Team Control Number

For office use only

T1 _____

T2 _____

T3 _____

T4 _____

74341

Problem Chosen

B

For office use only

F1 _____

F2 _____

F3 _____

F4 _____

2018

MCM/ICM

Summary Sheet

An analysis of the trend of global language

Summary

As we know, the number of language speakers changes over time. In this article, we try to start with existing data, develop appropriate models to describe these changes and provide reasonable advice to client companies.

First, we use a multivariate regression statistical model based on the related data of the previous ten years. For each language, we fit the “the number of language speaker growth rate” and the functional relations of the above factors. Moreover, we can get a direct picture of how the total number and proportion of people changes over time.

Second, according to the migration situation of each country in the previous decade, the migration situation of each language area is obtained; according to the neural network model, the data obtained above is used as training data to predict the migration from one language area to another. Through calculations of 100 groups based on the top ten language areas of the question, we can get the results and mark them on a map so that we can clearly see the changes in geographical distribution of different languages.

Third, in order to determine the reasonable number and address of offices, according to the AHP model, we select the size of the company, the company's industry, the market degree of dispersion of these factors. Furthermore, according to the top five hundred companies of the world related to the above factors to quantitative research, we can get the weight which completes the AHP model. According to the above analysis and results, we write a memo summarizing that the results and a reasonable suggestion are given to the service company's chief operating officer.

In the end, we analyze the superiority and sensitivity of the model, and verify that the modeling method and the results are in accordance with the expected requirements. After that, we analyze the strengths and weaknesses of the model method and give our relevant thinking and possible optimization scheme.

Contents

1 Introduce

1.1 Background description

Language, as a tool for transmitting information, performs the same function of achieving human communication. The distribution of the languages of the world and the population in use is very uneven, since a small number of languages are spoken by a very large population while most languages are spoken by a very small number of people. For example, a large number of languages are used by only a few hundred, dozens or even a few people, according to David Crystal, a famous world language researcher. As many as 96% languages are spoken only by 4% of the world's population. The total number of the world's top 15 speakers accounts for about half of the world's population. The trend of globalization has brought about profound changes in the world. The effects of this trend are, of course, economic and social in the first place, but the effects on language and culture have gradually emerged. That is, to have a common language that will allow all mankind to communicate freely and eliminate the Babel Tower, the language that has made mankind respond.

1.2 Restatement of the problem

Part I:

A. Consider the influences and factors described in the background paragraph above, as well as other factors your group may identify. Based on projected trends, and some or all of these influences and factors, model the distribution of various language speakers over time.

B. Use your model to predict what will happen to the numbers of native speakers and total language speakers in the next 50 years. Do you predict that any of the languages in the current top-ten lists.(either native speakers or total speakers) will be replaced by another language? Explain.

C. Given the global population and human migration patterns predicted for the next 50 years, do the geographic distributions of these languages change over this same period of time? If so, describe the change.

Part II:

A. Based on your modeling from Part I, and assuming your client company wants to open six new international offices, where might you locate these offices and what languages would be spoken in the offices? Would your recommendations be different in the short term versus the long term? Explain your choices.

B. Considering the changing nature of global communications, and in an effort to save your clientcompany resources, might you suggest that the company open less than

six international offices? Indicate what additional information you would need and describe how you would analyze this option in order to advise your client.

Part III:

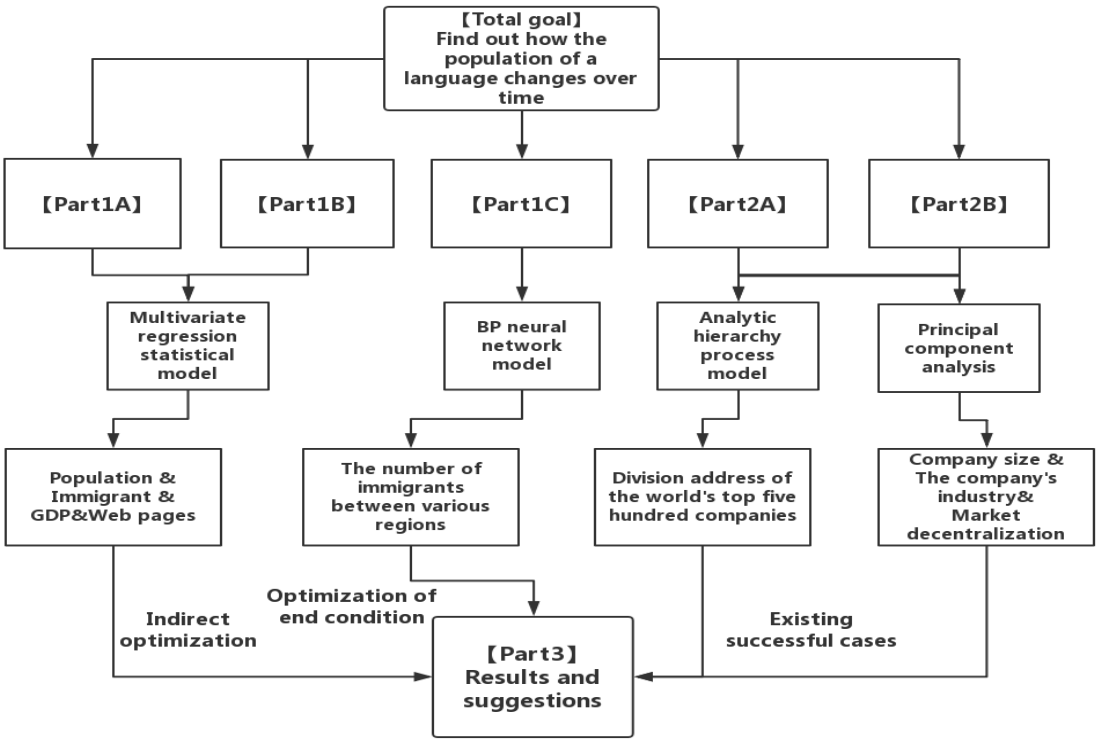
Write a 1-2 page memo to the Chief Operating Officer of the service company summarizing your results and recommendations.

2 Our Work

2.1 Problem Hypothesis

- In the forecast period, there is no big enough to affect the language development track and unpredictable changes, such as major natural disasters, large-scale wars, and so on.
- Only people who have mastered a language and often use this language are counted in the number of people used in the language.
- The staff recruited by the office of the company mastered the local average level of language reference;
- Customer company is a general company, there are no too special factors to interfere with the location and choice of language.

2.2 Analysis of Ideas



Graph 1

2.3 Symbolic Description

Chart 1

Symbol	Meaning
x_1	Natural population growth rate
x_2	Increase in the number of immigrants
x_3	Total GDP
x_4	Per capita GDP
x_5	The growth rate of the number of Web pages
W_{ij}	Weight between neuron i and j
O_j	The output of neuron j
E_j	The error of neuron j
T_j	Reference output of training samples
sigmod′(x)	First derivative of Sigmod

3 Multivariate Regression Statistical Model

3.1 Model Description

Multiple regression is a regression for variables, and two or more than two independent variables, also known as multiple linear regression, the number is a reflection of the phenomenon or things in a variety of phenomena or things change and corresponding changes in the laws, it is a statistical method for establishing linear or nonlinear mathematical model of the relationship between multiple variable types. As stated in the title, many factors can affect the world's language use number, including the national or regional people use habits, population migration result in the habit of language change, a language on behalf of the countries or regions of the technology and economy impact on the world. For example, the common language is English in the world nowadays, one of the reason is that compared to English other languages is more difficult, another more important reason is the use of English in two countries the United States and Britain's economic strength is very strong, the English is the key for people from other countries or regions to do business with enterprises or individuals in the United States or Britain. Therefore, the influence of economy and science technology will promote many people to learn to use some language. In addition, the factors that influence the number of speakers for a language include national policies and so on. These parameters cannot be measured in mathematical language. So these factors will be abandoned.

