# Prediction of Language Development and Establishment of International Offices

Fanfan Wu, Xiaofeng Ji, Hui Li
Qingdao University of Technology, Qingdao, 266520, China

**Abstract:** First, correlation analysis is conducted on the data of influencing factors.Based on the data of its number and influencing factors, multiple linear regression equation model 1 and model 2 are established; Secondly, a time series prediction model is established based on the 5 year data of these influencing factors,and the trend moving average method is used to predict the data of these influencing factors after 50 years; Then, using the algorithm represented by pseudo code and three kinds of network diagrams realized by e-chars,the agent model of language competition is established to describe the change of geographical distribution of language; Then,using the algorithm represented by pseudo code and three kinds of network diagrams realized by e-chars, the agent model of language competition is established to describe the change of geographical distribution of language.

**Keywords:** Multivariablelinear regression model; Time series model; Languagecompetitionagentmodel; Multi-objectiveprogrammingmodel; Improved Multi-objective programming model

## 1. Introduction

### 1.1. Multivariable Linear Regression Model

In order to find the data for each of factors,we make a table for the main countries in which languages are located.

**Table 1. Varies language native speakers**

| Language | The main country | Language | The main country |
|---|---|---|---|
| Mandarin | China,Singapore | Punjab. | Pakistan |
| English | Britain,America,Australia | Vietnamese | Vietnam |
| Hindi | India,Fiji | Tamil | India,Sri Lanka |
| Spanish | Spain,Mexico,Peru | Java language | Indonesia |
| Russian | Belarus,russia, | Telugu | India |
| Arabic | Arab countries | Turkish | Bulgaria,Turkey |
| German | Belgium,Germany | Korean | North Korea,South Korea |
| Bengali | Bangladesh | Malay | India |
| Portuguese | Brazil, Portugal | Italian | Croatia,Italy,Switzerland |
| French | France,Italy | Thai | Thailand |
| Japanese | Japan | Farsi | Afghanistan,Iran |
| Uhl. | India,Pakistan. | | |

Throughtheaccess to information [2],weselected ten important factors to analyze the language population distribution.

**Table 2. Definition and symbol description**

| Notation | Definition | Unit |
|---|---|---|
| xi | Factor value | N/A |
| y | Language speakers | Million |
| y2 | Native language speakers | Million |
| bi | Regression coefficient | N/A |
| ri | Risidual | N/A |
| S | Standard deviation | N/A |

We use Excel to test the correlation of ten factors and generate the correlation coefficient matrix.

In the correlation coefficient matrix,A value greater than 0.8 indicates that there is a strong correlation between the two.Therefore,we selected six out of 10 factors as the main factors,such as comprehensive national strength, immigrant population, unemployment rate,The index of informationdevelopment, the growth rate of the population and the number of tourists,and establish the multiple linear regression model 1.

$$
\begin{aligned}
&y = b_0 + b_1 x_1 + \mathbf{L} + b_m x_m \\
&r = N(0, s^2)
\end{aligned} \qquad (1)
$$

Taking six main factors as independent variable and the number of language speakers as the dependent variable.Then we get the values of each parameter by MATLAB.

**Table 3. Parameter value**

| | Value | bi |
|---|---|---|
| b0 | 0.446 | (-0.46451,1.35669) |
| b1 | -0.623 | (-1.92901,0.68219) |
| b2 | 0.340 | (-0.60517,1.28536) |
| b3 | 0.013 | (-0.87552,0.90184) |
| b4 | 0.066 | (-0.76399,0.89705) |
| b5 | 0.838 | (-0.14809,1.82507) |
| b6 | -0.354 | (-1.34566,0.63598) |

The resulting multiple regression equation is:

$$y = 0.446 - 0.623x_1 + 0.34x_2 + 0.013x_3 + 0.066x_4 + 0.838x_5 - 0.354x_6 \quad (2)$$

## 2. Time Series Model

Using the same method as before,we establish a multi-regression linear model 2 of native speakers and four influencing factors.

