Journal of Mathematics and Informatics
Vol. 14, 2018, 1-10
ISSN: 2349-0632 (P), 2349-0640 (online)
Published 6 August 2018
www.researchmathsci.org
DOI: http://dx.doi.org/10.22457/jmi.v14a1

Journal of
Mathematics and
Informatics

On Languages Distribution

Bi-chuan Jiang

Department of Applied Mathematics, Chongqing University of Posts and Telecommunications
Chongqing 400065, Chongqing, China. E-mail: 981937548@qq.com

Received 20 June 2018; accepted 31 July 2018

Abstract. In this paper, we studied the distribution of global languages; we set up a multivariate prediction model to predict the number distribution of languages. A multivariate dynamic rough model was used to predict the geographic distribution of languages, and predict the distribution of language in the next 50 years.

Keywords: Wavelet neural network; rough set; prediction; gray technology

AMS Mathematics Subject Classification (2010): 00A71

1. Introduction

There are currently about 6,900 languages spoken in the world. The trend of globalization not only profoundly changes the economy and society, but also affects the trend of language development. Currently, more than 90% of Internet information is in English; More than 60 countries make English the official or semi-official language; It is also the main language of diplomatic and international trade. These factors, coupled with economic development, geographical endowments, government promotion and other factors, Make English to world lingua franca. However, some languages have disappeared. The number of languages has been constantly decreasing. Therefore, in our globalized world, it is of great research significance that people study the changes in the distribution pattern of languages.

At present, people are studying more and more languages. In [4,5], the author studied the natural distribution of language, and the current distribution pattern of each country's language. In [8,11], the author further studied the evolution process, influencing factors and distribution of Chinese dialects. In [6,7,9], the author studied the current migration patterns, spatial distribution patterns and prediction methods of the population. For the current dynamic prediction, there are also many methods. In [10], the author uses a population migration algorithm based on artificial fish swarm algorithm to predict the population. In [13], the authors used a multi-factor predictive model of gray technology and wavelet network fusion to predict things. There are also many methods for the prediction of spatial distribution. In [12], the author used a multi-factor dynamic coarse prediction model to predict the dynamic tourism demand. In [14], the author studied the classification of spatio-temporal data models of geographic information systems.

In this paper, we set up a multivariate prediction model to predict the number distribution of languages. A multivariate dynamic rough model was used to predict the geographic distribution of languages. To better study the distribution of global language in changing trends.

2. Population distribution

In order to simulate the distribution of various speakers according to the forecasting trend, we describe the number distribution and spatial distribution. As the population growth rate of the birthplace of language directly affects the number of people, this factor has a direct relationship with the distribution. The more developed the economy of origin is, the more traffic there will be in the region and the closer the economic exchanges will affect the language distribution. The larger the number of language learners, the more closely it is distributed. The larger the number of people who move into the language in each year, the greater the number of people who study the language, thus these are affecting the distribution of the language. Large companies play a crucial role in international exchanges, language requirements are also very strict. Therefore, the proportion of the members of a large corporation is also related to the language distribution.

Through the above description, we extract the impact of five major factors on language distribution: population growth rate, GDP, number of language learners, number of residents who move in, and composition ratio of language of large-sized companies.

2.1. Improved wavelet neural network structure

We first consider the BP neural network three-tier structure, the data are trained to achieve the trend prediction of various languages. Because the historical data of the influencing factors have the characteristics of randomness and non-linearity, and the importance of each factor on the predictive objects is also different, with gray causality, and the original sequence of the predicting objects is concrete and has the ginkgo character. In the sample Small capacity and small sample information will reduce the prediction accuracy of the wavelet network model, so we embed the ashing layer at the input end to weaken the randomness of the observed data and enhance the regularity of the data.

In order to describe the quantitative distribution of various languages, we establish the following multi-factor prediction model of fusion of gray technology and wavelet neural network.

