For office use only
T1 _____
T2 _____
T3 _____
T4 _____

Team Control Number
74826

Problem Chosen
B

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

**2018**
**MCM/ICM**
**Summary Sheet**

# How many languages?

**Summary**

We formulate a multiple regression analysis model to simulate the distribution of various language speakers over time and identify population, economic factor, geographic environment and immigrant conditions as influencing factors. And then we integrate the relationship among geographic language distribution, population trends and immigration pattern based on Radiation Model to estimate the geographic language distribution over the next 50 years. We also construct a RBF Neural network to simulate the proper cities for Location of international service offices and this algorithm is based on AHP (Analytic Hierarchy Process) to decide the related weighted evaluation elements. And this process is optimized through genetic annealing algorithm. And the total model framework contains three main part.

In part I, in our basic model, we presume some factors remain constant such as geographic environment and social policy and others represent linear change relationship over time. We estimate the trends of these influencing factors over time and get the distribution of various language speakers based multiple regression analysis.

In part II, we introduce the immigrants factor to our basic model framework, and choose Radiation Model to describe the immigrant pattern. We choose the ten main countries based on the most popular languages. On the basis of related net migration data and immigration countries relation, we simulate the estimated geographic language distribution and infer the new top-ten languages in the next 50 years.

In part III, We select some proper cities as testing factors based on economic conditions, popularity, and feasibility levels to build international offices. This process in mainly finished under AHP. After this first selection process, we continue to use GASA to decide the six most appropriate cities. In conclusion, we confirm that six cities in short term are

$$\{Paris, Hong\ Kong, Los\ Angeles, Madrid, mumbai, Cairo.\}$$

And then after clustering these data, we can confirm that six cities in long term are

$$\{Berlin, Beijing, San\ Francisco, Rio, mumbai, Cairo.\}$$

In conclusion, we find that homoplasy in the language distribution changes is prominent and in the future the variety of world language will continue to decrease. The Matthew Effect have great performance and we think the importance of languages will weaken. Considering the increase of marginal cost, the international can be reduced to less 6.

Key words: language distribution, multiple regression analysis, RBF neural network, GASA, homoplasy in language distribution

**MEMO**

**Date:** February 12, 2018

**To: The Chief Operating Officer**

**From: Team#74826**

**Subject: Solve the problem of finding a proper city to build the international office**

I am writing this MEMO to inform you that I could provide a proper solution to help you build the international offices. And the details are listed as follows.

Firstly, our team uses the multiple regression model to build a proper model to predict the distribution of different languages in different regions. And the error checking is in the reasonable range. On this basis, our team uses the part I as a factor of influencing building company.

Secondly, as we all know, when we plan to build a company, we must consider lists of factors, such as geography, economic, policy and so on .

After checking a series of literature, finally our team divide the influencing factors into four main parts which includes natural environmental factors, economic and environmental factors, economic rationality factor, the change of the number of people who using a given language. Given the uniqueness of this issue, our team put forward a thought which uses the BPF Neural Networks and Analytic Hierarchy Process to consider all factors and in the end, we make our conclusion that will be chosen to build the office.

In our Analytic Hierarchy Process, these six places have the relatively good evaluation results. Regardless of the superiority of geographical location or the economic factor, these six regions have their own comparative advantages. And then, we seek more information to make an optimization to reduce the number of the company. So, our team use the Annealing algorithm to have an in-depth optimization analysis . After that, we reduce 6 offices to 5 offices, because of the population distribution of the world.

Thirdly, our team will list some comments about building the company .

①When building the company , we should consider many factors not only what our team list out but also others like environment and so on .

②Try to construct company dispersal , and try to keep most population of the world will have the opportunities to have a benefit from your company .

③ Try to update your information in a regular time .

If you have any questions, please do not hesitate to contact me for more information.

# Contents

# 1. Introduction

## 1.1 Problem Statement

Languages distribution are influenced by different geographical environment, cultural environment, economic environment and other factors. Although there are currently about 6,900 languages spoken on Earth, among these languages, only the following ten languages have the most native speakers: Mandarin, Spanish, English, Hindi, Arabic, Bengali, Portuguese, Russian, Punjabi, and Japanese. If we consider the population of second languages or particular languages (native speakers plus second or third, etc. language speakers), some data maybe changed but those factors still have influence. Considering the influences and factors described above, a large multinational service company plans to expand its business. It is investigating opening additional international offices and desires to have the employees of each office speak both in English and one or more additional languages. Our team's task is to simulate the trends of global languages and investigate locations for new offices. Based on the data sourcing from *Ethnologue, World Bank, etc.* , we consider how the effects would change as the 50 years passing, and the prediction about numbers of languages speakers as well as their distribution showed in world map. Based on our model, we listed several practical suggestions to the service company. The suggestions and analyses are also included in memo.

## 1.2 Analysis and Approach Overview

In section 2, firstly we identify three significant factors, geographical environment, economic development indicators, and migration indicators. Based on these factors, we construct a Multivariable Linear Regression Model to simulate the trends of world languages. Secondly, we also design two sub-models to support the regression model: one is the estimation of immigrant population, the other is to predict the economic development indicators. Finally, we visualize the data about immigrant, languages users and population to describe the geographic distributions in different countries.

In section 3, we construct a RBF Neural network to research location issues of international office. Six countries with least cost and most effects are chosen. Then considering how to minimize limited resources, we use Anneal Arithmetic to optimize the location issues. This process is based on the minimized cost and draw a conclusion that 5 offices is enough to cover about half the world's population.

In section 4, we test the sensitivity of our models in section 2 and 3, and analyze the strengths and weakness of our models in section 5.

## 1.3 Related Work

**Languages distribution** in the world has been researched for a long time by many people. We can get some basic models about population growth, human migration patterns. [3]

There is an equation about immigrants pattern.

$$\left\langle T_{ij} \right\rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \quad \textbf{(1)}$$

**World Language Evolution pattern** is also a constant research topic, we refer to the article '*Language Order in Contemporary World*'[1]. And we learn that language attrition plays an important role in the evolution process, otherwise there are striking similarity between language attrition and language extension. We use it for reference and make some optimization.

# 2. Assumptions and Notations

## 2.1 Assumptions

We make the following basic assumptions in order to simplify the problem. Each of our assumptions is justified and is consistent with the basic fact.

