Telephone Traffic Forecasting of Electric System Based on Multi-factor Decomposition

Bo HU[1,a], Xiao-Ling REN[2,b], Cai-Jun ZHANG[3,c], Xiao-Hui ZHANG[3], Hua-Qin QIN[4] and Chong-Hui GUO[2]

1State Grid Anshan Electronic Power Supply Company, Anshan 114000 China.

2Research Institute of System Engineering, Dalian University of Technology, Dalian 116024, China.

3 State Grid Customer Service Center, Tianjin 300000, China.

4 Beijing Kedong Electricity Control System of Limited Liability Company, Beijing 100085, China.

a)hb@ln.sgcc.com.cn

b)340589430@qq.com

c)caijun-zhang1@sgcc.com.cn

Keywords: telephone traffic forecasting; electric system; multi-factors decomposition.

Abstract. Many factors can affect the telephone traffic of power supply service center. And telephone traffic forecasting would help to estimate the population size, make schedule analysis and copy with the sudden situation. We use some province data of 95598 call center as an instance to analyze factors using time series data as well as extra domain knowledge in order to get the routine rules of telephone traffic. Combining those two to detect the noise and outliers and confirm three factors which respectively are scheduled outage, time, and weekday affecting the traffic flow. The developed time series decomposition model combined with the domain factors is introduced here to make a more accurate relative error, especially when it comes to the rush hour in day time.

Introduction

In order to improve the staff shift arrangement we always need to know the amount of telephone traffic in power supply service center. Thus we can adjust the number of stuff needed more flexible and save the cost. Accordingly, how to improve the prediction accuracy of telephone traffic is one of the most important problems of call center. For all we know, most of telephone traffic forecasting methods were out of time series forecasting[1] based on historical data. The main idea was using the weighted average or the improved edition of that to smooth the time line in order to build regression model to forecast the future telephone traffic. Ahmed NK[2] made use of machine learning which mainly based on their statistical characteristic[3] to forecast the time series. Zhou proposed the method of similar hierarchical prediction method to deal with factors affecting telephone traffic in short time period. The above methods of forecasting are all based on statistics. In another word, they try to improve the accuracy directly according to the numerical data and try to improve the time series forecasting algorithm. However power supply service system are affected by many particular factors, because of the particular domain restrained by the within laws. Besides, there are more and more information acquisition means which offers various kinds of data more than just numerical data, especially for text data. Thus different kinds of data can offer more information and domain knowledge which could be used to forecast telephone traffic.

Framework Constraction of Telephone Traffic Forecasting

According to investigation, there are six factors affecting telephone traffic, which respectively are territory, outrage information, power load, temperature and weather, special days namely holidays, workday, weekend, and the bill days, and the last factor is special event like rotation table plan approval, new policy and public statement.

The data we obtained here mainly contains two parts, the telephone traffic flow data and text data gathered by various ways including outrage notice information. We make use of the two types of data trying to mix them to mine something interesting which can gain better forecasting result and we develop a framework of forecasting in Figure 1. The framework mainly aims to make the forecasting based on understanding as much as possible of the domain knowledge instead of simply using numerical data. Firstly, in the preprocessing phase, the domain knowledge here meaning the outrage notice information must be matched and synthesized with the time series data in order to identify abnormal points and then smooth data. Secondly, we amylase the processed data on multi-granularity. Based on the cyclicity and volatility level, we summarize two significant factors, time periods and week days. Then we combine the results of first two steps into the final forecasting model. That is to say, the power failure time obtained from first step and the tow significant factors out of the second step are there factors we find affecting the telephone traffic and we mix there of them to make a better result. And the final experiment contrasted with AR model shows that the framework works better. In fact, we use those factors to develop the time-series decomposition model which introduces external influencing factor of power failure time and get the better accuracy.

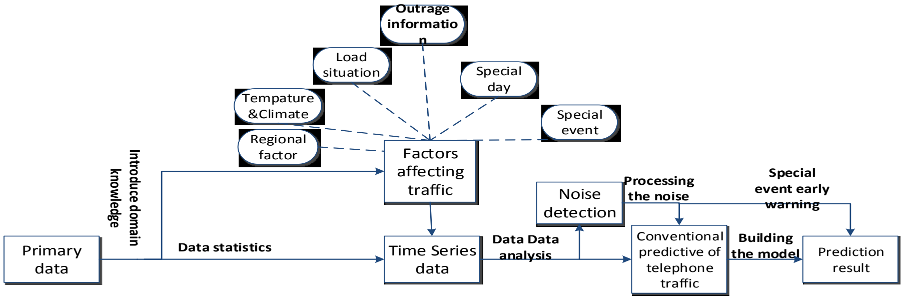

Figure 1. The forecasting framework.

