The Model of How Languages will Develop

W. Sun*, J. Wang, H. Zhang and X. Hou

Department of Mathematics, Shandong University of Technology, Zibo 255049, China

Corresponding author email: sunwenhua@sdut.edu.cn

Abstract. The total number of speakers of a language may increase or decrease over time because of various and complicated influences. In this paper, we build the quantitative prediction model based on gray estimation, FCM model and geographical distribution prediction model to predict the distribution of language users. In addition, we conducted a sensitivity analysis of the model to facilitate future research.

Keywords: Languages, gray estimation, FCM, geographical distribution, offices selection.

1. Introduction

In the process of globalization, strong language spread rapidly and firmly established hegemony in the world language jungle. However, with the reduction of linguistic functions and the decrease of the number of users, the weakened languages slowly lose their vigor and vitality and gradually become endangered or even extinct. The state limits or promotes a certain language due to national interest. The languages used in schools, migration and assimilation of cultural groups, and immigration and emigration with countries that speak other languages. Moreover, with the increasing development of economic globalization, the impact on culture and language has also gradually emerged. The rapid development of multi-media, the rapid popularization of the Internet and the promulgation and promotion of cultural policies in other countries all will aggravate the speed of language distribution changes and exert a tremendous influence on the distribution of the world's languages[1][2].

2. Symbols

Table 1. Variables and Symbols

| Symbols | Descriptions                  |
|---------|-------------------------------|
| $B$     | Coefficient Matrix            |
| $\rho$  | Latitude Coefficient          |
| $R$     | The total amount of predicting speakers |
| $Q$     | Factors Index                 |
| $S$     | Resource Utilization          |
| $K$     | Known Domain                  |
| $\eta$  | Correlation Coefficient       |
3. Predictive Model

3.1. Gray Prediction Model

Based on the gray estimation prediction model of the number of speakers in various languages. The gray prediction model can be used a first-order differential function to describe with respect to time variable. The corresponding differential equation is:

\[
\frac{dx^{(1)}}{dt} + ax^{(1)} = u
\]  

(1)

In the above formula (1) \(x\) is a cumulative generation sequence; \(t\) represents time; \(a,u\) represents the estimated parameters and the gray number for endogenous control, respectively. First, we create a cumulative number of columns generated. Let the original sequence be

\[
x^{(0)} = \{x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), \ldots, x^{(0)}(n)\}
\]

(2)

for \(t = 1, 2, \ldots, n\), according to

\[
x^{(1)}(t) = \sum_{m=1}^{t} x^{(0)}(m), t = 1, 2, \ldots, n.
\]

(3)

Once accumulated, we can get the number of generations which denoted by

\[
y_n = \begin{bmatrix} x^{(0)}(2), x^{(0)}(3), \ldots, x^{(0)}(n) \end{bmatrix}^T,
\]

(4)

then we can use the least squares method to find the parameters \(a, u\) as

\[
\widehat{a} = \begin{bmatrix} a \\ u \end{bmatrix} = (B^T B)^{-1} B^T y_n.
\]

(5)

(6)

Then we found the quantitative predictive model\[3\]

\[
x^{(1)}(t + 1) = (x^{(0)}(1) - \frac{u}{a}) e^{-at} + \frac{u}{a},
\]

(7)

\[
\begin{cases}
\hat{x}^{(0)}(1) = x^{(1)}(1), \\
\hat{x}^{(0)}(t) = x^{(1)}(t) - \hat{x}^{(1)}(t-1), t = 2, 3, \ldots, n.
\end{cases}
\]

(8)

We have date from 2008 to 2017 from Wikipedia on various languages populations including native speakers and total speakers as shown in the Table 1.
Table 2. Number of Mandarin Chinese in each year.

| Mandarin Chinese | 1997 | 2008 | 2012 | 2015 | 2017 |
|------------------|------|------|------|------|------|
|                  | 1120 | 1051 | 1013 | 1030 | 1090 |

\[
x^{(0)} = \{1120, 1051, 1013, 1030, 1090\}, \quad x^{(0)} = \{1217, 3184, 4214, 5304\}
\]

\[
B = \begin{bmatrix}
    -1645.5 & 1 \\
    -2677.5 & 1 \\
    -3699 & 1 \\
    -4759 & 1
\end{bmatrix}, \quad y_n = \begin{bmatrix}
    1051 \\
    1013 \\
    1030 \\
    1090
\end{bmatrix}
\]