1 The first language $L_1$ (native language) of a person depends only on his nationality at present. The second Language $L_2$ represents other languages that a person masters.

2 We just consider the influence of policy (social safety levels, immigrant policy) , education (literacy rate, the rate of college enrollment), economic development indicators (GDP, imports and exports, NIC) on the population of a specific language:

- **Population**: It is the leading factor of the total number of speakers of a language. For example, we can see the number of Chinese speakers have direct linear relation with the population of China. Through these years, we can see Chinese has become the most popular language. Obviously, the primary cause of the increase is the increase of the world population over these years.
- **Geographical Environment**: Density of population remain stable for a long time due to almost stable geographical and climate condition. We consider the number of different language speakers has direct relation with density of population and divide the world into nine kinds of gradient in accordance with climatic conditions, habitability etc.
- **Education**: School is the most important place for a person to learn language, especially the second language. Higher education is another important factor to influence the number of learners of a language. Generally, the more develop the higher education of a country is, the more people will be attracted to the country. So in order to get a better education, people are willing to learn the languages used in the countries which has good higher education.

- **Economic Development Indicator**: We use the data *national income per capita*(NIC) as the core indicator. Considering present world is connected more than before, everyone has more probability and motivation to chase higher earning's jobs. A higher NIC means wealthier national and greater employment opportunity. So the more wealthier the country is, the more people will be attracted to the country.
- **Migration**: After moving to a new country, people must learn the local language in order to satisfy the needs of work and life. Also their children will accept local language education.

3 It takes no time to learn a new language. This assumption is aimed at the migrants. That is, a migrant can master the language of the country they move to as soon as he arrives there.

4 We only consider the ten most popular languages that are used in the ten countries as official languages, described as follows:

| Rank(present) | Country | Language |
|---|---|---|
| 1 | China | Chinese, Mandarin |
| 2 | Spain | Spanish |
| 3 | United Kingdom | English |
| 4 | India | Hindi |
| 5 | Portugal | Portuguese |
| 6 | Bangladesh | Bengali |
| 7 | Russian Federation | Russian |
| 8 | Japan | Japanese |
| 9 | Indonesia | Javanese |
| 10 | Germany | German, Standard |

Table 1. Chosen country and languages

5 We only use AMPTRAC Retro-fit Kits AMPTRAC tool to simulate the relation between these factors and dependent variable (the number of special language users).

6 We assume that the trend of economics and education approximatively accord with our regression equation in the next 50 years. And the geographic environment remain stable over the next 50 years.

7 We assume that the location of international offices are affected by four influence parts, which include *Natural factors*, *Economic environmental factors*, *Economic rationality*, *and Use of population changes in some languages.* We select ten parameters to measure the levels of the four influence parts.

8 We assume each international office has the same capacity to deal with issues, and there is no different in merit and demerit.

9 Each international office is not affected by other influence factors except the ten parameters during the construction process.

## 2.2 Notations

**The notation table contains all the notations we use in this paper.**

| Symbol | Definition |
|---|---|
| $y_i$ | The number of speakers of languages investigated |
| $\beta_0$ | Irrelevances |
| $X_{2i}$ | The geographic influencing factor |
| $X_{3i}$ | National individual per capita |
| $X_{4i}$ | The total number of immigrants of the investigated area |
| $\mu_i$ | Error term |
| N | The number of speakers of languages based immigrant influence |
| K | The influence coefficient matrix of populations |
| P | The number of populations in countries investigated |
| $X_1$ | The number of speakers of languages as the first language |
| $X_2$ | The number of speakers of languages as the first language |
| $K_1$ | The influence coefficient matrix of populations of the first language |
| $K_2$ | The influence coefficient matrix of populations of the second language |
| *a* | The influence factor of literacy rate |
| *b* | The influence coefficient of economics |
| *c* | The influence coefficient of education |
| *d* | The influence coefficient of geographic environment indicator |
| LR | Literacy rate |
| $x=\left(x_{1,}x_{2,}\cdots,x_n\right)^T\in R^n$ | Input vector |
| $w=\left(w_1,w_2,\cdots,w_n\right)^T\in R^m$ | Weight matrix |
| $f(x)$ | Output vector |
| $\phi(x,c_i)$ | Radial basis function |
| $c_i$ | The *i*-th center of Radial basis function |

Table 2. Variables and parameters

# 3. Model Construction

## 3.1.1 Part I: Trends prediction and geographic distribution of Languages

- **Trends prediction of language population:**
  From article [1] and website [2], we can know that the language distribution is diverse and changing over time. And there some consensus on the languages' current distribution status and evolution trends:
- Some continents have more languages than others, which is mostly influenced by geographic factors and national policy. And not all continents are equally diverse in the number of spoken languages. Whereas Asia leads the statistic with 2,301 languages, Africa follows closely with 2,138.

- There are also about 1,300 languages in the pacific, and 1,064 in South and North America, Europe, despite its many nation-states, is at the bottom of the pack with just 286. And we also can see that there is significant quantity variance between different languages.
- The list of world's mainly spoken languages based on population is almost stabe and unchangeable. To some extent, present world language pattern and linguistic order are defined by these most popular languages.
- Endangered languages exist extensively around the world and concentrate upon five hot spot regions: Eastern Siberia, northern Australia, central South America, Oklahoma, Pacific northwest. At past human language experts usually consider population as the main parameter, for instance, if considering endangered languages mainly manifest in small population or little population and then we will take for the big population is in the safe status. At present its affecting factors is various and complicated. We divide it into 'endangered language' and 'decay language'. In the first type, languages are mainly by environmental pattern, which means this language directly extinct with its population spoken. In the latter type, languages are most influenced by economical society and the changes of social formation can result in their extinction. We call this pattern as 'socioeconomic pattern'.
- In conclusion, we can see that today's world language formation is experiencing a changing influence pattern. This is to say, the language threatened by the death of the natural ecology and the migration and other factors caused the change of the language endangered mode caused by the socio-economic factors.

And we mainly focus on this latter pattern, for those assumptions as above.

