Predicting LTE Throughput Using Traffic Time Series

Xin Dong[1], Wentao Fan[1], and Jun Gu[2]
(1. Beijing University of Posts and Telecommunications, Beijing, 100876, China;
2. ZTE Corporation, Shanghai 201203, China)

◀ Abstract

Throughput prediction is essential for congestion control and LTE network management. In this paper, the autoregressive integrated moving average (ARIMA) model and exponential smoothing model are used to predict the throughput in a single cell and whole region in an LTE network. The experimental results show that these two models perform differently in both scenarios. The ARIMA model is better than the exponential smoothing model for predicting throughput on weekdays in a whole region. The exponential smoothing model is better than the ARIMA model for predicting throughput on weekends in a whole region. The exponential smoothing model is better than the ARIMA model for predicting throughput in a single cell. In these two LTE network scenarios, throughput prediction based on traffic time series leads to more efficient resource management and better QoS.

◀ Keywords

ARIMA; exponential smoothing method; throughput prediction

1 Introduction

In recent years, there is a trend towards users accessing the Internet from a variety of applications and without restriction in terms of geographic location. This has resulted in an exponential increase of wireless traffic. In 2012, global wireless data traffic grew 70 percent year on year [1]. Thus, mobile network operators have to make a use of limited resources to meet ever-increasing traffic demands. To plan and run networks efficiently, it is important to understand the statistical characteristics of data traffic by analyzing the real traffic.

In [2], the authors use the throughput measured from a real-work cellular network to statistically model time-varying throughput per cell and the distribution of instantaneous throughput per cell over different cells. The proposed statistical models can be used to simulate the time-varying and location-varying throughput of cells. In [3], the authors analyze several widely accepted throughput network-performance indicators in LTE. Their analysis is based on counters and call traces of a live network. However, neither [1] nor [2] describe a scenario where throughput in a whole region changes over time. In [4], the authors estimate this throughput using a formula that expresses the behavior of TCP throughput. We consider throughput data as a time series that can be predicted using data measured in the past.

In this paper, we consider two practical scenarios: whole region and single cell. In the first scenario, we constructed a better model than both the individual ARIMA model and exponential smoothing model for predicting downlink throughput on weekdays and weekends in a whole region. In the second scenario, the traffic load in a single cell is uncertain and varying over time. We construct a model for predicting the instantaneous downlink throughput in a single cell of a large urban cellular network.

2 Data Set and Modeling Methodology

2.1 Data Description

Our data set includes records of Internet downloads and uploads in Hong Kong. The data was collected from 1352 cell sites across the city over 21 days between February and March 2014. Each data session includes the throughput of the downlink and uplink, timestamp, and cell ID. Each cell ID is also associated with the GPS coordinates of the corresponding cell. In this paper, LTE throughput is modeled as a time series and then predicted using an ARIMA model and exponential smoothing method.

2.2 Time Series Analysis

Time series data is an important class of data. Any change of an attribute value as a function of time can be considered time series data. Such data may derive from the atmosphere, commodity production, geography, sensors, the stock market, or inventory control. The throughput data in an LTE network can also be viewed as a time series. Prediction of time series is based on the idea that historical data related to past behavior can be used to predict the future behavior.

2.2.1 ARIMA Model

The autoregressive integrated moving average (ARIMA) model was introduced by Box-Jenkins [5]. ARIMA (p, d, q) is an autoregressive moving average (ARMA) model based on differenced time series data. The original time series data is differenced on the order d to make the data stationary. A stationary time series can be modeled as an ARMA model of order (p, q), where p is the order of the AR process and q is the order of

Predicting LTE Throughput Using Traffic Time Series
Xin Dong, Wentao Fan, and Jun Gu

the MA process. ARMA-modeled current time series data is given by:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + b_1 e_{t-1} + \\ b_2 e_{t-2} + \ldots + b_q e_{t-q} + e_t \tag{1}$$

where $y_{t-1}, y_{t-2} \cdots y_{t-p}$ are the data at past time points, $e_{t-1}, e_{t-2} \cdots e_{t-2}$ are the errors at past time points, e_t is a present error (ARMA assumes this error is Gaussian-distributed), $a_1, a_2, \ldots, a_p$ are the AR coefficients, and $b_1, b_2, \ldots, b_q$ are the MA model coefficients [6].

