrt{n}}, \tag{11}$$

where $\hat{x}$ represents the estimated value and s represents the standard deviation. $|t| > 2$ indicates that the coefficient of the variable is significantly not equal to 0. That means the variable has a strong explanatory power for explanatory variables. On the contrary, if $|t| \leq 2$, this variable has almost no explanatory power. Then, we can choose to remove this variable, which not only optimizes the model but also simplifies the calculation. As shown in Table 3 above, the significance of the coefficient in this paper is tested by the t test. In the last column, all t values satisfy $|t| > 2$. So it proved that these variables have strong explanatory to the model.

3.3.2 Stationarity and reversibility test

Stability and reversibility of the model are judged using the unit root test. The stationary condition of model is that inverted roots of function $\Phi(B_s)\phi(B) = 0$ are within the unit circle. And the reversible condition of model is that inverted root of function $\Theta(B_s)\theta(B) = 0$ is within the unit circle. Here, the distribution of the eigenvalue for the model ARMA$(0, 4) \times (1, 0)_{24}$ is detected. The inverted root of characteristic equation is shown in Table 4. The former twenty-four values are inverted roots of $1 - 0.91B^{24} = 0$, and the latter four values are inverted roots of $1 + 1.01B + 0.92B^2 + 0.82B^3 + 0.25B^4 = 0$. The points in Fig. 2 correspond to the values in Table 4. The blue hollow circles represent
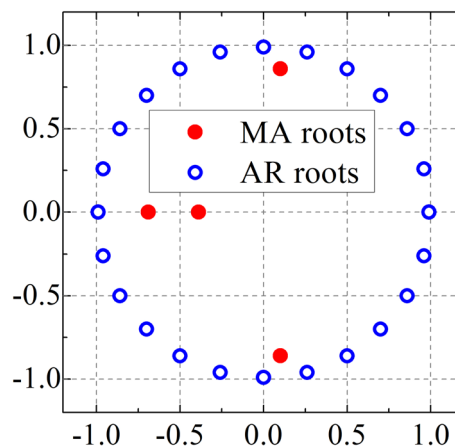

Fig. 2 Distribution diagram of the inverted roots for the model ARMA$(0, 4) \times (1, 0)_{24}$

reciprocal eigenvalues of the autoregressive polynomial, and the red solid circles represent reciprocal eigenvalues of the moving average polynomial. Each root is within the unit circle, which indicates that the model ARMA$(0, 4) \times (1, 0)_{24}$ has the stability and reversibility.

3.3.3 Residual sequence test

The residual sequence can test the sufficiency of the model fitting. The model is considered to be well fitted if the residual sequence is close to the white noise sequence. Using the ARMA$(0, 4) \times (1, 0)_{24}$ to fit the data on former 120 h, the residual sequence with mean value 0.0001 and variance 0.0005 is obtained. The residual sequence has almost non-correlation and non-biased autocorrelation. It is preliminarily judged that the residual sequence is close to the white noise sequence. Furthermore, the accurate results are detected by Ljung–Box test. The p value equals to 0.9364, greater than 0.05. The Ljung–Box test result tells us that the residual sequence is considered as a white noise sequence.

3.4 Data prediction

In order to measure the prediction effect, the MSE and MAE are used as evaluation performance indicators.

Table 4 Distribution of the inverted roots for the model ARMA$(0, 4) \times (1, 0)_{24}$

Inverted AR roots	0.99	$0.96 - 0.26i$	$0.96 + 0.26i$	$0.86 + 0.50i$	$0.86 - 0.50i$	$0.70 + 0.70i$
	$0.70 - 0.70i$	$0.50 - 0.86i$	$0.50 + 0.86i$	$0.26 - 0.96i$	$0.26 + 0.96i$	$0.00 - 0.99i$
	$0.00 + 0.99i$	$-0.26 - 0.96i$	$-0.26 + 0.96i$	$-0.50 + 0.86i$	$-0.50 - 0.86i$	$-0.70 + 0.70i$
	$-0.70 - 0.70i$	$-0.86 - 0.50i$	$-0.86 + 0.50i$	$-0.96 + 0.26i$	$-0.96 - 0.26i$	-0.99
Inverted MA roots	$0.10 + 0.86i$	$0.10 - 0.86i$	-0.39	-0.69		

They are often used to evaluate prediction accuracy. For instance, the prediction accuracy of weather forecast is evaluated [24], the prediction performance of surface temperature prediction model is evaluated [25], and the effect of tuberculosis incidence prediction model is evaluated [26]. MSE is defined by the mean of the sum of squared errors and can evaluate the degree of variation of data. The smaller value of MSE indicates that the accuracy of predictive model in describing experimental data performances is excellent. The occurrence of abnormal value will make the MSE value larger. MAE is the mean of the absolute error, which can reflect the actual situation of the prediction error better.

