For office use only	Team Control Number	For office use only
T1 _____	**74867**	F1 _____
T2 _____		F2 _____
T3 _____	Problem Chosen	F3 _____
T4 _____	**B**	F4 _____

2018

MCM/ICM

Summary Sheet

Models to fit and predict language trends : application on selecting international offices

The trend of development of languages is worth of research and is a crucial factor to be considered for international companies.

We study the change of distribution of language speakers from six aspects: economy, immigration, education, trade, tourism and language families. When selecting locations for new international offices, besides the language effect mentioned above, we also take into account the competency of the city itself, the cost and relative geographical relations between distinct cities.

The first stage, to establish the model 1 for number of native speaks and model 2 for total language speakers and the distribution of various language speakers over time, we use PCA method to pick out five factors which are of most importance. And we use the ratio of people speaking each language as the probability for a person to that language. We use polynomial model to fit the ratio with the five factors and use Lasso method to choose to use linear regression. Also, including the extra factor called language families , we use linear regression to predict the factors, thus the probability to learn one language. Besides, after analysis, it is suitable to use the population of one country to predict the number of native speakers.

The second stage, we construct model 3 to predict what will happen to the first and second language in the next 50 years. We build a multi-agenda model to simulate human behavior which focuses on age structure, and various factors including birth and death rate and the possibility to learn a different new language. Each unit in this model iterates in pre-decided way which mimics nature process of language learning, first or second. One great strength of this model is its ability to give reasonable results without involving complex theoretical analysis and mathematical derivation. Its accuracy can be refined by more detailed initial value and more precise factors. We based on the result of this model to analyze the trend and give our suggestions.

The third stage, we construct model 4 to quantify the total profit of opening new offices in different cities and our goal is to maximize the profit. We consider four major indicators of profit: native language of the city, the competency of the city itself, the living price index of a city and its distances with other cities we have chosen. With regard to the competency of a city, we draw on the experience of HDI (Human Development Index) and consider from three respects: economy, education and immigration, which implies the attraction and potential of the city. As done in the previous stages, we quantify these three factors and convert them into numbers between 0 and 1 and then define competency as the geometric mean of those three values. Similarly, we extract data of living price index of all the cities and then divide them by the maximum one to get a value between 0 and 1. Finally, we consider the average distances between all the cities we choose as a punishment term, which suggests that we hope the number to be properly bigger so that the distribution of the offices can be more uniform. Last, we select a comprehensive linear model of all the four factors mentioned above and endow the competency of the city with a relatively heavier weight. From a short-term perspective, we choose the average rank of ten years and choose the average rank of fifty years from a long-term then we can get a final rank of all the cities.

Finally, we conduct sensitivity analysis on model 2 and model 4. In model 2, we consider Polynomial fitting instead of Linear fitting (a special case) i.e. we observe the change of the weight of five factors when increasing the highest degree of the model. For fear of overfitting, we set the highest degree $n<6$. In model 4, we gradually change the coefficient of the linear model in front of the term of competency to see the change of ranks. After summarizing our results, we analyze the strengths and weaknesses of our models and write a letter to our client to inform them of our results and suggestions.

Contents

1 Introduction
1.1 Background
1.2 Problem Restatement
1.3 Data resources
2 Analysis of Overall and Key Points
3 Assumptions and Justification
4 Symbols and Definitions
5 The Model Theory
5.1 Model 1: To predict the numbers of native speakers for each listed language
5.2 Model 2: To fit the numbers of each language with some preeminent factors
5.3 Model 3: To predict the numbers of each language with multi-agenda model(MA model)
5.4 Model 4: To determine the numbers and locations of international offices
6 Model Results
7 Sensitivity Analysis
8 Further Discussion
9 Strengths and Weaknesses
9.1 Strengths
9.2 Weaknesses
10 Conclusions
11 A letter to client company
References
Appendices

1.Introduction

1.1 Background

There are approximately 6,900 distinct languages spoken on Earth at present and according to statistics, the following ten languages are the most popular which are spoken by about half of human beings as a native language: Mandarin(incl. Standard Chinese), Spanish, English, Hindi, Arabic, Bengali, Portuguese, Russian, Punjabi, and Japanese. Simultaneously, a great number of people also speak a second language, a third language, etc. When it comes to the development of the trend of the total numbers of speakers of a certain language (native speakers plus second or third, etc. language speakers), there are several significant factors ought to be taken into consideration: social factors, political factors, economic factors, cultural factors and so forth.

With international cooperation becoming more and more significant, it is very necessary and challenging for those international companies to choose offices worldwide in order to be more international and gain more profits. Besides other factors, distribution of the number of people speaking each language is also an important factor when considering making plans.

1.1 Data Required and Resources

We require the following data for our modeling and predicting :

(1) The population of all countries in recent ten years

(2) The number of speakers as native speakers of the fifteen most spoken languages in recent ten years

(3) The number of speakers as non-native speakers of the fifteen most spoken languages in recent ten years

(4) The birth rate and death rate of all countries in recent ten years in the sub-age perspective in recent ten years

(5) The Per Capita GDP of all countries in recent ten years

(6) Net immigration number of people to a particular country in recent ten years

(7) The Human Development Index (HDI) of all countries in recent ten years

(8) The number of International Tourists Arrival to a particular country in recent ten years

(9) Net foreign direct investment to a particular country in recent ten years

(10) Total public expenditure on education of all countries in recent ten years

We extract all data from Official World Bank Database.