ARIMA (p, d, q) modeling involves making the data stationary, then identifying suitable values for the model order, then predicting the time series data from the model.

2.2.2 Exponential Smoothing Model

Exponential smoothing is a trend-analysis and prediction method based on the moving average method. Exponential smoothing method has three main submethods—linear exponential smoothing, secondary exponential smoothing and cubic exponential smoothing—that differ in terms of smoothing times [7]–[8]. The most common of these methods is secondary exponential smoothing, given by:

$$F_t^{(1)} = \alpha Y_t^{(1)} + (1-\alpha) F_{t-1}^{(1)}, F_t^{(2)} = \alpha Y_t^{(1)} + (1-\alpha) F_{t-1}^{(2)} \tag{2}$$

$$Y_{t+m} = a_t + b_t m \tag{3}$$

$$a_t = 2 F_t^{(1)} - F_t^{(2)} \tag{4}$$

$$b_t = \frac{\alpha}{1-\alpha} \left(F_t^{(1)} - F_t^{(2)} \right) \tag{5}$$

where $F_t^{(1)}$ is the smoothed value of period t, $F_t^{(2)}$ is the second smoothed value of t, $F_{t-1}^{(2)}$ is the second exponential smoothed value of $t-1$, and α is the smoothing factor [9].

2.3 Metrics

Root-mean-square error (RMSE) and R-squared are used to determine how well the model fits. RMSE represents the mean-squared error statistics of the output model. These statistics show the difference between the model's predictions and real values, i.e., the standard deviation of the residuals. The unit of measure is consistent with the original data. The RMSE is given by [10]:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2} \tag{6}$$

where $\hat{y}_i$ is the real value, and y_i is the predicted value.

R-squared [11] is the square of the correlation between the measured (empirical) value and the predicted value. A higher R-squared means a better-fitting model. The maximum R-squared value is 1. When the time series contains seasonal trends, a stationary R-squared statistic is better than a normal R-squared statistic.

In this paper, we use stationary R-squared as the evaluation

index for data with obvious seasonal trends. We use RMSE as the evaluation index for data with no obvious seasonal trend, such as throughput data from a single cell.

3 Modeling and Results

Here, we analyze two practical scenarios. In the first scenario, each cell is divided into regions, and the throughput of an entire region is predicted. In the second scenario, the throughput of a single cell is predicted according to historical data.

The reason for creating these two scenarios is that network operators are constantly constructing, adjusting, and optimizing their network, and single cell throughput prediction alone is not enough. If a new cell is built next to cell A, then the throughput of cell A is bound change, and the former data is discarded. Therefore, the first scenario is proposed. QoS can be improved by knowing the network throughput in advance.

3.1 Throughput Prediction for a Whole Region

We first investigate how downlink throughput in a whole region changes over time. **Fig. 1** shows the mean throughput in a region on weekdays and weekends. The weekday mean throughput was obtained by averaging the throughput in the whole region over 10 consecutive weekdays, and the weekend mean throughput was obtained by averaging the throughput over two consecutive weekends (four days). For both weekdays and weekends, the mean throughput in the whole region was at its lowest at 05:00. On a weekday, the mean throughput peaked at 09:00 and 19:00. On the weekend, throughput peaked at 13:00. We divided the throughput in the whole region in weekdays and weekends for further statistical analysis

To analyze the throughput on weekdays, we used the hourly data of ten consecutive weekdays. Five days of this data was used for modeling, and the other five days was used to determine the accuracy of the prediction.