Descriptie Analysis of Telephone Traffic

Description analysis is a primary approach of data which can sees some attributes of the data, especially for time series data. In the practice of time series analysis, description analysis plays a vital role in noise detection, affecting factors detection and model selection. Here according to what we got, in this paper we introduce the power outrage notice information which is a important section of power system, to improve the forecasting accuracy.

Here is two part of data which could be used to detect the noise. The first is numerical data of telephone traffic and the other is the outrage notice information which is used to inspect how the special event affects the traffic. The time series out of the original data set shows in Figure 2. And based on the method of noise detection in, to spot the obvious noise. Besides, we got the external information of power outrage notice information which can be used to measure the affection of power failure onto the telephone traffic.

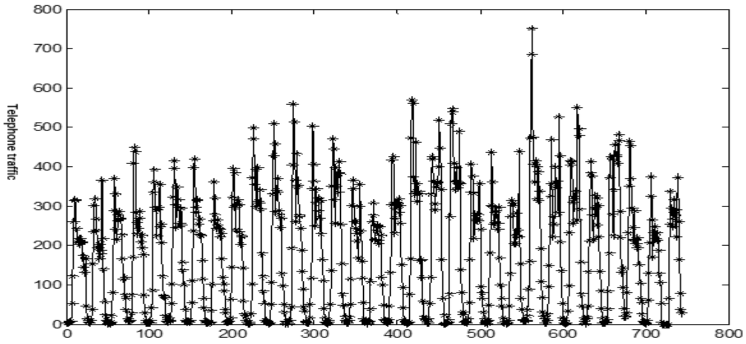

Figure 2. Time series volatility.

According to statistic result of telephone traffic amount, those points whose value are larger than 1.5IQR are treated as noise. In Figure 2 the points of March 24 we can see are deviated a lot, and can be treat as noise. In order to as much as possible make use of data, the outrage notice information of March and April are included in the stage of preprocessing. The data sets contain the time of the calls, duration of the calls, accept content, handle option, reply content, contact address, customer name and election address. Under that situation, we can measure the duration of outrage's affection onto the amount of the telephone traffic, and then we get the index of outrage duration to further detect noise. In order to simplify the model we divide the time series data into two parts,the outrage day and no outrage day, and then we can measure the affection of outrage duration to get the index. The index can be get by formula 1).

$$I_d = \frac{x_d}{\frac{1}{N}\sum_{i=1}^{N} x_{n_i}} \tag{1}$$

Id is the index of outrage duration. xd is the amount of telephone traffic in outrage day in contrast with x_{n_i} in the normal day (i=1..N). N means the number of normal days. According to the date match and computation, the distribution result of duration index shows in Figure 3 based on which we set the threshold at the value of 1.3. That means we the index is larger than 1.3, the duration has obvious affection on the telephone traffic and the detailed result of duration index is the table 1. In this table 1, the bold Figures represent those whose value beyonds our pinning down threshold. And the index is clustered into three groups in which the duration time between 4 to 5 hours has obvious affection. Thus, by the basic noise detection model as well as the duration index method, the total noise in March can be decided.

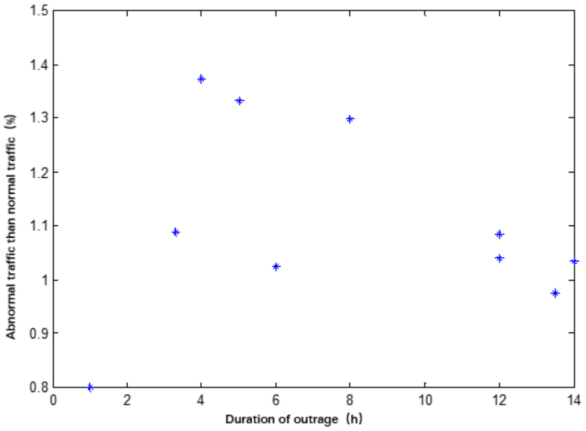

Figure 3. Distribution of duration index.

Table1 Outrage date, dutation time and index

The outrage date	Duration(h)	Index
March 20	5	1.332597697
March 21	13.5	0.975500957
March 26	8	1.298554944
March 27	14	1.034430136
March 28	4	1.373214223
March 29	12	1.040064799
March 30	1	0.801295972
April 9	3.3	1.088709677
April 16	6	1.024193548
April 24	12	1.084677419

Time series are often being affected by various factors, especially of seasonal factors which means the data itself always shows similar behavior after S time intervals. Those factors are cyclical factors which are basic factors of time series. And mining seasonal factors are of vital important for time series analysis.