2.1.1. Build wavelet neural network

We are considering the three-layer structure of BP neural network, thus we construct a neural network that uses wavelet bases instead of neuron activation functions, input variable is $X = (x_1, x_2, \cdots, x_n)$ and output variable Y ,the neuron activation function of the input layer is set as a linear function $f_1(x) = x$, input layer and hidden layer connection rights for the $v_{ij}(i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, m)$, the hidden layer neuron activation function is set to Morlet function $f_2(x) = \cos(1.75x)e^{(-\frac{x^2}{2})}$, a_i is scalability

factor of the hidden layer neurons and b_j is translation factor of the hidden layer neurons, hidden layer and output layer connection rights for the $u_j (j = 1, 2, \cdots, m)$, output layer neuron activation function is $f_3(x) = \frac{1}{1+e^{-x}}$, thus the entire neural network model is

$$y = f_3(\sum_{j=1}^{m} u_j f_2(\frac{\sum_{i=1}^{n} v_{ij} x_i - b_j}{a_j})) \tag{1}$$

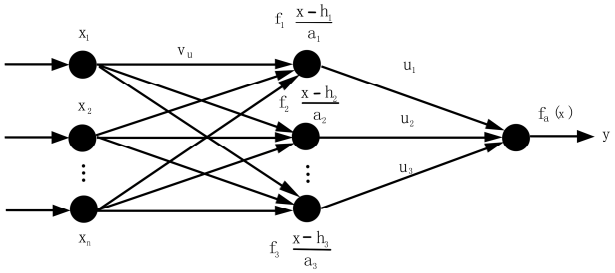

Figure 1: Wavelet neural network structure

By studying sample $(x_{1k}, x_{2k}, \cdots, x_{nk}), k = 1, 2, \cdots, L$, we can get

$$\hat{y} = f_3(\sum_{j=1}^{m} u_j f_2(\frac{\sum_{i=1}^{n} v_{ij} x_{ik} - b_j}{a_j})) \tag{2}$$

Learning principle is

$$\min E = \sum_{k=1}^{L} (y_k - \hat{y}_k)^2 \tag{3}$$

2.1.2. Establish data ashing model

Suppose that the observation of the *ith* $(i = 1, 2, ..., n)$ influencing factor satisfies $X_{i,t} \geq 0, t = 1, 2, \cdots, N$, the data is ashed using the first-order accumulation technique, the data ashing model is A.

$$X_{i,t}^{(1)} = \sum_{k=1}^{t} X_{i,k} \quad i = 1, 2, ..., N$$

Then we can get

$$X_{i,t}^{(1)} = \sum_{k=1}^{t} X_{i,k} \quad i = 1, 2, \cdots, n; \quad t = 1, 2, \cdots, N$$

$$X_{i,1}^{(1)} = X_{i,1} \quad i = 1, 2, \cdots, n$$

$$X_{i,t}^{(1)} = X_{i,t-1}^{(1)} + X_{i,t} \quad i = 1, 2, \cdots, n; \quad t = 2, 3, \cdots, N$$

the prediction object original data sequence $Y_t \quad t=1, 2, \cdots, N$ remains unchanged.

2.1.3. Establish wavelet neural network model

The gray data of the influencing factors and the original value of the forecasting object $(X_{1,t}^{(1)}, X_{2,t}^{(1)}, \cdots, X_{n,t}^{(1)}; Y_t) \quad (t=1, 2, \cdots, N)$ are used as learning samples to train the above wavelet neural network and we can get the wavelet network model:

$$Y^{(1)} = f_3(\sum_{j=1}^{m} \hat{u}_j f_2(\frac{\sum_{i=1}^{n} \hat{v}_{ij} X_i^{(1)} - \hat{b}_j}{\hat{a}_j})) \tag{4}$$

2.1.4. Establish a predictive model

Using the ashing model, ashing of the raw data $X = (x_{1,N+1}, x_{2,N+1}, \cdots, x_{n,N+1})$ of the influencing factor N + 1 is as follows:

$$\hat{X}_{i,N+1}^{(1)} = X_{i,N}^{(1)} + X_{i,N+1} \quad i=1, 2, \cdots, n$$

The ashed data $(\hat{X}_{1,N+1}^{(1)}, \hat{X}_{2,N+1}^{(1)}, \cdots, \hat{X}_{n,N+1}^{(1)})$ is substituted into the wavelet neural network to obtain the gray prediction value:

$$\hat{Y}_{N+1} = f_3(\sum_{j=1}^{m} \hat{u}_j f_2(\frac{\sum_{i=1}^{n} \hat{v}_{ij} \hat{X}_{i,N+1}^{(1)} - \hat{b}_j}{\hat{a}_j})) \tag{5}$$

That Y in the N + 1 period of the forecast value.