The idea of this paper is to find out these characteristics by finding the data in the first ten years of the ten languages in the list, and fitting ten multivariate equations with

the actual data of ten years, and predict the distribution of the number of language users in the next 50 years by these ten equations. And predict the future parameters of the ten languages, and find the relationship with time by MATLAB, then get the parameters for predicting the number of future language users.

For the number of factors mentioned above using a certain language, we use five parameters to measure, we use the natural population growth rate to measure the language speaking population change in countries. For the impact on migration flows, this paper uses a language area's immigrants change rate per year to measure. For some language influence technology in this paper, using the number of web pages to measure. And for some language economic influence, this paper uses the language of the region GDP and per capita GDP as criterions, the reason why we choose these two GDP as parameters in the formula is that the number of Asian population is relatively large, because the amount of GDP in some countries are high, so they has a certain economic influence in the world such as China. However, the amount of GDP in some countries is not high, but their population is small, with a higher per capita GDP, they still have greater economic influence, such as Portugal and Spain. So it is necessary to put both of the two into the formula to get more accurate results.

To sum up, the following formula is obtained.

$$y = k_0 + k_1 {x_1}^n + k_2 {x_2}^n + k_3 {x_3}^n + k_4 {x_4}^n + k_5 {x_5}^n$$

x_1—— Natural population growth rate
x_2—— Increase in the number of immigrants
x_3—— Total GDP
x_4—— Per capita GDP
x_5—— The growth rate of the number of pages
k_0~$k_5$are constant coefficients

For the future prediction of the five parameters, we use Matlab to fit and take time as an independent variable.

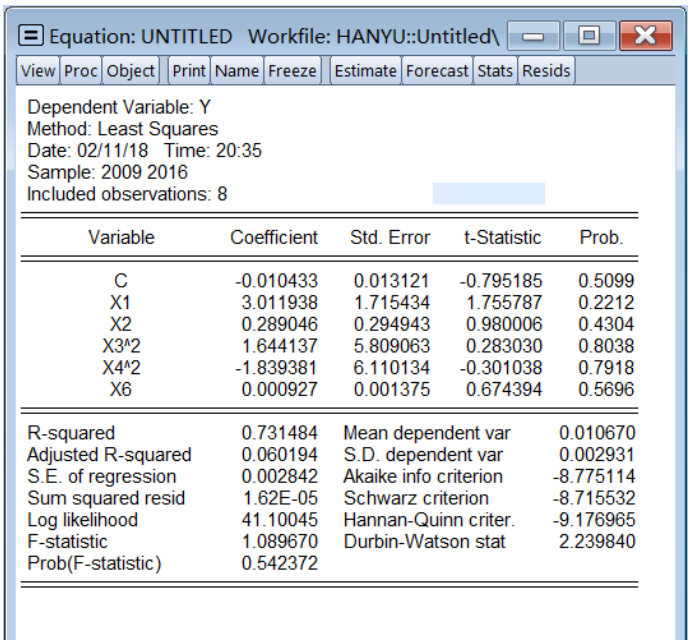
3.2 Data Preprocessing

In order to more clearly display the status of languages users' number changing and avoid of different parameters of different units, the huge difference between the number of problems, this paper use the parametric growth rate instead of parameters. The growth rate of five parameters obtained by predicting can calculate growth rate of a language speakers' number. Finally the number of a language speakers of the year can be obtained by number of last year's growth rate being multiplied by the specific value.

3.3 Regression Statistics and Fitting

First of all, the above formula is fitted with the data from the first ten years. Take Chinese as an example. Use the commonly used software Eviews of economic statistics to carry out one-dimensional multivariate fitting. It is found that the results are not

satisfactory, so we try to improve the order of the parameters. After many experiments, the results are shown below. The results obtained in Akaike info criterion and Schwarz criterion tests are -8.77 and -8.71 respectively, so the results are more reliable.



Equation: UNTITLED Workfile: HANYU::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 02/11/18 Time: 20:35
Sample: 2009 2016
Included observations: 8

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.010433	0.013121	-0.795185	0.5099
X1	3.011938	1.715434	1.755787	0.2212
X2	0.289046	0.294943	0.980006	0.4304
X3^2	1.644137	5.809063	0.283030	0.8038
X4^2	-1.839381	6.110134	-0.301038	0.7918
X6	0.000927	0.001375	0.674394	0.5696

R-squared	0.731484	Mean dependent var	0.010670
Adjusted R-squared	0.060194	S.D. dependent var	0.002931
S.E. of regression	0.002842	Akaike info criterion	-8.775114
Sum squared resid	1.62E-05	Schwarz criterion	-8.715532
Log likelihood	41.10045	Hannan-Quinn criter.	-9.176965
F-statistic	1.089670	Durbin-Watson stat	2.239840
Prob(F-statistic)	0.542372		

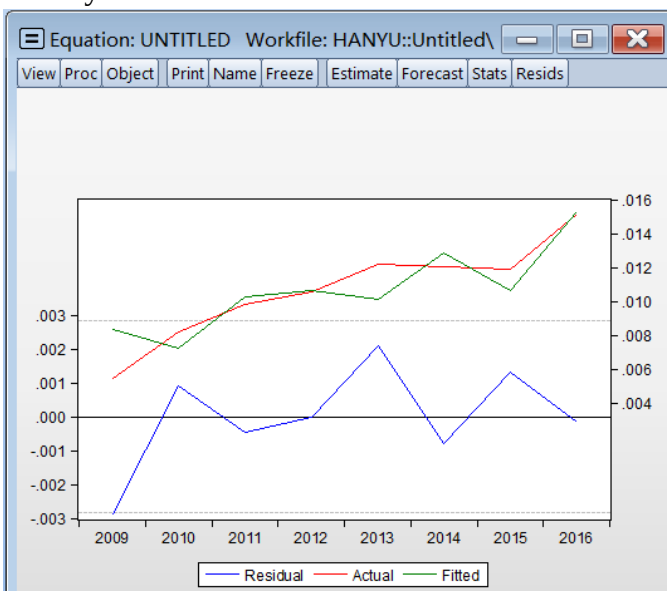
Graph 2

So we get the value of k_0~k_5:

k_0=-0.010433 k_1=3.011938 k_2=0.289046 k_3=1.644137

k_4=-1.839381 k_5=0.000927

The curve of the fitting curve and the original data can be seen by the following graph, which is basically close.



Graph 3

Finally, we get the growth rate function that calculates the number of people who use a language each year:

$$y = -0.010433 + 3.011938x_1 + 0.289046x_2 + 1.644137{x_3}^2 + -1.839381{x_4}^2 + 0.000927x_5$$

The following is the fitting process of each parameter in each language, using the cftool toolbox in MATLAB to fit the natural growth rate of Chinese population. The dependent variable is the natural growth rate, and the independent variable is the time. The results with high accuracy can be obtained by using the second order Fourier fitting.