We formulate a Multivariable Linear Regression Model to simulate the number of top ten language speakers.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \mu_i \tag{2}$$

The three factors data satisfy the following equation.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{3}$$

For the assumptions (1,2,6), considering the geographic environment factors has the same basic influence on the numbers of language users with population. And it remains stable. We refer to *World Bank* to get the world climate-distribution and analyze it to visualize a table that can be used to describe the degree of attraction to

language users as weighting factors. This process is under analytic Hierarchy process.

The symbol $X_{2i}$ represent the weighting factors about the influence between geographic environment and language users. We divide it into nine kinds of gradient, from 0.1 to 0.9.

| Language | Climate type | Weighting rate | Main country |
|---|---|---|---|
| Chinese, Mandarin | Temperate monsoon climate | 0.9 | China |
| Spanish | Temperate marine climate | 1 | Spain |
| English | Temperate marine climate | 1 | United Kingdom |
| Hindi | Tropical monsoon climate | 0.6 | India |
| Portuguese | Mediterranean climate | 0.8 | Portugal |
| Bengali | Subtropical monsoon climate | 0.9 | Bangladesh |
| Russian | Temperate continental climate | 0.4 | Russian Federation |
| Japanese | Temperate monsoon climate | 0.8 | Japan |
| Javanese | Tropical rainy climate | 0.5 | Indonesia |
| German, Standard | Temperate marine climate | 1 | Germany |

Table 3. the weighting factors of geographic environment

We plug the historical data in data fitting tools to find that the error is limited to 5%, which proves these data have pretty well linear relation.

### 3.1.2 Immigrant

To predict human migration patterns, we apply radiation model. This model can predict average migration flux from one location to another location. Suppose we are considering the average migration flux from location *i* to *j,* then the estimation is

$$\left\langle T_{ij} \right\rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{1}$$

Where $m_i$ and $n_j$ are populations of i and j respectively, and $r_{ij}$ is the distance between *i* and *j*.

Back to our model we apply radiation model to estimate the number of migrants from country *i* to country *j*.

$$T_{ij}(n) = \sum_{j \neq i} \frac{l^{(i)}\left[P^{(j)}(n-1)\right]^2 * P^{(i)}(n-1)}{\left[P^{(j)}(n-1) + S^{(ij)}(n-1)[P^{(j)}(n-1) + P^{(i)}(n-1) + S^{(ij)}(n-1)\right]} \tag{4}$$

Then the number of emigrants of $E_i(n)$ is

$$E_i(n) = \sum_{j \neq i} T_{ij}(n) \tag{5}$$

The number of immigrants of $I_i(n)$ is

$$I_i(n)=\sum_{j\neq i} T_{ji}(n) \tag{6}$$

The net migrants that move to country I in the *n*-th year is

$$M_i(n) = I_i(n) - E_i(n) \tag{7}$$

Using the above formula, we can get estimation of the population of country *i* in the *n*-th year from the population in the *n*—1-th year.

$$P_i(n) = B_i(n) + M_i(n) \tag{8}$$

### 3.1.3 Prediction of number of speakers on the geographic distribution

From assumptions 1,2,3,4 we can get

$$N=KP \tag{9}$$

Where X, P are vectors, K is a matrix, X and P represent the total speakers of languages investigated and the population of countries investigated plus immigrants. K is the coefficient matrix, where K[*i*] [*j*] is a positive real number related to the influence factors such as economics and education and a function of time.
From assumptions about influence factors(3,4,5,6), we can get the following formulas：

$K = K_1+K_2$

$X= X_1 + X_2 = K_1P + K_2P$

$$K_1[i][j]=\begin{cases} 0 & \text{if language i is the official language of country j} \\ a*LR & otherwise \end{cases}$$

$$K_2^{(n)}[i][j]=\left(C_{ij}^{(n)}*ECO+d_{ij}^{(n)}*EDU\right)*b_{ij}^{(n)}$$

Where a, b, c can be regarded as linear change over time, and d remain constant. They can be gotten from the data known.

$$X_{1(i)}^{(n)}=\sum_{j\in S} a_j LR_j^{(n)} P_j^{(n)}$$

$$= \left[ LR_{j1}^{(n)} P_{j1}^{(n)} \cdots LR_{jm}^{(n)} P_{jm}^{(n)} \right] * \begin{bmatrix} a_{j1} \\ a_{j2} \\ \vdots \\ a_{jm} \end{bmatrix}$$

$$X_{1(i)} = \begin{bmatrix} LR_{j_1}^{(n_1)} & \cdots & LR_{j_m}^{(n_1)} \\ \vdots & \ddots & \vdots \\ LR_{j_1}^{(n_q)} & \cdots & LR_{j_m}^{(n_q)} \end{bmatrix} * \begin{bmatrix} a_{i_{j1}} \\ a_{i_{j2}} \\ \vdots \\ a_{i_{jm}} \end{bmatrix}$$

$$b_{ij}^{(n)} = \begin{cases} 0 & language\ is\ the\ national\ language\ of\ country\ j \\ \dfrac{\sum_{\in} T_{jk(n)}}{E_{j(n)}} & otherwise \end{cases} \tag{10}$$

Assume that $c_{ij}^{(n)}, d_{ij}^{(n)}$ do not change with time, then we have

$$X_{2(i)}^{(n)} = \sum_{j=1}^{N} K_{2ij}^{(n)} * P_j^{(n)}$$

$$= \sum_{j=1}^{N} b_{ij}^{(n)} c_{ij} ECO_j^{(n)} P_j^{(n)} \cdots b_{iN}^{(n)} ECO_N^{(n)} P_N^{(n)} \begin{bmatrix} c_{i1} \\ c_{i2} \\ \vdots \\ c_{iN} \end{bmatrix} +$$

$$\left[ b_{i1}^{(n)} ECO_1^{(n)} P_1^{(n)} \cdots b_{iN}^{(n)} EDU_N^{(n)} P_N^{(n)} \right] \begin{bmatrix} d_{i_1} \\ d_{i_2} \\ \vdots \\ d_{i_N} \end{bmatrix}$$

$$X_{2i} = \begin{bmatrix} b_{i_1}^{(n)} ECO_1^{(n_1)} P_1^{(n_1)} & \cdots & b_{i_N}^{(n)} ECO_N^{(n_1)} \\ \vdots & \ddots & \vdots \\ b_{i_1}^{(n)} ECO_1^{(n_q)} & \cdots & b_{i_N}^{(n_q)} ECO_N^{(n_q)} P_N^{(n_q)} \end{bmatrix} \begin{bmatrix} c_{i1} \\ \vdots \\ c_{i_N} \end{bmatrix} +$$