In this manuscript, the $ARMA(0, 4) \times (1, 0)_{24}$ is used to make short-term prediction for the next twenty-four-hour traffic, as shown in Fig. 3. The black box indicates the real value sequence, and the red ring indicates the prediction value sequence. The changing trend of the two curves is consistent, and the absolute value of error at each hour less than 0.1. In addition, the prediction accuracy is MSE = 0.0018 and MAE = 0.0424. From the error results, we can see that there is no abnormal value of the traffic volume, and the prediction deviation per hour is relatively small. The prediction accuracy is high. The model can accurately predict the future traffic volume.

The main purpose of modeling is to predict the possible situation in the future. Inferring the future based on the history and current situation is the basic idea for implementing predictions. Therefore, it is necessary to assume the history and current situation are representative or sustainable. The time series shows the history and current situation of random variables. Maintaining the basic characteristics of random variables requires the essential characteristics of time series to continue into the future. The mean, variance and covariance of time series can be used to describe its essential characteristics. Thus, the time series whose values of these statistics can remain unchanged in the future is stable. That is, if the time series generated by the random process satisfies the following: the mean is independent of time t, the variance is independent of time t, and the covariance is independent of

time t, then the time series is called as stationary. Vividly, the stationarity requires that the fitted curve can continue along the existing form within a period of time. It can be seen that the stationarity of time series is the basic assumption of classical regression analysis. The prediction based on stationary time series is effective.

The above prediction better shows the bimodal characteristics of time series (in Fig. 3). Furthermore, the predicted peaks are consistent with the actual traffic volume, appearing around 12:00 and 18:00, respectively. Observing the traffic volume time series in Fig. 1 directly, the similar characteristics appear in the time series after the same period interval. Obviously, the time series is periodic. The traffic volume is similar every twenty-four hours. Every day, the traffic volume starts to increase at 8:00 and reaches the maximum value at 11:00 and 12:00 for the first time; then, the traffic volume starts to fall and reaches the trough at 14:00; after that, it continues to increase and reaches the maximum value at 18:00 for the second time; then, the traffic volume drops almost linearly; between 0:00 and 8:00, the traffic volume is very small. Although the traffic volume on weekends is significantly lower than that on weekdays, it is still a bimodal distribution. This periodic bimodal feature of the time series reflects the characteristic of human communication behavior. The peak of communication traffic occurs at 12:00 and 18:00, corresponding to the noon peak and evening peak.

4 Discussion

The rapid development of mobile communication technology has promoted the explosive growth of the mobile communication market. With the rapid increase in the mobile communication network load, the mobile communication network will face greater challenges, for example the rapid growth of mobile communication data, the continuous increase in mobile network connections, the continuous emergence of new services, the diversification of business scenes and so on. Mining mobile communication data can provide suggestions for operation optimization, protocol development and architecture design. Prediction can provide support for decision making, and it is the basis for planning. The ARMA model and the modified ones are very important methods in the current prediction field. Based on the time series data from February 22 to February 26, the stability and the seasonality are obtained. By means of the parameter estimation and the hypothesis testing method, the optimal prediction model is determined to be $ARMA(0, 4) \times (1, 0)_{24}$, which has characteristics of significant parameters, stability, reversibility and white noise residual sequence. The significant parameters mean that variables give a strong explanatory
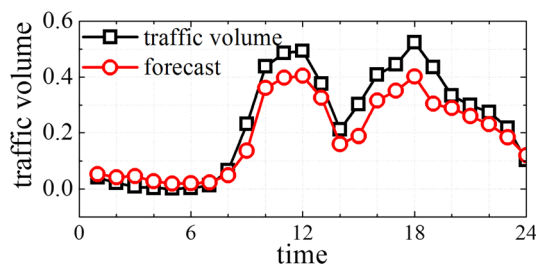
Fig. 3 Forecast sequence of traffic volume on February 27, 2017