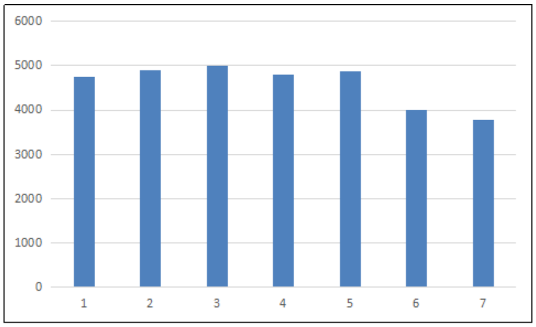

Figure 4. Histogram of traffic amount of different days in a week (x-axis is the 7 days in a week, and y-axis is the traffic amount).

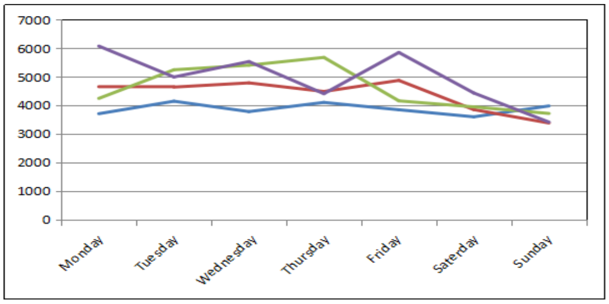

Figure 5. Traffic fluctuation of 4 weeks in March 2014 (x-axis is the 7 days in a week, y-axis is the traffic amount and four colors of lines represent 4 weeks).

Figure 4 draws a histogram of average traffic amount of weekdays and weekend in March 3, 2014. It's easy to see that the traffic amount of the day is related to week. The telephone traffic amount between Monday to Friday is similar in which the amount is between 4,500 to 5, 000. However, the telephone traffic amount in weekend is obvious lower than in weekday and Saturday is slightly higher than Sunday. Therefore, week is a factor affecting the traffic amount owing to the difference between weekday and weekend. And Figure 5 shows the telephone traffic fluctuation of four different weeks in March, 2014. And this Figure has been smoothed after filtering noise.

Based on different time periods, Figure 6 shows the box Figure of the telephone traffic amount in March 2014. And of course, it's easy to see that the amount of traffic is different in different time period of a day no matter the day belonging to weekday or weekend. The amount is higher in between 7:00 to 22:00 than that in between 23:00 to 6:00 the next day. And that situation is similar in different, the amount is higher in daytime than in the night which also is related to human lifetime. People often calls in the daytime and do not care about the power problem in the night.

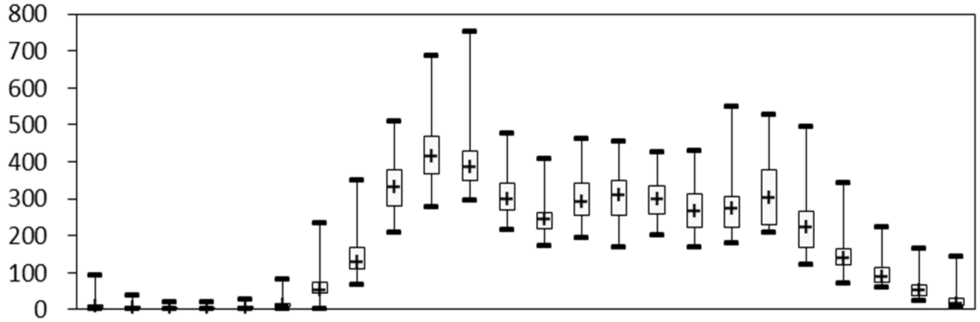

Figure 6. Box plots of the telephone traffic in a day.

Telephone Traffic Forecasting Model

Let't be the sampling period, N be the number of time points, and the number of calls construct the time series data set which can be represented by the data set of { x1, x2, x3,…, xN }.

Based on the above analysis of telephone traffic, the telephone traffic amount is affected mainly by three factors to sum up. That is the time period in a day, the day in a week and the outrage duration. In turn, we propose a method called multi-factors decomposition which decomposes the original time series after noise filtering to get seasonal factors and introduces the outrage duration index in to form the complex model. Firstly, according to the seasonal factors get the normal trends. Secondly, use AR model to smooth to predict the seasonal trends. And then use the addictive model to repair back the seasonal factors. Besides, the outrage duration index must be added in the final forecasting step.