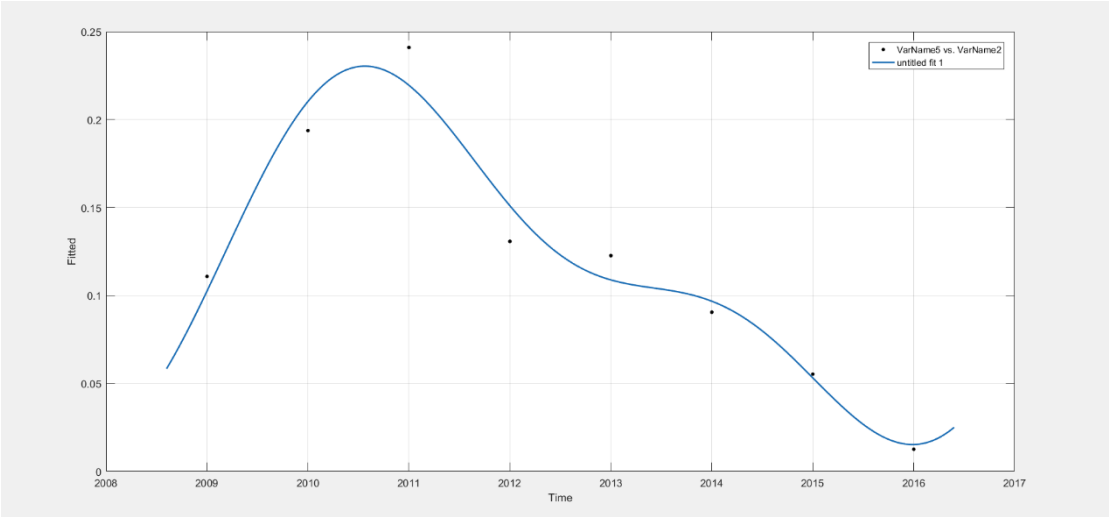
Results

General model Fourier2:
f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) +
a2*cos(2*x*w) + b2*sin(2*x*w)
Coefficients (with 95% confidence bounds):
a0 = 0.1171 (0.05415, 0.18)
a1 = -0.05212 (-56.93, 56.83)
b1 = -0.0697 (-42.58, 42.44)
a2 = 0.03169 (-32.87, 32.94)
b2 = 0.02017 (-51.71, 51.75)
w = 0.7612 (0.3554, 1.167)

Goodness of fit:
SSE: 0.001456
R-square: 0.9605
Adjusted R-square: 0.8617
RMSE: 0.02698

Graph 4

R-square is close to 1 as SSE is approaching 0, which proves that the model is in line with the actual data.

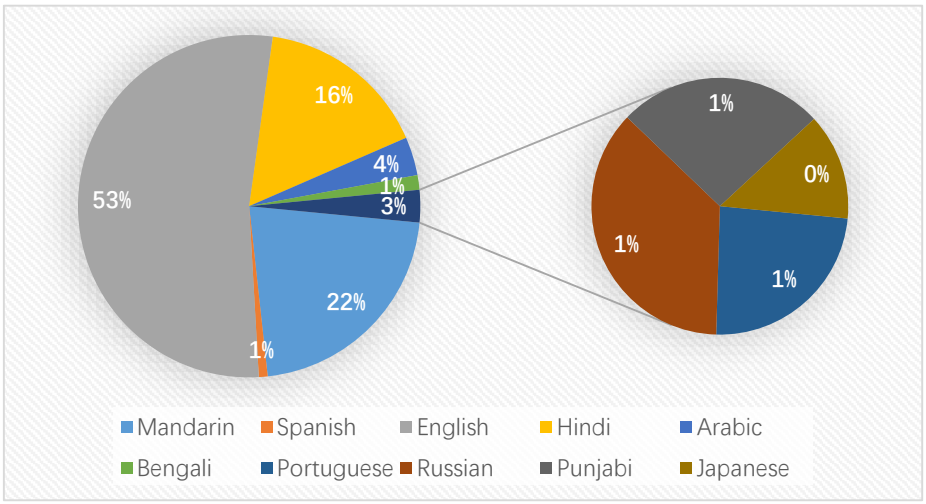


Graph 5

The image below also proves that the fitted function accords with the actual data, the black point is the actual value, and the blue curve is the curve fitted according to the actual value.

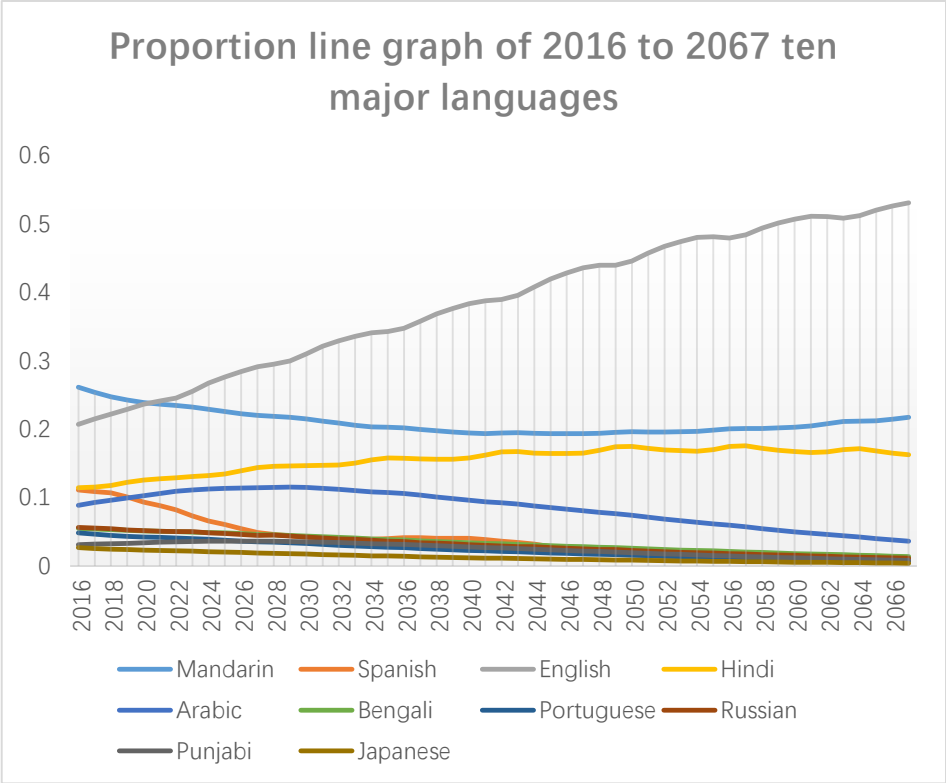
3.4 Analysis and Results

By calculating the formula in Section 3.1, we can get the following results: After 50 years English will be the world's largest language, accounting for more than half, with Chinese and Hindi occupying the second and third place respectively.



Graph 6

The following is a broken line diagram of the proportion of the ten languages drawn from the results of the forecast in more than 50 years (actually 51 years, with more data from 2016 for reference). It can be seen from the graph that the proportion of English is rising rapidly at a very high speed; The proportion of Chinese decreased first and then increased slowly, while Hindi, the third language, rose slowly, while the other languages showed a slow downward trend.



Graph 7

The reasons can be summarized as follows：

1、Because of its great economy, English can be seen as one of the most promising among all languages. Although it does not currently rank first in the number of speakers, it has a strong momentum for growth. In the future will become the world's largest language.

