The Multiple Linear Regression Model and Time Series Model of Language Prediction

Zhiqun Zhang[1], Changming Song[2], Weilian Zheng[3], Zhihong Ma[4]*
[1]College of Humanities, Tianjin Agriculture University, Tianjin, 300384, China
[2]College of Engineering and Technology, Tianjin Agriculture University, Tianjin, 300384, China
[3]College of Food Science and Bioengineering, Tianjin Agriculture University, Tianjin, 300384, China
[4]College of basic science, Tianjin Agricultural University, Tianjin, 300384, China

Abstract: There are currently about 6,900 languages spoken on Earth. At the same time, the total number of speakers of a language may increase or decrease over time because of a variety of influences. We choose the quantity of the migration population, net number of tourist entry and exit, GDP per capita as three general factors which can influence the distribution of various language speakers. Then we develop a MLR model and Time series model to predict trends of global languages in the next 50 years.

Keywords: Multiple Linear Regression; Time Series Model; Language prediction

1. Introduction

At present, the languages still being used can add up to 6,900 categories throughout the world, but this cultural diversity is rapidly disappearing, especially in areas with less population. However, most language extinctions are caused by language transfer rather than by population extinction. So if a kind of language wants to survive, use it or lose it, it is exactly important that been spoken by a dynamic speaker community and plays a critical role in the global economy.

In this paper, we build a MLR model of the distribution of various language speakers on a set of parameters. In addition, we choose the top 20 native speakers and total language speakers at five-year intervals as the research object from 1995 to 2015, and then we use MATLAB to calculate in various languages through the time series model of the trend of moving average.

Multiple Linear Regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. While time series forecasting is the use of a model to predict future values based on previously observed values.

In the process of simulation, we choose a few of typical countries and regions of each language speakers, such as English spoken by four main countries: United States, United Kingdom, Canada and Australia etc. Then we select the quantity of the migration population, net number of tourist entry and exit, GDP per capita as three general factors which can influence the distribution of various language speakers. Each factor is determined by several metrics which can be easily detected.

2. Model 1: Multiple Linear Regression

2.1. Parameters

2.1.1. Quantity of the migration population

Obviously, the number of migration population is a priority in considering the distribution of various language speakers over time. It measures the attractiveness of a language by immigrating or emigrating the typical countries or regions. The index can be calculated by using the number of immigrants minus the number of emigrants, and we conclude the quantity of the migration population about typical countries and regions in 2012.

2.1.2. Net number of tourist entry and exit

As culture exchanges between countries getting closed, the increased global tourism is another critical factor. Therefore, we choose the net number of tourist entry and exit of represented countries and regions in 2012 to reflect the languages geographically distant to interact brought by tourism industry.

2.1.3. GDP per capita

GDP per capita is often considered an indicator of a country's standard of living, which values all final goods and services produced within a nation in a given year, converted at market exchange rates to current U.S. dollars, divided by the average population for the same year[1]. However, these figures should be used with caution, so we choose the average GDP per capita of the typical countries and regions represented by each language speakers in 2012.

Table 1. Structure of Asphalt Pavement of Test Section Statistics for each Language Index in 2012

Native Language	Country	Quantity of the Migration Population (Ten thousand)	Net Number of Tourist Entry and Exit (Ten thousand)	GDP Per Capita($)	Native Speakers (Million)
Mandarin Chinese (incl. Standard Chinese)	China	-158.121	-7278.7	20647.7	842
Spanish	Spain	-57	4504.2	28976	323
	Mexico	-30	857	10123	
	Total	-87	5361.2	19549.5	
English	United Kingdom	99	-2725.6	38591	320
	United States	450	596	49802	
	Canada	114.5721	-1593.2	50826	
	Australia	91.3107	-218	46330	
	Total	754.8828	-3940.8	46387.25	
Hindustani (Hindi/ Urdu)	India	-212.457	-834.2	1447	259
Arabic	Egypt	-27.5026	551.8	3109	293
Bengali	Bangladesh	-252.648	-214.8	791	193
Portuguese	Portugal	-14	614.2	19768	175
	Brazil	1.5924	-283.5	12340	
	Total	-12.4076	330.7	16054	
Russian	Russia	101.7884	-1963.6	13765	147
Punjabi	Punjab	-7.3	-4.7	1468	145
	Delhi	5.1	-7.1	2746	
	Total	-2.2	-11.8	2107	
Japanese	Japan	35.8133	-1013.3	46896	128