The model to predict the amount of telephone traffic $\theta(x_k)$ in time period k is

$$\theta(x_k) = f(\overline{x}_k(hour), \overline{x}_k(weekday), d(k), I(k)) \tag{2}$$

$\theta(x_k)$ is the prediction result of k time, $\overline{x}_k(hour)$ is the time affecting factor and $\overline{x}_k(weekday)$ is factor of day of the week. d(k) is the outrage duration factor and I(k) is residua. Thus the multi-factor decomposition is based on the basic formula (2) can be get as following:

$$\theta(x_k) = d(k) \times (\overline{x}_k(hour) + \overline{x}_k(weekday)) + I(k) \tag{3}$$

$\overline{x}_k(hour)$ is the average value of 24 hours after AR smoothing, and there are 24 values to get. $x_k(weekday)$ can be get by the same way which is the average of days of week after normalization based on different weeks and there are 7 values in total. $d(k)$ represents the external a`ffecting factors, hers meaning the outrage duration index. Concerning the real situation, the duration time lies between 4 to 5 hours comes out in contrast with others which we treat their index as 1 meaning no obvious affection. And here the expectation of $d(k)$ is 0 which is $E(I(k)) = 0$.

Experiment and Result Analysis

The power system company of some province offered us the data set about 95598 power service call center and the data time lies in March 2014. In order to make better use of the original data and make further use of the forecasting result, we counted the numerical data of number of calls in two granularities, 24 hours and week days. Therefore, we get 744 values out of the original 135, 492 calls.

In order to predict the normal telephone traffic amount, we make use of the method in what has been mentioned in section 2.1 to detect, filter and smooth the noise caused by emergency situation and power outrage notice information. In detail, we use the average value of traffic amount of before and after three days to replace the value of every point. Figure 7 is the one noise has been filtered and the detection result shows that March 20, March 24 and March 28 are the noise.

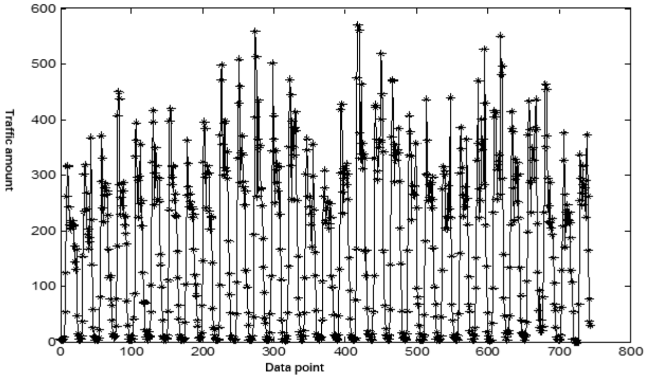

Figure 7. The time series fluctuation after filtering the noise.

After preprocessing we divide the data set into two parts, the training set and the test set. Based on the real condition of March 2014, the training set contains 28 days which is four complex weeks, and the data in the last three days as the test data set.

We use the model mentioned in section to train on training set and make the forecasting on the test set. And compare the forecasting result to the result obtained by the AR model. In the experiment set stage, in order to keep the validation useful, the data we try on two different models all has been filtered the noise. Based on the multi-factor decomposition model, the time period trend and the week day trend have been settled and the fluctuation line charts showing in Figure 8a and Figure 8b. And the forecasting is to add the $x_k(weekday)$ to the $\overline{x}_k(hour)$ to adjust the value.

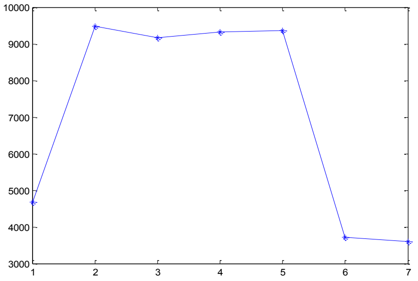

Figure 8a. time period factor line chart.

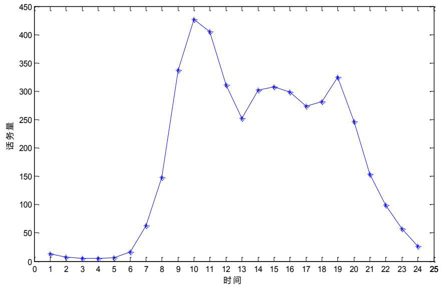

Figure 8b. Bweek day factor line chart.