$$\begin{bmatrix} b_{i_1}^{(n)} EDU_1^{(n_1)} P_1^{(n_1)} & \cdots & b_{i_N}^{(n)} EDU_N^{(n_1)} \\ \vdots & \ddots & \vdots \\ b_{i_1}^{(n_q)} EDU_1^{(n_q)} & \cdots & b_{i_N}^{(n_q)} EDU_N^{(n_q)} P_N^{(n_q)} \end{bmatrix} \begin{bmatrix} c_{i_1} \\ \vdots \\ c_{i_N} \end{bmatrix} \tag{11}$$

$$= E_{c_i} c_i + E_{d_i} d_i$$

$$= \begin{bmatrix} E_{c_i} & E_{d_i} \end{bmatrix} \begin{bmatrix} c_i \\ d_i \end{bmatrix}$$

We use this equation set to predict the distribution of language users, the data all can be simulated or known.

## 3.2 Part II: Location Issue of New international offices based on RBF Neural network

As the assumptions 7,8,9 above, we select the ten countries to make comparative analyses. From the first model we can get the number of special language users, and we can process geographic distribution simulation through the second model.
Consider that it is one-sided based on the single data, we introduce feasibility of construction and cost performance analysis to prove that the international offices can used for a long time.
From above, we can list all the four part influence factors:

| |
|---|
| I.Natural environment factors: meteorology, geology, terrain |
| II.Economic environment : regional economic policy, customer demand, competitive pressure |
| III. Economic rationality factor: cost benefit, labor factor, land resource factor. |
| IV. The use of population in a language: changes in the population of the language used in the region. |

Table 4. The influence aspects of the Location of international offices

We choose Analytic Hierarchy Process to evaluate these influence factors above.

### 3.2.1 Model Description

### I.Establish a hierarchical structure

In the AHP model, elements at the same level play a dominant role in certain elements of the next level, and at the same time it is also governed by the last level of elements. The value of the judgment matrix manifests the significance of different elements, we use proportional scale 1~9 to measure its significance.

| Scale | Definition |
|---|---|
| 1 | It represents compare two elements, and they are of equal importance. |
| 3 | It represents compare two elements, and the former is moderate importance than the latter. |
| 5 | It represents compare two elements, and the former is obviously important. |
| 7 | It represents compare two elements, and the former is highly important |

| | |
|---|---|
| 9 | It represents compare two elements, and the former is absolute important. |
| 2,4,6,8 | They represents the intermediate value of the above adjacent judgment. |
| Reciprocal | If the significance of element *i* compare with *j*'s is $a_{ij,}$ which means the importance ration is $a_{ji}$ =1/ $a_{ij.,}$ |

Table 5. The matrix scale and its definition.

## II.Constructive judgment matrix

## III. Sequencing by exponents(It is based on the model of II)

## IV.Total ordering

## V. Consistency test(Standard as follows)

During the calculation process, according to the consistency test of the judgment matrix, we use

$C.I = \frac{\lambda_{\max} - n}{n-1}$, $n$ represents the matrix order.

When $C.R = \frac{C.I}{R.I} < 0.10$, we assume the judgment matrix satisfy the consistency, otherwise it is need to be adjusted. We use Stationary random index *R.I* to judge it.

| Order of Matrix | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| *R.I* | 0.00 | 0.00 | 0.58 | 0.90 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.49 |

Table 6. Stationary random index *R.I*

### 3.2.2 RBF Neural network

1.Model diagram

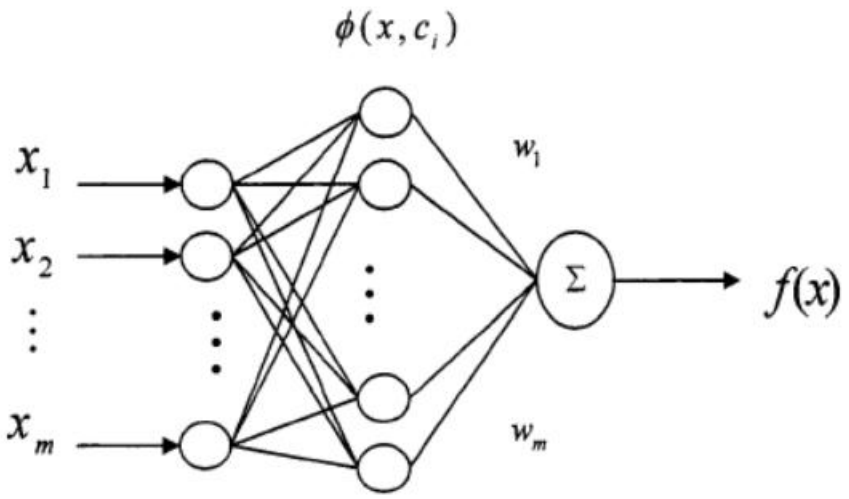


Figure 1: The model diagram of RBF Neural network

We can formulate this model from these process.

$$f(x) = \sum_{i=1}^{m} w_i \phi(x, c_i) \tag{12}$$

We consider that the international offices are in a indiscriminate status. And based on this case ,we choose a typical international service company as our example to simulate the proper location.

We formulate the influencing factor distribution table as follows:

| | | | |
|---|---|---|---|
| International office location evaluation system. | Natural environment factors | Climate condition | $u_{11}$ |
| | | Geographic condition | $u_{12}$ |
| | | Topographic condition | $u_{13}$ |
| | Economic environment | Regional economic policy | $u_{21}$ |
| | | Customer demand | $u_{22}$ |
| | | Competitive pressure | $u_{23}$ |
| | Economic rationality factor | Cost | $u_{31}$ |
| | | Labor link | $u_{32}$ |
| | | Land resource factor | $u_{33}$ |
| | The use of population in a language | The trends of the number of language users in this area | $u_{41}$ |

Table 7. Influencing factor distribution table

We use the data known to process this algorithm and can simulate the proper location. The data we add to the appendix as the first table.