In **Fig. 2**, the real throughput on weekdays in the whole re-
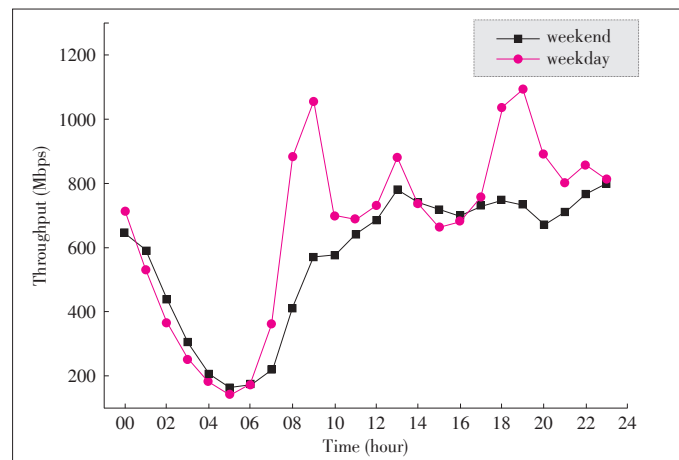
▲Figure 1. Weekday and weekend mean throughput in a whole region over 24 hours.

gion is seasonal. Therefore, we use the ARIMA (2, 0, 1) model and exponential smoothing with $\alpha = 0.600$ to predict throughput on weekdays in the whole region. Although there are gaps between the measured and predicted throughput in the whole region, the predictions by both models are highly accurate. The ARIMA model is more accurate in the valleys of the real throughput curve, which occur at around 05:00, 11:00 and 15:00 every weekday.

Table 1 shows the degree of fit statistics for the prediction models. Both the fit of the curve and the stationary R-squared statistic indicate that the ARIMA model is better than the exponential smoothing model for predicting throughput on weekdays in a whole region.

To study the throughput on weekends, we used hourly throughput data from two consecutive weekends. Two days of this data was used for modeling, and the other two days of data was used to determine the accuracy of the prediction.

The prediction models for throughput of weekends in a whole region is ARIMA (1, 0, 2) and exponential smoothing method with $\alpha = 0.500$. Fig. 3 shows predicted weekend throughput in a whole region using the ARIMA model and exponential smoothing model separately. The throughput predicted using the exponential smoothing model is closer to actual throughput that that predicted using the ARIMA model on a weekend in a whole region (Table 2). The degree of fit statistics supports this. Hence, we obtain the result, that exponential smoothing method is better to predict the weekends' through-
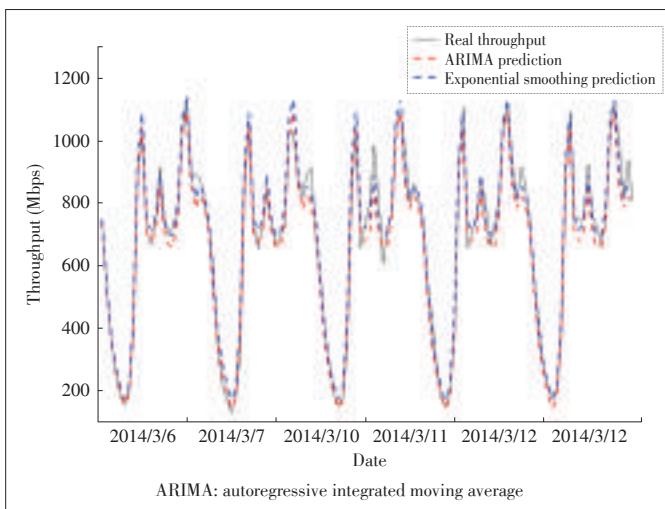
▲Figure 2. ARIMA model and exponential smoothing model are used to predict the throughput on weekdays in a whole region.

▼Table 1. Degree of fit statistics for models used to predict throughput on weekdays in a whole region

	Stationary R-squared
ARIMA model	0.880
Exponential smoothing model	0.664
ARIMA: autoregressive integrated moving average	
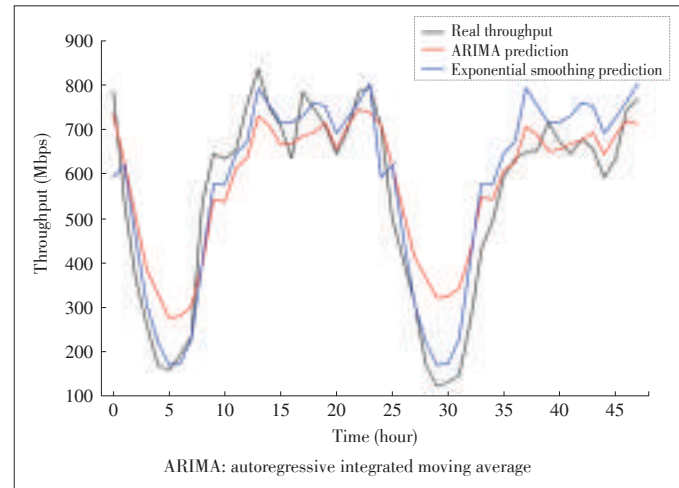
▲Figure 3. ARIMA model and exponential smoothing model for predicting the throughput on weekends in a whole region.