2. Analysis of Overall and Key Points

2.1 Overall Analysis

We generally select the fifteen most spoken languages and thirty most typical countries which speak those languages as official languages as the object of study, which is showed in the following table 2.1. (where the question mark means the language is not the official language of any country or the data can’t be reached officially)

In the table below, after each language and each country, there exists one number to represent them in order to make the coding simple and uniform.

Language							
Mandarin(incl. Standard Chinese)(4)	China(0)						
English(11)	United States of America(18)	United Kingdom (19)	Canada(80%) (6)				
Hindustani(5)	India(7)						
Spanish(10)	Argentina (24)	Colombia (27)	Mexico(9)	Spain(20)	Peru (17)	Chile (15)	Uruguay (1)
Arabic(0)	Saudi Abaria(25)	Yemen(2)	Sudan(26)	Syria(22)	Iraq (3)	Egypt (23)	
Malay(6)	Malaysia(28)						
Russian(2)	Russia(5)						
Bengali(7)	Bangladesh (10)						
Portuguese(8)	Brazil(12)	Portugal (21)					
French(3)	France(16)	Canada（20%）(6)					
Hausa	?						
Punjabi	?						
Japanese(9)	Japan(13)						
German(1)	Germany(13)	Austria (11)					
Persian(12)	Iran(4)	Tajikistan (8)					

Table 2.1

3. Assumptions and Justification

A. **About the definition of first language**: we assume that the definition of first language is **native language or official language**.

B. **About the data of numbers of people**: as far as one particular language is concerned, we ignore some countries whose number of speakers of that **language account for less than 1%** of the total number of speakers.

C. **About data deficiency**: when some data (less than 1%) is missing or can't be reached officially, we supply the missing data **with existing data by linear regression**. We make sure that the R square of the linear regression is approximately 0.99 so that the data is relatively precise and predictable.

D. **About the cellular automation**: we assume that in the model to fit and predict the numbers of speakers with respect to each language, namely in the process of cellular automation, the **gender**

ratio amongst people we are concerned is 1:1. And the **total birth rate is considered to only impact on the group of people aged 20-50**.

4. Symbols and Definitions

In the section, we use symbols for constructing the model as follows:

In order to conveniently show each language, each country and each year, we use the notation "language j or country j", "year i", "factor k" to respectively standing for the j-th language or country, the i-th year and the j-th factor.

Symbols	Definitions
$x_{i,j,k}$	The value of factor k with respect to the country j in the year i
$Y_{i,j}$	The number of people speaking language j in the year i
$z_{i,j,k}$	The value of factor k with respect to language j in the year i.
$y_{i,j}$	The value of the number of speakers with respect to language j in the year i.
$a_{i,j}$	The coefficient of factor j with respect to language i
c_i	The constant in the linear regression of language i.
$b_{i,j}$	The birth rate of people speaking language j in the year i
$I_{i,j}$	Total profit of opening a new office in country i in the year j
$p_{i,j}$	The competitive power of country i in the year j
w_i	The price index of country i
d_{j_1,j_2}	The distance between location j_1 and j_2
n	The number of new offices
$d_{i,j}$	The birth rate of people speaking language j in the year i

Table 4.1

5. The Model

5.1 Model to predict the numbers of native speakers for each language

5.1.1 Establishment of the model

As the assumption we have pointed above and according to the UNESCO, we define native

languages as one is born to learn to speak and a nation claims to use officially. Therefore, it is convenient and of considerably precise that we use the prediction of the population of each country excluding the immigrants who get the country's nationality as a model to predict the numbers of native speakers for each language in the next 50 years.

5.1.2 Parameters and data of the model

As we describe in the part 5.1.1, we collect date of the predicted population of each country we are interested in for the next 50 years according to Official World Bank Database and sum up the numbers of people when their native languages are the same to use as the numbers of native speakers for each language predicted in the next 50 years, called $nx_{i,j}$, meaning the number of native speakers with respect to language j in the year i.

5.2 To predict the numbers of each language with some preeminent factors

5.2.1 Establishment of the model

Considering the problem to predict the numbers of speakers for each language in the next 50 years, there should exists a model to determine the probability of a person tending to learn and speak some language(s). Combining all the factors that we have considered, after using principal component analysis, namely PCA, to reduce and choose some of the best-behaved factors as following:

A. Net Immigration Flow, the factor to show the immigration into the country. Considering the probability of a person who immigrates to the country to learn the corresponding language is very large, we choose this data from Official World Bank Database as a factor, called $x_{i,j,0}$, meaning the value of factor 0 with respect to the country j in the year i, and the same for the following as well.

B. Net Foreign Direct Investment Inflow, a factor to show the fund and other kinds of investment into the country. Considering the investment from foreign countries showing a considerably high probability for people from foreign countries, especially amongst fields like business, financial affair, industry, international cooperation and so on to learn the corresponding language, we choose this data from Official World Bank Database as a factor, called $x_{i,j,1}$.

C. Per Capita GDP, a factor to show the economic strength for the people living in the country. Considering the economic strength showing the happiness index of the people in the country partially but importantly, the probability to learn the corresponding language, even to immigrate to that country ought to be large, thus serving as factor using the data from Official World Bank Database, called $x_{i,j,2}$.