According to the above procedure, the results can be obtained

\[
\hat{a} = \begin{bmatrix}
    a \\
    u
\end{bmatrix} = \left( B^T B \right)^{-1} B^T y_n = \begin{bmatrix}
    -0.01 \\
    -1004.3
\end{bmatrix},
\]

\[
x^{(1)}_{10} (t+1) = 101550 e^{0.0001t} - 100430.
\]

The following nine languages prediction can be drawn from the above relationship

- **Spanish**: \( x^{(0)}_{11}(t+1) = 2864 e^{0.121t} - 2544 \)
- **English**: \( x^{(0)}_{12}(t+1) = 1553 e^{0.257t} - 1073 \)
- **Hindustani(Hindi)**: \( x^{(0)}_{13}(t+1) = 5267 e^{0.0750t} - 5017 \)
- **Arabic**: \( x^{(0)}_{14}(t+1) = 7257 e^{0.0686t} - 7036 \)
- **Bengali**: \( x^{(0)}_{15}(t+1) = 4432 e^{0.0477t} - 4247 \)
- **Portuguese**: \( x^{(0)}_{16}(t+1) = 1906 e^{0.0930t} - 1718 \)
- **Russian**: \( x^{(0)}_{17}(t+1) = 10670 e^{0.0234t} - 10385 \)
- **Punjabi**: \( x^{(0)}_{18}(t+1) = 514 e^{0.2099t} - 426 \)
- **Japanese**: \( x^{(0)}_{19}(t+1) = 6128 e^{0.0198t} - 5995 \)

### 3.2. FCM Model

Due to the uncertainty of the distribution of language users in different regions, a country will generally restrict or promote a certain language for the sake of national interests. In addition, the environmental capacity of a certain area, the standard language of a certain area and the dialect all influence the languages. Many languages make "mother tongues" complicated. For example, a native speaker of Chinese who immigrated to the United States, then English may be the second language, and their descendants may become native speakers of English. The percentage of world population growth also affects the distribution of language users. Due to the complexity of various factors, we have to screen out some of the more important factors to facilitate the next study.

Let \( d(x, y) \) be a distance function, \( A^p \) and \( A^q \) represent the state value vector at any two moments in the FCM model; \( g(A^p) \), \( g(A^q) \) denote the vector of system state values at \( A^p \) and \( A^q \) at the next moment obtained by the FCM derivation function, then we define:

If the vector of states \( A^p, A^q \) at any time of the fuzzy cognitive graph model, the

\[
d(g(A^p), g(A^q)) \leq \mu d(A^p, A^q), 0 < \mu < 1
\]

is constant established. Then, the FCM model will converge to the steady state. That is, for the state value vector \( A^p, A^q \), at any time
\[
\frac{d[g(A^p)g(A^q)]}{d[A^p, A^q]} = \left\| g(A^p)g(A^q) \right\| \leq \mu < 1
\]

is constant established. That is, if the FCM model of the adjacency matrix \( W \) can make the slope of the formula (13) less than 1, the fuzzy cognitive graph model will converge to the only stable state. When using the sigmoid function as the threshold function, let \( \lambda = 1 \) we know that the slope of the threshold function is equal to \( \mu \leq 1 \). Let \( w_i = [w_{i1}, w_{i2}, ..., w_{in}] \), then for the whole system, combined the knowledge of FCM derivation function and matrix theory, if FCM model is stable, there are

\[
\left| f \left( W_iA^p \right) - f \left( W_iA^q \right) \right| < \mu \left| A^p - A^q \right|, 0 < \mu < 1.
\]

Therefore, the formula for calculating the entire system can be derived

\[
\left\| g \left( A^p \right) - g \left( A^q \right) \right\| = \sqrt{\sum_{i=1}^{n} \left( \left| f \left( W_iA^p \right) - f \left( W_iA^q \right) \right| \right)} < \sqrt{\sum_{i=1}^{n} \left( \mu \left\| W_i \right\| \left\| A^p - A^q \right\| \right)^2},
\]

\[
\left\| g \left( A^p \right) - g \left( A^q \right) \right\| < \sqrt{\sum_{i=1}^{n} \left( \left\| W_i \right\| \right)^2} \left\| A^p - A^q \right\|.
\]