2.1.4. Data collection

We find the data from www.imf.org/en/Data and www.un.org/zh/data bases[1], after that, we sort out data and draw the Table1.

2.2. Model description and solution

Two or more influencing factors as independent variables to explain the change of dependent variable, called multiple regression, when the relationship between self and dependent variables is a linear change, called multiple linear regression. We suppose X is an independent variable, Y is a dependent variable, and establish MLR model of factors influencing languages as follows[2]:

$$Y=b_0+b_1X_1+b_2X_2+b_3X_3$$

Among them, X_1 represents the quantity of the migration population, X_2 represents the net number of tourist entry and exit, X_3 represents GDP per capita, while Y is the number of native language speakers, b_0 is a constant term, b_1, b_2, b_3 are regression coefficients[3]. Based on the previously obtained data (10 languages, 3 main factors), we use MATLAB to get the results: b_0=19271, $b_1=-32.3892$, $b_2=-4.1901$, $b_3=0.2843$ and the multiple linear regression equation is $y=-32.389x_1-4.1901x_2+0.2843x_3+19271$

3. Model 2: Time Series Model

3.1. Extra Symbols

Signs and definitions indicated above are still valid. Here are some extra signs and definitions.

Table 2. Symbols

Items	The cost time
t	Current period
m	Number of periods
a_t	Intercept
M_t	Moving average

3.2. Predicting future change

Considering obvious trend changes in the time series[4] between mother tongue users and total language users in each country in this study, we need to amend the lag deviation caused by the simple moving average method and

the weighted moving average method. The correction method is following Quadratic moving average, the use of moving average lag deviation rule to establish a straight line trend forecasting model.

Average of one move is:

$$M_t^{(1)} = \frac{1}{N}(y_t + y_{t-1} + \mathbf{L} + y_{t-N+1})$$

A moving average on the basis of a moving average is the second moving average, which is calculated is

$$\begin{aligned} M_t^{(2)} &= \frac{1}{N}(M_t^{(1)} + \mathbf{L} + M^{(1)}_{t-N+1}) \\ &= M^{(2)}_{t-1} + \frac{1}{N}(M^{(1)}_t - M^{(1)}_{t-N}) \end{aligned} \tag{1}$$

Next we discuss how to use the moving average hysteresis bias to establish the prediction model:

First set the time series $\{y_t\}$ from a certain period of time that has a straight line trend, and that the latter period are based on this trend changes. So the straight line trend forecasting model is

$$\hat{y}_{t+m} = a_t + b_t m, \quad m = 1, 2, 3\mathbf{L} \tag{2}$$

Then, we determine the smoothing factor based on the moving average. From (2), we can see that is

$$a_t = y_t$$

$$y_{t-1} = y_t - b_t$$

$$y_{t-2} = y_t - 2b_t$$

$$\mathbf{L}$$

$$y_{t-N+1} = y_t - (N-1)b_t$$

Then you get

$$M_t^{(1)} = \frac{1}{N}(y_t + y_{t-1} + \mathbf{L} + y_{t-N+1})$$

$$= \frac{1}{N}y_t + (y_t - b_t) + \mathbf{L} + [y_t - (N-1)b_t]$$

$$= y_t - \frac{N-1}{2}b_t$$

So:

$$y_t - M^{(1)}_t = \frac{N-1}{2}b_t \tag{3}$$

From (2), we can see that is

$$y_{t-1} - M^{(1)}_{t-1} = \frac{N-1}{2}b_t \tag{4}$$

So :