D. Total public expenditure on education, a factor to show educational strength especially advanced education. Considering the probability of the overseas students to learn the corresponding language to be very large, we choose this factor as a index to show the attraction for foreign students. Besides, education is high related to the probability of a person tending to learn other languages. Therefore, we use it a factor and collect data from Official World Bank Database, called $x_{i,j,3}$.

E. International Tourists Arrival Flow, a factor to show the numbers of tourists in the country. Considering the important need for a tourist, especially not short term, to master or at least know about the corresponding language, we choose it as factor to show the attraction for people in the foreign countries and the probability of their learning the language. Therefore, we collect the data

from Official World Bank Database calling the factor $x_{i,j,4}$.
Those above are the preeminent factors that we are interested in. And we, from Official World Bank Database, collect the total numbers of people speaking each language using it as a index to serve as the actual number to learn each language, called $Y_{i,j}$, meaning the number of people speaking language j in the year i. After gathering the number and data of the factors and the actual probability to learn each language, we use linear regression to fit the probability with factors in order to give a formula to predict the numbers in the next 50 years, which will be shown in the next part.

5.2.2 Implement and detail for the model

A. To process the data and make some reduction

In order to run the algorithm, we make those data divide by the corresponding sum to get them into the interval of 0 to 1. Considering the fact that some countries share the same kind of languages, we sum up other factors $x_{i,j,k}$ when their countries j share the same kind of languages-j and we get $z_{i,j,k}$, meaning the value of factor k with respect to language j in the year i.

Correspondingly , we do the same reduction with respect to the $Y_{i,j}$, namely dividing it by the corresponding sum, thus getting $y_{i,j}$, meaning the value of the number of speakers with respect to language j in the year i.

B. Detailed algorithm and formula

Using linear regression with machine learning algorithm, we denote $a_{i,j}$ as the coefficient of factor j with respect to language i, and c_i as the constant in the linear regression of language i. And show them in a formula as what follows:

$$y_{k,i} = \sum_j a_{i,j} \times z_{k,i,j} + c_i$$

We collect ten groups of data in the years between 2007 to 2016, meaning that k equals 0 to 9 and j equals 0 to 4(because we have 5 kinds of factors) to train the model in Python with scikit-learn and get the results for $a_{i,j}$ and c_i.

5.3 MA model to predict the trend of numbers of speakers with each language

5.3.1 Linear regression to predict the value of learning rate called $y_{i,j}$

As the model shown in the part 5.2, in order to predict the number of people speaking each language with the cellular automation, besides the present condition, we should also predict the probability of their learning rate $y_{i,j}$, thus needing to predict $z_{i,j,k}$.

After evaluating the trend of the value of $z_{i,j,k}$, linear regression will just suitable to predict with much convenience and precise. Using Python with linear regression, with the data $z_{i,j,k}$ for k from 0 to 9, we can get the fitting linear function and predict $z_{i,j,k}$ for k meaning the next 50 years (detailed coding for the algorithm will be shown in the appendix for coding). Then we still use the formula:

$$y_{i,k} = \sum_j a_{i,j} \times z_{i,j,k} + c_i$$

With the coefficient and constant, namely $a_{i,j}$ and c_i remaining invariable, we can predict $y_{i,k}$ meaning the learning rate with respect to language k in the year i.

5.3.2 Data pre-processing

Before constructing our model, we need to collect and wash our required data. In order to predict, we require not only the present data but also the predicted data for the next fifty years. So we have to deal with the database first.

We divide age into four groups:

a. $[0,10)$, which stands for preschool children
b. [10,20) , which stands for teenagers
c. [20,50) , which stands for young adults
d. [50,100] , which stands for the old

Because of lack of detailed data, we simplify our model by assuming that all the parents are between age of twenty and fifty, in other words, the birth rate $b_{i,j}$ satisfies:

$$b_{i,j} = 0, \forall i, j \quad \text{if one's age doesn't lie in [20,50)}$$

When predicting future data, take birth rate for instance. We extract the birth rate of all countries in last ten years from World Bank Database and we need to predict the birth rate of next fifty years. At first, we just predict the data by linear regression, but the effect is not satisfying because it doesn't take social and political effects into account.

Then we predict the data by the following method: the birth rate of the next year is the simple average of the birth rate of all the years before, i.e.

$$b_{i+2017} = \frac{\sum_{2007 \le j < i+2017} b_j}{i+10}$$

Where b_i denotes the birth rate of year i.

We find that in this way the birth rate converge in a fast rate and is more reliable.

5.3.3 Multi-agenda model to predict the number of people speaking each language

A Multi-agenda model is adopted to simulating the dynamic process of distribution of various language speakers. In this model, a single unit represents 10,000 people sharing roughly the same characters, such as age, nationality and language, which is a tradeoff between computational complexity and model accuracy. The time interval in the iteration is one year. In each step of the iteration, a unit can face four different situations:

1. it will die
2. it will give birth to a new unit
3. it will learn a second language if it hasn't master one yet
4. nothing will happen

In order to improve the accuracy of this model, various factors are considered to calculate the

possibility of different situations, including date, age, first language, international business relations and so on. The possibility of birth and death is equal to the corresponding birth rate and death rate and the possibility of learning a second language is the transfer-factor, whose details will be illustrated in the next part.