to the model. The stability indicates that the mean and the variance of the model do not change with time and are less affected by the disturbance term. The reversibility tells that the model is convergent. The white noise residual sequence indicates that the model can extract relevant information sufficiently. A good prediction can minimize the impact of uncertainty on the object and provide a service for scientific decision making. The ARMA$(0, 4) \times (1, 0)_{24}$ is used to predict the situation of last twenty-four hours on February 27. The prediction curve accurately depicts the double-peak trend of the actual data. The performance of the proposed method is evaluated by MSE and MAE. The experimental result shows that the method has high prediction accuracy.

Using the time series analysis method, we studied the mobile communication traffic data. The effective and accurate prediction can provide operators with capacity expansion basis and resource allocation support. The capacity expansion refers to increasing base stations or increasing carrier frequencies. The carrier frequency of a base station can be understood as the capacity of the base station, which determines the processing capacity of the base station. When the predicted peak value is higher than the designed capacity, operators need to consider capacity expansion. In addition, operators can determine the scale of capacity expansion according to the predicted values. The prediction for hourly traffic volume can provide support for operators to allocate network resources reasonably. When the traffic volume is low, operators will close some network resources (such as carrier frequency resources and software resources) to achieve the purpose of saving energy and cost. According to the prediction results of time series, operators can accurately preset the time to turn on or off network resources. For example, operators can preset to turn on the network resources during the predicted peak periods such as 11 o'clock and 18 o'clock and preset to turn off the network resources during the predicted trough periods. In addition, operators can preset how many network resources to turn on or off according to the prediction value.

In this article, we use traffic volume of all cellular cells, whose geographical locations are different. Occasionally, accidental power failure may cause temporary equipment failure, resulting in part data loss in few geographical locations. When generating the hourly traffic volume time series, the traffic volume is defaulted to 0 to participate in the calculation if the traffic volume of a certain cellular cell is "none." We focus on the analysis and prediction of time series in this article. In the future, the spatiotemporal analysis and prediction of communication data is one of our focuses. The estimation of loss data in certain cellular cell can be completed by spatial data analysis methods. It may contribute to prediction when using spatial interpolation or spatiotemporal methods to estimate the loss data in some geographical locations. Furthermore, the spatiotemporal clustering is a work we are considering in another manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest

References

1. Büyükçorak S, Kurt GK, Toprakkiran G. User behavior modeling of voice communications: an empirical study. Wirel Commun Mobile Comput 2016, 16(1): 29–46. https://doi.org/10.1002/wcm.2491
2. Mao H, Shuai X, Ahn YY et al (2015) Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Cote d'Ivoire. EPJ Data Sci 4(1):15. https://doi.org/10.1140/epjds/s13688-015-0053-1
3. Peng Y, Lei M, Li JB et al (2014) A novel hybridization of echo state networks and multiplicative seasonal ARIMA model for mobile communication traffic series forecasting. Neural Comput Appl 24(3–4):883–890. https://doi.org/10.1007/s00521-012-1291-9
4. Joyoung L, Slobodan G, Branislav D et al (2017) Deployment and field evaluation of in-vehicle traffic signal advisory system (ITSAS). Information 8(3):72. https://doi.org/10.3390/info8030072
5. Wang H, Zhou Y, Sha W (2017) Research on wireless coverage area detection technology for 5G mobile communication networks. Int J Distrib Sens Netw 13(12):155014771774635. https://doi.org/10.1177/1550147717746352
6. Zeng Yu, Zhou T Hu H (2018) Weight based channel selection towards 5G in the unlicensed spectrum. China Commun 15(8):54–66. https://doi.org/10.1109/CC.2018.8438273
7. Peng Y, Lei M, Guo J et al (2013) Multiresolution analysis and forecasting of mobile communication traffic. Chin J Electron 22(2):373–376. https://doi.org/10.1016/j.image.2012.01.018
8. Ren F, Li SP, Liu C (2017) Information spreading on mobile communication networks: A new model that incorporates human behaviors. Phys A 469:334–341. https://doi.org/10.1016/j.physa.2016.11.027
9. Dong Y, Chawla NV, Tang J et al (2017) User modeling on demographic attributes in big mobile social networks. ACM Trans Inf Syst 35(4):1–33. https://doi.org/10.1145/3057278
10. Wang T, Wang M (2018) Communication network time series prediction algorithm based on big data method. Wirel Pers Commun 102(2):1041–1056. https://doi.org/10.1007/s11277-017-5138-7
11. Schwenk G, Pabst R, Mueller KR (2019) Classification of structured validation data using stateless and stateful features. Comput Commun 138:54–66. https://doi.org/10.1016/j.comcom.2019.02.007
12. Wu B, Duan T (2017) A performance comparison of neural networks in forecasting stock price trend. Int J Comput Intell Syst 10(1):336–346. https://doi.org/10.2991/ijcis.2017.10.1.23