$$y_t - y_{t-1} = M^{(1)}_t - M^{(1)}_{t-1} = b_t \tag{5}$$

Similarly, From (3), we can deduce that is

$$M^{(1)}_t - M^{(2)}_t = \frac{N-1}{2}b_t \tag{6}$$

By (3) and (6), we can calculate smoothing coefficient formulas

$$\begin{aligned} a_t &= 2M^{(1)}_t - M^{(2)}_t \\ b_t &= \frac{2}{N-1}(M^{(1)}_t - M^{(2)}_t) \end{aligned} \tag{7}$$

3.3. Predict the number of users

In order to test the reliability of the results, we combine the number of native speakers and the total number of linguistic users found in question A and the related extranet to select the number of native speakers and the total number of users from 1995 to 2015 that every five years Of the number of linguistic users ranked in the top 20, and then predict the number of native speakers and total linguistic users in the next 50 years by combining the trend-moving averages[5] in the time series model. The data of the two are shown in the following Table 3 and 4 respectively.

Table 3. The Number of Native Speakers in 1995-2015 (million)

Languages	1995	2000	2005	2010	2015
Mandarin Chinese	835	845	873	842	900
English	330	322	309	320	339
Spanish	300	332	222	323	323
Haidi	250	182	181	259	260
Arabic	200	155	293	288	295
Portuguese	160	170	174	175	230
Russian	160	170	120	147	150
Bengali	185	189	171	293	205
French	75	80	64	167	80
Urdu	53	56	59	60	68
Punjabi	90	85	61	138	146
Japanese	125	110	121	128	128
Persian	56	56	62	58	60
Swahili	13	20	15	13	16
Telugu	62	65	69	67	80
Vietnamese	65	63	66	60	78
Javanese	80	81	75	84	84
Tamil	73	76	66	64	74
Korean	56	58	65	69	77
Italian	60	60	64	63	65

Table 4. The Number of Total language Users in 1995-2015 (million)

Languages	1995	2000	2005	2010	2015
Mandarin Chinese	1120	1200	1066	1020	1090
English	480	390	508	825	942
Spanish	320	388	548	338	570
Hindi	260	405	301	270	385
Arabic	221	452	375	370	380
Portuguese	188	193	208	207	262
Russian	285	250	230	257	260
Bengali	188	250	211	283	224
French	265	120	114	180	220

Urdu	150	148	152	154	162
Punjabi	100	86	68	146	147
Japanese	133	123	200	128	130
Persian	112	102	105	108	110
Swahili	86	87	86	95	98
Telugu	68	83	74	79	92
Vietnamese	65	68	76	87	91
Javanese	93	96	90	100	86
Tamil	63	68	76	66	85
Korean	76	80	70	74	79
Italian	86	96	86	124	78

Taking Chinese as an example, we predict the number of native speakers in 2020:

Taking N = 3, we could calculated that once and twice the moving average is

$M^{(1)}{}_{21} = 0.333$, $M^{(2)}{}_{21} = 320$

Then by the formula (7) you can get that is

$a_{21} = 2M^{(1)}{}_{21} - M^{(2)}{}_{21} = 325.333$

$b_{21} = \frac{2}{-1}(M^{(1)}{}_{21} - M^{(2)}{}_{21}) = 2.6667$

So when has $t = 2015$, the prediction model of the straight-line trend is

$\hat{y}_{5+k} = a_{21} + kb_{21}$

The number of predicted native speakers in 2020 is

$\hat{y}_{2020} = \hat{y}_{22} = 328$

By the same token, we use MATLAB to get the number of native speakers and the total number of linguistic users in the next 50 years. The summary data is shown in the following Table 5.