2、1. The main speakers of Chinese are China, and the main speakers of Hindi are India. These two developing countries have less economic and technological influence than the United States and Britain, which use English, so the growth rate of these two major languages is slower than that of English. But countries using languages such as Bangladesh are still growing faster than other countries, such as Arabic, and Portuguese-speaking countries have a larger population base than Spain. So it can be the second and third language in the next 50 years.

3、Most of the countries using Spanish and Portuguese are developing countries, and the rate of development is fast. The corresponding Spanish and Portuguese have a higher growth rate.

4、Arabic, Bengal and other languages have less influence on science and technology that represent national and regional economies. So the growth rate of these languages mainly depends on the natural rate of population growth.

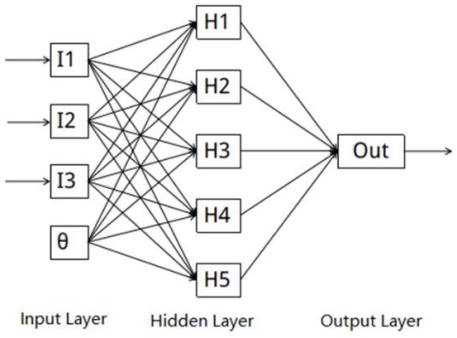
5、The Punjab language is at the bottom of the list because of the low economic and technological influence of the countries and regions that use the language, and the general natural population growth rate. Japan, a country corresponding to Japanese, although highly developed, has an international economy. Technology is highly influential, but it is also at the bottom because its population base is too low for the top countries, and Japan's population sometimes grows negatively.

And for now, according to data from Wikipedia and the Languages of the World, French and German have outnumbered Punjab and Bengali, and according to our model projections, In the coming decades, the number of people in French and German as a second language will grow faster than the natural rate of population growth in the Punjabi and Bengali communities. So there is reasonable to believe that. French and German will have a greater chance of outnumbering Punjab, Bengali and even Hindi.

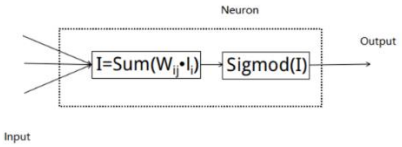
4 BP Neural Network Model

4.1 Model Description

The neural network model introduced in this paper belongs to MLP in structure and is trained by BP algorithm. The neural network consists of three layers: input layer, hidden layer and output layer. The number of neurons in input layer is related to the characteristic number. The number of output layers is the same as the number of categories, and the number of layers and neurons of the hidden layer can be customized, which needs to be reasonably selected from the operation time and the result optimality. The basic working principle is shown in the following figure, and each neuron processes the data once:



Graph 8



Graph 9

The functional relationship between the output and the input of each hidden and output layer neuron is as follows:

$$I_j = \sum_i W_{ij} O_i$$

$$O_j = sigmod(I_i) = \frac{1}{1 + e^{-I_i}}$$

Where Wij denotes the weight of the connection between neuron i and neuron j, Oj represents the output of neuron j, sigmod is a special function used to map any real number to an interval. The sigmod function in the formula is called the excitation of the neuron. In addition to the sigmod function, you can also use tanh and ReLU functions.

In this case, we use a trained neural network to deal with the regression problem, where each sample has n inputs, and accordingly, The neural network has n input-neurons and one output neuron. We add an extra bias neuron to the input layer to provide a controllable input correction. N features are fed into the input neuron in turn. The hidden layer neuron obtains the output of the input layer and calculates its own output value. The output layer neuron calculates the regression value according to the hidden layer output. The feedforward Feed-Forward process is basically completed. In this process, the input and output of the neural network are the same as the multidimensional functions. We began to train the neural network.

The training process of BP neural network as a supervised learning algorithm is the process of adjusting the connection weight W_{ij} according to the error and comparing the predicted value and the reference value. In this back propagation process, the data flow is exactly the same as the previous one. The feed process is opposite. First of all, We initialize the connection weight at random.We get the output of each neuron by a feedforward process of a training sample, and calculate the error of the output layer.

$$E_j = signmod'(O_j) * (T_j - O_j) = O_j(1 - O_j)(T_j - O_j)$$

Where:

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{(1+e^{-x})-1}{(1+e^{-x})^2}$$
$$= \frac{1+e^{-x}}{(1+e^{-x})^2} - (\frac{1}{1+e^{-x}})^2 = \sigma(\mathrm{x}) - \sigma(\mathrm{x})^2$$
$$\sigma' = \sigma(1-\sigma)$$

Based on the calculation of the hidden layer error, there are

$$\mathrm{E}_j = sigmod'(O_j) * \sum_k E_k W_{jk} = O_j(1-O_j)\sum_k E_k W_{jk}$$

holds。Hidden layer output does not have a reference value. Use the weighted sum of the next level of error take place of ($T_i - O_j$), $W_{ij}$and $\theta_j$can be updated after error calculation:

$$W_{ij} = W_{ij} + \lambda E_j O_i$$

Where λ is a parameter called learning rate, which is generally valued in interval of (0,0.1). In fact, in order to speed up the efficiency of learning, we also introduce a mechanism called the correction matrix, which records the $\mathrm{E}_j O_i$ values during the last backpropagation. Therefore W_j updates the formula to:

$$W_{ij} = W_{ij} + \lambda E_j O_i + \mu C_{ij}$$

μ is a parameter called correction rate. The correction matrix is updated thereafter:

$$C_{ij} = E_j O_i$$

In this way, each training sample updates the parameters of the entire network once. We need to set additional conditions for training termination. You can set the simplest training termination condition to set the maximum number of iterations, such as stopping the training after 1000 iterations on the dataset. However, Simply setting the maximum number of iterations can not guarantee the accuracy of the training results. A better method is to use the loss function as the basis for the termination of the training. In this paper, the loss function can select the variance of each node in the output layer.

$$\mathrm{L} = \sum_j (T_j - O_j)^2$$

The new problem is that neural networks may perform pointless iterations. In order to avoid this situation. We withdraw part of the training data set for verification and terminate the training ahead of time when the prediction error is higher than the threshold value.

4.2 Analysis and Results

First, we define a stochastic function and matrix transformation function, then we define a BPNeuralNetwork class, using three list maintenance input layer, hidden layer and output layer neurons, output elements in the list represents the current value of each neuron. Two two-dimensional lists are used to maintain the connection weights between the input layer and the hidden layer, the hidden layer and the output layer in the form of adjacency matrix, and save the correction matrix in the same form. Next,

the setup method is defined to initialize the neural network, and then the predict method is defined for a feedforward.

Here is the feedback algorithm of the hidden layer. The input and output layer is analogous to that. After a previous neuron iterated the data and sent data to the neurons in the latter process, the latter neuron received the message and processed it, then processed the data back to the previous neuron.