▼Table 2. Degree of fit statistics for models used to predict throughput on weekends in a whole region

	Stationary R-squared
ARIMA model	0.460
Exponential smoothing model	0.723
ARIMA: autoregressive integrated moving average	

put in a whole region.

3.2 Throughput Prediction for a Single Cell

A single-cell traffic time series is highly unpredictable and has no obvious seasonal trend. Even within the same cell, throughput changes greatly on different days. Although there are gaps between the real and predicted throughput curves, a time series model for a single cell still has some use in network optimization. Here, we use the throughput data of an LTE network over eight consecutive days. Seven days of this data is used for modeling, and the other day of data is used to determine how well the model fits.

The stationary R-squared statistic is usually used as an evaluation index when the time series contains seasonal trends. Because there is no significant seasonal trend in the throughput of a single cell, we use RMSE as an evaluation index.

Fig. 4 shows the throughput prediction for single cell. The prediction models are ARIMA (1, 1, 1), and exponential smoothing with $\alpha = 0.100$. Fig. 4 shows that these two models do not accurately predict abrupt changes of throughput in the single cell. The exponential smoothing model is a little more accurate between 17:00 and 23:00. Table 3 shows the accuracy statistics of the two models.

We chose 100 cells randomly and modeled them. Then we obtained the RMSE statistics for these cells. Fig. 5 shows the distribution of RMSE for prediction using the ARIMA model and exponential smoothing model in 100 cells. The RMSE of

Predicting LTE Throughput Using Traffic Time Series
Xin Dong, Wentao Fan, and Jun Gu
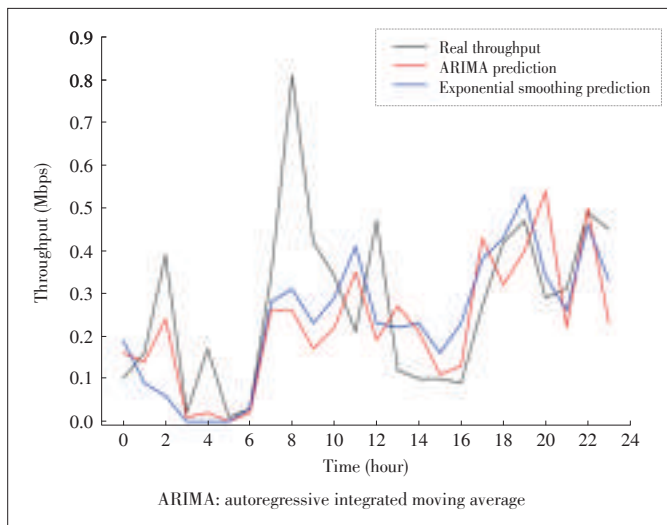
▲Figure 4. ARIMA model and exponential smoothing method for predicting the throughput of a single cell in an LTE network.

▼Table 3. RMSE of the prediction model (single-cell throughput)

	RMSE
ARIMA model	0.193
Exponential smoothing model	0.153
ARIMA: autoregressive integrated moving average RMSE: root-mean-square error	