$$y_2 = 0.119 - 1.213x_1 - 0.061x_2 + 0.759x_3 + 0.698x_4 \quad (3)$$

We consider to set up time series model to obtain the value of each factor of native English 50 years later.For each of the influencing factors $x_i$（$i = 1, \mathbf{L}\, k$） ，theobservation sequence is $x_{i1}, \mathbf{L}, x_{iT}$ .We take the moving average of the number of items $N < T$ .The moving average is calculated as:

$$\begin{aligned} M_{it}^{(1)} &= \frac{1}{N}(x_{it} + x_{it-1} + \mathbf{L} + x_{it-N+1}) \\ &= M_{t-1}^{(1)} + \frac{1}{N}(x_{it} - x_{it-N}) \end{aligned} \quad (4)$$

Because of the slow changes of various influencing factors,the moving average method can be used to establish the predictive model:

$$\overset{\wedge}{x}_{it+1} = M_{it}^{(1)} = \frac{1}{N}(x_{it} + x_{it-1} + \mathbf{L} + x_{it-N+1}), \quad (5)$$

$$t = N, N+1, \mathbf{L}, \mathrm{T}$$

The standard deviation of prediction is:

$$S = \sqrt{\frac{\sum_{t=N+1}^{T} (\overset{\wedge}{x}_{it+1} - x_{it})^2}{T - N}} \quad (6)$$

After iteratively running in MATLAB,we obtain the number of native speakers of English 50 years later and the number of English speakers after 50 years.Sort into the following table.

**Table 4. Number Comparison（million）**

| | Native English speakers | English speakers |
|---|---|---|
| Now | 371 | 983 |
| After 50 years | 484.4 | 1497.5 |

The result shows that both of the two will increase after 50 years.

We use the trend moving average method to predict the number of language users from 50 years.Then rank them against the current rankings of language users.So we can analyze which language will be replaced.Sort the data and get Table 5.

**Table 5. Comparison of language usage rankings**

| The current rankings | Ranking 50 years later |
|---|---|
| Mandarin | English |
| Spanish | Mandarin |
| English | French |
| Hindi | Spanish |
| Arabic | Russian |
| Bengali | German |
| Portuguese | Japanese |
| Russian | Bengali |
| Punjabi | Portuguese |
| Japanese | Korean |

Hindi,Arabic and Punjabi will be replaced by French,German and Korean after 50 years.The number of speakers in English,Japanese and Russian will rise sharply,while the number of speakers in Putonghua and Arabic will plummet.

## 3. Language Competition Agent Model

We establish the language competition agent model.The pseudo-code and network diagram are as follows.

**Table 6. Variable definitions**

| Notation | Definition |
|---|---|
| W | Network density |
| X | Monolingual |
| Y | Mono 2 |
| Z | Bilingual |
| RS | Vulnerable language status |
| YS | Advantageous language status |
| Step | Steps |
| F | Fragmented mixed |
| D | Independent settlement |
| WF | No individual step back |
| FS | Return individual steps |

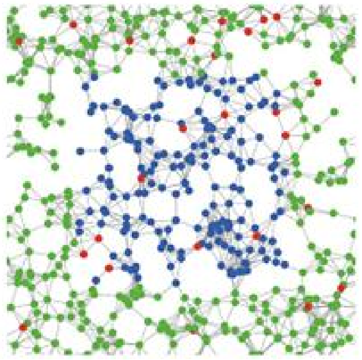
### 3.1. Mixed structure pseudo-code

```
1: Create 1000 agents and randomly distributed in 200 × 200 two-dimensional space.
2: W =0.62%;
3: X =the number of monolingual 1 people in the country as a percentage of the total population;
```

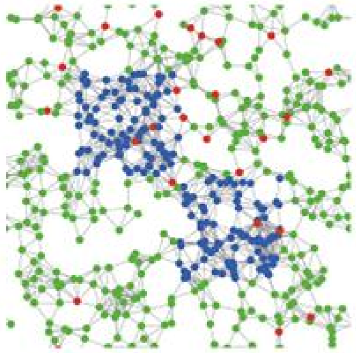
```
4: Y =the number of monolingual 2 people in the country as a percentage of the total population;
5: Z =the bilingual number of the country as a percentage of the total;
6: YS =0.6;
7: RS =0.4;
8: YS + RS =1 ;:
9: Set the social radius of individuals SR, the establishment of social circle network;
10: Establish social networks based on social circle theory;
11: Set a certain proportion of individual attributes in daily life for short-term flow in the established social network model of mixed structure;
12: Step =1;
13: If (individuals have daily short-term flow properties within each time step)
{Individual move one step;}
14: If (time step increase by 1)
{Individual moves one step}:
15: While (individual network structure in line with)
{Rebuild the network individual structure;
Goto 10;}
```