Then we define the back_propagate method to carry out a back-propagation and update weight, and return the final prediction error. The specific process is:

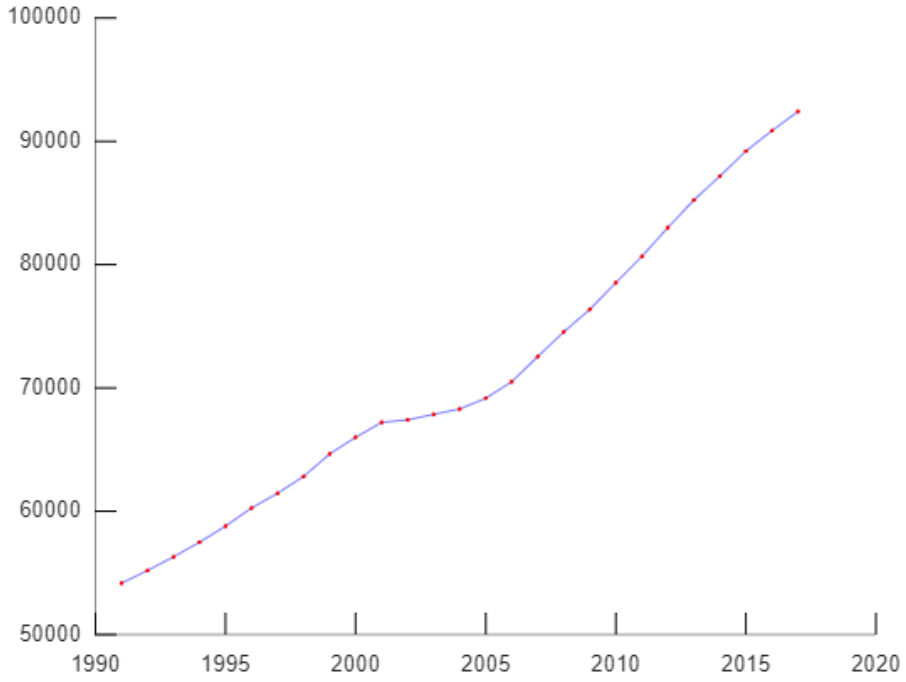
1. Get the error of the output layer and analyze it.

2. The hidden layer is taken to take the wrong processing, then the above data is corrected, and the weight of the output layer and the input layer are updated in time.

3. Get the global error and remove the error.

The implementation of the output layer adjustment and back propagation update weights. The principle of the hidden layer is the same. After many training, we set up the conditional function of the iteration to reach the desired condition.

This method can modify the maximum iteration number m, the learning rate v and the correction rate u three parameters. After that, we can make a reasonable prediction for the next fifty years based on the original data. From 1990 to 2017, "the number of Chinese area to English area" is calculated as follows, and the image is obtained:



Graph 10

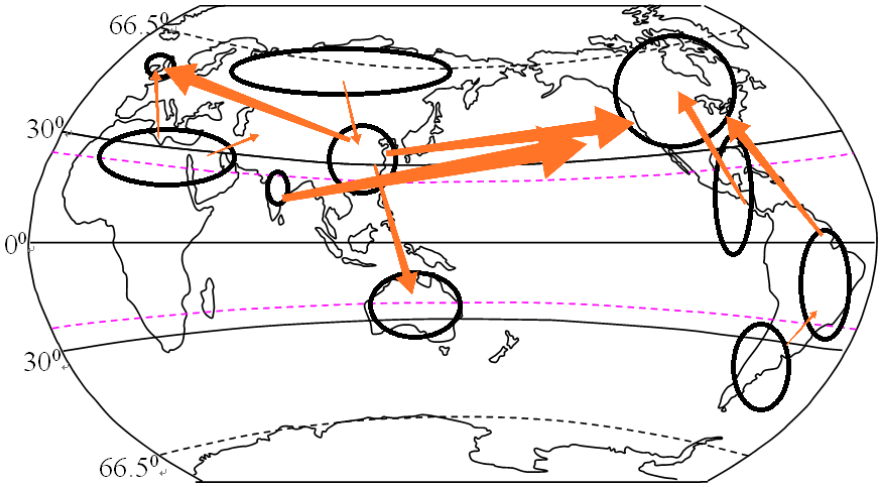
Similarly, using the above method, the other 99 sets of data are processed and some of the original data lists are listed as follows:

Chart 2

Prediction of the number of immigrants in the Chinese region to other areas in the next fifty years (part)									
Year\Language area	English	Spanish	Hindi	Arabic	Bengali	Portuguese	Russian	Punjab	Japanese

2018	410000	82000	8500	12000	213	23000	12000	24	11000
2019	413000	85000	8600	12300	203	23200	11800	25	11500
2020	432000	88000	8400	12500	205	24000	12300	21	11800
2021	452000	84000	8800	12700	214	25000	12500	20	12000
2022	436000	89000	8900	12800	218	24800	12600	19	12300
2023	428000	92000	9200	13100	229	25300	12800	20	12200
2024	422000	94000	9400	13400	234	26400	13400	25	12400
2025	447000	95000	9000	14200	239	27700	13500	34	12600

Finally, we tagged it on the map to get a general change in the location of the ten major languages for the next fifty years:

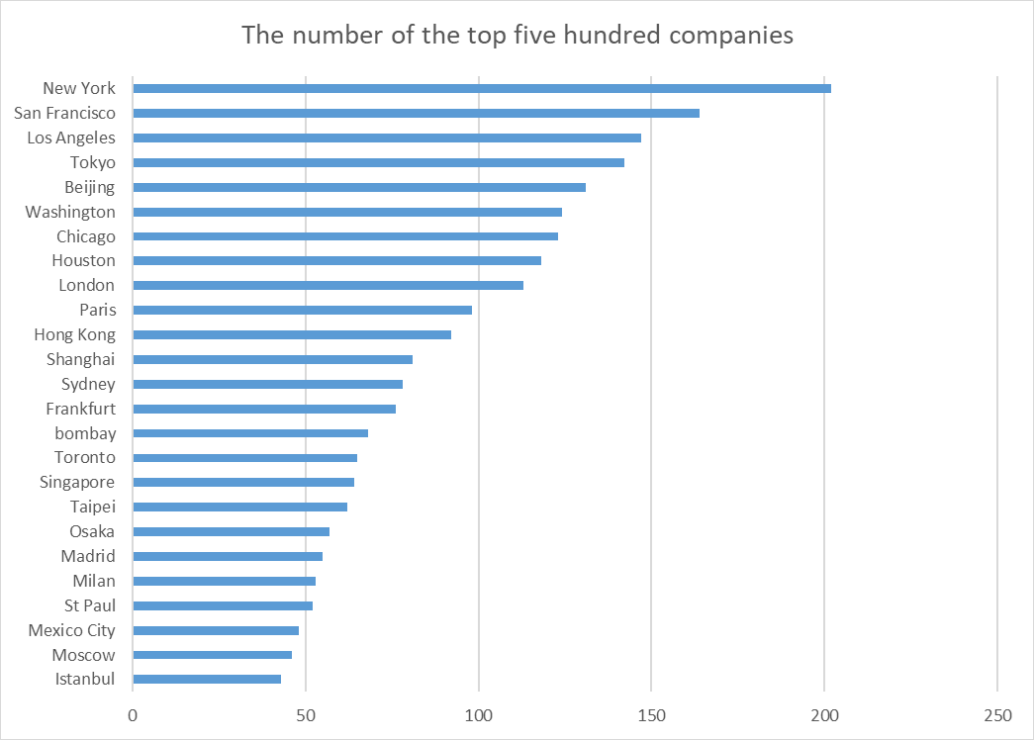


Graph 11

The size of the arrow represents the number of language migration and diffusion indices for a period of time. From this, the Chinese area, area and area to the English Hindi Spanish region migration amount is larger, more diffuse to English area trend; Arabic and Portuguese to English Zone diffusion area trend is small; in addition, the Russian language and Arabic region has spread to the area of the district is still almost English trend; English migration in the region; not indicated in the drawings (Bengali language area of Punjab District, the language area and the Japanese region) outward diffusion trend is not obvious, mainly concentrated in the area within the original still.