13. Zhu H, Lian W, L L, et al. An improved forecasting method for photo-voltaic power based on adaptive BP neural network with a scrolling time window. Energies 2017, 10(10): 1542. DOI: 10.3390/en10101542

14. Lopez JC, Rider MJ, Wu Q (2019) Parsimonious short-term load forecasting for optimal operation planning of electrical distribution systems. IEEE Trans Power Syst 34(2):1427–1437. https://doi.org/10.1109/TPWRS.2018.2872388

15. Zhang L, Peng X (2016) Time series estimation of gas sensor baseline drift using ARMA and Kalman based models. Sens Rev 36(1):34–39. https://doi.org/10.1108/SR-05-2015-0073

16. Mohammadpour M (2017) A note on backward prediction for multivariate ARMA processes. Iran J Sci Technol Trans A Sci 41(A1):231–235. https://doi.org/10.1007/s40995-017-0207-z

17. Zhang Y, Li C, Li L (2017) Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods. Appl Energy 190:291–305. https://doi.org/10.1016/j.apenergy.2016.12.130

18. Bernardi M, Petrella L (2015) Multiple seasonal cycles forecasting model: the Italian electricity demand. Stat Methods Appl 24(4):671–695. https://doi.org/10.1007/s10260-015-0313-z

19. Xie T, Zhang G, Liu H et al (2018) A hybrid forecasting method for solar output power based on variational mode decomposition, deep belief networks and auto-regressive moving average. Appl Sci 8(10):1901–1909. https://doi.org/10.3390/app8101901

20. Mehdizadeh S, Sales AK, Tsakiris G (2018) A comparative study of autoregressive, autoregressive moving average, gene expression programming and Bayesian networks for estimating monthly stream flow. Water Resour Manag 32(15):1–22. https://doi.org/10.1007/s11269-018-1970-0

21. Xing L, Guo M, Liu X et al (2018) Identification and prioritization of differentially expressed genes for time-series gene expression data. Front Comput Sci 12(4):813–823. https://doi.org/10.1007/s11704-016-6287-7

22. Chen AZ, Shi ZP, He ZQ (2018) A robust blind detection algorithm for cognitive radio networks with correlated multiple antennas. IEEE Commun Lett 99:1–1. https://doi.org/10.1109/LCOMM.2017.2789184

23. Bogusz J, Klos A, Figurski M et al (2016) Investigation of long-range dependencies in the stochastic part of daily GPS solutions. Emp Surv Rev 48(347):140–147. https://doi.org/10.1179/1752270615Y.0000000022

24. Kendzierski S, Czernecki B, Kolendowicz L et al (2018) Air temperature forecasts' accuracy of selected short-term and long-term numerical weather prediction models over Poland. Geofizika 35(1):67–85. https://doi.org/10.15233/gfz.2018.35.5

25. Zhang X, Zhang Q, Zhang G et al (2018) A novel hybrid data-driven model for daily land surface temperature forecasting using long short-term memory neural network based on ensemble empirical mode decomposition. Int J Environ Res Public Health 15(5):1032. https://doi.org/10.3390/ijerph15051032

26. Adeboye A, Davies O, Akinwumi O et al (2016) Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape, South Africa, using a hybrid model. Int J Environ Res Public Health 13(8):757. https://doi.org/10.3390/ijerph13080757