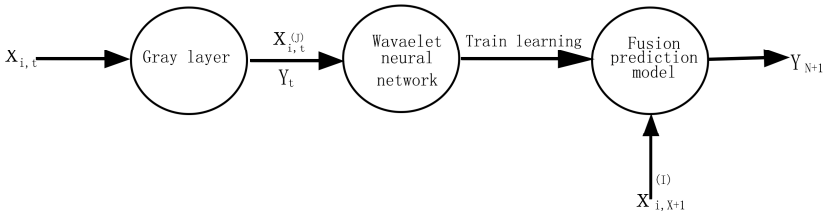

Figure 2: Multi-factor Forecasting Model on the Integration of Grey Technology and Wavelet Network

Using the multivariate prediction model established above, we can predict the 26 languages using MATLAB. First, we predict the influencing factors and then substitute the predicted influencing factors into the model, so we can predict the number of native speakers and total language speakers in the next 50 years respectively, and we have carried on the screening to the earliest language kind, to some commercial communicative language and the local dialect because they have not the language form essentially. Therefore, we eliminated these languages and did not participate in the rankings, that is, the number distribution trend in 19 languages was obtained.

We extracted nine years of data, that is, from 2010 to 2017, five years of data for the training function, the last three years of data for the results of the test, Predicted the number of native speakers in each language, we get the following result(Unit: million)：

Table 1: The total number of native speakers of English and Mandarin each year

years	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028
number1	387	418	415	474	482	481	492	499	521	535
number2	896	902	913	911	923	934	931	974	958	997
years	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038
number1	576	572	610	647	641	684	685	721	741	762
number2	1018	1023	1075	1103	1175	1215	1347	1399	1405	1426
years	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048
number1	755	784	799	797	801	827	833	829	855	867
number2	1445	1472	1497	1503	1529	1542	1554	1558	1562	1577
years	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058
number1	873	884	892	901	923	944	956	949	959	967
number2	1578	1642	1588	1692	1726	1842	1833	1876	1992	1997
years	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068
number1	987	996	1001	1018	1026	1084	1160	1247	1289	1302
number2	2103	2127	2247	2274	2310	2341	2394	2394	2401	2441

Number1 indicates the number of native speakers in English, number 2 indicates the number of native speakers in Mandarin Chinese. Other results can be obtained in the same way. Thus, we can get the change of the number of native speakers in the next 50 years.

From the above results we can see that most languages will increase over time, with fewer languages going down. Due to local economic conditions and the war and other reasons, leading to fewer and fewer speakers. And we can conclude that in the short term, any of the top ten lists will not be replaced by other languages, but for a long time there will be situations in which the language will be replaced by the latter.

We can then use the same model to get the total number of speakers in each language (Unit: million):

Table 2: The total number of people who speak English and mandarin each year

Years	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028
Number 1	997	1047	1154	1237	1374	1441	1512	1589	1601	1654
Number 2	1097	1113	1178	1175	1214	1226	1237	1249	1249	1251
Years	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038
Number 1	1678	1711	1745	1747	1810	1841	1887	1879	1903	1942
Number 2	1312	1315	1329	1347	1344	1349	1382	1441	1471	1446
Years	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048
Number 1	1999	2025	2033	2079	2112	2138	2139	2117	2149	2184
Number 2	1492	1518	1542	1556	1579	1592	1634	1648	1655	1668
Years	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058
Number 1	2210	2239	2241	2258	2247	2239	2275	2271	2284	2263
Number 2	1679	1706	1793	1826	1859	1899	1926	1985	1998	2013
Years	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068
Number 1	2298	2341	2374	2485	2547	2577	2647	2741	2818	2974
Number 2	2147	2195	2241	2283	2348	2472	2523	2599	2674	2777

Number 1 indicates the number of native speakers in English, number 2 indicates the number of native speakers in Mandarin Chinese.