### 3.2. Extroverted and intrusive pseudo-code

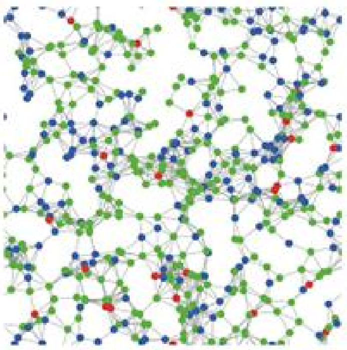
```
1: The establishment of 1000 agents distributed in 200 × 200 two-dimensional space
2: W =0.62%
3: Demarcate the spatial area by coordinates and divide the network into F and D
4: F coordinate range:x belongs to [-40,40], y belongs to [-100,100], as the concentration area of weak language
5: D coordinate range: x belongs to [-120,120], y belongs to [-120,120] as the concentration area of weak language
6: Y =the number of monolingual 1 people in the country as a percentage of the total population;
7: X =the number of monolingual 2 people in the country as a percentage of the total population;
8: Z =The bilingual number of the country as a percentage of the total;
9: YS =0.6;
10: RS =0.4;
11: YS + RS =1;
12: Set the social radius R for each agent
13: Establishing social networks based on social circle theory SNET;
14: In the already established network of spatial distribution structure
15: set D and F return to the proportion of individuals and non-return to the proportion of individuals were set to return to the mobile property and no return
16: Set WF> SR and FS> SR
17: stopped moving across the area
18: Set the time interval returned by the individual
19: If (to return to individuals in different regions of residence time is full)
{Back to the flow;}
20: If (time step increase by 1)
{Different individuals return and no return flow separately;}
21: While (network disconnected or rebuilt)
{Goto builds social networks based on social circle theory;}
```



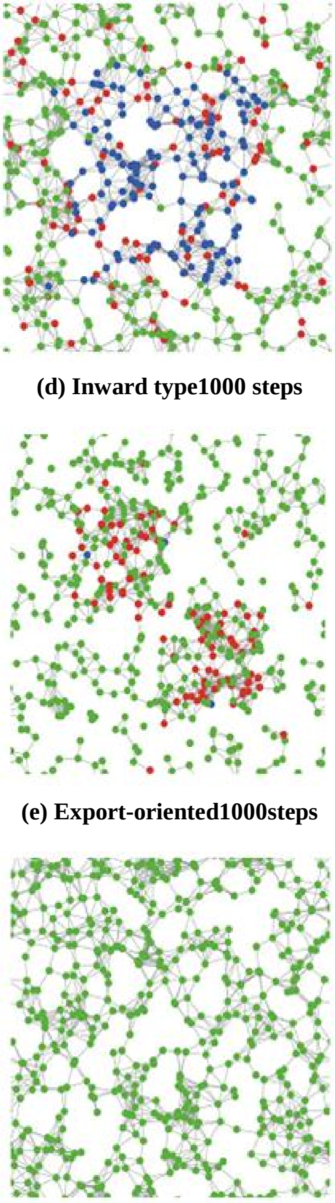
**(a) Inward type initial**



**(b) Export-oriented initial**



**(c) Mixed initial**



**(d) Inward type1000 steps**

**(e) Export-oriented1000steps**

**(f) Mixed 1000 steps**

Wefind that Bengali might disappear over time.The monolington of Bangladesh is Bengali and English,so Bangladesh should be a segmented mixed structure and the dominant language is English. Bangladesh's native language may be the only English language 50 years later.

In the same way,we find the other nine languagesis increasing.Therefore,these languages expand on the basis of previous countries and regions.

## 4. Multi-Objective Programming Model

The choice of offices should cover as many languages as possible and the distance between offices is as large as possible.Then we create a multi-objective programming model.