Table 5. Top-ten Native Speakers or Total Language Speakers List in 2065 (million)

	1	2	3	4	5	6	7	8	9	10
NATIVE	MAN DARIN CHINESE	HINDI	BENGALI	PORTUGUESE	ENGLISh	PUNJABI	RUSSIAN	ARABIC	SPANISH	FRENCH
TOTAL	ENGLISH	MAN DARIN CHINESE	SPANISH	PORTUGUESE	RUSSIAN	ARABIC	PUNJABI	FRENCH	BENGALI	HINDI

3.4. Judgment

As the trend moving average method for the existence of linear trends and periodic fluctuations of the sequence is a kind of both reflect the trend changes, but also can be effectively separated from the cycle of changes in the method. So the verification of this model is based on the number of native speakers of English from 1995 to 2015. Then we use MATLAB software to make the scatter plot of the object and scatter plot in Figure 1, for determining whether the trend of moving average method can predict the number of native speakers of English. In the end, we verify that the model is used correctly.

As Figure 1 shows, the number of native speakers in English decreased linearly from 1995 to 2015, and increased from 1995 to 2015. Therefore, the number of native speakers of English in the next 50 years is predicted to be a time series model.

Then in MATLAB, the number of L1 native speakers in 2017 was calculated using the time series model. The result was 380 million. The actual statistical result was 371 million. The theoretical prediction data is similar to the actual statistical results, which shows that the prediction result is correct.

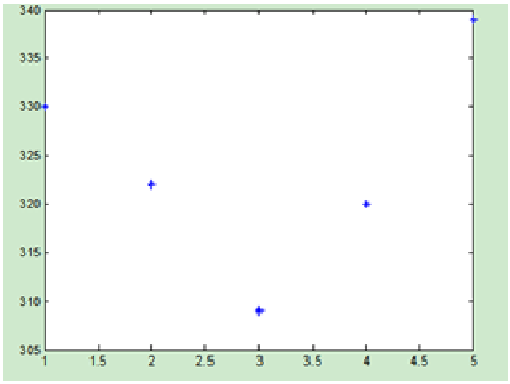

Figure 1. The Number of Native Speakers of English in 1995-2015

4 . Conclusions

In this paper, based on the data from UN and IMF website, we set up a MLR model to simulate the distribution of various language speakers and time series model to forecast language substitution in Top-ten Total language speakers List in the next 50 years. we are required to offer the recommendation as to investigate trends of global languages and location options for new offices by this way.

Acknowledge

Tianjin Agricultural University student innovation and entrepreneurship training project (No. 201710061127).

References

[1] https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita

[2] An R L, Hou J C.Characterization of derivations on tri-angular rings: Additive maps derivable at idempotents[J].Linear Algebra Appl, 2009, 431: 1070-1080.

[3] Jing W.On Jordan all-derivable points of B(H)[J]. LinearAlgebra Appl, 2009, 430: 941-946.

[4] Lei Wu.Parameter estimation and prediction of time series models under censored data [D]. East China Normal University, 2011.

[5] Wang Guixin, Pan Zehan. Spatial Distribution of Floating Population in China and Its Influencing Factors - Based on the Analysis of the Sixth Census Data [J] .New Urban Studies, 2013, 28(3): 4-13. [Wang Guixin, Pan Zehan. China's floating population spatial distribution and influencing factors: Evidence fromyear 2010 population census of China. Modern Urban Research, 2013, 28 (3): 4-13.

[6] Atta.R.T.Boutraa, and A, Akhha.2011. Smart imigation System for Wheat in Saudi Arbia Using Wireless Sensors Networks Technololy. International Journal.478-B2.

[7] Wang Xuejun.The combination of spatial analysis technology and geographic information system[J].Geography research,1997.16(3):70-74.

[8] Zhu chuangeng, Gu qiuling, Ma ronghua. Factors and spatial distribution of the floating population in China[J]. Journal of Geography, 2001,56(3)L549-560.

[9] Heider F. The psychology of interpersonal relationships[M]. [S.I]: Psychology Press,1982:1-317.