From the above results, we can see that most of the languages will increase over time, fewer languages will be declining, and the number of native speakers will be much smaller than that of the second language. For example, the number of people who predict English is more and more. Although the number of people who learn Mandarin is also steadily rising, those who may learn English may get more and more after long-term prediction More, leading to more people than mandarin. And we can also conclude that in the short term, any of the top ten lists of languages will not be replaced by other languages, but for a long time there will be cases where the language will be replaced by a later language.

To sum up, we can conclude that, the top ten list languages will be replaced by other languages, we have to consider the above factors, to find out why:

(1) Due to the respective policies of the country, or the social environment of the country, the conditions of war and other factors, the population growth rate in the area will cause significant differences or negative growth, resulting in a decrease in the number of people learning the language.

(2) For GDP, showing the country's economic status, etc., it has a direct impact on the country's exchange and development.

(3) The rate of studying abroad and the rate of population migration also lead to cultural exchanges and affect the ranking of languages.

2.2. Multifactor dynamic prediction models based on rough sets

Due to the great incompleteness and uncertainty of the geographical distribution itself, we introduce a rough set to predict the geographical distribution trend of the language. Combining the dynamic characteristics of geographical distribution and the data analysis techniques of the rough set, we obtain the following geographical distribution Multi-factor Dynamic Rough Prediction Model:

We assume that the ratio of a language in the region is y, influenced by m factors, which is $X=(x_1, x_2, \cdots, x_m)$. $u_t=(x_{1,t}, x_{2,t}, \cdots, x_{m,t}; y_t), t=1, 2, \cdots, n$ is composed of the ratio value of language in the t period and the value of its influencing factors.

Based on the principle of pattern recognition, we can assume that there is an indeterminate relationship between variables X and Y:

$$Y=F(X)$$

X and Y are discrete random variables, and we can get

$$Y_M=y_M, X_N=(x_1, x_2, \cdots, x_N)$$

2.2.1. Create a decision data table

We regard the factors that influence the geographical distribution of language as the condition attributes of rough set theory, that is, $C=\{x_1, x_2, \cdots, x_m\}$;Language area ratio y is the decision attribute, namely $D=\{y\}$; $U=\{u_1, u_2, \cdots, u_m\}$ is the rough set theory domain.

Thus, we can get $x_{i,t}$ as the condition attribute x_i on the attribute value of the object u_t. y_t is the decision attribute y on the attribute value of the object u_t. The decision table $S =< U, C, \cup D >$ is obtained, which is the decision data table model.

2.2.2. Attribute value characterization, reduction condition attribute set C

According to the attribute dependency calculation formula, you can get the dependency degree of the decision attribute dependent on the condition attribute:

$$\gamma_C(D) = \frac{card(POS_C(D))}{card(U)} \tag{6}$$

if x_0 is the redundant attribute D of C. As a result, we have achieved the impact of the geographical distribution of abbreviated languages.

2.2.3. Extract decision rules

Suppose that the condition attribute reduction set is C^*. From the corresponding decision table U available classification set and basic decision rule set, which

$$U/C^* = \{X_1, X_2, \cdots, X_N\}, U/D = \{Y_1, Y_2, \cdots, Y_M\}$$

$$G_{(0)} = \{r_{ij}^0 : Des_{C^*}(X_i) \to Des_D(Y_j)(V_i^{(0)}, \mu_{ij}^{(0)}) \mid i = 1, 2, \cdots, N; j = 1, 2, \cdots, M\}$$

$Des_{C^*}(X_i)$ is the pattern description of attribute C^* for pattern X_i, $Des_D(Y_j)$ is the pattern description of attribute D for pattern Y_j, $\mu_{ij}^{(0)}$ represents the confidence of a decision rule established estimates, $v_i^{(0)}$ represents the coverage of the corresponding conditional attribute pattern in the decision table.