## 2. AHP evaluation decision

Firstly, stratify the data into different levels

① The first level: $U=\{u_1,u_2,u_3,u_4\}$

② The second level: $u_1=\{u_{11},u_{12},u_{13},u_{14}\},u_2=\{u_{21},u_{22},u_{23},u_{24},u_{25}\}$
$u_3=\{u_{31},u_{32},u_{33},u_{34}\},u_4=\{u_{41},u_{42},u_{43},u_{44},u_{45}\}$

And then formulate the judgement matrix：
For the elements in the delivery hierarchy, the elements in the hierarchy can be compared with each other in turn, so as to establish a series of judgment matrixes. The judgement matrix $M=(a_{ij})_{n\times n}$ has the following nature:

$$a_{ij}>0,a_{ij}=\frac{1}{a_{ji}},a_{ii}=1(i,j=1,2,\cdots,n) \tag{13}$$

Suppose the 'criteria - sub-criteria' judgment matrix respectively as M1,M2,M3,M4, and then:

$$M1=\begin{pmatrix}1&2&1/5&1/7\\1/2&1&1/4&1/5\\5&4&1&1/3\\7&5&3&1\end{pmatrix},\ M2=\begin{pmatrix}1&3&5&6&4\\1/3&1&4&5&3\\1/5&1/4&1&2&1/3\\1/6&1/5&1/2&1&1/5\\1/4&1/3&3&5&1\end{pmatrix}$$

$$M3=\begin{pmatrix}1&3&7&5\\1/3&1&5&2\\1/7&1/5&1&1/4\\1/5&1/2&4&1\end{pmatrix},M4=\begin{pmatrix}1&3&7&9&5\\1/3&1&5&7&3\\1/7&1/5&1&2&1/3\\1/9&1/7&1/2&1&1/5\\1/5&1/3&3&5&1\end{pmatrix} \tag{14}$$

We can calculate these parameters to generate the evaluation matrix as follows:

| | Feature vector | | | | | $\lambda_{max}$ |
|---|---|---|---|---|---|---|
| M1 | 0.0865 | 0.0704 | 0.2812 | 0.5618 | — | 4.2065 |
| M2 | 0.4711 | 0.2618 | 0.0705 | 0.0458 | 0.1509 | 5.2805 |
| M3 | 0.5705 | 0.2355 | 0.0525 | 0.1415 | — | 4.1231 |
| M4 | 0.5153 | 0.2621 | 0.0575 | 0.0357 | 0.1293 | 5.1848 |

Table.8 evaluation matrix

According to the weight factors, we can make the weighted sum to get the evaluation indicator. We can continue to use this algorithm to iterate

all the ten factors to get the final results. At last relist all the results, we get the proper six office locations.

### 3.3 Part III: Optimize the Location Issue based on minimized cost

We choose the simulated annealing algorithm to develop the last location model. In this process, we try to minimize the cost and to achieve its target about language users.

Using simulated annealing algorithm, one of the regions with highest economic radiation power can be selected from the coverage of several languages as the center of gravity of the region. It will be the primary consideration when siting. The principle of the annealing algorithm is that when the temperature drops to a very low value, the material enters the minimum energy state with a great probability. After many times of transfer, the material reaches the goal of lowering the temperature, that is, the optimal state.

We can use the Asia as the example to analyze, there are Chinese, Japanese, Hindi and Russian. We can select proper cities based on the four parts of influencing factors. Firstly, under the evaluation pattern, we select 15 proper cities in Asia and consecutively number as {1,2,3, ⋯, 15} , randomly select one city as the original point

$$\sigma_t = (x_i, y_i); i = 1,2,3\cdots 14,15; t = 1,2,3,\cdots; \tag{15}$$

And then define the next point as $\sigma_{t+1} = |\sigma_t|(1+\vec{e}_i);$

In the equation, parameter $\vec{e}_i = \dfrac{Q_i - \sigma_i}{|Q_i - \sigma_i|}$

We can get the total change values.

$$\Delta W = \sum_{1}^{14} E_{t+1} - E_t;$$

It conforms to the equation

$$P = \begin{cases} 1 & \Delta W < 0 \\ \exp\left(-\dfrac{\Delta W}{T}\right) & \Delta W > 0 \end{cases} \tag{16}$$

Then we choose a proper cooling parameter to end the process. At last we get the final coordinate and confirm the proper city. In this example, we get the city 'Hong Kong'.

In the same way, we can dispose other continent's data to confirm the proper locations.

## 4. Model Extension and Simulation Analysis

We construct two models to support our simulation and decompose our simulation into three section which correspond to the two problems. The first problem A is based on the Multivariable Linear Regression Model and we choose time trend

analysis to simulate the factors' time data. It concentrates on the prediction of the language distribution. To the second problem B, we consider the economic factors, geographic factors, related policies and feasibility levels etc. that can affect the location selection as the influencing factors. Firstly, our objective level is to seek how to make the location's clients master English and other language as much as possible. Secondly, it is based on the last model to optimize this location decisions, which means using offices as little as possible to make the final objective and minimize cost.

## 4.1 Problem A-1: The Model of the distribution of various language speakers over time.

Considering all these influencing factors of assumption, we can dispose these data into matrix, as the education as an example:

$$\beta_1 = (\ \ 0.9843,\ 0.8967,\ 0.6253,\ \ 0.9341\ ,\ 0.9407\ \ )$$

In the similar ways, we can recalculate all other factors, and respectively get each equation about the number of different language users.
For the Chinese as the example, the regression equation is as follows:

$$Q(t) = [\int \left(\frac{\partial M(t)}{\partial t} \cdot \beta_i\right) dt + 1] * Q_0 = \frac{18.5 * e^{0.15\ \langle t-2017 \rangle}}{2431+39*e^{0.15\ \langle t-2017 \rangle}} - \frac{1}{39+23*e^{-2*(t-2017)}}$$

For other languages, we just repeat this process to simulate the final equation of each language