The result of 50 iterations reveals the prediction of the next 50 years. This process is repeated several times to ensure the validation of the result.

5.4 To determine the numbers and locations of international offices

5.4.1 Establishment of the model

As we have finished in the previous sections, we determined the competitive power of all the fifteen languages, which is one major factor to be considered into account when the client company choose the place to open new international offices.

Now we start to consider other two major factors : the competitive power of a city and relative geographical locations of different cities. The final goal of our plan is to maximize the total profit of opening new offices. After obtaining the total profit of opening offices in different cities, we can rank the cities and take their fifty year's average rank as a final rank.

5.4.2 Parameters and data of the model

From the US studies, the design of our total profit of country i in year j : $I_{i,j}$ is generally based on four factors:

(1) The actual probability to learn language i in year j : $y_{i,j}$

(2) The competitive power of country i in the year j : $p_{i,j}$

(3) The price index of country i : w_i

(4) The total distance between city i and other cities we choose : $\sum_{j \neq i} d_{i,j}$

(5) The international rate of city i : v_i

Where all parameters $y_{i,j}$, $p_{i,j}$, d_{j_1,j_2} will be converted into a number between 0 and 1.

We notice that has been calculated before and w_i and $d_{i,j}$ can be obtained from World Bank Database, so it suffices for us to calculate $p_{i,j}$.

We draw on the experience of HDI (Human Development Index) to get the formula of $p_{i,j}$, which calculates the geometric mean of all factors. We need three factors altogether:

A. Net Immigration Flow : $x_{i,j,0}$

B. Per Capita GDP : $x_{i,j,2}$

C. Total public expenditure on education: $x_{i,j,3}$

Then we get $p_{i,j}$ according to the following formula:

$$p_{i,j} = \sqrt[3]{x_{i,j,0} \times x_{i,j,2} \times x_{i,j,3}}$$

We denote the final profit of plan x as I_x, the numbers of new offices as $n\ (2 \le n \le 6)$

$$I_{xj} = \sum_{i=1}^{n} (5 \times y_{ij} + 5 \times p_{ij} - w_i + 5 \times v_i) - \frac{\sum_{j \neq j} d_{jj}}{n}$$

The last one of the formula represents the average distances between the cities we choose and we hope the number to be properly bigger so that the distribution of the offices can be wider. Because the company wants to be more international and gains more profits, we give a heavier weight to $p_{i,j}$, $y_{i,j}$ and v_i.

5.4.3 Implement and detail for the model

Table 5.4.3.1

Table 5.4.3.1 is partly the cost of price index of part of countries on Earth. New York ranks the first whose value is 100 so we divide the index by 100 for each country that we can obtain a number between 0 and 1.

When dealing with data of distances , we divide all the distances by the maximum one so that we can obtain a number between 0 and 1. i.e. $\hat{d}_{i,j} = \frac{d_{i,j}}{\max_i (d_{i,j})}$

Then we calculate the index of each city based on the model 2 and model 3, namely Beijing, Hongkong, Moscow, Berlin, Bombay, Chicago, Tokyo, Paris, Abu Dhabi, Madrid and so on, to find out the best plan for the number of the cities and the locations with the maximum of I_x.

6.Model Results

6.1 Model 1

We will see the results in the table 6.3.1 in the results of model 3.

6.2 Model 2

Table 6.2.1 shows the result of linear regression.

	coefficient0	coefficient1	coefficient2	coefficient3	coefficient4	shift
0	7.17645E-05	-0.008314868	0.149661505	-0.179492013	-0.121204708	0.0823028
1	-3.2946E-05	0.004859621	0.031161045	0.071846918	-0.071884038	0.011279129
2	-7.46099E-05	0.001459027	0.033535879	0.317148214	-0.082189975	0.016623044
3	-3.09362E-07	-0.008807816	-0.04696279	-0.071692383	-0.042982283	0.043681011
4	-2.73057E-05	-0.002758631	-1.086662701	0.048885612	-0.02367377	0.227798336
5	4.75329E-05	0.026390646	1.468202575	0.256981464	0.005604349	0.187048744
6	0.000131081	0.073371464	0.004637407	0.008801017	-0.006167652	0.004070425
7	-7.85657E-07	0.043366979	0.097891654	0.026780233	0.044152322	0.023931375
8	-0.000367548	0.0025119	-0.038991164	0.034642876	0.112396289	0.037842695
9	0.000436761	0.038372884	-0.008771676	0.225208364	-0.095294044	0.016453111
10	-2.43677E-05	0.000216636	0.008682135	0.002519827	-0.00562582	0.059370355
11	-5.27418E-05	-0.000600807	-0.032348244	0.182049862	-0.433064559	0.389734093
12	2.57671E-06	0.002387262	-0.009076655	-0.000690242	0.018710237	0.01208911

Table 6.2.1

This is a matrix of the coefficients of model 2.

The element (i, j) of the first five columns of the matrix is the coefficient with respect to the language i and the factor j, which is $a_{i,j}$.

The last column stands for the shift with respect to the language i and the factor j. which is c_j.

