Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Prediction of language population model using regression theory

Junjie Chen

Department of Electrical Engineering, North China Electric Power University, Hebei, China

imchenjunjie@126.com

Abstract: Most opinion in the literature believe that language population growth cannot be predicted. In this paper, multivariate nonlinear regression demonstrated its advantages in forecasting population of language people in near future. All the factors that may affect the development of a language were took into account and were classified into quantitative indicators. We take Mandarin as an example to determine the value of parameters. The result shows that the data for English match quite well in previous years, and indicates the number of English speakers in 50 years will reach 4 billion.

Keywords: Language prediction, Multivariate nonlinear regression theory, language population.

---

1. INTRODUCTION

There are currently about 6,900 languages spoken on Earth. However, 10 of these languages cover about half of the world's population. If we add the number of second languages speakers, the top 15 languages make up about 75% of the world's population.

The number of speakers of a language may change over time because of a variety of influences including politics, policy, economy, culture, globalization and so on. By studying these influencing factors, we can predict the world's population distribution and the trend of changes in the number of people in each language.

2. PREDICTION

2.1 Analysis of elements and indicators

Language population is an abstract concept, where a variety of complex factors stand behind.

a relatively objective prediction will not come before we find the right metrics to quantify them.

Fig. 1 shows what we take into account in this language prediction. For example, we use the number of books translated from a language to represent culture, GDP as an economic indicator, and edits in Wikipedia as social media indicator.

Policy Group: Policy is almost unpredictable and policies usually have a powerful promoting or restricting effect on things. We consider once the measures are conducted, population of the corresponding language will change dramatically in a short period of time.

Therefore, we take this influence as a constant and adjust it according to the actual situation.

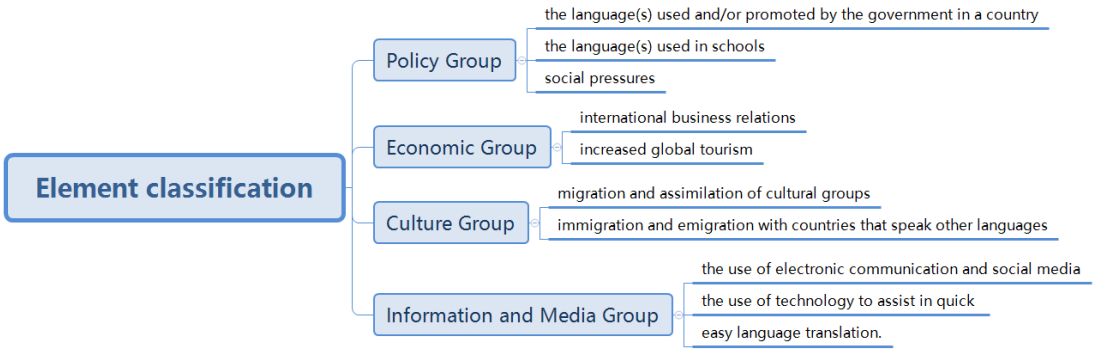

Fig. 1: Element classification

Economy Group: Under the background of knowledge economy and globalization, language is becoming a huge new potential for promoting and restricting the competitiveness of economy. GDP is the core indicator of the national economy, so we use it to indicate prosperity of economic activity.

Culture Group: What is considered here is the international influence of a specific language. When we are considering the external influence of a language, national cultural export is an objective measure of the extent to which migration and cultural assimilation affect language.

Information and Media Group: Accompanied by above-mentioned factors, the prosperity of information and media can have a significant impact on the language's population. We use the number of wiki entries to indicate how active a language is in terms of socialization and communication

2.2 Multivariate nonlinear regression theory

In terms of traditional methods, the main operation is to perform a least squares method on the linear regression model to fit the regression equation.