Therefore, if the FCM model has a unique equilibrium state, there is

\[
\mu \sqrt{\sum_{i=1}^{n} \left( \left\| W_i \right\| \right)^3} < 1.
\]

Then the FCM model can be written as[4]:

\[
\begin{align*}
&f_N(x \mid \xi_{N-1}) = \sup_{d \in D_N} R_N(x, d \mid \xi_{N-1}) \\
&f_N(x \mid \xi_{N-1}) = \sup_{d \in D_N} R_N(x, d \mid \xi_{N-1}) \\
&\quad + \sum_{i=1}^{n} \frac{1}{P_n} (x + \xi_i - d)P_n (\xi_i \mid \xi_{N-1})
\end{align*}
\]

\(n \leq N - 1\)
\(f = 1, 2, 3, ..., k\)

The relationship between the various factors shown in Figure 1. \(C, D, E, F, G, H\) represents environmental capacity, social pressure, cultural groups, schools, electronic and social media respectively, and \(\rho\) represents latitude coefficient. We can get the relationship between the distribution of the number of users in various languages \(Q\) and each factor is

\[
Q = \frac{cC + dD + (eE + fF + hG)H}{\rho},
\]

where \(c, d, e, f, g, h\) are the coefficients of each factor.

From the formula we saw that the above factors also caused the change of the number of users, while the correlation between languages would also change with these factors. With the migration of population and the increase in the global population, the number of languages moved out transformed into the number of languages moved in. Then, the relevance of these two languages would increase. From this we introduced a correlation coefficient \(\eta\) to represent the correlation between languages and described the distribution of the number of users in each language[5].
\[ R = \left( x^{(0)}(t+1) + Q \right) \eta. \]  

(20)

### 4. Sensitivity Analysis

To test the gray predictive model, we define

\[ L_0 = \sqrt{\sum_{r=1}^{n} \left[ x^{(0)}(t) - x^{(0)}_{-} \right]^2} / (n-1), \]

\[ x^{(0)}_{-} = \frac{1}{n} \sum_{r=1}^{n} x^{(0)}(t). \]  

(21)

The mean square error

\[ L_1 = \sqrt{\sum_{t=1}^{n} \left[ e^{(0)}(t) - e^{(0)}_{-} \right]^2} / (n-1), \]

(22)

where \[ e^{(0)}_{-} = \frac{1}{n} \sum_{t=1}^{n} e^{(0)}(t) \] and \[ e^{(0)}(t) = x^{(0)}(t) - \hat{x}^{(0)}(t). \] So the variance \[ \sigma = L_1 / L_0. \] The error probability

\[ p = \left\{ \left| e^{(0)}(t) - \hat{e}^{(0)}(t) \right| < 0.6745 \times L_0 \right\}. \]  

(23)

### 5. Conclusion

The accuracy of the gray model results obtained from the formula was qualified. But the Mandarin Chinese prediction accuracy is about 0.7, so our model needs a further improvement, which we will study in the future.

From the above test, it is possible to predict the number distribution of language users by the gray model. However, errors may also exist. The reason for this may be that the data amount is not large enough or the model has not reached a better predictive result\[6-7\].

### References

[1] Kamiliangjiang Abdulkarm. *On the status of the world language and its future trends*. Journal of Jiamusi Education Institute, 2013(4) P 181-189.

[2] "Summary by language size". Ethnologue. Retrieved 2016-04-06.

[3] Qiu Yali. *Prediction and Analysis of Tourist Attraction Population Based on Gray Model*. Statistics and Decision, 2013 (17): 114-117.

[4] Chen ZR. *Research on site selection of garbage power plants based on multi-objective optimization model*. North China Electric Power University, 2014.

[5] Wu Shanshan. *Data stream frequent item mining and correlation analysis algorithm*. Zhejiang University, 2017.

[6] ZHANG Lina, LI Chunlan, YAN Guangzhou, ZHANG Yajing. *Note on the Convergence of Weighted Sums of Random Sequences*. Mathematics in Practice and Theory, 2010(40) )169-174.

[7] Vasile Drăgan, Ivan G. Ivanov, Ioan-Lucian Popa. *On the closed loop Nash equilibrium strategy for a class of sampled data stochastic linear quadratic differential games*. Chaos, Solitons & Fractals, August, 2020(187) id.109877.