Objective function:

$$\max \sum x_i \cdot f(i) \tag{7}$$

In the formula,f (i) indicates the number of languages used by country i.

Objective function:

$$\max \sum_{i,j} d\left(p(i), p(j)\right) x_i x_j \tag{8}$$

In the formula, P (i) represents the specific location of the office i, p (j) represents the specific location of the office j, and d (p (i), p (j)) represents the distance between the office i and the office j.

The target planning constraints are:

There are only two options for each country.One is to become an office and the other is to not become an office.

There must be 6 offices to choose from, so $\sum x_i = 6$

The objective programming equation is as follows:

$$\max \begin{cases} \sum x_i \cdot f(i) \\ \sum_{i,j} d\left((i),(j)\right) \cdot x_i x_j \end{cases} \qquad s.t. \begin{cases} x_i = 0,1 \\ \sum x_i = 6 \end{cases} \tag{9}$$

We select 16 countries,from which we select six offices.Then,we use the capital of each country as a candidate for the office,Calculate the distancesd (p (i), p (j)) between the longitude and latitude of the capitals of each country.Finally, we use Matlab to solve the specific location of the six offices,and select the office language.

**Table 7. Office location and language used**

| Office location | Use language |
|---|---|
| Australia | English,Spanish |
| United Kingdom | English |
| India | English,Hindi,French,Bengali |
| Brazil | Portuguese,English |
| Canada | French,English |
| South Africa | English |

The long-term presence of offices also depends on social safety factors and traffic indices.The lower the social security factor and the traffic index,the less suitable as an office.

**Table 8. Office of the national security index and traffic index**

| Office location | National Security Factor | traffic index |
|---|---|---|
| Australia | 56.88 | 8.5 |
| United Kingdom | 57.23 | 9 |
| India | 53.41 | 5.6 |
| Brazil | 28.77 | 6.9 |
| Canada | 60.8 | 7.8 |

| | | |
|---|---|---|
| South Africa | 21.57 | 6.2 |

By analyzing the table above,we find that social security and traffic indices in South Africa and Brazil are relatively low.We replaced Brazil with Argentina,where security index is 40.72 and traffic index is 7.2.South Africa is still chosen as an office because of the generally low safety factor in African countries.

## 5. Improved Multi-objective Programming Model

Taking changes in the nature of communications into account,based on the multi-objective programming model,we add the sum of the communication indicators between every two countries as the minimum objective function.

$$\max \begin{cases} \sum x_i \cdot f(i) \\ \sum\limits_{i,j} d\left((i),(j)\right) \cdot x_i x_j \end{cases}$$

$$Min\ sum(tx(i,j) \cdot x(i) \cdot x(j)) \qquad (10)$$

$$s.t. \begin{cases} x_i = 0,1 \\ \sum x_i = 6 \\ ceq2 = sum\left(x(i)\right) - 6 \end{cases}$$

In the formula, $tx(i,j)$ represent the communication between countries indicators.

Then use MATLAB to write the program and get the location of the office as shown in the following table.

**Table 9. New office location and language**

| Office location | Use language |
|---|---|
| Australia | English,Spanish |
| India | English, Hindi, French, Bengali |
| Brazil | Portuguese, English |
| Canada | French, English |
| South Africa | English |

Of the six previously obtained countries,Australia and the United Kingdom have higher communication targets.One office can be chosen between the two,while Australia is more than the UK to meet the improved model.

## 6. Conclusion

After 50 years,English still occupies a dominant position.With the development of Japan,Russia and other countries,Japanese and Russian have a place in the world.At the same time, some languages are gradually losing or even disappearing.

## References

[1] https://wenku.baidu.com/view/469322a20722192e4436f60f.html

[2] http://www.un.org/zh/index.html

[3] http://www.ilo.org/global/lang--en/index.htm

[4] https://migrationdataportal.org/?i=stock_abs_&t=2017

[5] http://www.askci.com/news/hlw/20161227/14221885057.shtml

[6] DaquanGan.Research on Global Language Trends [M].Shanghai Normal University.2007.5

[7] https://www.bls.gov/home.htm

[8] https://esa.un.org/unpd/wpp/