5 Level Analysis Office Model

First, there are many practical factors that need to be considered for the location of the company. For example, the factors that need to be considered in this topic may be: GDP of each city, GDP of the country, the market size of the company in these cities, the population of the city, the education level of the city, etc. These factors are not only dispersive but not easy to measure, so we use the principal component analysis method. According to the previous data and the reference papers, we have finally screened the number of the components that the world's top five hundred companies have established in each city. We found the relevant data, statistics, lists, and sorting, and the final results are as follows:



Graph 12

According to the conclusion of our Part1, excluding the city of language change is bigger; while considering other factors such as politics, economy, science and technology, culture, geographical location and corporate earnings, the final choice we make is: Beijing (or Shanghai), Singapore, Paris, London, Tokyo, New York (or San Francisco). The selected language according to the five hundred companies with language analysis, according to the geographical distribution of the Part1 language and the number of people using the conclusion, here we conclude that the use of language: the office of local language, language and regional headquarters offices where influenced by language. For example, for an office in South Korea, the company headquarters in Spain, it set up an office in Seoul, so the first language is Korean, second language is Spanish, In addition, the economy, politics and culture of Korea are greatly influenced by Japan and China, so the third and fourth languages need to consider Chinese or Japanese.

According to the Part1, in the short term, language changes in the world will not be too large, especially in the cities we choose. Therefore, this language proposal will not change in a short time. But in the long run, English and Chinese will occupy more and more widely used space. So in the long run, we will think more about English and Chinese.

For Part2 B, first, determine the criteria layer, the number of the scale of the company, the industry, the main market place; the size of the company to measure the size range from 0 to 100, depending on the amount of the company's revenue, market capitalization (if not listed with reference to the valuation of the company) and the number of Companies in the industry company; the five hundred companies are divided according to the industry, the number of branch statistics for each industry to determine a measure index, the index ranges from 0 to 10, such as retail value of more than investment banking financial industry; main market where the number of set values range from 0 to 1, the specific value the company accounted for the proportion of the company in all city market in the city market. We need to determine the limiting factor: the maximum acceptable cost and guarantee the minimum number of market segments and headquarters communication smoothly; the maximum acceptable cost by estimating the enumeration and estimate the 6,5,4,3,2,1 offices required cost is about how much, in order to determine an upper limit, if the upper limit is greater than 6 is set: the office number the upper limit of 6, at the same time, the value of not less than 1; the minimum number of successful communication headquarters division shall not be less than 1. The final selection layer is 1,2,3,4,5,6. Finally, the world top five hundred companies are selected as samples, and the corresponding standard level data and their actual selection data of the company are brought into the model to determine the weight of the standard layer in this model.

For example, for Berkshire Hathaway, the company scale is 93, the main business for the company's industry investment index was 5, the main market is located in developed countries and emerging countries economic and financial center of the financial district, for the top 6 markets accounted for the total market is about 63%, this value is 0.63. Finally, through most of the validation, our conclusion is:

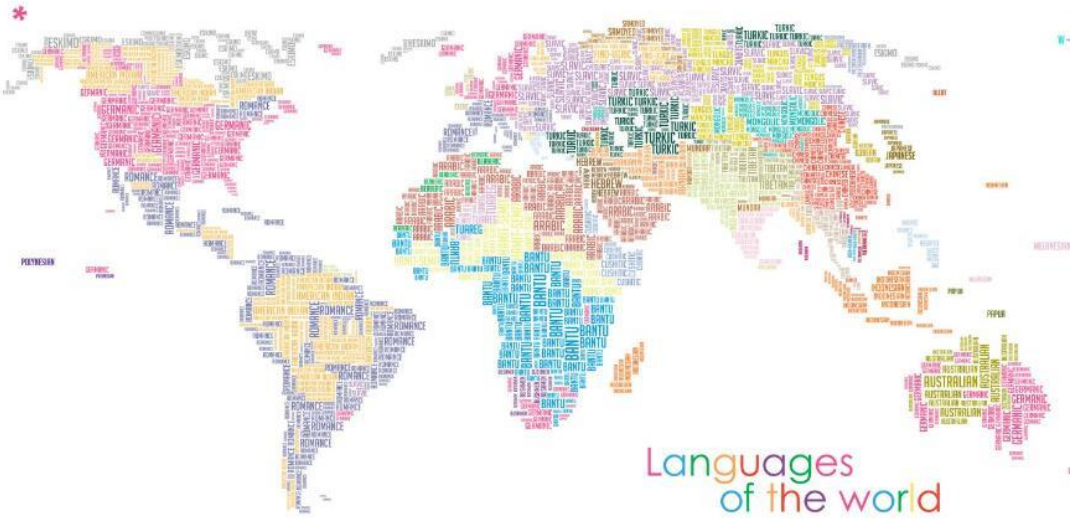
$$M = 0.478 \cdot [Company\ \ size] + 0.782 \cdot [The\ \ industry's\ \ index]^2 + 5.72 \cdot [Top\ \ six\ \ share\ \ index]^3$$

The relationship between M and the number of m in the office is:

Chart 3

M	m
<5	1
5-10	2
10-20	3
20-35	4
35-55	5
55-80	6
>80	6

After determining the size of the company, the industry and the market dispersion, we can make reasonable suggestions for the company through calculation.



Graph 13

6 Model Analysis

6.1 Analysis of Goodness of Fit and Sensitivity

First of all, for the prediction model, when we analyze the model excellence, for the multivariate regression statistical model, Goodness of fit refers to the degree of fitting of regression line to observed value. The statistical measure of fit goodness is that the value of the maximum value of R^2 For example, for the Spanish-speaking region, in our calculation, the overall prediction formula is as follows:

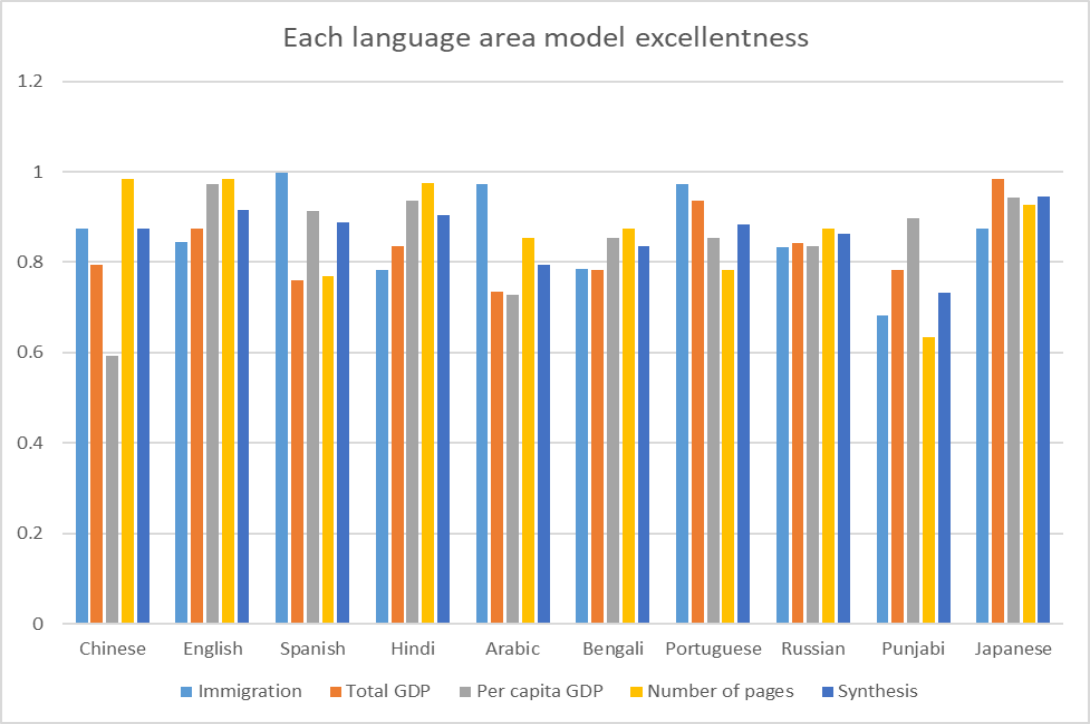
$$y = 0.122708 - 5.410921 \cdot x_1 - 6.425129 \cdot x_2 - 12.20033 \cdot x_3^2 + 11.41579 \cdot x_4^2 + 0.266982 \cdot x_5$$