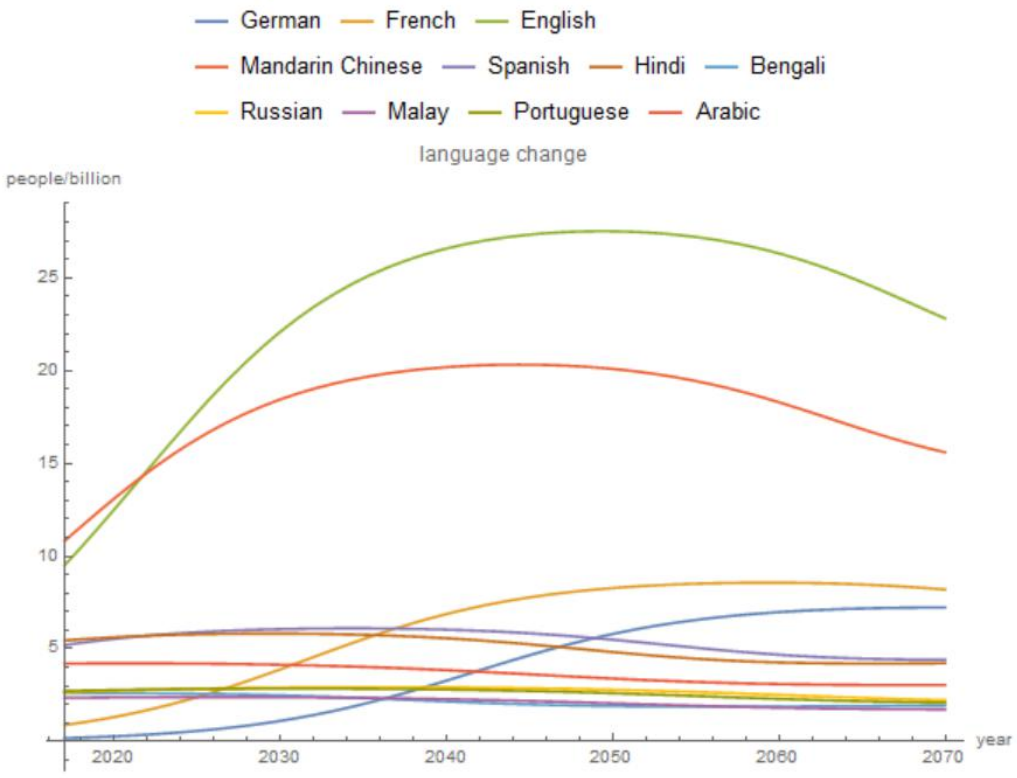


Figure 2. The distribution of various language speakers over time

Then we can get each language's trend predication curves, and put them in the same diagram. We can get the conclusion.

## 4.2 Problem A-2: The top-ten lists' changes in the next 50 years

We use the model above to simulate the number of language users over the next 50 years. And to get the new list as follows:

| Rank (original) | Rank (future prediction) | Country | Language |
|---|---|---|---|
| 3 | 1 | United Kingdom | English |
| 2 | 2 | Spain | Spanish |
| 1 | 3 | China | Chinese, Mandarin |
| 5 | 4 | Portugal | Portuguese |
| 4 | 5 | India | Hindi |
| 6 | 6 | Bangladesh | Bengali |
| 7 | 7 | Russian Federation | Russian |
| 9 | 8 | Indonesia | Javanese |
| 10 | 9 | Germany | German, Standard |
| 8 | 10 | Japan | Japanese |

Table 9. The top-ten lists' changes in the next 50years

From the recent years' language research, we can know that world language structure is mainly based on the most popular languages. And all the language evolution can be divided into two patterns. One is direct related with the population and the population play the most important role in influencing this kind of language. The other is based on socioeconomic factors, which represents more complicated and various factors. And it includes social policy, economic conditions etc. In conclusion, the pattern is developing from simple to complicated.

As the above new table, we can see that Chinese and Japanese respectively degrade from 1 to 3 and 8 to 10. We consider that Chinese and Japanese both are relatively single language family, especially for Japanese. Although Chinese consist of the Sino-Tibetan as leading part, its language also exist in a limited region and keep itself unrelated with other language families. Because of these elements, the two languages are also mainly in natural pattern, and population has the biggest impact. As population growth slowing during these years, they cannot remain its original ranks.

On the contrary, Indonesia still keep its population growth velocity in a high condition, and its rank is promoted by this impact. It represents in the table as from 9 to 8.

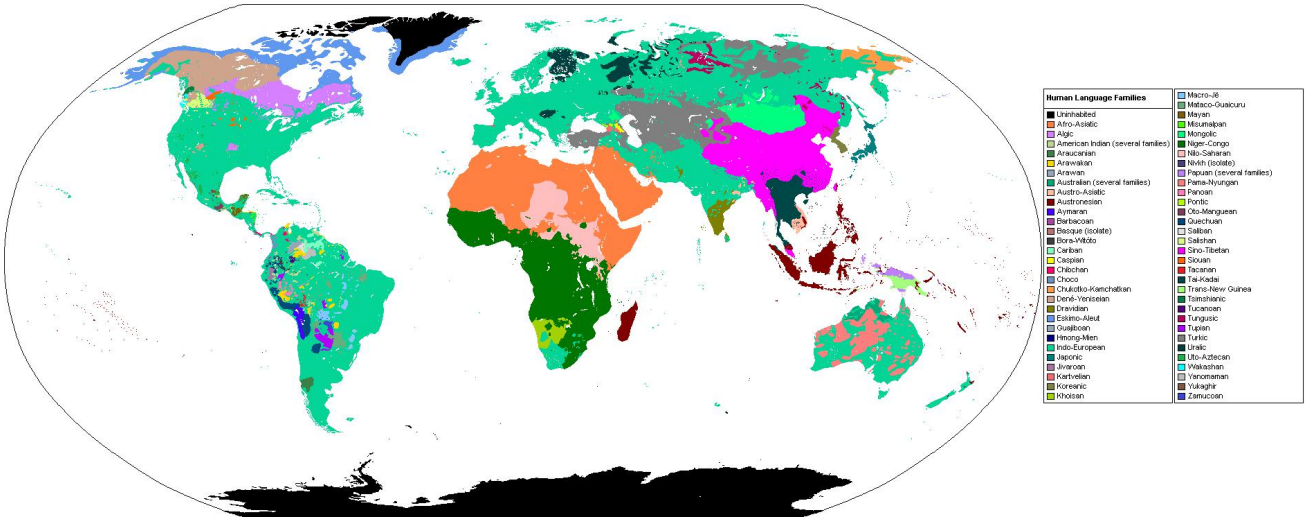


Figure 3: The distribution diagram of the world language family

For the English and Spanish, they respectively upgrade from 3 to 1 and remain its position at the rank 2. We can focus on the broadest colors of the distribution diagram. It reflects its geographic distribution and span of the language family. Analyze that level changes conform with this case. We consider that these languages have great relation with other language families and mainly affected by social factors, economic factors and other complicated factors.