ARIMA: autoregressive integrated moving average
RMSE: root-mean-square error
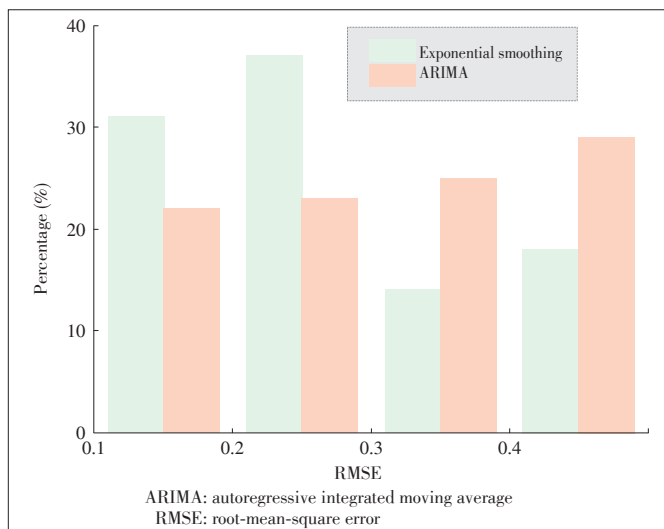
▲Figure 5. RMSE statistics for throughput prediction in 100 cells.

the exponential smoothing method is mainly distributed between 0 and 0.3, and that for the ARIMA model is mainly distributed above 0.3. In general, the exponential smoothing model is better for predicting throughput in a single cell.

4 Conclusion

In this paper, LTE throughput is modeled as a time series, and future values of the traffic time series are predicted using the ARIMA model and exponential smoothing model. Using different time series models, we studied throughput in both a single cell and a whole region within an LTE network. When studying throughput in a whole region, we considered weekday and weekend separately because their throughput patterns were different. The ARIMA model is better than exponential smoothing for predicting throughput on weekday in a whole region, and exponential smoothing model is much better than the ARIMA model for predicting throughput on weekends in a whole region. Exponential smoothing is more accurate than the ARIMA model for predicting throughput in a single cell. Throughput prediction based on time series models can be used in the design, management, planning, and optimization of networks.

References
[1] C. V. N. Index, "Global mobile data traffic forecast update, 2012‑2017," Cisco White Paper, 2013.
[2] E. Nan, X. Chu, W. Guo, and J. Zhang, "User data traffic analysis for 3G cellular networks," in Proc. CHINACOM, Guilin ,China, Aug. 2013, pp. 469–472. doi: 10.1109/ChinaCom.2013.6694641.
[3] V. Buenestado, J. Ruiz‑Aviles, M. Toril, et al., "Analysis of Throughput Performance Statistics for Benchmarking LTE Networks," IEEE Communications Letters, vol. 18, no. 9, pp. 1607–1610, Sept. 2014.
[4] M. Mirza, J. Sommers, P. Barford, and X. Zhu, "A machine learning approach to TCP throughput prediction," in Proc. SIGMETRICS, New York, USA, 2007, pp. 97–108.
[5] G. E. P. Box and G. M. Jenkins, Time Series Analysis Forecasting and Control, 2nd ed. San Francisco, CA: Holden‑Day, 1976.
[6] C. Babu and B. Reddy, "Predictive data mining on average global temperature using variants of ARIMA models", in Proc. ICAESM, Tamil Nadu, India, 2012, pp. 256–260.
[7] Q. Chen and X. Li, System Engineering‑Theory and Practice. Beijing, China: National Defense Industry Press, 2009.
[8] X. Shang, W. Lin, and Y. Tang. "Development and application of a combined water quality prediction model based on exponential smoothing and GM(1,1)," Environmental Science&Technology, vol. 34, no. 1, pp. 191–195, May 2011, doi: 10.3969/j.issn.1003-6504.2011.01.046.
[9] W. Sun and R. Yang, Economic Forecast. Beijing, China: Agricultural University Press, 2005.
[10] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International Journal of Forecasting, vol. 22, pp. 679–688, 2006.
[11] J. R. Taylor, An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. Mill Valley, USA: Univ. Science Books, 1996.

Biographies

Xin Dong (dongxin2014@gmail.com) is pursuing her master's degree in telecommunications at Beijing University of Post and Telecommunications (BUPT). Her research interests include data mining and time series analysis. She has previously researched the prediction of time serials of traffic flow.

Wentao Fan (ffantastic@126.com) is pursuing his master's degree in telecommunications at BUPT. His research interests include data mining, and network analysis and optimization based on mobile devices. He has researched the prediction of time serials of traffic flow using the SVR method.

Jun Gu (gu.jun@zte.com.cn) is a chief engineer of 4G radio network planning at ZTE Corporation. He has 10 years' research and field experience in network principles, standardization, simulation, algorithm design, and planning and optimization.