6.3 Model 3

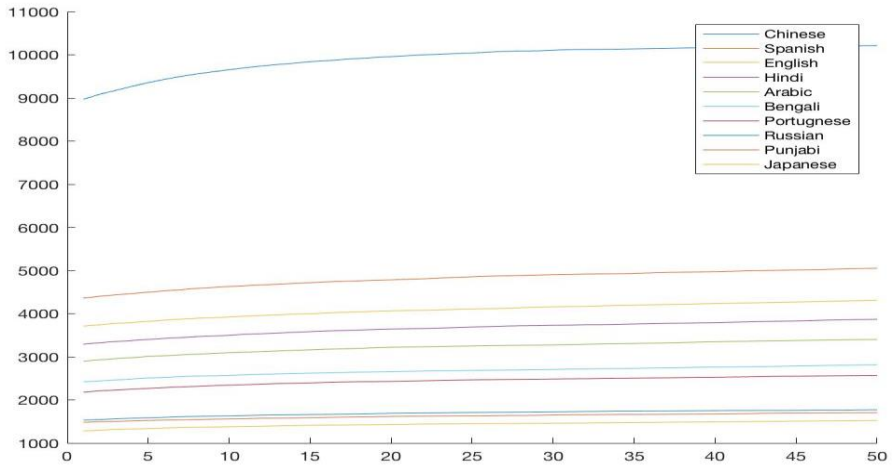
We obtain the prediction for the next 50 years. What follows is the part results of all . The number in the left refers to some language with the meaning showed in the Table 2.1. And 1-st means the number of native language; 2-nd means the number of second language. Total means the total number of that language. The value of 1 in the table means 100 thousand people.

years from 2018 to 2020	1st	2nd	Total
0	8970	1851	10821
1	4360	971	5331
2	3710	6090	9800
3	3290	2127	5417

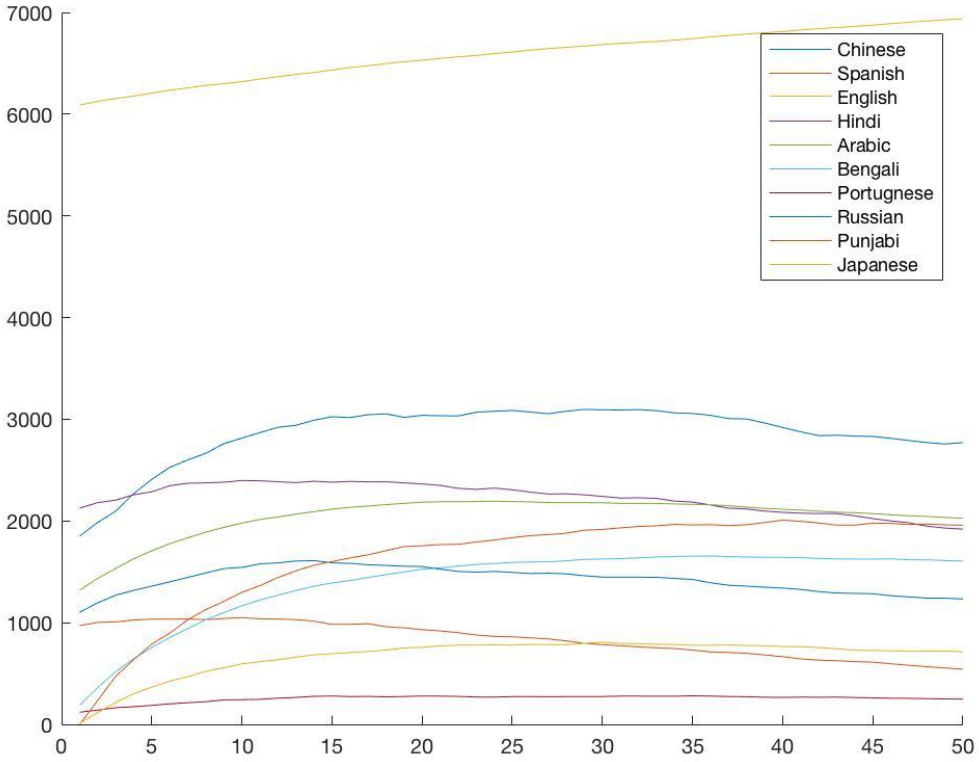
4	2900	1320	4220
5	2420	188	2608
6	2180	120	2300
7	1530	1101	2631
8	1480	0	1480
9	1280	10	1290
years from 2021 to 2023			
0	7686	2070	9756
1	3901	1035	4936
2	3341	5502	8843
3	2916	2287	5203
4	2539	3250	5789
5	2129	3031	5160
6	1888	187	2075
7	1346	1359	2705
8	1335	789	2124
years from 2024 to 2026			
0	6849	2186	6849
1	3539	1041	4580
2	3060	5044	8104
3	2657	2381	5038
4	2304	4404	6708
5	1951	4739	6690
6	1681	241	1922
7	1212	1532	2744
8	1235	1209	2444
9	978	557	1535
years from 2027 to 2029			
0	6285	2216	8501
1	3284	1030	4314
2	2838	4588	7426
3	2425	2377	4802
4	2121	5058	7179
5	1804	5833	7637
6	1518	265	1783
7	1128	1606	2734
8	1153	1510	2663
9	886	661	1547

Table 6.3.1

And we get the pictures about the number of native speakers and total language speakers as what follows:



Picture 6.3.1



Picture 6.3.2

6.4 Model 4

In the short term (10 years), the values of I_x of London,Beijing, Hongkong, Moscow, Berlin, Bombay, Chicago, Tokyo, Paris, Abu Dhabi and Madrid are respectively : 15.67882134,6.2096511, 10.18830237, 6.51647223, 9.64447677, 4.30845307, 7.39078022, 15.57887436, 12.23504, 8.68701994, 6.1075511.