2.2.4. Establish a multi-factor dynamic rough forecasting model

We characterize $x_{1,(n+1)}, x_{2,(n+1)}, \cdots, x_{m,(n+1)}$. The corresponding object in the decision data table is set to u_{n+1}, we are using the closeness between the object $u_{(n+1)}$ and the influencing factors of the decision rules $\gamma_{C^*,\ i}(u_{(n+1)})$, and make use of the uncertain law between the mode of influence and the problem, we can build predictive models. So we set up the following forecasting model:

$$\hat{y}_{(n+1)} = \sum_{k=1}^{N} \lambda_{C^*,k}(u_{(n+1)}) \sum_{j=1}^{M} \mu_{k,j}^{(0)} \times Des_D(Y_j) \tag{7}$$

and

$$\lambda_{C^*,i}(u_{(n+1)}) = \frac{\gamma_{C^*,k}(u_{(n+1)}) v_k^{(0)}}{\sum_{i=1}^{N} \gamma_{C^*,i}(u_{(n+1)}) v_i^{(0)}}, \quad k = 1, 2, \cdots, N \tag{8}$$

Indicates the weight of the forecast mean $\sum_{j=1}^{M} \mu_{k,j}^{(0)} \times Des_D(Y_j)$ to which the influencing factor pattern X_k responds.

$$\gamma_{C^*,i}(u_{(n+1)}) = \frac{1}{1+\left\| Des_{C^*}(u_{(n+1)}) - Des_{C^*}(X_i) \right\|} \tag{9}$$

Through the above-mentioned multi-factor dynamic rough forecasting model, we can predict the geographic distribution of languages and get the geographical distribution trend of languages. We conclude the following factors: population growth, GDP , number of language users, area moved into the number and area moved out.

We consider two aspects of the geographical distribution of languages: the description of the proportion of language in seven continents and the proportion of languages in a country. So as to carry on the description of the linguistic geographical distribution tendency through these two aspects.

In the first case, we choose to analyze English and use the data from 2010 to 2017 to divide the continents into regions. We predict the proportion of the total number of people who use English in each of the next few years. Based on the above multi-factor dynamic The rough prediction model has the following results:

Table 3: The distribution of English in seven continents in 2018

island	Asia	Europe	North America	South America	Africa	Oceania	Antarctica
proportion	0.75	0.90	0.78	0.62	0.37	0.93	0

Table 4: The distribution of English in seven continents in 2019

island	Asia	Europe	North America	South America	Africa	Oceania	Antarctica
proportion	0.72	0.89	0.78	0.62	0.35	0.92	0

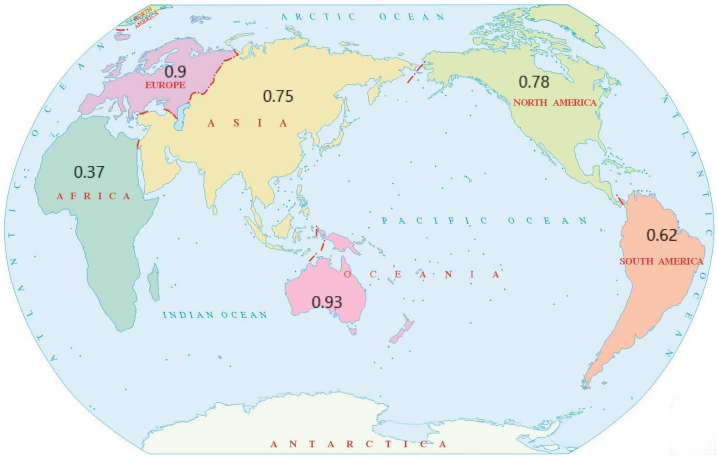

Figure 4:

On Languages Distribution

For the second case, we take the region to conduct an analysis for China. Based on the estimates of the languages in the top 10 lists in 2018, we predict the proportion of the total number of people in each of the next few years and get the following results:

Table 5: The distribution of the top ten languages in China in 2018

Mandrin Chinese	Spanish	English	Hindustani	Arabic
0.97	0.21	0.72	0.23	0.12
Bengali	Portuguese	Russian	Punjabi	Japnese
0.14	0.09	0.29	0.1	0.23

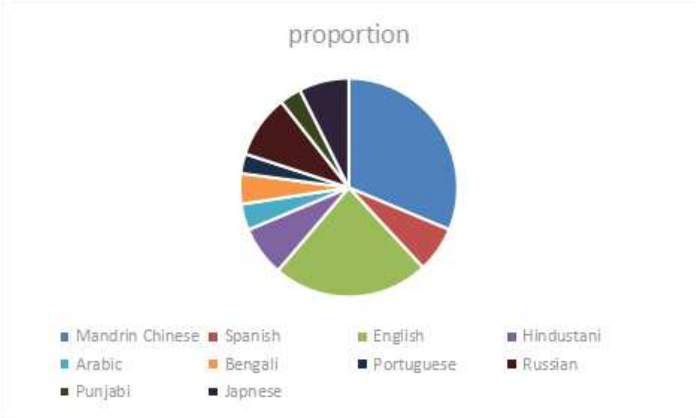

Figure 5: The distribution of the top ten languages in China in 2018

We can get the above chart, most commonly used in China is the Mandarin, followed by the most commonly used is English, the other ten languages on the list of languages in China there is a certain percentage, with a certain diversity, and more diverse languages.