We can see that here the value of R is 0.8871, which we think is more appropriate as a composite formula. In addition, in some parameters, the value of Spanish-speaking region R is shown in the following table:

Chart 4:Spanish region regression prediction superiority table

Parameters	R^2	Superiority evaluation
Number of immigrant	0.9986	excellent
Total GDP in language area	0.7604	common
Per capita GDP	0.9120	fine
Number of web pages	0.7690	common
Sum	0.8871	preferable

Analyze the R data for other language areas and chart the following：



Graph 14

It can be seen from the above chart that the R value of all regression prediction fitting is above 0.65, and the overall average fitting goodness is 0.8674. The overall result is in line with the expectation, and the language change and distribution of each language area in the future is well predicted. But we also see that, In some language areas, such as Punjab, the data goodness of fit is not ideal. The analysis is mainly due to the small amount of data and the greater contingency. There are good reasons to make predictions based on this data.

Secondly, we analyze the sensitivity of the above model. We simulate the effect on the overall result when the value of a parameter changes. Take the immigration parameter of Spanish speaking area as an example to analyze. Use the data in 2017 as the base, adjust to 105%, 95% and 80% of the number of immigrants. And observe the change of the final result, as shown in the table below：

Chart 5

Number of immigrant(taking2017 as 1)	Number of speakers in language area	Ratio to 2017	appraisal
80%	13147180	0.8321	good
90%	14671880	0.9286	good
95%	15411320	0.9754	well
105%	16966040	1.0738	good
110%	17805020	1.1269	good
120%	19266520	1.2194	well

Similarly, we analyzed the sensitivity of other language regions. The results showed that the sensitivity of the model was very high for the number of immigrants, the sensitivity for total GDP and GDP per capita, and the sensitivity for the number of web pages. In addition, in different language areas, The sensitivity is also different. In general, the sensitivity is higher for a large number of language areas, such as the English area, the Chinese area, and so on, while in other areas, such as Punjab and Bengali, The reason why the sensitivity of individual regions is not good is that the data is not large enough, because from the language area where the data is large and the sample is large, the sensitivity is very high, which completely meets our expectation.

6.2 Advantages and Disadvantages

Strengths:

(1). The multivariate regression analysis method can accurately measure the degree of correlation and the degree of regression fitting among various factors, and improve the effect of prediction equation. Compared with the method of single regression analysis, multivariate regression analysis is more suitable for practical economic problems.

(2). The advantage of BP neural network is that it has strong nonlinear mapping ability and flexible network structure. Even if the system is subject to local damage, it can still work normally.

(3). The weight setting of each layer in AHP will affect the result directly or indirectly, and the influence degree of each factor on the result in each level is quantitative, very clear and clear.

Weaknesses:

(1). Regression statistical model is still difficult to depict the actual influence of many factors.

(2). For AHP models, the increase in indicators means that we have to construct a deeper, more numerous, larger judgment matrix. Then we need to compare a lot of indicators in pairs. What is more, because of the complexity of objective things or the one-sidedness of understanding things, the eigenvector (weights) obtained by the constructed judgment matrix is not necessarily reasonable.

Reference

[1] Xu Jie, Lin Jun. The model of the city competitiveness of the commercial complex [C]. Shanghai: Xu Jie, 2005.

[2] Geng Jin Hua, Gao Qi Sheng, Zhang Siying. Community satisfaction evaluation system based on analytic hierarchy process and factor analysis [C]. Qingdao, 2007.

[3] Han Linshan, Li Xiangyang, Yan [C]. exams. Several mathematical methods of sensitivity analysis: Zhengzhou, 2008.

[4] MBA think tank. The analytic hierarchy process [EB/OL]. http://wiki.mbalib.com/wiki/%E5%B1%82%E6%AC%A1%E5%88%86%E6%9E%90%E6%B3%95.

[5] Harald, Hammarstrom. Ethnologue 16/17/18th editors: A comprehensive review[EB/OL]. http://www.ethnologue.com/15/ethno_docs/distribution/family/.

[6] Jiang Hua, Zhao Jie. Evaluation model of learning behavior based on BP neural network and implementation of [D]. Shanghai, 2008.

[7] Gao Hongshen, Tao Youde. BP neural network model improvement [D]. Beijing, 1996.

Respectable Mr. CEO:

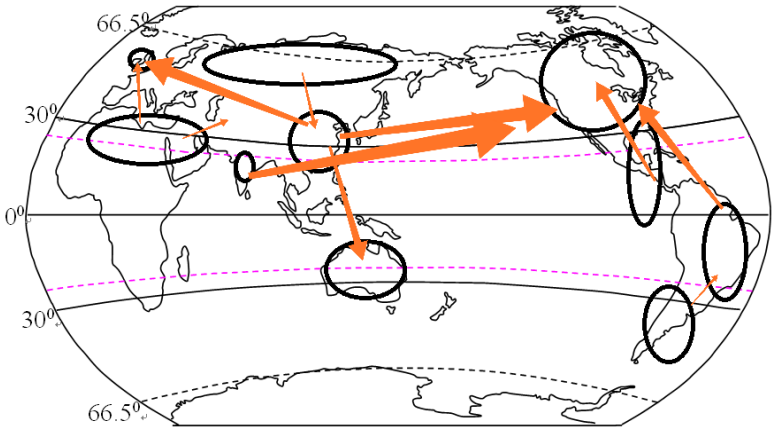
Hello! As a language analysis team you hired, we have obtained relevant data and suggestions after fully researching and calculating prediction. Now, as a memorandum, I submit to you as follows:

First of all, in the world the numbers of the ten languages we focus on are increasing year by year, German and French also have large bases which are increasing yearly. In terms of the data, the bases of English and Mandarin users are maxima, their growth rate are faster as well. So in the foreseeable decades, English and Mandarin will increasingly become the world universal language. Thanks to the natural population growth rate, the number of people who use Hindi, Bengali and Punjab will grow quickly and large base. Arabic, Spanish, Portuguese, Russian and Japanese areas have steady progresses but little changes. So, our suggestion is that for choosing languages, we prefer to use more widely distributed languages in the future, such as English and Mandarin.