So we choose six cities with the highest values: London, Tokyo, Paris, Hongkong, Berlin and Chicago.(The next two are Abu Dhabi and Bombay)

But considering the factor about the location, we will choose Bombay and Abu Dhabi instead of Berlin and Chicago because Berlin is near Paris and London and Chicago is near New York.

And considering the fact that the company has already chooses Shanghai in China, we take out Hongkong. So we can have less than six cities for the plan: London, Tokyo, Paris, Abu Dhabi and Bombay.

In the long term (50 years), with the same method, we choose six cities with the highest values of the plan and the distance factor: London, Tokyo, Paris, Berlin, Abu Dhabi and Bombay.

7.Sensitivity Analysis

7.1 Sensitivity Analysis of model 2

In model 2, we used the following formula to obtain the weight of each factor:

$$y_{k,i} = \sum_j a_{i,j} \times z_{k,i,j} + c_i$$

We consider $z_{k,i,j}^n (n > 1)$ instead of $z_{k,i,j}$, i.e. consider Polynomial fitting instead of linear fitting, then we have the formula:

$$y_{k,i} = \sum_j \hat{a}_{i,j} \times z_{k,i,j}^n + \hat{c}_i$$

Where $\hat{a}_{i,j}$ and and $\hat{c}_i$ denotes new coefficients and shifts.

For fear of overfitting, we choose $n \le 5$.

And we get the conclusion shown in table 7.1

Table 7.1

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
$n_{\max} = 1$	Total public expenditure on education	Per Capita GDP	Net Foreign Direct Investment Inflow	Net Immigration Flow	International Tourists Arrival Flow
$n_{\max} = 2$	Total public expenditure on education	Per Capita GDP	Net Immigration Flow	Net Foreign Direct Investment Inflow	International Tourists Arrival Flow
$n_{\max} = 3$	Per Capita GDP	Total public expenditure on education	Net Immigration Flow	Net Foreign Direct Investment	International Tourists Arrival Flow

				Inflow	
$n_{\max}=4$	Per Capita GDP	Total public expenditure on education	Net Immigration Flow	Net Foreign Direct Investment Inflow	International Tourists Arrival Flow
$n_{\max}=5$	Per Capita GDP	Net Immigration Flow	Total public expenditure on education	Net Foreign Direct Investment Inflow	International Tourists Arrival Flow

In table 7.1, $n_{\max}$ denotes the highest degree of Polynomial fitting

From table 7.1, we can see clearly that

A. Tourism always ranks last despite the change of n

B. Total public expenditure on education and Per Capita GDP always have strong dominance among the five factors

C. With the increase of n, the rank of Net Immigration Flow increases, which implies the potential problem of overfitting if n is too large, so we think the range: $n \le 5$ is reasonable.

7.2 Sensitivity Analysis of model 4

In model 4, we calculate $I_{x,j}$ with the following formula:

$$I_{x,j}=\sum_{i=1}^{n}(y_{i,j}+5\times p_{i,j}-w_i)-\frac{\sum_{j_1\neq j_2} d_{j_1,j_2}}{n}$$

Now we try to change the coefficient k before $p_{i,j}$ and set $k=1,2,...,10$ respectively.

And we find that when $k\le 4$: London, Paris ,Tokyo and Chicago ranks higher than other cities.

When $k\ge 6$: Tokyo, Bombay and Chicago ranks higher than other cities.

8. Further Discussion

8.1 Discussion for model 1

As for model 1, we should know exactly the difference between the population of one country and the number of people speaking corresponding native languages.

Besides, we should use other precise model to predict the population trend instead.

8.2 Discussion for model 2

As for model 2, we should use some more precise function model to fit the factors and the ratio of people speaking each languages. And we should consider more factors to precisely predict the results.

8.3 Discussion for model 3

As for the cellular automation algorithm, we should give more factors that will impact the probability for one person to learn the languages. And more training date should be collected to make the results more correct.

8.4 Discussion for model 4

We should use more factors to actually show the value of I_x. And as for the factor about distances between cities, we should use quantified data instead to see them in the map to determine the distances.

9. Strengths and Weaknesses

9.1 Strengths

9.1.1 Model 1

Considering the fact that it is very hard to define which language is a native language with respect to one person, even for UNESCO, we assume, accordingly, that a native language is an official language that one country claims to be. Therefore, to handle that problem, we choose to use the size of population standing for the number of people speaking corresponding native language and sum them up to get the results. And there are some obvious advantages:

A. It is very convenient to obtain the data of the prediction for population.
B. It is reasonable as we evaluate the ratio of the number of native speakers with respect to each language and the ratio of corresponding countries population.
C. We have reduced people who immigrate to other country with different native language and get the nationality.