To sum up, we predict and study the number distribution and geographical distribution of languages respectively, and get the development trend of each language in the world.

3. Test: test difference test

1. The basic test process is as follows: $x^{(0)}$ is the original sequence, $\hat{x}^{(0)}$ is the model simulation sequence, and ε is the residual sequence. In which $\varepsilon(k) = x^{(0)}(k) - \hat{x}^{(0)}(k)$, the relative error sequence is $\Delta = (\Delta_1, \Delta_2, \cdots, \Delta_n)$, and $\Delta_k = \left| \frac{\varepsilon(k)}{x^{(0)}(k)} \right|$ and Δ_k are the relative error of the K point, and the $\overline{\Delta} = \frac{1}{n}\sum_{k-1}^{n} \Delta_k$ is the average relative error, and $p = 1 - \Delta$ is defined as the prediction accuracy.

We use the data after the test for three years to get $C = 0.1790969, \quad P = 1$, due to $C < 0.35, \quad P > 0.95$, so the prediction result is excellent.

2. We also get from the above residual analysis $C = 0.3217, \quad P = 0.9514$, because of $C < 0.35, \quad P > 0.95$, so the forecast result is qualified.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 70471057) and the Natural Science Foundation of Education Department of Shaanxi Province (No. 03JK065).

REFERENCES

1. Lei An, Optimization of corporations in enterprise value assessment, *Capital University of Economics and Business*, 72 (2017) 51-61.
2. X.C.Deng, Vector optimization problems: all-round stability, good qualitative and sensitivity analysis, *Guizhou University*, 191(6) (2016) 37-68.
3. Jiang Guo, Based on the distribution of cross-language cross-task natural language analysis, *Harbin Institute of Technology*, 391(2) (2017) 324-681.
4. Changzhu Huang, Global languages in the context of globalization: changes in the use and distribution pattern, *Foreign Social Sciences*, 7(6) (2009) 4-17.
5. Jie Jie, 75 years of China's population distribution of time and space and changes. *Beijing University of Architecture*, 3(4) (2015) 384-395.
6. Ruliang Liu, Renan Jia and Qiuxian Dong, Improvement of population migration model and simulation of system dynamics simulation, *Mathematics and Practice*, 2(18) (2008) 128-133.
7. Baojia Li, On the historical evolution of the dialect pattern of Chinese dialect, *Journal of Jilin Normal University*, 7(3) (2004) 53-58.
8. Deqin Liu, Yu Liu and Xinyu Xue, Population distribution and spatial correlation analysis in China, *Science and Geomatics*, 2004 (S1) 76-79.
9. Bin Li, Population migration algorithm based on artificial fish swarm algorithm, *Wuhan University of Technology*, 2(4) (2008) 28-75.
10. Xiaoqing Su, Jinggong Xu, The influence of geographical changes on the dialect distribution pattern-take pizhou dialect in Jiangsu as an example, *Journal of Xuzhou Institute of Technology*, 7(7) (2015) 293-304.
11. Zhi Xiao and Yulan Ye, A multi-factor dynamic prediction model for tourism demand, *Statistics and Decision*, 7(12) (2005) 33-34.
12. Lei Yan, Bo Zhong, Huiliang Luo and Bangjun Lei, Multifactor prediction model of fusion of gray technology and wavelet network, *Journal of Shanxi Normal University*, 28(6) (2012) 215-210.
13. Shanshan Zhang, Classification of spatio-temporal data model of geographic information system, *Proceedings of Surveying and Mapping Science*, 37(4) (2012) 215-217.
14. Qun Zong, Model checking based on residual analysis Chinese, *Journal of Control Theory*, 2(4) (2008) 28-75.