Secondly, for the proportion of language use, more and more people will choose English, Mandarin, Arabic, Spanish, Portuguese, Japanese, Russian as a second language. The population of choosing English as a second language is the largest and growing relatively fast, the population of choosing Mandarin as the second language has the fastest growth rate, which means that the degree of internationalization of English and Mandarin will be higher and growing faster compared to other languages. In addition, the growth rate of people who learn the second language is much higher than the natural growth rate of the population, which means that the proportion of people who have two languages in the world will be higher and higher. In the ten languages the population of choosing Punjab language, Bengali and Hindi as a second language is relatively small, in the next few years the growth rate is also not fast, so the rate between language native population and the total population will be increased. In contrast, the population of choosing Mandarin, English French, German, Japanese, Spanish and Portuguese as a second language accounts for the proportion of the total population will rise rapidly. Therefore, when India and Bangladesh natural population growth rate decrease, it’s more likely to be altered by French and German, then they’ll drop out of the top ten languages. It’s also clear that English, Spanish, French and German native speakers due to the natural growth rate is low or even negative, the ratio between native speakers and the world's population will further reduce. Therefore, for making the choice of office addresses we suggest that try to choose to use those who has a larger population speaking widely used languages, such as the English area of London and New York, the Mandarin area of Beijing and Shanghai and so on, this can minimize the cost of communication caused by linguistic factors.

Thirdly, in terms of geographical distribution, it is the most obvious trend that Mandarin and Hindi area erodes English area. Furthermore, the migration from Mandarin area, Hindi area and Spanish area to English area is larger, which means likely a trend of diffusing to English area. Arabic and the Portuguese region also have smaller trend of spreading to English region. In addition, the Russian and Arabic region has a trend spreading to the Mandarin area. People in English area almost still migrate in English area; the languages areas which are not indicated in the following pictures

(Bengali language area and Punjab dialect area and Japanese area) has little outward diffusion trends, people there mainly concentrated in the original area.



Additionally, the advice we give about choosing your company offices' addresses is that 6 offices are in Beijing (or Shanghai), Singapore, Paris, London, Tokyo, New York (or San Francisco). With the corresponding, the six offices of the use of languages all contains English other languages are Mandarin, Mandarin and Japanese, French, Mandarin and French or German, Mandarin and Japanese, Mandarin or Hindi. This proposal will not change in a short period of time, but in the long term, taking the language change into account, the requirement of English will not change, but offices may should increase the proportion of Mandarin use. Furthermore, if your company set up offices in India, Bangladesh or other countries, using the local language as more as possible is a good choice. This is because in these areas, the proportion of using mother tongue is great, and look furtherly, the natural growth rate of population in these countries will continue to be high for a long time in the future. Although these languages do not very widely, the use of morther tongue and it's growth rate will be continually very high.

Finally, for the problem of office number, though the current global communication increasingly developed, we suggest you to set up offices as more as possible in the controllable scope of cost and profit. It's benefit for both the influence of headquarters and the worldly communication to provide better service. We established the relevant evaluation model, in order to draw a more scientific conclusion accurately, our team requires a detailed description of your company's industry, market size, degree of dispersion data, then we will finally draw the conclusion according to the comparison with the world's top five hundred enterprises.

I hope that in the next time, our team will be able to keep in touch with you better and know more about your company, to provide you with better services and suggestions.

Your employment team, XXX

Appendix

```
import math
import random

random.seed(0)


def rand(a, b):
    return (b - a) * random.random() + a


def make_matrix(m, n, fill=0.0):
    mat = []
    for i in range(m):
        mat.append([fill] * n)
    return mat


def sigmoid(x):
    return 1.0 / (1.0 + math.exp(-x))


def sigmoid_derivative(x):
    return x * (1 - x)


class BPNeuralNetwork:
    def __init__(self):
        self.input_n = 0
        self.hidden_n = 0
        self.output_n = 0
        self.input_cells = []
        self.hidden_cells = []
        self.output_cells = []
        self.input_weights = []
        self.output_weights = []
        self.input_correction = []
        self.output_correction = []

    def setup(self, ni, nh, no):
        self.input_n = ni + 1
        self.hidden_n = nh
```

```
        self.output_n = no
        # init cells
        self.input_cells = [1.0] * self.input_n
        self.hidden_cells = [1.0] * self.hidden_n
        self.output_cells = [1.0] * self.output_n
        # init weights
        self.input_weights = make_matrix(self.input_n, self.hidden_n)
        self.output_weights = make_matrix(self.hidden_n, self.output_n)
        # random activate
        for i in range(self.input_n):
            for h in range(self.hidden_n):
                self.input_weights[i][h] = rand(-0.2, 0.2)
        for h in range(self.hidden_n):
            for o in range(self.output_n):
                self.output_weights[h][o] = rand(-2.0, 2.0)
        # init correction matrix
        self.input_correction = make_matrix(self.input_n, self.hidden_n)
        self.output_correction = make_matrix(self.hidden_n, self.output_n)

    def predict(self, inputs):
        # activate input layer
        for i in range(self.input_n - 1):
            self.input_cells[i] = inputs[i]
        # activate hidden layer
        for j in range(self.hidden_n):
            total = 0.0
            for i in range(self.input_n):
                total += self.input_cells[i] * self.input_weights[i][j]
            self.hidden_cells[j] = sigmoid(total)
        # activate output layer
        for k in range(self.output_n):
            total = 0.0
            for j in range(self.hidden_n):
                total += self.hidden_cells[j] * self.output_weights[j][k]
            self.output_cells[k] = sigmoid(total)
        return self.output_cells[:]

    def back_propagate(self, case, label, learn, correct):
        # feed forward
        self.predict(case)
        # get output layer error
        output_deltas = [0.0] * self.output_n
        for o in range(self.output_n):
            error = label[o] - self.output_cells[o]
```

```
            output_deltas[o] = sigmoid_derivative(self.output_cells[o]) * error
        # get hidden layer error
        hidden_deltas = [0.0] * self.hidden_n
        for h in range(self.hidden_n):
            error = 0.0
            for o in range(self.output_n):
                error += output_deltas[o] * self.output_weights[h][o]
            hidden_deltas[h] = sigmoid_derivative(self.hidden_cells[h]) * error
        # update output weights
        for h in range(self.hidden_n):
            for o in range(self.output_n):
                change = output_deltas[o] * self.hidden_cells[h]
                self.output_weights[h][o] += learn * change + correct *
self.output_correction[h][o]
                self.output_correction[h][o] = change
        # update input weights
        for i in range(self.input_n):
            for h in range(self.hidden_n):
                change = hidden_deltas[h] * self.input_cells[i]
                self.input_weights[i][h] += learn * change + correct *
self.input_correction[i][h]
                self.input_correction[i][h] = change
        # get global error
        error = 0.0
        for o in range(len(label)):
            error += 0.5 * (label[o] - self.output_cells[o]) ** 2
        return error

    def train(self, cases, labels, limit=10000, learn=0.05, correct=0.1):
        for j in range(limit):
            error = 0.0
            for i in range(len(cases)):
                label = labels[i]
                case = cases[i]
                error += self.back_propagate(case, label, learn, correct)

    def test(self):
        cases = [
                                        54167,
                                        55196,
                                        56300,
                                        57482,
                                        58796,
                                        60266,
```

```
                                        61465,
                                        62828,
                                        64653,
                                        65994,
                                        67207,
                                        66207,
                                        65859,
                                        67295,
                                        69172,
                                        70499,
                                        72538,
                                        74542,
                                        76368,
                                        78534,
                                        80671,
                                        82992,
                                        85229,
                                        87177,
                                        89211,
                                        90859,
                                        92420,
            [0, 0],
            [0, 1],
            [1, 0],
            [1, 1],
        ]
        labels = [[0], [1], [1], [0]]
        self.setup(2, 5, 1)
        self.train(cases, labels, 10000, 0.05, 0.1)
        for case in cases:
            print(self.predict(case))


if __name__ == '__main__':
    nn = BPNeuralNetwork()
    nn.test()
```