In conclusion, the simulation of these trends of language distribution can be into two kind of evolution pattern. And some is mainly related with population and the stable geographic environment factors. The other is leaded by the more complicated factors. And we deem that the both pattern will develop towards the same direction.

## 4.3 Problem A-3: The geographic distributions of these languages change over next 50 years

From our assumption (2,4,5), Immigration will only directly influence the distribution of second language.

We select China as the representative, and we can get the other countries that have immigrant relations with China. According to the radiation model, we can simulate the future language distribution over time.
Repeat this process, we can roughly simulate the geographic distribution over 50 years.

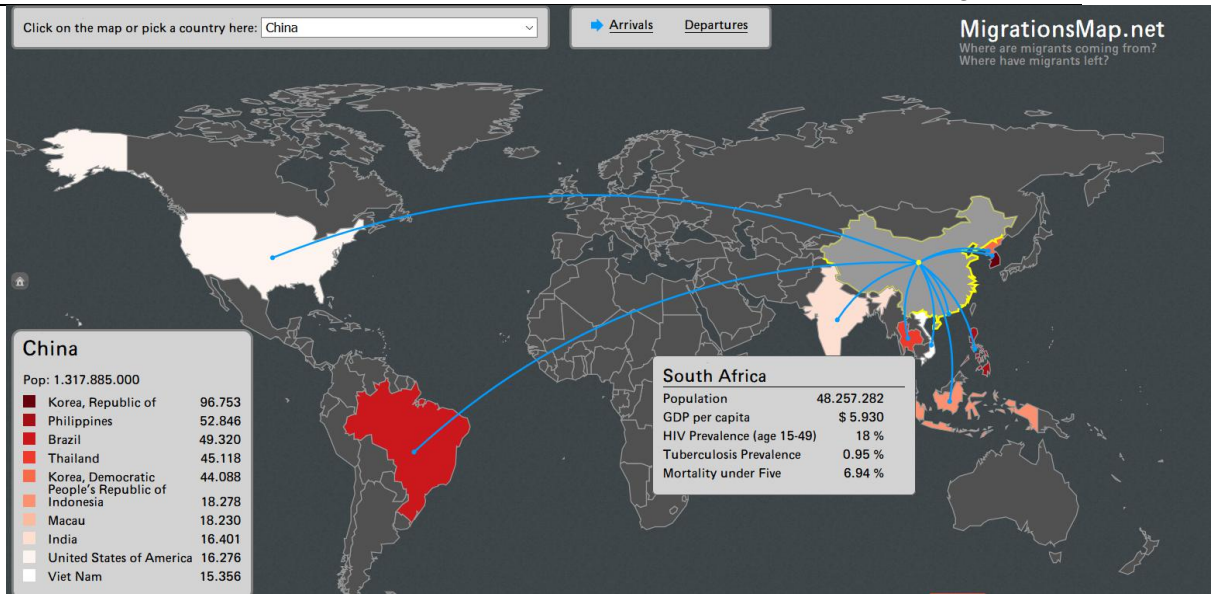


Figure 4. The immigrant of China

And we can only get a rough result that the geographical distribution of language shows convergence and Matthew effect. It is to say, that the scope of the use of English and Chinese will be larger and larger, while other languages will gradually be replaced or disappeared.

## 4.4 Problem B-1: The proper locations of international offices and recommendations

Based on the model constructed, we can analyze the cities' weighted value under the analysis of AHP algorithm. And plug the data of influencing factors in this model

$$f(x)=\sum_{i=1}^{m} w_i \phi(x,c_i)$$

We get the evolution list of the world's major cities. And then we can get the weighted indicators of these cities based on economic factor, social policy factor, feasibility level.

$$R_1=\begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{31} & r_{32} & r_{33} & r_{34} & r_{35} \\ r_{41} & r_{42} & r_{43} & r_{44} & r_{45} \end{bmatrix} R_2=\begin{bmatrix} r'_{11} & r'_{12} & r'_{13} & r'_{14} & r'_{15} \\ r'_{21} & r'_{22} & r'_{23} & r'_{24} & r'_{25} \\ r'_{31} & r'_{32} & r'_{33} & r'_{34} & r'_{35} \\ r'_{41} & r'_{42} & r'_{43} & r'_{44} & r'_{55} \end{bmatrix}$$

$$\mathrm{R}_3=\begin{bmatrix} r''_{11} & r''_{12} & r''_{13} & r''_{14} & r''_{15} \\ r''_{21} & r''_{22} & r''_{23} & r''_{24} & r''_{25} \end{bmatrix}$$

The four criteria and the 10 indicators in the comprehensive evaluation system are not the same in the overall evaluation. The important status should be given more weight, on the contrary, the less weight should be given.

Using the simulated annealing algorithm to process these data, and after clustering these data, we can get the final diagram that can represent the long-term selection plan.