9.1.2 Model 2

In the model 2, we use linear regression to fit the ratio of people speaking each language with some preeminent factors and there are some obvious advantages:

A. It is convenient and reasonable to regard the ratio of people speaking each language as the probability of one person to learn the corresponding language, which also serves as the learning rate used in the model 3 with cellular automation.
B. It is of much precise to just use as many as five factors amongst so many factors, for the reason that after PCA (principal component analysis) method we choose the factors attributing 85% of the model so that we can not only make a simple method but not cost so much accuracy as well.
C. After using the polynomial method to fit the model, we find, using Lasso method, that almost all the coefficients of different degrees of polynomial approximately equal to zero and it is great to choose linear functions to fit the model with much precise and convenience.
D. We use some reduction methods such as dividing by the sum to make the value lie in the interval of 0 to 1, to make it very convenient to compute and it is with good probability meaning.

9.1.3 Model 3

As for the prediction of $z_{i,j,k}$, we use linear regression to handle it with some obvious strengths:

A. It is very convenient to use linear regression to predict the trend and it is just in accordance with the linear regression used in the model 2.
B. It is considerably reasonable to choose linear function to fit and predict $z_{i,j,k}$, with the evidence that from the historical data we can get, those data do not change rapidly.
C. We have done some reduction to make sure that value lies in the in the interval of 0 to 1, so that it has some probability meaning.

As for the prediction of $y_{i,j}$, we calculate it by using the linear regression model in model 2 and regard it as the learning rate to predict the probability to learn other languages. There are several advantages:

A. It is in good accordance with the model 2 to use linear regression and convenient to compute.
B. Form the meaning of $y_{i,j}$, namely the ratio of people speaking each language, it is reasonable to regard it as the probability to learn other languages, which is a good reduction from abstract conception to mathematical conception with data to analysis and predict.

As for the cellular automation, we capture the preeminent factors that will determine the probability for one person to learn each language. And by using the birth rate and death rate, we establish a simple and convenient model for the computer to run.

9.1.4 Model 4

In the model 4, we start to consider other two major factors : the competitive power of a city and relative geographical locations of different cities. There are some strengths:

A. We introduce a multi-agenda system to evaluate each city we are interested in so as to be comprehensive.
B. Our goal is to choose the location with the aims to be international and gain more profits. So we use some main factors:

   International rate;
   Comprehensive scores;
   geographical factors to avoid wasting;

   So we can not only try to be located in the more international city and gain more profits with less waste.
C. One specific factor we have taken into consideration is the cost of living for each city. Because big cities will also tend to have high cost of living.
D. We use linear regression to make that model in order to make it a simple and convenient way to take into practice without so much inaccuracy.

9.2 Weaknesses

9.2.1 Model 1

Considering the fact that it is very hard to define which language is a native language with respect to one person, even for UNESCO, we assume, accordingly, that a native language is an official language that one country claims to be. Therefore, to handle that problem, we choose to use the

size of population standing for the number of people speaking corresponding native language and sum them up to get the results. Therefore, the weaknesses are:

A. The population will account some people who do not speak or do not choose their official language as native language.
B. The population will miss people who speak their native language but lose their nationality for some reason.
C. The survey of population is sometimes done in every several years, so it is not very suitable to show the number every year.

9.2.2 Model 2

In the model 2, we use linear regression to fit the ratio of people speaking each language with some preeminent factors. But there exist some weaknesses:

A. The rest of the factors, although not consisting so much, still matter when solving this problem.
B. Linear regression has some kind of error.
C. During the process of some reduction such as dividing by the sum, we have assumed each factor is of equal importance.

9.2.3 Model 3

As for the prediction of $z_{i,j,k}$, we use linear regression to handle it. As for the prediction of $y_{i,j}$, we calculate it by using the linear regression model in model 2 and regard it as the learning rate to predict the probability to learn other languages. So the weaknesses are:

A. Linear regression has some kind of error.
B. We assume the coefficients of the factors remain invariable , which will definitely change with time.
C. $y_{i,j}$ means the ratio of people speaking each language and we use it as the probability to learn other languages. So the error is that the whole ratio cannot exactly represent the probability of each person to learn other language.

As for the cellular automation model , we ignore some other factors which are not so important based on the model 3, and because of the weaknesses listed about model 2, all those weaknesses will contribute to the inaccuracy of the model 3. And we should try to ignore some minor features to make the code run faster.

9.2.4 Model 4

In the process of making model 4, we use some main factors:

International rate;
Comprehensive scores;
geographical factors to avoid wasting;

But there exist some weaknesses:

A. There are other factors which will also contribute to the problem.
B. Linear regression has some kind of error.
C. Coefficients of the factors are determined by the present situation, although appearing to be not changing rapidly.

D. geographical factors are considered roughly, not the exact distances.

10. Conclusions

Part I.A

Based on the factors as what follows:

0.Net Immigration Flow

1.Net Foreign Direct Investment Inflow

2. Per Capita GDP

3. Total public expenditure on education

4. International Tourists Arrival Flow

We call them respectively as $x_{i,j,k}$ meaning the value of the factor k with respect to the language j in the year i. And $y_{i,j}$ means the ratio of people speaking language j in the year i amongst all the people.

Then the model of distribution is what follows:

$$y_{k,i} = \sum_{j} a_{i,j} \times z_{k,i,j} + c_i$$

With regard to the exact result of regression, please refer to table 6.2.1 .