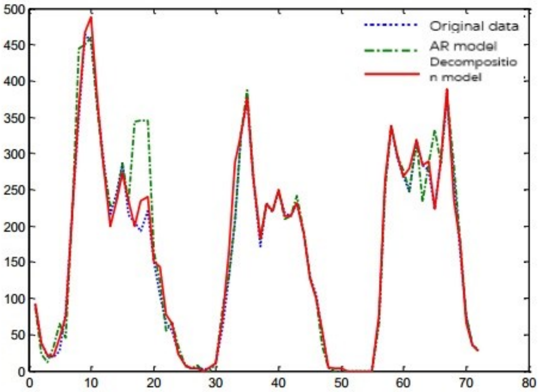

Figure 9. The forecasting results.

And the detailed values show in table 2 and table 3.

Table 2 Time period factor values

Time period	0:00-0:59	1:00:00-1:59:59	2:00:00-2:59:59	3:00:00-3:59:59	4:00:00-4:59:59	5:00:00-5:59:59
x(hour)	13	8	6	6	7	17
Time period	6:00:00-6:59:59	7:00:00-7:59:59	8:00:00-8:59:59	9:00:00-9:59:59	10:00:00-10:59:59	11:00:00-11:59:59
x(hour)	63	149	338	427	405	311
Time period	12:00:00-12:59:59	13:00:00-13:59:59	14:00:00-14:59:59	15:00:00-15:59:59	16:00:00-16:59:59	17:00:00-17:59:59
x(hour)	253	302	308	300	274	282
Time period	18:00:00-18:59:59	19:00:00-19:59:59	20:00:00-20:59:59	21:00:00-21:59:59	22:00:00-22:59:59	23:00:00-23:59:59
x(hour)	325	247	156	100	58	27

Table 3 Week day factor values

Week day	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
x(weekend)	-27	-31	13	16	21	13	13

To get the final result, we sum up components results and make the rounding amount to be the final forecasting value. The outrage duration index of duration time being 4 to 5 hours we set d(k)=1.352, and the other duration time the index being 1. Because there is no negative value of telephone traffic, we set those below zero to be zero. Compare the multi-factor decomposition model and the result of AR model, the result of last three days shows in Figure 9. And the comparison result on prediction accuracy of those two measures is as following in table 4.

Table 4 The comparison result of AR with multi-factor decomposition model

Model **Time Relative erroe**	**0:00-5:59**	**6:00-11:59**	**12:00-17:59**	**18:00-23:59**
AR	13.11%	29.12%	30.98%	39.22%
Multi-factor ecomposition	13.12%	21.22%	24.19%	26.73%

We can see that the forecasting result is similar in the time valley zone when it comes to 0:00-5:59. But in the rush hour when people are busy in the daytime, the model we propose called multi-factor decomposition in this paper dose a better job than the model which runs merely on the numerical time series data by AR method.

Conclusion

We propose a telephone traffic forecasting framework which combines numerical data of time series and the domain knowledge to make full use of the information we can get in desire of a better result. Based on that framework, to mix those two kinds of data, we further propose a model called multi-factor decomposition to predict the traffic result. Therefore, the model can include more affecting factors that indeed affect the traffic amount in the real world to reflect the real situation. Besides, this model is a open model which you can introduce more information we can obtain as much related information as you can. In another word, the more information you can get from the unstructured data, the better the model would be to tell you more about the real situation. Through the comparison we can tell that the framework on the one hand can embrace multiple heterogeneous data, on the other hand can get a better prediction result. The multi-factor decomposition model also can make use of the external knowledge, and decompose the useful affecting factors and can offer a early warning of the telephone traffic because this model mix the knowledge of outrage situation. When the duration outrage is going to be around 5 hours and also in the weekday of daytime, the call center must arrange more staff to waiting by the table.

When it comes to short time telephone traffic forecasting, seasonal factors are the main affecting factors to affect the fluctuation of the time series. The workday and weekend have different influence on the traffic, And of course the daytime and night hour also influence the telephone traffic.

Targeting to the prediction issue, the original data base also contains other unstructured data and lots of information reflecting the background and human behavior information. By the framework of mix these information can provide a better way to know the information as much as possible instead of ignoring these information. And as a result, how to measure the usefulness of all the external information will be the future work.

References

1. CHEN R. Communication traffic analysis with the study and comparison of multiple prediction models[D]. Beijing, Beijing University of Posts and Telecommunications.

2. Ahmed N K, Atiya A F, Neamat E G, et al. An empirical comparison of machine learning models for time series forecasting[J]. Econometric reviews, 2011(29): 594-621.

3. Time series analysis: forecasting and control (4th edition)[M]. Beijing University of Technology Press, 2011.