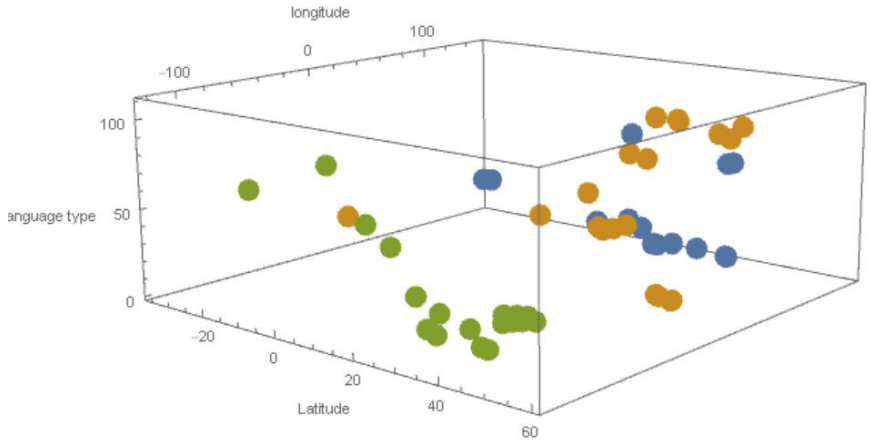


Figure 5. The Clustering of proper cities data in long term

In conclusion, we simulate data to choose the six cities in short term as

$$\{Paris, Hong\ Kong, Los\ Angeles, Madrid, mumbai, Cairo.\}$$

In the long term, the location selection should be

$$\{Berlin, Beijing, San\ Francisco, Rio, mumbai, Cairo.\}$$

### 4.5 Problem B-2: To optimize the location Issue based on minimized cost

As the changing nature of global communications, and substantial changes in global communications will impact all walks of life, the maturity of translation software, the trend of language itself, the global market increase, etc. All these elements make the language is no longer so important. In multinational corporations, there is no need to consider the language issue as the first factor to consider when selecting a site.

Otherwise, company need to maximize its profits, any investment about selecting a new site which must can get positive profit. As the present worth equation

$$NPV = \sum_{t=0}^{n} \frac{I_t}{(1+i)^t} - \sum_{t=0}^{n} \frac{O_t}{(1+i)^t}$$

If the world language condition has changed as so, we must focus on the decreased marginal income. With topic in an effort to save client company resources, we consider to reduce the 6 offices to 5 offices at least.

As the world's language distribution on the world geographic map in the future, we consider 5 offices can also cover more than half-population in main language users.

## 5. Strengths and Weaknesses

### 5.1 Strengths

1. Regression analysis is simpler and more convenient when analyzing multi-factor model;

2. Using the regression model, as long as use the same model and data, we can calculate the only result by using standard statistical method, but in the form of figure and table, the interpretation of the relationship between data often varies from person to person, different analysts draw a fitting curve is probably not the same;
3. Regression analysis can accurately measure the correlation degree of each factor and the degree of regression fitting degrees, and improve the effect of the prediction equation; Due to actual a variable in the regression analysis method is rarely influenced by individual factors, only to pay attention to the suitable range of model, so a yuan regression analysis for the influence that there is a dependent variable, significantly higher than the other factors of variable is used. Multivariate regression analysis is applied to practical economic problems and is used when multivariate effects are combined.

4. RBF network is based on the data provided, learning and training, find out the internal relations of input and output, so as to calculate the problem of the solution, rather than according to the experience knowledge. Therefore, it has the adaptive function, can weaken the influence of artificial factors in determining the index weight.

## 5.2 Weaknesses

1. Sometimes, in regression analysis, choosing which factor and which expression is a kind of speculation, which influenced the diversity of power utilization factor and some factors of unpredictable, regression analysis, in some cases was limited.
2. We cannot visualize the geographic distribution of various languages based on radiation model, and only can estimate the clustering condition in different regions

## 5.3 Future work

From these models above, we get the result that the world's language is towards the same way. And these effect manifests great in the future. We think the trends may be right and the more work need to tessellate this process. Try to minimize the error through bigger database. And then based on this condition, the influence of language will decrease acutely. Each one will be influenced and need to make changes as soon as possible. And the international companies can make proper adjustments of development project in the long run.

Otherwise, the combination between AHP and Genetic Simulated Annealing algorithm is a little optimize. We think that the rating analysis in AHP can be used in wider areas and fields, such as climate estimation, agriculture etc.

# References

[1] Wang Chunhui. Language Order in Contemporary World. , 2016, 1(4): 69-82.

[2] Noack, R. and Gamio, L. (April 23, 2015). The World's Languages in 7 Maps and Charts. The Washington Post. Retrieved from https://www.washingtonpost.com/news/worldviews/wp/2015/04/23/the-worlds-languages-in-7-mapsand-charts/?utm_term=.a993dc2a15cb

[3] A universal model for mobility and migration patterns Filippo Simini, Marta C. Gonza´lez, Amos Maritan & Albert-La´szlo´ Baraba´si.
Nature 484 (7392) | doi:10.1038/nature10856

[4] *World Bank*

[5] *Ethnologue*

# Appendix

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $U_{11}$ | 0.91 | 0.71 | 0.88 | 0.98 | 0.88 | 0.66 | 0.73 | 0.92 | 0.72 | 0.92 | 0.85 | 0.75 |
| $U_{12}$ | 0.76 | 0.68 | 0.85 | 0.93 | 0.90 | 0.91 | 0.88 | 0.93 | 0.68 | 0.68 | 0.65 | 0.81 |
| $U_{13}$ | 0.88 | 0.71 | 0.78 | 0.86 | 0.61 | 0.96 | 0.99 | 0.68 | 0.66 | 0.69 | 0.61 | 0.86 |
| $U_{21}$ | 0.85 | 0.82 | 0.81 | 0.61 | 0.77 | 0.61 | 0.61 | 0.79 | 0.81 | 0.98 | 0.71 | 0.82 |
| $U_{22}$ | 0.87 | 0.99 | 0.62 | 0.64 | 0.63 | 0.87 | 0.97 | 0.63 | 0.84 | 0.69 | 0.83 | 0.76 |
| $U_{23}$ | 0.85 | 0.96 | 0.84 | 1.00 | 0.97 | 0.60 | 0.97 | 0.63 | 0.84 | 0.69 | 0.83 | 0.76 |
| $U_{31}$ | 0.81 | 0.76 | 0.96 | 0.67 | 0.61 | 0.84 | 0.83 | 0.83 | 0.74 | 0.82 | 0.94 | 0.93 |
| $U_{32}$ | 0.71 | 0.89 | 0.96 | 0.80 | 0.82 | 0.95 | 0.94 | 0.84 | 0.79 | 0.60 | 0.95 | 0.83 |
| $U_{33}$ | 0.70 | 0.89 | 0.77 | 0.72 | 0.96 | 0.88 | 0.90 | 0.71 | 0.84 | 0.67 | 0.93 | 0.87 |
| $U_{41}$ | 0.90 | 0.88 | 0.84 | 0.65 | 0.76 | 0.87 | 0.92 | 0.64 | 0.71 | 0.61 | 0.90 | 0.81 |

Table x. Location historical data of one certain service company