We can draw the conclusion that among the five factors, Total public expenditure on education ranks first and Per Capita GDP second, which implies that education and economy are two most important indicators for the country’s language competency and it conforms to our common sense. Besides, International Tourists Arrival Flow ranks last, which suggests that tourism doesn’t have an significant influence on language relatively and it’s also understandable because of the development of technology, an increasing number of people rely on translation software so they don’t have to learn another language by themselves.

Part I.B

From the pictures, we can predict the trend of the number of native speakers is that all the number of native speakers will increase with different speed and the rank will not change in 50 years. How ever, when total numbers are taken into consideration, sharp change can be seen in our prediction: First, Chinese will become the second largest second language in the world in place of Hindustani; Second, Punjabi and Japanese will stand out and surpass their opponents.

Part I.C

According to the prediction of population and the number of people in each country and speaking each language, we draw the conclusion that there is no obvious change over the same period of time.

Part II

In the short term (10 years), we choose six cities with the highest values of the plan and the distance factor: London, Tokyo, Paris, Hongkong, Abu Dhabi and Bombay.

And considering the fact that the company has already chooses Shanghai in China, we take out Hongkong. So we can have less than six cities for the plan: London, Tokyo, Paris, Abu Dhabi and

Bombay.

In the long term (50 years), we choose six cities with the highest values of the plan and the distance factor: London, Tokyo, Paris, Berlin, Abu Dhabi and Bombay.

Reference

[1] Lane, J. (2017). The 10 Most Spoken Languages in the World. Babbel Magazine. Retrieved from https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world
[2] List of Languages by Total Numbers of Speakers
https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers
[3] data of population, birth rate, death rate, Net Immigration Flow, Net Foreign Direct Investment Inflow, Per Capita GDP, Total public expenditure on education, and International Tourists Arrival Flow from
https://data.worldbank.org/
[4] "Summary by language size". Ethnologue. Retrieved 2016-04-06.
[5] "Världens 100 största språk 2010" (The World's 100 Largest Languages in 2010), in Nationalencyklopedin
[6] 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
[7] ”Estimating global migration flow tables using place of birth data” by Guy J.Abel
Data of index and population from http://www.renkou.org.cn/
[8] trading data from https://zh.tradingeconomics.com

Dear sir:

The purpose of the letter is to inform you of our team's results and to advice you on the design of the location of new international offices and its corresponding languages. We have analyzed the popularity and competency of the fifteen most spoken languages including Mandarin, English, Hindustani, Spanish, Arabic, Malay, Russian, Bengali, Portuguese, French, Hausa, Punjabi, Japanese, German and Persian. We choose six main factors and quantify them to determine the concrete weight of each factor. Then we make a prediction of their future trend and distribution based on predicted data of birth rate, death rate and other basic factors. When it comes to choosing new locations of international offices, besides the language effect we've mentioned above, we also considered three other overwhelming factors: the competency of the city itself, the cost of living index of the city and the relative geographical location of the city with other cities.

The first is the determination of the development and future trend of different languages. We select five most official and typical indicators as major contributing factors to language: Net Immigration Flow, Net Foreign Direct Investment Inflow, Per Capita GDP, Total public expenditure on education, and International Tourists Arrival Flow, which include economic, cultural, educational, political effects in the round to make the analysis more comprehensive and scientific. Besides, we also calculate the correlation of different languages and we assume that if two languages belong to the same language family, the native speaker of one language is more prone to learn the other as a second language or third language. For each language, we regard the number of the total speaker of that language as the actual popularity of that language. After processing data, we draw the conclusion that among the five factors, Total public expenditure on education ranks first and Per Capita GDP second, which implies that education and economy are two most important indicators for the country's language competency and it conforms to our common sense. Besides, International Tourists Arrival Flow ranks last, which suggests that tourism doesn't have an significant influence on language relatively and it's also understandable because of the development of technology, an increasing number of people rely on translation software so they don't have to learn another language by themselves.

The second is to determine the best location of new offices and their corresponding languages. The final goal of our plan is to maximize the profit so we construct a model to quantify the total benefit of setting a new office in certain city. We assume that once a country is chosen, we select the capital or the most prosperous city in the country automatically. And we hope to lessen the probability that two offices lie in the same country or too close so that the distribution of new offices can be wider. Simultaneously, we select the living price index as the indicator of the cost of opening a new office at a certain area. From a short-term perspective, we choose the average rank of ten years and come to the conclusion that the following six cities are the most optimal choice:

1.Tokyo 2. Paris 3.Hongkong 4.Berlin 5. Chicago 6.Abu Dhabi .(The next is Bombay)

From a long-term perspective, we choose the average rank of fifty years and come to the conclusion that the following six cities are the most optimal choice:

1.London 2.Tokyo 3.Paris 4. Berlin 5. Abu Dhabi 6. Bombay.

The third is to analyze whether the number of six is optimal and should the number of new offices to be fewer or not. When considering the factor of the location, we choose Bombay instead of Berlin because Berlin is near Paris and London. And considering the fact that the company has already chosen Shanghai in China and New York in USA, we take out two cities: Hongkong and Chicago. So we can have less than six cities for the plan: London, Tokyo, Paris, Abu Dhabi and

Bombay.

In a nutshell, we suggest that you should open five new offices and set them in London, Tokyo, Paris, Abu Dhabi and Bombay respectively, which combines profit and relative geological relations.

Hope that our model can be helpful to you!

Sincerely

MCM Team Members