The linear regression model[1] of the random variable y and the independent variable x is:

$$\mathrm{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon$$

When there are n sets of observation data, the regression model can be expressed as a matrix form:

$$\mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathrm{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

According to the principle of least squares, the chosen estimation method should minimize the residual between the estimated value and the observed value at all sample points.

$$\mathrm{Q} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \varepsilon\varepsilon'$$

$$\varepsilon = \mathrm{y} - \mathrm{X}\hat{\beta}$$

According to Principle of calculus extremum,when ε_{min},

$$\frac{\partial Q}{\partial \hat{\beta}} = \frac{\partial (y - X\hat{\beta})'(y - X\hat{\beta})}{\partial \hat{\beta}} = 0$$

In the processing of multivariate nonlinear equations, x and y can be appropriately changed to construct a linear equation, and then solved by the linear regression equation processing method. Therefore, on the basis of multiple regression, the sum of squares of residuals can be minimized, and the optimal solution can be obtained.

After many attempts and analyzes,we get the following formula.

$$y = \beta_0 + \beta_1 x_1^{1.5} + \beta_2 x_2^{0.5} + \beta_3 x_3 + \varepsilon$$

2.3 Parameter determination

Table 1: Quantitative indicators of Mandarin[2][3][4]

Year	WiKi Edits/104	GDP/109	Cultural export/106
2010	30	6.101	12.52
2011	35	7.573	21.7
2012	45	8.561	22.7
2013	70	9.607	27.9
2014	80	10.482	31.7

Table 1 shows the level of indicators in Mandarin in recent years. Then we use matlab to fit the above data to get the following formula:

$$y = 26.4 + 3.77x_1^{1.5} + 2.427x_2^{0.5} + 0.135x_3 + 50$$

（x_1: the annual GDP; x_2: the number of Wikipedia edits; x_3: the total amount of cultural exports.）

3. HOW MANY ENGLISH SPEAKERS?

Then we apply the model to test English-speaking countries.Putting data in **Table 2** into the existing model for testing and comparison.

Table 2: Quantitative indicators of English[2][3][4]

Year	WiKi Edits/104	GDP/109	Cultural export/106
2010	1000	14.964	20.15
2011	2250	15.518	25.04
2012	3200	16.155	60.1
2013	4000	16.692	78.5
2014	4800	17.393	100.02

As we can see in Fig. 2 ,It fit well, too.

When we set the year in fifty years,we get the number is 4.0114×10^9. That is to say, nearly 4 billion people will speak English in fifty years.

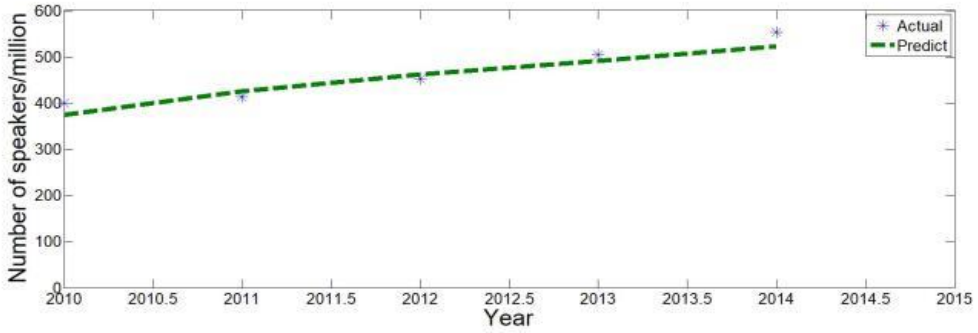

Fig. 2: Comparison of English

4. CONCLUSION

There is a strong argument against attempting forecasting in a sphere of life in which cultural and political factors are so salient [6]. But if we do want to, for example, there have been models to analyze the future of languages in Ethnologue [5]. All of them require input data from demographic and economic forecasts in order to predict demand for languages.

However it is inefficient to analyze all the factors given and there is also overlap between the factors given above. We Identified the most simple and effective factors in this paper that represent the trends of a language in several decades. And apply the method to predict the number of English speakers in fifty years. The verification proves the model fit good. At least in the next few years, this prediction method is still very informative.

REFERENCES

[1] Rencher, Alvin C.; Christensen, William F. (2012), Methods of Multivariate Analysis, Wiley Series in Probability and Statistics. ISBN 9781118391679.
[2] https://www.worldbank.org.
[3] The Globalisation of Cultural Trade: A shift in consumption—International flows of cultural goods and services 2004-2013.
[4] https://www.wikipedia.org/
[5] Ethnologue (2017 20th edition)
[6] David Graddol. The future of English? The British Council 1997, 2000.